

1

Variables and their measurement

This book aims to help people analyze quantitative information. Before detailing the ‘hands-on’ analysis we will explore in later chapters, this introductory chapter will discuss some of the background conceptual issues that are precursors to statistical analysis. The chapter begins where most research in fact begins; with **research questions**.

A **research question** states the aim of a research project in terms of **cases** of interest and the **variables** upon which these cases are thought to differ.

A few examples of research questions are:

‘What is the age distribution of the students in my statistics class?’

‘Is there a relationship between the health status of my statistics students and their sex?’

‘Is any relationship between the health status and the sex of students in my statistics class affected by the age of the students?’

We begin with very clear, precisely stated research questions such as these that will guide the way we conduct research and ensure that we do not end up with a jumble of information that does not create any real knowledge. We need a clear research question (or questions) in mind before undertaking statistical analysis to avoid the situation where huge amounts of data are gathered unnecessarily, and which do not lead to any meaningful results. I suspect that a great deal of the confusion associated with statistical analysis actually arises from imprecision in the research questions that are meant to guide it. It is very difficult to select the *relevant* type of analysis to undertake, given the many *possible* analyses we could employ on a given set of data, if we are uncertain of our objectives. If we don’t know why we are undertaking research in the first place, then it follows we will not know what to do with research data once we have gathered them. Conversely, if we are clear about the research question(s) we are addressing the statistical techniques to apply follow almost as a matter of course.

We can see that each of the research questions above identifies the entities that I wish to investigate. In each question these entities are students in my statistics class, who are thus the **units of analysis** – the **cases** of interest – to my study.

A **case** is an entity that displays or possesses the traits of a variable.

In this example, as in many others, the cases are individual people. It is important to bear in mind, however, that this is not always so. For example, if I am interested in retention rates for high schools in a particular area, the cases will be high schools. It is individual high schools that are ‘stamped’ with a label indicating their respective retention rate.

In the research questions listed above, *all* the students in my statistics class constitute my target **population** (sometimes called a **universe**).

A **population** is the set of all possible cases of interest.

In determining our population of interest, we usually specify the *point in time* that defines the population – am I interested in my currently enrolled statistics students, or those who completed my course last year as well? We also specify, where relevant, the *geographic region* over which the population spreads.

For reasons we will investigate later, we may not be able to, or not want to, investigate the entire population of interest. Instead we may select only a sub-set of the population, and this sub-set is called a **sample**.

A **sample** is a set of cases that does not include every member of the population.

For example, it may be too costly or time consuming to include every student in my study. I may instead choose only those students in my statistics class whose last name begins with 'A', and thus be only working with a sample.

Suppose that I do take this sample of students from my statistics class. I will observe that these students differ from each other in many ways: they may differ in terms of sex, height, age, attitude towards statistics, religious affiliation, health status, etc. In fact, there are many ways in which the cases in my study may differ from each other, and each of these possible expressions of difference is a **variable**.

A **variable** is a condition or quality that can differ from one case to another.

The opposite notion to a variable is a **constant**, which is simply a condition or quality that *does not* vary between cases. The number of cents in a United States dollar is a constant: every dollar note will always exchange for 100 cents. Most research, however, is devoted to understanding variables – whether (and why) a variable takes on certain traits for some cases and different traits for other cases.

The conceptualization and operationalization of variables

Where do variables come from? Why do we choose to study particular variables and not others? The choice of variables to investigate is affected by a number of complex factors, three of which I will emphasize here.

1. *Theoretical framework*. Theories are ways of interpreting the world and reconciling ourselves to it, and even though we may take for granted that a variable is worthy of research, it is in fact often a highly charged selection process that directs one's attention to it. We may be working within an established theoretical tradition that considers certain variables to be central to its world-view. For example, Marxists consider 'economic class' to be a variable worthy of research, whereas another theoretical perspective might consider this variable to be uninteresting. Analyzing the world in terms of economic class means not analyzing it in other ways, such as social groups. This is neither good nor bad: without a theory to order our perception of the world, research will often become a jumble of observations that do not tie together in a meaningful way. We should, though, acknowledge the theoretical preconceptions upon which our choice of variables is based.
2. *Pre-specified research agenda*. Sometimes the research question and the variables to be investigated are not determined by the researchers themselves. For example, a consultant may contract to undertake research that has terms of reference set in advance by the contracting body. In such a situation the person or people actually doing the research might have little scope to choose the variables to be investigated and how they are to be defined, since they are doing work for someone else.
3. *Curiosity-driven research*. Sometimes we might not have a clearly defined theoretical framework to operate in, nor clear directives from another person or body as to the key concepts to be investigated. Instead we want to investigate a variable purely on the basis of a hunch, a loosely conceived feeling that something useful or important might be revealed if we study a particular variable. This can be as important a reason for undertaking research as theoretical imperatives. Indeed, when moving into a whole new area of research, into which existing theories have not ventured, simple hunches can be fruitful motivations.

These three motivations are obviously not mutually exclusive. For example, even if the research agenda is specified by another person, that person will almost certainly be operating within some theoretical framework. Whatever the motivation, though, social inquiry will initially direct us to particular variables to be investigated. At this initial stage a variable is given a **conceptual definition**.

The **conceptual definition** (or **nominal definition**) of a variable uses literal terms to specify the qualities of a variable.

A conceptual definition is much like a dictionary definition; it provides a working definition of the variable so that we have a general sense of what it ‘means’. For example, I might define ‘health’ conceptually as ‘an individual’s state of well-being’.

It is clear, though, that if I now instruct researchers to go out and measure people’s ‘state of well-being’, they would leave scratching their heads. The conceptual definition of a variable is only the beginning; we also need a set of rules and procedures – **operations** – that will allow us to actually ‘observe’ a variable for individual cases. What will we look for to identify someone’s health status? How will the researchers record how states of well-being vary from one person to the next? This is the problem of **operationalization**.

The **operational definition** of a variable specifies the procedures and criteria for taking a measurement of that variable for individual cases.

A statement such as ‘a student’s health status is measured by how far in meters they can walk without assistance in 15 minutes’ provides one operational definition of health status. With this definition in hand I can start measuring the health status of students in my statistics class by observing distance covered in the set time limit.

The determination of an operational definition for a variable is a major, if not *the* major, source of disagreement in research. Any variable can usually be operationalized in many different ways, and no one of these definitions may be perfect. For example, operationalizing health status by observing a student’s ability to complete a walking task leaves out an individual’s own subjective perception of how healthy they feel.

What criteria should be used in deciding whether a particular operational definition is adequate? In the technical literature this is known as the problem of **construct validity**. Ideally, we look for an operationalization that will vary when the underlying variable we think it ‘shadows’ varies. A mercury thermometer is a good instrument for measuring changes in daily temperature because when the underlying variable (temperature) changes the means of measuring it (the height of the bar of mercury) also changes. If the thermometer is instead full of water rather than mercury, variations in daily temperature will not be matched by changes in the thermometer reading. Two days might actually be different in temperature, without this variation being ‘picked up’ by the instrument. Coming back to our example of health status, and relying on an operational definition that just measures walking distance covered in a certain time, we might record two people as being equally healthy, when in fact they differ. Imagine two people who each walk 2200 meters in 15 minutes, but one of these people cannot bend over to tie their shoelace because of a bad back. Clearly there is variation between the two people in terms of their health – their state of well-being. But this variation will not be recorded if we rely solely on a measure of walking ability.

To illustrate further the ‘slippage’ that can occur in moving from a conceptual to an operational definition, consider the following example. A study is interested in people’s ‘criminality’. We may define criminality conceptually as ‘non-sanctioned acts of violence against other members of society or their property’. How can a researcher identify the pattern of variation in this variable? A number of operational definitions could be employed:

- counting a person’s number of criminal arrests from official records;
- calculating the amount of time a person has spent in jail;

- asking people whether they have committed crimes;
- recording a person's hair color.

Clearly, it would be very hard to justify the last operationalization as a valid one: it is not possible to say that if someone's level of criminality changed so too will their hair color! The other operational definitions seem closer to the general concept of criminality, but each has its own problems: asking people if they have committed a crime may not be a perfect measure because people might not be truthful about such a touchy subject. Counting the number of times a person has been arrested is not perfect – two people may actually have the same level of criminality, yet one might have more recorded arrests because they are a member of a minority group that the police target for arrest. This operationalization may thereby actually be measuring a different variable from the one intended: the biases of police rather than 'criminality'. Using any of these operational definitions to measure a person's criminality may not perfectly mirror the result we would get if we could 'know' their criminality.

A number of factors affect the extent to which we can arrive at an operational definition of a variable that has high construct validity.

1. *The complexity of the concept.* Some variables are not very complex: a person's sex, for example, is determined by generally accepted physical attributes. However, most variables are rarely so straightforward. Health status, for example, has a number of **dimensions**. At a broad level we can differentiate between physical, mental, and emotional health; two people might be physically well, but one is an emotional wreck while the other is happy and contented. If we operationalize health status by looking solely at the *physical* dimension of its expression, important differences in this variable may not be observed. Indeed each of these broad dimensions of health status – physical, mental, and emotional – are conceptual variables in themselves, and raise problems of operationalization of their own. If we take physical health as our focus, we still need to think about all of its particular forms of expression, such as the ability to walk, carry weight, percentage of body fat, etc.
2. *Availability of data.* We might have an operationalization that seems to capture perfectly the underlying variable of interest. For example, we might think that number of arrests is a flawless way of 'observing' criminality. The researchers, though, may not be allowed, for privacy reasons, to review police records to compile the information. Clearly, a less than perfect operationalization will have to be employed, simply because we cannot get our hands on the 'ideal' data.
3. *Cost and difficulty of obtaining data.* Say we were able to review police records and tally up the number of arrests. The cost in doing so, though, might be prohibitive, in terms of both time and money. Similarly, we might feel that a certain measure of water pollution is ideal for assessing river degradation, but the need to employ an expert with sophisticated measuring equipment might bar this as an option, and instead a subjective judgment of water 'murkiness' might be preferred as a quick and easy measure.
4. *Ethics.* Is it right to go looking at the details of an individual's arrest record, simply to satisfy one's own research objectives? The police might permit it, and there might be plenty of time and money available, but does this justify looking at a document that was not intended to be part of a research project? The problem of ethics – knowing right from wrong – is extremely thorny, and I could not even begin to address it seriously here. It is simply raised as an issue affecting the operationalization of variables that regularly occurs in research dealing with the lives of people. (For those wishing to follow up on this important issue, a good starting point is R.S. Broadhead, 1984, Human rights and human subjects: Ethics and strategies in social science research, *Sociological Inquiry*, 54, pp. 107–23.)

For these (and other) reasons a great deal of debate about the validity of research centers around this problem of operationalization. In fact, many debates surrounding quantitative research are not actually about the methods of analysis or results of the research, but rather whether the variables have been ‘correctly’ defined and measured in the first place. Unless the operational criteria used to measure a variable are sensitive to the way the variable actually changes, they will generate misleading results.

Scales of measurement

We have, in the course of discussing the operationalization of variables, used the word ‘measurement’, since the purpose of deriving an operational definition is to allow us to take measurements of a variable.

Measurement (or observation) is the process of determining and recording which of the possible traits of a variable an individual case exhibits or possesses.

The variable ‘sex’ has two possible traits, female and male, and measurement involves deciding into which of these two categories a given person falls. This set of categories that can possibly be assigned to individual cases makes up a **scale of measurement**.

A **scale of measurement** specifies a range of scores (also called points on the scale) that can be assigned to cases during the process of measurement.

When we construct a scale of measurement we need to follow two particular rules. First, *the scale must capture sufficient variation to allow us to answer our research question(s)*. Take the research questions that I posed at the start of this chapter regarding my statistics students. I may use ‘number of whole years elapsed since birth’ as the operational definition for measuring the ages of students in my statistics class. This produces a scale with whole years as the **points** on the scale, and will yield a variety of scores for age, given the fact that my students were born in different years. Imagine, though, that I was teaching a class of primary school students; measuring age with this scale will be inadequate since most or all of the students will have been born in the same year. Using a measurement scale for age that only registers whole years will not pick up enough variation to help me meet my objectives; every student in the class will appear to be the same age. I might consider, instead, using ‘number of whole *months* elapsed since birth’ as the scale of measurement. Age *in months* will capture variation among students in a primary class that age *in years* will miss.

This example of the age for a group of students highlights an intrinsic problem when we try to set up a scale to measure sufficient variation for a **continuous** variable that does not arise when we try to measure a **discrete** variable.

A **discrete variable** has a finite number of values.

A **continuous variable** can vary in quantity by infinitesimally small degrees.

For example, the sex of students is a discrete variable with only two possible categories (male or female). Discrete variables often have a unit of measurement that cannot be subdivided, such as the number of children per household. Other examples of discrete variables are the number of prisoners per jail cell, the number of welfare agencies in a district, and the number of industrial accidents in a given year.

The age of students, on the other hand, is a continuous variable. Age can conceivably change in a *gradual* way from person to person or for the same person over time. Because of this, continuous variables are measured by units that can be infinitely subdivided. Age, for example, does not have a basic unit with which it is measured. We may begin by measuring age in terms of years. But a year can be divided into months, and months into weeks, weeks

into days, and so on. The only limit is exactly how precise we want to be: years capture less variation than months, and months less than weeks. *Theoretically*, with a continuous variable we can move gradually and smoothly from one value of the variable to the next without having to jump. *Practically*, though, we will always have to ‘round off’ the measurement and treat a continuous variable *as if* it is discrete, and this causes the scale of measurement to ‘jump’ from one point on the scale to the next. The *scale* is by necessity discrete, even though the underlying variable is continuous.

The use of a discrete measurement scale to measure age, whether we do it in years or months, causes us to cluster cases together into groups. The points on the scale act like centers of gravity pulling in all the slight variations around them that we do not want to worry about. We may say that two people are each 18 years old, but they will in fact be different in terms of age, unless they are born precisely at the same moment. But the slight difference that exists between someone whose age is 18 years, 2 months, 5 days, 2 hours, 12 seconds... and someone whose age is 18 years, 3 months, 14 days, 7 hours, 1 second... might be irrelevant for the research problem we are investigating and we treat them the same in terms of the variable, even though they are truly different. No measurement scale can ever hope to capture the full variation expressed by a continuous variable. The practical problem we face is whether the scale captures *enough* variation to help us answer our research question.

The second rule of measurement is that *a scale must allow us to assign each case into one, and only one, of the points on the scale*. This statement actually embodies two separate principles of measurement. The first is the **principle of exclusiveness**, which states that no case should have more than one value for the same variable. For example, someone cannot be both 18 years of age and 64 years of age. Measurement must also follow the **principle of exhaustiveness**, which states that every case can be classified into a category. A scale for health status that only had ‘healthy’ and ‘very healthy’ as the points on the scale is obviously insufficient; anyone who is less than healthy cannot be measured on this scale.

Levels of measurement

A scale of measurement allows us to collect **data** that give us information about the variable we are trying to measure.

Data are the measurements taken for a given variable for each case in a study.

Scales of measurement, however, do not provide the same amount of information about the variables they try to measure. In fact, we generally talk about measurement scales having one of four distinct **levels of measurement**: nominal, ordinal, interval, and ratio.

We speak of *levels* of measurement because *the higher the level of measurement the more information we have about a variable*. These levels of measurement are a fundamental distinction in statistics, since they determine much of what we can do with the data we gather. In fact, when considering which of the myriad of statistical techniques we can use to analyze data, usually the first question to ask is the level at which a variable has been measured. As we shall see there are things we can do with data collected at the interval level of measurement that we cannot do with data collected at the nominal level.

Nominal scales

The lowest level of measurement is a **nominal** scale.

A **nominal** scale of measurement classifies cases into categories that have no quantitative ordering.

For example, assume I am interested in people’s religion. Operationally I define a person’s religion as the established church to which they belong, providing the following range of categories: Muslim, Hindu, Jewish, Christian, Other.

Notice that to ensure the scale is exhaustive this nominal scale, like most nominal scales, has a catch-all category of 'Other'. Such a catch-all category, sometimes labelled 'miscellaneous' or 'not elsewhere counted', at the end of the scale often provides a quick way of identifying a nominal scale of measurement.

Another easy way to detect a nominal scale is to rearrange the order in which the categories are listed and see if the scale still 'makes sense'. For example, either of the following orders for listing religious denomination is valid:

Christian	Muslim	Jewish	Hindu	Other
or				
Muslim	Jewish	Hindu	Christian	Other

Obviously, the order in which the categories appear does not matter, provided the rules of exclusivity and exhaustiveness are followed. This is because there is no sense of rank or order of magnitude: one cannot say that a person in the 'Christian' category has more or less religion than someone in the 'Hindu' category. In other words, *a variable measured at the nominal level varies qualitatively but not quantitatively*: someone in the Christian category is qualitatively different to someone in the Hindu category, with respect to the variable 'Religion', but they do not have more or less Religion.

It is important to keep this in mind, because *for convenience* we can assign numbers to each category as a form of shorthand (a process that will be very useful when we later have to enter data into SPSS). Thus I may **code** – assign numbers to – the categories of religion in the following way:

1	2	3	4	5
Muslim	Jewish	Hindu	Christian	Other

These numbers, however, are simply category labels that have no quantitative meaning. The numbers simply identify different categories, but do not express a mathematical relationship between those categories. They are used for convenience to enter and analyze data. I could just as easily have used the following coding scheme to assign numerical values to each category:

1	5	6	8	9
Muslim	Jewish	Hindu	Christian	Other

Ordinal scales

An **ordinal** scale of measurement also categorizes cases. Thus nominal and ordinal scales are sometimes collectively called **categorical scales**. However, an ordinal scale provides additional information.

An **ordinal** scale of measurement, in addition to the function of classification, allows cases to be ordered by degree according to measurements of the variable.

Ordinal scales, that is, enable us to **rank** cases. Ranking involves ordering cases in a quantitative sense, such as from 'lowest' to 'highest', from 'less' to 'more', or from 'weakest' to 'strongest', and are particularly common when measuring attitude or satisfaction in opinion surveys. For example, assume that in trying to measure age I settle on the following scale:

<i>18 years or less</i>	<i>19 to 65 years</i>	<i>Over 65 years</i>
-------------------------	-----------------------	----------------------

This scale clearly does the task of a nominal scale, which is to assign cases into categories. In addition to this it also allows me to say that someone who is in the '19 to 65 years' category is *older* than someone in the '18 years or less' category. Put another way, the '19 to 65 year-

old' is **ranked above** someone who is '18 years or less'. Unlike nominal data, a case in one category is not only different to a case in another, it is 'higher', or 'stronger', or 'bigger', or more 'intense': *there is directional change*.

Unlike nominal scales, we *cannot rearrange the categories* without the ordinal scale becoming senseless. If I construct the scale in the following way, the orderly increase in age as we move across the page from left to right is lost:

19 to 65 years Over 65 years 18 years or less

As with nominal data, numerical values can be assigned to the points on the scale as a form of shorthand, but with ordinal scales *these numbers also need to preserve the sense of ranking*. Thus either of the following sets of numbers can be used:

1 = 18 years or less 23 = 18 years or less
2 = 19 to 65 years or 88 = 19 to 65 years
3 = Over 65 years 99 = Over 65 years

Either coding system allows the categories to be identified and ordered with respect to each other, but the numbers themselves do not have any quantitative significance beyond this function of ranking.

One common mistake in statistical analysis is to treat scales that allow either a 'No' or 'Yes' response as only nominal, when they are almost invariably ordinal. Consider a question that asks participants in a study 'Do you feel healthy?' We can say that someone who responds 'Yes' is not only different in their (perceived) health level, but they also have a higher health level than someone who responds 'No'. Practically any question that offers a Yes/No response option can be interpreted in this way as being an ordinal scale.

Interval/ratio scales

Ordinal scales permit us to rank cases in terms of a variable; we can, for example, say that one case is 'better' or 'stronger' than another. But an ordinal scale does not allow us to say *by how much* a case is better or stronger when compared with another. If I use the above age scale, I cannot say how much older or younger someone in one category is relative to someone in another category. It would be misleading for me to use the second of the coding schemes above and say that someone in the oldest group has 76 more units of age than someone in the youngest group (i.e. $99 - 23 = 76$). The distances – **intervals** – between the categories are unknown.

Consider, however, if we measure age in an alternative way, by asking each person how many whole years have elapsed between birth and their last birthday. Clearly, I can perform the task of assigning people into groups based on the number of years. I can also perform the task of rank ordering cases according to these measurements by indicating who is older or younger. Unlike nominal and ordinal scales, however, I can *also measure the amount difference* in age between cases. In this measurement scale the numbers we get do really signify a quantitative value: number of years. It is this ability to measure the distances between points on the scale that makes this method of observing age an **interval/ratio** scale.

An interval scale has units measuring intervals of equal distance between values on the scale.

A ratio scale has a value of zero indicating cases where no quantity of the variable is present.

In other words, not only can we say that one case has more (or less) of the variable in question than another, but we can also say how much more (or less). Thus someone who is 25 years old has 7 years more age than someone who is 18 years old; we can measure the interval between them. Moreover, *the intervals between points on the scale are of equal value over its whole range*, so that the difference in age between 18 and 25 years is the same as the difference in age between 65 and 72 years.

Clearly the numbers on an interval scale do have quantitative significance. Hence these numbers are termed the **values** for the variable. (In the following chapters we will also refer to the numbers used to represent the categories of nominal and ordinal data as ‘values’ or ‘scores’, so that the terms ‘values’, ‘scores’ and ‘categories’ are used interchangeably. For the reasons we have just outlined this is, strictly speaking, incorrect. However, if we take note that for nominal and ordinal data such values are simply category labels without real quantitative significance, such terminology is not too misleading.)

Notice that an observation of 0 years represents a case which possesses no quantity of the variable ‘age’. Such a condition is known as a **true zero point** and is the defining characteristic of a ratio scale, as opposed to an interval scale. For example, heat measured in degrees Celsius does not have a ‘true’ zero. There is a zero point, but 0°C does not indicate a case where no heat is present – it is cold but not that cold! Instead, 0°C indicates something else: the point at which water freezes. However, this fine distinction between interval and ratio scales of measurement is not important for what is to follow. We can generally perform the same statistical analyses on data collected on an interval scale that we can on data collected on a ratio scale, and thus we speak of one interval/ratio level of measurement.

The importance of the distinction between nominal, ordinal, and interval/ratio scales is the amount of information about a variable that each level provides (Table 1.1).

Table 1.1 Levels of measurement

Level of measurement	Examples	Measurement procedure	Operations permitted
Nominal (lowest level)	Sex Race Religion Marital status	Classification into categories	Counting number of cases in each category; comparing number of cases in each category
Ordinal	Social class Attitude and opinion scales	Classification plus ranking of categories with respect to each other	All above plus judgments of ‘greater than’ or ‘less than’
Interval/ratio (highest level)	Age in years Pulse rate	Classification plus ranking plus description of distances between scores in terms of equal units	All above plus other mathematical operations such as addition, subtraction, multiplication, etc.

Source: J.F. Healey, 1993, *Statistics: A Tool for Social Research*, Belmont, CA: Wadsworth, p. 14.

Table 1.1 summarizes the amount of information provided by each level of measurement and the tasks we are thereby allowed to perform with data collected at each level. Nominal data have the least information, ordinal data give more information because we can rank cases, and interval/ratio data capture the most information since they allow us to measure difference.

Before concluding this discussion of levels of measurement there are two important points to bear in mind. The first is that *any given variable can be measured at different levels*, depending on its operational definition. We have seen, for example, that we can measure age in whole years (interval/ratio), but we can also measure age in broad groupings (ordinal). Conversely, *a specific scale can provide different levels of measurement depending on the particular variable we believe it is measuring*; it can be, to some degree, a matter of interpretation. For example, we may have a scale of job types broken down into clerical, supervisory, and management. If we interpret this scale as simply signifying different jobs, then it is measuring job *classification* and is nominal. If we see this scale as measuring job *status*, however, then we can hierarchically order these categories into an ordinal scale.

Univariate, bivariate, and multivariate analysis

We have just spent some time discussing the notion of levels of measurement, since the nature of the scales we use to measure a variable affects the kinds of statistical analysis we can perform (as we will see in later chapters). The other major factor involved in determining the analysis we perform is the *number* of variables we want to analyze. Take, for example, the first research question listed at the start of this chapter, which asks ‘What is the age

distribution of the students in my statistics class?’ This question is only interested in the way that my students may differ in terms of age; age is the only variable of interest to this question. Since it analyzes differences among cases for only *one* variable, such a question leads to **univariate statistical analysis**.

The next two questions are more complex; they are not interested in the way in which students vary in terms of age alone. The second links differences in age with health status, and the third throws the sex of students into the mix. A question that addresses the possible relationship between two variables leads to **bivariate statistical analysis**, while a question looking at the interaction among more than two variables requires **multivariate statistical analysis**.

This distinction between univariate, bivariate, and multivariate analyses replicates the way in which statistical analysis is often undertaken. In the process of doing research we usually collect data on many variables. We may collect data on people’s weekly income, their age, health levels, how much TV they watch, and a myriad number of other variables that may be of interest. We then analyze each of these variables individually. Once we have described the distribution of each variable, we usually then build up a more complex picture by linking variables together to see if there is a **relationship** among them. Everyone probably has a common-sense notion of what it means for two variables to be ‘related to’, or ‘dependent on’, each other. We know that as children grow older they also get taller: age and height are related. We also know that as our income increases the amount we spend also increases: income and consumption are related. These examples express a general concept for which we have an intuitive feel: as the value of one variable changes the value of the other variable also changes.

To further illustrate the concept of related variables, assume, for example, that we believe a person’s income is somehow related to where they live. To investigate this we collect data from a sample of people and find that people living in one town tend to have a low income, people in a different town have a higher level of income, and people in a third town tend to have an even higher income. These results suggest that ‘place of residence’ and ‘income level’ are somehow related. If these two variables are indeed related, then when we compare two people and find that they live in different towns, they are also likely to have different income levels. As a result we do not treat income as a wholly distinct variable, but as somehow ‘connected’ to a person’s place of residence. To draw out such a relationship in the data we collect, we use bivariate descriptive statistics that do not just summarize the distribution of each variable separately, but rather *describe the way in which changes in the value of one variable are related to changes in the value of the other variable*.

If we do believe two variables are related we need to express this relationship in the form of a **theoretical model**.

A **theoretical model** is an abstract depiction of the possible relationships among variables.

For example, the second research question with which I began this chapter is interested in the relationship between the sex and health status of my students. Before analyzing any data I may collect for these variables, I need to specify the causal structure – the model – that I believe binds these two variables together. For this example the model is easy to depict: if there is a relationship it is because a student’s sex somehow affects the student’s health level. It is not possible for the relationship to ‘run in the other direction’; a student’s sex will not change as a result of a change in their health level. In this instance we say that sex is the **independent variable** and health status is the **dependent variable**.

The variation of an **independent variable** affects the variation of the **dependent variables** in a study. The factors that affect the distribution of the independent variable lie outside the scope of the study.

Determining the model that characterizes any possible relationship between the variables specified by our research question is not always so easy. Consider again the example of income and place of residence. We can model the possible relationship between these two variables in many different ways. The simplest way in which two variables can be causally related is through a **direct relationship**, which has three possible forms.

1. *One-way direct relationship with income as dependent.* This models the relationship as a one-way street running from place of residence to income (Figure 1.1). We may have a theory that argues job and career opportunities vary across towns and this affects the income levels of people living in those towns. In this case we argue that there is a pattern of dependence with income as the dependent variable and place of residence as the independent variable.

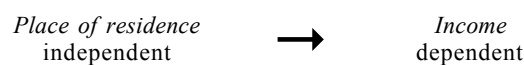


Figure 1.1 One-way direct relationship with income as dependent

2. *One-way direct relationship with place of residence as dependent.* Another group of social researchers may disagree with the previous model; they come from another theoretical perspective that agrees there is a pattern of dependence between the two variables, but it runs in the other direction. People with high incomes can choose where they live and will move to the town with the most desirable environment. Thus place of residence is the dependent variable and income is the independent variable (Figure 1.2).

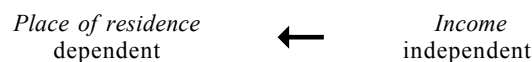


Figure 1.2 One-way direct relationship with place of residence as dependent

3. *Two-way direct relationship with place of residence and income mutually dependent.* A third group of researchers may agree that the two variables are related, but believe that both types of causality are operating so that the two variables affect each other. In this model, it is not appropriate to characterize one variable as the independent and the other as the dependent. Instead they are **mutually dependent** (Figure 1.3).

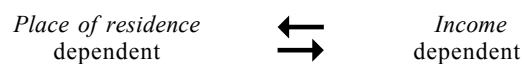


Figure 1.3 Two-way direct relationship with place of residence and income mutually dependent

The important point to remember is that we choose a model based on particular theoretical views about the nature of the world and people's behavior. These models may or may not be correct. Statistical analysis cannot prove any of the types of causality illustrated above. All it can show is some statistical relationship between observed variables based on the data collected. The way we organize data and the interpretation we place on the results are contingent upon these theoretical presuppositions. The same data can tell many different stories depending on the theoretical preconceptions of the story-teller. For instance, we have presented the three simplest models for characterizing a relationship between two variables.

There are more complex models that involve the relationship between three or more variables. To explore more complex relationships would take us into the realm of **multivariate analysis** – the investigation of relationships between more than two variables, which we explore in later chapters. However, it is important to keep in mind when interpreting bivariate results the fact that any observed relationship between two variables may be more complicated than the simple cause-and-effect models described above.

Descriptive statistics

We have discussed some conceptual issues that arise when we plan to gather information about variables. The rest of this book, however, is concerned with **data analysis**; what do we do with measurements of variables once we have taken them? Usually the first task of data analysis is the calculation of **descriptive statistics**.

Tolstoy's *War and Peace* is a very long book. It would not be possible to do such a book justice in any way other than to read it from cover to cover. However, this takes a lot of time and concentration, each of which may not be readily available. If we want simply to get a gist of the story, a shorter summary is adequate. A summary reduces the thousands of words that make up the original book down to a few hundred, while (hopefully) retaining some of the essence of the story. Of course, the summary will leave out a great deal of detail, and the way the book is summarized for one purpose will be different from the way it is summarized for another. Nevertheless, although much is lost, something is also gained when a book so large is summarized effectively.

The same holds true with research. Most research projects will generate a wealth of information. Presenting the results of such research in their complete form may be too overwhelming for the reader so that an 'abridged version' is needed; **descriptive statistics** provide this abridged version.

Descriptive statistics are the numerical, graphical, and tabular techniques for organizing, analyzing, and presenting data.

The great advantage of descriptive statistics is that they make a mass of research material easier to 'read'. By *reducing* a large set of data into a few statistics, or into some picture such as a graph or table, the results of research can be clearly and concisely presented.

Assume we conduct a survey that gathers the data for the age of 20 students in my statistics class, and obtain the following results:

18, 21, 20, 18, 19, 18, 22, 19, 20, 18, 19, 22, 19, 20, 18, 21, 19, 18, 20, 21

This arrangement of the measurements of a variable is called a **distribution**. I could present this distribution of the **raw data** as the results of the research, which, strictly speaking, they are. It is not difficult to see, however, that very little information is effectively communicated this way. It is evident that the raw data, when presented in this 'naked form', do not allow us to make any meaningful sense of the variable we are investigating. It is not easy to make any sense about the way age is distributed among this group of students.

We can, alternatively, take this set of 20 numbers and put them through a statistical 'grinder', which produces fewer numbers – **statistics** – that capture the relevant information contained in the raw data. Descriptive statistics tease out some important feature of the distribution that is not evident if we just present the raw scores. One such feature we will focus on in later chapters is the notion of average. For example, we might calculate a single figure for the 'average' age and present this single number as part of the results of the research. The measure of 'average' chosen will certainly not capture all the information contained in the primary data – no description ever does that – but hopefully it will give a general notion of what the 20 cases 'look like' and allow some meaningful interpretation.

We have just introduced the notion of 'average' as an important feature of a distribution of scores in which we might be interested. In more technical terms this is one of many **numerical techniques** for describing data since it involves the use of mathematical formulas for making calculations from the raw data. There are also a variety of **graphs** and **tables** in which data can be represented visually to make the information easier to read. The chapters following Chapter 2 will explore these various methods for describing data.

In all of these chapters we will see that regardless of whether we are using graphs, tables, or numerical techniques as the descriptive statistics we are using to summarize our data, the specific choice among these broad classes of statistics is largely determined by the level of measurement for each variable and whether we are undertaking univariate, bivariate, or

multivariate analysis of the variables. This is why we spent some time in the previous sections discussing these concepts. All of these various ways of describing data are summarized in Table 1.2.

Table 1.2 Types of descriptive statistics

Type	Function	Examples
Graphs	Provide a visual representation of the distribution of a variable or variables	Pie, bar, histogram, polygon (univariate) Clustered pie, clustered/stacked bar (bivariate, nominal/ordinal scales) Scatterplot (bivariate, interval/ratio scales)
Tables	Provide a frequency distribution for a variable or variables	Frequency table (univariate) Crosstabulations (bivariate/multivariate)
Numerical measures	Mathematical operations used to quantify, in a single number, particular features of distributions	Measures of central tendency (univariate) Measures of dispersion (univariate) Measures of association and correlation, regression (bivariate/multivariate)

Given the array of descriptive statistics available, how do we decide which to use in a specific research context? The considerations involved in choosing the appropriate descriptive statistics are like those involved in drawing a map. Obviously, a map on the scale of 1 to 1 is of no use (and difficult to fold). A good map will be on a different scale, and identify only those landmarks that the person wanting to cover that piece of terrain needs to know. When driving we do not want a roadmap that describes every pothole and change of grade on the road. We instead desire something that will indicate only the major curves, turn-offs, and distances that will affect our driving. Alternatively a map designed for walkers will concentrate on summarizing different terrain than one designed for automobile drivers, since certain ways of describing information may be ideal for one task but useless for another.

Similarly, the amount of detail to capture through the generation of descriptive statistics cannot be decided independently of the purpose and audience for the research. Descriptive statistics are meant to *simplify* – to capture the essential features of the terrain – but in so doing they also leave out information contained in the original data. In this respect, descriptive statistics might hide as much as they reveal. Reducing a set of 20 numbers that represent the age for each of 20 students down to one number that reflects the average obviously misrepresents cases that are very different from the average (as we shall see).

In other words, just as a map loses some information when summarizing a piece of geography, some information is lost in describing data using a small set of descriptive statistics: it is a question of whether the information lost would help to address the research problem at hand. In other words, *the choice of descriptive statistics used to summarize research data depends on the research question we are investigating.*

Exercises

1.1 Consider the following ways of classifying respondents to a questionnaire.

- (a) Voting eligibility:
- Registered voter
 - Unregistered but eligible to vote
 - Did not vote at the last election
- (b) Course of enrolment:
- Physics
 - Economics
 - English
 - Sociology
 - Social sciences

- (c) Reason for joining the military:
- Parental pressure
 - Career training
 - Conscripted
 - Seemed like a good idea at the time
 - No reason given

Do any of these scales violate the principles of measurement? If so, which ones and how?

1.2 What is the level of measurement for each of the following variables?

- (a) The age in years of the youngest member of each household
- (b) The color of a person's hair
- (c) The color of a karate belt
- (d) The price of a suburban bus fare
- (e) The years in which national elections were held
- (f) The postcode of households
- (g) People's attitude to smoking
- (h) Academic performance measured by number of marks
- (i) Academic performance measured as fail or pass
- (j) Place of birth, listed by country
- (k) Infant mortality rate (deaths per thousand)
- (l) Political party of the current Member of Parliament or Congress for your area
- (m) Proximity to the sea (coastal or non-coastal)
- (n) Proximity to the sea (kilometers from the nearest coastline)
- (o) Relative wealth (listed as 'Poor' through to 'Wealthy')
- (p) The number on the back of a football player

1.3 Find an article in a journal that involves statistical analysis. What are the conceptual variables used? How are they operationalized? Why are these variables chosen for analysis? Can you come up with alternative operationalizations for these same variables? Justify your alternative.

1.4 For each of the following variables construct a scale of measurement:

- | | |
|----------------------|-----------------------|
| (a) Racial prejudice | (e) Voting preference |
| (b) Household size | (f) Economic status |
| (c) Height | (g) Aggressiveness |
| (d) Drug use | |

For each operationalization state the level of measurement. Suggest alternative operationalizations that involve different levels of measurement.

1.5 Which of the following are discrete variables and which are continuous variables?

- | | |
|---------------------------------------|--|
| (a) The numbers on the faces of a die | (d) The number of cars in a carpark |
| (b) The weight of a new-born baby | (e) Household water use per day |
| (c) The time at sunset | (f) Attitude to the use of nuclear power |

2

Setting up an SPSS data file

This chapter will introduce SPSS, which is a widely used statistical package for analyzing data.

Obtaining a copy of SPSS

To conduct the procedures detailed in this book for yourself, using the data files on the CD, you need a copy of SPSS. At the time of printing the latest version of SPSS was Version 13, but this book is relevant for all versions from 10 and after. SPSS is sold as a Base system for an annual license fee plus annual renewal charge, plus add-on modules that require additional cost and extend the functionality of the Base. This text will cover the functions that are available as part of the Base so that it is relevant to all users of this program, regardless of the configuration. Those who do have add-on modules should explore these, however, to see if they provide alternative and better options for obtaining the statistical results we describe in the rest of the book. There are several options to obtain a copy of SPSS Base:

1. *Purchase a commercial version of SPSS.* This can be done through a software retailer or on-line (www.spss.com). The initial annual fee is substantial, although the annual upgrade is much cheaper. A demonstration copy can be downloaded for free from the SPSS website, but this has a limited period of use.
2. *Access a site license copy.* If you belong to a large organization such as a university or public sector department, your organization may have negotiated a site license with SPSS for installation and use of the program. You should check with the relevant people who manage software licenses to see if you can obtain a copy of the program through such an arrangement and what the licensing conditions include.
3. *Purchase an SPSS Graduate Pack.* If you are a university or college student you may be able to purchase a Graduate Pack from your campus bookstore, which includes a manual and copy of the software at a much lower price than the commercial version. As with any software you purchase, however, you should check the licensing details before purchase.
4. *Purchase a Student Version.* A Student Version of the program is available at a relatively inexpensive price. This does not have the full functionality of the commercial version or Graduate Pack; it is limited to 1500 cases and 50 variables. However, it is suitable for most of the needs of an introductory user for non-commercial purposes. Prentice-Hall distributes SPSS Student Version through university and college bookstores around the world. Simply present your valid student identification. If your campus bookstore does not carry SPSS Student Version, order the software on-line at www.prenhall.com, or ask the bookstore manager to contact the local Prentice-Hall distribution office.

Alternatives to SPSS

This text details SPSS procedures for statistical analysis because it is the most widely used statistics package (other than common spreadsheet programs such as Excel, which can be used for complex statistical analysis but are really designed for other purposes). This text does

not use SPSS because it is the best; readers will note my frustration with this program and its peculiarities as they read the following chapters. It is only appropriate therefore to draw your attention to alternatives that are available and which you may choose rather than SPSS. To assist this the CD accompanying this book contains all the data files for the following examples in tab-delimited ASCII format so that they can be imported into a wide range of alternative software. There are three broad classes of alternatives to SPSS.

1. *Other comprehensive commercial programs.* There are many commercial alternatives to SPSS such as GB-Stat, InStat, JMP, Minitab, SAS, and StatA. A full list of such packages is available at www.statistics.com/content/commsoft/fulllist.php3
2. *Free comprehensive programs.* An exciting development in recent years is the amount of free software that are available, usually produced under the open source license, and this includes some useful statistical analysis software. A listing of such software is available at freestatistics.altervista.org/stat.php and members.aol.com/johnp71/javasta2.html. Of these, three are particularly worth mentioning. Epi Info has been developed by the US Center for Disease Control (www.cdc.gov/epiinfo) mainly for the use of epidemiologists and other health scientists, although researchers from other disciplines will find this program suitable to their needs. An open source version of Epi Info, called OpenEpi, which runs on all platforms and in a web browser, is available from www.openepi.com. Second, the program StatCrunch (www.statcrunch.com) allows users to load, analyze, and save results using nothing more than a standard web browser, but with much of the functionality (and sometimes more) of SPSS. Last, an extremely powerful and comprehensive open source program called R is available for all platforms for free (www.r-project.org). At present it requires some knowledge of the R programming language, but graphical user interfaces are being developed so that users can select commands from a drop-down menu; a phase of development similar to that which SPSS experienced in the early 1990s.
3. *Calculation pages for specific statistical pages.* These are web pages that provide tools for conducting specific analysis, including many that we will cover in later chapters. We will refer to some of these below, but a general listing of these pages is available at the web page members.aol.com/johnp71/javastat.html or StatPages.net.

Options for data entry in SPSS

In setting up an SPSS file to undertake statistical analysis we encounter in a very practical way many of the conceptual issues introduced in Chapter 1. Assume that in order to answer the research questions I posed at the start of Chapter 1 I survey 200 of my statistics students. In this hypothetical survey I am interested in three separate variables: age, sex, and health level. Sex is measured by classifying cases into male or female (nominal). The survey respondents also rate their respective health level as 'Very healthy', 'Healthy', or 'Unhealthy' (ordinal, but with a 'Don't know' option). Finally, I ask students their respective age in whole years on their last birthday (interval/ratio). This chapter will detail how we can record this information directly in SPSS so that we can use it as an example when we learn the techniques for statistical analysis in later chapters. However, there are other means by which data can be imported into SPSS other than by entering it directly, as we will do here.

1. *Importing data from database, spreadsheet, or statistics programs.* SPSS recognizes files created by other popular data programs such as Excel, Systat, Lotus, dBase, and SYLK. The range of programs and file extensions that SPSS recognizes are listed under the **File/Open/Data** command; click on the **File of Type:** (Windows) or **Enable** (Macintosh)

option when the **Open File** box appears to see the full list. This is a convenient option for large-scale data entry, since such programs are widely available and well-known. Thus only one copy of SPSS needs to be purchased and located on the computer where the analysis will be conducted, with data entry performed on these other programs and then imported into the copy of SPSS. The limitation is that data have to be entered into these other programs in a specific way so that problems and errors are not encountered when importing into SPSS. These problems can sometimes be avoided by simply copying the block of data from the other program and then pasting the data into the SPSS **Data Editor** and defining the variables within SPSS using the procedures we detail below.

2. *Importing text files (*.txt)*. If you are not sure whether SPSS will read the 'native' version of the data file you create in another program, you may be able to save the file as a tab-delimited text (ASCII) file. SPSS will import such a file through the **File/Read Text Data** command. An Import Wizard will then appear to assist you to bring the data across.
3. *Importing from data entry programs*. There are many programs available that require no special knowledge of SPSS (or any other data program) to facilitate data entry. These programs have many advantages: for example, they often restrict the numbers that can be entered to only those that are considered valid, thus avoiding errors. They also save the data set directly in SPSS format, or else in ASCII format that can be imported into SPSS. SPSS Inc has its own product called SPSS Data Entry, but other commercial services exist such as Quest (www.dipolar.com). Similarly, some of the free programs listed above, such as Epi Info also provide data entry facilities, as do on-line survey programs such as PHPSurveyor (phpsurveyor.sourceforge.net).

The SPSS Data Editor

When you launch SPSS a window first appears asking **What would you like to do?** Select **Type in data** and then **OK**, which is the option for directly entering new data, rather than opening a file that already has data. You will then see the **SPSS Data Editor** window (Figure 2.1); make sure that the **Data View** tab at the bottom-left of the window is selected.

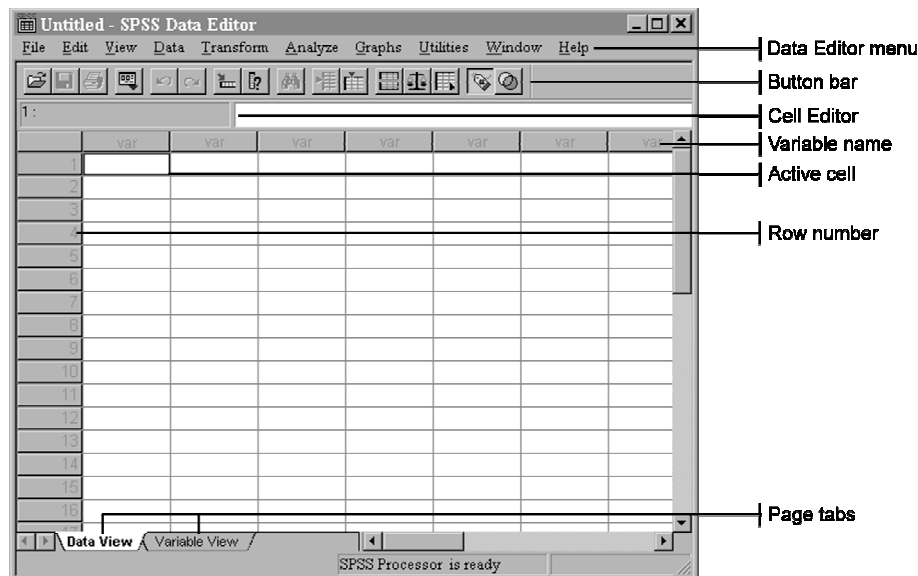


Figure 2.1 The SPSS Data Editor window open on Data View

Note the **Data Editor** menu at the top of the window. We define and analyze data by selecting commands from this menu. Usually selecting commands from the menu will bring up on the screen a small rectangular area called a **dialog box**, from which more specialized options are available, depending on the procedure we want to undertake. By the end of this and later chapters this way of hunting through the **Data Editor** menu for the appropriate commands will be very familiar. In fact, it is very similar to many other software applications that readers have encountered, such as word processing and spreadsheet software.

In Figure 2.1, below the **Data Editor** menu bar is a **Button bar** that provides an *alternative* means for activating many of the commands contained within the menu. Generally we will concentrate on using the **Data Editor** menu to activate SPSS commands, even though sometimes clicking on the relevant button on the **Button bar** may be quicker. We will concentrate just on the use of menu options simply to ensure that we learn one method consistently; after some level of proficiency readers can then decide whether selecting commands through the menu or by clicking on the buttons is preferable.

You should also observe that the unshaded cell at the top left of the page in Figure 2.1 has a heavy border, indicating it is the **active cell**. The active cell is the cell in which any information will be entered if I start typing and then hit the enter-key on the keyboard. Any cell can be made active simply by pointing the cursor at it and clicking the mouse. You will then notice a heavy border around the cell on which you have just clicked, indicating that it is the active cell.

The **Data Editor** window consists of two pages, indicated by the page tabs at the bottom-left of the **Editor** window. The first is the **Data View** page on which we enter the data for each variable. The **Data View** is the 'data page' on which all the information will be entered. Think of it as a blank table without any information typed into it. The **Data View** page is made up of a series of columns and rows, which form little rectangles called **cells**. Each column will contain the information for each one of the variables, and each row will contain the information for each case. The first row of cells at the top of the columns is shaded and contains a faint **var**. This row of shaded cells will contain the names of the variables whose information is stored in each column. Similarly, the first column is shaded and contains faint **row numbers** 1, 2, 3, etc.

The second page is the **Variable View** page on which we define the variables to be analyzed. A *column* in the **Data View** page stores data for a single variable, whereas each *row* in the **Variable View** contains the definition for a single variable.

We can switch from the **Data View** page to the **Variable View** page in one of two ways:

1. click on the **Variable View** tab at the bottom of the window, or
2. point the cursor at the shaded cell at the top of the relevant column (**Data View**) or left-edge of the relevant row (**Variable View**) and double-click the mouse button (Figure 2.2).

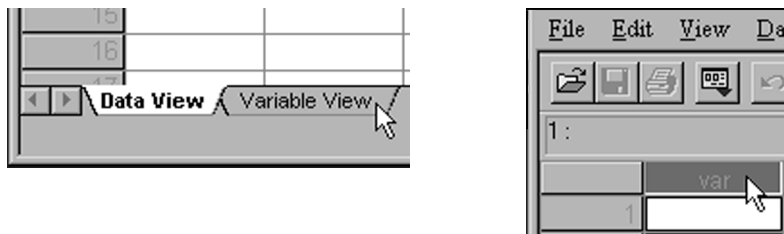


Figure 2.2 Switch from **Data View** to **Variable View** by tabbing or double-clicking column head

Try both methods to see that the result will be the same: the **Variable View** page moves to the front of the **Data Editor** window (Figure 2.3).

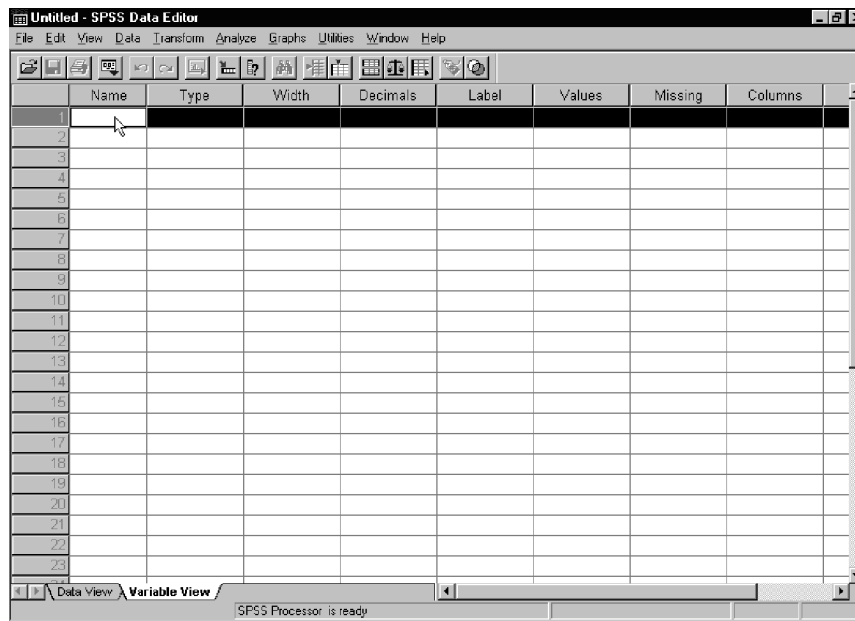


Figure 2.3 The Variable View page

Each row in this page defines a single variable. We will illustrate the process of defining a variable in the **Variable View** page using students' sex.

Assigning a variable name

The first task is to give the variable a name. If we make the cell below **Name** in the first row active by clicking on it, we can type in a variable name, which in this instance is **sex**.

There are some limitations imposed by SPSS on the names we can assign to our variables:

- In SPSS version 12 or later, a variable name can have a maximum of 64 characters made up of letters and/or numbers. In earlier versions, only eight characters are permitted.
- A variable name must begin with a letter.
- A variable name cannot end with a period.
- A variable name cannot contain blanks or special characters such as &, ?, !, ' , * , or ,
- A variable name must be unique. No other variable in a data file can have the same name.
- A variable name appears in lower-case letters, regardless of the case in which it is typed.

Given these specific limitations, there are two schemes for naming variables in SPSS. One scheme uses **sequential names** indicating where on the research instrument (the questionnaire, interview schedule, record sheet, etc.) the variable appears. An example of this might be to name variables q1, q2, q3a, q3b, and so on, to indicate which question number on a questionnaire generated the data for a given variable. This provides a quick and easy way of assigning variable names and allows you to link a name directly to the research instrument on which the data are recorded. Its disadvantage is that the individual variable names do not give an impression of the contents of the variable.

The other variable naming scheme that is commonly adopted, and which we are using here, is **descriptive names**. This is a more time-consuming method, but the individual variable name, such as **sex**, gives a direct impression as to what the data in a given column are about.

It is also possible to use a combination of these two naming schemes. For example, we might use sequential names for the bulk of responses to a questionnaire, but also use descriptive names for key demographic variables such as sex and age.

Setting the data type

You should notice that as soon as you enter the variable name and strike the return key, information is also automatically entered in the subsequent cells in the first row. These are termed **default** settings; things about the variable's definition that are pre-set unless we choose to change them.

For example, in the second column headed **Type** the word **Numeric** appears. This is the most common form of data type, whereby numbers will be entered to indicate the category that each case falls into for the specific variable. In this instance we plan to enter 1 for female and 2 for male. Since most data are of a numeric type, SPSS sets this as the default so we don't need to change it. If we did want to change the data type we click on the small shaded square next to **Numeric**. This brings up the **Type** dialog box in which we can select other data types (Figure 2.4).

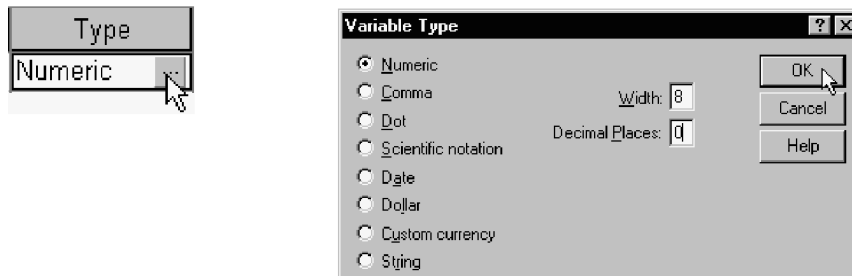


Figure 2.4 Setting the data Type

There are a number of other choices available for data type listed below **Numeric**. The following is a brief description of some of these items (a useful feature of SPSS is the contextual help available; if you right-click the mouse button on an item in any dialog box for which you require more information, such as the list of data types in the **Variable** Type box, a contextual help option appears which if selected will give details about that item):

- *Comma*. This defines a numeric variable whose values are displayed with a comma for every three places and with a period as the decimal delimiter.
- *Scientific notation*. A numeric variable whose values are displayed with an imbedded E and a signed power-of-ten exponent. The **Data Editor** accepts numeric values for such variables with or without an exponent. The exponent can be preceded either by E or D with an optional sign, or by the sign alone.
- *Date*. A numeric variable whose values are displayed in one of several calendar-date or clock-time formats. You can enter dates with slashes, hyphens, periods, commas, or blank spaces as delimiters. The century range for 2-digit year values is determined by your **Options** settings. This data type can be useful, for example, where a person's birth date needs to be recorded, or the date on which a survey was completed needs to be included with the data set.
- *Custom currency*. A numeric variable whose values are displayed in one of the custom currency formats that are defined in the **Currency** tab of the **Options** dialog box. Defined custom currency characters cannot be used in data entry but are displayed in this format in the **Data View** page.

- *String*. Values of a string variable are not numeric, and hence not used in calculations. They can contain any characters up to the defined length. Upper and lower case letters are considered distinct. Also known as alphanumeric variable. An example is ‘m’ and ‘f’ for male and female respectively. This data type is often used for typing responses to open-ended questions that may be different for each case, and therefore cannot be precoded.

Setting the data width and decimal places

The **Width** of the data is the maximum number of characters that can be entered as a datum for each case. The default setting is eight, so if we had values for a variable with more than eight digits we would need to change this. For example, if we were entering the populations of various countries, we would not be able to include data for countries such as the USA or China, which have populations greater than 99,999,999. We would need to change the default data **Width** from 8 to a higher number such as 10.

The number of **Decimal Places** is a ‘cosmetic’ function, in that it alters the way data are displayed once they are entered but does not affect what we can do. If we do not change the default setting of 2 decimal places, 1 will show up as 1.00 on the **Data View** page.

There are two ways by which the data **Width** and **Decimal Place** settings can be changed from the default settings.

1. In the **Variable Type** dialog box that we brought up to enter the data **Type** we also have the option to change the variable width and the number of decimal places.
2. Another way to change these aspects of the variable definition is in the columns headed **Width** and **Decimals** on the **Variable View** page. Clicking on either of these cells produces up and down arrows on the right edge of the cell, which can be used to change values (Figure 2.5). Alternatively, you can highlight over the number 8 and type in the desired value.

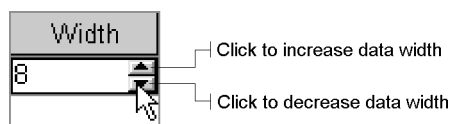


Figure 2.5 Setting the data Width

Defining variable labels

The next column is headed **Label**. A **Variable Label** is a longer description of the variable (up to 120 characters) than can be included in the **Name** column. It also permits formatting that is more suitable for presentation purposes, such as the use of capital letters and spaces between words. Although the short variable name **sex** is fairly self-explanatory, to get into the habit of providing variable labels we will type **Sex of student** in the **Label** column. If we do not provide a label, any tables we generate for this variable, for example, will be headed by ‘sex’; by providing a longer and better formatted label, the table will instead be headed by ‘Sex of student’, which is a much better way of presenting results.

There are a couple of tips for providing **Variable Labels**:

- With interval/ratio data it is very useful to include the unit of measurement in the label. Thus, even though the variable ‘age’ does not seem to require a **Variable Label** to explain its meaning, it is useful to type ‘Age in years’ in the label for that variable.
- It is often helpful to use the exact wording of a questionnaire/interview question as a **Variable Label** so that it will be presented in any output.

Defining value labels

The **Value Labels** function allows us to specify our coding scheme; the way in which responses will be transformed into ‘shorthand’ codes (numbers) that allow us to perform statistical analysis, and especially to make data entry quicker. In SPSS the numerical codes are called **values** and the actual responses are called **value labels**. Thus sex has two value labels: female and male, and we link each label to a specific code number or **value**:

- 1 = female
- 2 = male

Instead of typing in male or female as our data, we type in the *codes* assigned to these labels which is a much faster procedure.

With a nominal scale such as ‘sex’ the actual numerical code given to each value label is arbitrary: we can just as easily reverse the order and assign 1 to male and 2 to female. In fact, we could assign 3 to female and 7 to male, or any other combination of values. But, generally, the simpler the coding scheme the better. The procedure for defining the value labels for sex is provided in Table 2.1 and Figure 2.6.

Table 2.1 Assigning Value Labels in SPSS

SPSS command/action	Comments
1 In the column headed Values click on the small shaded square next to None	This brings up the Value Labels dialog box so that we can assign labels
2 In the box next to Value: type 1	
3 In the box next to Value Label: type female	You will notice that as soon as you start typing Add suddenly darkens, whereas it was previously faint
4 Click on Add	This pastes the information into the adjacent area so that 1=“female” . The cursor will automatically jump to the box next to Value:
5 Type 2	
6 In the box next to Value Label: type male	
7 Click on Add .	2=“male” will be added to the list
8 Type 3	
6 In the box next to Value Label: type Did not answer	
7 Click on Add .	3=“Did not answer” will be added to the list
8 Click on OK	

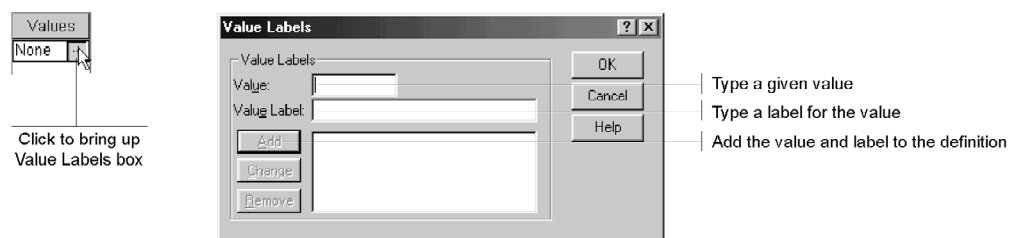


Figure 2.6 Assigning value labels in SPSS

If you receive an error message when you click on **Continue** stating ‘**Any pending Add or Change operations will be lost**’, it is because you have forgotten to click the **Add** button after typing in a value and value label. If this happens click on **OK** which will return you to the **Value Labels** dialog box and click on **Add**.

This variable is an easy one for providing value labels, but for other instances there are some simple rules to follow in determining value labels that can prevent problems during analysis.

1. *Give separate codes for all possible types of 'non-valid' responses.* In my hypothetical survey of students, I asked students to rate their own health. The scale had three options, plus a 'Don't know' option at the end of the scale for students who did not feel competent to answer. For this variable I would like to separate students that answered 'Don't know' and exclude them from analysis. In other words, I am treating 'Don't know' as a non-valid response. Some students, however, may simply not answer this question, and such cases also represent non-valid responses. In other words, there are at least two reasons why a student may not give 'Unhealthy', 'Healthy', or 'Very healthy' as the response; either they responded 'Don't know' or they simply did not answer. In order to allow us to later distinguish between these types of non-valid responses, they are each assigned a unique code as follows:

1 = Unhealthy
 2 = Healthy
 3 = Very healthy
 4 = Don't know
 5 = Did not answer

2. *Give 'null' values a code of zero.* We often encounter survey questions that have a No/Yes response set. In such instances we would code '0 = No'. The value of 0 is assigned to No for a deliberate reason; it is a type of response that indicates a null response: a case where no quantity of the variable is present. Other examples of null responses are responses of 'Never' and 'Not at all'. Giving these responses a value of zero is very useful in later analysis, such as scale construction where values for a set of variables are added together.
3. *For ordinal variables, make sure that the numeric scale reflects the measurement scale.* Variables measured at the ordinal level have categories that indicate the relative strength in which the variable is present. Our scale for measuring the health of students is an example. We can say that a student who answers 'Very healthy' *possesses more* of the variable 'health' than does a person who responds 'Unhealthy'. This quantitative increase should be reflected in the numerical codes assigned to the categories: 'Very healthy' should get the biggest number, and 'Unhealthy' the smallest. Our coding scheme might therefore be:

1 = Unhealthy
 2 = Healthy
 3 = Very healthy

This is particularly helpful in later data analysis, where correlations might be calculated with this variable.

4. *Do not give interval/ratio scales complete value labels.* With interval/ratio scales we do not need to code the responses since the values 'speak for themselves', especially if we have included the units of analysis in the **Variable Label**. Thus we don't need to indicate that case number '1 = 1', Case number '2 = 2', and so on. Nor do we have to state for age that '20 = 20 years of age', '21 = 21 years of age', and so on. The only situation in which we might use the Value Labels function for interval/ratio data is where there are specific categories of missing data that we want to identify (as is the cases with the age variable in our example of student survey responses).

Setting missing values

The next option in defining a variable is the **Missing** values option. A missing value is a number that indicates to SPSS that the response is not valid (as we discussed in the previous section) and should not be included in the analysis. Missing values can arise for many

reasons. For any survey question there is always the possibility that someone simply did not answer the question, or else wrote with illegible handwriting. In another instance, a survey question, such as the Health rating question in our student survey, may allow for 'Don't know' at the end of the scale and as such should not be used in analysis. A third type of missing data occurs where we have skip or filter questions that result in some questions being not applicable to all respondents. For whatever reason, when we do not have a useful datum for an individual case for a specific variable we need to enter a missing value into the relevant cell, indicating that a valid response was not provided in that instance and therefore should not be included in any analysis of that variable.

SPSS has a default setting called the **system missing value** that appears as a period in a cell where no valid response is entered. We can also provide **user-defined missing values**, whereby we specify a particular number to indicate invalid responses. We need to be careful to select a value for the missing value that is one that the variable cannot possibly take. In this example, we can choose 3 to be the missing value, since it is impossible for the variable 'sex' to take on this value. Obviously if we were measuring the age of pre-school children, 3 would not be an appropriate choice for the missing value; 99 might be better because such a score could not actually represent a real case. The procedures for assigning user-defined missing values to the 'Sex of student' variable in my survey are presented in Table 2.2 and Figure 2.7.

Table 2.2 Assigning user-defined missing values on SPSS

SPSS command/action	Comments
1 In the column headed Missing click on the small shaded square next to None	This brings up the Missing Values: box. The radio button next to No missing values is selected indicating that no user-defined missing values is the default setting
2 Click on the radio button next to Discrete missing values	The radio button to the left of Discrete missing values is selected, and the cursor will be flashing in the adjacent rectangle
3 Type 3	
4 Click on OK	

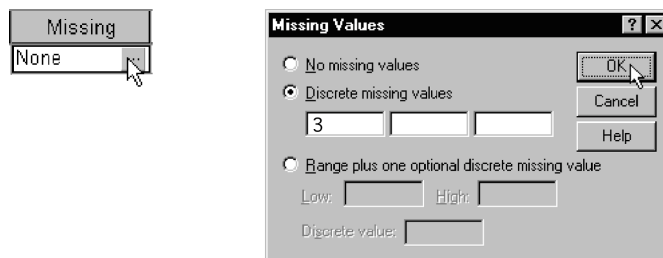


Figure 2.7 Assigning missing values in SPSS

Having defined the missing value when we set up the data file, we can then enter 3 during the data entry process if we encounter a respondent for whom their sex cannot be determined.

Setting the column format and alignment

The next columns on the **Variable View** page are headed **Column** and **Align**. These options are basically cosmetic, in so far as they only affect the appearance of the **Data Editor** without affecting the type of analysis we can do. These columns affect two particular settings. The first is the column width as it appears in the **Data Editor**, and the other is the alignment of the data contained in each column. The default settings are a column width of 8, and right alignment, which are both adequate for our purposes, so we will change any these settings.

Specifying the level of measurement

The last step in defining a variable for data entry is to specify the level of measurement for the variable. The default setting is **Scale**, which is SPSS's unfortunate name for interval/ratio data. In this example, sex is a nominal variable, so to change the setting to nominal we click on the down-arrow next to **Scale** and select **Nominal** (Figure 2.8).

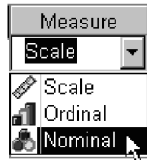


Figure 2.8 Assigning the level of measurement

We can now go through the same variable definition procedure for our remaining two variables. I leave it to you to go through the steps we have just followed, but adapting them to record the relevant information for Health rating and Age. To help you along you should follow the coding scheme in Table 2.3. In fact, before undertaking data entry it is very helpful to write out a coding scheme such as this to clarify the definitions that you will follow. Constructing a coding scheme is especially helpful if more than one person is involved in the data entry process, so that everyone follows the same scheme.

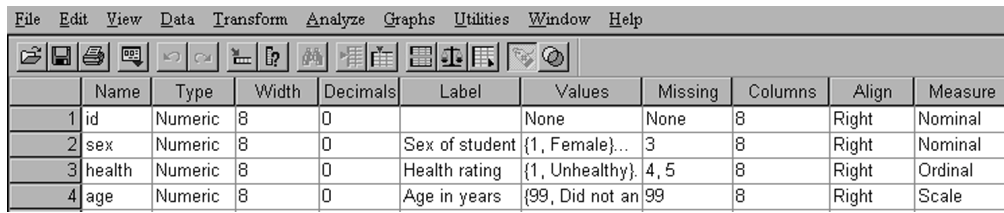
Table 2.3 Coding scheme

Variable name	Variable label (optional)	Values = Value labels	Missing values	Level of measurement
sex	Sex of respondent	1 = Female 2 = Male 3 = Did not answer	3	Nominal
health	Health rating	1 = Unhealthy 2 = Healthy 3 = Very healthy 4 = Don't know 5 = Did not answer	4, 5	Ordinal
age	Age in years	99 = Did not answer	99	Interval/ratio

One of the big advantages of the **Variable View** page is the ease with which we can copy aspects of one variable's definitions to another. For example, if we had two or more variables with exactly the same value labels, we could specify these labels for each variable individually, or we can simply define one variable's value labels, click on this cell with the definitions, select **Edit/Copy** from the menu, click on the relevant cell(s) for subsequent variables into which the same value labels will be entered, and select **Edit/Paste**. You might do this here to copy 0 decimal points specified for sex down to the other two variables.

Another consideration that we often make when setting up a data file is to first insert a column for the ID number of each case. We do not need to provide any definitions for this column, since it is simply used to keep track of where the data came from. It ensures that if the data file is sorted so that the cases appear in a different order, or new cases are inserted, each case retains its unique identification number.

If you follow these procedures correctly the **Variable View** page will look like Figure 2.9. Notice that the order of the variables across the columns in the **Data View** page is the same as the order of the variables down the rows in the **Variable View** page. We can easily change this order so that we have particular groups of variables sitting side-by-side in the **Variable View** page.



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	id	Numeric	8	0		None	None	8	Right	Nominal
2	sex	Numeric	8	0	Sex of student	{1, Female}...	3	8	Right	Nominal
3	health	Numeric	8	0	Health rating	{1, Unhealthy}	4, 5	8	Right	Ordinal
4	age	Numeric	8	0	Age in years	{99, Did not an	99	8	Right	Scale

Figure 2.9 Variable View with complete data definitions

For example, rather than have age appear as the last variable, we might want to have age appear in the second column between id and sex. To move this variable, on the **Variable View** page, we follow these steps:

1. Click on the shaded, numbered cell that is the left-edge of the row containing the variable definition for age. The whole row should then be highlighted.
2. Click and hold on the same shaded, numbered cell.
3. Move the cursor up so that the red line that appears sits between the two variables where you want age to appear.
4. Release the mouse button.

Generating variable definitions in SPSS

There are 3 ways in which we can obtain variable definitions during the course of the analysis (without, that is, actually going back into **Variable View**). One method by which we can ask SPSS to provide the coding scheme we have used in the variable definitions is by selecting from the menu **Utilities/File Info** (SPSS versions before 12) or **File/Display Data File Information/Working File** for SPSS version 12 and later (Figure 2.10).

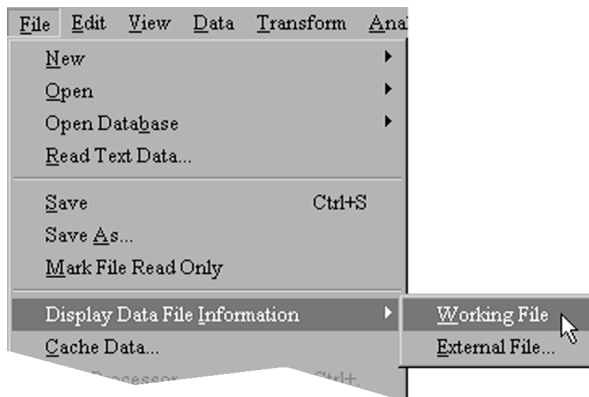


Figure 2.10 Generating file information, SPSS V.12 or later

This command will generate information in the **Viewer** window (Figure 2.11) that can be printed off for reference during the analysis. The numbers for **Position** next to each variable name indicate the column in which the variable appears. Printing out this file information and having it close at hand can be very useful when actually analyzing data. It also provides a record for anyone else who may need to work with the data set and therefore will need to know the coding scheme used for data entry.

File Information

List of variables in the working file

```

Name (Position) Label

Id (1)
  Measurement Level: Nominal
  Column Width: Unknown Alignment: Right
  Print Format: F8
  Write Format: F8

sex (2) Sex of student
  Measurement Level: Nominal
  Column Width: Unknown Alignment: Right
  Print Format: F8
  Write Format: F8
  Missing Values: 3

      Value      Label
        1      Female
        2      Male
        3 M     Did not answer

health (3) Health rating
  Measurement Level: Ordinal
  Column Width: Unknown Alignment: Right
  Print Format: F8
  Write Format: F8
  Missing Values: 4, 5

      Value      Label
        1      Unhealthy
        2      Healthy
        3      Very healthy
        4 M     Don't know
        5 M     Did not answer

age (4) Age in years
  Measurement Level: Scale
  Column Width: Unknown Alignment: Right
  Print Format: F8
  Write Format: F8
  Missing Values: 99

```

Figure 2.11 SPSS File Info output

The **File Info** command allows you to obtain variable definitions for *all* your variables at once. You may, however, occasionally want to quickly obtain the definition for one variable on the screen. This can be done through the **Utilities/Variables** command. This command will bring up the **Variables** dialog box, from which you can select the variable for which definitions are required (Figure 2.12).

The **Utilities/Variables** command provides one additional function, which is specially useful when working with large data files with lots of variables. Say I want to find out what the response to variable **age** is for case number 15 in a particular data set. If I select the row containing the data for case number 15, and in the **Variables** dialog box then select **age** and click on the **Go To** button, the active cell will jump to the one containing this particular datum. This eliminates the need to scroll across and down a large file of numbers which can easily blend into each other and make it difficult to find the particular entry we are trying to locate.

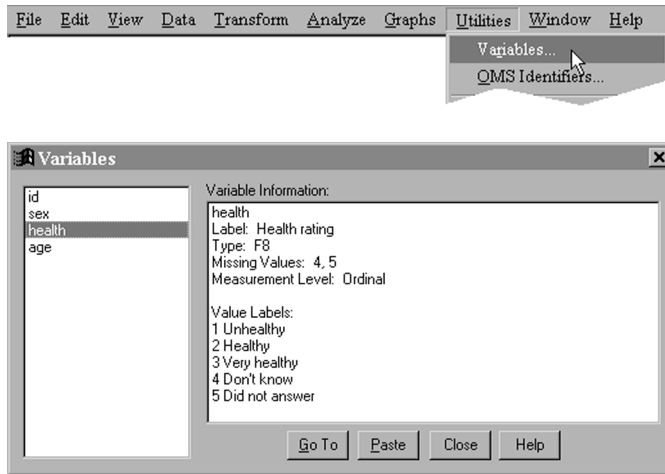


Figure 2.12 SPSS Utilities/Variables command and dialog box

The remaining way by which variable information can be obtained is in the various dialog boxes that appear when we are conducting particular types of analysis. For example, in Chapter 4 we will work with the **Analyze/Descriptive Statistics/Frequencies** command. As with other dialog boxes, we are given a source list of variables from which we can choose variables to actually analyze. If we want to quickly find the definitions for a variable in the list, rather than cancelling and going back to the **Data View**, SPSS provides a contextual menu that we can bring up on screen. To bring up this contextual menu right-button click the mouse on the desired variable. A small menu will appear from which you select **Variable information**. A window appears providing the basic definitions for the chosen variable (Figure 2.13).

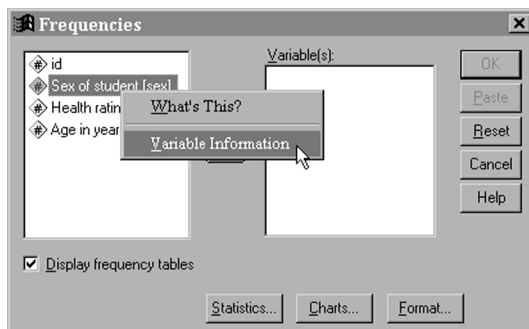


Figure 2.13 SPSS Utilities/Variables command

The SPSS Viewer window

You will notice that SPSS printed the file information we requested on a separate **Viewer** window (Figure 2.14). You will become familiar with this process of requesting output from the menu, based on the data in the **Data Editor**, and viewing the output in the **Viewer** window. To switch back and forth we simply select the relevant 'page' from the **Window** option on the menu bar, or select either window from the task bar at the bottom of the screen (Windows), or right-click on the SPSS icon on the dock (Macintosh).

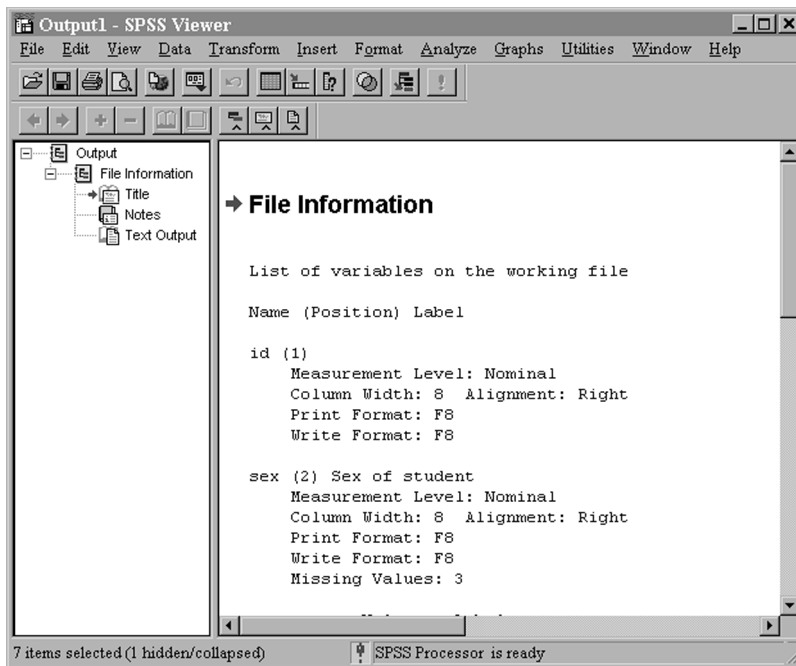


Figure 2.14 The SPSS Viewer window

The major part of the **Viewer** window is a **Display** frame that contains the information we have requested. Two points are worth noting about the **Display**:

- SPSS does not print over existing output whenever we request new information. Instead it adds new output to the bottom of the existing output. When undertaking a lot of analysis this can create a lengthy **Display**.
- ‘Old’ output is not automatically updated whenever we change information on the **Data Editor** window. For example, if we went back and changed the value labels for the variables we are working with, the variable information we have just generated will no longer apply, but will still be recorded on the **Viewer** window. We will need to again run the **File Info** command to generate the information for the updated variable labels.

It is often very helpful to add text to explain the output listed in the **Display**. For example, the file information we generated just states **File Information** as a title. You can double-click on this title (or any other text in the **Display**) and add/change the text, or also change the formatting. Thus you may add text to the title so that it reads **File Information for Ch.02.sav** so that when you print this output you know to which data file it applies. Alternatively, you can use the **Insert** command to insert particular types of preformatted text into the output. The options available for adding new text (rather than editing existing text) are:

- *New Heading*. This appears in the navigator window and acts like a new heading in a Table of Contents.
- *New Title*. This is a first-level heading that normally appears in large, bold font.
- *New Page Title*. A page-break is automatically inserted before a page title, and the typed text appears at the top-center of the page.
- *New Text*. This is small, plain text.

To make it easier to navigate through the output in the **Display** frame, on the left-hand side is a narrower **Outline** frame that provides a ‘Table of contents’. This is a list of the output we have generated during the course of an SPSS session. We can alter the relative size of the two frames that make up the **Viewer** by clicking and dragging the line that separates them. The **Outline** frame provides a number of handy features that can make data analysis much easier, especially when working with a lot of results.

- *Move around the output page.* Sometimes we generate a large amount of output that spreads down several (virtual) pages of the **Display** frame. We may find ourselves constantly referring back and forth to different parts of the output. We can do this by using the vertical scroll-bar on the right-edge of the window. A more direct way is to click on the desired bit of output in the **Outline** list.
- *Hide or show different parts of the output.* We can simplify the **Display** by hiding parts of the output that we don’t think we will be referring to for a while, but still want to keep for possible later use or printing. To hide an item we double-click on its label in the **Outline** frame. This will also hide all other items below this one in the hierarchy of output. An item that is hidden is indicated by the closed-book icon next to its name in the **Outline** frame (by default the Notes that are generated with any output are hidden). To reveal an item we again double-click on it; the icon appearing next to it will now be an open-book.
- *Move or delete selected parts of the output.* I can reorder the way output appears in the **Outline** by first selecting its title in the list, and then dragging-and-dropping it on the point at which I want it to appear. Similarly, by clicking on an item in the **Outline** and striking the delete-key I can permanently remove it from the output. You can also select multiple items in **Outline** by holding down the shift-key for consecutive items in the list, or by holding down the control-key (Windows) or apple-key (Macintosh) for non-consecutive items in the list.
- *Select items for printing.* When it comes to printing SPSS is an extremely wasteful program. The output is not very compact, and the **Display** frame can accumulate a great deal of output during the course of analysis. To avoid waste when printing, it is possible to select a sub-set of output to print. By using the control-key (Windows) or apple-key (Macintosh) you can select individual items from the **Outline** frame before printing. Using the **File/Print** command will then default to printing only the selected items rather than the entire contents of the **Display** frame.

Saving a data file

The remaining action in setting up the data file is to save the file (we need to switch from the **Viewer** window to the **Data Editor** containing the data before we save them) (Figure 2.15).

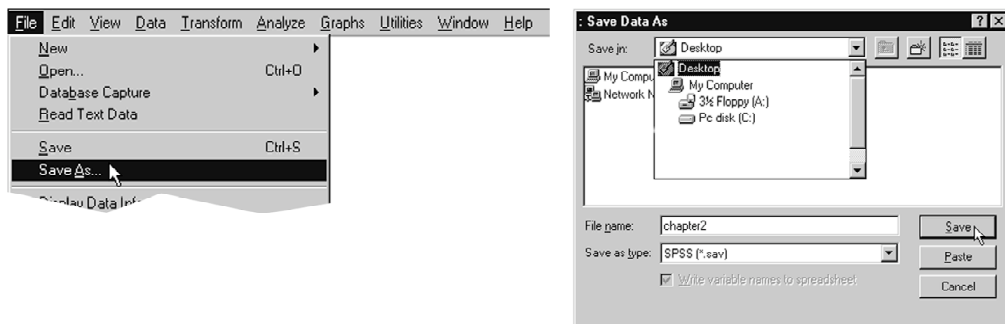


Figure 2.15 The Save Data As command and dialog box

We need to decide where we want to store the file. This is usually a choice between somewhere on the computer's hard drive, or on a disk that we place in the disk drive. Wherever you choose to store data it is very important to *make a regular backup*. No storage medium is immune to errors. Get into the habit of making a copy of your data files to guard against any unforeseeable problems.

Once this has been done the name of the active data file will appear in the bar at the top of the **Data Editor** (Figure 2.16).

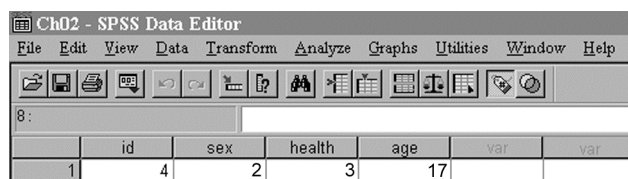
chapter2 - SPSS Data Editor

Figure 2.16 SPSS File name display

After using the **File/Save As** command, a file can be quickly resaved in the same location and with the same name with the **File/Save** command instead of **File/Save As**. In fact, you should not wait until the end of your data entry session to save the file. Mishaps can happen, often at the worst time. Losing data, after spending a considerable amount of time entering them, can be very demoralizing. We should get into the habit of saving work every 15 minutes or so (and also making a backup).

Data entry

Once the variables have been defined, the data can be entered. If we switch to the **Data Editor** window and select the **Data View** page we will see that the first four columns are headed sequentially with the variable names we have just defined (Figure 2.17). The first column in the **Data View** corresponds with the first row in the **Variable View**, the second column in the **Data View** corresponds with the second row in the **Variable View**, etc.



The screenshot shows the SPSS Data Editor window titled 'Ch02 - SPSS Data Editor'. The menu bar includes File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Window, and Help. Below the menu bar is a toolbar with various icons. The main area shows a data grid with the following structure:

	id	sex	health	age	var	var
1	4	2	3	17		

Figure 2.17 The **Data View** page with defined variables

We can now enter the data into each of these columns. To see how this is done, we will enter the data for the student with ID number 1, who is male, very healthy, and 17 years old. We click on the top-left cell below **id** so that it is active and type **1**, which is this student's ID number, and then the tab-key. The cursor will jump across to the cell below **sex** where we type **2**, which is the code value for male. We tab across to the next column and type **3** for 'Very healthy', and then tab across to enter 17 as under **age**.

SPSS will either display the value labels or the values on the data page, and we can switch between these options by selecting **View** from the menu and then **Value Labels** (Figure 2.18).

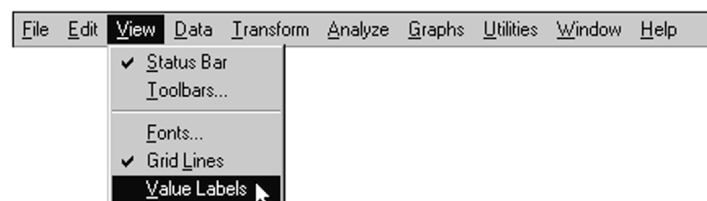


Figure 2.18 The **Value Labels** command

The advantage of viewing data as labels rather than numeric codes is that when we enter data that have been assigned value labels into a cell, we are presented with a contextual menu from which we can select the appropriate response. For example, to enter the fact that Case 1 is male, we can either type 2, or click on the down-arrow in the cell and scroll to Male in the list.

You may notice that, as you type, the data initially appear on the bar just above the data fields, called the **Cell Editor** (Figure 2.19).



Figure 2.19 The Cell Editor

This is where information is entered until you hit the **tab** key, which then enters the data into the active cell. The ‘address’ of the active cell is indicated on the left of the **Cell** editor. This address is defined by the combination of the row number and column name that intersect at the active cell. If at any point we make a mistake, or we need to change the information in any particular cell, we simply make that cell active and type in the new information. On hitting the **return** key the new value will replace the old.

For your convenience, the data for the hypothetical survey of 200 students have already been entered into SPSS and saved in the file **Ch02.sav** on the CD that comes with this book. To open a data file we select **File/Open** and then select the appropriate directory and filename in the dialog box. The files that are used in this and later chapters can be selected by highlighting the appropriate file once the CD has been selected as the location where the files reside. If you open the **Ch.02.sav** file the **Data View** page will look like Figure 2.20. It is important to note that, unlike most other computer programs you may be familiar with, SPSS will not allow more than one data file to be open at any one time. Thus when you choose to open another file, the currently active file will be closed. SPSS will prompt you to save your data before it closes a file, but this limitation should be borne in mind, especially when copying data to and from files.

Figure 2.20 The SPSS Data Editor after data entry

Checking for incorrect values: Data cleaning

Whether we enter data into SPSS ourselves or receive data from other people, before we actually analyze the data, we need to check that the data are 'clean'. Clean data are data that do not contain any invalid or nonsensical values e.g. we don't have someone with an age of 234 years, or someone coded 3 for sex, when we have coded 'females = 1' and 'males = 2'.

We check data to see if they need cleaning by generating frequency tables on all the variables (see Chapter 4). Frequency tables allow us to check to see that there are no values that are outside the permissible range. We can also use more elaborate analysis such as crosstabulations to assess whether particular groups within the data set only have values that are valid for them.

If we discover unusual responses during the data cleaning process, we make either of the following changes to the data set:

- *if it is a data entry mistake*, go back to the original data gathering instrument for that case, by cross-referencing the id number, and find the appropriate value to enter; or
- *if it is an invalid response*, type in an assigned missing value.

Summary

We have worked through the process of setting up an SPSS data file. Needless to say we have only skimmed the surface with respect to the full range of options available. I leave it to the reader to play around with SPSS and learn the full range of features that it provides. The program comes with its own tutorial and sample data files that will guide you through the various features. The program also comes with a very useful help facility, which is available from the menu bar. In addition, the CD that accompanies this book includes a number of extra chapters that illustrate some of the more advanced functions available in SPSS. These include the **Recode** command, the **Multiple Response** command, the **Compute** command, and the **Select Cases** command.

One last point needs to be made about working with SPSS. Many of the default settings for entering, presenting, and analyzing data in SPSS are less than ideal. For example, displaying numeric data by default to 2 decimal places is annoying when most data are entered in whole numbers. This does not actually affect any results, but makes the **Data View** page unnecessarily cluttered. The **Decimals** setting on the **Variable View** page can be changed whenever we work with a new file as we described above. An alternative is to use the **Edit/Options** (Windows) or the **SPSS/Preferences** (Macintosh) command to change the default setting once and for all, so that when data are entered they appear with 0 decimal places (i.e. whole numbers). The settings that can be changed under this command are far too numerous to even list here; you should explore these at your leisure, but with one word of warning: do not alter the default setting on someone else's version of SPSS without their permission.

Exercises

2.1 A survey gathers the data for the weekly income of 20 people, and obtains the following results:

\$0, \$0, \$250, \$300, \$360, \$375, \$400, \$400, \$400, \$420, \$425, \$450, \$462, \$470, \$475, \$502, \$520, \$560, \$700, \$1020

Create an SPSS data file and enter these data, entering all the necessary labels. Save the file with an appropriate filename.

- 2.2** The following data represent time, in minutes, taken for subjects in a fitness trial to complete a certain exercise task.

31	39	45	26	23	56	45	80	35	37
25	42	32	58	80	71	19	16	56	21
34	36	10	38	12	48	38	37	39	42
27	39	17	31	56	28	40	82	27	37

Each subject's heart rate is also recorded in the same sequence as their time score:

63	89	75	80	74	65	90	85	92	84
74	79	98	91	87	76	82	90	93	77
74	89	85	91	102	69	87	96	83	72
92	88	85	68	78	73	86	85	92	90

The first 20 of these scores (reading from left to right) are taken from males and the second 20 from females. Create an SPSS data file and enter these data, entering all the necessary labels. Save the file with an appropriate filename.

- 2.3** A research project has collected data from 10 people on the following variables:

Television watched per night (in minutes)	Main channel watched	Satisfaction with quality of programs
170	Commercial	Very satisfied
140	Public/government	Satisfied
280	Public/government	Satisfied
65	Commercial	Very satisfied
180	Commercial	Not satisfied
60	Commercial	Not satisfied
150	Public/government	Satisfied
160	Commercial	Not satisfied
200	Public/government	Satisfied
120	Commercial	Not satisfied

Prepare an SPSS data file for these data, creating variables and variable labels, values and value labels.

3

The graphical description of data

The most striking method of summarizing a distribution is often a **graph**. A graph (sometimes called a **chart**) provides a quick visual sense of the main features of a distribution. This chapter will begin with univariate graphs, which are used to describe the distribution of a single variable. It will then discuss more complex graphs in which two or more variables can be displayed. The graph that can be constructed in any given context is determined largely by whether the variable is discrete (usually measured on a nominal or ordinal scale) or continuous (usually measured on an interval/ratio scale), and the number of variables to be described by the graph, as indicated in Table 3.1.

Table 3.1 Graphs by type and number of variables

Type of variable	Univariate graph	Bivariate/multivariate graph
Discrete	Pie graph Bar graph	Clustered pie graph Clustered bar graph Stacked bar graph
Continuous	Histogram Polygon	Scatterplot

Some general principles

In this chapter, to illustrate the procedures and principles involved in constructing graphs, we will use the hypothetical student survey introduced in Chapter 2 (contained in the SPSS file **Ch03.sav** on the CD that comes with this book).

Before we use SPSS to create graphs, we need to be aware of some general rules that apply to their construction. Most importantly a graph should be a *self-contained* bundle of information. A reader should not have to search through the text in order to understand a graph (if one does have to search the text this may be a sign that the graph is concealing information rather than illuminating it). In order for a graph to be a self-contained description of the data we need to:

- give the graph an **appropriate title**;
- clearly identify the **categories or values** of the variable;
- indicate, for interval/ratio data, the **units of measurement**;
- indicate the **total number of cases**;
- indicate the **source of the data**.

SPSS provides two alternative means for generating the same types of graphs, both of which are accessed from the **Graphs** option on the command menu (Figure 3.1):

1. A range of chart types is listed on the menu that immediately drops down when **Graphs** is selected. In Version 12 of SPSS, the construction and format of these graphs changed substantially and, in my opinion, not for the better (a guide to generating graphs on earlier versions of SPSS is contained on the CD, called **GraphsVersion10-11.pdf**).
2. The alternative means for creating graphs, and the one we will detail in this chapter, appears when we scroll down to the **Interactive** option on the **Graphs** menu.

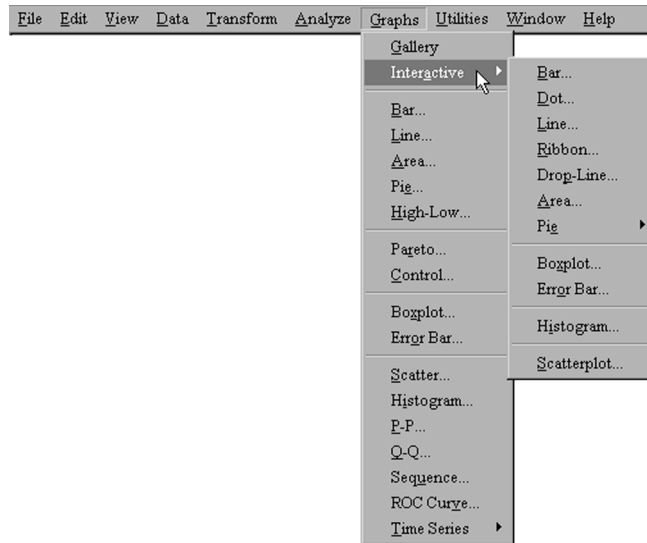


Figure 3.1 The SPSS Graphs command

Pie graphs

The simplest graph we can construct for data organized into discrete categories is a pie graph.

A **pie graph** presents the distribution of cases in the form of a circle. The relative size of each slice of the pie is equal to the proportion of cases within the category represented by the slice.

To generate a pie graph on SPSS we select **Graphs/Interactive/Pie/Simple** from the menu. The **Create Simple Pie Chart** dialog box appears, shown in Figure 3.2.

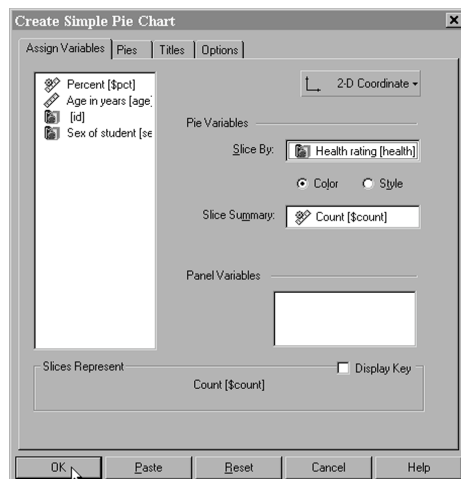


Figure 3.2 The Create Simple Pie Chart dialog box

To generate a pie chart for the health rating of the students in my survey, I select the following options from this dialog box (note, however, that it is not possible to detail all the options available under this command; you are encouraged to explore them at your leisure).

- Select the variable we want to analyze by dragging it from the **source list** on the left to the **Slice By:** area on the right. In this instance we drag the **Health rating** label across.
- Determine what the slices of the pie will represent in the **Slice Summary:** area. The default setting is for the slices to represent the *number* of cases in each category, as indicated by the **Count [Scout]** label. We could replace this with the **Percent [Spct]** label that is listed in the source list, but with a pie chart this would make no difference; the slices of the pie will look exactly the same. We therefore leave **Count [Scout]** as the slice summary.
- Vary the format of the pie graph and the labels attached to each slice of the pie by selecting the **Pies** sub-command. Under **ScaleLabels** select **Count** and **Percent**, so that the number and percentage of cases in each category of health rating is displayed.
- Add titles to the chart by selecting the **Titles** sub-command. Three levels of heading are available; we will type 'Health rating of students' as the **Chart Title:** that will appear at the top of the graph, and 'Source: Hypothetical student survey, 2004' as the **Caption:** to appear at the bottom of the graph.
- Vary other features of the graph by selecting the **Options** sub-command. Here we will not alter any of the default settings.

The output shown in Figure 3.3 will be generated when we click the **OK** button. Note that the total is less than 200 as 22 students did not provide valid responses for this variable.

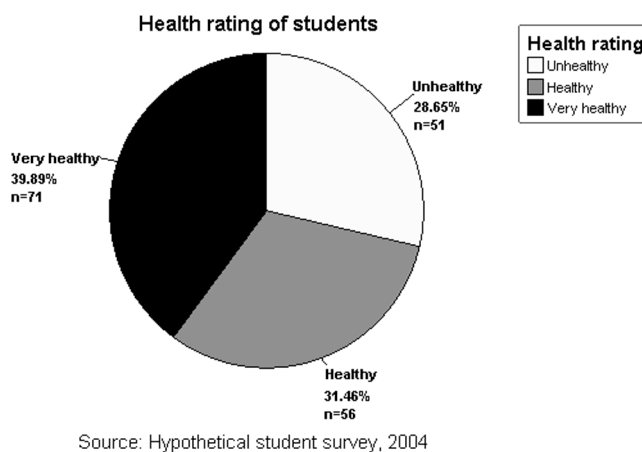


Figure 3.3 SPSS Interactive Pie chart

Once we generate a graph, we can alter its appearance and content. For example, we may not be happy with the title we have specified, or the font used to display labels, or would prefer to have each of the categories of health appear with the slice labels rather than in the boxed legend on the right of the chart. To edit a graph in these and other ways, we click on the graph in the **Viewer** window and select **Edit/SPSS Interactive Graphic Object/Edit** from the menu at the top (we can also simply double-click on the graph). The chart is now editable, as indicated by the cross-hatched border around its edge and the new button-bar that appears around the chart.

Once in 'edit-mode' we have a range of options for improving the quality of our graph. There are, as before, too many options to work through here, and you should explore these at your leisure. As a general rule, though, if you want to change a particular element of a chart, you can click the right mouse-button on it and a contextual menu will appear on the screen from which you select an appropriate option. For example, if I wish to change the font size of

the chart tile I right-click on **Health rating of students** and select **Text...** A box will appear from which I can select an alternative font face, style, size, and color, and the effect these changes will have will be displayed in the **Sample:** box.

Now that we have the pie chart what does it tell us about the distribution of the variable we are investigating? *Pie graphs emphasize the relative importance of a particular category to the total. They are therefore mainly used to highlight distributions where cases are concentrated in only one or two categories.* In this instance we can immediately see the high proportion of cases that rated themselves as ‘Very healthy’.

Pie graphs begin to look a bit clumsy when there are too many categories for the variable (about six or more categories, as a rule of thumb, is usually too many slices for a pie graph). More than five slices to the pie will cause the chart to look a little cluttered.

A pie chart is actually a specific case of a more general type of graph. Any shape can be used to represent the total number of cases, and the area within it divided up to show the relative number of cases in various categories. For example, the United Nations has used a champagne glass to illustrate the distribution of world income (Figure 3.4). Rather than using a simple circle, the metaphor of the champagne glass, a symbol of wealth, makes this graph a powerful illustrative device for showing the inequality of world income.

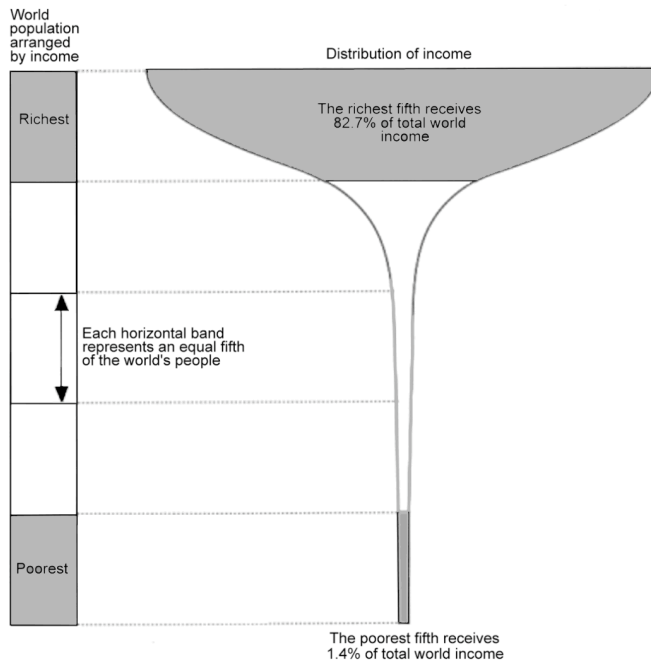


Figure 3.4 World distribution of income

Source: UNDP, 1992, *Human Development Report 1992*, Oxford: Oxford University Press.

Bar graphs

A **bar graph** is another type of chart that can be produced from the same set of data as that used to generate a pie chart. While these two types of graph can be generated from exactly the same set of data, they describe different aspects of the distribution for those data. We have noted that a pie graph emphasizes the relative contribution of the number of cases in each category *to the total*. On the other hand, *bar graphs emphasize the frequency of cases in each category relative to each other*.

To generate an interactive bar graph on SPSS we choose the **Graphs/Interactive/Bar** command from the menu bar. This brings up the **Create Bar Charts** dialog box (Figure 3.5).

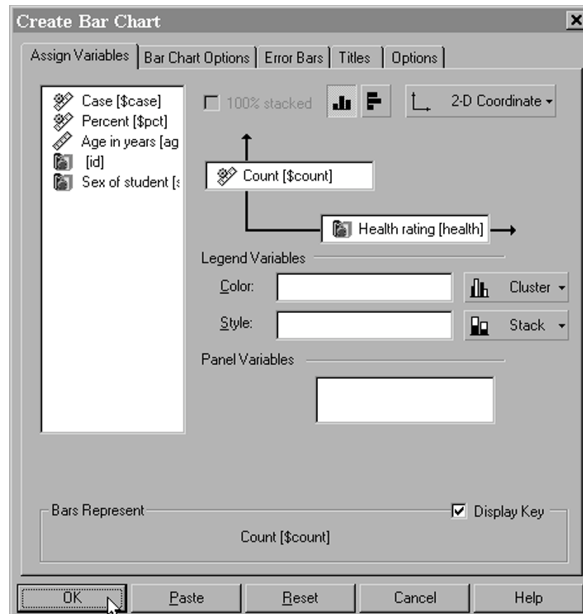


Figure 3.5 The SPSS **Create Bar Chart** dialog box

A bar graph has two sides or **axes** and we specify the information displayed by these axes in the section of the dialog box that has the two intersecting lines with arrow-heads.

- Along one axis of the graph are the *categories* of the variable. This axis is called the **abscissa**, and is usually the horizontal base of the graph. We can drag the **Health rating** label from the source list into the empty box on the horizontal arrow.
- Along the other axis are the *frequencies*, expressed either as the raw count or as percentages of the total number of cases. This axis is known as the **ordinate**. This is usually the left vertical axis, and SPSS sets the count of cases in each category as the default setting, as indicated by the **Count [\$count]** label in the box on the vertical arrow.
- Note that we can ‘flip’ the graph around by clicking the button with horizontal bars that sits just above the two arrows. This will shift the abscissa to the vertical axis and the ordinate to the horizontal. Here, though, we will use the default setting with vertical bars.
- As with pie graphs, we can select the **Titles** sub-command to add relevant information.
- It is worth exploring the other sub-commands that are available when creating a bar chart. One we can use here is in the **Bar Chart Options** sub-command, which allows us to include the count of cases *in each category* represented by the bars of the graph.

The bar graph that results from these choices is presented in Figure 3.6. As with the pie graph we generated above we can ‘tidy’ up this bar graph by double-clicking on the graph (we can alternately right-click on the chart and select **Edit**). Again I emphasize that SPSS provides many options for improving the layout and appearance of a graph and the amount of information it contains. Often these choices (such as the wording and placement of titles) are matters of aesthetic judgment so you need not follow my preferences exactly. Play around with the options so that you produce something that you like, but always keep in mind the general rules of graph construction that we discussed above.

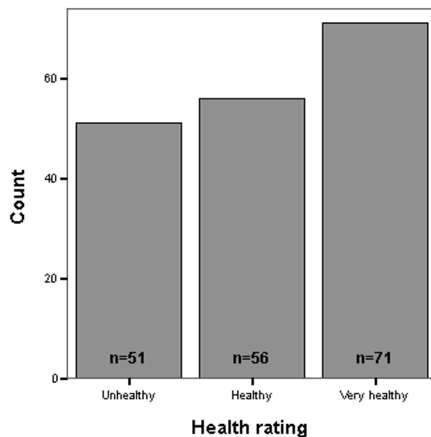


Figure 3.6 An SPSS Interactive Bar chart

Histograms and polygons

Bar graphs constructed for discrete variables always have gaps between each of the bars: there is no gradation between male and female, for example. A person's age, on the other hand, is a continuous variable, in the sense that it progressively increases. Even though we have chosen discrete values for age (whole years) to organize the data, the variable itself actually increases in a continuous way (as we discussed in Chapter 1). As a result, the bars on a **histogram**, unlike a bar graph, are 'pushed together' so they touch. The individual values (or class mid-points if we are working with class intervals) are displayed along the horizontal, and a rectangle is erected over each point. The width of each rectangle extends half the distance to the values on either side. *The area of each rectangle is proportional to the frequency of the class in the overall distribution.*

Using the example of the age distribution of students we can construct the following histogram. We select **Graphs/Interactive/Histogram** from the SPSS menu. This brings up the **Create Histogram** dialog box (Figure 3.7). This dialog box is very similar to the bar chart command.

- We select the relevant variable from the source list and transfer it to the empty box on the horizontal arrow representing the abscissa of the graph.
- The default setting is for the count of cases with each value of the desired variable to be displayed, as indicated by the **Count [\$count]** label on the vertical arrow.
- We add titles and footnote captions through the **Titles** sub-command.
- The **Histogram** sub-command gives us some control over the size of the age groups into which cases will be grouped in the histogram. The default setting is for SPSS to determine these intervals automatically. In this instance we want SPSS *not* to collect cases into broader age groups. We therefore deselect the **Set interval size automatically** option under **Interval Size**, and select instead **Width of intervals:** and type in **1**. This will ensure that for this histogram we get a separate bar for each age value.

Selecting **Age in years** and clicking on **OK** will produce the histogram in Figure 3.8. We can see that the area of the rectangle over 20 years of age (which is the tallest bar immediately before the 20 tick-mark on the abscissa), as a proportion of the total area under the histogram, is equal to the proportion of all cases having this age.

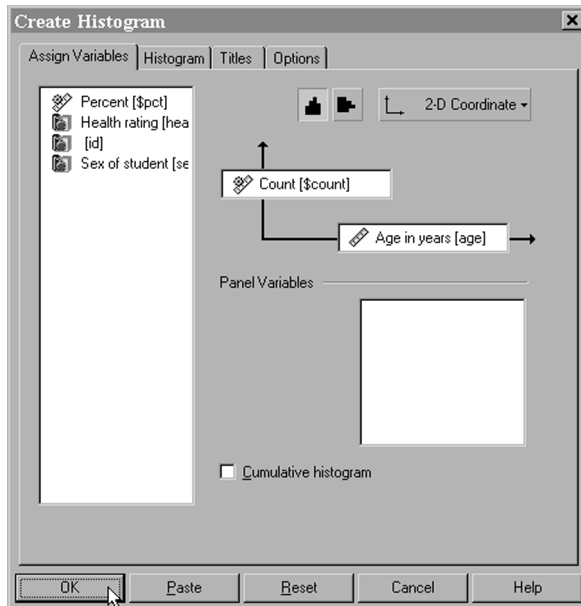


Figure 3.7 The SPSS Create Histogram dialog box

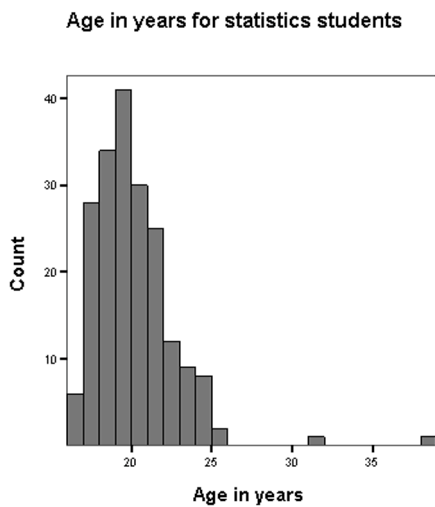


Figure 3.8 An SPSS Interactive Histogram

When working with continuous interval/ratio data, we can also represent the distribution in the form of a **frequency polygon**.

A **frequency polygon** is a continuous line formed by plotting the values or class mid-points in a distribution against the frequency for each value or class.

If we place a dot on the top and center of each bar in the histogram, and connect the dots, we produce a frequency polygon, such as the SPSS line graph in Figure 3.9.

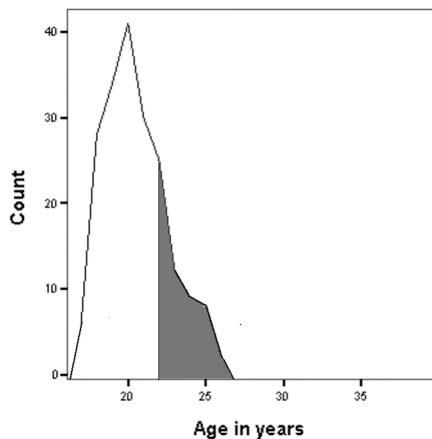


Figure 3.9 An SPSS Interactive line graph

I have made some slight changes to the SPSS graph. First, notice that the polygon begins and ends on a frequency of zero. To explain why we ‘close’ the polygon in this way we can think of this distribution as including the ages 16 and 27, even though there were no students with these ages. Since there are no such respondents with either of these ages the line at these values is on the abscissa, indicating a frequency of zero. Notice also, that I have excluded the ‘outliers’ that were present in the histogram, since the line will simply run along the abscissa until it reaches these values.

One aspect of frequency polygons (and histograms) that will be of utmost importance in later chapters needs to be pointed out, even though its relevance may not be immediately obvious. A polygon is constructed in such a way that *the area under the curve between any two points on the horizontal, as a proportion of the total area, is equal to the proportion of valid cases in the distribution that have that range of values*. Thus the shaded area in Figure 3.9 (which I have added), as a percentage of the total area under the curve, is equal to the percentage of cases that are aged 22 or above. This is 56/197, or 30% of all valid cases.

Another way of looking at this is to say that the **probability** of randomly selecting a student aged 22 years or more from this group is equal to the proportion of the total area under the curve that is shaded. This probability is 0.30, or a nearly 1-in-3 chance of selecting a student in this age group.

Interpreting a univariate distribution

We have spent some time working through the technical, and sometimes tedious, process of generating graphs for single variables in SPSS. We will now stop and discuss the more important issue of what sense we make of a graph that we generate; how do we **interpret** it? When we look at a histogram, such as that in Figure 3.9 for example, we generally try to identify four aspects of the distribution:

- Shape
- Center
- Spread
- The existence of outliers

There are certain common **shapes** that appear in research. For example, Figure 3.10 illustrates the bell-shaped (symmetric) curve, the J-shaped curve, and the U-shaped curve.

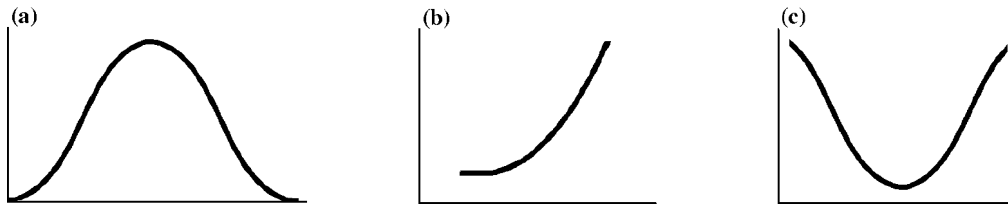


Figure 3.10 Three polygon shapes: (a) bell-shaped; (b) J-shaped; and (c) U-shaped

The bell-shaped curve is one we will explore in much greater length in later chapters. For distributions that have this ‘bell shape’, such as our distribution for age, we also describe two further aspects of its shape. The first is the **skewness** of the distribution. If the curve has a long tail to the left, it is **negatively** (or **left**) skewed. In Figure 3.8 the age distribution of students has a long tail to the right, which means it is **positively** (or **right**) skewed. The second aspect of a ‘bell-shaped’ curve is **kurtosis**. Kurtosis refers to the degree of ‘peakedness’ or ‘flatness’ of the curve. We can see in Figure 3.11(a) a curve with most cases clustered closed together. Such a curve is referred to as **leptokurtic**, while curve (b) has a wide distribution of scores, which ‘fatten’ the tails, and is referred to as **platykurtic**. Curve (c) lies somewhere in between, and is called **mesokurtic**.

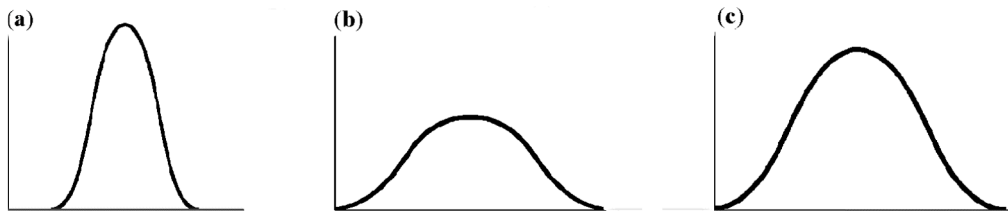


Figure 3.11 A (a) leptokurtic, (b) platykurtic, and (c) mesokurtic curve

Looking at the polygon for age of students in Figure 3.8 we can see that it is leptokurtic.

To gauge the **center** of a distribution, imagine that the bars of the histogram for the age distribution of students are lead weights sitting on a balance beam. Where would we have to locate a balance point along the bottom edge of the graph to prevent it from tipping either to the left or right? To the naked eye this point is probably around 20 years of age. This, in a loose fashion, identifies the average or typical age.

We can also observe that the students represented by Figure 3.8 are fairly tightly clustered around the central point. In other words, most students do not deviate much from the average. The **spread** of scores is not very wide. We also note, however, the existence of two **outliers**, that are not just at the upper or lower end of the tails, but are disconnected from the rest of the group in terms of age: one is 32 years old and the other is 39 years old.

We have just used a graph to help us identify key features of a particular distribution, namely its shape, center, spread, and whether it includes any outliers. Later chapters will present more precise numerical measures for identifying these features of a distribution, but as we have just observed, sometimes a well-constructed graph can do this task almost as effectively, and in a visually striking form.

Graphing two variables

The previous sections discussed the means of displaying the distribution of a single variable, and we used as an example the health rating of students. Assume that we believe the way in which students rate their health is affected by their respective sex; it is not uncommon for men

to rate their own health more highly than do woman. Here we are interested in whether there is a **relationship** between two variables: sex and health rating. Graphs can give us a very immediate sense as to whether such a relationship exists between two variables. It may help at this point to read again the section in Chapter 1 on the general conceptual issues involved in analyzing the relationship between variables. In particular, we need to decide which is the **independent** variable and which is the **dependent** variable in the relationship. *The independent variable generally forms the groups to be compared*; it is the variable we believe somehow affects or determines the state of the other variable. Here we clearly suspect that a student's sex affects how they rate their own health, so that the sex of students is the independent variable and health rating is the dependent variable. Thus we want to compare how male and female students differ in terms of their health rating.

When generating each of the graphs we discussed above in SPSS, you may have noticed an area of each dialog box labelled **Panel Variables**: We drag the independent variable that will form the groups we wish to compare into this area of the dialog box. For example, if we wanted to compare separate pie charts for males and females, we move **Sex of student** from the source list to the **Panel Variables**: list in the **Create Simple Pie Chart** dialog box. The effect will be to present two pie charts in separate panels, as shown in Figure 3.12. We can immediately see that indeed males do rate their own health more highly than female students.

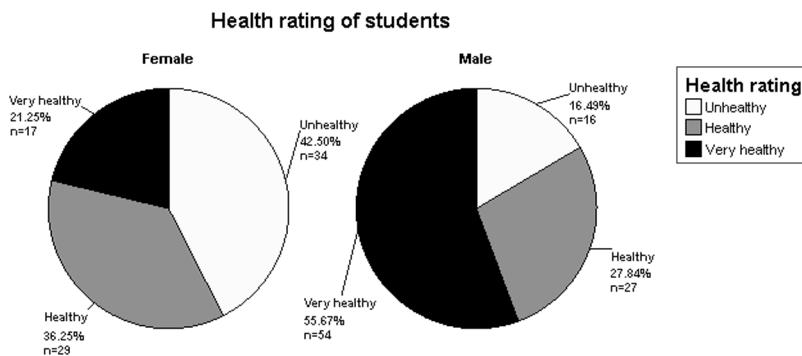


Figure 3.12 SPSS Interactive Pie graphs of Health rating with Sex of student as a panel variable

The **Panel Variables** option is available for all graph types above, so that we can generate separate graphs of any type based on the categories of the independent variable. For bar graphs, there are, in addition to the **Panel Variables** option, alternative ways of illustrating the relationship between two variables. One is to stack the bars on top of each other.

A **stacked bar graph** (sometimes called a **component bar graph**) layers in a single bar the number (or proportion) of cases in each category of a distribution.

Each bar is divided into layers, with the area of each layer proportional to the frequency of the category it represents. It is very similar in this respect to a pie chart, but using a rectangle rather than a circle.

To generate a stacked bar graph in SPSS we use the **Graphs/Interactive/Bar** command. In this example:

- we place **Sex of student** into the area along the horizontal arrow, since we want to display separate bars for males and females.

- we place **Health rating** in either the **Color:** or **Style:** areas and select **Stack** from the adjacent drop-down menu. This instructs SPSS to stack on top of each other the bars representing each of the categories of health rating.
- we click on the **100% stacked** check-box, so that the two stacked bar charts are both presented in percentage terms. This will compensate for the fact that there are different numbers of men and women in the survey, making comparisons based on counts difficult.

The effect of these commands is displayed in Figure 3.13. We can clearly see, as with the pie graphs in Figure 3.12, that male students rate their own health more highly than females.

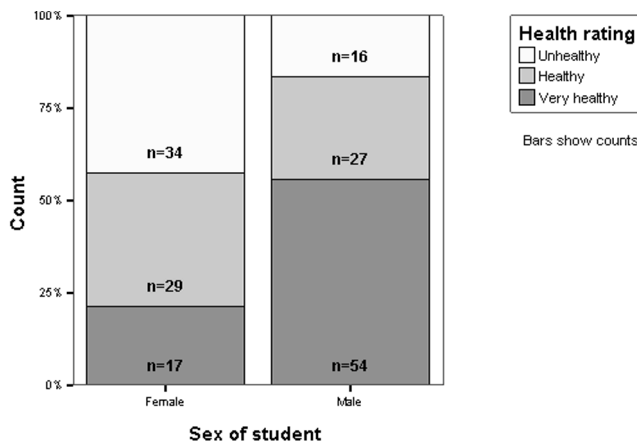


Figure 3.13 An SPSS Interactive Stacked Bar graph

Stacked bar graphs have the same limitations as pie graphs, especially when there are too many values stacked on top of each other. Figure 3.13 contains only three values stacked on top of each other to form each bar, but when there are more than five categories, and some of the layers are very small, the extra detail hinders the comparison of distributions that we want to make, rather than illuminating it.

You may have noticed that an alternative called **Cluster** was available when we selected **Stack** in the dialog box. A **clustered bar graph** is an alternative to a stacked bar chart in that it places the bars for each category side by side along the horizontal.

A **clustered bar graph** displays, for each category of one variable, the distribution of cases across the categories of another variable.

A clustered bar graph is particularly useful when we want to observe *how the frequency of cases in individual categories of the dependent variable differs across the comparison groups*. If we choose **Cluster** rather than **Stack** for the **Legend Variables** in the **Create Bar Chart** command, the bar chart in Figure 3.14 will appear.

There is one further graph type for displaying the relationship between two variables we could discuss at this point. This is a scatter plot, which is a means of displaying the relationship between two variables that are each measured on interval/ratio scales with many points on each scale. We will, however, leave the discussion of this graph type until Chapter 12, since scatter plots are rarely generated for their own sake, but are usually created as a preliminary to more elaborate statistical procedures covered in that Chapter. The reader is welcome to skip ahead to the start of Chapter 12 if they do wish to complete their understanding of graphs at this point.

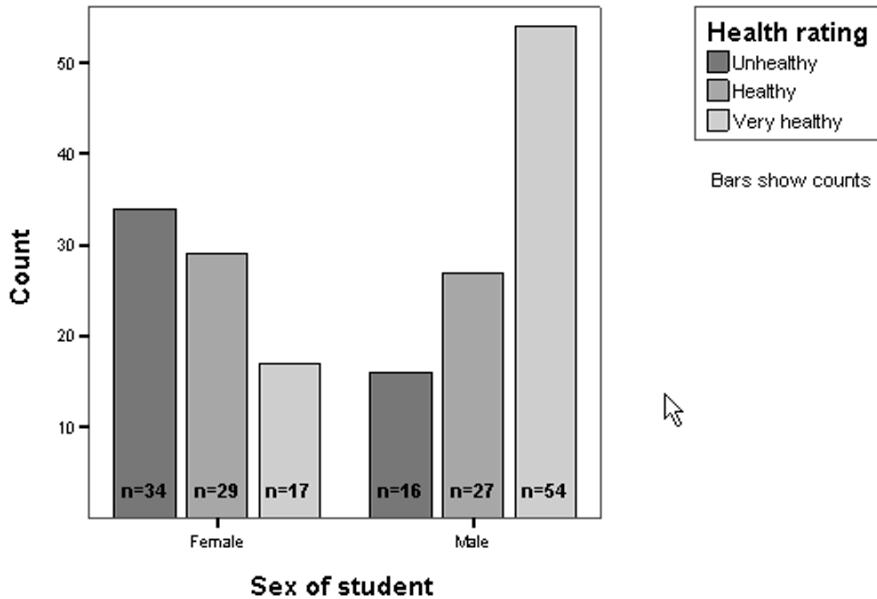


Figure 3.14 An SPSS Interactive Clustered Bar graph

Common problems and misuses of graphs

Unfortunately graphs lend themselves to considerable misuse. Practically every day in newspapers we can find one (if not all) of the following ‘tricks’, which can give a misleading impression of the data. As an exercise you may wish to flick through a copy of your local daily paper and tick off the number of following tricks that you find. The starting point for those wishing to look at this issue further is Darrell Huff’s cheap, readable, and entertaining classic, *How to Lie with Statistics*, New York: W.W. Norton, various printings.

Relative size of axes

The same data can give different graphical ‘pictures’ depending on the relative sizes of the two axes. By stretching either the abscissa or the ordinate the graph can be ‘flattened’ or ‘peaked’ depending on the impression that we might want to convey.

Consider for example the data represented in Figure 3.15. This Figure provides two alternative ways of presenting the same data. The data are for a hypothetical example of the percentage of respondents who gave a certain answer to a survey item over a number of years. Graph (a) has the ordinate (the vertical) ‘compressed’ relative to graph (b); similarly graph (b) has the abscissa (the horizontal) relatively more ‘compressed’. The effect is to make the data seem much smoother in graph (a) – the rise and fall in responses is not so dramatic. In graph (b) on the other hand, there is a very sharp rise and then fall, making the change appear dramatic. Yet the two pictures describe the same data!

Which of the two alternatives in Figure 3.15 is correct and which represents a distortion? There are no hard and fast rules about the relative size of the two axes in a graph, although there are some common conventions. In order to avoid the distortion in Figure 3.15 the convention in research is to construct graphs, wherever possible, such that *the vertical axis is around two-thirds to three-quarters the length of the horizontal*.

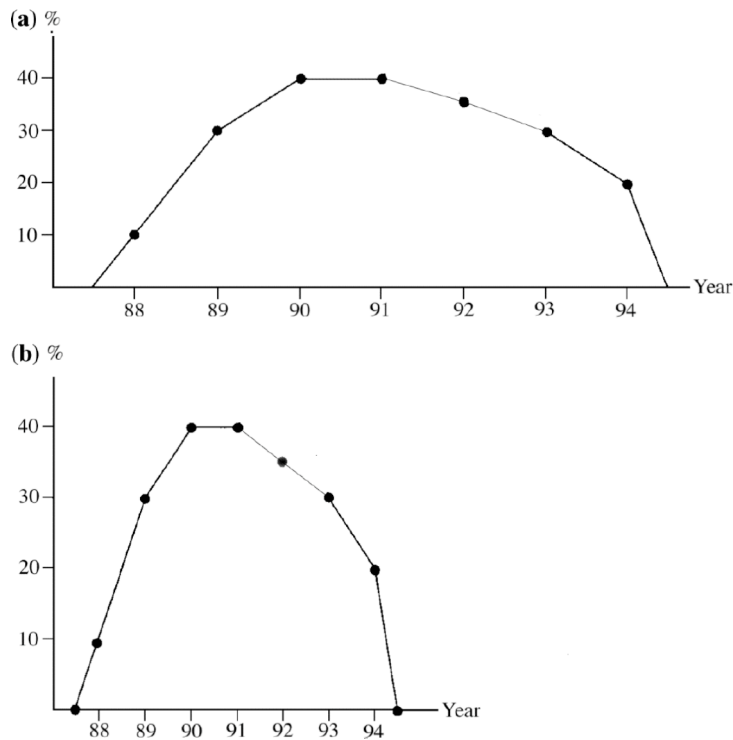


Figure 3.15 Two ways of presenting data

Truncation of the ordinate

A similar effect to the one just described can be achieved by cutting out a section of the ordinate. Consider the data on the relative pay of women to men in Australia between 1950 and 1975, shown in Figure 3.16. Graph (a) is complete in the sense that the vertical axis goes from 0 through to 100 percent. The vertical axis in graph (b) though has been truncated: a large section of the scale between 0 percent and 50 percent has been removed and replaced by a squiggly line to indicate the truncation. The effect is obvious: the improvement in women’s relative pay after 1965 seems very dramatic, as opposed to the slight increase reflected in graph (a). If we wanted to emphasize continued inequality graph (a) will serve our purposes, whereas if we wanted to emphasize increasing equality, graph (b) will illustrate this point very dramatically.

Selection of the start and end points of the abscissa

This is especially relevant with data that describe a pattern of change over time. Varying the date at which to begin the data series and to end it can give different impressions. For example, we may begin in a year in which the results were unusually low, thereby giving the impression of increase over subsequent years, and vice versa if we choose to begin with a year in which results ‘peaked’. Consider, for example, the time series of data over a 15 year period shown in Figure 3.17. If we wanted to emphasize steady growth we could begin the graph where the dashed line is displayed, which would completely transform the image presented by the graph.

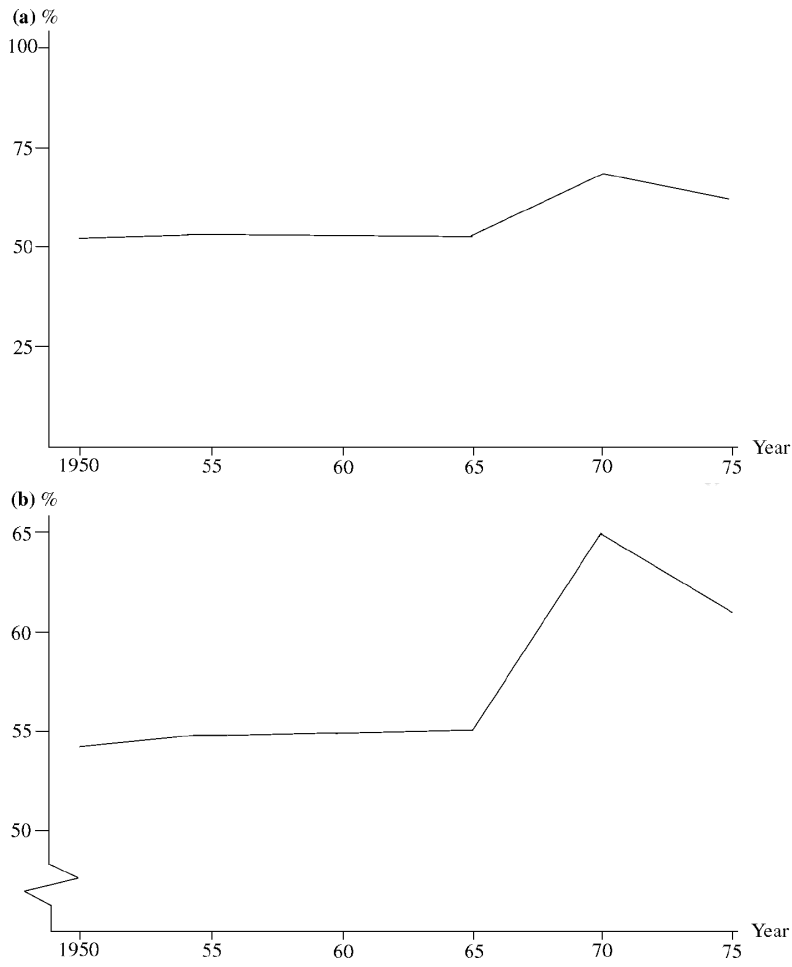


Figure 3.16 Women's income as a percentage of men's income, 1950–1975

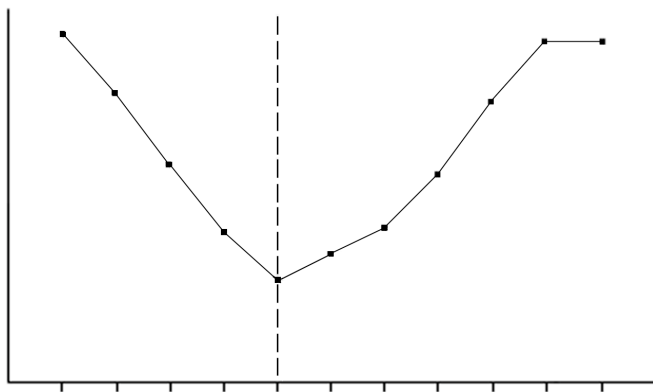


Figure 3.17 The choice of start and end points of the abscissa

Use of 3-dimensional shapes

One of the most misused techniques when presenting graphs is to present the figure representing each category as a 3-dimensional shape. With pie charts the rotation of such 3-D shapes about the horizontal axis (as if the bottom of the pie is being tilted towards the reader) can make the slice of the pie at the bottom of the page look relatively larger than the frequency it seeks to represent. Similarly, 3-D bars in a bar graph can distort the relative frequencies by allowing some bars to be partially obscured by others. By presenting relatively more of the shape for one bar and less of those that are hidden, some categories can be made to seem larger than is warranted by the actual distribution of scores. For example, Figure 3.18 presents a 3-D bar graph for Health rating, which you can compare to Figure 3.6.

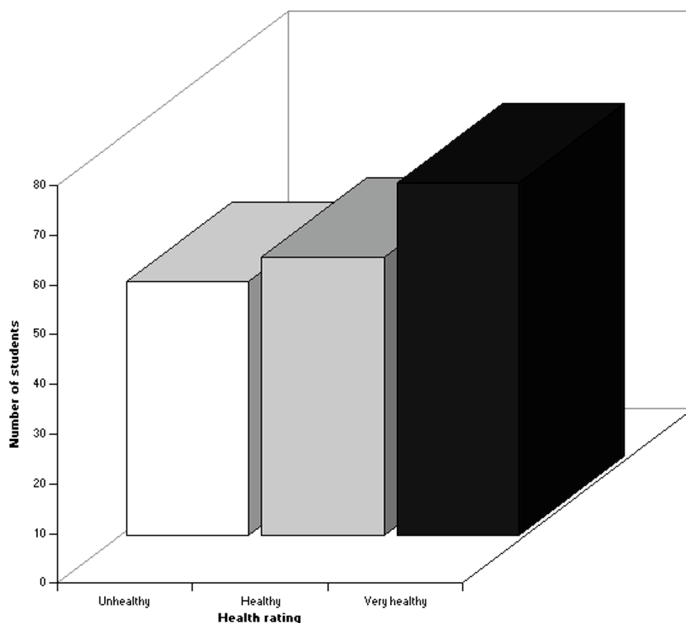


Figure 3.18 A 3-dimensional bar graph of Health rating

By ‘hiding’ large parts of the bars for the other two categories, we can make the Very healthy group appear to be more prevalent in the distribution than the proportion they actually represent. For these reasons, despite the ‘wow-factor’ computer graphics can create when used to create elaborate graphs, we should always bear in mind the principle behind graphs, which is that they should accurately present the relative size of each category in terms of the frequency of scores in each.

Exercises

- 3.1** Which aspects of a distribution do pie graphs emphasize and which aspects do bar graphs emphasize?
- 3.2** Explain the difference between a bar graph and a histogram.
- 3.3** Survey your own statistics class in terms of the variables age, sex, and health rating. Use the graphing techniques outlined in this chapter to describe your results.

- 3.4** Enter the following data into SPSS (time, in minutes, taken for subjects in a fitness trial to complete a certain exercise task):

31	39	45	26	23	56	45	80	35	37
25	42	32	58	80	71	19	16	56	21
34	36	10	38	12	48	38	37	39	42
27	39	17	31	56	28	40	82	27	37

Using SPSS select an appropriate graphing technique to illustrate the distribution. Justify your choice of technique against the other available options.

- 3.5** Consider the following list of prices, in whole dollars, for 20 used cars:

11,300	9200	8200	8600	10,600
11,100	7980	12,900	10,750	9200
13,630	9400	11,800	10,200	12,240
11,670	10,000	11,250	12,750	12,990

From these data construct a histogram using these class intervals:

7000–8499, 8500–9999, 10000–11499, 11500–12999, 13000–14499.

- 3.6** Construct a pie graph to describe the following data:

Migrants in local area, place of origin

Place	Number
Asia	900
Africa	1200
Europe	2100
South America	1500
Other	300
Total	6000

What feature of this distribution does your pie graph mainly illustrate?

- 3.7** From a recent newspaper or magazine find examples of the use of graphs. Do these examples follow the rules outlined in this chapter?
- 3.8** Use the **Employee data** file to answer the following problems with the aid of SPSS.
- I want to emphasize the high proportion of all cases that have clerical positions. Which graph should I generate and why? Generate this graph using SPSS, and add necessary titles and notes.
 - Use a stacked bar graph to show the number of women and the number of men employed in each employment category. What does this indicate about the sexual division of labor in this company?
 - Generate a histogram to display the distribution of scores for current salary. How would you describe this distribution in terms of skewness?