Statistics I: data and correlations

Anthony McCluskey BSc, MB, ChB, FRCA
Abdul Ghaaliq Lalkhen MB, ChB, FRCA

Statistics is the mathematical science dealing with the presentation, analysis, and interpretation of numerical information (data). In descriptive statistics, raw data are simplified as tables, graphs, and summary statistics such as mean and standard deviation. Inferential statistics is used to analyse and draw conclusions about a population of interest using data taken from a sample of the population, according to the laws of probability theory. Investigators are usually interested in populations rather than samples. For example, it is more useful to know the average arterial pressure in the UK adult population as a whole rather than in a much smaller sample in whom it is actually measured during the course of an investigation. As the study of entire populations is generally not possible, statistical methods are used to extrapolate from measured (known) sample characteristics to unmeasured (unknown) population characteristics.

Three key terms are often used in statistics: parameter, variable, and sample statistic. A parameter is a measurable characteristic or attribute of a population or model (e.g. the average height or weight of the UK population). It has a fixed and usually unknown value. A variable is a measurable attribute of a sample (e.g. height or weight in a sample data set); variables vary from individual to individual. Variables may be either quantitative (interval or ordinal data) or qualitative (nominal data). A dependent variable is simply measured, whereas an independent variable is manipulated or controlled experimentally. For example, in an investigation of a new neuromuscular blocking agent, the dose of neuromuscular blocking agent received by patients is an independent variable, whereas the length of time before return of the first twitch in a train-of-four count is a dependent variable. A sample statistic is a mathematical quantity calculated from sample data variables, and is used as an approximation to a corresponding population parameter.

Types of data

It is important to appreciate the different types of data generated during the course of an investigation. Interval and categorical data are the two main types of data. Interval data are continuous and quantitative (e.g. height, weight). A subtype of interval data is integer data; such data are continuous, but may be assigned only integer values (e.g. age of patients). Non-interval data are placed into different groups or categories, hence the term categorical data. Categorical data are discrete and qualitative (or pseudo-quantitative) and comprise two subtypes: nominal and ordinal. Examples of nominal categorical data are gender, hair colour, and patients' preferences for postoperative analgesia. There is no sense of a mathematical relationship or hierarchy between the categories.

When categorical data are stratified into groups with an implied rank order, the data are termed ordinal. A common example of this is the use of verbal rating pain scores (e.g. no pain, 0; mild pain, 1; moderate pain, 2; severe pain, 3). There is a definite hierarchy of categories that have pseudo-quantitative characteristics. A common source of confusion with ordinal categorical data is to treat them as being continuous and quantitative. However, the values in an ordinal scale have no real mathematical relationship to each other. It is meaningless to say that a patient with a pain score of 2 is in twice as much pain as another with a score of 1. It is also statistically invalid to state that the average (mean) pain score is 2.4; an ordinal pain score of 2.4 has no definition.

Presenting data

In general, data may be presented in one of three formats: numerical, tabular, or graphical. Numerical and tabular representations of data are precise and provide the reader with all of

Key points

Data may be classified as either interval (continuous, quantitative) or categorical (discrete, qualitative).

Common methods of summarizing and presenting data are tables, pie charts, bar charts, histograms, frequency and cumulative frequency curves, dot plots, and $x-y$ scatterplots.

A high statistical correlation between two variables does not imply cause and effect.

A Bland–Altman plot is used to compare two different methods of measuring the same variable rather than determination of a correlation coefficient.

Anthony McCluskey BSc, MB, ChB, FRCA
Department of Anaesthesia
Stockport NHS Foundation Trust
Stepping Hill Hospital
Stockport SK2 7JE
UK
Tel: +44 0161 419 5869
Fax: +44 0161 419 5045
E-mail: a.mccluskey4@ntlworld.com
(for correspondence)

Abdul Ghaaliq Lalkhen MB, ChB, FRCA
Department of Anaesthesia
Royal Lancaster Infirmary
Ashton Road
Lancaster LA1 4RP
UK

doi:10.1093/bjaceaccp/mkm012 Advance Access publication May 2, 2007
Continuing Education in Anaesthesia, Critical Care & Pain | Volume 7 Number 3 2007

the data, so that they are able to perform any statistical calculations for themselves. Graphical representations of data have more visual impact and are useful in identifying patterns in the data. The different methods compliment each other.

Categorical data

The simplest way of summarizing categorical data is to present them as a table. For example, a coin is tossed 20 times, and the following data are obtained: HHTHTHHTTHTTHHHHTHTT. A summary table of these data describing frequencies, percentages or both must contain the size of the data set (Table 1).

A contingency table is a special type of frequency table that may be used to summarize and analyse categorical data by cross-classifying two or more independent sample variables. The rows represent sample groups and the columns represent outcomes. In a 2×2 contingency table, the upper row usually represents the active group, the lower row the control group, the left column represents the 'positive' outcome (e.g. those with the disease) and the right column the 'negative' outcome (e.g. those without the disease). A 2×2 contingency table may be drawn up from a study of 100 critically ill patients with acute respiratory distress syndrome. Two ventilatory strategies, A (a newer, experimental approach) and B (routine management), are used on patients in the study, chosen by random allocation. ICU mortality is the primary end point (Table 2). Each cell in the table must contain the actual count of the data, which must not be expressed as a proportion or percentage. The cells must also be both exhaustive and mutually exclusive.

A table may be considered a gold standard for summarizing and presenting data. All of the data from a study may be included in a table as the precise numerical values obtained, and readers are able to analyse the data for themselves. However, as the number of groups increases, it can become increasingly difficult to appreciate the data and the relationship between the groups. Therefore, a pictorial or graphical representation of the data is often used (e.g. pie charts, bar charts). Although pie charts are useful graphical representations of categorical data, bar charts are more commonly used as most people find them better for comparing the frequency distribution of the different categories. Also they are a very good way of showing how the categorical frequency distributions of two or more different groups compare.

In a line graph, the bars are replaced by points or symbols joined together by a straight line. The individual points may be omitted entirely, showing only the line. Because the nature of the line suggests a continuum along the x-axis as the categories are run through one by one, a line graph is most useful when the categories depicted on the x-axis exist as a continuum.

Table 1 Distribution of frequency data

Category	Frequency (%)
Heads	11 (55)
Tails	9 (45)

Table 2 ICU mortality between a newer, experimental approach (strategy A) and routine management (strategy B)

Ventilatory strategies	Non-survivors	Survivors
A	18	32
B	12	38

Interval data

The main methods used to present interval data are tables, histograms, cumulative frequency curves, dot plots, and $x-y$ scatterplots.

Tables

The principles underlying the construction of tables for interval data are similar to those already described for categorical data. When preparing a table, interval data must first be divided arbitrarily into ranges known as class intervals. The number of class intervals chosen should neither be too small nor be too large, as meaningful comparisons of the spread of the data are then difficult. The class intervals should be of equal size and arranged in a rank order. However, when dealing with outliers, it is permissible for the first and last class interval to be enlarged to avoid multiple occurrences of empty, or nearly empty, class intervals. Table 3 shows frequency data obtained for the weights of a sample of 100 adult males.

Histograms

Histograms are graphical displays of tabulated frequencies, and are often used to represent interval data. Unlike a bar chart, the x-axis of a histogram is not dimensionless as it represents interval data. Although the class intervals are ideally the same size, it is permissible to have differing sizes, because it is the area of the rectangles that is proportional to the frequency count of each class interval. The data for the weights of the sample of 100 males are shown in Figure 1. A histogram may be examined to show how the data are distributed from its overall shape, whether or not it is symmetric, skewed, or has multiple peaks (multimodal), and where the central tendency of the data appears to be located.

Frequency curves

Frequency curves present data in a similar way to histograms. The rectangles are replaced by points positioned at the mid-point of each class interval on the x-axis and the appropriate height

Table 3 Frequency data obtained for the weights of a sample of 100 adult males

	Weight (kg)						
	<50	50–60	60–70	70–80	80–90	90–100	>100
Frequency (f)	1	5	20	30	25	16	3

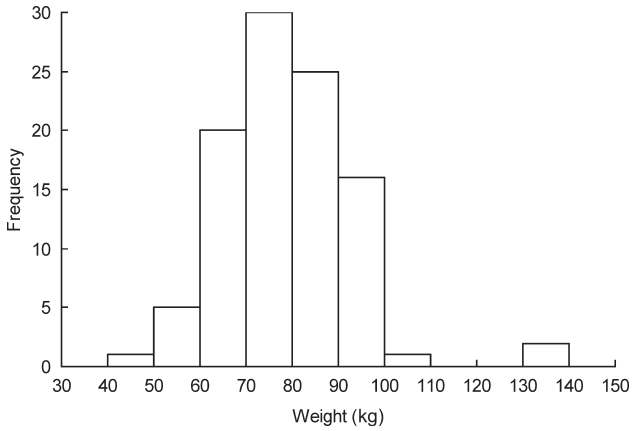

Fig. 1 Histogram showing the distribution of body weight in 100 males.

[frequency or relative frequency (%)] on the *y*-axis. The points are joined by straight lines. Frequency curves are useful for comparing two or more distributions.

Cumulative frequency curves

In a cumulative frequency curve, a running total of the frequency of occurrence of interval data on the *x*-axis is plotted on the *y*-axis. Datum points are placed at the upper end of each class interval. A cumulative relative frequency curve simply converts the scale on the *y*-axis to either 1 or 100%. A cumulative frequency curve can be used to estimate any percentile (discussed later) from the distribution. Figure 2 shows the data for the weights of the sample of 100 males; the 50th percentile (median) is indicated on the graph.

Dot plots

In a histogram, continuous interval data on the *x*-axis is subdivided into different class intervals. In a dot plot, the raw data is presented without modification; it is therefore most useful with relatively small data sets.

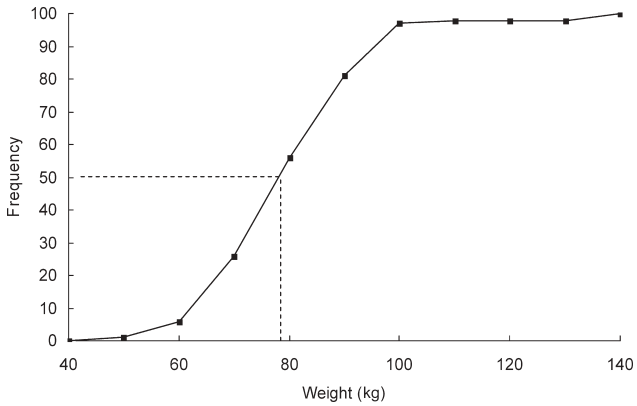

Fig. 2 Cumulative frequency curve showing the distribution of body weight in 100 males.

Scatterplots

A scatterplot is useful when looking at relationships or associations between two sample variables, *x* and *y* (bivariate data). In a scatterplot, both the *x* and *y* axes represent interval data. Each datum comprises an *x*–*y* pair of values represented by a dot or other symbol on the plot. If an association between two variables in more than one group is under investigation (e.g. variables height and weight in groups male and female), different symbols or colours may be used.

It may be apparent from looking at a scatterplot that there may be a linear relationship between the two variables, which may be positive (*y* increases with *x*) or negative (*y* decreases with *x*). The strength of this possible linear relationship may be tested statistically by calculating the correlation coefficient between the two variables. If the correlation coefficient is statistically significant, then linear regression analysis may be used to determine the approximate mathematical relationship between the two variables, and an equation may be derived.

Although there may be a strong correlation between two variables, it must never be assumed that this is a result of cause and effect. No statistical test can prove or negate this; before ascribing any association between two variables, because of cause and effect, all other potential reasons for the association must be considered and excluded.

Mathematical relationships between two variables

Correlation

A correlation refers to the relationship between two sets of paired interval data (e.g. height and weight of patients or the measurement of cardiac output using both a pulmonary artery catheter and a pulse contour analysis monitor). A linear correlation may exist if, when eyeballing the data, there looks to be a straight line relationship between the two sets of paired data. If both data sets are normally distributed, the Pearson correlation coefficient (r) is calculated; otherwise, either the Spearman (ϱ) or Kendal (τ) correlation coefficient is calculated. The correlation coefficient lies within the range -1 (for a perfect straight line negative correlation) to $+1$ (for a perfect straight line positive correlation). When $r = 0$, there is no correlation at all and eyeballing the data on a scatterplot reveals a completely random pattern. The statistical significance of r may be tested, and a *P*-value determined. When data are normally distributed, r^2 value may be quoted and lies between 0 and $+1$. When expressed as a percentage, it equals the amount of variance between the two sample variables that is shared. For example, if $r^2 = 0.8$ between *x* and *y*, 80% of the variance in *y* is because of variation in *x*, and *vice versa*.

Linear regression

If a statistically significant correlation exists between two variables, linear regression analysis may be used to calculate the equation for

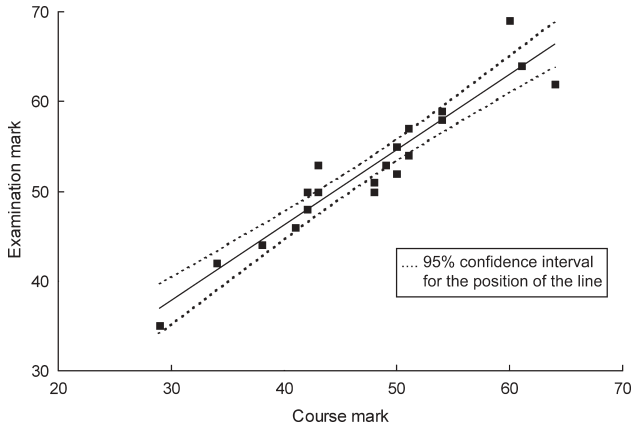

Fig. 3 Scatterplot of a preparatory course mock examination mark against actual performance in the examination itself. The correlation coefficient is very high ($r = 0.95$). The linear relationship between the two examination marks has been determined by linear regression analysis.

the straight line relationship. Confidence intervals for the slope and position of the line may also be determined. Figure 3 shows a scatterplot of the marks obtained in a preparatory course mock examination against performance in the actual examination itself. The two variables are well correlated ($r = 0.95$). The equation for the regression line is $y = 0.45x + 12.5$, with the 95% confidence interval shown by the dotted lines.

Bland–Altman plots

The determination of a correlation coefficient is a valid approach when comparing two entirely different (independent) variables that we believe may be associated (e.g. height and weight of patients or mean alcohol consumption and serum γGT concentrations). However, when comparing two different methods of measuring the same variable, this approach may be misleading and unreliable. This is because even if a statistically significant correlation is observed (which we would expect anyway as the two methods are measuring the same thing—unless one of the methods was entirely inaccurate and unreliable), there may still be clinically unacceptable differences between the output of one method and the other for individual measurements.

The standard statistical approach in this situation is to construct a Bland–Altman plot. In clinical research, we may wish to compare the performance of a new monitor with a gold standard or routine monitor. Thus, the performance of a pulse contour analysis cardiac output monitor may be compared with that of a pulmonary artery catheter, the acknowledged gold standard for measuring cardiac output. In a Bland–Altman plot, the mean of each individual data pair is shown on the *x*-axis and the difference between each data pair on the *y*-axis.

Another example comparing the two sets of examination marks, one from a preparatory course mock examination and the other from the actual examination itself, can also be analysed in this way, as the two examinations are essentially two different ways of measuring the same thing—the knowledge base of examination candidates. Clearly, the gold standard must be the actual examination itself.

The Bland–Altman plot shown in Figure 4 is a scatterplot of the mean examination mark of each candidate from the two examinations plotted on the *x*-axis against the difference between the two marks of each candidate (actual examination mark − preparatory mock course mark) on the *y*-axis. The average (mean) difference between all of the data pairs is shown by the solid line. It can be seen that, on average, candidates score five more marks in the actual examination than they do on the course. This average discrepancy between the two methods of measurement is termed the bias, and in this case the bias is +5. However, the bias is, by definition, an average discrepancy of the sample studied, but how does this apply to an individual candidate who wants to know how well he might perform in the actual examination from a knowledge of his score on the course? The 95% confidence interval for the range of differences between individual data pairs (limits of agreement) is indicated by the dotted lines, and equals twice the standard deviation of the distribution of the differences. The candidate now knows that on average his actual examination mark will be five marks better than his performance on the course, but may vary within the range −0.5 and +10.5 (with 95% confidence—however, there is a 5% chance that his mark will fall even outside this range). This is a large range that might well make the difference between a pass and a fail. The correlation coefficient between the two examinations is almost perfect ($r = 0.95$); yet, the Bland–Altman analysis has given a quite different impression.

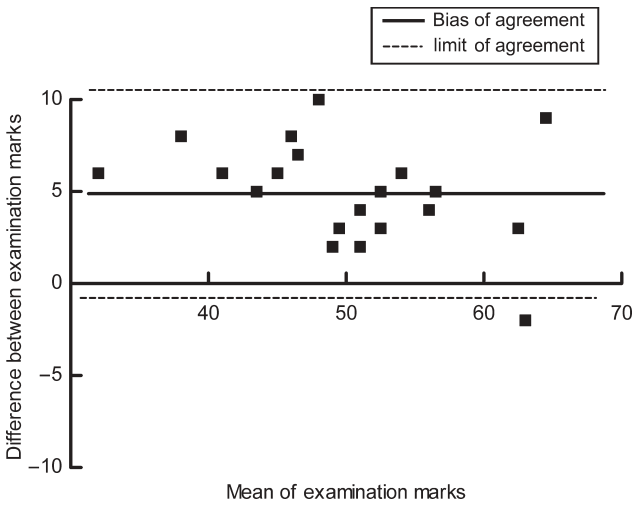

Fig. 4 Bland–Altman plot comparing performance in a preparatory course mock examination against performance in the actual examination.

Acknowledgement

The authors are grateful to Professor Rose Baker, Department of Statistics, Salford University for her valuable contribution in providing helpful comments and advice on this manuscript.

Bibliography

Bland M. *An Introduction to Medical Statistics*, 3rd Edn. Oxford: Oxford University Press, 2000

Altman DG. *Practical Statistics for Medical Research*. London: Chapman & Hall/CRC, 1991

Rumsey D. *Statistics for Dummies*. New Jersey: Wiley Publishing Inc., 2003.

Swinscow TDV. Statistics at square one. http://bmj.bmjjournals.com/collections/statsbk/index.shtml (accessed on 23 October 2006)

Lane DM. Hyperstat online statistics textbook. http://davidmlane.com/hyperstat/ (accessed on 23 October 2006)

SurfStat Australia. http://www.anu.edu.au/nceph/surfstat/surfstat-home/surfstat.html (accessed on 23 October 2006)

Greenhalgh T. *How to Read a Paper*. London: BMJ Publishing, 1997

Elwood M. Critical Appraisal of Epidemiological Studies and Clinical Trials, 2nd Edn. Oxford: Oxford University Press, 1998

Please see multiple choice questions 23–26