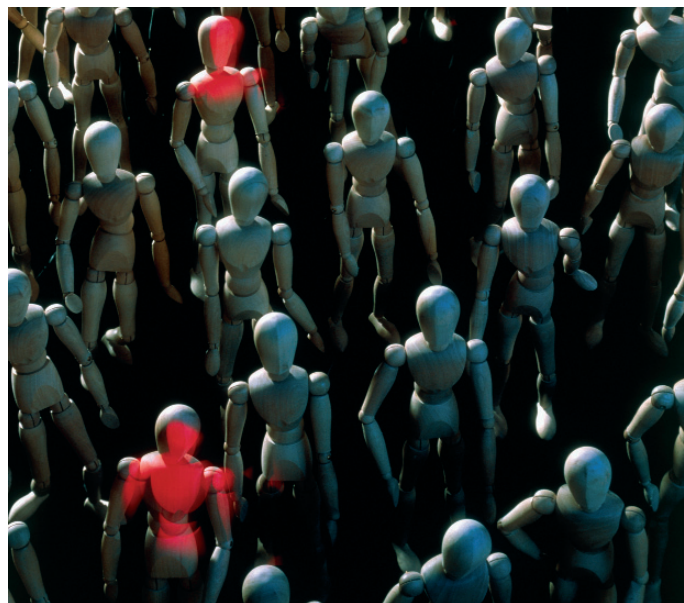


Analysing data

— choosing appropriate statistical methods

By Sarah L. Vowler, MSc

There are many different statistical tests that can be used to analyse data. When reporting results it is important that the tests used are appropriate for the type of data that have been collected. This article, the first in a special feature on statistics, describes the different types of data and the tests used to analyse them



R. MAISONNEUVE, PUBLIPHOTO DIFFUSION/SPL

The choice of which statistical test to use to analyse a set of data is completely dependent on the type of data that have been collected. This article describes how different types of data are categorised and which tests should be used to analyse them.

The following article in this feature (p47) will describe how to assess and interpret clinical papers, depending on the way results have been analysed and presented.

There are many different statistical methods that can be used in different situations. Each test makes particular assumptions about the data, as described in Panel 1 (p40). These assumptions should be taken into consideration when deciding which is the most appropriate test.

When analysing data it is useful first to get a feel for the data, what distributions the variables have and what inter-relationships exist. This can be done pictorially using scatterplots for comparing two continuous variables, matrix scatterplots for several continuous variables, histograms for individual variables and dot plots or boxplots

Sarah Vowler is a medical statistician at the Centre for Applied Medical Statistics, Department of Public Health and Primary Care, University of Cambridge

for continuous variables between groups (see Panel 2, p40 and Panel 3, p43). Categorical data can be plotted using pie charts or bar charts.

If the data to be analysed are arranged in groups and the comparison between groups is of interest then answering the four questions below may help to decide which test to use.

1. How many groups are there? The number of groups often determines which analysis to carry out. Different tests are used depending on whether the data are in one group, two groups or more than two groups. The tests for larger numbers of groups are often extensions of the tests for smaller numbers of groups, and thus their assumptions may be similar.

2. What type of data have been collected? The analysis chosen usually differs depending on whether the data collected are categorical or continuous. Categorical data are defined as data whose values are non-numeric (eg, gender). Continuous data are defined as data whose values are on a continuous scale of real numbers (eg, height). There are four different types of data: nominal, ordinal, interval and ratio. Nominal and ordinal data are categorical and interval and ratio data are continuous.

Nominal data In the most simple case, a binary nominal variable, (ie, one that has two categories), can be written as “yes” or “no” (eg, gender male — yes/no). Other examples of nominal variables are blood type and ethnicity. Nominal variables have the following three characteristics:

- The same value is given to all members of one level of a variable
- The same number cannot be assigned to different levels
- Each observation is in one level

Nominal data tend to be summarised using frequencies and percentages or the mode. Bar charts and pie charts can be constructed to display nominal data.

Ordinal data Ordinal data are also categorical but, unlike nominal data, there is an implicit ordering to the categories. Examples of ordinal data include socio-economic status, any ranked data and Likert scales. (Likert scales are a type of psychometric response scale often used in questionnaires. A statement is made and subjects can rate their response as, for example, 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree.) Ordinal data have the same three characteristics as nominal data.

With ordinal variables it is possible to say that one observation is higher than another in some sense, but it is not possible to say how much higher. Ordinal data can be summarised with frequencies and percentages or using the median. Again, bar charts and pie charts can be used to display this type of data.

Interval data Interval data is similar to ordinal data but have equidistant values. Examples of interval scales are visual analogue scales, a depression scale from 0–100 and temperature in Fahrenheit. Interval data are continuous, therefore it is possible to have decimal places between the equidistant values. With interval data it is possible to say how much higher one observation is than another, but distances between values do not necessarily carry the same meaning. Interval data can be summarised by use of means and standard deviations, or medians and interquartile ranges, modes or ranges. They can be displayed graphically by using scatterplots, dot plots, boxplots or means with 95 per cent confidence interval bars.

Ratio data The final type of data is ratio data. This is also a continuous measure. It is similar to the interval scale of measurement but has a meaningful zero point which represents a complete absence of the attribute. Examples include height, weight and temperature in Kelvin. Ratio data are summarised and displayed in the same way as interval data.

There are also some other data types that are special cases of these four types of data. For example, ranked data are a form of ordinal data in which data are ranked from first to last (eg, in a league table, or from the smallest to the largest observation).

3. Are the data normally distributed? The main assumption of many parametric tests is that the data follow a normal distribution. Non-parametric methods do not assume normality, and usually have fewer assumptions. The best way to decide if the assumption of normality holds is to “eyeball” the data using a histogram, which should form a bell-shaped curve.

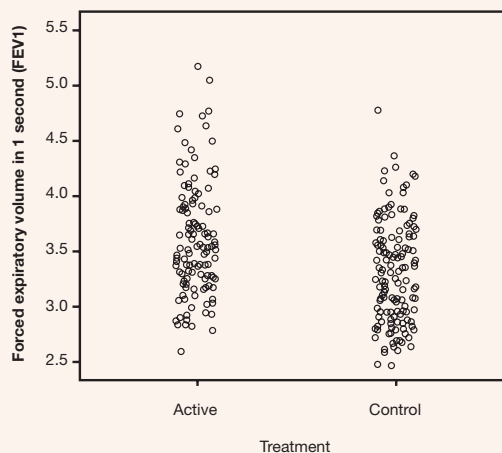
Panel 1: Assumptions

An assumption is an assertion that is assumed to be true for the test to be valid. Each statistical test has its own set of assumptions. It is important to check what the assumptions of a particular test are before carrying out the analysis, and to check that the data meet the assumptions. For example, the chi-squared test assumes that the data are counts, that the two variables are nominal, observations are independent of each other, 80 per cent of the expected counts exceed five and all exceed one. Standard texts will list the assumptions of statistical tests and methods.

Knowing the assumptions of the particular test should help you decide whether the test is appropriate for the data in question. If the assumptions are not met, the test is invalid and any conclusions drawn from the results of the test may not be correct.

Panel 2: Dot plots

A dot plot is a graph that plots a point (dot) for each observation, as shown in the example below. It is used to display continuous data within groups. The plots are often jittered (ie, a small random amount is added to each x value to separate the points so that they do not overlap).



If using paired tests, such as the paired *t*-test, the differences between variables should be normally distributed. When analysing data that are in independent groups, the data in each group should be normally distributed. For analysis of variance (ANOVA) (a method that uses variances to test for a difference in means) and regression (a term for models of variables that predict or explain an outcome variable) the residuals (ie, the difference between the observed and predicted value) should be normally distributed. Slight departures from normality can often be tolerated by tests if the sample size is large, and thus should not affect the result.

Statistical tests such as the Kolmogorov-Smirnov test can be used to test whether data are normally distributed, but tests of this type tend to be underpowered for small samples (ie, less likely to reject null hypothesis when it is false) and too sensitive for larger samples. Carrying out these tests has the effect of adding unnecessary hypothesis tests to the analysis (a hypothesis test is a statistical test that tests a hypothesis and produces a *P*-value) and so adds to the type I error rate (the probability of rejecting the null hypothesis when it is true). They are therefore usually best avoided.

4. Are the data independent? Different statistical tests should be used depending on whether the data are independent, paired, related or repeated measures data (see Panel 4, p43). Independent data form distinct groups — an example of this might be comparing age between men and women. Each person would either be in the male group or the female group with no overlap. Paired, related or repeated measures data can arise when something is measured repeatedly over time, or before and after an intervention. Measurements taken on the same person would always be considered to be paired.

Answering the four questions above will categorise the type of data being dealt with, making it easier to decide which test to use to analyse the data.

The following section describes the methods that can be used for analysis of data, depending on the type and distribution of the data and the number of groups involved.

Analysis of continuous data

Data in one group If the data are continuous, are in one group and are sufficiently normally distributed, it is possible to compare the mean value to a particular value using the one-sample *t*-test. However, if the data are not sufficiently normally distributed or if the sample size is small it may be more appropriate to compare the median value to a proposed median. This can be carried out using the sign test.

Data in two groups When continuous data are in two groups, the groups may be independent (unpaired), or paired. If the groups are independent and the data are sufficiently normally distributed and have

similar variances, then the unpaired *t*-test (also called the independent samples test) can be used to compare the means of the two independent groups. If the assumption of similar variances does not hold, then Welch's test can be used to compare the means.

If the groups are paired, the paired *t*-test can be used to compare the means between the two groups. The paired *t*-test requires the differences to be plausibly normally distributed and independent of each other. If the data are not sufficiently normally distributed or the sample size is small, the Mann-Whitney U test (also called the Wilcoxon rank sum test) can be used. If the two groups have a similar distribution, the Mann-Whitney U test can be used to compare medians between the two independent groups. If the two groups do not have similar distributions then the Mann-Whitney U test compares the shape and spread of the data between the two groups.¹

If the two groups are paired and not sufficiently normally distributed, the Wilcoxon signed rank test can be used to compare medians between the paired groups. The difference for each pair is calculated and plotted, and for this test to be valid this plot should be symmetric about the median. If this assumption does not hold then the sign test should be used to compare medians between the two groups.

More than two groups (independent)

If the data are continuous, there are more than two independent groups and the data

are sufficiently normally distributed, then ANOVA can be used to compare means between the groups. There are some non-parametric alternatives that can be used if the data are not sufficiently normally distributed or if the sample size is small.

If the groups all have similar distributions, the Kruskal-Wallis test can be used to test for an overall difference in medians between the groups. If the data within the groups do not have similar distributions, the Kruskal-Wallis test tests for a difference in shape and spread between the groups.

Panel 4: Paired, related and repeated measures data

Paired data Paired data are observations that are not independent — observations from two groups are paired in some way. This may be because randomisation was carried out in pairs, with one member being randomly assigned to one treatment, or there may be natural pairs being compared, such as measurements from two ears on the same person. Another form of pairing is looking at measurements before and after an intervention in the same person, for example, looking at urine production before and after a diuretic drug is given.

Related data Related data is another term for paired data and the terms can be used interchangeably. However, related data can mean there are more than two groups. It is therefore also interchangeable with repeated measures.

Repeated measures data Repeated measures data are where measurements are taken at two or more time points on the same patients (who may be in different groups) and are therefore not independent. For example, in a trial of a drug used to treat hypertension, blood pressure measurements might be taken weekly to examine what effect the drug has on blood pressure over time. Measurements on the same patient will be more similar than measurements on different patients and it is important that this is taken into account.

If the groups do not have a similar distribution, the median test can be used to test for a difference in medians.

ANOVA, the Kruskal-Wallis test and the median test can all be used as overall tests between the groups rather than doing pairwise tests between the groups straight away. If one of these tests shows that there is an overall difference between the groups, then post-hoc tests can be carried out to see where any pairwise differences between the groups lie. This is a way of cutting down the type I error rate (mistakenly rejecting the null hypothesis when it is true).

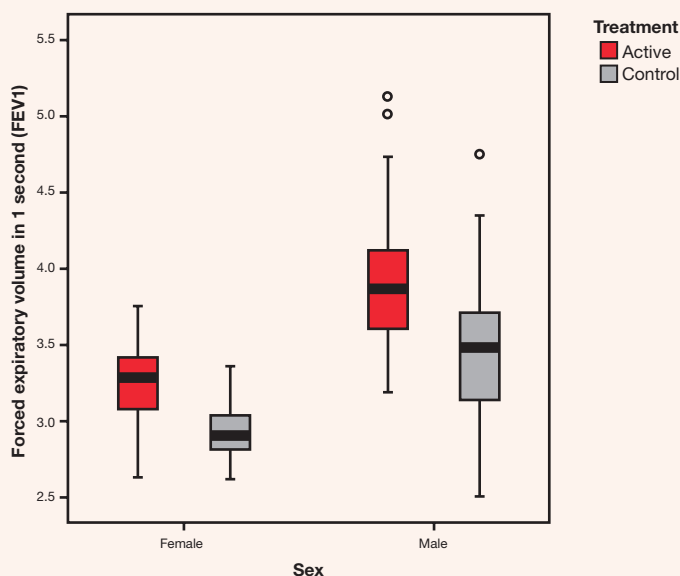
More than two groups (related) If the data are continuous, in more than two groups, the groups are related or repeated measures data, and the data are sufficiently normally distributed, repeated measures ANOVA can be used to analyse the data. Once repeated measures ANOVA has been carried out, a check that the residuals are sufficiently normally distributed should be carried out.

If the data are not sufficiently normally distributed, then there are some alternative non-parametric methods that can be carried out. First, there is the Friedman test. This test is an extension of the sign test and tests for an overall difference in medians between the related groups. If the Friedman test proves significant it is possible to carry out post-hoc tests to see where any differences between the groups lie. There is a specific post-hoc test method available for the Friedman test.

Alternatively, pairwise Wilcoxon signed rank tests or sign tests can be used. The Quade test is an extension of the Wilcoxon signed rank test. It also tests for a difference in medians between the related groups. The Quade test may be more powerful when there are three or four groups and the Friedman test when there are five or more groups.

Panel 3: Boxplots

An example of a boxplot (also called a "box and whisker plot") is shown below. The box displays the interquartile range (ie, the middle half of the data). The black bar within the box is the median value and the "whiskers" extend to the range (minimum and maximum) of the data unless outliers or extremes (unusually large or small values) are present.



If the data are not arranged in groups then other types of analysis might be appropriate, details of these are given below.

Correlation

If the association between two continuous variables is of interest then correlation should be used. For normally distributed data, Pearson's correlation coefficient (r) can be used. The coefficient of determination (r^2) is the proportion of variance explained by the association.

There are two non-parametric correlation coefficients that can be used when the data are not normally distributed. The first is Spearman's rank correlation coefficient, ρ (ρ). The second is Kendall's tau (τ) which is often used when there are many ties (identical values) in the data.

Categorical data: methods

The method of analysis most commonly used to analyse categorical data is the chi-squared test. The chi-squared test can be used to look at the association between two categorical variables if the data are independent. The chi-squared test requires that 20 per cent of the expected counts are greater than five and none are less than one. (An expected count is obtained by multiplying the row total by the column total and dividing by the total number of observations.) If this assumption is not met then the Fisher's exact test can be used instead. The Fisher's exact test is used to analyse data arranged in a 2 x 2 table. There is an extension of this called the Fisher-Freeman-Houlton test that can be used for larger tables.

If the categorical data are ordinal rather than nominal, it can be more powerful to carry out a test for trend, such as the chi-squared test for trend. This is also known as linear-by-linear association or the Mantel-Haenszel test for trend. This technique looks for a trend of increasing proportions and is therefore more powerful than the ordinary chi-squared test if the categories have an order.

If the data are categorical, in a 2 x 2 table, and data are paired, the McNemar test should be used to analyse the data. There are two extensions to the McNemar test, one for data collected at more than two time points, (called the Cochran's Q test), and one for data that have more categories than a binary outcome (the Stuart-Maxwell test).

Agreement

If two measurements of values are identical they are said to be in perfect agreement. Assessing agreement between variables or raters (who rate patients on a scale) is often carried out incorrectly. Agreement between two continuous measures is often assessed by

use of a correlation coefficient. However, this is inappropriate. Correlation measures whether one variable tends to increase as the other variable does. This is not what is being assessed with agreement, which is measuring how similar two variables are.

Correlation is also dependent on the range of the values. To measure agreement between two continuous measures, a method called limits of agreement, derived by Bland and Altman is used.² This method assumes a normal distribution of the differences between the two measurements on a subject. Furthermore, for this method to be appropriate, the differences plotted against the average of the two measurements should form a random scatter of points. The average difference and the 95 per cent limits of agreement are calculated (ie, the range in which the agreement will fall 95 per cent of the time) and added as reference lines to the plot. Limits of agreement can be used to assess agreement between two methods of measurement, two observers taking measurements from the same patients or the same observers taking measurements from a patient twice.

When looking at agreement between categorical data, analysis is often carried out by calculating the percentage agreement between the raters. If there are two raters rating the same subjects on the same nominal scale, the kappa coefficient, also called Cohen's kappa is a better measure of agreement. Kappa assesses the amount of agreement beyond chance, ie, how much better two raters agree than if the assignment was made by simply tossing a coin.

If the rating scale is ordinal then the weighted kappa can be used. This adjusts for however many categories away from agreement any disagreement lies.

Time to event data

Time to event data are a special sort of data. They tend to be highly skewed as times are not always fully observed. Such times are called censored observations — where a particular length of time has passed without the event happening. For example, when a RCT closes before all the patients have died, the patients still alive are referred to as being censored. Standard statistical methods therefore cannot be used, and special methods that take the censoring into account should be used instead.

The log-rank test can be used to compare time to event data between groups. Cox regression can be used for continuous data or adjusting for other variables, provided the proportional hazards assumption is met. The proportional hazards assumption is the assumption that the hazard in one group relative to the other does not change with time. Kaplan-Meier plots are used to display the time to event data and censoring.

Regression models

If the data are in the form of outcome and predictor variables then a regression analysis should be used for the analysis. For example, a regression model might be used to see if the predictor variables age, sex and blood glucose can be used to predict the outcome of cholesterol level. If the outcome is continuous and plausibly linear, then linear regression should be used. This requires the residuals to be normally distributed and a plot of predicted against residual values should show a random scatter of points. If the outcome is categorical with two outcomes (eg, yes/no) then binary logistic regression can be used to analyse the data. This also requires a normal distribution of residuals. If the outcome is an ordinal variable then ordinal regression can be used. If the outcome is count data (ie, data that are in the form of counts or frequencies) then Poisson regression can be used.

Conclusion

The most important part of choosing the correct test is to ensure that the test is appropriate for the type of data that have been collected. It is essential to check that the data meet the assumptions of whichever test is to be used, otherwise any conclusions drawn from the test may not be valid. When reporting the results of the test it should be clear which test has been used and the data should be summarised in the correct way and shown graphically if possible. Wherever possible, 95 per cent confidence intervals should be reported with the results of analysis. If in doubt, expert statistical advice should be sought for the analysis of data and interpretation of results, papers and meta-analyses.

References

1. Hart A. Mann-Whitney test is not just a test of medians: differences in spread can be important. *BMJ* 2001;323:391-3.
2. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet* 1995;346:1085-7.

Further reading

1. Altman DG, Machin D, Bryant TN, Gardner MJ. *Statistics with confidence*. 2nd edition. London: BMJ Books;2000.
2. Pett MA. *Nonparametrics for health care research*. Thousand Oaks: Sage;1997.
3. Siegel S, Castellan NJ. *Nonparametric statistics for the behavioural sciences*. New York: McGraw-Hill;1988.
4. Swinscow TDV, Campbell MJ. *Statistics at square One*. 10th edition. London: BMJ books;2002.

Other further reading materials are listed on p51.