Fielding and Gilbert (2006)

FREQUENCIES, PROPORTIONS AND PERCENTAGES

3

CONTENTS

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

FREQUENCY DISTRIBUTIONS

One of the first things you might want to do with data is to count the n occurrences that fall into each category of each variable. This provides **frequency distributions**, allowing you to compare information betwee

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 nuclear testing should be stopped'. They also allow you to see what are the highest and lowest values and the value at 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 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. Thus the total of all the percentages in any particular group (boys or girls) equals 100 per cent. Thus at

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

Exhibit 3.1 *A levels studied in a hypothetical sixth-form college*

A-level subject	Boys		Girls	
	Frequency (*f*)	Proportions (*p*)	Frequency (*f*)	Proportions (*p*)
Science	26	$\frac{26}{43} = 0.605$	23	$\frac{23}{67} = 0.34$
Arts/ social science	17	$\frac{17}{43} = 0.395$	44	$\frac{44}{67} = 0.65$
Totals	43	1.0	67	1.0

Exhibit 3.2 *Frequencies and proportions for boys and girls*

this sixth-form college, 60.5 per cent of boys study science subjects, cor to 34.3 per cent of girls.

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

RATIOS

Ratios are another way of expressing the different numbers studying scien arts/social science subjects. The ratio of boys studying science to boys 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

$$\frac{1.53}{1}$$

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

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 science, that is, there is one girl studying science for every two studying arts/social science. Alternatively, we

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

Proportion:

$$p = \frac{f_1}{N} \quad \text{or} \quad \frac{f_2}{N}$$

Percentage:

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

Ratio:

$$= \frac{f_1}{f_2} \quad \text{or} \quad \frac{f_2}{f_1}$$

could arrive at the same conclusion by turning the ratio round and expres 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 stu science.

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

CODING VARIABLES FOR COMPUTER ANALYSIS

Before you can use SPSS to help you calculate a frequency distributior need to give each category of a variable a numeric code. In addition, you n give each variable a variable name of no more than 8 characters, as descril Chapter 2.

Exhibit 3.3 shows the data for sex, marital status, age and social class fc

Case	Sex	Marital status	Social class	Age
1	Male	Single	SOCIAL CLASS IIIN	68
2	Male	Cohabiting	SOCIAL CLASS IIIN	38
3	Female	Single	SOCIAL CLASS II	18
4	Female	Cohabiting	SOCIAL CLASS IIIN	29
5	Female	Married	SOCIAL CLASS IIIN	62
6	Female	Widowed	SOCIAL CLASS IV	81
7	Male	Widowed	SOCIAL CLASS II	74
8	Female	Divorced	SOCIAL CLASS II	41
9	Female	Married	SOCIAL CLASS IIIN	59
10	Male	Married	SOCIAL CLASS IIIM	59
11	Female	Divorced	SOCIAL CLASS IIIM	74
12	Female	Married	SOCIAL CLASS IIIN	39
13	Male	Married	SOCIAL CLASS I	78
14	Female	Divorced	SOCIAL CLASS IV	45
15	Female	Married	SOCIAL CLASS IIIN	64
16	Male	Married	SOCIAL CLASS II	45

Exhibit 3.3 *Data for sex, marital status, social class and age for 16 respondents*

16 people, before numeric codes have been assigned to each category of each variable. This data set is in a file called GHS95_16CASES.SAV. For example, person 1 (case 1) is male, single, in social class III non-manual (IIIN) and aged 68.

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 in a completely different order.

Social class, the third variable, has been given the SPSS variable name SOCLASE, 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 than someone in class IV. There are only two ways you can code an ordinal variable, in either ascending order or descending order, and it generally does not matter which way you choose. So SOCLASE could be coded either as in scheme A or as in scheme B in Exhibit 3.6.

Finally, AGE, the respondents' age, is a variable measured at the interval level.

SEX	Coding scheme
Males	1
Females	2

Exhibit 3.4 *Coding for the variable SEX*

MARSTAT - Marital status	Coding scheme
Married	1
Cohabiting	2
Single	3
Widowed	4
Divorced	5

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

SOCLASE - SOCIAL CLASS	Scheme A	Scheme B
Social class I	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 (SOCLASE)*

Here, instead of assigning codes to each person's response we will use tl actual response. So we will use their actual age in the data.

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

FREQUENCY DISTRIBUTIONS IN SPSS

In order to get SPSS to carry out the frequency procedure you need to sel **Analyze|Descriptive statistics ▸|Frequencies**... and select the variables fr the variable list before clicking on **OK** (see Chapter 2, Exhibit 2.17). T resulting frequency table for SEX is seen in Exhibit 3.8.

The first column shows the numeric codes that have been assigned to e category, labelled by the respective value label. The columns headed **Frequer** and **Percent** show the number of cases in the category and the percentages the whole data set in the category respectively. The columns headed **Va Percent** and **Cumulative Percent** will be explained in a later section in t chapter.

Exhibit 3.9 shows the frequency distribution of the variable MARSTAT, a Exhibit 3.10 the distribution for SOCLASE.

If SPSS were asked for a frequency distribution for a variable which has ma categories such as AGE, one would get a very, very long table, with a row each different age. In the GHS data set the youngest respondent is 16 and oldest 97, therefore there would be 82 rows in the table. This table is too large comprehend easily and not very useful. The conclusion is that an SP frequency distribution is only suitable for variables which have a moder number of categories. If you do have a variable such as age which has ma categories, it is best to divide the variable first into a small number of groupi (for example, group **age** into age bands) and then find the frequency distribut of the grouped categories.

	sex	marstat	soclase	age
1	1	3	3	68
2	1	2	3	38
3	2	3	2	18
4	2	2	3	29
5	2	1	3	62
6	2	4	5	81
7	1	4	2	74
8	2	5	2	41
9	2	1	3	59
10	1	1	4	59
11	2	5	4	74
12	2	1	3	39
13	1	1	1	78
14	2	5	5	45
15	2	1	3	64
16	1	1	2	45

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

SEX SEX

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Male	6	37.5	37.5	37.5
	2 Female	10	62.5	62.5	100.0
	Total	16	100.0	100.0	

Exhibit 3.8 *SPSS frequency output for* SEX

MARSTAT MARITAL STATUS

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Married	7	43.8	43.8	43.8
	2 Cohabiting	2	12.5	12.5	56.3
	3 Single	2	12.5	12.5	68.8
	4 Widowed	2	12.5	12.5	81.3
	5 Divorced	3	18.8	18.8	100.0
	Total	16	100.0	100.0	

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

SOCLASE SOCIAL CLASS OF INDIVIDUAL

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 SOCIAL CLASS I	1	6.3	6.3	6.3
	2 SOCIAL CLASS II	4	25.0	25.0	31.3
	3 SOCIAL CLASS IIIN	7	43.8	43.8	75.0
	4 SOCIAL CLASS IIIM	2	12.5	12.5	87.5
	5 SOCIAL CLASS IV	2	12.5	12.5	100.0
	Total	16	100.0	100.0	

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

GROUPED FREQUENCY DISTRIBUTIONS

In order to group a continuous, interval variable, respondents are divided appropriate or convenient intervals. The first task is to decide the boundarie the intervals to be used. If we group AGE into ten-year age bands using categories and codes in Exhibit 3.11, SPSS will produce the frequency distr tion shown in Exhibit 3.12 using the GHS95_16case data.

Creating intervals in this way seems quite straightforward. However, we l profited by the fact the survey only recorded respondents' ages in whole y 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 u the coding scheme in Exhibit 3.11. Code 2 is assigned to those aged betwee and 29 and code 3 is assigned to those aged between 30 and 39. It appears those aged between 29 and 30 do not belong in either code. In order to cate all possible codes we need to close up all the gaps between the intervals. need to abut the intervals by creating **real class intervals** with **real class lim**

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*

AGEGROUP

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	1	6.3	6.3	6.3
	2	1	6.3	6.3	12.5
	3	2	12.5	12.5	25.0
	4	3	18.8	18.8	43.8
	5	2	12.5	12.5	56.3
	6	3	18.8	18.8	75.0
	7	3	18.8	18.8	93.8
	8	1	6.3	6.3	100.0
	Total	16	100.0	100.0	

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

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.

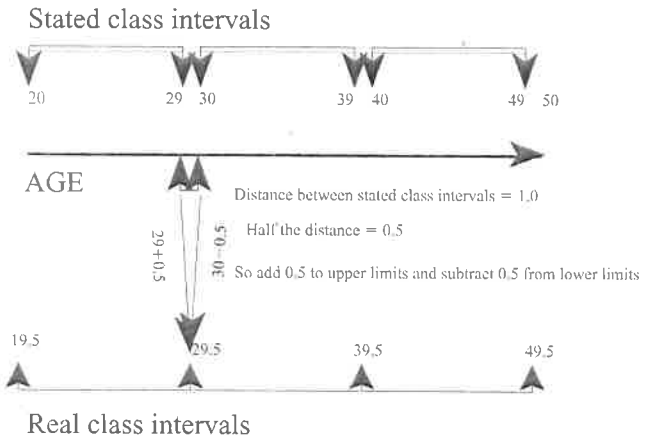

Exhibit 3.13 *The relationship between stated class intervals and real class intervals*

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* `agegroup`

MIDPOINTS

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

Midpoint of interval = real lower class limit

+ one half of size of class inte