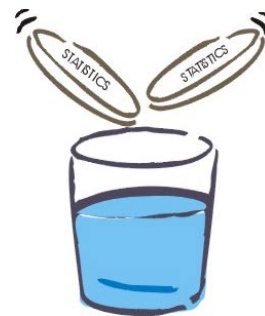


Statistics in Divided Doses



July 2001 Number 1

Making observations and taking measurements

Contents

- Type and definition of data
- Precision in data measurement
- Summarising data – use and abuse of mean, median and percentage

Type and definition of data

What does the term data mean?

In clinical trials, data are observations or measurements e.g. blood pressures, number of tablets taken daily, white cell counts or scores of disease severity.

The type of data measured influences the choice of statistical test used for analysis.

Data can be broadly classified into *categorical* or *numerical*.

What are categorical data?

These are data that can be readily allocated to categories or groups.

Categories may include:

- simple "either/or" groups, also known as *dichotomous* or *binary* data
e.g. male/female
- multiple groups also known as *nominal* data
e.g. blood groups

If categorical data are arranged in ascending order of severity or magnitude they are described as *ordinal* e.g. staging of disease severity or levels of pain.

What are numerical data?

Data that have a particular numerical value e.g. ratios, percentages. They can be further classified into *discrete* or *continuous* data.

What are discrete numerical data?

These are values that describe numbers of events e.g. number of pregnancies in an individual, number of migraine attacks within a given period.

What are continuous numerical data?

This is a potentially confusing term. It is applied to data derived by a process of numerical measurement. For example, the heights of a sample of patients, which will range from the shortest to the tallest in a *continuous* manner. They can be recorded to any desired level of precision.

Precision in data measurement

What is meant by the precision of data?

We normally think of precision as the level of numerical accuracy applied in making and summarising observations. In measuring height, for example, it is usual to report findings to the nearest centimetre since this level of precision is sufficient for most clinical purposes. In statistical terms precision implies reproducibility of the measurement method i.e. will it give consistent results when repeated?

Why is precision in measurement important?

Although the classification of data is sometimes arbitrary, knowledge of the methods and precision of measurement used in studies is essential for the use of appropriate statistical analyses. It is pointless and potentially misleading to apply sophisticated statistical analysis to data that have not been generated in a systematic and reproducible manner. For example, measurement of blood pressure is subject to a number of variables that can profoundly influence outcome. In the context of a clinical trial, variables such as observer variation, patient characteristics etc should be allowed for.

What is meant by spurious accuracy?

This implies a level of accuracy in reporting results that is not justified by the assessment methods. For example, in assessing pain response, a crude scale of assessment such as "mild, moderate or severe" may be used. If we score these responses numerically as 1, 2 or 3, it is possible to calculate a total score and an average to the first decimal place. However, to report the average pain response as, say, 2.7 units

would be misleading and imply a level of accuracy unjustified by the method of pain assessment.

Summarising data – use and abuse of mean, median and percentage

Why do we summarise study data?

Data are summarised to provide a reliable and typical measure of response that is meaningful to the reader. For example, it is confusing and unhelpful to detail every blood pressure measurement in a study of a large number of subjects. Instead, *summary* values that characterise the results allow the reader to draw conclusions about the severity of hypertension in the sample of patients.

How can we numerically summarise a set of observations?

One approach is to calculate the *mean* (or average) value and present this with the *range* of values i.e. the lowest and highest value.

How is the mean calculated?

The mean of a set of observations is calculated by adding all the values together and dividing by the number of values.

When is the mean value misleading?

When a set of observations contain values that are atypical (i.e. much larger or smaller than the majority). These atypical values are known as outliers. A mean value will be distorted by outliers and not be representative of most values in the sample. Also, the mean value on its own gives no information on how narrowly or widely the individual observations are scattered.

Consider these two sets of numbers:

A 8, 3, 7, 6, 4, 5, 7, 5, 6, 8, 6, 7
Total = 72, Mean = 6

B 1, 1, 1, 2, 2, 1, 3, 4, 2, 5, 44, 5.
Total = 72, Mean = 6

The means are identical. However, the scatter of values and the reliability of the mean as a typical value vary between the two samples.

How does the range define a sample?

Very poorly. It includes the extreme values at either end of a sample, but gives no information about the extent of scatter within the range.

Is there an alternative to the mean?

In some circumstances, the *median* can provide a more useful measure.

How is the median calculated?

The values are set out in ascending order. The middle value of the resulting series is the median. When there is an odd number of values, the middle value is obvious. If the number of values is even, the middle two are added together and divided by two.

Setting out the values used above in ascending order gives:

A 3, 4, 5, 5, 6, 6, 6, 7, 7, 7, 8, 8
Median = 6 (i.e. same as mean)

B 1, 1, 1, 1, 2, 2, 2, 3, 4, 5, 5, 44
Median = 2.

Which is used more frequently the mean or the median?

The mean is used more frequently but there are situations in which the median is more useful.

When should the median be used in preference to the mean?

To remove the distorting influence of outliers. A study of haemoglobin levels might find that although the majority of patients in the sample were within normal limits, there were several with extremely high or low levels. These outliers may distort the clinical and/or laboratory implications for the group and give a misleading mean value. As an everyday example, using the median scores of judges in international competitions is an effective way of removing bias.

Another use for the median is with censored data. When we can measure some observations exactly but only know that others are greater than a certain value, we term this censored data. For example, in measuring survival times, we may decide not to follow-up patients for longer than ten years. However, later entrants into the study will be followed up for less than 10 years and the likelihood they are still alive is greater than it is for early trial entrants. In this situation, providing that at least half the survival times had been recorded, the median (but not the mean) can be calculated.

Why can percentages be misleading?

Percentages can be used to disguise small numbers. A report of response to drug treatment in 7 of 12 patients can be reported as a 58.3% level of response. This looks impressive but gives no clue as to the small number of patients. Percentages are frequently reported in advertising literature.

What does the term 'mode' mean?

The *mode* is the value occurring most frequently within a sample but has little application in statistical analysis. In example B above, the mode value is 1.