

# Statistics III: Probability and statistical tests

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

## Probability theory

The probability of an event may be determined empirically (by observation) or mathematically (using probability theory). Probability theory is fundamentally important to inferential statistical analysis. Predicting population parameters from sample data is based on the assumption that the sample data are 'typical' of the population data. The laws of probability govern just how typical the data are. For example, we may toss a coin 20 times to determine the likelihood of obtaining heads on a single throw. Common sense tell us that, provided the coin is unbiased with heads just as likely to fall as tails, the ratio of heads:tails should be 1:1 and therefore the 'expected' outcome after 20 tosses would be 10 heads. However, the actual outcome may well be different. If we were to repeat the experiment by tossing the coin 1000 times, it is likely that the ratio heads:tails would be very close to 1:1 and if the coin was tossed an infinite number of times, the ratio would be exactly 1:1.

We may consider the population of interest in this scenario to be the outcome of an infinite number of coin tosses. A sample drawn from this population is an experiment in which the coin is tossed a finite number of times. Returning to the experiment in which the coin is tossed 20 times, probability theory may be used to determine mathematically the likelihood of obtaining any combination of heads and tails.

The probability ( $P$ ) of an event occurring is an expression of the relative frequency that the event occurs in an infinite number of trials.  $P$  ranges from 0 (the event never occurs) to 1 (the event always occurs). Let us consider eye colour and suppose that eyes are either blue, brown, grey, or green. The probability that an individual's eyes are blue is given by the expression  $P(\text{blue})$ . The four categories are exhaustive and mutually exclusive. Accordingly,  $P(\text{blue and brown}) = 0$ . The probability that an individual's eyes are coloured either blue or brown is given by the expression  $P(\text{blue or brown}) = P(\text{blue}) + P(\text{brown})$ . Because the above categories are exhaustive,  $P(\text{blue or brown or grey or green}) = P(\text{blue}) + P(\text{brown}) + P(\text{grey}) + P(\text{green}) = 1$ .

Generally, if we consider two mutually exclusive events A and B, then:

$$P(A \text{ and } B) = 0.$$
$$P(A \text{ or } B) = P(A) + P(B).$$

Additionally, if the events A and B are exhaustive,  $P(A \text{ or } B) = 1$ .

If two events are not mutually exclusive, they be independent (the probability of one event occurring is not affected by whether or not another event occurs). If two events A and B are not mutually exclusive but are independent, then:

$$P(A \text{ and } B) = P(A) \times P(B).$$
$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B).$$

## The binomial distribution

Returning to the coin tossing experiment, if we toss the coin three times, the following outcomes are possible: TTT, TTH, THT, HTT, THH, HTH, HHT, HHH. If we assume that the probability of obtaining either a head or a tail is equally likely on each toss, then each of the eight possible combinations listed above is equally likely to occur with a probability of  $1/8$ . Accordingly, the probability of obtaining no heads in three trials is  $1/8$ , one head is  $3/8$ , two heads is  $3/8$ , and three heads is  $1/8$ . The same exercise could be repeated by tossing the coin a 100 times (Fig. 1).

It is observed that as  $n$  increases, the number of heads obtained tends towards a normal distribution. This type of distribution of two independent outcomes is termed a binomial distribution. In this particular example, the probability of one of the outcomes (heads) is 0.5 per trial, but a binomial distribution may be defined for any probability, e.g. obtaining sixes after throwing a die a 100 times.

Generally, a binomial distribution may be used to describe any situation where there are  $n$  independent trials with two mutually exclusive, independent outcomes, the outcome of interest occurring with a probability of  $p$  on each trial. It follows a normal distribution provided  $n$  is

## Key points

The laws of probability dictate how typical a sample dataset is of the population from which it is drawn.

Which statistical test to use to analyse a dataset depends on a number of considerations including the type of data being analysed (e.g. interval or categorical), whether interval data are normally distributed or not and whether data are independent or paired.

Student's unpaired and paired  $t$ -tests are used to compare two groups of normally distributed independent and matched groups, respectively.

Analysis of variance (ANOVA) and repeated measures ANOVA are used to compare three or more groups of normally distributed independent and matched groups, respectively.

There are non-parametric equivalents of all the above tests.

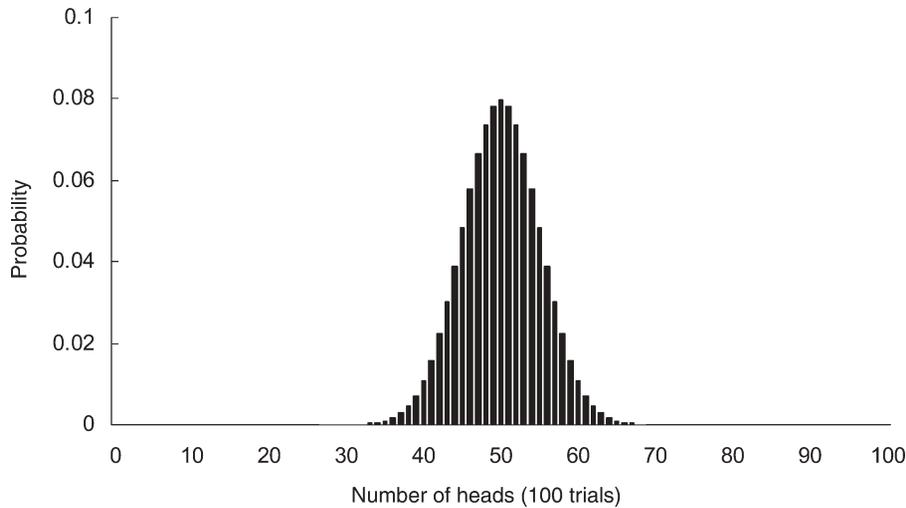
Categorical data are compared by drawing-up a contingency table and applying either Fisher's exact or  $\chi^2$  tests.

Anthony McCluskey BSc MB ChB FRCA

Consultant  
Department of Anaesthesia  
Stockport NHS Foundation Trust  
Stepping Hill Hospital  
Stockport, SK2 7JE  
UK  
Tel: +44 161 419 5869  
Fax: +44 161 419 5045  
E-mail: a.mccluskey4@ntlworld.com  
(for correspondence)

Abdul Ghaaliq Lalkhen MB ChB FRCA

Specialist Registrar  
Department of Anaesthesia  
Royal Lancaster Infirmary  
Ashton Road  
Lancaster, LA1 4RP  
UK



**Fig. 1** Binomial distribution: the probability of obtaining heads after tossing a coin 100 times. The area under the curve (which follows a normal distribution) = 1.

reasonably large and  $p$  does not take too extreme a value (close to 0 or 1). It can be shown mathematically that:

The mean of a binomial distribution =  $np$

The standard deviation of a binomial distribution =  $\sqrt{np(1-p)}$

## Statistical inference

### One group of interval data

Sometimes, we may wish to analyse just one dataset. For example, we may wish to infer a population parameter (e.g. mean) from a sample of the population data. Or, we may wish to determine whether the mean (or median) of a sample dataset differs from either a known value (e.g. population parameter) or a theoretical value.

#### Estimation of the population mean from sample data

Suppose that we want to know the average IQ of all UK trainees in anaesthesia. We will assume that IQ follows a normal distribution. As testing the entire population is impractical, we decide to test a random sample of 200 trainees. The data are analysed and the sample mean and sample standard deviation are calculated. How accurate is the sample estimate of the population mean (the mean IQ of all UK trainees in anaesthesia)?

If we were to repeat the same investigation numerous times, we would obtain a series of sample means that would follow a normal distribution. This is the central limit theorem and it applies even when the population data are not normally distributed. The mean of this sampling distribution is equal to the population mean  $\mu$ . The standard deviation of the sampling distribution equals  $SD/\sqrt{n}$  [i.e. the standard error of the mean (SEM)]. As we do not know the

population standard deviation, the sample standard deviation is used instead.

From previous discussions, we know that  $\sim 95\%$  of a sample of normally distributed data lies within  $\pm 1.96$  SD. Thus, the 95% confidence interval for the population mean IQ is given by the expression  $\bar{x} - 1.96 \times (SD/\sqrt{n}) \leq \mu \leq \bar{x} + 1.96 \times (SD/\sqrt{n})$ .

#### One sample t-test

Suppose now that we wish to know how the average IQ of UK trainees in anaesthesia compares with the 'known' data for the UK adult population as a whole. The estimate of the UK population data for trainees in anaesthesia obtained in the above investigation is used and compared with the known (published) data on the IQ of the adult UK population as a whole using a one-sample  $t$ -test.

A  $t$ -value analogous to the  $z$ -value previously discussed in relation to the standard normal distribution is calculated according to the equation:  $t = (\bar{x} - \mu)/SEM$ . For reasonably large samples, a  $t = 1.96$  returns a  $P$ -value of  $\sim 0.05$ . The  $t$ -value refers to the  $t$ -distribution, which is used in this situation, rather than the  $z$ -distribution because the population standard deviation (of UK trainees in anaesthesia) is unknown and values from the sample data are substituted. In fact, the  $t$ -distribution comprises a family of curves depending on sample size; the  $t$ -distribution used for a given sample size is specified by the number of degrees of freedom (equal to  $n - 1$ ).

#### Wilcoxon rank sum test

In this test, a sample median is compared against a known or hypothetical population median in a non-parametric distribution. Each sample datum is assigned a rank depending on how far it is from the median. Datum values lower than the median are given negative values. All of these signed ranks are summed to produce a  $W$ -value. If the null hypothesis is true,  $W$  is near to zero.

## Comparing two groups of interval data

In clinical studies, we often want to compare two sample groups. Two key criteria must be specified: are the data normally distributed and are the data paired?

### Unpaired (independent) normally distributed data: Student's unpaired two-sample *t*-test

For example, the efficacy of a new hypotensive drug A may be compared with an established drug B. The study has  $n_A$  patients in treatment Group A with sample mean  $\bar{x}_A$  and standard deviation  $SD_A$  and  $n_B$  patients in treatment Group B with sample mean  $\bar{x}_B$  and standard deviation  $SD_B$ ; ( $n_A$  and  $n_B$  do not have to be equal). We need to calculate the difference between the two sample means and the standard error of this difference between the two means, from which we can calculate a confidence interval for the difference between them.

For Student's *t*-test to be valid, the standard deviations of both groups must be similar. This is often the case, even when the sample means are significantly different. Most statistics software programs will routinely check that this is true. If the two sample standard deviations are observed to be unequal, Welch's correction to Student's *t*-test should be applied.

The standard error of the difference between the two means is given by  $\sqrt{(s^2/n_A + s^2/n_B)}$ , where  $s$  is the pooled sample standard deviation. It follows from previous discussions that the confidence interval for the difference between the two means is given by  $CI = (x_A - x_B) \pm t \times \sqrt{(s^2/n_A + s^2/n_B)}$  where the specific *t*-value depends on the confidence interval of interest (e.g.  $t = 1.96$  for the 95% confidence interval of a large sample). If the calculated confidence interval excludes zero, then we can be 95% confident that the difference between the two treatment groups is statistically significant (did not arise by random chance).

In the equation used to estimate the 95% confidence interval, a value of  $t = 1.96$  was used. In order to calculate a *P*-value for the observed difference between two study groups, assuming the null hypothesis to be true, the equation may be re-arranged:

$t = (x_A - x_B) / \sqrt{(s^2/n_A + s^2/n_B)}$ . The resulting *t*-value may be looked up in tables or calculated by a statistics programme. As discussed previously, because the *t*-distribution is a family of curves, the number of degrees of freedom has to be taken into account (equal to  $n_A + n_B - 2$ ).

### Paired normally distributed interval data: Student's paired two-sample *t*-test

The study comparing two hypotensive agents could be designed differently. Instead of having two independent groups, all of the patients recruited could be treated with one of the two study drugs (decided upon by random allocation) and the effect of treatment measured after a period of stabilization. Drug treatment is then stopped and after a washout period during which arterial blood pressure levels return to baseline levels, treatment with the other drug is commenced and its effect determined. This type of study

in which all of the subjects receive both drugs under investigation is called a crossover study. Each subject acts as his or her own control.

The design of a crossover study involves the analysis of matched pairs of data. In this situation, the appropriate statistical test is Student's paired *t*-test. Instead of analysing the data of two pooled groups, the effects of drug treatment on each individual in either arm of the study is separately analysed. As we shall see later, this form of analysis is more powerful.

### Non-parametric interval data

Student's *t*-test is not used for data that does not follow a normal distribution. The analogous statistical test to the unpaired *t*-test is the Mann–Witney *U*-test; the analogous test to the paired *t*-test is the Wilcoxon matched pairs test. Both tests analyse the data by comparing the medians rather than the means, and by considering the data as rank order values rather than absolute values.

### Three or more groups of interval data

The *t*-tests and their non-parametric equivalents are only used to compare two groups. When there are three or more groups under investigation, the appropriate test for normally distributed interval data is analysis of variance (ANOVA). If ANOVA testing suggests the groups are different, we are usually interested in knowing between which specific groups the differences exist. Thus, if we have three study groups A, B, and C with unequal means, is A different from B, A different from C, B different from C? One of several so-called *post hoc* tests may then be used to determine which differences are significant. This approach is inherently more robust than simply performing three two-sample *t*-tests as we only proceed to compare pairs of data once we have evidence of a significant difference between all of the study groups. ANOVA may also be used to compare just two study groups, when it is equivalent to Student's unpaired *t*-test.

When three or more normally distributed datasets are matched, the repeated measures ANOVA test is equivalent to Student's paired *t*-test. For data that is not normally distributed, the Kruskal–Wallis ANOVA by ranks test is used for independent groups and the Friedman test for matched datasets.

### Categorical data

When data are classified into groups, either the Fisher exact or the  $\chi^2$  test is used to determine whether the sample proportions in

**Table 1** Contingency table

	Outcome 1	Outcome 2
Group 1	A	B
Group 2	C	D

Relative risk =  $(A/(A + B))/(C/(C + D))$ ; Odds ratio =  $(A/B)/(C/D)$

**Table 2** Choosing a statistical test

Analysis required	Statistical test	
	Normally distributed data	Non-normally distributed data
Compare mean or median of one sample group against a known value	One sample <i>t</i> -test	Wilcoxon Rank Sum test
Compare means or medians of two sample groups (unpaired data)	Unpaired <i>t</i> -test	Mann–Whitney <i>U</i> -test
Compare means or medians of two sample groups (paired data)	Paired <i>t</i> -test	Wilcoxon Matched Pairs test
Compare means or medians of $\geq$ three sample groups (unpaired data)	ANOVA	Kruskal–Wallis ANOVA
Compare means or medians of $\geq$ three sample groups (paired data)	Repeated measures ANOVA	Friedman test
Compare two sample proportions of categorical data (unpaired data)		Fisher exact test
Compare two sample proportions of categorical data (paired data)		McNemar's test
Compare $\geq$ three sample proportions of categorical data		$\chi^2$ test

each group are significantly different. A contingency table containing the data is produced as previously described. A  $2 \times 2$  contingency table is best analysed using the Fisher exact test. For larger tables, the  $\chi^2$  test is used. Although the  $\chi^2$  test can also be used for analysis of  $2 \times 2$  tables (when Yates' correction is usually applied), it gives a less accurate result. It was used in the past, as it is easier to calculate than the Fisher exact test. However, with the widespread availability of computer software packages for statistical analysis, the Fisher exact test is preferable.

Two statistical measures of the relative likelihood of an event or outcome occurring in two sample groups may be defined: the relative risk and odds ratio. When the data in a contingency table relate to a prospective study, both these measures may be calculated. Only the odds ratio may be calculated for retrospective case–control studies. Calculation of relative risk and odds ratio are summarized in Table 1.

The relative risk of obtaining a given outcome after one intervention compared with another is equal to the ratio of the observed risk of the outcome after the first intervention divided by the observed risk of the outcome after the second intervention. The key factor in calculating relative risk is knowing the actual number of individuals at risk in each group. In a prospective study, the number of patients at risk of an outcome is known, whereas in a retrospective study, the outcome is the starting point and the number of patients at risk is not known.

The odds ratio is defined as the odds of the outcome of interest occurring after the first intervention divided by the odds of the outcome of interest occurring after the second intervention, where there are two mutually exclusive outcomes. The odds of an event or outcome in each of the two study groups  $= p/(1 - p)$ , where  $p$  is the probability of the outcome of interest and  $(1 - p)$  the probability of the alternate outcome.

Both relative risk and odds ratio are statistically valid approaches. Relative risk is usually preferred as it accords more to the commonsense notion of how we view relative risks between two groups when the data are eyeballed.

Table 2 summarizes which statistical test to use depending on the data to be analysed.

## Acknowledgements

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

1. McCluskey A, Lalkhen AG. Statistics I: data and correlations. *CEACCP* 2007; **7**: 95–9
2. McCluskey A, Lalkhen AG. Statistics II: central tendency and spread of data. *CEACCP* 2007; **7**: 127–30
3. Bland M. *An Introduction to Medical Statistics*, 3rd Edn. Oxford: Oxford University Press, 2000.
4. Altman DG. *Practical Statistics for Medical Research*. London: Chapman & Hall/CRC, 1991.
5. Rumsey D. *Statistics for Dummies*. New Jersey: Wiley Publishing Inc, 2003.
6. Swinscow TDV. Statistics at square one. Available from <http://www.bmj.com/statsbk/> (accessed 11 August 2007).
7. Lane DM. Hyperstat online statistics textbook. Available from <http://davidmlane.com/hyperstat/> (accessed 11 August 2007).
8. SurfStat Australia. Available from <http://www.anu.edu.au/nceph/surfstat/surfstat-home/surfstat.html> (accessed 11 August 2007).
9. Greenhalgh T. *How to Read a Paper*. London: BMJ Publishing, 1997.
10. Elwood M. *Critical Appraisal of Epidemiological Studies and Clinical Trials*, 2nd Edn. Oxford: Oxford University press, 1998.