UNIT 4 DESCRIBING DATA

PART II – MEASURES OF SPREAD

AIMS

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

OBJECTIVES

At the end of Week 4 you should be able to:

- Explain the meaning of, and the differences between, the range, the interquartile range and the standard deviation.
- Calculate the most appropriate measure of spread for the sample data in question.
- Explain the area properties of the Normal distribution in terms of the mean and standard deviation.

Reading: Bland: Section 4.8.
or Bowers: Chapter 6 (ignoring computer applications).

Introduction

In Unit 3, we discussed one of two important summary descriptive measures - measures of location. In this unit we want to discuss the second important summary descriptive measure - measures of spread. Measures of location provide summary measures of the value around which the values tend to congregate. Measures of spread or dispersion provide summary measures of the extent to which the values are spread out.

There are three principal measures of spread: the range, the interquartile range (iqr), and the standard deviation (s.d.). The two range-based methods are used with numeric ordinal data or with skewed metric data; the standard deviation

only with metric data. There are no commonly used measures of spread for nominal data (although some measures do exist - for example, see Bowers-1, pp. 118-23).

The range

The range is the distance from the smallest value to the largest value in the data set, i.e.

Range = (smallest value to largest value)

Q. 4.1 What is the range of the duration of the index episode given in Figure 1.3?

Q. 4.2 (a) The range is highly sensitive to outliers. Why? (b) Calculate the range for the age data in Table 3.1 (Unit 3).

The interquartile range

The interquartile range is the distance from the first to the third quartiles and written as (Q1 to Q3), i.e.

iqr = (Q1 to Q3)

So the iqr measures the spread of the middle 50% of values.

Q. 4.3 (a) Using the results you obtained in Q. 3.5 for the quartiles Q1 and Q3 of blood cholesterol, what is the inter-quartile range for the cholesterol level in the control patients. (b) The inter-quartile range is not sensitive to outliers. Why? .

Q. 4.4 Estimate the iqr for age in Table 3.1 (you should have already calculated Q1 and Q3 in answering Q 3.7).

Q. 4.5 What is: (a) the range; and (b) the inter-quartile range; for the time to give an analgesic injection using the "traditional" method, shown in the boxplot of Figure 2.10 (Unit 2)?

Q. 4.6 Figure 4.1 shows the baseline characteristics for a sample of patients in a study to control post-operative phantom stump pain. The "blockade" group is the one receiving the treatment. In the control group, interpret the iqrs for: (a) pain in week before amputation; (b) daily opioid consumption at admission;?

Characteristics of patients	Blockade group (n=27)	Control group (n=29)
Men/women	15/12	18/11
Mean (SD) age in years	72·8 (13·2)	70·8 (11·4)
Diabetes	10	14
Concurrent treatment because of cardiovascular disease	18	19
Previous stroke	3	2
Previous contralateral amputation	7	3
Median (IQR) pain in week before amputation (VAS, 0–100 mm)	51 (23·8–8–78)	44 (25·3–68)
Median (IQR) daily opioid consumption at admission (mg)	50 (20–68·8)	30 (5–62·5)
Level of amputation		
Below knee	15	16
Through knee-joint	5	2
Above knee	7	11
Reamputations during follow-up	3	2
Died during follow-up	10	10

Table 1: **Baseline characteristics of patients**

Figure 4.1 Baseline characteristics of subjects in post-operative stump pain study. The Lancet, 350, 1997.

Standard deviation

The standard deviation (often abbreviated as s.d. or sd) is the most widely used measure of spread (but is only appropriate for metric data). We can loosely interpret s.d. as a measure of the distance, on average, of all of the data values from the mean. The larger the value of s.d. the further away the values are collectively, from the mean, i.e. the more spread out are the data values. If you have a calculator with an s.d. function then you can use it. If not, to calculate the sample s.d. by hand follow these steps:

Step 1. Calculate the mean.

Step 2. Subtract the mean from every value.

Step 3. Square each of the values obtained in Step 2. Add these squared values together to give the *sum of the squares* value.

Step 4. Divide the final result obtained from Step 3 by the total number of values in the sample minus 1, i.e. divide the sum of squares by (n-1).

Step 5. Take the square root of the result obtained in Step 4. This is the s.d. Note that the s.d. is measured in the same units as the original data.

Q. 4.7 (a) Do you think the s.d. is sensitive to the presence of outliers in the data? Why? (b) Calculate the s.d. for the age data in Table 3.1. (c) How would you interpret the result?

Q. 4.8 In Figure 4.1, which of the two groups(blockade or control) is: (a) older on average; (b) has a wider spread of ages?

Area properties of the Normal distribution

In Unit 2 we referred to some important area properties of the Normal distribution. Now that we know about the standard deviation we can have a look at what these properties are.

If data is distributed Normally then about:

68% (or two-thirds) of the values will lie between the mean and one s.d. either side of the mean;

95% of the values will lie between the mean and two s.d.s either side of the mean; and,

99% of the values will lie between the mean and three s.d.s either side of the mean.

For example, referring back to the distribution of sample birthweight data shown in Figure 1.6 (Unit 2). Suppose the sample mean birthweight is 3261g and the s.d. is 577g. Then the area properties tell us that about two thirds of the babies in the sample will have a birthweight of between 3261g ± 1 × 577g, i.e. from 2684g to 3438g. About 95% will have a birthweight between 3261g ± 2× 577gHg, i.e. from 2107g to 4415g. Finally, virtually the whole of the sample (99%) will have a birthweight between 3261g ± 3× 577g, i.e. from 1530g to 4992g.

But remember that this is only true for perfectly Normal distributions. The further away from Normal the distribution is, the more approximate the area properties become.

Q. 4.9 In Figure 4.1, if we assume that age in the blockade group is Normally distributed, what ages will include 95% of the blockade patients?

Choosing an appropriate measure of spread

Table 4.1 is a guide to choosing the most appropriate measure of spread.

Type of variable	Measure of spread		
	Range	iqr	s.d.
Nominal	NO	NO	NO
Ordinal	YES	YES	NO
Metric (discrete)	YES	YES	YES
Metric (continuous)	YES	YES	YES

Table 4.1 Guide to choosing an appropriate measure of spread

In general, if you have decided the mean is an appropriate measure of location, then the s.d. is usually the most appropriate measure of spread. If the median is appropriate as a measure of location, then the inter-quartile range is usually the most appropriate measure of spread.

Q. 4.10 What measures of spread 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; and (d) initial disability questionnaire score? With the help of Table 4.1 comment on the appropriateness or otherwise of their choice of measure. (You answered a similar question on the suitability of the chosen measures of location in Q. 3.10).

Solutions to coursebook questions

Unit 4: - Descriptive statistics
Part II Measures of spread

Q. 4.1 Range is (1.5 to 70.0) hours.

Q. 4.2 (a) Because the range is calculated using the smallest and largest values in the sample data. If either or both of these are outliers the range will be correspondingly distorted. (b) Largest value is 83, smallest is 31 years so range = (31 to 83) years.

Q. 4.3 (a) Q1 = 5.5mmol/l and Q3 = 7.0mmol/l, so inter-quartile range is (5.5 to 7.0) mmol/l. (b) Because the ends of the sample data are discarded, along with possible outliers.

Q. 4.4 From Q. 3.6 Q1= 40.5 years and Q3 = 63.75 years, thus the interquartile range is (40.5 to 63.75) years. So the middle 50% of patients cover an age range between these two values.

Q. 4.5 (a) range is about (6.5 to 15.2) minutes. (b) inter-quartile range is about (8 to 12.5) minutes.

Q. 4.6 (a) (25.3 to 68); (b) (5 to 62.5) mg.

Q. 4.7 (a) Yes, because to calculate sample s.d. we use the sample mean which is itself sensitive to outliers. (b) s.d. = 15.6 years. On average, each sample value is 15.6 years from the mean.

Q. 4.8 (a) Blockade group is older, mean = 72.8 years compared to control group mean of 70.8 years; (b) Ages more spread out in blockade group, s.d. = 13.2 years compared to 11.4 years in control group.

Q. 4.9 Using the area properties of the Normal distribution: $72.8 \pm 2\times 13.2 =$ 46.4 to 99.2 years.

Q. 4.10 (a) s.d. (b) range (minimum to maximum); (c) s.d. (d) s.d.

Appropriateness (using Table 4.1): (a) age (years) is continuous metric, so s.d. is appropriate; (b) duration of index episode (hours) is continuous metric so range is appropriate; (c) initial visual analogue pain score is ordinal so s.d is inappropriate (inter-quartile range would be more appropriate); (d) initial

disability score is ordinal so s.d. is inappropriate (inter-quartile range would be more appropriate).