Statistics in medicine

Ian Kestin

Abstract

This article covers the basic principles of statistics in medicine. Topics covered include types of data, descriptive statistics (mean, median, mode, percentiles), the normal distribution, confidence intervals and the standard error of the mean, hypothesis testing and the choice of statistical tests, type I and II errors, contingency tables, correlation and regression, and meta-analysis.

Keywords Clinical trials; correlation; hypothesis tests; normal distribution; regression; statistics

Introduction

All clinicians should understand the correct use of research data, and that statistics are the tools used to describe and analyse numbers. The complete data set from a study may comprise many thousands of observations, and it is not practical to give the full results in a published paper. Descriptive statistics are used to summarize this numerical information. We also use statistics to infer properties about a wider population of subjects beyond those actually studied, and this is called inferential statistics. Uncertainty, probability and error are crucial concepts for understanding the limitations of statistics.

Types of data

The types of data obtained in a research project determine the methods used to describe and analyse the data. There are three main types.

- **Categorical (nominal) data:** each of the subjects in the study is allocated to one of two or more mutually exclusive categories, for example sex (male/female), blood group (A, B, AB, O) or social class. The categories have no ranking or numerical relationship to each other.
- **Ordinal data (ordered categories or ranked data):** each of the subjects in the study is allocated to one of several mutually exclusive categories, and these categories have an intrinsic ranking or ordering. Examples would be grades of oedema (mild/moderate/severe), or American Association of Anaesthetists (ASA) scores (1, 2, 3, 4 or 5). The categories may be numbered, but this numbering only defines the ordering of the categories, and does not 'scale' the relative magnitude of one category to another. For example, head-injured patients are allocated using the Glasgow coma score (GCS) to one of 13 possible categories denoted by a whole number between 3 and 15. A patient with a GCS of 4 is worse than a patient with a GCS of 8, but is not 'twice as bad', whereas a patient weighing 80 kg is exactly twice as heavy as one weighing 40 kg. Misusing ordinal data by treating the numbers as if real measurements had been made is a common mistake.
- **Numerical data:** this type of data describes actual numerical properties of the subjects. The measurements can be either discrete or continuously variable. Discrete numerical data can only take certain values, usually integers (e.g. number of children, hospital deaths per year); continuously variable data can theoretically take any numerical value, but usually occur within a certain range (e.g. heart rate, weight).

Descriptive statistics

Descriptive statistics are required to summarize large data sets. Categorical data are easily described by histograms or pie charts; a visual illustration of the data clearly shows the frequency of the categories (Figure 1).

Two essential properties describe ordinal or numerical data:

- the central location — where the bulk of the observations lie
- the variability — how closely the observations are clustered about the central location.

The central location of a series of observations is usually described using the mean, median or mode (Table 1).

Misuse of the mean is a common error, which properly should only be used with continuously variable numerical data. For ordinal data the median or mode must be used (e.g. it is quite wrong to quote a mean GCS of 7.5).

The variability can be described by the range, percentiles or the standard deviation. The range gives the maximum and minimum values of the observations and is useful if there is some particular interest in the maximum or minimum response (e.g. the lowest respiratory rate recorded would be of clinical importance in patients given opiates). However, the range can give a misleading impression of the variability if there are single extreme results in the data.

A percentile is that observation which is greater than the appropriate percentage of all the observations in the data set, so the 10th percentile is the observation that is greater than 10% of all the observations; the median is the 50th percentile; and the 90th percentile is the observation that is greater than 90% of the observations. Commonly used percentiles are the interquartile range (the 25th to 75th percentile) and the 2.5th to 97.5th percentile (containing 95% of the observations). Percentiles can be used for any type of data, but the standard deviation is only applicable to data that are continuously variable and normally distributed (see later). A common graphical way of summarizing information is the 'box and whisker' plot (Figure 2).

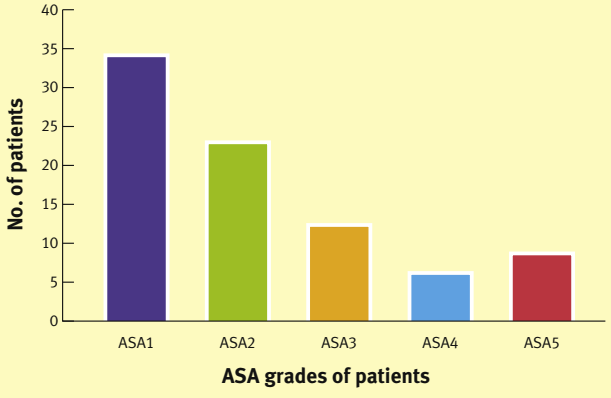
Frequency distribution curves

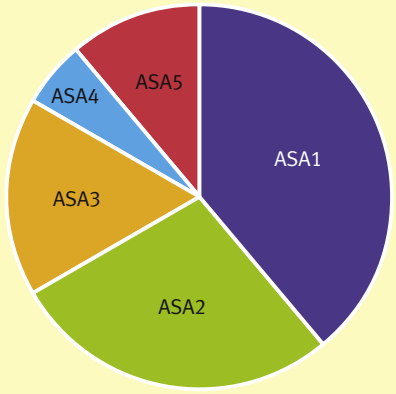
A graph showing the probability of obtaining any particular observation is called a frequency distribution (Figure 3).

The normal distribution is a specific frequency distribution pattern that is common in biological data for which many statistical tests have been designed (e.g. *t*-test, analysis of variance).

Ian Kestin *FRCA is Consultant Anaesthetist at the Western Infirmary, Glasgow, UK. He trained in anaesthesia in Southampton, Bristol, UK, and the USA. His interests include education in anaesthesia and mathematical applications in medicine.* *Conflicts of interest: none declared.*

A pie chart and histogram, two ways of illustrating the frequency distribution of categorical or ordinal data





ASA, American Society of Anesthesiologists.

Figure 1

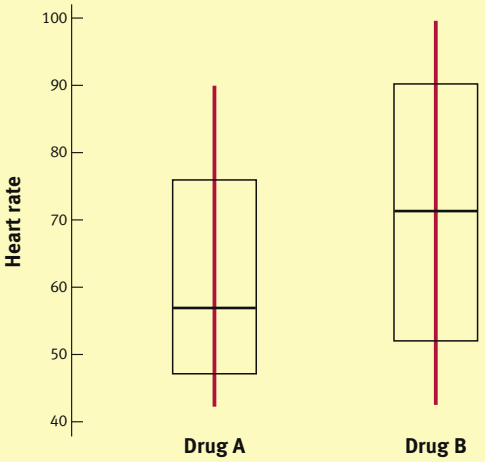
The central location can be described by the mean (which is the same as the mode and median), and the variability is described by the standard deviation. Multiples of the standard deviation about the mean always contain the same proportion of the observations (Figure 4).

Common measures of central location

Measure of central location	Type of data	Definition
Mean	Continuously variable	Sum of all observations/ number of observations
Median	Ordinal and numerical	The observation with half the observations above and half below, i.e. 50th percentile
Mode	Ordinal and numerical	The most commonly occurring observation

Table 1

Box and whisker plot of heart rates after two different drugs



The horizontal line shows the median, the 'box' shows percentiles, commonly the 2.5th to 97.5th, and the vertical line shows the range of the sample data.

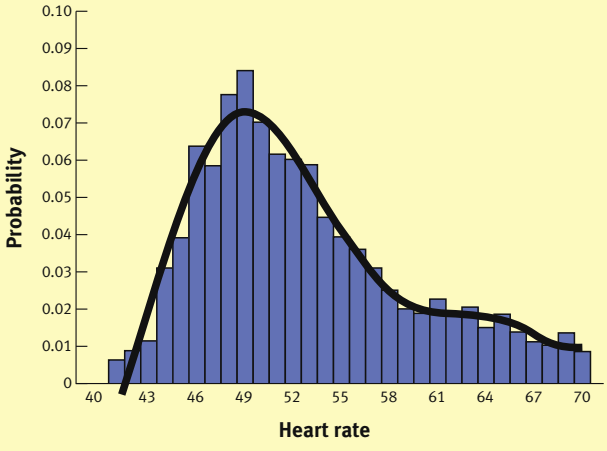
Figure 2

Not all symmetrical frequency distributions are normal (Figure 5).

Skewed distributions are a common pattern in biological data, when the frequency distribution curve is not symmetrical (Figure 6).

The frequency of hospital stay after an operation is commonly skewed (see Figure 6); most patients have similar lengths of stay, but some have complications and stay much longer. This is an example of positively skewed data; negatively skewed data is the reverse pattern and is less common. The mean, median and mode

Frequency distribution of heart rates



The probability of any given heart rate is shown by the histogram. The histogram can be replaced by a continuous curve as the intervals on the *x*-axis become smaller.

Figure 3