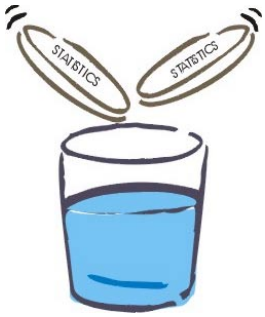
Statistics in Divided Doses



August 2001 Number 2

Describing a sample

Contents

Samples and populations

What is a "sample"?

A sample is a group of individuals and/or observations selected from a larger group for purposes of analysis. Opinion polls are an example. The views of a selected group are analysed in order to predict those of the population.

What is a "population"?

In statistical terms, the population is an entire group of individuals and/or observations about which information is sought. It does not necessarily refer to people; a population may be a collection of blood pressure or height measurements.

In clinical trials, what do these terms mean and how are they used?

Selected groups of patients (samples) are exposed to various treatments to assess their response. From the results, conclusions are drawn regarding the efficacy of such treatments in general use (i.e. within the population from which the samples are drawn).

Why is choice of sample important?

Not surprisingly, a sample representative of its population is more likely to yield results that predict the response of that population than an unrepresentative sample will.

For example, if a new antihypertensive drug were to be used in an elderly population with moderate hypertension, a trial with a sufficiently large sample of elderly patients with this condition would constitute a representative sample. A trial involving a sample of young patients with malignant hypertension would be unrepresentative as neither the age group nor indication are representative of the population in which the drug will be used.

This is why the inclusion and exclusion criteria of a trial should be specified to allow the reader to judge whether a suitable sample has been selected

Intersubject variation

What does intersubject variation mean?

It is unlikely that all patients in a sample will respond similarly to an intervention. Some will respond better than others, and some may fail to respond at all. This is known as *intersubject variation*.

How relevant is intersubject variation?

The degree of such variation will influence the reliability of measurements and subsequent analysis of results. For example, for antihypertensive drugs, if the range of blood pressure responses is wide and the sample size small, the *mean* blood pressure reading may be an unreliable index of response.

Although differences in individual responses to a drug may relate to recognised patient characteristics, most differences cannot be explained and are therefore attributed to *random variation*. This causes problems, as it is a factor that may significantly influence results. To overcome this, we use statistical analyses to give the reader an indication of the potential influence of random variation on results. This will be discussed further in later issues.

How is the spread of values, or variability, within a sample described?

An obvious measure is the *range*, the difference between the highest and lowest values. This was discussed in the first issue of this series of bulletins. However, the range is influenced by extreme values and will vary from sample to sample. One way to overcome this problem is to use the *standard deviation*.

NHS

Produced for healthcare professionals by North West Medicines Information Service,
The Pharmacy Practice Unit, 70 Pembroke Place, Liverpool, L69 3GF.
Editor: Frank Leach. Telephone: 0151 794 8117 E-mail: druginfo@liv.ac.uk

UKMi
UK Medicines Information

However, don't be misled, there is nothing standard about it!

Standard deviation

How do we calculate the standard deviation of a sample?

The formula for calculating the standard deviation is:

$$SD = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n-1}}$$

SD = standard deviation
x = the individual values within the sample
$\bar{x}$ = the mean value
n = the number of values
Σ = sum

For example, imagine that we have measured the heights of 20 men to the nearest centimetre. The respective heights are:
165, 170, 170, 172, 175, 167, 170, 167, 150, 155, 170, 172, 187, 180, 177, 167, 167, 172, 162, 180.

The mean value ($\bar{x}$ = 170) is derived from adding all the values together (to give 3395) and dividing by the number of values (n = 20).

If we subtract the mean value from each value in turn, we would get both negative and positive numbers i.e. -2, 0, +1 etc. and if we added these together we would get 0! But if each of the numbers is squared, we get a positive number.
Hence, Σ [$(165-170)^2$, $(170-170)^2$, $(170-170)^2$, $(172-170)^2$….etc], gives Σ [25, 0, 0, 4…etc] which equals 1341.

What is the variance of a sample?

The variance is a statistical term which describes the variability within a sample. Using the above example, to calculate the variance, we divide 1341 by [n-1] i.e. 19, to give 70.6 (the variance).

Why is n-1 used rather than n in the formula for standard deviation?

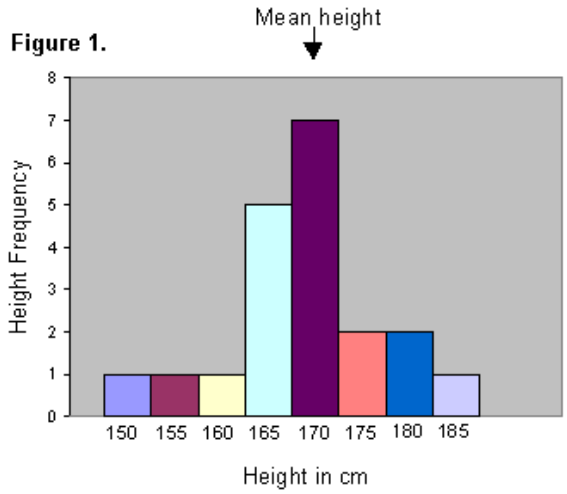
The use of n-1 involves a rather elusive statistical concept known as degrees of freedom. We shall return to this in a future bulletin. By using n-1 we obtain a closer estimate of the variability around the mean within the population from which the sample is taken. Therefore, as n (the sample) gets larger, the difference between n and n-1 is reduced.

Use of graphs to illustrate data

How can the spread of values in a sample be illustrated graphically ?

One method is to construct a histogram as in Figure 1 which is based on our fictitious height sample.

Figure 1.



The histogram plots the heights (in 5cm intervals) against the number of men who had a particular height i.e. the height frequency.

What does the histogram describe?

A histogram describes *frequency distribution*. This shows the relationship between individual values in a sample and the frequency with which those values occur. Thus, figure 1 illustrates that 7 men had a height between 170 and 174cm while only one had a height in the range 155 to 159cm.

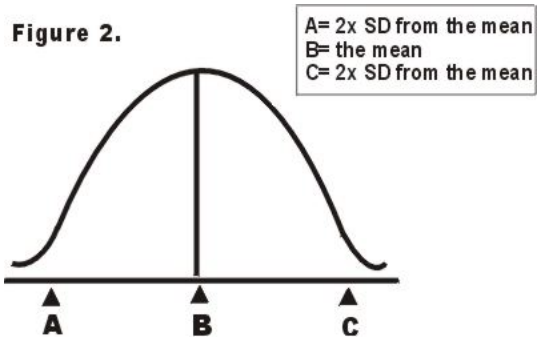
What if the sample size is larger?

As the sample size increases, the distribution of the samples will approximate more closely to that of the parent population.

Does the shape of the frequency distribution matter?

In figure 1, the values are roughly symmetrical around the mean value of 170cm. If the sample size is larger the histogram will become even more symmetrical. This type of frequency distribution within a population is known as Normal (or Gaussian) distribution. It follows a curve which is symmetrical around the mean with a characteristic bell-shape. This is illustrated in figure 2.

Figure 2.



What is meant by the term Normal distribution'?

The Normal distribution is a key concept and many statistical tests are based upon it. It is described by two

parameters – the mean and the standard deviation. Its curves are always symmetrical and bell-shaped - the extent to which the bell is flattened or compressed depends on the standard deviation of the population.

If measurements are 'Normally distributed' around the mean, you can be confident that use of the standard deviation, as an estimate of distribution, is likely to be valid. However, be careful. Just because the distribution of the sample is not symmetrical, you cannot assume the parent population does not exhibit Normal distribution. It may be that the sample size is too small.

What does the standard deviation tell us about a sample from a population with a Normal distribution?

In such a sample, or one in which there is reasonably symmetrical distribution around the mean, one standard deviation value either side of the mean will represent about 68% of observations within the sample. In addition, 2 SDs either side of the mean will represent approximately 95% of measurements within the sample (represented by points A and C in figure 2). Going back to our fictitious height example, we can assume that approximately 95% of height measurements within our sample lie between the limits of 170 - (2 x 8.4) and 170 + (2 x 8.4) cm (i.e. approximately 153 to 187cm). The mean ± 3 SDs includes 99.7% of the observations.

It is important to remember that in a Normal distribution, the mean and median values will be identical.

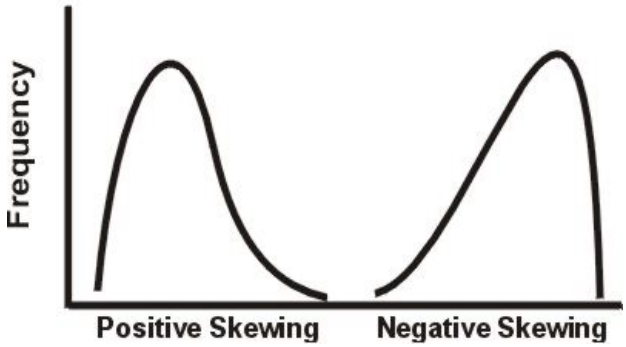
What does a negative standard deviation value imply?

When calculating the range of values representing 95% of the sample you would get negative and positive values if the mean and standard deviation values are of a similar size. This would suggest that the data are skewed (see below). In this case, use of standard deviation to calculate sample range would be invalid.

Do all frequency distributions exhibit Normal distribution?

No, there are a number of different frequency distributions (e.g. Poisson and Binomial). *Skewed distributions* are those in which the distribution of values is weighted towards one extreme. They can be described as negative or positive depending on the shape of the curve (Figure 3). Positive is more common than negative skewing. Different statistical tests are used for analysis of skewed distribution, compared to those used for Normally distributed data. These will be covered in future issues of *Statistics in Small Doses.*

Figure 3.



Centiles and ranges

What does 'centile' mean in statistical terms?

A *centile* is an alternative way to describe a frequency distribution. For example, if a child is on the 60th centile for height it can be assumed 60% of children, at the same age, have the same or a lower height than the individual child does.

What do the terms 'interquartile range' and 'central range' mean?

The *interquartile range* describes the range of values between the 25th and 75th centiles. It will be quite variable from sample to sample but is useful to describe measurements that have an asymmetrical distribution. The *central range* is that within which 90% of values lie.