Descriptive statistics

B. K. Dassanayake, MBBS (Peradeniya)

Department of Surgery, Faculty of Medicine, University of Peradeniya, Kandy.

Key words: Descriptive statistics; Inferential statistics.

This second section of the series will deal with which type of descriptive statistics to use on data that we collect

To restate from the last section, Descriptive Statistics only try to 'describe' or 'summarise' data. This is achieved by trying to represent all the values in the data collected by a single, representative measure. Any data we collect can be described or summarised using two such measures:

a. Measures of Central Tendency: A single value which will give a good idea of all the data that has been collected. eg. Mean, Median etc.

b. Measures of Spread: A value describing how close or far from the Measure of Central Tendency the actual data points are: eg: Range, Variance, Standard Deviation, Coefficient of Variation

Measure of central tendency

The commonest measures available are Mean, Median, Mode

a. Mean

This is the measure of choice for Ratio or Interval Data (see Section 1). There are many types of Means and thus we need to choose which type of Mean most appropriately represents our data:

Arithmetic Mean (Average): This is the commonest type of Mean we use, commonly called the Average. The Average describes most types of ratio/interval data very well.

$$\frac{120 + 130 + 140 + 150}{4} = 135 \text{ cm}$$

Eg. We have the heights of four students:

120 cm, 130 cm, 140 cm, 150 cm

The average height of the group is

Geometric Mean: This is commonly used to prevent a few observations with very large values from giving misleading picture about the entire set of data we have collected

Eg. We are observing a bacterial colony on a plate and measuring its width every day. As we know, the growth rate will gradually increase. Last weeks' values for each day are:

Correspondence: B. K. Dassanayake
Department of Surgery, Faculty of Medicine, University of Peradeniya, Kandy, Sri Lanka.
E-mail: thraless@gmail.com

2 mm, 4 mm, 8 mm, 16 mm, 32 mm, 64 mm, 128 mm

$$\frac{2 + 4 + 8 + 16 + 32 + 64 + 128}{7} = 36.3 \text{ mm}$$

What is the best single value to represent last weeks' diameter?

The Arithmetic Mean would be:

$$\sqrt[7]{2 \times 4 \times 8 \times 16 \times 32 \times 64 \times 128}$$

However, for most of the week the colony was smaller than 36.3 mm. It only became larger than that on the sixth

$$= \sqrt[7]{268435456} = 16 \text{ mm}$$

day. Therefore, the Arithmetic Mean is being influenced more by the larger values, and not giving a fair picture.

The Geometric mean solves this problem by calculating the mean as follows:

Thus, the geometric mean is calculated by multiplying all 'n' number of observations and then finding the nth root of that value. It is useful in studying growth parameters and prevents larger values from giving a wrong description of the data.

Harmonic Mean: This has limited applications in biological studies. It goes one step further than the Geometric Mean and favours the smaller values in the observations. For example, if genetic variations of a population are studied over time, the critical periods will be the occasions when the population size was smallest (genetic bottlenecks).

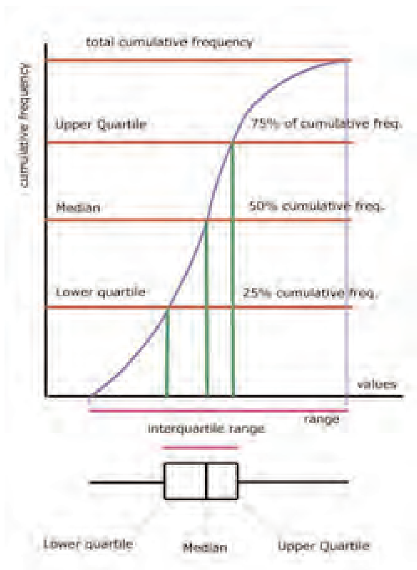
$$\frac{1}{H} = \frac{\frac{1}{20} + \frac{1}{15} + \frac{1}{30}}{3} = 0.05 \quad \therefore \quad H = 20$$

Eg. If a population size of mosquitoes was studied over three weeks, and the sizes were 20, 15 and 30 in each week.

The Harmonic mean (H) is calculated as follows:

b. Median

Conceptually, if all the values we have measured in a study are arranged in ascending order of magnitude, the median is the value which will separate the higher half of values from the lower half. The values may also be divided into four parts, called 'Quartiles'. In this case, the first and second Quartiles will contain values smaller than the Median, and the third and fourth Quartiles will contain values larg-



er than the Median. (Figure 1)

The Median is the statistic most appropriate to describe 'Interval data', or 'Ordinal Data' which has numerical values. Even with ratio data, the Median can be calculated.

Eg. Visual analog Pain Scores for five patients are: 3, 5, 1, 9 and 7.

Since Pain Scores are Ordinal Data, the appropriate statistic is NOT the mean, but median.

The first step is to arrange the values in ascending order, as 1, 3, 5, 7, 9. Since the value lying in the centre is five, the Median Pain Score of the patients is five.

Figure 1. Median, Quartiles and Interquartile Range

c. Mode

The mode is the value which occurs most often in a set of data (ie. the observation with the highest frequency). Nominal, Ordinal, Interval and Ratio data can all be described using a Mode.

Eg. Glasgow Coma Scale values for six patients are 3, 12, 4, 3, 15 and 8.

The Coma Scale represents an example for Ordinal Data. The Mode is therefore three.

Measure of Spread

The commonly used measures of spread are given below in increasing order of 'complexity'

a. Range
b. Interquartile Range
c. Standard deviation
d. Coefficient of variation

Range and Interquartile Range

Giving a 'range' is simply giving the difference between the smallest and largest values we have measured. It is appropriate for describing spread in ordinal, interval and ratio data.

The interquartile range excludes the largest and smallest values, and instead gives the range between the upper and lower Quartiles (Figure 1). The advantage over range is that it excludes extreme values. Therefore, a few very high or very low values which (may have been due to errors in the equipment used etc) will be prevented from giving a wrong idea about the population studied.

Standard Deviation

The standard deviation is appropriate for use in instances where a Mean has been calculated as a measure of central tendency. (Eg. in ratio and interval data).

The concept is simple: The mean gives a single value to represent all the data. Therefore, the spread is best shown by how far from the mean each actual value is. Let us approach this in a stepwise manner.

Eg. The weights of four patients with gastric cancer were 40 kg, 50kg, 60 kg, 70 kg.

Thus their average weight (Arithmetic mean) would be 55 kg.

1. How can we assess how far from 55 kg each weight is? (ie, the spread). What if we calculate the difference

$$(40 - 55) + (50 - 55) + (60 - 55) + (70 - 55) = (-15) + (-5) + 5 + 15 = 0$$

between the mean and each value and add all the differences together?

This is the easiest method that comes to mind. However, it has a problem

Difference of each value from mean: (40 - 55)kg, (50 - 55)kg, (60 - 55)kg, (70 - 55)kg

Therefore the 'total difference' will be

Thus, adding the differences of each value from the mean

$$(-15)^2 + (-5)^2 + (5)^2 + (15)^2 = 225 + 25 + 25 + 225 = 500$$

will always give zero. Therefore, we need a different method to assess spread.

$$(-15)^2 + (-5)^2 + (5)^2 + (15)^2 = 225 + 25 + 25 + 225 = 500$$

2. What if we square the differences to remove the negative sign?

$$\frac{500}{4} = 125$$

Adding the squares of the differences prevents the total from becoming zero

$$\text{Variance} = \frac{\text{Adding up the squares of the difference of each observation from the mean}}{\text{Number of observations}}$$

3. Now we can take an 'average' of the squares of the differences.

We had four observations. Therefore the average of the differences will be as follows:

The average of the square of the differences from the mean is called the Variance.

The main disadvantage of the Variance is that its units will be kg2 instead of kg. Thus the zz

4. What if we take the square root of the variance? Won't

$$\sqrt[2]{125} = 11.18 \text{ kg}$$

that give the same unit as the mean?

This is obviously the most appropriate way to assess 'spread' from the mean. It will give a good idea about how far each value is from the mean, and also be in the same units as the mean. The square root of the Variance is known as the Standard Deviation.

Thus the standard deviation of the patient's weights will be

Coefficient of Variation

The co-efficient of variation can be calculated for Ratio data. It is useful when comparing two groups of data with very different mean values.

$$\text{Standard Deviation of Infants' weights} = \sqrt[2]{1 \text{ (Variance)}} = 1 \text{ kg}$$

$$\text{Standard Deviation of Mothers' weights} = \sqrt[2]{25 \text{ (Variance)}} = 5 \text{ kg}$$

eg. Weight of three infants: 4 kg, 5 kg and 6 kg.

Weight of their mothers: 50 kg, 55 kg and 60 kg.

Which of these two groups (infants and mothers) have a wider 'spread' of weights?

First, let us calculate the Average weights of the two groups.

Now let us calculate the Standard Deviations

Therefore, it appears as if the mothers' weights have a larger 'spread'. However, we know that the difference in spread shown here is really is due to the mothers having larger weight values than the infants.

The Coefficient of Variation overcomes this problem by Dividing the Standard Deviation by the Mean of the group.

Therefore, the Infants actually have a wider 'spread' of weights than the mothers. Note that the co-efficient of variation will also have no Units. The discussion thus far has been summarised in Table 1.

Table 1. Types of descriptive statistics appropriate for each type of Data

Type of Data	Measure of Central Tendency	Measure of Spread
Nominal	Frequency, Mode	-
Ordinal	Median, Mode	Range, Interquartile Range
Interval	Mean, Median, Mode	Standard Deviation
Ratio	Mean, Median, Mode	Standard Deviation, Co-efficient of Variation

References (also to be used as further reading)

1. Thattil, RO, and Gunaratne LHP. Basic Statistics for Advanced Level Students. Godage & Bros. Colombo, Sri Lanka. 2000.
2. Thattil, RO. Handbook on Survey Design and Analysis. PGIA Publication. 1999.
3. Dodge Y. The Oxford Dictionary of Statistical Terms, Oxford University Press. 2003.
4. Stevens, SS. "On the Theory of Scales of Measurement." Science New Series.
   1946; 103(2684): 677-80
   PMID: 20984256