Types of data

Derek Richards

Centre for Evidence-based Dentistry, Oxford, UK

In most, if not all, studies we collect data to obtain information about an area of research in which we have an interest. For example, we might want to know the level of dental caries in our area. In order to discover this we might need to observe a number of different variables, which could include, age, sex, number of teeth, cavities, fillings, extraction, pain, sepsis and quality of life. This information or data is normally obtained from a sample of the population which can then be summarised, analysed and conclusions drawn. This collection, summarising and analysis of data are what statistics and statistical technique are all about.

Evidence-Based Dentistry (2007) **8**, 22-23. doi:10.1038/sj.ebd.6400501

For part of the process of deciding what statistical techniques are most appropriate for a given task, we need to know what type of data or variable we are dealing with. There are two main types of data, categorical or numerical (Figure 1), but within these broad groups are various different types of data.

Categorical (qualitative) data

When an individual can only be allocated to one of a number of mutually exclusive

categories the data are categorical, eg, male/female, married/ single, smoker/ nonsmoker. Allocation to one of two categories is the simplest of situations. Often, however, there is more than one category available: married/ single/ divorced/ separated/ widowed, or blood group A/B/AB/O. These types of data where categories are unordered are termed *nominal data*.

If there is an obvious ordering of the categories, the data are termed *ordinal data*.

In pain studies, for instance, people may classify pain as minimal, moderate, severe or unbearable. Likewise, in the Index of Orthodontic Treatment Need (IOTN), out of IOTN categories 1/2/3/4/5, 1 is considered to have the lowest need for treatment and 5 the highest. Although numbers may sometimes be used to indicate the categories, such as in the IOTN example, these numbers merely indicate the order or ranking of results. The numbers are not measurable on a scale. So in the case of the IOTN, we know that a patient with a score of 4 has a higher treatment need than someone with a score of 3, but we cannot quantify how much higher is that need.

Numerical (quantitative) data

Where a variable takes a measurable numerical value it is said to be numerical. There are two types of these data, discrete and continuous.

Discrete data. These occur where the observations can only take certain numerical values, such as the number of visits to the dentist, or the number of episodes of mouth ulcers.

Continuous data. These data have no limitation in the values the variable can take, eg, height, weight or age.

The statistical methods used to analyse data are often dependent on whether data are categorical or numerical. On the whole, the distinction between the two is clear but in some circumstances it is less so. In analysis, continuous data may be reduced to several categories, eg, age or blood pres-

Figure 1. Different types of variable

Figure 1. Visual analogue scale

sure, but do not be tempted to record numerical data as categorical at the outset (“age range, 40–49 years” instead of actual age “22/07/1965”) because, although it is easy to convert a date of birth to a category, the raw data cannot be retrieved later if only the categories are recorded.

Presenting data

Categorical and numerical data are common in dental research and they may be analysed, presented or summarised by a variety of methods including:

Proportions (eg, percentages). These may arise when considering improvements in patients following a treatment, eg, the prevented fraction following a caries prevention programme.

Rates. Disease rates are calculated where the number of disease events is divided by the number of people at risk over the

time period under consideration, eg, oral cancer rates.

Scores/indices. When it is not appropriate to take direct measurements it is often possible to grade individuals in some way, eg, quality of life scores, or indices such as the oral hygiene index¹ or simplified oral hygiene index² can be used to create plaque scores. There are issues here to be noted: for most scoring systems a degree of subjectivity is present, and numerical coding can be misinterpreted to imply (often inappropriately) that differences between scores are equally important and that the scores are equally important.

Visual analogue scales. Where patients are asked to assess unmeasurable variables such as pain or hunger the visual analogue scale (VAS) or linear analogue is an improvement upon ordered categories. A patient is shown a straight line (often 10 cm long) anchored at each end with extreme states (eg, no pain

and unbearable pain). They are then asked to mark on the line a point which represents their current state (Figure 2). This is clearly subjective and is of most use when looking at change within individuals.

Censored data

If we cannot measure an observation precisely but we know that it is beyond some time limit we refer to it as censored. For example, if we are measuring the level of virus in a sample the level may be below the level that the test or machine can detect even though the value should be greater than zero. These values are said to be censored at the limit of detectability. Another more common situation in dentistry is where we are following people in a trial of implants or post crowns and we are looking at survival. Here the trial will end at a fixed point after recruitment so that implants or post crowns surviving at the end of the trial have censored survival times, with the censoring varying with the period that the participant has spent in the trial.

1. Greene JC, Vermillion JR. Oral hygiene index: a method for classifying oral hygiene status. *J Am Dent Assoc* 1960; 61:172–177.
2. Greene JC, Vermillion JR. The simplified oral hygiene index. *J Am Dent Assoc* 1964; 68:7–13.