1

First things first – the nature of data

Learning objectives

When you have finished this chapter, you should be able to:

- Explain the difference between nominal, ordinal and metric data.
- Identify the type of any given variable.
- Explain the non-numeric nature of ordinal data.

Variables and data

Let's start with some numbers. Have a look at Figure 1.1.

These numbers are actually the birthweights of a sample of 100 babies (measured in grams). We call these numbers *sample data*. These data arise from the variable *birthweight*. To state the blindingly obvious, a variable is something whose value can vary. Other variables could be blood type, age, parity and so on; the values of these variables can change from one individual

Medical Statistics from Scratch: An Introduction for Health Professionals, Third Edition. David Bowers.

2240	4110	3590	2880	2850	2660	4040	3580	1960	3550
3050	3130	2660	3150	3220	3990	4020	3040	3460	4230
4110	2780	2840	3660	3580	2780	3560	2350	2720	2460
3200	2650	3000	3170	3500	2400	3300	3740	2760	3840
3740	2380	3300	3480	3740	3770	2520	3570	3400	3780
3040	3170	3300	3560	3180	2920	4000	2700	3680	2500
2920	2980	3780	2650	2880	4550	3570	1620	3000	3700
4080	3280	3800	2800	2560	2740	3180	3200	3120	4880
2800	3640	4020	3080	2590	3360	3630	3740	2960	3300
3090	3600	3720	2840	3320	2940	3640	2720	3220	4140

Figure 1.1 Some numbers. Actually, the birthweight (g) of a sample of 100 babies. Data from the Born in Bradford Cohort Study. Born in Bradford, Bradford Institute for Health Research, Bradford Teaching Hospitals NHS Foundation Trust

to the other. When we measure a variable, we get data – in this case, the variable birthweight produces birthweight *data*.

Figure 1.2 contains more sample data, in this case, for the *gender* of the same 100 babies.

M	M	F	F	M	M	F	F	M	M
M	M	M	F	M	M	F	F	M	M
F	F	M	M	F	F	M	F	F	F
M	M	F	F	M	M	M	M	F	F
M	M	F	F	M	F	F	F	F	F
F	M	F	M	M	M	F	F	M	F
F	F	M	M	M	F	M	M	M	F
M	F	M	M	M	M	M	M	M	M
M	F	M	M	M	F	F	M	M	F
M	F	M	F	M	M	F	F	M	F

Figure 1.2 The gender of the sample of babies in Figure 1.1

Moreover, Figure 1.3 contains sample data for the variable *smoked while pregnant*.

The data in Figures 1.1, 1.2 and 1.3 are known as *raw* data because they have not been organised or arranged in any way. This makes it difficult to see what interesting characteristics or features the data might contain. The data cannot tell its story, if you like. For example, it is not easy to observe how many babies had a low birthweight (less than 2500 g) from Figure 1.1, or what proportion of the babies were female from Figure 1.2. Moreover, this is for only 100 values. Imagine how much more difficult it would be for 500 or 5000 values. In the next four chapters, we will discuss a number of different ways that we can organise data so that it can tell its story. Then, we can see more easily what is going on.

No	No	No	No	No	No	No	No	Yes	No
No	No	No	No	No	No	No	No	Yes	No
No	No	No	No	No	Yes	No	No	No	No
Yes	No	Yes	No	No	No	No	No	No	No
No	No	No	No	No	No	No	No	No	No
No	No	No	Yes	Yes	No	No	No	No	No
No	No	Yes	No	No	Yes	No	No	No	No
Yes	No	No	No	Yes	Yes	No	No	Yes	No
No	No	No	No	No	No	No	No	No	No
No	Yes	No	No	No	Yes	No	No	Yes	No

Figure 1.3 The variable 'smoked while pregnant?' for the mothers of the babies in Figure 1.1

Exercise 1.1. Why do you think that the data in Figures 1.1, 1.2 and 1.3 are referred to as 'sample data'?

Exercise 1.2. What percentage of mothers smoked during their pregnancy? How does your value contrast with the evidence which suggests that about 20 per cent of mothers in the United Kingdom smoked when pregnant?

Of course, we gather data not because it is nice to look at or we've got nothing better to do but because we want to answer a question. A question such as 'Do the babies of mothers who smoked while pregnant have a different (we're probably guessing lower) birthweight than the babies of mothers who did not smoke?' or 'On average, do male babies have the same birthweight as female babies?' Later in the book, we will deal with methods which you can use to answer such questions (and ones more complex); however, for now, we need to stick with variables and data.

Where are we going ... ?

- This book is an introduction to medical statistics.
- Medical statistics is about doing things with data.
- We get data when we determine the value of a variable.
- We need data in order to answer a question.
- What we can do with data depends on what type of data it is.

The good, the bad, and the ugly – types of variables

There are two major types of variable – *categorical* variables and *metric* variables; each of them can be further divided into two subtypes, as shown in Figure 1.4.

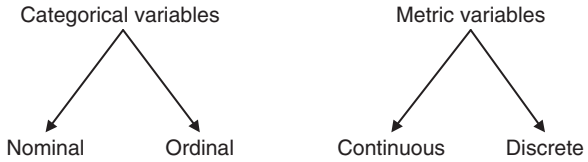

Figure 1.4 Types of variables

Each of these variable types produces a different type of data. The differences in these data types are of great importance – some statistical methods are appropriate for some types of data but not for others, and applying an inappropriate procedure can result in a misleading outcome. It is therefore critical that you identify the sort of variable (and data) you are dealing with *before* you begin any analysis, and we need therefore to examine the differences in data types in a bit more detail. From now on, I will be using the word 'data' rather than 'variable' because it is the data we will be working with – but remember that data come from variables. We'll start with categorical data.

Categorical data

Nominal categorical data

Consider the gender data shown in Figure 1.2. These data are *nominal categorical* data (or just nominal data for short).

The data are 'nominal' because it usually relates to *named* things, such as occupation, blood type, or ethnicity. It is particularly *not* numeric. It is 'categorical' because we allocate each value to a specific category. Therefore, for example, we allocate each M value in Figure 1.2 to the category Male and each F value to the category Female. If we do this for all 100 values, we get:

Male	265
Female	235

Notice two things about this data, which is typical of all nominal data:

- The data do not have any units of measurement.[1]
- The ordering of the categories is *arbitrary*. In other words, the categories cannot be ordered in any meaningful way.[2] We could just as easily have written the number of males and females in the order:

[1] For example, cm, seconds, ccs, or kg, etc.
[2] We are excluding trivial arrangements such as alphabetic.

Female	235
Male	265

By the way, allocating values to categories by hand is pretty tedious as well as error-prone, more so if there are a lot of values. In practice, you would use a computer to do this.

Exercise 1.3. Suggest a few nominal variables.

Ordinal categorical data

Let's now consider data from the Glasgow Coma Scale (GCS) (which some of you may be familiar with). As the name suggests, this scale is used to assess the level of consciousness after head injury. A patient's GCS score is judged by the sum of responses in three areas: eye opening response, verbal response, and motor response. Notice particularly that these responses are *assessed* rather than measured (as weight, height or temperature would be). The GCS score can vary from 3 (deeply unconscious) to 15 (fully conscious). In other words, there are 13 possible categories of consciousness.[3]

Suppose that we have two motor-cyclists, let us call them Wayne and Kylie, who have been admitted to the Emergency Department with head injuries following a road traffic accident. Wayne has a GCS of 5 and Kylie a GCS of 10. We *can* say that Wayne's level of consciousness is *less* than that of Kylie (so we can order the values) but *we can't say exactly by how much*. We certainly cannot say that Wayne is exactly half as conscious as Kylie. Moreover, the levels of consciousness between adjacent scores are not necessarily the same; for example, the difference in the levels of consciousness between two patients with GCS scores of 10 and 11 may not be the same as that between patients with scores of 11 and 12. It's therefore important to recognise that we cannot quantify these differences.

GCS data is *ordinal categorical* (or just *ordinal*) data. It is ordinal because the values can be meaningfully ordered, and it is categorical because each value is assigned to a specific category. Notice two things about this variable, which is typical of all ordinal variables:

- The data do not have any units of measurement (so the same as that for nominal variables).
- The ordering of the categories is *not* arbitrary, as it is with nominal variables.

The *seemingly* numeric values of ordinal data, such as GCS scores, are not in fact real numbers but only *numeric labels* which we attach to category values (usually for convenience or for data entry to a computer). The reason is of course (to re-emphasise this important point) that GCS data, and the data generated by most other scales, are *not properly measured* but *assessed* in

[3]The scale is now used by first responders, paramedics and doctors, as being applicable to all acute medical and trauma patients.

some way by a clinician or a researcher, working with the individual concerned.[4] This is a characteristic of all ordinal data.

Because ordinal data are not real numbers, it is not appropriate to apply any of the rules of basic arithmetic to this sort of data. You should not add, subtract, multiply or divide ordinal values. This limitation has marked implications for the sorts of analyses that we can do with such data – as you will see later in this book. Finally, we should note that ordinal data are almost always integer, that is, they have whole number values.

Exercise 1.4. Suggest a few more scales with which you may be familiar from your clinical work.

Exercise 1.5. Explain why it would not really make sense to calculate an average GCS for a group of head injury patients.

[4]There are some scales which may involve *some* degree of proper measurement, but these still produce ordinal values if even one part of the score is determined by a non-measured element.

Metric data

Discrete metric data

Consider the data in Figure 1.5. This shows the parity[5] of the mothers of the babies whose birthweights are shown in Figure 1.1.

0	0	2	0	0	3	3	1	0	3
0	0	0	0	1	0	3	2	3	1
2	2	3	1	10	0	1	0	1	5
1	0	1	0	0	0	0	0	0	0
2	0	0	0	2	1	0	2	2	0
1	0	0	0	0	0	1	0	0	0
0	0	2	2	3	2	2	0	3	1
0	4	0	0	2	1	0	0	0	1
3	3	0	3	0	0	6	0	1	0
2	2	1	2	4	1	0	2	1	0

Figure 1.5 Parity data (number of viable pregnancies) for the mothers whose babies' birthweights are shown in Figure 1.1

Discrete metric data, such as that shown in Figure 1.5, comes from *counting*. Counting is a form of measurement – hence the name 'metric'. The data is 'discrete' because the values are in discrete steps; for example, 0, 1, 2, 3 and so on. Parity data comes from counting – probably by asking the mother or by looking at records. Other examples of discrete metric data would include number of deaths, number of pressure sores, number of angina attacks, number of hospital visits and so on. The data produced are real numbers, and in contrast to ordinal data, this means that the difference between parities of 1 and 2 is exactly the same as the difference between parities of 2 and 3, and a parity of 4 is exactly twice a parity of 2.

In short:

- Metric discrete variables can be *counted* and can have units of measurement – 'numbers of things'.
- They produce data which are real numbers and are invariably integers (i.e. whole numbers).

[5]Number of pregnancies carried to a viable gestational age – 24 weeks in the United Kingdom, 20 weeks in the United States.

Continuous metric data

Look back at Figure 1.1 – the birthweight data.

Birthweight is a *metric continuous* variable because it can be measured. For example, if we want to know someone's weight, we can use a weighing machine; we don't have to look at the individual and make a guess (which would be approximate) or ask them how heavy they are (very unreliable). Similarly, if we want to know their diastolic blood pressure, we can use a sphygmomanometer.[6] Guessing or asking is not necessary. But, what do we mean by 'continuous'? Compare a digital clock with a more old-fashioned analogue clock. With a digital clock, the seconds are indicated in discrete steps: 1, 2, 3 and so on. With the analogue clock, the hand sweeps around the dial in a smooth, continuous movement. In the same way, weight is a continuous variable because the values form a continuum; weight does not increase in steps of 1 g.

Because they can be properly measured, these data *are* real numbers. In contrast to ordinal values, the difference between any pair of adjacent values, say 4000 g and 4001 g is exactly the same as the difference between 4001 g and 4002 g, and a baby who weighs 4000 g is exactly twice as heavy as a baby of 2000 g. Some other examples of metric continuous data include blood pressure (mmHg), blood cholesterol (μg/ml), waiting time (minutes), body mass index (kg/m^2), peak expiry flow (l per min) and so on. Notice that all of these variables have units of measurement attached to them. This is a characteristic of all metric continuous data.

Because metric data values are real numbers, you can apply all of the usual mathematical operations to them. This opens up a much wider range of analytic possibilities than is possible with either nominal or ordinal data – as you will see later.

To sum up:

- Metric continuous data result from *measurement* and they have units of measurement.
- The data are real numbers.

These properties of both types of metric data are markedly different from the characteristics of nominal and ordinal data.

Exercise 1.6. Suggest a few continuous metric variables which you are familiar with. What is the difference between assessing the value of something and measuring it?

Exercise 1.7. Suggest a few discrete metric variables which you are familiar with.

Exercise 1.8. What is the difference between continuous and discrete metric data? Somebody shows you a six-pack egg carton. List (a) the possible number of eggs that the carton could contain; and (b) the number of possible values for the weight of the empty carton. What do you conclude?

[6] We call the device that we use to obtain the measured value, for example, a weighing scale, a sphygmomanometer, or a tape measure, etc. a *measuring instrument*.

How can I tell what type of variable I am dealing with?

The easiest way to tell whether data is metric is to check whether it has *units* attached to it, such as g, mm, °C, μg/cm^3, *number of* pressure sores and *number of* deaths. If not, it may be ordinal or nominal – the former if the values can be put in any meaningful order. Figure 1.6 is an aid to variable-type recognition.

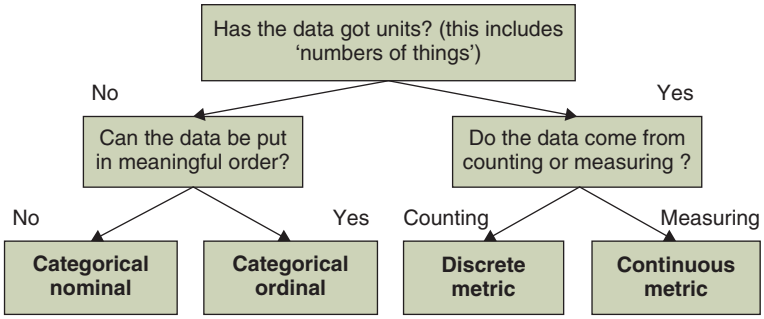

Figure 1.6 An algorithm to help identify data type

Exercise 1.9. Four migraine patients are asked to assess the severity of their migraine pain one hour after the first symptoms of an attack by marking a point on a horizontal line 100 mm long. The line is marked 'No pain' at the left-hand end and 'Worst possible pain' at the right-hand end. The distance of each patient's mark from the left-hand end is subsequently measured with an mm rule, and their scores are 25 mm, 44 mm, 68 mm and 85 mm. What sort of data is this? Can you calculate the average pain of these four patients? Note that this form of measurement (using a line and getting subjects to mark it) is known as a visual analogue scale (VAS).

The baseline table

When you are reading a research report or a journal paper, you will want to know something about the participants in the study. In most published papers, the authors will provide the reader with a summary table describing the basic characteristics of the participants in the study. This will contain some basic demographic information, together with relevant clinical details. This table is called the *baseline table* or the *table of basic characteristics*. In the following three exercises, we make use of the baseline tables provided by the authors.

Exercise 1.10. Figure 1.7 contains the basic characteristics of cases and controls from a case–control study[7] into stressful life events and the risk of breast cancer in women. Identify the type of each variable in the table.

[7] Do not worry about the different types of study; I will discuss them in detail in Chapter 6.

Variable	Breast cancer group ($n = 106$)	Control group ($n = 226$)	p value
Age	61.6 (10.9)	51.0 (8.5)	0.000*
Social class† (%):			
I	10 (10)	20 (9)	
II	38 (36)	82 (36)	
III non-manual	28 (26)	72 (32)	0.094‡
III manual	13 (12)	24 (11)	
IV	11 (10)	21 (9)	
V	3 (3)	2 (1)	
VI	3 (3)	4 (2)	
No of children (%):			
0	15 (14)	31 (14)	
1	16 (15)	31 (13.7)	0.97
2	42 (40)	84 (37)	
≥3	32 (31)†	80 (35)	
Age at birth of first child	21.3 (5.6)	20.5 (4.3)	0.500*
Age at menarche	12.8 (1.4)	13.0 (1.6)	0.200*
Menopausal state (%):			
Premenopausal	14 (13)	66 (29)	
Perimenopausal	9 (9)	43 (19)	0.000§
Postmenopausal	83 (78)	117 (52)	
Age at menopause	47.7 (4.5)	45.6 (5.2)	0.001*
Lifetime use of oral contraceptives (%)	38	61	0.000‡
No of years taking oral contraceptives	3.0 (5.4)	4.2 (5.0)	0.065§
No of months breastfeeding	($n = 90$)	($n = 195$)	
	7.4 (9.9)	7.4 (12.1)	0.990*
Lifetime use of hormone replacement therapy (%)	29 (27)	78 (35)	0.193§
Mean years of hormone replacement therapy	1.6 (3.7)	1.9 (4.0)	0.460*
Family history of ovarian cancer (%)	8 (8)	10 (4)	0.241§
History of benign breast disease (%)	15 (15)	105 (47)	0.000§
Family history of breast cancer¶	16 (15)	35 (16)	0.997§
Units of alcohol/week (%):			
0	38 (36)	59 (26)	
0–4	26 (25)	71 (31)	0.927‡
5–9	20 (19)	52 (23)	
≥10	22 (21)	44 (20)	
No of cigarettes/day:			
0	83 (78.3)	170 (75.2)	
1–9	8 (7.6)	14 (6.2)	0.383‡
≥10	15 (14.2)	42 (18.6)	
Body mass index (kg/m^2)	26.8 (5.5)	24.8 (4.2)	0.001*

*Two sample t test.
†Data for one case missing.
‡χ^2 test for trend.
§χ^2 test.
¶No data for one control.

Figure 1.7 Basic characteristics of cases and controls from a case–control study into stressful life events as risk factors for breast cancer in women. Values are mean (SD) unless stated otherwise. Source: Protheroe *et al.* (1999). Reproduced by permission of BMJ Publishing Group Ltd

Exercise 1.11. Figure 1.8 is from a cross-sectional study to determine the incidence of pregnancy-related venous thromboembolic events and their relationship to selected risk factors, such as maternal age, parity, smoking, and so on. Identify the type of each variable in the table.

Exercise 1.12. Figure 1.9 is from a study to compare two lotions, malathion and *d*-phenothrin, in the treatment of head lice in 193 schoolchildren. Ninety-five children were given malathion and 98 *d*-phenothrin. Identify the type of each variable in the table.

	Thrombosis cases ($n = 608$)	Controls ($n = 114\,940$)	OR	95%
Maternal age (y) (classification 1)				
≤19	26 (4.3)	2817 (2.5)	1.9	1.3, 2.9
20–24	125 (20.6)	23,006 (20.0)	1.1	0.9, 1.4
25–29	216 (35.5)	44,763 (38.9)	1.0	Reference
30–34	151 (24.8)	30,135 (26.2)	1.0	0.8, 1.3
≥35	90 (14.8)	14,219 (12.4)	1.3	1.0, 1.7
Maternal age (y) (classification 2)				
≤19	26 (4.3)	2817 (2.5)	1.8	1.2, 2.7
20–34	492 (80.9)	97,904 (85.2)	1.0	Reference
≥35	90 (14.8)	14,219 (12.4)	1.3	1.0, 1.6
Parity				
Para 0	304 (50.0)	47,425 (41.3)	1.8	1.5, 2.2
Para 1	142 (23.4)	40,734 (35.4)	1.0	Reference
Para 2	93 (15.3)	18,113 (15.8)	1.5	1.1, 1.9
≥Para 3	69 (11.3)	8429 (7.3)	2.4	1.8, 3.1
Missing data	0 (0)	239 (0.2)		
No. of cigarettes daily				
0	423 (69.6)	87,408 (76.0)	1.0	Reference
1–9	80 (13.2)	14,295 (12.4)	1.2	0.9, 1.5
≥10	57 (9.4)	8177 (7.1)	1.4	1.1, 1.9
Missing data	48 (7.9)	5060 (4.4)		
Multiple pregnancy				
No	593 (97.5)	113,330 (98.6)	1.0	Reference
Yes	15 (2.5)	1610 (1.4)	1.8	1.1, 3.0
Preeclampsia				
No	562 (92.4)	111,788 (97.3)	1.0	Reference
Yes	46 (7.6)	3152 (2.7)	2.9	2.1,3.9
Cesarean delivery				
No	420 (69.1)	102,181 (88.9)	1.0	Reference
Yes	188 (30.9)	12,759 (11.1)	3.6	3.0,4.3

Data presented as *n* (%).
OR, odds ratio; CI, confidence interval.

Figure 1.8 Table of baseline characteristics from a cross-sectional study of thrombotic risk during pregnancy. Source: Lindqvist *et al*. (1999). Reproduced by permission of Wolters Kluwer Health

Characteristic	Malathion ($n = 95$)	*d*-phenothrin ($n = 98$)
Age at randomisation (year)	8.6 (1.6)	8.9 (1.6)
Sex – no of children (%)		
Male	31 (33)	41 (42)
Female	64 (67)	57 (58)
Home no (mean)		
Number of rooms	3.3 (1.2)	3.3 (1.8)
Length of hair – no of children (%)*		
Long	37 (39)	20 (21)
Mid-long	23 (24)	33 (34)
Short	35 (37)	44 (46)
Colour of hair – no of children (%)		
Blond	15 (16)	18 (18)
Brown	49 (52)	55 (56)
Red	4 (4)	4 (4)
Dark	27 (28)	21 (22)
Texture of hair – no of children (%)		
Straight	67 (71)	69 (70)
Curly	19 (20)	25 (26)
Frizzy/kinky	9 (9)	4 (4)
Pruritus – no of children (%)	54 (57)	65 (66)
Excoriations – no of children (%)	25 (26)	39 (40)
Evaluation of infestation		
Live lice-no of children (%)		
0	18 (19)	24 (24)
+	45 (47)	35 (36)
++	9 (9)	15 (15)
+++	12 (13)	15 (15)
++++	11 (12)	9 (9)
Viable nits-no of children (%)*		
0	19 (20)	8 (8)
+	32 (34)	41 (45)
++	22 (23)	24 (25)
+++	18 (19)	20 (21)
++++	4 (4)	4 (4)

The two groups were similar at baseline except for a significant difference for the length of hair ($p = 0.02$; chi-square)

*One value missing in the *d*-phenothrin group square.

Figure 1.9 Baseline characteristics of the *Pediculus humanus capitis*-infested schoolchildren assigned to receive either malathion or *d*-phenothrin lotion. Source: Chosidow *et al.* (1994). Reproduced by permission of Elsevier

At the end of each chapter, you should look again at the chapter objectives and satisfy yourself that you have achieved them.