UNIT 1 SOME BASIC IDEAS

AIMS

To impart an understanding of some fundamental terms and concepts in statistics which together underpin much of what is to follow.

OBJECTIVES

At the end of Unit 1 you should be able to:

- Demonstrate that you appreciate the difference between samples and populations, and between sample statistics and population parameters.
- Explain the rationale behind statistical inference.
- Show that you understand the difference between (i) a nominal categorical, an ordered categorical and a metric variable; and (ii) a discrete and a continuous variable.
- Organise sample data using the most appropriate frequency and cumulative frequency distributions.
- Describe the shape of a frequency distribution in terms of skewness, symmetry, Normalness, etc.

Introduction

The rationale underlying all of the statistical methods described in this coursebook is the concept of **statistical inference**. Accordingly therefore we begin by examining what this concept is and what role it plays. To do that we need to define a few important terms.

A **population** is every single member of a defined group of interest. For example, it could be defined as, "every person in the UK who suffers from asthma" or "all children under the age of 5, living in Leeds and suffering from asthma" or "all children under the age of 5, living in Leeds, suffering from asthma, and registered with a particular GP", and so on. Our interest might be " the mean age at first diagnosis"*. But however a population is defined, we cannot generally study every member of a population (think of the difficulty of even identifying all the under 5's with asthma in Leeds, never mind locating and searching their records).

In practice therefore, if we want to discover something about any population we invariably have to use the information from a **sample** taken from that population. Provided the sample is reasonably *representative* of the population, we can then apply what we've discovered about the sample to the population from which the sample was taken. The most representative sample of all is what is called a **random** sample, for which every individual in a population has to have an equal chance of being picked for the sample.

Note that, while members of the populations (and hence of the samples) we study in medicine will most often be people (usually patients), they could also be cervical smears, or DNA material, or rats, and so on.

Back to our example. Suppose a random sample of 500 Leeds under 5's diagnosed with asthma gives an mean age of first onset equal to 5.5 years. We might then with reasonable confidence conclude that the **true** mean age of onset of **all** such Leeds under 5's (i.e. in the population) was also going to be *around* about 5.5 years. We say "around" because intuitively we know that no sample (even a random sample) will be *exactly* identical in every detail to its parent population, so we can't draw any *exact* conclusions about the population based on this (or any other) sample.

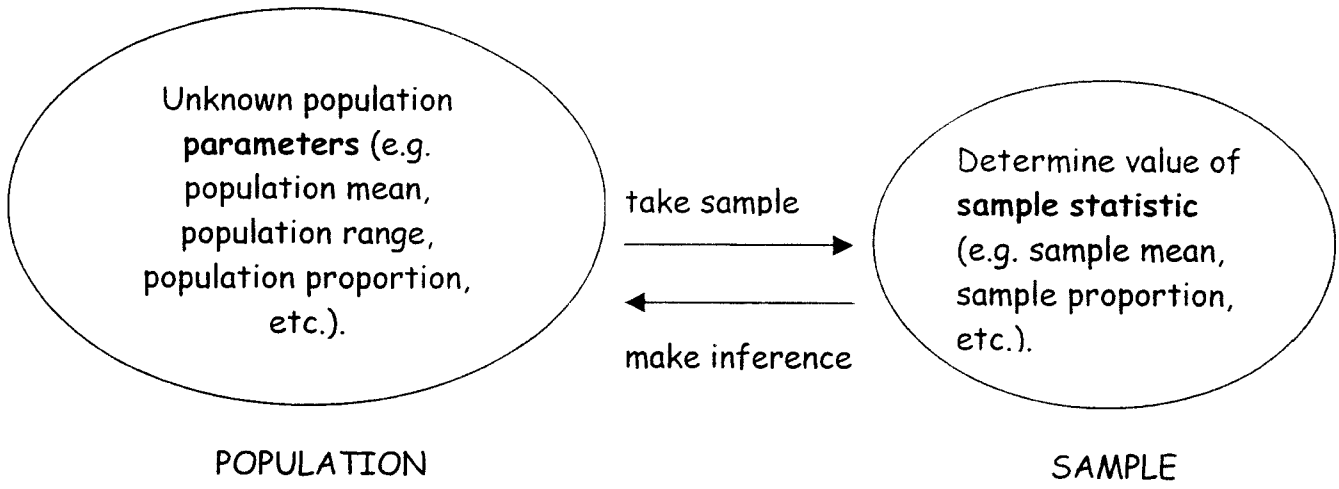
This process of making informed guesses or estimates of the characteristics of a *population* on the basis of sample results is known as **statistical inference**. The

* By "mean" we of course mean that which is commonly called the "average".

process of simply describing the principal features of a *sample* is known as **descriptive statistics.**

The particular value of the population which we wish to estimate (e.g. in the above example the mean age of first onset of asthma in Leeds under 5's) is called the **population parameter.** Other population parameters of interest might be the proportion of girls, or the range in the age of first diagnosis, and so on. It is important to note that we will never actually know the true value of any population parameter, the estimates we make are as close as we can get to the true value.

The sample value we use to estimate the value of the population parameter in question is called the **sample statistic.** For example, the sample mean age of first onset of 5.5 years in the example above is a sample statistic.



On its own **descriptive statistics** is a collection of methods to way of summarise the principal features of some sample data. However, this process also provides us with the sample statistic(s) we need for subsequent inferences about the population.

So statistical inference means using the value of a sample statistic to make an inform guess (called an "estimate") as to the true value of the corresponding population parameter.

Q. 1.1 In a random sample of 100 patients admitted to a psychiatric ward following an act of deliberate self-harm (dsh), 20 patients have a history of previous dsh.

(a) Which sample statistic can be calculated with this information? What is its value?
(b) Which population parameter can be estimated using this sample statistic? What would this estimate be?

(c) A second random sample of 200 patients is taken from the same population. Which sample is likely to lead to a more accurate estimate of the true value of the corresponding population parameter and why?

Q. 1.2 You want to estimate the mean age of menopause in a population of post-menopausal women prescribed HRT using a sample of 10000 such women. (a) What sample statistic would be the most useful? (b) How large would a sample have to be to provide the true population mean age?

Further reading: Bland, Section 3.3 or Bowers-1, pp. 5-8

Types of variable

A **variable** is a label we give to some attribute or property of the subjects in a study. The "value" of the attribute can change or *vary* from subject to subject and/or over time. For example, *blood type* is a variable whose value can vary from patient to patient, as is the variable *age* (which also changes over time). We can classify variables (and thus the data they produce) into two broad types, **categorical** and **metric**, and it is important to be able to identify the type of all of the variables in any study so that the most appropriate statistical method can be determined (we'll see why later).

I CATEGORICAL VARIABLES

Qualitative

There are two types of categorical variable (and thus tow corresponding types of data):

Nominal categorical

With **nominal** categorical (or nominal) data the "value" is one of a number of categories. For example, the variable "blood group" has four categories, O, A, B,

AB, into only one of which a patient can be classified. Crucially, the *ordering of the categories is arbitrary* (we could equally have ordered the blood groups AB, O, B, A, for example).

Ordinal categorical

With **ordered** categorical (or ordinal) data the "value" is again one of a number of categories. Now however the categories have an inherent *order*. For example, in response to the question, "How satisfied were you with your treatment?" patients might be required to respond "Very unsatisfied", "Unsatisfied", "Satisfied", or "Very satisfied". This is the natural **ordering** of these possible responses (although reverse order would be just as acceptable). Similarly, Glasgow Coma Scale scores (used to assess head injury) have the natural ordering of 3 (coma), 4, 5, 6, ... 14, to 15 (normal),. Notice that ordinal scores may be alphabetic, as in the first of these examples, or "numeric" (3, 4, 5, ... etc) as in the second.

The most important feature of such ordinal data is that the differences between adjacent scores are not necessarily the same, i.e. they cannot be exactly quantified. In other words although we know that a patient who is "very satisfied" with their treatment is *more* satisfied than a patient who is merely "satisfied", we don't know by exactly *how* much more.

Similarly the difference in wellbeing between a patient with a GCS score of 6 and one of 7 is not necessarily the same as the difference between patients with scores of 8 and 9 say. Nor is a patient with a score of 6 exactly half as well as a patient with a score of 12. So "numeric" ordinal scores are not proper numbers as such, and it is not appropriate therefore to apply the rules of arithmetic to them, i.e. they shouldn't be added, subtracted, multiplied, etc. This is a problem we will return to.

Categorical data is often referred to as *qualitative data*.

II METRIC VARIABLES

Quantitative

Metric variables and their data comes in two forms, discrete and continuous.

Discrete: outcome is one of a finite or countable number of possible values, e.g. parity (number of previous children), number of asthma attacks, number of deaths, etc. Discrete variables usually *count* things.

Continuous: outcome is one of a number of infinite possible values, e.g. birthweight (g), temperature (^{o}C), etc. Continuous variables usually *measure* things.

The crucial difference between metric data and ordinal data is that with metric data the difference between adjacent values is always the same. For example with birthweight (g), the difference between birthweights of say 3000g and 3001g is the *same* as that between 4568g and 4569g, and a baby weighing 4400g is *exactly* twice as heavy as one weighing 2200g. This allows us to use the rules of arithmetic with such data.

Note: Metric data always has **units** attached, e.g. length of femur (cm), birthweight (g), time on surgical waiting list (days), blood cholesterol concentration (mmol/ml), amount of blood transfused (ml), cervical smears (number of), patients (number of), etc. The algorithm shown in Figure 1.1 may help you identify variable types.

Q. 1.3 What type of variables are the following (and if metric, whether discrete or continuous?):

(a) number of visits to the GP in a year;
(b) marital status;
(c) size of tumour (cm);
(d) stages of breast cancer (I/II/III/IV);
(e) blood pressure (mmHg);
(f) age last birthday (years);
(g) day of week of road traffic accidents (Monday, Tuesday, etc.);
(h) grade for essay (A, B, C, etc.);
(i) occupational class (professional, administrative/ clerical, skilled manual, etc.);
(j) age group last birthday (0-19 years, 20-29 years, etc.);
(k) Pressure Sore Risk Assessment Scale Scores (ranges from 0=little or no risk of developing pressure sores to 20=high risk of developing pressure sores).

Further reading: Bland p.46; or Bowers-1, Chapter 2.

Q. 1.4 Identify the type of each variable (shown in **bold** typeface - but excluding Pruritus and Excoriations) in Figure 1.2, taken from a study comparing two lotions for the treatment of nits?

Q. 1.5 Identify the type of the following variables listed in Figure 1.3 taken from a study into the prognosis for acute back pain: (a) age; (b) sex; (c) duration of index episode; (d) initial visual analogue pain scale score*: (e) initial disability questionnaire score.

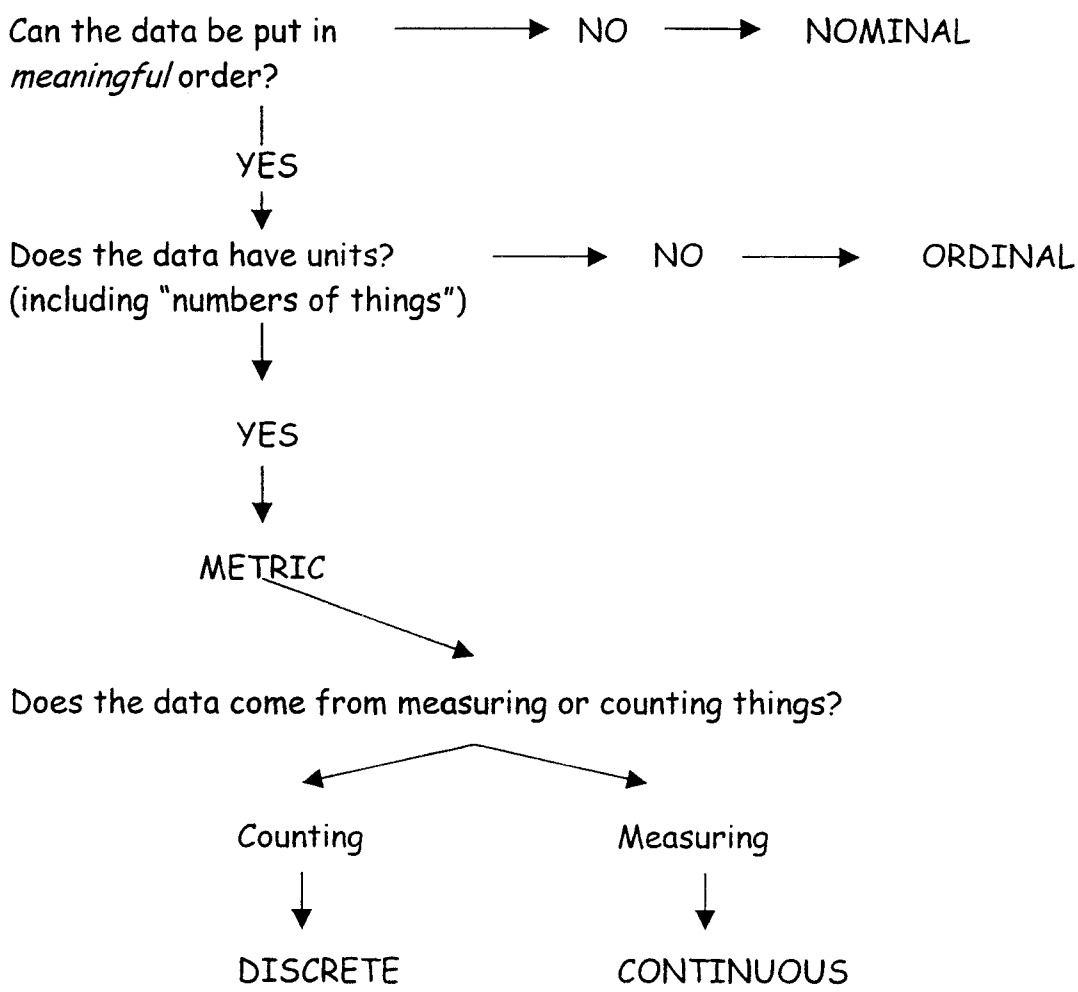


Figure 1.1 Algorithm for identifying type of data

* A visual analogue scale is a horizontal line (typically 10cm long) drawn on a piece of paper with "no pain" at the left-hand end and "the worst possible pain" at the right-hand end. A patient is asked to mark the line at the point which corresponds to their pain. Afterwards a 10cm rule is placed over the line and the "size" of the pain, in cm bands is measured. So a mild pain, for example, might score 1, a severe pain 8.

Characteristic	Malathion (n=95)	d-phenothrin (n=98)
Age at randomisation (yr)	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 (45)
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 2 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.

Table 2: **Baseline characteristics of the *P humanus capitis*-infested schoolchildren assigned to receive malathion or d-phenothrin lotion***

Figure 1.2 Baseline characteristics of subjects in nit lotion study. The Lancet, 344, 1994.

TABLE I—*Baseline characteristics of subjects (n=103) at entry to study. Except where stated otherwise, values are numbers (percentages) of subjects*

	Value
Sociodemographic variables:	
Mean (SD) age (years)	46·5 (14·3)
Male sex	62 (60)
French nationality	92 (89)
Manual worker	29 (28)
Employed at entry	75 (73)
Back pain history:	
One or more previous acute episodes	63 (61)
Previous chronic (>3 months) episode of low back pain	8 (8)
Prior back surgery	0
Median (minimum, maximum) duration of index episode (hours)	26 (1·5, 70)
Sudden onset (<2 minutes)	36 (35)
Pain and disability variables:	
Mean (SD) initial visual analogue scale score	6·6 (1·8)
Constant pain at night	16 (16)
Pain aggravated by impulsion	44 (43)
Pain aggravated by moving back	99 (96)
Pain worse on standing	67 (65)
Pain worse on lying	27 (26)
Unable to stand even briefly	18 (17)
Mean (SD) initial disability questionnaire score†	12·1 (5·6)
Physical findings:	
Limited passive movements	72 (70
Catch	61 (59)
Straight leg raising <75°	31 (30)
Psychosocial variables:	
DSM-III-R diagnosis	12 (12)
Depression	5 (5)
Generalised anxiety	7 (7)
Compensation status‡	9 (9)
Job difficulty (heavy labour)	16 (16)
Poor job satisfaction	34 (33)

†If able to stand.
‡Invariably awarded in France for pain occurring at work.

Figure 1.3 Baseline characteristics of subjects in back pain study. BMJ, 1994, 308, 577-80.

Frequency distributions

A frequency distribution is a description of the way in which the values of a variable are distributed across its possible range. A frequency **table** is a simple way of presenting this information succinctly. It records the number (i.e. the **frequency**), or the percentage (**relative frequency**), of values which lie in each category or class. As an example, Table 1.1, based on Figure 1.2, shows the frequency and relative (or %) frequency distributions of hair colour.

Hair colour	Frequency	Relative frequency (%)
Blond	15	16
Brown	49	52
Red	4	4
Dark	27	28
Totals	95	100

52=(49/95) × 100

Table 1.1 Frequency distribution of hair colour in nits study

In addition, we can also calculate **cumulative** and *relative* **cumulative** frequencies, i.e. the numbers (or %) of observations *less than or equal* to specified levels. To find cumulative frequencies we add up (i.e. cumulate) the frequencies starting with the first (top) value and working down the frequency column. For example, Figure 1.4 shows the frequency (and relative frequency) distributions for the Disability Rating Scale scores of 28 patients with traumatic brain injury (low scores good, high scores bad).

Table 3
Functional Status as Shown by the Disability Rating Scale

Disability Rating Scale Score*	Frequency	Percentage
0	1	3.6
1	9	32.1
2	2	7.1
3	5	17.9
4	5	17.9
5	3	10.7
8	2	7.1
9	1	3.6

Note. The Disability Rating Scale scores were assigned at the time of the interview.
*0 = none, 1 = mild, 2–3 = partial, 4–6 = moderate, 7–11 = moderately severe.

Figure 1.4 Disability Rating Scale scores for 28 patients with Traumatic Brain Injury. Amer J Occup Therapy, 48, 1994

Table 1.2 shows in addition to the information in Figure 1.4, the cumulative frequency and % cumulative frequency DRS scores. Table 1.2 tells us, for example, that 5, or 17.8%, of the patients had a DRS score of 3, and further that 17, or 60.7%, had a DRS score of *3 or less*.

DRS score	Frequency	Relative (or %) frequency	Cumulative frequency	Cumulative relative (or %) frequency
0	1	3.7	1	3.7
1	9	32.1	10	35.8
2	2	7.1	12	42.9
3	5	17.8	17	60.7
4	5	17.8	22	78.5
5	3	10.7	25	89.2
6	0	0.0	25	89.2
7	0	0.0	25	89.2
8	2	7.1	27	96.3
9	1	3.7	28	100.0
totals	28	100.0		

Table 1.2 Disability Rating Scale scores for a sample of 28 patients with traumatic brain injury (from Table 1.4)

Q. 1.6 What % of patients in Table 1.2 had a DRS score of: (a) 5 or less; (b) more than 5?

With ordinal or metric data, tables of **grouped** frequencies can be calculated. These provide a record of the number (and/or percentage) of observations within certain intervals or *groups*.

Table 1.3 shows such a grouped frequency distribution for the age of 138 women attending a family planning clinic and diagnosed as having endometriosis.

Q. 1.7 What number and percentage of women in the endometriosis sample (Table 1.3) are aged: (a) less than 40? (b) 45 or more?

Age group (years)	Frequency (number of cases)	Relative (%) frequency	Cumulative frequency	Relative (%) cumulative frequency
25-29	3	2.2	3	2.2
30-34	14	10.1	17	12.3
35-39	42	30.4	59	42.8
40-44	58	42.0	117	84.8
45-49	18	13.0	135	97.8
≥ 50	3	2.2	138	100.0
totals	138	99.9*		

Table 1.3 Frequency (& %) & cumulative frequency (& %) for age of endometriosis women. Source: BMJ, 306, 1993.

Q. 1.8 The data in Table 1.4 contains values for the % mortality rate (deaths ÷ admissions) × 100, in a sample of 26 intensive care units in the UK. Construct a grouped frequency table for this data, showing frequency, relative frequency, and cumulative relative frequency columns (use groups 10.0-14.9, 15.0-19.9, etc.

ICU number	% mortality		ICU number	% mortality		ICU number	% mortality
1	15.2		10	29.4		19	18.9
2	31.3		11	21.1		20	13.7
3	14.9		12	20.4		21	17.7
4	16.3		13	13.6		22	27.2
5	19.3		14	22.4		23	19.3
6	18.2		15	14.0		24	16.1
7	20.2		16	14.3		25	13.5
8	12.8		17	22.8		26	11.2
9	14.7		18	26.7			

Table 1.4 Mortality rates in 26 intensive care units. BMJ, 307, 1993

Q. 1.9 Why does it make no sense to calculate cumulative frequencies for nominal categorical data (such as that for hair colour)?

* Without rounding errors this total would equal 100.

Further reading: Bland, pp. 47-50
or Bowers-1, pp. 37-46; 48

Shapes of distributions

In addition to the type of data involved, the "shape" of the distribution of ordinal and metric data often influences the type of analysis performed. We can best judge the shape of a particular distribution from an appropriate graph - we'll see how in Unit 2, but some terms used to describe shapes are:

Positively skewed	Most values on the low side with fewer high values.
Negatively skewed	Most values on the high side with fewer low values.
Symmetric	The distribution of values has a fairly similar pattern either side of the "middle" of the values.
Mound-shaped	A symmetrical distribution with a dome-like shape.
Normal	A symmetric distribution with a special, smooth bell-shaped curve (very important in statistics).
Bimodal	Values form two distinct peaks.

Q. 1.10 (a) How would you describe the shape of the distributions in Figure 1.5, which displays the age distribution of NHS nursing and non-nursing staff. (b) The data represented is grouped metric continuous for which a histogram is the appropriate chart type. Why do you think the authors of this study chose to treat the age groups as if they were categorical and plot them using a multiple bar chart? (c) What alternative charts might they have used?

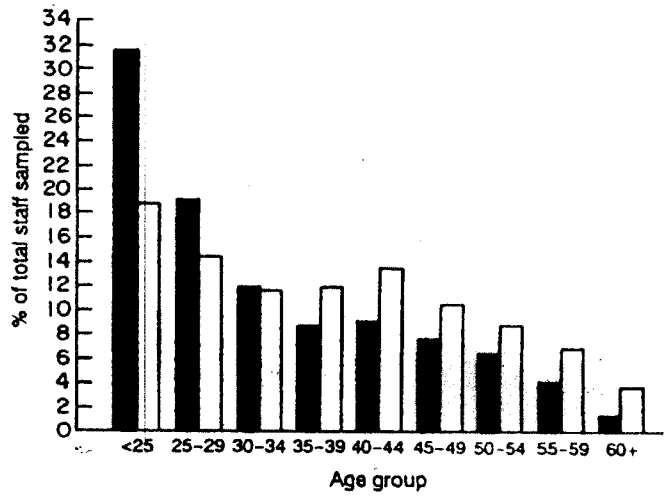


Figure 1 The age distribution of National Health Servic labour, comparing nursing and non-nursing staff. ■, Nursin staff; □, non-nursing staff.

Figure 1.5 Clustered bar chart showing age distribution of NHS workforce. J Advanced Nursing 1994.

The Normal distribution

The shapes of the distributions of many human clinical variables, such as height, hemoglobin levels, blood pressure, etc., often have the special bell shape we call the Normal* distribution. Other variables can often be made to follow a Normal distribution by applying a simple transformation (for instance by taking logarithms of the original data). The Normal distribution plays a very important role in statistical inference (which we will come to later).

The Normal distribution is a symmetric distribution, for which the mean, median and mode are all equal (in the middle of the distribution). Most of the values cluster around this middle of the distribution, with progressively fewer and fewer values below and above the centre.

The histogram in Figure 1.6 shows the birthweights (g) of a sample of 952 babies born to Jamaican women in 1997. A Normal curve is superimposed on the distribution and shows a reasonably good fit. On this evidence, we would describe the distribution of birthweights as being Normal.

* We generally use a capital N for Normal to distinguish the word from its non-statistical counterpart.

If data is distributed Normally it has some very useful area properties which we will return to in Unit 4.

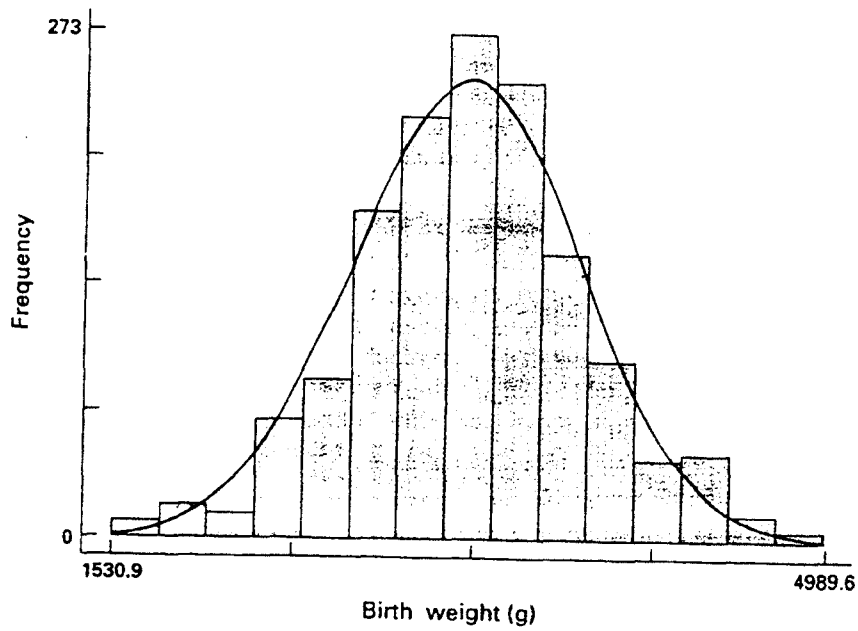


Figure 1 Distribution of birth weights in Jamaican babies.

Figure 1.6 Histogram of birthweights in Jamaican babies with a Normal curve superimposed. J Epid & Community Health, 51, 1997

Solutions to Coursebook Questions

UNIT 1: Basic Statistical Ideas

Q. 1.1 (a) The sample proportion who have a history of previous dsh. 0.20 (20%)
(b) True (population) proportion of those with previous history of dsh. 0.2.
(c) The second with n=200. The larger the sample the more likely it is to be representative of the population.

Q. 1.2 (a) The sample mean age at menopause. (b) It would have to be the same size as the population, i.e. it would have to *be* the population.

Q. 1.3 (a) metric (discrete)
(b) nominal
(c) metric (continuous)
(d) ordinal
(e) metric (continuous)
(f) metric (discrete)
(g) nominal
(h) ordinal
(i) ordinal
(j) ordinal (underlying distribution is metric continuous)
(k) ordinal

Q. 1.4 (a) age is metric continuous.
(b) sex is nominal categorical.
(c) number of rooms is metric discrete.
(d) hair length is ordinal.
(e) hair colour is nominal.
(f) hair texture is nominal.
(g) live lice is ordinal.
(h) viable nits is ordinal.

Q. 1.5 (a) age is metric continuous.
(b) sex is nominal.
(c) duration of index episode is metric continuous.
(d) initial visual analogue pain score is ordinal.
(e) disability questionnaire score is ordinal.

Q. 1.6 (a) 89.2% (b) 10.8%.

Q. 1.7 (a) 59 or 42.8%; (b) 21 or 15.2%.

Q. 1.8

% mortality	Frequency	% frequency	cumulative frequency	% cumulative frequency
0.0-9.9	0	0	0	0
10.0-14.9	9	34.6	9	34.6
15.0-19.9	8	30.8	17	65.4
20.0-24.9	5	19.2	22	84.6
25.0-29.9	3	11.5	25	96.2
30.0-34.9	1	3.8	26	100.0

Q. 1.9 Because the ordering of the categories is arbitrary.

Q. 1.10 (a) There are more nursing than non-nursing staff up to age of 34, but situation is reversed for those aged more than 34; (b) They wanted to be able to compare nursing and non-nursing staff in each age category; (c) Could have used two histograms, dotplots or boxplots.