Statistics IV: Interpreting the results of statistical tests

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

Key points

The null hypothesis proposes that there is no difference between the study groups with respect to the variables(s) of interest.

Choosing a statistical test depends on the type of data being analysed (e.g. interval or categorical), whether interval data are normally distributed, whether data are paired or unpaired and whether or not to perform a one- or two-tailed calculation.

Interpreting statistical analyses relies on an understanding of *P*-values, confidence intervals, statistical versus clinical significance, type I and II errors, and the hazards of multiple comparisons.

A type I error occurs when the null hypothesis is incorrectly rejected. A type II error occurs when the null hypothesis is incorrectly accepted.

Anthony McCluskey BSc MB ChB FRCA

Consultant Anaesthetist
NHS Foundation Trust Stockport
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

This is the fourth in a series of articles in this journal on the use of statistics in medicine. In the previous issue, we described how to choose an appropriate statistical test. In this article, we consider this further and discuss how to interpret the results.

More on choosing an appropriate statistical test

Deciding which statistical test to use to analyse a set of data depends on the type of data (interval or categorical, paired *vs* unpaired) being analysed and whether or not the data are normally distributed. Interpretation of the results of statistical analysis relies on an appreciation and consideration of the null hypothesis, *P*-values, the concept of statistical *vs* clinical significance, study power, types I and II statistical errors, the pitfalls of multiple comparisons, and one *vs* two-tailed tests before conducting the study.

Assessing whether a data set follows a normal distribution

It may be apparent from constructing a histogram or frequency curve that the data follow a normal distribution. However, with small sample sizes ($n < 20$), it may not be obvious from the graph that the data are drawn from a normally distributed population. The data may be subjected to formal statistical analysis for evidence of normality using one or more specific tests usually included in computer software packages, such as the Shapiro–Wilkes test. Such tests are fairly robust with larger sample sizes ($n > 100$). However, the choice between parametric and non-parametric statistical analysis is less important with samples of this size as both analyses are almost equally powerful and give similar results. With smaller sample sizes ($n < 20$), tests of normality may be misleading. Unfortunately, non-parametric analysis of small samples lacks statistical power and it may be almost impossible to generate a *P*-value of <0.05, whatever the differences between the groups of sample data.

When in doubt as to the type of distribution that the sample data follow, particularly when the sample size is small, non-parametric analysis should be undertaken, accepting that the analysis may lack power. The best solution to avoiding mistakes in choosing the appropriate statistical test for analysis of data is to design a study with sufficiently large numbers of subjects in each group.

Unpaired *vs* paired data

When comparing the effects of an intervention on sample groups in a clinical study, it is essential that the groups are as similar as possible, differing only in respect of the intervention of interest. One common method of achieving this is to recruit subjects into study groups by random allocation. All subjects recruited should have an equal chance of being allocated into any of the study groups. Provided the sample sizes are large enough, the randomization process should ensure that group differences in variables that may influence outcome of the intervention of interest (e.g. weight, age, sex ratio, and smoking habit) cancel each other out. These variables may themselves be subjected to statistical analysis and the null hypothesis that there is no difference between the study groups tested. Such a study contains independent groups and unpaired statistical tests are appropriate. An example would be a comparison of the efficacy of two different drugs for the treatment of hypertension.

Another method of conducting this type of investigation is the crossover study design in which all subjects recruited receive either treatment A or treatment B (the order decided by random allocation for each patient), followed by the other treatment after a suitable 'washout' period during which the effects of the first treatment are allowed to wear off. The data obtained in this study would be paired and subject to paired statistical analysis. The

Continuing Education in Anaesthesia, Critical Care & Pain | Volume 7 Number 6 2007 doi:10.1093/bjaceaccp/mkm042

effectiveness of the pairing may be determined by calculating the correlation coefficient and the corresponding *P*-value of the relationship between data pairs.

A third method involves defining all those characteristics that the researcher believes may influence the effect of the intervention of interest and matching the subjects recruited for those characteristics. This method is potentially unreliable, depending as it does on ensuring that key characteristics are not inadvertently overlooked and therefore not controlled.

The main advantage of the paired over the unpaired study design is that paired statistical tests are more powerful and fewer subjects need to be recruited in order to prove a given difference between the study groups. Against this are pragmatic difficulties and additional time needed for crossover studies, and the danger that, despite a washout period, there may still be an influence of the first treatment on the second. The pitfalls of matching patients for all important characteristics also have to be considered.

The null hypothesis and *P*-values

Before undertaking statistical analysis of data, a null hypothesis is proposed, that is, there is no difference between the study groups with respect to the variable(s) of interest (i.e. the sample means or medians are the same). Once the null hypothesis has been defined, statistical methods are used to calculate the probability of observing the data obtained (or data more extreme from the prediction of the null hypothesis) if the null hypothesis is true.

For example, we may obtain two sample data sets which appear to be from different populations when we examine the data. Let us consider that the appropriate statistical test is applied and the *P*-value obtained is 0.02. Conventionally, the *P*-value for statistical significance is defined as $P < 0.05$. In the above example, the threshold is breached and the null hypothesis is rejected. What exactly does a *P*-value of 0.02 mean? Let us imagine that the study is repeated numerous times. If the null hypothesis is true and the sample means are not different, a difference between the sample means at least as large as that observed in the first study would be observed only 2% of the time.

Many published statistical analyses quote *P*-values as ≥ 0.05 (not significant), < 0.05 (significant), < 0.01 (highly significant) etc. However, this practice resulted from an era before the widespread availability of computers for statistical analysis when *P*-values had to be looked up in reference tables. This approach is no longer satisfactory and precise *P*-values obtained should always be quoted. The importance of this approach is illustrated by the following example. In a study comparing two hypotensive agents, drug A is found to be more effective than drug B and $P < 0.05$ is quoted. We are convinced and immediately switch all our hypertensive patients to drug A. Another group of investigators conduct a similar study and find no significant difference between the two drugs ($P \geq 0.05$). We immediately switch all our hypertensive patients back onto drug B as it is less expensive and seems to be equally effective. We may also be somewhat confused by the apparently contradictory conclusions of the two studies.

In fact, if the actual *P*-value of the first study was 0.048 and that of the second study was 0.052, the two studies are entirely consistent with each other. The conventional value for statistical significance ($P < 0.05$) should always be viewed in context and a *P*-value close to this arbitrary cut-off point should perhaps lead to the conclusion that further work may be necessary before accepting or rejecting the null hypothesis.

Another example of the arbitrary nature of the conventional threshold for statistical significance may be considered. Suppose a new anti-cancer drug has been developed and a clinical study is undertaken to assess its efficacy compared with standard treatment. It is observed that mortality after treatment with the new drug tends to be lower but the reduction is not statistically significant ($P = 0.06$). As the new drug is more expensive and appears to be no more effective than standard treatment, should it be rejected? If the null hypothesis is true (both drugs equally effective) and we were to repeat the study numerous times, we would obtain the difference observed (or something greater) between the two study groups only 6% of the time. At the very least, a further larger study needs to be undertaken before concluding with confidence that the new drug is not more effective—as we shall see later, the original study may well have been under-powered.

Statistical *vs* clinical significance

Statistical significance should not be confused with clinical significance. Suppose two hypotensive agents are compared and the mean arterial blood pressure after treatment with drug A is 2 mm Hg lower than after treatment with drug B. If the study sample sizes are large enough, even such a small difference between the two groups may be statistically significant with a *P*-value of < 0.05. However, the clinical advantage of an additional 2 mm Hg reduction in mean arterial blood pressure is small and not clinically significant.

Confidence intervals

A confidence interval is a range of sample data which includes an unknown population parameter, for example, mean. The most commonly reported is the 95% confidence interval (CI 95%), although any other confidence interval may be calculated. If an investigation is repeated numerous times, the CI 95% generated will contain the population mean 95% of the time.

Confidence intervals are important when analysing the results of statistical analysis and help to interpret the *P*-value obtained. They should always be quoted with the *P*-value. Consider an investigation comparing the efficacy of a new hypotensive agent with standard treatment. The investigator considers that the minimum *clinically* significant difference in mean arterial blood pressure after treatment with the two drugs is 10 mm Hg. If $P < 0.05$, three possible ranges for CI 95% may be considered (Fig. 1). If

$P \geq 0.05$, four possible ranges for CI 95% may be considered (Fig. 2). These ranges for the CI 95% are summarized in Table 1.

Study power and types I and II statistical errors

After statistical analysis of data, the null hypothesis is either accepted or rejected on the basis of the *P*-value. As the null hypothesis may be either true or false in reality and the *P*-value obtained may be statistically significant ($P < 0.05$) or not, four possible outcomes need to be considered, as shown in Table 2.

If the null hypothesis is really true (i.e. there is no difference in reality between the groups) and the *P*-value obtained is ≥ 0.05, the conclusion based on the statistical analysis accords with reality. Similarly, if the null hypothesis is really false (i.e. there is a difference in reality between the groups) and the *P*-value obtained is <0.05, the conclusion based on the statistical analysis once again accords with reality.

However, if the null hypothesis is true and a *P*-value of <0.05 is obtained, the incorrect inference is drawn that the sample groups of data are different. This is termed a type I statistical error. A difference is found statistically where none exists in reality. The difference between the groups of sample data is not due to any intervention but rather by random chance. It is a fact of statistical life that whatever the value of *P*, there will always be a random chance of making a type I error, although the lower the *P*-value is, the smaller this becomes.

The final possibility to consider is that the null hypothesis is false in reality but the *P*-value obtained is ≥ 0.05. We have incorrectly concluded that the sample groups are similar—we have missed a real difference. This is a type II statistical error. The main cause of type II errors is inadequate sample size—the study lacks power. The power of a test is defined as $(1 - \beta) \times 100\%$, where β is the probability of a type II error. In order to be acceptable for publication, most editors of scientific journals require the power of a study to be at least 80%. The relationship between sample size and study power is shown in Figure 3.

It is good practice to perform a power calculation before commencing the clinical study proper in order to minimize the risk of obtaining a type II error and most journals and ethics committees require this to be explicitly defined in the methodology section. For example, in planning an investigation into the effect of a new inotrope on cardiac output, the investigator must decide the minimum difference between the cardiac output of controls *vs* active treatment that would be considered to be *clinically*

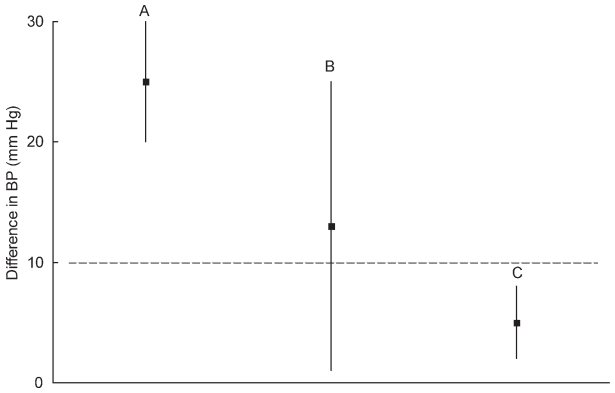

Fig 1 Statistical significance of the 95% confidence interval when $P < 0.05$ (Table 1).

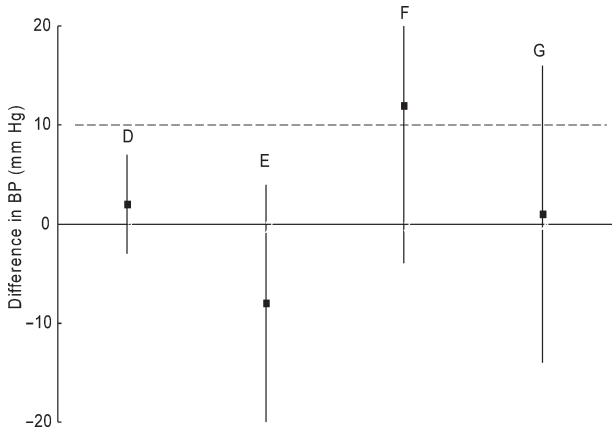

Fig 2 Statistical significance of the 95% confidence interval when $P \geq 0.05$ (Table 1).

Table 1 Interpreting confidence intervals

P-value <0.05	Is the difference between sample means clinically significant?		Interpretation
	At the lower end of CI range	At the upper end of CI range	
Yes	Yes	Yes	A: There is a clinically important difference between the study groups
Yes	No	Yes	B: Cannot reach a final conclusion—more data required
Yes	No	No	C: There is a clinically unimportant difference between the sample groups
No	No	No	D: There is no clinically important difference between the two groups
No	Yes	No	E: Cannot reach a final conclusion—more data required
No	No	Yes	F: Cannot reach a final conclusion—more data required
No	Yes	Yes	G: Meaningless range of CI—more data required

Table 2 Types I and II errors

P-value significant (<0.05)?	Null hypothesis true or false in reality	
	True	False
Yes	Type I error (α)	Analysis is correct
No	Analysis is correct	Type II error (β)

significant. Once this difference has been defined, the investigator needs access to data obtained from previously published work or an initial pilot study detailing the mean and standard deviation of the control data.

The dangers of multiple comparisons

Consider an investigation in which 20 different herbal remedies are studied for their effects on the amount of sleep obtained by subjects with insomnia. There is also a placebo group against which each of the 20 active treatment groups are compared. Multiple *t*-tests are performed for each of the herbal remedies and it is observed that one of them does appear to promote increased sleep when compared with placebo, with a *P*-value of <0.05. How valid is this conclusion?

In fact, the probability of any one of the 20 herbal remedies giving a statistically significant result at the level $P < 0.05$ is 1 in 20. Therefore, it would not be surprising if statistical analysis of one of the 20 remedies under investigation produced a $P < 0.05$ just by random chance. The correct approach when undertaking multiple comparisons such as this is to employ a correction factor. The most well known is Bonferroni's correction in which the *P*-value for significance is adjusted from $P < 0.05$ to $P < 0.05/n$ where n is the number of comparisons being made. Alternatively, an analysis of variance (ANOVA) between all 21 study groups should be performed, followed by *post hoc* individual comparisons calculated only if the *P*-value for the ANOVA is <0.05.

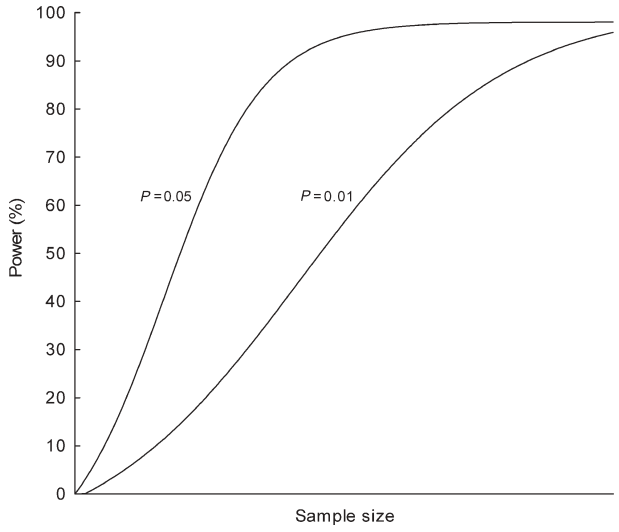

Fig 3 Relationship between power and sample size.

One *vs* two-tailed tests

All statistical tests start with the premise of the null hypothesis. This is then tested by calculating the probability that the differences observed between the sample groups are due to chance (the *P*-value). Let us consider an investigation comparing two sample means (e.g. mean arterial blood pressure after treatment of hypertension with two different drugs). When analysing such data, we obviously do not know whether the drugs are equally effective, if drug A is more effective than drug B or vice versa. Accordingly, when calculating the *P*-value, the key question is: what is the probability of obtaining the difference observed between the two sample means (or something more extreme) by random chance given that *either* group may have the higher mean? The two-tailed unpaired *t*-test answers this question.

It is almost always appropriate to conduct statistical analysis of data using two-tailed tests and this should be specified in the study protocol before data collection. A one-tailed test is usually inappropriate. It answers a similar question to the two-tailed test but crucially it specifies in advance that we are only interested if the sample mean of one group is greater than the other. If analysis of the data reveals a result opposite to that expected, the difference between the sample means must be attributed to chance, even if this difference is large.

For example, the organizer of a statistics course subjects the candidates to an MCQ test both before and after the course. The course marks are then analysed using a paired *t*-test (as the data are matched pairs of pre- and post-course marks for each candidate). The organizer decides to use a one-tailed test as he is certain that candidates' knowledge must improve after the course and discounts the possibility that candidates will score less well after it. Somewhat surprisingly, after the data are analysed, the mean MCQ scores post-course are worse than pre-course with a *P*-value of 0.01. The correct statistical interpretation of this result is to attribute the observed difference as due to random chance. However, it may be indeed be true that candidates do perform less well after the course. Perhaps, the course is confusing or contains numerous errors of fact. The course organizer was wrong to use a one-tailed test in this situation—a two-tailed test would have been appropriate.

One-tailed tests should always be viewed with some suspicion. It is actually quite difficult to think of examples in clinical research where a one-tailed test is appropriate. One example might be a study of a neuromuscular blocking drug in which two different intubating doses are given to patients and the time taken for the train-of-four ratio to recover to ≥0.8 recorded. It is probably justifiable to discount the possibility that the lower dose of drug results in a longer recovery time.

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. *Contin Educ Anaesth Crit Care Pain* 2007; **7**: 95–9
2. McCluskey A, Lalkhen AG. Statistics II: Central tendency and spread of data. *Contin Educ Anaesth Crit Care Pain* 2007; **7**: 127–30
3. McCluskey A, Lalkhen AG. Statistics III: Probability and statistical tests. *Contin Educ Anaesth Crit Care Pain* 2007; **7**: 167–70
4. Bland M. *An Introduction to Medical Statistics*, 3rd Edn. Oxford: Oxford University Press, 2000
5. Altman DG. *Practical Statistics for Medical Research*. London: Chapman & Hall/CRC, 1991
6. Rumsey D. *Statistics for Dummies*. New Jersey: Wiley Publishing Inc, 2003
7. Swinscow TDV. Statistics at square one. http://www.bmj.com//statsbk/ (accessed 21 October 2007)
8. Lane DM. Hyperstat online statistics textbook. http://davidmlane.com/hyperstat/ (accessed 21 October 2006)
9. SurfStat Australia. http://www.anu.edu.au/nceph/surfstat/surfstat-home/surfstat.html (accessed 21 October 2006)
10. Greenhalgh T. *How to Read a Paper.* London: BMJ Publishing, 1997
11. Elwood M. *Critical Appraisal of Epidemiological Studies and Clinical Trials*, 2nd Edn. Oxford: Oxford University press, 1998

Please see multiple choice questions 26–28