Data sets and data matrices

A matrix is an arrangement of numbers in rows and columns. In psychology it is customary to write people in the rows and variables in the columns. Thus, the matrix below would involve five people (five rows), for each of whom data were available for three variables (three columns).

Person	Variable 1	Variable 2	Variable 3
1	7	9	4
2	6	12	8
3	19	7	7
4	12	14	10
5	2	4	3

Sample v. Population

Suppose we wished to conduct an investigation concerning the intelligence of university students in your country. We might undertake to -obtain the IQ score of every individual enrolled in all universities in the entire country. This set of scores would be a complete or exhaustive set of observations relevant to our investigation. With such a complete set of observations we could make statements about the IQ of University students with perfect confidence. For example, you could make an exact statement about the range of variation or about the average IQ of all students in that particular year (or moment) of investigation.

When we have selected a sample from a population, we usually compute certain characteristics of that sample - for example an average IQ level.

Characteristics of samples are called **STATISTICS**. However, when we are dealing with populations, characteristics such as average IQ levels are called **PARAMETERS**. Thus, if we define our population as all people taking 1st Year Psychology in Bendigo, the average height or average IQ of these people would be a parameter. But if we put all the names of these

students into a barrel and randomly draw out 10 names, the average height or IQ of this sample would be a statistic.

2. ORGANISING AND PRESENTING DATA

2.1Frequency Distributions

The collection of research data results in an accumulation of numbers. Most research projects yield a great many numbers and as researchers we must have techniques by which we can put order into the masses of numbers we collect. Clearly it is impractical simply to write out all the numbers obtained. For example, suppose a researcher interested in human ability obtains a sample of 100 University students and gives each one 10 problems to solve - a subject's score may therefore fall between 0 (none correct) and 10 (all correct). Let us say the scores for the 100 students are as shown in Table 2.1:

Table 2.1: Number of problems solved correctly by 100 University students (hypothetical data)

5	8	3	6	5	8	3	7	4	7
6	8	7	4	6	5	8	7	10	6
5	3	6	1	8	6	8	5	7	10
8	9	6	9	6	5	9	9	6	3
8	5	8	4	8	6	7	4	10	5
8	7	6	8	5	7	6	9	6	10
4	8	6	8	6	5	4	7	9	6
7	4	5	5	9	6	6	7	4	5
10	3	5	7	9	10	6	7	6	6
6	9	8	7	8	5	7	4	6	8

As you can see the table of 100 numbers is rather confusing and difficult to comprehend. **A frequency distribution,** on the other hand, will present a clearer picture. The first step in constructing such a distribution is to list every possible score value in the first column of a table (this column is denoted)9_7 with the highest score at the top. Second, the frequency (denoted f) of each score, or the number of times a given score was obtained, is listed to the right of the

score in the second column of the table. To arrive at the figures in the "frequency" column for the data in Table 2.2 you go through the data and count all the 10s, go through the data again, and count all the 9s and so on, until all frequencies are tabulated. The completed frequency distribution is shown in Table 2.2.

Table 2.2: Frequency distribution for data in Table 5.1

Score (X_i)	Frequency (f)
10	6
9	9
8	17
7	15
6	23
5	15
4	9
3	5
2	
1	1
0	0

22 Cumulative Frequency Distribution

Table 2.2 reveals at a glance how often each score was obtained. This makes it easier to form impressions as to the performance of subjects since you can conveniently ascertain (among other things) that 6 was the most frequently obtained score, that scores distant from 6 tended to occur less frequently than scores close to 6, and that the majority of subjects got more than half the problems correct. This latter observation is made clearer by adding a third column to our table - the cumulative frequency column (denoted cf). To construct the cumulative frequency column, we first construct a frequency distribution. We then start with the lowest score in the distribution and fouli a new column of cumulative frequencies by adding up the frequencies as we go. For example, as shown in Table 2.3 the cumulative frequency of 4 is 15. This value was obtained by adding the frequencies of 0, 1, 2, 3 and 4.

Table 2.3: Frequency and cumulative frequency distributions for the data in Table 5.1

Score (X_1)	Frequency (f)	Cumulative Frequency (cf)
10	6	100
9	9	94
8	17	85
7	15	68
6	23	53
5	15	30
4	9	15
3	5	6
2	0	1
1	1	1
0	0	0

The cumulative frequency distribution is interpreted as follows: the cumulative frequency of 15 for the score of 4 means that 15 people obtained a score of 4 or less. This can be readily verified by looking at the frequencies: nine people scored 4, five scored 3, one person scored 1, a total of 15 people with scores at or below 4. Note that we include in our table all possible scores, even if some of them were not achieved by anyone. The "frequency" of these scores is of course 0. In Table 2.2, it can be seen that no students achieved 0 or 2 correct problems.

The cumulative frequency for the highest score always equals N, the number of subjects. How many people obtained a score of 6 or more? The cumulative frequency distribution reveals that 30 people obtained scores of 5 or less, so there must be 70 people who scored 6 or more, since there were 100 people in all. The cumulative frequency distribution, then, allows us to arrive at certain needed information more quickly and conveniently than is possible using the frequency distribution.

23 Grouped Frequency Distributions

When the number of different scores to be listed in the score (X) column is not too large, frequency distributions are an excellent way of conveniently summarising a set of data. If, however,

there are more than 15 or 20 possible values of X to be written, constructing a frequency distribution is likely to prove very tedious. One possibility is to use a grouped frequency distribution. Instead of listing single scores in the score column (for example, 0, 1, 2, 3, 4), several score values are grouped together into a class interval (for example 0-4, 5-9), and frequencies are tallied for each interval.

For example, the data in Table 2.4 represent the scores of 85 students in a mid - term exam. The highest score is 50. If a regular frequency distribution were to be used, some 50 separate scores and corresponding frequencies would have to be listed. To avoid such a tiresome task, a grouped frequency distribution has been formed in Table 2.5. An interval size of three has been chosen, meaning that there are three score values in each class interval. Frequencies for each class interval are tabulated by treating all scores falling into the same interval as equal. When a tabulation is completed, the frequency opposite a given class interval indicates the number of cases with scores in that interval.

Table 2.4: Scores of 85 students on a 50 point midterm exam (hypothetical data)

39	42	30	11	35	25	18	26	37	15
29	22	33	32	21	43	11	11	32	29
44	26	30	50	13	38	26	39	45	21
31	28	14	35	10	41	15	39	33	34
46	21	38	26	26	37	37	14	26	24
32	15	22	28	33	47	9	22	31	20
37	40	20	39	30	18	29	35	41	21
26	25	29	33	23	30	43	28	32	32
34	28	38	32	31					

Table 2.5: Grouped and cumulative frequency distributions for data in Table 2.4

Class interval	Frequency (f)	Cumulative Frequency (cf)
51-53	0	85
48-50	1	85
45-47	3	84
42-44	4	81
39-41	6	77
36-38	7	71
33-35	9	64
30-32	14	55
27-29	8	41
24-26	10	33
21-23	8	23
18-20	4	15
15-17	3	11
12-14	3	8
9-11	5	5
6- 8	0	0

When grouping data, one of the important issues is how wide (large) the class interval should be, since grouped frequency distributions lose information because they do not provide the exact value of each score. Naturally, the wider the interval, the more information is lost, but if it is too narrow, we encounter the same problems as with individual scores. Thus, there is a trade-off between losing information and presenting a meaningful display. To have the best of both worlds, we must choose a class interval size which is not too wide or too narrow. In practice, it has been found that dividing the distribution into about 10 equal intervals (depending on the number and ranges of raw scores) usually works well. In general, up to about 20 class intervals is satisfactory. The size of the class interval is thus determined by dividing the range of scores (the difference between the highest and lowest obtained scores) by the number of intervals to be used.

Summary of procedures for constructing a grouped frequency distribution.

1. Find the range of the scores.
2. Determine the width of each class interval.
3. List limits of each class interval, placing the interval containing the lowest score at the bottom.
4. Tally the raw scores into the appropriate class intervals.
5. Add the tallies for each interval to obtain the interval frequency.

2.3.1 Exact limits and midpoint of a grouped frequency distribution

It is important to notice that each class has an exact lower and upper limit. This is a fictitious score lying halfway between each class interval and the one above it (exact upper limit) or below it (exact lower limit). Take the class interval 33-35 in Table 2.5. Its exact lower limit is halfway between the lowest score in the interval (33) and the highest score in the class interval below it (30-32). What value lies halfway between 32 and 33? The answer is 32.5; the exact lower limit of the class interval 33-35 is 32.5. In a similar way, the exact upper limit of 33-35 lies halfway between 35 and the lowest value of the next interval above 32-35. This interval is 36-38. The score halfway between 35 and 36 is 35.5. Thus, the exact limits of the class interval 33-35 are 32.5 (lower limit) and 35.5 (upper limit).

A class interval also has a midpoint (the value exactly in the middle of the class interval). This is obtained by adding the exact lower limit of the class interval and the exact upper limit and dividing by 2. To continue with the example of the class interval of 33-35 in Table 2.5, the midpoint is (32.5 + 35.5)/2.

This equals 68 = 34.0. The midpoint of the interval is 34.0. 2

In certain calculations with grouped data it is assumed that the cases lying within the class interval were equally distributed between the exact lower limit and the exact upper limit. In other calculations, it is assumed that all the cases lay at the midpoint of the class interval: for this reason, it is important to know how to calculate these values.