UNIT 3 DESCRIBING DATA

PART I - SUMMARY MEASURES OF LOCATION

AIMS

To show how sample data may be summarised in terms of its location.

OBJECTIVES

At the end of Unit 3 you should be able to:

- Explain the differences between the mode, the median and the mean as measures of location.
- Identify, calculate and interpret the most appropriate measure of location for the sample data in question.
- Demonstrate an understanding of and be able to calculate and interpret centile and quartile values.

Reading: Bland: Section 4.5, 4.6.
or Bowers-1: Chapter 5 (ignoring computer applications).

Introduction

In Unit 1 we saw that we can use a sample statistic to make intelligent guesses or *estimates* of the true value of the corresponding population parameter. This is the process of statistical inference. Provided the sample is reasonably representative of the population from which it was taken, our estimate of the true value of the population parameter should be reasonably close. Since we will never in general know the actual value of the population parameter we must hope that this is true.

So we might, for example, use the proportion of a sample of schoolchildren who suffer from asthma as a guide to the true proportion, i.e. the population proportion, with this illness. When we calculate values for the sample statistics

we are effectively describing the main features of the sample. This process is called *descriptive statistics*.

There are two summary descriptive measures in particular that are useful. The first provides information about the value around which most of the sample values tend to congregate. Measures such as these are known as *measures of location or central tendency*. The second tells us how spread out the sample values are. These measures are known as *measures of spread or dispersion*. We will consider measures of location in this unit and measures of spread in the next.

In Units 1 we saw that there are three types of variable (nominal, ordinal and metric), and in Unit 2 that the shape of the sample frequency distribution can be described in terms of symmetry, skewness, Normalness, etc. As we will see, the choice of the most appropriate measures of location and spread depend (sometimes crucially) on variable type and distributional shape.

There are three commonly-used summary measures of location to choose from: the mode; the median; and the mean.

The mode

Although the mode appears rarely in the literature, it can on occasion be a useful measure. The mode is the value which *occurs most often* in the sample data set, and in this sense is a measure of *typical-ness*. The mode is most appropriately applied to categorical variable data (nominal or ordinal). One disadvantage of the mode is that it may not be a unique value, i.e. there may be more than one mode for a given set of sample data. If so then this is clearly not very helpful as a summary measure. Also we can't combine two or more modal values to find the overall mode. However, the mode has the advantage of not being sensitive to outliers*. Moreover the value of the mode is always equal to one of the original sample values.

The easiest way to determine the mode is to arrange the sample values in ascending order and then count how many times each value occurs (in other words construct a frequency table).

Q. 3.1 What is the modal DRS score in Table 1.2 (Unit 1)?

---

* Recall that an *outlier* is a sample value (there may be more than one) which lies a long way (either much smaller or much larger, or both) from the general mass of values.

The median

The median offers a summary measure of *central-ness*. When the data is arranged in ascending order, the median is the middle value. So half of the sample values will be smaller than the median, and half larger. It can used as a measure of location with both ordinal and metric data, particularly when the latter has a skewed distribution (we will see why shortly).

For example, suppose the pulse rates (in beats per min) of 7 patients are arranged in ascending order thus:

66 72 72 84 88 95 96

The median is 84bpm since this is the middle value. Notice that there are three values less than 84 and three greater.

If there are an even number of values, the median is the average of the two central values. For example, with an eighth patient:

66 72 72 84 88 95 96 99

The two centre values (the 4th and 5th values) are 84 and 88, whose average is 86, so the median is 86 bpm.

The median has the advantage of not being sensitive to outliers, because only the middle value(s) determine its value. Extreme values are simply ignored. The median does not therefore use all of the information in the sample data. Moreover, unlike the mode, the median has a unique value. A disadvantage is we can't combine two or more medians to find an overall median of medians.

For the mathematically minded, after putting the data in ascending order, the median is equal to the $\frac{1}{2}(n + 1)$th value, where n is the number of values. Thus in the last example where n was 8, $\frac{1}{2}(n + 1) = \frac{1}{2}(8 + 1) = 4\frac{1}{2}$. So the median is the value of the $4\frac{1}{2}$th value, i.e. the average of the 4th and 5th values, as we found above.

For example, let's find the median time to onset for the data in Table 3.1, taken from a study into the possible relationship between the use of gangliosides (a form of glycolipid found in nerve tissue and marketed in several countries as a treatment for many neurological diseases) and the development of Guillain-Barré syndrome, researchers gathered data from a sample of 24 Italian Guillain-Barré sufferers. Table 3.1 contains sample data for each patient on three variables: their age (years); the time from the start of ganglioside treatment to the onset of Guillain-Barré syndrome (days); and their residual neurological deficit at follow-up some months later.

Patient	Age (years)	Time from gangliosides to onset of syndrome (days)	Residual neurological deficit
1	64	11	Moderate
2	44	10	Moderate
3	36	14	Moderate
4	45	10	Moderate
5	61	13	Severe
6	40	11	Moderate
7	61	10	None
8	70	6	Dead
9	58	11	Moderate
10	60	4	Severe
11	58	8	Severe
12	40	8	Severe
13	69	18	Moderate
14	31	6	Moderate
15	49	15	Severe
16	42	10	None
17	39	7	Moderate
18	31	15	None
19	58	5	None
20	80	14	Moderate
21	83	13	Dead
22	62	6	Severe
23	83	16	Moderate
24	63	4	Severe

Table 3.1 Data on Guillain-Barré disease in 24 patients. BMJ, 307, 1993.

There are 24 values, an even number, so the median is the average of the 12^{th} (=10 days) and 13^{th} (=10 days) values. That is, the average of 10 and 10, which is 10 days. So half the patients had a time to onset of 10 or fewer 10 days, while half had a time to onset of more than 10 days.

Q. 3.2 Determine the sample median DRS scores shown in Table 1.2 (Unit 1)?

Q. 3.3 Use the cumulative frequency curve in Figure 2.11 (Unit 2) to estimate the sample median blood cholesterol of the control patients.

Q. 3.4 Calculate the median age from the age data in Table 3.1.

Centiles and quartiles

As we've seen, the median divides the frequency distribution into two equal parts with half the values below it and half above. Centiles divide the distribution into 100 equal parts.

For example the 23th percentile has 23% of the sample values below it (and thus 77% above). The 25th percentile has 25%, or a quarter, of the sample values below it (and 75% above). The 25th percentile is also known as the 1st quartile and denoted Q1. The 75th percentile, or third quartile Q3, has 75 per cent (three-quarters) of the values below it (and 25% above it).

The 50th percentile has 50% of the values below it, and is also known as the second quartile, Q2. In other words *the median is also the 2nd quartile.*

Quartiles and percentiles can be estimated from cumulative frequency curves, or by calculation using the same methods we used with the median described above. To find say the 25th percentile (the second quartile) we need the value below which a quarter of the values lie, and above which three quarters of the values lie (after arranging the values in ascending order).

Calculations of exact percentiles and quartiles can be a little problematic, especially with small samples. No consensus exists on the correct procedures. In view of this it is perhaps better to adapt the formulae we encountered above. At least this offers a consistent method.

We have already seen that the median is the $\frac{1}{2}(n + 1)$th value, which we could write as:

$$\frac{50}{100}(n+1)$$

Similarly, the 25th percentile (or the first quartile) is the $\frac{25}{100}(n+1)$ th value, and so on for other percentiles.

Q. 3.5 Estimate (a) the 20th percentile; (b) the 1st and 3rd quartiles, levels of cholesterol in the control patients using the ogive in Figure 2.11.

Q. 3.6 Calculate Q1 (25th percentile) and Q3 (75th percentile) for the age data in Table 3.1

The mean

The *mean* (strictly speaking the arithmetic mean) is a *measure of average-ness* and is calculated in the usual way for any average: by adding up all of the values and dividing the sum by the number of values. It is only appropriate with metric data and shouldn't be used with ordinal data (although often is). Although the mean is sensitive to the presence of outliers in the data, is does use all of the information in the sample data set, and we can combine two or more means to find an overall mean.

Q. 3.7 Why does it make no sense to calculate the mean of the hair colour data in Table 1.1?

Q. 3.8 Why is the median sometimes a more representative measure of location than the mean when the frequency distribution contains outliers?

Relationship between measures of location and skew

If we know the values of the mean and median for a distribution this can provide insight into the skewness or otherwise of the distribution, thus:

If mean > median then the distribution positively skewed
(i.e.
long tail to right).

If mean < median then the distribution negatively skewed
(i.e.
long tail to left).

If mean = median then the distribution symmetric.

Q. 3.9 Calculate the mean and median times to onset of G-B syndrome using the data in Table 3.1. What do the two values tell you about the skewness of the time-to-onset distribution?

Q. 3.10 What measures of location did the authors use in Figure 1.3 (Unit 1) to summarise: (a) age; (b) duration of index episode; (c) initial visual analogue pain scale score; (d) initial disability questionnaire score?

Choosing an appropriate measure of location

Table 3.2 is an advisory guide to choosing the most appropriate measure of location for data, depending on variable type and the shape of the distribution. Thus it is not wrong (nor impossible) to calculate the mode for continuous metric data, but it is unlikely to produce a useful measure. Neither is it wrong to calculate the mean for skewed metric data, but the value obtained may produce an unrepresentative (even misleading measure) influenced as the mean is by outliers.

Type of variable ↓	Measure of location		
	Mode	Median	Mean
Nominal	YES	NO	NO
Ordinal	YES	YES	NO
Metric (discrete)	NO (unless no. values small)	YES	YES*
Metric (continuous)	NO	YES	YES*

Table 3.2 Guide to choosing appropriate measure of location

* If the distribution is skewed, the median may be more representative.

Q. 3.11 With the help of the guidelines set out in Table 3.2, how appropriate do you think was the authors' choice of summary measures of location in Question 3.11? Remember that you have already identified variable type in Q. 1.5.

Solution to coursebook questions

Unit 3: Descriptive Statistics I -- Measures of Location

Q. 3.1 The modal DRS score is 1 (the largest frequency of 9 or 32.1%). Seems a reasonable summary value although 64.2% of patients have a higher score.

Q. 3.2 The first step is to write the individual DRS values in ascending order:

Value	0	1	1	1	1	1	1	1	1	2	2	3	3	3	3	3	4	4	4	4	4	5
Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23

Value	5	5	8	8	9
Position	24	25	26	27	28

There are an even number of values, 28, and so the median is the average of the two in the centre, i.e. the 14th and 15th values. The 14th value is 3, the 15th value is also 3, so the average (and the median) is 3.

Or we could have used the formulae: $\frac{1}{2}(28 + 1) = \frac{1}{2} \times 29 = 14.5$. So the median is the 14.5th value, i.e. 3, as before.

Q. 3.3 About 6 mmol/l (this is of course the same answer as Q2.10(b)).

Q. 3.4 Arranging ages in ascending order:

Value	31	31	36	39	40	40	42	44	45	49	58	58	58	60	61	61	62	63
Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

Value	64	69	70	80	83	83
Position	19	20	21	22	23	24

mode = 58 years; median = average of 12th and 13th observations = 58.

Q. 3.5 (a) 5.2mmol/l (b) 5.5mmol/l and 7.0mmol/l

Q. 3.6 We first need to order the age data, which we have already done for Q. 3.4.

To calculate percentiles it is easier to use the formulae rather than trying to count values. We know that the 25th percentile, i.e. Q1, is the $\frac{25}{100}(n+1)$ th value, where in this example n = 24.

So Q1 = 25/100× (24 + 1) = 6.25th value. That is, it is the value a quarter of the way from the 6th value to the 7th value. The sixth value is 40 and the 7th value is 42. So Q1 = 40.5 years.

Similarly, Q3 = 75/100× (24 + 1) = 18.75th value. The 18th value is 63 and the 19th value is 64, so Q3 = 63.75 years.

Q. 3.7 Because the data is not metric and the category ordering is arbitrary.

Q. 3.8 Because the mean is influenced by the presence of outliers and might therefore be thought to give an unrepresentative picture of location. The median is not influenced by outliers and might therefore be a more representative figure.

Q. 3.9 After arranging the time-to-onset data in ascending order the median is found to be 10 days (the value of the 12.5th score - check). The mean time to onset = 245/24 = 10.2 days. Since the mean (10.2) > median (10) then the distribution appears to be almost symmetric.

Q. 3.10 (a) mean; (b) median; (c) mean; (d) mean.

Q. 3.11 (a) appropriate; (b) appropriate (could have used mean since data is metric, but might have chosen median if distribution is skewed); (c) inappropriate, visual analogue scale produces ordinal data; (d) inappropriate, disability questionnaire produces ordinal data.