Eur Respir J 2001; 18: 393–401
Printed in UK – all rights reserved


European Respiratory Journal
ISSN 0903-1936

*REVIEW*

# Statistics for the *European Respiratory Journal*

S. Chinn

*Statistics for the* European Respiratory Journal. *S. Chinn. ©ERS Journals Ltd 2001.*
**ABSTRACT: This review summarizes statistical methods likely to be needed by researchers. It is not a replacement for a statistics book, and almost no symbols or mathematics are used. It seeks to guide researchers to the appropriate methods and to make them aware of some common pitfalls.**

**Sections deal with methods for quantitative outcomes, both basic and more advanced, and parallel methods for qualitative or categorical outcomes. Reference is made to papers using the more advanced methods in the *European Respiratory Journal* in order that their relevance may be appreciated.**

**The paper seeks to improve the quality of papers submitted to the *European Respiratory Journal*, to reduce the revisions to papers required of authors, and to enable readers of the journal to gain more insight into the interpretation of results.**
*Eur Respir J 2001; 18: 393–401.*

Dept of Public Health Sciences, King's College London, London.

Correspondence: S. Chinn, Dept of Public Health Sciences, King's College London, 5th floor Capital House, 42 Weston Street, London SE1 3QD.
Fax: 44 2078486605

Keywords: Biostatistics
statistical tests
trial design

Received: March 19 2001
Accepted: March 20 2001

No researcher can ignore statistical methods, either in reporting their own results or in reading the literature. The December 2000 issue of the *European Respiratory Journal* (*ERJ*) contained 20 original articles, of which 19 had at least some statistical summary, and 13 used methods of greater complexity than that described below as "basic". As in every other subject, new methods are continually being developed and standards in use change. It is not possible in one article to explain details of statistical methods; rather this article will try to explain when particular methods are required, give useful references, and highlight some common pitfalls in analysis and presentation. Where reference is made to articles in the *ERJ* this is to highlight the relevance of the method to respiratory medicine. The pitfalls have not been referenced, as examples in print are a result of a failure of the reviewing and editorial process.

The subject of statistics is about all stages of research, not just the analysis. In any study the analysis should follow the design, and no amount of analysis can rescue a study with a bad, or the wrong, design for the question being examined. Major issues in design are therefore, presented first.

## Design

### *Randomized controlled trials*

Not too much needs to be said here about parallel group trials in which patients are randomized individually, as a summary of necessary procedure and how results should be presented are given succinctly elsewhere [1, 2]. Authors should read the guidelines while preparing a grant application, and certainly before starting a study. Failure to record full recruitment details, for example, may lead to difficulty in publishing the results. Some details which may not seem necessary to reporting an individual trial become relevant to researchers seeking to include all trials in a meta-analysis [3].

Ethics committees demand sample size or power calculations, as well as editors for publication [1]. There is no such thing as an "exact" sample size calculation, as the information taken from a previous or pilot study will not be precise, and a few per cent more or less subjects make little difference to the power to detect a given treatment difference. What is important is not to have a sample size that is too small or too large by a factor of, say, 50% or more. It is never easy to decide on the minimum difference that the study should detect, and prior information on the variability of the relevant outcome may not be available. Equivalence studies, in which researchers seek to establish comparability of two treatments within given limits, generally require more subjects than those aiming to show a difference [4]. A medical statistician is used to discussing such issues and may be able to suggest an alternative design or outcome when difficulties occur, as well as perform the required calculations. The most commonly needed sample size calculations are described by CAMPBELL *et al.* [5].

Every effort should be made to obtain data on all randomized individuals so that an "intention to treat" analysis can be carried out. Otherwise the benefits of randomization, that the groups will on average be balanced on unknown prognostic factors, is lost. If compliance with treatment is not 100% then an "on treatment" analysis may also be presented, but should not replace the "intention to treat" analysis, as this may be biased.

Cross-over trials [6], in which patients with a chronic illness are given two or more treatments in random order, have a number of problems. Although in theory they require fewer patients than the corresponding parallel group trial, as subjects act as their own control, the necessary data on within-patient variation required to calculate sample size are often lacking, and selective drop-out can make the cross-over biased, or carry-over effects of treatment render the analysis invalid. Such trials should only be undertaken when clearly warranted [7].

### *Experimental laboratory studies*

A randomized controlled trial (RCT) is an experiment on people. When the experimental units are nonhuman animals the same principles should apply. It is less usual to randomize animals to treatment groups, as they are often inbred and assumed to be genetically identical. However, conditions of housing may vary in subtle ways, and if there is no reason against it, other than inconvenience, randomization should be used. Sample size justification, particularly for large animal studies, is becoming necessary. Compliance will not be an issue, but death of animals prior to sacrifice may prevent an "intention to treat" analysis. The biggest problem with animal experiments seems to be in over-complicated designs and in the analysis of serial measurements on a small number of animals. A researcher should have an analysis plan before starting the experiment, as is required for drug-licensing trials in humans and increasingly, in other RCTs.

### *Observational studies*

Studies may be prospective, cross-sectional or retrospective. Animal studies are nearly always prospective, albeit over short periods. In human studies the different designs have advantages and disadvantages. Prospective, also known as cohort or longitudinal studies, are optimal for studying risk factors for disease, survival or disease progression. However, particularly for the study of the incidence of rare diseases, they require follow-up of large numbers of people, and so are expensive, take time to do, and may have administrative problems and selective drop-out of subjects. Hence, retrospective case-control studies [8] are often used, in which controls are matched with cases of the disease and data on risk factors obtained by recall or searching medical records. The main issues here are selection of a suitable control group and whether to individually or group match the cases and controls, for example, on age. Repeatability studies are a special form of prospective study, as data are necessarily collected at two different time points. The main distinction is that the order of data collection should have no bearing on the result in a true repeatability, or reproducibility study, and also that the time scale is usually short. Particular methods of analysis apply, as referenced below. In development of models to predict disease progression two prospective studies are required, one to develop the model, the other to validate it, although one large study randomly divided into two may be used.

### *Cross-sectional surveys*

Cross-sectional surveys, in which all data are collected at the same point in time, are used for a variety of purposes. Many are comparisons of different patient groups. Those in which the aim is to assess disease prevalence in a population must be based on a sampling frame, *i.e.* a representative list of the population. Multicentre cross-sectional studies have been used to study variation in the prevalence of asthma and atopy [9, 10].

Case series may be cross-sectional or prospective, depending on whether the patients are followed-up. They may be used for hypothesis generation, but lacking controls or comparison group they rarely enable hypothesis testing, and will not be considered further.

### *Other types of study*

Although not yet very common in respiratory research, two other types of study deserve mention. The first is meta-analysis [3], in which no new data collection is undertaken but results from several studies are combined to give an overall result. Guidelines are now available for the reporting of such studies [11]. The other type of study to experience a large increase in popularity is that of cluster randomized trials (US terminology is cluster randomization trial). Instead of individuals being randomized to different interventions, whole family practices, geographical areas or other distinct units are randomly allocated. This may be because the intervention is at the cluster level [12], to avoid "contamination" between individuals [13] or to estimate the total community benefit [14]. A draft extension [15] of the CONSORT (Consolidated Standards of Reporting Trials) statement [1] for individually randomized trials has been published. Meta-analysis and cluster randomized trials were developed independently, but methods of analysis share some common features, as in meta-analyses data are clustered within studies.

The above is not an exhaustive list of types of study, but covers the ones most likely to be encountered.

## Distinction between outcome and explanatory variables

Before presentation and analysis can be discussed, the distinction between outcome, or dependent, variables and explanatory variables, also called independent or exposure variables, needs to be clarified. Usually there will be no confusion. In a randomized controlled trial, survival or recovery of the patient may be the outcome of interest and the treatment group is the explanatory variable. There may be additional explanatory variables, such as age

and sex, and these should include any variable used to stratify the patients in an RCT. However, in some circumstances there is ambiguity. In a case-control study, subjects are selected as having the disease, the cases, or not having the disease, the controls, and the measured potential risk factors are the outcomes of the study. The data analysis proceeds by treating "caseness" as the outcome and the risk factors as explanatory variables, but strictly speaking the opposite is true. In a study of asthmatic patients presenting in Accident and Emergency it is possible to compare the ages of patients that do or do not require admission or to analyse the risk of admission by age. In the first analysis, age is treated as the outcome and admission the independent variable, but more logically in the second, admission is the outcome. Although a conclusion that increasing age is associated with lower risk of admission might be found from either analysis, the second leads to results in a more useful form and also enables adjustment for risk factors other than age.

## Basic statistics

### *Descriptive statistics*

The first task is to describe the data, whether characteristics of groups being compared or baseline data in a prospective study [16] (table 1). The methods depend on whether the data to be described are continuous quantitative (ratio or interval scale in alternative terminology), such as forced expiratory volume in one second ($FEV_1$), discrete quantitative, for example, number of visits by a patient to his doctor, ordered categorical (ordinal), such as grading of severity of disease, or unordered categorical (nominal), a type which includes diagnosis, and also many binary variables, such as whether the patient survived or not. Both a measure of "central tendency", such as a mean, and one of variation need to be given, as shown in the first column of table 1. When data are skewed, medians and interquartile ranges may be more informative than mean and standard deviation. A separate row is not shown in the table for discrete quantitative data; when there are sufficient values they can be treated as if continuous, or when few values as ordered categorical.

### *Hypothesis tests and estimation*

The simplest hypothesis tests concern comparison of two groups and are classified in two ways. One is the nature of the outcome variable. The other classification is whether the subjects in the two groups are matched. The appropriate method of analysis depends primarily on these two features of the design and data, and are set out in table 1. The simplest example of matching is of subjects before and after treatment. Data are then said to be paired. A cross-over trial of two treatments is another example when the paired t-test can be used, but only if it is safe to assume that carry-over and time effects are negligible.

Comparison of means using a t-test assumes an underlying Normal distribution, and in the case of the unpaired t-test that the underlying standard deviations of the two groups are the same. t-tests are "robust" to non-Normality, *i.e.* they give quite accurate p-values and confidence intervals even when the distributions are skewed, and so researchers should not worry about this too much. It is never possible to "prove" Normality, and in small samples impossible to examine it. t-tests are preferable to the equivalent nonparametric tests as they are more powerful and give related confidence intervals more easily. It should always be made clear whether a t-test is paired or unpaired; the term "Student's" is unnecessary.

When data are positively skewed log transformation reduces skewness. Figure 1 shows serum total immunoglobulin-E (IgE) for a sample of males and females before and after taking logs to base 10. As tables 2, 3 and 4 show, the standard deviations before transformation were much bigger than the mean, and both mean and standard deviation are greater for males than for females. After transformation the standard deviations are almost equal. Although the log-distributions are not quite Normal they are close enough to allow comparison using an unpaired t-test. All calculations are carried out on the log values, but for presentation the means should be antilogged to give geometric means, and the difference in means and confidence interval to give the ratio of the geometric means and its confidence interval (tables 2, 3 and 4). The p-value quoted is the one derived from the log values. The base of the logarithms used does not

Table 1. – Basic statistical methods for two-group comparisons

| Type of data and summary statistics | Paired design | Unpaired design |
|---|---|---|
| Continuous quantitative data<br>Summary: mean±SD<br>*median and interquartile range* | Paired (one-sample) t-test<br>*Wilcoxon signed rank test* | Unpaired (two-sample/independent) t-test<br>*Wilcoxon rank sum test** |
| Ordered categorical data<br>Summary: median and interquartile range | Wilcoxon signed rank test | Wilcoxon rank sum test* |
| Unordered categorical data<br>Summary: proportions or percentages | McNemar's test if two categories | Chi-squared test#<br>*Fisher's exact test* |

Less usual methods are indicated in italics. *: =Mann-Whitney U-test; #: =z-test for proportions if two categories.

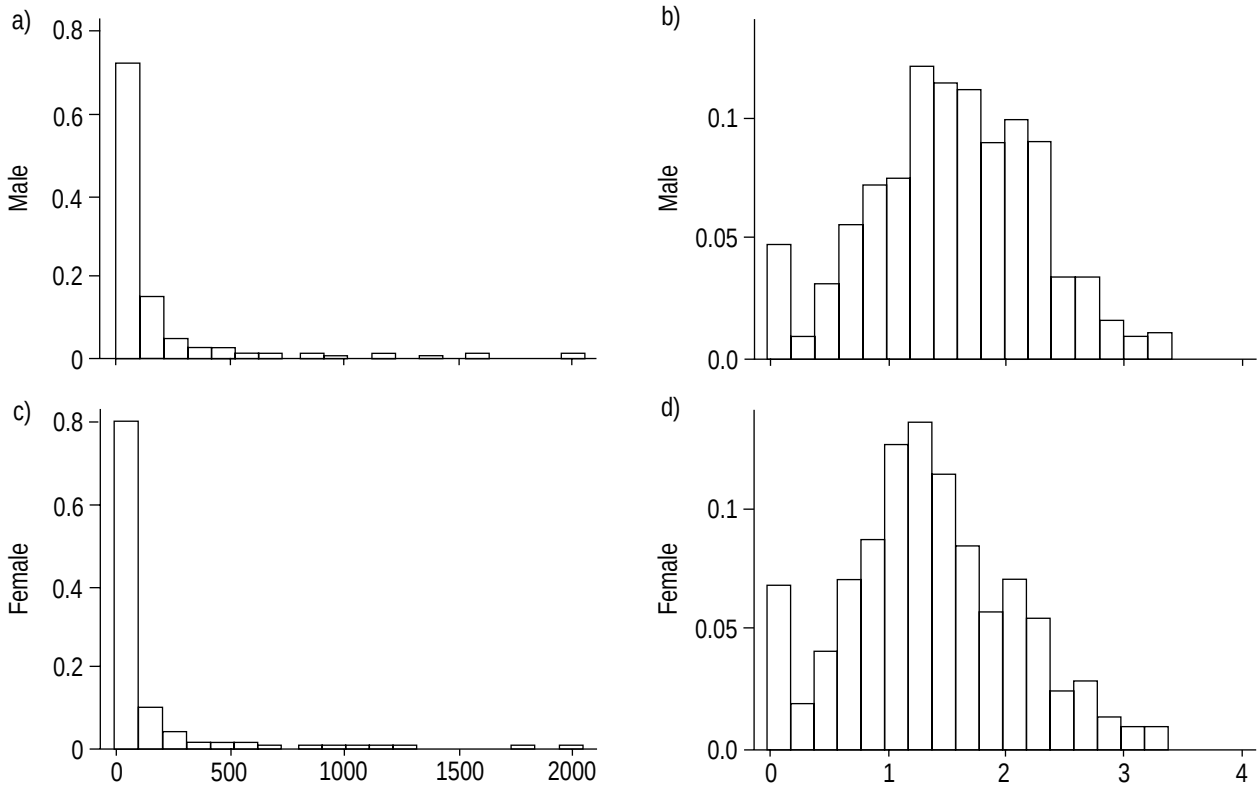


Fig. 1. – The distribution of serum total immunoglobulin-E a) before and b) after $\log_{10}$ transformation in males and c) before and d) after $\log_{10}$ transformation in females in the British arm of the European Respiratory Health Survey.

matter provided the antilogging (exponentiation) corresponds. Due to the extreme skewness of the raw IgE values, the geometric means are much less than the arithmetic means and much more descriptive of central tendency.

Chi-squared tests should not be used when numbers in cells are very small. The most quoted criterion is that "80% of expected numbers should be at least 5". In a two by two table this is satisfied if all the observed numbers are $\geqslant 5$. The z-test for difference in proportions gives the same p-value and is more informative as a confidence interval for the difference in proportions can be derived. Both the Chi-squared test and the corresponding z-test can be "corrected for continuity", which gives a slightly larger p-value. Statisticians still argue over whether this is necessary. If numbers are small, Fisher′s exact test can be used to give a p-value. McNemar′s test is a simple form of Chi-squared test used when binary data are paired, for example, in looking at changes in allergic sensitization in a sample of people over time. However, for paired data on variables with more than two categories there is no simple generalization.

Table 2. – An example of the analysis of positively skewed data: raw data-serum total immunoglobulin-E $kU \cdot L^{-1}$

| | Number in sample | Mean±SD |
|---|---|---|
| Males | 406 | 115.2±242.4 |
| Females | 509 | 92.8±230.4 |

Table 3. – An example of the analysis of positively skewed data: data transformed-$\log_{10}$(total immunoglobulin-E)

| | Mean | SD | SE | 95% CI | p-value |
|---|---|---|---|---|---|
| Males | 1.559 | 0.695 | 0.034 | 1.491–1.626 | |
| Females | 1.380 | 0.709 | 0.031 | 1.318–1.441 | |
| Difference | 0.179 | | 0.047 | 0.087–0.271 | 0.0001 |

95% CI: 95% confidence interval.

### *p-Values and confidence intervals*

Whatever the design of the study and estimate of interest a confidence interval is more informative than a p-value. In the case of a statistically significant difference between two treatments it enables the reader to judge the clinical importance of the difference; a small difference may be statistically

Table 4. – An example of the analysis of positively skewed data: results transformed back to $kU \cdot L^{-1}$

| | Geometric mean | Ratio of geometric means | 95% CI | p-value |
|---|---|---|---|---|
| Males | 36.22 | | 30.97–42.27 | |
| Females | 23.99 | | 20.80–27.61 | |
| Difference | | 1.51 | 1.22–1.87 | 0.0001 |

95% CI: 95% confidence interval.

significant if the study is large. Conversely, and more commonly, a p-value >0.05 neither precludes the data being compatible with a difference of clinical importance nor proves equivalence [4]; a large p-value may be the result of too small a study. The 95% confidence interval shows the range of plausible values for the estimate and should always be given if possible for the main comparison of interest. It is usual also to provide the associated p-value, which is the probability of getting the observed result (or one more extreme) if the "null hypothesis", usually a statement of a chance finding, is true. The p-value is a measure of strength of evidence against the null hypothesis. It should be quoted as an actual value to two decimal places, and not as, for example, "p<0.05" or "NS" implying p>0.05. Values between 0.001–0.01 should be given to three decimal places. "p<0.001" is acceptable, and occasionally, for brevity in the text, a statement such as "no other factor was significantly associated with outcome (p>0.3)".

p-Values should always be "two-sided", *i.e.* the possibility of a difference occurring in either direction needs to be allowed for. Only if the researcher can truly say that a difference in the opposite direction would be equivalent to no difference is a one-sided p-value appropriate. This is rarely the case.

Only descriptive statistics should be used to describe baseline data in a clinical trial. Provided the randomization has been performed correctly, the null hypothesis must be true and any imbalance is due to chance. Baseline data on any factor likely to be associated with outcome should be taken into account in the analysis whether or not imbalance is evident at baseline, as the precision of the treatment difference in outcome can be increased [17]. Analysis of change in the outcome from baseline to final value should be justified if used rather than the preferred analysis of final value adjusted for baseline.

At the other extreme, a hypothesis test should not be used to compare groups on any variable that is included in the definition of the groups, as then by definition the null hypothesis cannot be true. This applies, for example, to component parts of a score used to define disease groups.

Another misuse of hypothesis tests is to claim that a variable showing a baseline difference in means between two groups of patients, one of which is found to have better prognosis, is "predictive". Any small difference in means can be shown to be statistically significant if the samples are large enough. Only if the distributions of the measurements do not overlap, or only to a small degree, can the measurement be validly claimed to be predictive [18]. The degree of overlap can be described using the index of separation, the difference in means divided by the within-group standard deviation, sometimes known as the "effect size". However, when a new diagnostic test is proposed it is more useful to estimate the sensitivity and specificity [19] of the measurement for the optimal cut-off point; one study or random half should be used to establish the cut-off point and the other for the estimation. Estimation of positive and negative predictive values is of even greater value [20].

### *Range or confidence interval, standard deviation or standard error*

In most studies it is appropriate to quote a range or standard deviation when describing baseline data or patient groups, but a confidence interval or standard error when describing the main results, although there are exceptions to the latter. A confidence interval is preferred to a standard error, as the latter gives too reassuring a picture of the accuracy of the results. Similarly, a 95% range is more descriptive than a standard deviation; the full range depends on the sample size and is therefore, less useful [18]. Exceptions to giving a confidence interval in relation to the main results are in reporting reproducibility, in comparison of methods of measurement, or in reporting degree of prediction of a continuous outcome. In each of these some measure of variation of the individual values is appropriate.

### *Regression and correlation*

Equally as basic as t-tests are methods to relate one quantitative variable to another. Simple regression analysis provides an estimate of the linear increase in the outcome variable for unit increase in an explanatory variable, known as the regression coefficient, with associated confidence interval and p-value, and is usually more informative than the associated (Pearson) correlation coefficient, which gives a measure of linear association between two variables. The hypothesis tests of no linear relation between the two variables based on the regression coefficient and the correlation coefficient, are equivalent in that the p-values are the same. Linear regression assumes Normality and constant standard deviation of the outcome variable for given values of the explanatory variable. The Pearson correlation coefficient is based on a Normal distribution of both variables and is heavily influenced by outliers. Nonparametric correlation coefficients, Spearman′s or Kendall′s, can be used when the assumptions are violated. Data should always be plotted first, as only if the relation is at least approximately linear is it sensible to use either linear regression or Pearson′s correlation. Spearman′s rank correlation coefficient will show the degree of any monotone relation.

## Extensions to basic methods

Unfortunately the above methods rarely suffice. Fortunately most of them generalize to more complicated designs, so only a little more effort is required once the above have been mastered. There may be more than two groups, or more than one explanatory variable, in any of the cases so far mentioned.

## Continuous outcomes

### *Analysis of variance*

The unpaired t-test is a comparison of two means in relation to the within-group variation. The bigger the

variation, the more the two means are expected to differ by chance. One-way analysis of variance is an extension to more than two groups; the p-value provides evidence against equality of all group means. It should be used when the difference between any two groups is of interest, followed by a test of pair-wise group differences in means only if the analysis of variance suggests that some difference does exist. The test of pair-wise group differences should be a test specific for this, such as Duncan′s multiple range test, Neuman-Keuls test [21], Tukey′s test or Scheffé′s test. If t-tests are used the p-values will be too small, as they do not allow for the fact that the k(k-1)/2 comparisons of k groups are not independent. These tests should not be confused with the Bonferroni adjustment of p-values when a number of different independent outcomes are analysed. The Bonferroni correction is not recommended [22], although debate continues.

A two-way analysis of variance is one in which two explanatory variables are cross-classified, for example, different inhaled steroids may be compared at the same time as different inhaler devices. The effect of each on, say, peak flow variability as an outcome would need to be known. Analysis of variance would tell whether mean effects of each steroid differ when the inhaler device is kept constant, and whether the mean effects of devices differ when the steroid is kept the same. Provided the study was planned with a large enough sample, whether there is an "interaction" effect, *i.e.* does the difference between steroids differ between devices, can be investigated. In this case interaction is unlikely but not impossible.

The paired t-test is in fact a special case of the two-way analysis of variance. The factor of interest, before and after treatment or two different treatments on the same subject, has only two categories, while the subject, the other level, is usually not of interest. The treatment categories may be extended to three or more occasions or treatments, but can no longer use the simple paired t-test approach if all comparisons are of interest and should use a two-way analysis of variance. The exception is, as above, that if one treatment is a control group, each of the others may be compared with the control.

In a parallel group RCT it is common to follow patients for some time and obtain multiple observations. There are three factors, treatment, subject and time, and a three-way analysis of variance could be performed. It is usually the treatment-time interaction that is of interest, *i.e.* do the treatments have a different effect on the outcome over time, given that at time zero they were randomized to be equal. However, this may only tell us that the treatments differ not how they differ, and once there are more than, say, four time points, this approach becomes increasingly unhelpful. The repeated measurements must not be analysed as if from different subjects. Researchers may be tempted to compare treatments at each time point, but the tests are not independent. "Repeated measures analysis", which takes account of the correlation of repeated measurements on the same subject over time, can mean several things, so must be fully described. However, MATTHEWS *et al.* [23] have suggested a pragmatic approach to the analysis of serial measurements which gives more informative results.

### *Multiple regression*

Frequently, a continuous outcome needs to relate not to just one continuous explanatory variable but several [24]. Multiple regression estimates the increase in mean outcome per unit increase in each explanatory variable for fixed levels of each of the others. This can be used to estimate the regression coefficient of interest, "controlling" for other variables. It works provided the intercorrelation, or "colinearity", of the explanatory variables is not too great. When it is, the standard errors of the regression coefficients increase enormously. Only one of two very highly correlated variables should be included. Results should be presented as the regression coefficients with standard error or confidence interval. Where prediction is of interest the standard deviation of the differences between actual and predicted values should be reported (sometimes misleadingly termed "SEE").

### *Equivalence of multiple regression and analysis of variance*

Traditionally, before the advent of flexible statistical computing programs in the 1970s, analysis of variance was used for analysing a continuous outcome with categorical explanatory variables. When a single explanatory variable was continuous and was being used to adjust the relation of outcome to a categorical explanatory variable of primary interest, the term "analysis of covariance" was used. However, since statisticians recognized that analysis of variance and linear regression were just slightly different forms of a linear model, "analysis of covariance" has become an obsolete term. A linear model can contain as many explanatory variables as the data can support, both continuous and categorical. Counting one for the overall mean, one for each continuous variable and (k-1) for each k-level categorical variable, the total is the number of estimates required, which should not be more than $\sim$20% of the size of the data set, or leaving at $\geqslant$25 "degrees of freedom" remaining to estimate the residual variance. A multiple regression program can be used to analyse a k-level categorical explanatory variable by creating "dummy variables" for the (k-1) independent differences between categories. The major computer programs do this automatically. Any reader for whom this is a new idea should compare the effect of analysing 2-level categorical variables in an analysis of variance program with that of a multiple regression program, with the two levels of the variables coded as 0 and 1. At the simplest, perform an unpaired t-test and use simple linear regression and compare the results. It will be seen that the "regression coefficient" is the difference in means and that the p-values are identical.

*Stepwise analysis*

Stepwise regression [25] can be used to select variables associated with outcome, but should be used with caution. If there is a prior hypothesis to be tested then adjustment should be made for all variables which, based on the literature, may be associated with outcome, including stratifying variables in an RCT, even if the relations are not statistically significant in the current study. The loss of degrees of freedom is usually outweighed by the reduction in residual standard deviation, so that the confidence interval for the estimate of interest is narrowed. Only when a parsimonious model is required, perhaps in the development of a new diagnostic or prognostic scale, should a stepwise analysis be used. Backwards stepwise, in which all variables are included at first and eliminated in order of least statistical significance, is preferable to forwards stepwise, in which variables are entered in order of greatest statistical significance. The latter should only be used when there are too few data for backwards elimination. Neither approach guarantees that the final equation will be optimal.

*Repeatability and comparison of methods of measurement*

When continuous measurements to be compared are on the same scale, the methods of BLAND and ALTMAN [26] should be used. Estimation of repeatability for continuous outcomes is also described. If the methods produce categorical results which should be the same, the kappa statistic is appropriate [27]. When the measurements to be compared are on different scales their repeatability can be compared using the intraclass correlation coefficient [28]. Any monotone relation implies that one measurement could be calibrated in terms of the other.

## Further analysis of categorical outcomes

Before the analysis of categorical outcomes can be extended beyond Chi-squared tests several other summary statistics, which can be derived from two-by-two tables, need description. Consider a prospective or cohort study in which healthy subjects are followed-up to the relation of disease outcome to a risk factor; it may help to think of smoking and lung cancer. Table 5 shows the notation to be used here. Of those with the risk factor present at the start of the study (*e.g.* smokers) a number "a" are found at follow-up to have the disease (*e.g.* lung cancer), while "b" do not. So the risk of the disease in those positive to the factor is the proportion "a/(a+b)". Similarly, in those without the disease (nonsmokers) the risk (of lung cancer) is "c/(c+d)". The difference in risk and associated confidence interval can be calculated. This is a measure of absolute effect of the risk factor (smoking). The ratio of these two risks can also be taken, which is called the relative risk or risk ratio (RR) and as the first name implies, is a relative measure that may be less dependent on disease incidence from one population to another or over different time periods. The RR is one when the "risk" factor has no effect, while the difference in risks is zero.

## Multiple logistic regression

In the analysis [24] it is probable to want to include adjustment for some variables, such as age and sex. The outcome (disease incidence) is a binary categorical variable for each subject. Either a person gets the disease or they do not. The appropriate analysis is multiple logistic regression. What is estimated in such analysis is the "odds ratio" (OR) associated with each unit increase in a continuous explanatory variable, or between the (k-1) categories of a k category explanatory variable. "Odds" is a betting term, the ratio of the chances for an event to the chances against, so reduces to the simple formula shown in table 5. The OR is one if there is no effect of the "risk" factor. Otherwise it is always further from one than RR and the difference between OR and RR is greater, the bigger the disease incidence. Unfortunately, OR is often loosely interpreted as RR and this may be misleading. Testing any of the null hypotheses, for OR, RR or difference in risk is approximately equivalent in the simple case to using the Chi-squared test, but it should be noted that a 95% confidence interval formulae for risk difference, OR or RR may include the null hypothesis value when the p-value is close to 0.05 or *vice versa*. Logistic regression actually produces an estimate of $\log_e(OR)$, but this and the related confidence interval can be antilogged and most programs do this automatically.

*Survival analysis*

In a cohort study, provided follow-up time is the same for those with and without the risk factor, and follow-up is complete, both the incidence of the disease and the initial prevalence of the risk factor can be estimated without bias. If the outcome is mortality then this can be achieved. If follow-up time is not constant then other methods, known as survival analysis, are required. This is appropriate when patients in a cohort are recruited at different times and allows data on date of death or disease incidence to be analysed, not just whether or not death or disease occurred. Results are displayed using a Kaplan-Meier

Table 5. – Summary statistics for cohort and case-control studies: cohort study

| Explanatory variable (risk factor) | Outcome variable (future disease) | | Total |
|---|---|---|---|
| | +ve | -ve | |
| +ve | a | b | a+b |
| -ve | c | d | c+d |
| Total | a+c | b+d | N=a+b+c+d |

Relative risk=(a/(a+b))/(c/(c+d)). Odds ratio=(a/b)/(c/d)=ad/bc.

survival curve [16, 29]. The association of survival with a single risk factor can be tested using a nonparametric test, the logrank test. This allows for the fact that not only do survival times have a very non-Normal distribution, but that for patients still alive the survival time is known only to be at least as long as current follow-up; their survival time is said to be censored. When several risk factors are to be analysed, or adjustment for other explanatory variables is required, the most common method of analysis is Cox proportional hazards regression, which estimates the ratio of the rate of dying or disease incidence between the two groups [16, 27]. This depends on the ratio being constant over time, hence the "proportional hazards" in the full name. The Kaplan-Meier survival curves may show that this is not the case, so the method should not be automatically applied. In the case of a single risk factor, similar p-values are often obtained from the logrank test and Cox regression.

*Analysis for case-control studies*

As already mentioned, when a disease is rare it is likely that a case-control study will be carried out rather than a cohort study, and a case-control study may also be the initial study to examine the plausibility of a new hypothesis. The link between smoking and lung cancer was first examined in this way. It is important to realize that the disease incidence can no longer be estimated, as a fixed number of cases and controls are selected (table 6), unless the study is "nested" in a cohort study from which cases and controls are drawn. If the controls are not individually matched with cases the OR can be estimated as shown, which, when the disease is rare, is a good approximation to RR. An OR adjusted for other explanatory variables can also be estimated using logistic regression with "caseness" as the dependent variable. If individual matching has been used then an OR can be calculated [27], and conditional logistic regression used to adjust for covariates.

## Meta-analysis

Meta-analysis is primarily a method for combining results from different RCTs in a systematic review [3, 30], but can also be used to combine results from observational studies [31] or across centres in a multicentre study [32]. The estimate are combined and weighted according to the amount of information provided by each study. The actual weights differ slightly between the different methods of meta-analysis. One reason that few meta-analyses have been published in respiratory disease may be that the analysis requires a common outcome to be reported from each study. Two systematic reviews [3, 30] found a mixture of continuous and categorical outcomes and in each two separate meta-analyses were performed. This is undesirable [33]; it is intended that this will be reported further elsewhere in relation to bronchial responsiveness. Meta-analysis is not without problems. This is a relatively new field and much is still being published.

Table 6. – Summary statistics for cohort and case-control studies: case-control study

| Past risk factor | Present disease | |
|---|---|---|
| | +ve | -ve |
| +ve | a | b |
| -ve | c | d |
| Total | a+c (fixed) | b+d (fixed) |

Odds ratio=ad/bc.

## Cluster randomized trials

If meta-analysis is relatively new, cluster randomized trials are all the rage but may be overused [13]. Again the literature is growing rapidly.

## Graphical methods

Graphs can illuminate the results and show whether the method of analysis was appropriate. Bar charts should be reserved for frequencies. Means should normally be displayed with two-sided error bars, which should always be defined [18].

## Software and reference to methods

This article deliberately says nothing about specific computer programs. There are many around, and all statistical software should be able to cope with descriptive and basic methods as described above, without error, if used correctly. Reference to the program used is not necessary when commonly used methods are reported, as the reference does not guarantee that the program has been used correctly or necessarily tell the reader exactly what has been done. "Analysis was carried out using a t-test (STATMAGIC)" is not informative; "mean $FEV_1$ was compared between the two patient groups using an unpaired t-test" is sufficient. Blanket statements about statistical methods repeated from one paper to another should never be used. The statistical analysis section should always be particular to the paper. Analysis of variance/multiple linear regression or multiple logistic regression can now be regarded as standard, so only methods beyond these need to be referenced or described in detail. As far as possible, reference should be to papers or books in print, as a reference to an out-of-print book is irritating for the reader (and referee) and may mean that the method has been superseded. A software reference is helpful for methods not implemented in the major packages.

## Further reading

Other methods not referred to here may at times be required. This article tries as far as possible to give accessible references, in both senses of the word. Some of these are to the excellent series of *British Medical Journal* articles by J.M. Bland and D.G. Altman, of which many more are available on a variety of topics. The book already referenced [27] is one of the best, and would meet most researchers′ needs.

## Final advice

Remember the audience and do not use methods more complicated than necessary. They will not impress this statistical editor!

### References


1. Moher D, Schulz KF, Altman DG, for the CONSORT group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Lancet* 2001; 357: 1191–1194.
2. Chaouat A, Weitzenblum E, Kessler R, *et al.* A randomized trial of nocturnal oxygen therapy in chronic obstructive pulmonary disease patients. *Eur Respir J* 1999; 14: 1002–1008.
3. Gøtsche PC, Hammarquist C, Burr M. House dust mite control measures in the management of asthma: meta-analysis. *BMJ* 1998; 317: 1105–1110.
4. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. *BMJ* 1996; 313: 36–39.
5. Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons (published erratum appears in *BMJ* 1996 312: 96). *BMJ* 1995; 311: 1145–1148.
6. Criqui GI, Solomon C, Welch BS, Ferrando RE, Boushey HA, Balmes JR. Effects of azithromycin on ozone-induced airway neutrophilia and cytokine release. *Eur Respir J* 2000; 15: 856–862.
7. Sibbald B, Roberts C. Understanding controlled trials. Crossover trials. *BMJ* 1998; 316: 1719.
8. Bodner C, Godden D, Brown K, Little J, Ross S, Seaton A. Antioxidant intake and adult-onset wheeze: a case-control study. Aberdeen WHEASE Study Group. *Eur Respir J* 1999; 13: 22–30.
9. Asher MI, Keil U, Anderson HR, *et al.* International study of asthma, and allergies in childhood (ISAAC); rationale and methods. *Eur Respir J* 1995; 8: 483–491.
10. Burney PG, Luczynska C, Chinn S, Jarvis D. The European Community Respiratory Health Survey. *Eur Respir J* 1994; 7: 954–960.
11. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF, for the QUOROM Group. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *Lancet* 1999; 354: 1896–1900.
12. Premaratne UN, Sterne JA, Marks GB, Webb JR, Azima H, Burney PG. Clustered randomised trial of an intervention to improve the management of asthma: Greenwich asthma study. *BMJ* 1999; 318: 1251–1255.
13. Togerson DJ. Contamination in trials: is cluster randomisation the answer? *BMJ* 2001; 322: 355–357.
14. Hayes RJ, Alexander NDE, Bennett S, Cousens SN. Design and analysis issues in cluster-randomized trials of interventions against infectious diseases. *Stat Meth Med Res* 2000; 9: 95–116.
15. Extending the CONSORT statement to cluster randomized trials: for discussion. *Stat Med* 2001; 20: 489–496.
16. Aurora P, Wade A, Whitmore P, Whitehead B. A model for predicting life expectancy of children with cystic fibrosis. *Eur Respir J* 2000; 16: 1056–1060.
17. Roberts C, Torgerson TJ. Understanding controlled trials. Baseline imbalance in randomised controlled trials. *BMJ* 1999; 310: 185.
18. Chinn S. Ranges, confidence intervals, and related quantities; what they are and when to use them. *Thorax* 1991; 46: 391–393.
19. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. *BMJ* 1994; 308: 1552.
20. Altman DG, Bland JM. Diagnostic tests 2: Predictive values. *BMJ* 1994; 309: 102.
21. Gupta M, Hernadez-Juviel JM, Waring AJ, Bruni R, Walther FJ. Comparison of functional efficacy of surfactant protein B analogues in lavaged rats. *Eur Respir J* 2000; 16: 1129–1133.
22. Perneger TV. What′s wrong with Bonferroni adjustments. *BMJ* 1998; 316: 1236–1238.
23. Matthews JNS, Altman DG, Campbell MJ, Royston P. Analysis of serial measurements in medical research. *BMJ* 1990; 300: 230–235.
24. Black PN, Scicchitano R, Jenkins CR, *et al.* Serological evidence of infection with *Chlamydia pneumoniae* is related to the severity of asthma. *Eur Respir J* 2000; 15: 254–259.
25. Neder JA, Nery LE, Castelo A, *et al.* Prediction of metabolic and cardiopulmonary responses to maximum cycle ergometry: a randomised study. *Eur Respir J* 1999; 14: 1304–1313.
26. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; i: 307–310.
27. Altman DG. Practical Statistics for Medical Research. London, Chapman and Hall, 1991.
28. Chinn S. Repeatability and method comparison. *Thorax* 1991; 46: 454–456.
29. Bland JM, Altman DG. Survival probabilities (the Kaplan-Meier method). *BMJ* 1998; 317: 1572–1580.
30. Abramson M, Puy R, Weiner J. Immunotherapy in asthma: an updated systematic review. *Allergy* 1999; 54: 1022–1041.
31. Cook DG, Strachan DP. Parental smoking, bronchial reactivity and peak flow variability in children. *Thorax* 1998; 53: 295–301.
32. Chinn S, Burney P, Sunyer J, Jarvis D, Luczynska C, on behalf of the European Community Respiratory Health Survey. Sensitization to individual allergens and bronchial responsiveness in the ECRHS. *Eur Respir J* 1999; 14: 876–884.
33. Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. *Stat Med* 2000; 19: 3127–3131.