PART 2

UNIVARIATE ANALYSIS

Univariate Statistics 3

Contents

The first step on the path to understanding a data set is to look at each variable, one at a time, using **univariate statistics**. Even if you plan to take your analysis further to explore the linkages, or relationships, between two or more of your variables you initially need to look very carefully at the distribution of each variable on its own.

This chapter sets out to give you an understanding of how to:

- Start exploring data using simple proportions, frequencies and ratios
- Code data for computer analysis
- Group the categories of a variable for more convenient analysis
- Use SPSS to create frequency tables which contain percentages
- Understand the difference between individual and household levels of analysis.

Frequency distributions

One of the first things you might want to do with data is to count the number of occurrences that fall into each category of each variable. This provides you with **frequency distributions**, allowing you to compare information between groups of individuals. They allow you to answer questions like, 'how many married people are there in the data' and to calculate 'what percentage of people think that it is safe to walk around in their neighbourhood after dark'. They also allow you to see what are the highest and lowest values and the value around which most scores cluster.

For instance, you might be interested in the take-up of science and arts/social science subjects at A (advanced) level in a particular sixth-form college. After asking each boy what subjects he is studying at A level you could divide the boys into those taking mainly science subjects and those taking mainly arts/social science subjects.

It would be clearer if we counted up the number of boys in each category. This would give the **frequency** of occurrence in each category (see Exhibit 3.1). There are 26 boys studying science and 17 studying arts/social science at this college. We might be interested in comparing these numbers with the girls' choice of subjects. There are 23 girls studying science and 44 girls studying arts/social science at the same college. So 26 boys and 23 girls study science. Does this mean that boys and girls are about equally interested in

Subject studied	**Frequencies of boys (*f*)**
Science	♂♂♂♂♂♂♂♂♂♂♂♂♂♂♂♂♂♂♂♂♂♂♂♂♂♂ = 26
Arts/social sciences	♂♂♂♂♂♂♂♂♂♂♂♂♂♂♂♂♂ = 17
Total	43

Exhibit 3.1 'A' levels studied in an Hypothetical Sixth Form College

A level subject	**Boys:** ***f***	**Boys:** **proportions** **(p)**	**Girls:** ***f***	**Girls:** **proportions** **(*p*)**
Science	26	$\frac{26}{43} = 0.605$	23	$\frac{23}{67} = 0.343$
Arts/ social sciences	17	$\frac{17}{43} = 0.395$	44	$\frac{44}{67} = 0.657$
Totals	43	1.0	67	1.0

Exhibit 3.2 Frequencies and proportions for boys and girls

science subjects? No, because there are more girls than boys. Twenty-six of a total of 43 boys are studying science compared with 23 of a total of 67 girls.

We need to give these figures a common base for comparison. The calculation of proportions provides this common base.

Proportions

Proportions are the number of cases belonging to a particular category divided by the total number of cases. The sum of the proportions of all the categories will always equal one. Exhibit 3.2 expresses the frequencies of girls' and boys' subject choices in terms of proportions: 0.605 of the boys study science, but only 0.343 of the girls.

Percentages

Percentages are proportions multiplied by 100. The total of all the percentages in any particular group (boys or girls) equals 100 per cent.

Thus, at this sixth-form college, 60.5 per cent of boys study science subjects compared with 34.3 per cent of girls.

If you want to **round** a percentage to the nearest whole percentage point, then look at the digits after the decimal point. If these are .499 or below, then round the figure down – for example, 23/67 = 34.328 per cent, or 34 per cent to the nearest whole number. If you have .500 or above, then round the figure up – for example, 17/43 = 39.535 per cent, which is 40 per cent to the nearest whole number.[1]

[1] There are other methods of rounding, for example just truncating the number at the decimal point or numbers ending in .5 rounding alternately up and down. However, these rules are hard to remember and so for simplicity in this book we will always round up numbers ending in .5.

Ratios

Ratios are another way of expressing the different numbers studying science and arts/social science subjects. The ratio of boys studying science to boys doing arts/social science A levels is

$$\frac{\text{frequency of boys studying science}}{\text{frequency of boys studying arts/social science}} = \frac{26}{17}$$

If we divide by the denominator (17), this becomes $\frac{1.53}{1}$

This can be written as 1.53 : 1. There are about 1.5 boys studying science subjects for every 1 boy studying arts. Since we normally like to express numbers like this as whole numbers, both the denominator and the numerator can be multiplied by 2 to show that there are three boys studying science subjects for every two boys studying arts/social science:

$$\frac{1.53}{1} \times \frac{2}{2} = \frac{3.06}{2}$$

Looking at the girls, the ratio of girls studying science to girls studying arts/social science is

$$\frac{\text{frequency of girls studying science}}{\text{frequency of girls studying arts/social science}} = \frac{23}{44} = \frac{0.52}{1}$$

Once again dividing by the denominator (44), there are about 0.5 girls studying science for every girl studying arts/social sciences, that is, there is one girl studying science for every two studying arts/social science. Alternatively we could arrive at the same conclusion by turning the ratio round and expressing it as follows:

$$\frac{\text{frequency of girls studying arts/social science}}{\text{frequency of girls studying science}} = \frac{44}{23} = \frac{1.91}{1}$$

There are 1.9 girls (2 if we round up) studying arts for every one studying science.

Proportions, percentages and ratios are alternative ways of comparing the relative amounts of something (in this example, the relative numbers of boys and girls taking science). Proportions and percentages are easy to convert from one to another and, while there is no hard rule, social scientists tend to prefer to use percentages. In this case, the percentages show clearly that the arts and social sciences subjects are more popular among girls, and that science is slightly more popular than arts/social science among boys.

Summary of notation for proportions, percentages and ratios

The following list summarizes the statistical concepts introduced so far this chapter.

Frequency:

The number of observations with attribute 1, f_1
The number of observations with attribute 2, f_2
Total number of observations, N

$$\text{Proportion: } P = \frac{f_1}{N} \text{ or } \frac{f_2}{N}$$

$$\text{Percentage: } = \frac{f_1}{N} \times 100\% \text{ or } \frac{f_2}{N} \times 100\%$$

$$\text{Ratio: } = \frac{f_1}{f_2} \text{ or } \frac{f_2}{f_1}$$

Coding variables for computer analysis

Before you can use SPSS to help you calculate a frequency distribution you need to give each category of a variable a numeric code. In addition you need to give each variable a variable name, as described in Chapter 2.

Exhibit 3.3 shows the data for sex, marital status, age and social class for just 20 people before numeric codes have been assigned to each category of each variable. This data set is in a file called **GHS2002subset.sav**.

For example, person 1, case 1, is male, is married, in social class III manual (IIIM) and aged 75.

The first variable, sex, is an example of a nominal variable which we can give the variable name SEX, and one possibility of coding this variable would be to assign codes as in Exhibit 3.4. Note that these codes have been assigned arbitrarily, so a code of 1 for males could equally have been 2 and vice versa for females.

The second variable, marital status, which we will call MARSTAT, could be coded as in Exhibit 3.5.

Once again, since marital status is a nominal variable we could have coded this variable is a completely different order.

Case	Sex	Marital status	Social class	Age
1	Male	Married and living with husband/wife	IIIM	75
2	Female	Divorced	IIIN	59
3	Male	Married and living with husband/wife	IIIM	55
4	Male	Single, never married	IV	18
5	Female	Married and living with husband/wife	IIIN	60
6	Female	Single, never married	IIIM	37
7	Female	Divorced	IIIN	66
8	Female	Widowed	IIIN	33
9	Male	Married and living with husband/wife	II	32
10	Female	Married and living with husband/wife	II	47
11	Female	Widowed	IIIN	67
12	Male	Single, never married	IV	20
13	Male	Married and living with husband/wife	IIIM	54
14	Female	Married and living with husband/wife	V	49
15	Female	Married and separated from husband/wife	IIIN	33
16	Male	Single, never married	IIIN	18
17	Male	Married and living with husband/wife	II	39
18	Male	Single, never married	II	48
19	Female	Married and living with husband/wife	II	60
20	Female	Widowed	I	84

Exhibit 3.3 Data for sex, marital status, social class and age for 20 respondents

Social class, the third variable, has been given the SPSS variable name NEWSC, and is an example of an ordinal variable where it is possible to rank or order the categories of the variable. As an ordinal variable, you know that someone in class I possesses more of whatever it is – salary, prestige, status – that goes to measure class, but you do not know *how much more* of these qualities they possess over someone in class V. There are only two ways you can code an ordinal variable, either in ascending order or descending order and

SEX	Coding scheme
Males	1
Females	2

Exhibit 3.4 Coding for the variable, **SEX**

MARSTAT – Marital status	Coding scheme
Single, never married	1
Married and living with husband/wife	2
Married and separated from husband/wife	3
Divorced	4
Widowed	5

Exhibit 3.5 Coding for the marital status **(MARSTAT)**

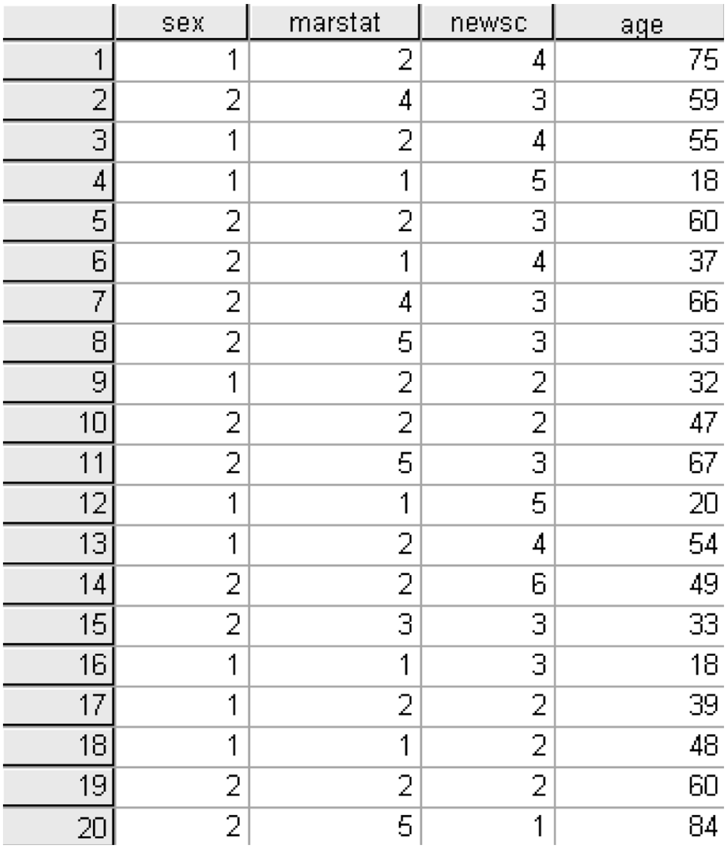
NEWSC – Social class	Scheme A	Scheme B
Social class 1	1	6
Social class II	2	5
Social class IIIN (Non-manual)	3	4
Social class IIIM (Manual)	4	3
Social class IV	5	2
Social class V	6	1

Exhibit 3.6 Coding for social class **(NEWSC)**

it generally does not matter which way you choose. So NEWSC could be coded either as in scheme A or as in scheme B in Exhibit 3.6.

Note that nominal and ordinal variables are often referred to as **categorical variables**.

Finally, AGE, the respondents' age, is a variable measured at the interval level. Here, instead of assigning codes to each person's response we will use their actual response. So we will use their actual age in the data.



	sex	marstat	newsc	age
1	1	2	4	75
2	2	4	3	59
3	1	2	4	55
4	1	1	5	18
5	2	2	3	60
6	2	1	4	37
7	2	4	3	66
8	2	5	3	33
9	1	2	2	32
10	2	2	2	47
11	2	5	3	67
12	1	1	5	20
13	1	2	4	54
14	2	2	6	49
15	2	3	3	33
16	1	1	3	18
17	1	2	2	39
18	1	1	2	48
19	2	2	2	60
20	2	5	1	84

Exhibit 3.7 The **Data Editor** in SPSS showing the variables SEX, MARSTAT, NEWSC and AGE.

Exhibit 3.7 shows these four variables in the SPSS **Data Editor** after they have been coded.

Frequency distributions in SPSS

In order to get SPSS to carry out the frequency procedure you need to select **Analyze|Descriptive statistics ▸| Frequencies** … and select the variables from the variable list before clicking on **OK** (see Chapter 2, Exhibit 2.16). The resulting frequency table for SEX is seen in Exhibit 3.8.

The first column shows the numeric codes that have been assigned to each category, with their respective value labels. The columns headed **Frequency** and **Percent** show the

sex Sex

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Male	9	45.0	45.0	45.0
	2 Female	11	55.0	55.0	100.0
	Total	20	100.0	100.0	

Exhibit 3.8 SPSS frequency output for sex

marstat Legal marital status

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Single, never married	5	25.0	25.0	25.0
	2 Married and living with husband/wife	9	45.0	45.0	70.0
	3 Married and separated from husband/wife	1	5.0	5.0	75.0
	4 Divorced	2	10.0	10.0	85.0
	5 Widowed	3	15.0	15.0	100.0
	Total	20	100.0	100.0	

Exhibit 3.9 SPSS frequency output for marital status **(MARSTAT)**

number of cases in the category and the percentages of the whole data set in the category respectively. The columns headed **Valid Percent** and **Cumulative Percent** will be explained in a later section in this chapter.

Exhibit 3.9 shows the frequency distribution of the variable MARSTAT, and Exhibit 3.10, the distribution for NEWSC.

If SPSS were asked for a frequency distribution for a variable which has many categories such as AGE, one would get a very, very long table, with a row for each different age. In the GHS data set the youngest respondent is 16 and the oldest 96, therefore there would be 81 rows in the table. This table is too large to comprehend easily and not very useful. The conclusion is that an SPSS frequency distribution is only suitable for variables which have a moderate number of categories. If you do have a variable such as AGE which has many categories, it is best to group the variable first into a small number of groupings (for example, group AGE into age bands) and then find the frequency distribution of the grouped categories.

newsc New social class

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 I	1	5.0	5.0	5.0
	2 II	5	25.0	25.0	30.0
	3 IIIN	7	35.0	35.0	65.0
	4 IIIM	4	20.0	20.0	85.0
	5 IV	2	10.0	10.0	95.0
	6 V	1	5.0	5.0	100.0
	Total	20	100.0	100.0	

Exhibit 3.10 SPSS frequency output for social class **(NEWSC)**

Age band	Code
10–19	1
20–29	2
30–39	3
40–49	4
50–59	5
60–69	6
70–79	7
80–89	8
90–99	9

Exhibit 3.11 Coding for recoding age into ten year age groups

Grouped frequency distributions

In order to group a continuous, interval variable, respondents are divided into appropriate or convenient intervals. The first task is to decide the boundaries of the intervals to be used. If we group AGE into 10-year age bands using the categories and codes in Exhibit 3.11, SPSS will produce the frequency distribution shown in Exhibit 3.12 using the data in **GHS2002subset.sav**.

agegroup

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	10–19	2	10.0	10.0	10.0
	20–29	1	5.0	5.0	15.0
	30–39	5	25.0	25.0	40.0
	40–49	3	15.0	15.0	55.0
	50–59	3	15.0	15.0	70.0
	60–69	4	20.0	20.0	90.0
	70–79	1	5.0	5.0	95.0
	80–89	1	5.0	5.0	100.0
	Total	20	100.0	100.0	

Exhibit 3.12 SPSS output for the age group (**AGEGROUP**)

Creating intervals in this way seems quite straightforward. However, we have profited by the fact the survey only recorded respondents' ages in whole years, for example as 29 years, not as 29 years, 4 months and 5 days or 29.342.

It is worth considering what would happen to a person aged 29.5 years using the coding scheme in Exhibit 3.11. Code 2 is assigned to those aged between 20 and 29 and code 3 is assigned to those aged between 30 and 39. It appears that those aged between 29 and 30 do not belong in either code. In order to cater for all possible codes we need to close up all the gaps between the intervals. We need to abut the intervals by creating **real class intervals** with **real class limits**. Note that, in practice, real class limits always have one more decimal place than the raw data and therefore never actually appear in the data. Thus if we *had* actually measured age to one decimal place, then the real class intervals would be defined using two decimal places.

Real class intervals

To create real class limits around real class intervals, divide the distance between the stated class intervals by 2, subtract this from the lower limit and add it to the upper limit (see Exhibit 3.13).

Exhibit 3.14 displays all the stated and real class limits for the variable AGE as previously coded. Of course, AGE could be recoded in many different ways with correspondingly different real class intervals.

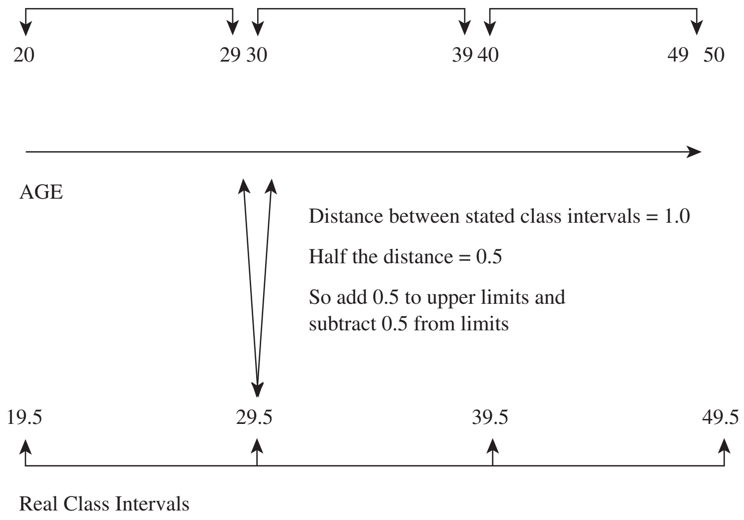


Exhibit 3.13 The relationship between stated class intervals and real class intervals

Stated class intervals	**Real class intervals**
10–19	9.5–19.5
20–29	19.5–29.5
30–39	29.5–39.5
40–49	39.5–49.5
50–59	49.5–59.5
60–69	59.5–69.5
70–79	69.5–79.5
80–89	79.5–89.5
90–99	89.5–99.5

Exhibit 3.14 Stated class intervals and real class intervals for age group

Midpoints

Another important statistic when creating real class intervals is the midpoint of the real class interval. The midpoint of a real class interval is defined as the point exactly half-way between the lower and upper real class limit.

$$\text{midpoint of interval} = \text{real lower class limit} + \text{one half of size of class interval}$$

For example,

$$\text{size of the class interval} = 29.5 - 19.5 = 10$$

$$\text{midpoint of the real class interval } [19.5 - 29.5] = 19.5 + \frac{10}{2} = 24.5$$

Exhibit 3.15 shows the midpoints of all the real class intervals for the variable AGE.

Stated class intervals	Real class intervals	Midpoints
10–19	9.5–19.5	9.5 + 0.5(10) = 14.5
20–29	19.5–29.5	19.5 + 0.5(10) = 24.5
30–39	29.5–39.5	34.5
40–49	39.5–49.5	44.5
50–59	49.5–59.5	54.5
60–69	59.5–69.5	64.5
70–79	69.5–79.5	74.5
80–89	79.5–89.5	84.5
90–99	89.5–99.5	94.5

Exhibit 3.15 Midpoints for real class intervals for age group

Procedure for grouped frequency distribution

1. Decide how many intervals to use. Between seven and ten intervals is a reasonable number. Too few and you lose too much information.
2. Find the size and number of the class intervals. Round the highest score up and the lowest score down to a convenient number. This will give you a range of scores. Then divide the range of the scores by the number of intervals to arrive at a convenient interval size. For example, to create age bands from AGE, round the highest age, 97 up to 100 and the lowest, 16, down to 10; this gives a distribution of scores from 100 to 10, a range of 90. Divide this range by a convenient number of intervals so that you also get a convenient class interval size (usually a round number like 10 rather than an awkward number like 9). This naturally would yield nine intervals of size 10.
3. Count the number of cases in each stated interval and report these as frequencies (f). Report the total number of cases (N).
4. Calculate percentages for each interval.

Frequency tables from the 2002 GHS

Finally, let us look at the frequency distributions using the large GHS 2002 data set. Exhibits 3.16 and 3.17 show the frequency tables for SEX and MARSTAT. If we report the percentages to the nearest whole number by rounding, we see that 47 per cent are male and 53 per cent are female. Looking at legal marital status, 54 per cent are married, 28 per cent are single and the rest (19 per cent) are widowed, divorced or separated. Notice that rounding in this way now means that the total percentage exceeds 100 ($54+28+19 = 101$), something you may come across in published articles but which is usually acknowledged in a footnote.

sex Sex

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Male	1943	47.3	47.3	47.3
	2 Female	2168	52.7	52.7	100.0
	Total	4111	100.0	100.0	

Exhibit 3.16 SPSS frequency output for sex from the GHS 2002

marstat Legal marital status

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 single, never married	1141	27.8	27.8	27.8
	2 married and living with husband/wife	2198	53.5	53.5	81.2
	3 married and separated from husband/wife	115	2.8	2.8	84.0
	4 divorced	316	7.7	7.7	91.7
	5 widowed	341	8.3	8.3	100.0
	Total	4111	100.0	100.0	

Exhibit 3.17 SPSS frequency output for marital status from the GHS 2002

Exhibit 3.18 demonstrates a frequency table for the grouped variable, AGEGROUP. This variable was derived from the original data by recoding the variable AGE according to the scheme in Exhibit 3.11.

The frequency table of the final variable, NEWSC, is described below and illustrates how one deals with non-response codes.

agegroup

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00 10–19	243	5.9	5.9	5.9
	2.00 20–29	608	14.8	14.8	20.7
	3.00 30–39	796	19.4	19.4	40.1
	4.00 40–49	671	16.3	16.3	56.4
	5.00 50–59	676	16.4	16.4	72.8
	6.00 60–69	513	12.5	12.5	85.3
	7.00 70–79	393	9.6	9.6	94.9
	8.00 80–89	200	4.9	4.9	99.7
	9.00 90–99	11	.3	.3	100.0
	Total	4111	100.0	100.0	

Exhibit 3.18 SPSS frequency output for age group from the GHS 2002

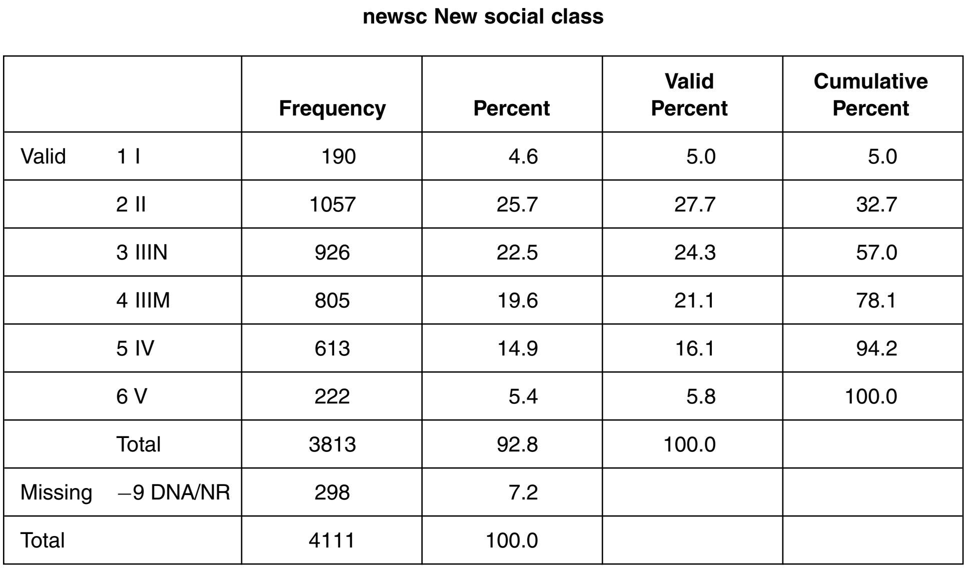
Missing values in SPSS

When you are coding the responses to a survey prior to computer analysis, *every* response must be coded,[2] even if the respondent refused to answer or if the question was inapplicable. For instance, there may be a two questions in a survey, the first asking if the respondent is employed and the second asking about the respondent's occupation. Clearly, if the respondent is not employed, then the next question is inapplicable. However, you must assign a code to this non-response but you can indicate that this code is special by assigning it as a **missing value.**

Consider the variable age measured on an interval scale. Some people may refuse to answer this question and be given a non-response code of −9. If you then wanted to work out the average age of all the respondents, you would want to make sure that the −9 code is not included in the analysis. Flagging a value as a missing value would ensure that the −9 code is excluded from any calculations.

In Exhibit 3.19, the frequency table for NEWSC, notice that some people have never worked and the question does not apply (DNA) or they did not respond (NR) and have been given the code −9. We do not know their social class and therefore would want to exclude them from any further consideration in our analysis. This has been done by declaring the code they have been assigned (−9) as a **missing value** in SPSS. The effect of this has been that the 298 respondents who have been given code −9 have been excluded from

[2] You can leave the data as a blank in SPSS. The data value will be treated as a missing value – a **system missing** value.



newsc New social class

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 I	190	4.6	5.0	5.0
	2 II	1057	25.7	27.7	32.7
	3 IIIN	926	22.5	24.3	57.0
	4 IIIM	805	19.6	21.1	78.1
	5 IV	613	14.9	16.1	94.2
	6 V	222	5.4	5.8	100.0
	Total	3813	92.8	100.0	
Missing	−9 DNA/NR	298	7.2		
Total		4111	100.0		

Exhibit 3.19 SPSS frequency output for social class from the GHS 2002

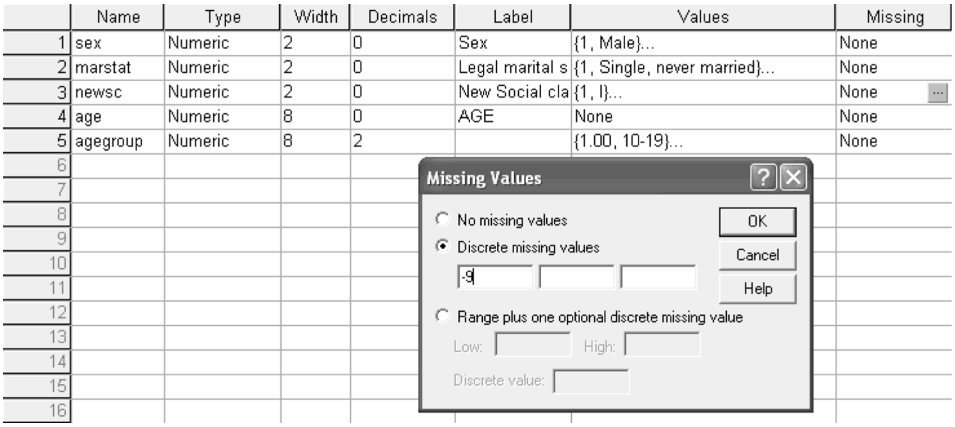
the percentage calculation seen in the column headed **Valid Percent**. Valid percentages calculated in this way are usually the ones that you are interested in and the ones that you report in your analysis.

Defining missing values in SPSS

To define a code as a missing value code for any particular variable is done in the SPSS **Data Editor** window (see Exhibit 3.20). Here, the grey square in the column headed **Missing** for the variable NEWSC has been clicked so that the **Missing Values** dialog box has appeared. The button marked **Discrete Missing Values** is selected and the value you wish to define as missing is typed in, as in Exhibit 3.20. Note that you are allowed a maximum of three discrete missing values.

Exploring the data set and creating a codebook

Sometimes you do not know the names of the variables you are interested in using in your analysis. This is especially the case when you are conducting a secondary analysis of data that was collected by someone else. For instance, you might be interested in how many people have a telephone and know that a question on telephone use was asked in the General Household Survey, although you do not know what variable name has been assigned to that question. There are two ways you could go about discovering this information. In the first place you could use an on-screen facility in SPSS to help find the



	Name	Type	Width	Decimals	Label	Values	Missing
1	sex	Numeric	2	0	Sex	{1, Male}...	None
2	marstat	Numeric	2	0	Legal marital s	{1, Single, never married}...	None
3	newsc	Numeric	2	0	New Social cla	{1, I}...	None
4	age	Numeric	8	0	AGE	None	None
5	agegroup	Numeric	8	2		{1.00, 10-19}...	None
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							

Missing Values

○ No missing values

● Discrete missing values

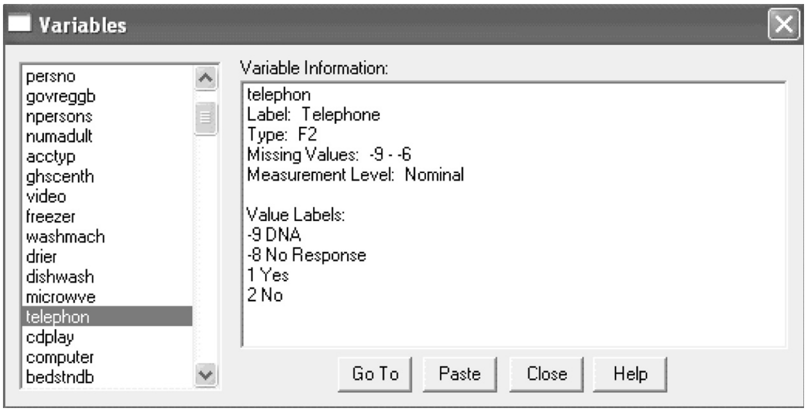
-9

○ Range plus one optional discrete missing value

Low: High:

Discrete value:

Exhibit 3.20 Defining **missing values** in SPSS for social class



Variables

persno
govreggb
npersons
numadult
acctyp
ghscenth
video
freezer
washmach
drier
dishwash
microwve
telephon
cdplay
computer
bedstndb

Variable Information:

telephon
Label: Telephone
Type: F2
Missing Values: -9 - -6
Measurement Level: Nominal

Value Labels:
-9 DNA
-8 No Response
1 Yes
2 No

Exhibit 3.21 Variable information in SPSS

variable name for this question and the codes that have been assigned to the answers. In the long term, though, it is more useful to create a codebook with the same information, but for all the variables, which you can print out to refer to later.

To use the on-screen facility to discover the name of the variable that has asked the respondents about telephone use, select **Utilities | Variables** … from the main menu to see the dialog box in Exhibit 3.21. A search of the list of variables in the left-hand side of the dialog box reveals a variable called TELEPHON. Selecting this variable displays its variable information in the right-hand side of the dialog box. Here you can see that those people who have a phone have been coded with a 1 and those who do not have a phone have been given code 2.

Question/variable label	Variable name	Values	Value labels	Missing values	Range of valid values
Sex	SEX	1 2	Male Female	–9	1, 2
Legal marital status	MARSTAT	1 2 3 4 5 –9	Single, never married Married Separated Divorced Widowed No response	–9	1–5
Age	AGE	–9	No Response	–9	16–96
Social class	NEWSC	1 2 3 4 5 6 –9	Social Class I Social Class II Social Class III (NM) Social Class III (M) Social Class IV Social Class V DNA/NR	–9	1–6

Exhibit 3.22 Codebook for the GHS2002subset datafile

A quick way of getting a basic, but often very extensive, codebook using SPSS is to select **File|Display Data File Information ▶| Working File** while the data file is open. This provides you with text output with the details of all the variables in the file, which can be saved or printed. However, it is often better to spend some time to create a more useful codebook of information of just the relevant information from your data file. Exhibit 3.22 is an example of a such a codebook showing the variable label (which corresponds to the question asked in the survey), the variable name in SPSS, the values assigned to each response with accompanying labels, the missing value(s) and finally the range of valid values. Such a codebook could be constructed in your word processor and would provide the key to using the data.

Households and individuals in the General Household Survey

As described in Chapter 1, the GHS consists of data about households and individuals within those households. Information about each household, such as car ownership, has been *spread* to each individual in that household. Therefore if three people live in one household and own a car, then the car will be counted three times at the individual level. To get round this problem and to gain a correct percentage for car ownership at the household level, we need to select just one person from each household. A variable, HOUSEHLD,

cars Number of cars or vans

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00 no car or van	852	20.7	20.7	20.7
	2.00 1 car or van	1672	40.7	40.7	61.4
	3.00 2 car or vans	1251	30.4	30.4	91.8
	4.00 three or more cars or vans	336	8.2	8.2	100.0
	Total	4111	100.0	100.0	

Exhibit 3.23 Frequency of car ownership without selecting households

cars Number of cars or vans in household

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00 no car or van	607	27.5	27.5	27.5
	2.00 1 car or van	937	42.5	42.5	70.1
	3.00 2 car or vans	544	24.7	24.7	94.7
	4.00 three or more cars or vans	116	5.3	5.3	100.0
	Total	2204	100.0	100.0	

Exhibit 3.24 Frequency of car ownership after selecting households

has been created to facilitate this task. It is coded one for one person in each household and zero for all other members of the household.

Therefore you select **Data | Select cases…** and click on **If condition is satisfied** and then **If**… . The condition you type in the next dialog box is HOUSEHLD = 1. This selects one person per household and all you now need to do is to carry out a frequency procedure for the variable CARS. Exhibits 3.23 and 3.24 demonstrate the different results obtained when running a frequency procedure with and without selecting households. If households are *not* selected prior to the analysis, the number of households without a car or van is shown as 20.7 per cent (Exhibit 3.23). If households are selected, however, this percentage increases to the truer value of 27.5 per cent (Exhibit 3.24). Clearly, we need to be aware of whether our analysis is at the individual or at the household level.

Summary

This chapter has introduced frequencies, proportions and percentages for single variables. How to code nominal, ordinal and interval variables has been described as a preliminary procedure before using statistical software such as SPSS.

You should now be familiar with the following topics:

- How to describe the distribution of a variable using simple, descriptive statistics
- How to code variables for computer analysis and create a codebook
- How to group continuous variables for analysis
- How to deal with missing information
- The difference between individual- and household-level data

and be able to perform the following tasks in SPSS:

- Create frequency tables and calculate the percentages of categories of single variables
- Add missing values to variables in SPSS.

Exercises

Using the General Household Survey data (**GHS2002.sav**), answer the following questions. Remember to select only household data if the response is required at the household level.

1. What percentage of the households have a video recorder?
2. What percentage of households own more than three cars or vans?
3. What percentage of respondents are Asian?
4. What percentage of households live in a semi-detached house?
5. What percentage of respondents live in Scotland?
6. What percentage of respondents are students?
7. What percentage of respondents are retired?
8. What percentage of respondents left school under 15 years of age?
9. What percentage of households consist of a couple and no children?
10. What percentage of households have only one colour TV?