MCQ 1 2 16

SQ 1

1 Types of data

Data and statistics

The purpose of most studies is to collect **data** to obtain information about a particular area of research. Our data comprise **observations** on one or more variables; any quantity that varies is termed a **variable**. For example, we may collect basic clinical and demographic information on patients with a particular illness. The variables of interest may include the sex, age and height of the patients.

Our data are usually obtained from a **sample** of individuals which represents the **population** of interest. Our aim is to condense these data in a meaningful way and extract useful information from them. **Statistics** encompasses the methods of collecting, summarizing, analysing and drawing conclusions from the data: we use statistical techniques to achieve our aim.

Data may take many different forms. We need to know what form every variable takes before we can make a decision regarding the most appropriate statistical methods to use. Each variable and the resulting data will be one of two types: **categorical** or **numerical** (Fig. 1.1).

Categorical (qualitative) data

These occur when each individual can only belong to one of a number of distinct categories of the variable.

- **Nominal data** – the categories are not ordered but simply have names. Examples include blood group (A, B, AB and O) and marital status (married/widowed/single, etc.). In this case, there is no reason to suspect that being married is any better (or worse) than being single!
- **Ordinal data** – the categories are ordered in some way. Examples include disease staging systems (advanced, moderate, mild, none) and degree of pain (severe, moderate, mild, none).

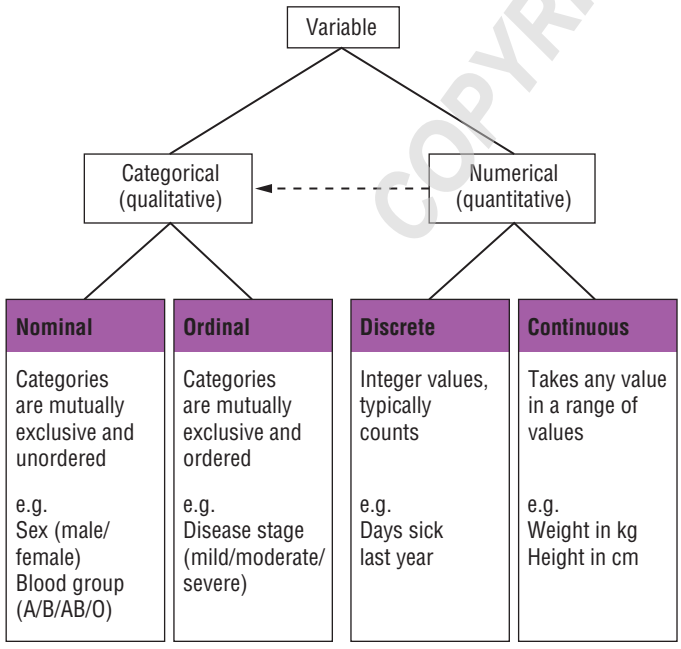

Figure 1.1 Diagram showing the different types of variable.

A categorical variable is **binary** or **dichotomous** when there are only two possible categories. Examples include 'Yes/No', 'Dead/Alive' or 'Patient has disease/Patient does not have disease'.

Numerical (quantitative) data

These occur when the variable takes some numerical value. We can subdivide numerical data into two types.

- **Discrete data** – occur when the variable can only take certain whole numerical values. These are often counts of numbers of events, such as the number of visits to a GP in a particular year or the number of episodes of illness in an individual over the last five years.
- **Continuous data** – occur when there is no limitation on the values that the variable can take, e.g. weight or height, other than that which restricts us when we make the measurement.

Distinguishing between data types

We often use very different statistical methods depending on whether the data are categorical or numerical. Although the distinction between categorical and numerical data is usually clear, in some situations it may become blurred. For example, when we have a variable with a large number of ordered categories (e.g. a pain scale with seven categories), it may be difficult to distinguish it from a discrete numerical variable. The distinction between discrete and continuous numerical data may be even less clear, although in general this will have little impact on the results of most analyses. Age is an example of a variable that is often treated as discrete even though it is truly continuous. We usually refer to 'age at last birthday' rather than 'age', and therefore, a woman who reports being 30 may have just had her 30th birthday, or may be just about to have her 31st birthday.

Do not be tempted to record numerical data as categorical at the outset (e.g. by recording only the range within which each patient's age falls rather than his/her actual age) as important information is often lost. It is simple to convert numerical data to categorical data once they have been collected.

Derived data

We may encounter a number of other types of data in the medical field. These include:

- **Percentages** – These may arise when considering improvements in patients following treatment, e.g. a patient's lung function (forced expiratory volume in 1 second, FEV1) may increase by 24% following treatment with a new drug. In this case, it is the level of improvement, rather than the absolute value, which is of interest.
- **Ratios** or **quotients** – Occasionally you may encounter the ratio or quotient of two variables. For example, body mass index (BMI), calculated as an individual's weight (kg) divided by her/his height squared (m^2), is often used to assess whether s/he is over- or underweight.
- **Rates** – Disease rates, in which the number of disease events occurring among individuals in a study is divided by the total number of years of follow-up of all individuals in that study (Chapter 31), are common in epidemiological studies (Chapter 12).
- **Scores** – We sometimes use an arbitrary value, such as a score, when we cannot measure a quantity. For example, a series of responses to questions on quality of life may be summed to give some overall quality of life score on each individual.

All these variables can be treated as numerical variables for most analyses. Where the variable is derived using more than one value (e.g. the numerator and denominator of a percentage), it is important to record all of the values used. For example, a 10% improvement in a marker following treatment may have different clinical relevance depending on the level of the marker before treatment.

Censored data

We may come across **censored** data in situations illustrated by the following examples.

- If we measure laboratory values using a tool that can only detect levels above a certain cut-off value, then any values below this cut-off will not be detected, i.e. they are censored. For example, when measuring virus levels, those below the limit of detectability will often be reported as 'undetectable' or 'unquantifiable' even though there may be some virus in the sample. In this situation, if the lower cut-off of a tool is x, say, the results may be reported as '$<x$'. Similarly, some tools may only be able to reliably quantify levels below a certain cut-off value, say y; any measurements above that value will also be censored and the test result may be reported as '$>y$'.
- We may encounter censored data when following patients in a trial in which, for example, some patients withdraw from the trial before the trial has ended. This type of data is discussed in more detail in Chapter 44.