Statistics in Biomedical Laboratory and Clinical Science: Applications, Issues and Pitfalls

John Ludbrook

Department of Surgery, The University of Melbourne, Melbourne, Australia

Introduction

For most of us, acquiring a working understanding of statistics is a slow and painful learning process. My own case is probably typical [1]. I did my first statistical test (Pearson’s χ² test) as a medical student. I was introduced to Student’s t test some 5 years later. In each case, I did the tests on my own data. This process of acquiring new data, learning new tests, has continued all my life as an academic surgeon, then full-time cardiovascular researcher, and now consultant biostatistician with accreditation from the Statistical Society of Australia Inc. (AStat).

I hope this review will help readers acquire a working knowledge of statistical analysis less painfully and more expeditiously than I did. I assume that most readers have at least a basic understanding of statistics and have put this into practice. For the minority who merely want to make sense of the statistics in published papers, I strongly recommend a popular monograph [2]. I shall refer to a wide range of statistical procedures, indicate when they are (or are not) appropriate and concentrate on issues when certain statistical procedures should not be used. I shall refer to statistical texts and original articles in which detailed descriptions of tests are given, and to statistical software (Appendices 2 and 3). I also give a glossary of statistical symbols used in this review (Appendix 1).

Biomedical investigators should not make the mistake of supposing that the discipline of statistics is one in
which there are hard and fast rules. The fact is that there are a great many ‘schools’ of statistics, often with greatly differing opinions. I apologize in advance for the sin of self-citation, but I do not adhere to any single school of statistics and try to steer a middle course.

Designing the Study

First, one has to have an idea, a question, a puzzle to solve. Following this, a formal hypothesis is constructed. Note that this is a scientific hypothesis, not a statistical null hypothesis. Then comes the process of devising a method for testing the scientific hypothesis. Unless the proposed study follows on from previous work, it is essential to conduct a pilot study. This serves two purposes. It refines and validates the methods to be used, but also serves an important statistical function. Nowadays, institutional ethics committees and research funding bodies insist that investigators estimate the minimal sample (group) size (sometimes, and somewhat illogically, called ‘power analysis’). This requires the investigators to have made a pilot study or to find sufficient detail in published work, then to decide what size of difference from existing (control) values or proportions could be regarded as important. There is a dual purpose for this exercise. Investigators must ensure that their sample size is sufficiently large to minimize the risk of making false-negative errors – that is, declaring that there is no effect when in reality there is. At the same time they must ensure that the sample is not unnecessarily large, to warrant economy in the use of humans or animals.

But I should go back a step or two. The aid of a qualified biomedical statistical consultant should be enlisted in the early stages of planning a study, and certainly before writing the ethics or grant application [3, 4]. The biostatistician will help refine the experimental design and suggest the best method for analyzing the results when they are obtained. Then, guided by the investigator’s estimate of an effect that is important, either in the clinical or scientific sense, the statistician will estimate the minimal size for the experimental groups that will have a power of 80% (or sometimes 90%) to reject the statistical null hypothesis (no difference or effect) at a two-sided level of significance of $p = 0.05$. Experienced investigators may be able to make estimates of group size themselves, either by way of published tables [5] or by using computer software (see Appendix 3). It is important to allow for possible dropouts from a clinical trial, or missing observations in a laboratory study.

My own practice has been to write a detailed draft of the Introduction and Methods sections of a manuscript before starting a study. Both are essential components of applications for ethics approval or research grants, so this is time saved, not wasted. In addition, setting out the methods in detail, not least the statistical methods, ensures that everyone involved in a study has these to refer to. It is also useful to prepare a printed worksheet or laboratory workbook to receive the results as they come in, and to enter the results regularly into a computer spreadsheet.

Interim Statistical Analyses

It is usually dangerous to undertake statistical analysis of data while they are still being accumulated [6]. In the usual trial design, the investigators have decided in advance on the size of the groups they will need, and they must continue the experiment until this size has been reached before they analyze their data. There are two circumstances in which it is permissible to breach this general rule. They are sequential analysis and data monitoring.

Sequential Analysis

In laboratory, and sometimes clinical, research a special technique known as sequential analysis can be used [6]. There must be a large pool of experimental units (cell cultures, animals or humans) which can be drawn on, and the outcome for each experimental unit must be available within a short time of its having entered the study. In this design, two sets of boundaries are constructed in advance [6]. If the latest recruit to the trial crosses one set, it can safely be inferred that there is a difference in the effects of the two conditions or treatments at $p \leq 0.05$. If the latest recruit crosses the other set of boundaries, it can safely be inferred that there is no significant difference between the effects of the treatments (i.e. $p > 0.05$).

Data Monitoring

Nowadays, it is essential that for every clinical trial there should be a Data Monitoring Committee (DMC) [7]. This is composed of several experts in the field of study, and a statistician. The DMC and its members are completely independent of the investigators. It meets at predetermined intervals while the study is in progress. The statistician’s function is to present to the DMC the uncoded data, and to draw attention to any important
disproportion of adverse events between control and test groups, or proof that a ‘significant’ difference already exists between the groups with respect to the principal outcomes. The DMC can advise the investigators that the trial should be terminated prematurely on either of these grounds, or the DMC’s statistician can use the results that have accumulated to re-estimate the minimal group size. If this exceeds the original estimate, the DMC can give appropriate advice to the investigators. It may be that a larger sample is required or, occasionally, that the trial should be abandoned because the newly estimated minimal group size is simply unattainable.

**Statistical Analysis of the Results**

Investigators should decide in advance what sort of statistical procedures should be applied that will best test the hypothesis(es). But the first step is to assemble the data in formats that are compatible with the chosen tests. The best way to inspect continuous data is by plotting it out as scattergrams. Categorical data is best viewed by constructing tables of frequencies. If the statistical consultant is readily accessible, it is a good idea to ask him/her to inspect the data, in case there are obvious reasons to modify the planned methods of statistical analysis. The first thing to do is to look for missing values and outlying observations. David Finney [8], a distinguished and experienced biometrician, has recently given an excellent account of how to handle these, and Little [9] has written an exhaustive monograph on the subject.

**Missing Values**

Missing values are indicated by blank cells in the worksheet. The investigators’ task is to ascertain how these came about. In a clinical trial, they may be due to loss of patients to follow-up or because patients have withdrawn from the trial. The former case is catered for by censoring – the recognition that the outcome for the individual is unknown. This is taken into account in the analysis of survival [10, 11].

The investigators must try to track down the origin of missing values, for this determines what to do about them. Some of the more obvious causes are failure of equipment, human failure to make the observation, or unexpected death of an animal. Sometimes, for instance when dealing with two or more independent groups, missing values are just a minor nuisance. However, in longitudinal studies of the sort that might, for instance, be analyzed by repeated-measures analysis of variance (RM-ANOVA), missing values can be disastrous. Just one missing value for an individual at one point means that all the observations on that individual are ignored in the RM-ANOVA. The situation is not so catastrophic in the case of two-dimensional rectangular matrices for two-way ANOVA, or multidimensional matrices for multiway ANOVA, because the degrees of freedom can be reduced to take account of missing values.

When missing values are likely to have an important effect, they can be inserted by a process called ‘imputation’. For small data sets, the mean of adjacent values may be inserted by hand. Conservative statisticians, among whom I include myself, would introduce the caveat that insertion by hand is licit provided there is only one missing value for the individual, and provided it is not the first or last of the series, or that there is not more than one missing value per row or column in a rectangular matrix. However, for large data sets such as those found in surveys, it is common practice to use a computer routine to impute missing values even if they constitute 10, 20 or even 30% of the data set. Imputation routines are available in most general-purpose statistical software packages (see Appendix 2). Most statisticians would require proof that the missing values are randomly distributed, so that bias is not introduced by the process of imputation.

**Outlying Observations**

Outlying observations (outliers, in statistical shorthand) are best recognized by inspecting a scatterplot of the values within groups, or an x, y plot in the case of paired observations or observations that are serial in time, dose or space.

One possible cause for apparent outliers is that they are just that: apparent. We are used to thinking of populations as being normally distributed, because that is what most statistical texts tell us to expect. But in biomedical work the distribution of values, especially after some stimulus, is often log-normal – that is, skewed to the right. Thus an extremely high value may merely reflect the log-normal distribution of the population that was sampled; log transformation of the data may make an outlier less apparent (fig. 1).

For analyzing large data sets, such as those in surveys or epidemiological studies, some statisticians favor so-called ‘robust’ methods of statistical analysis in order to cope with outliers. One example of this approach is ‘trimming’. This usually consists of trimming off the top and bottom 10% of values. The 10% trimmed means are then analyzed by conventional statistical techniques. Robust
regression consists in minimizing the sum of the squared deviations of the trimmed set of x, y values from the regression line.

‘Real’ outliers may result from human transcription error – for instance, impossible values such as 370 °C for body temperature or 910 mm Hg for mean blood pressure. Or they may result from a sudden disorder of measuring equipment or aberrant behavior of the human measurer. Provided that the investigators can find a satisfactory explanation for an outlier, they are entitled to delete it and declare it to be a missing value. In turn, they can insert a replacement by one of the methods described above.

Choosing the Best Statistical Procedures
As desktop computers and statistical software have become ubiquitous, it has become even more essential to choose wisely. Statistical tyros are faced with an enormous menu of statistical procedures to choose from. Statistical software of the sort listed in Appendix 3 can be relied on to execute tests correctly – but this is of no use if the wrong tests are chosen. Here again, professional biostatistical advice is invaluable. A description of some important issues follows.

Hypothesis Testing versus Estimation
Since R.A. Fisher's heyday, tests of significance (a term he invented) have resulted in p values, where p refers to the probability of a null hypothesis being true. However, the alternative approach of estimation has become very popular over the past 20 years or so [12]. Estimation refers to the calculation of a confidence interval (CI), usually the 95% CI. This indicates the range of values that a given statistic (for instance, the difference between two means) is estimated to assume in the parent population(s). Whether p values or CIs are to be preferred is contested [3, 13]. My personal preference is for p values because they indicate the strength of the evidence against the null hypothesis. A way round the debate is to give both. For calculating the CI for continuous variables, Altman [12] gives the correct formula.

The process of hypothesis testing starts with proposing a null hypothesis – that there is no effect, no difference. Then p is the probability of falsely rejecting that null hypothesis, or the type I error rate (sometimes called α). In passing, p values should almost always be two-sided – that is, we look for an effect in either direction. Investigators must resist the temptation to use one-sided p values because they are smaller – this borders on dishonesty. There clearly must also be a type II error rate. This is the probability of falsely accepting the null hypothesis (sometimes called β). The power to reject the null hypothesis is defined as 1 – β.

The Multiple Comparison Problem
If the analysis of one's data results in a single p value or CI, there is no problem. But if multiple tests are performed on the same data set, or if a single test results in multiple p values – beware! The so-called familywise type I error rate is normally set at p = 0.05. It will be intuitively apparent that if enough hypotheses are tested, then the odds that for at least one hypothesis the p value will be <0.05 are greatly increased, and therefore the risk of false-positive inferences. There is a massive literature...
on how to guard against this, summarized in Curran-Even Everett and Benos [3] and Ludbrook [14]. An excellent preventive measure is to test global hypotheses whenever possible [15]. If the global test (for instance, one-way ANOVA) is ‘significant’, then post hoc pairwise contrasts can be done to identify precisely where the inequality of means lies. But a multiple comparison procedure must be applied. Following ANOVA, the Tukey-Kramer procedure for all possible pairwise contrasts between group means, or the Dunnett procedure for all other group means versus control, provide complete control of the familywise type I error rate (table 1) [14]. The two most versatile corrective measures, validated by Monte Carlo simulations, are the Ryan-Holm step-down Bonferroni procedure and Hochberg’s false discovery rate [3, 14]. These can be used in any setting in which multiple hypotheses are tested and they provide complete control over the familywise type I error rate.

**Sampling Methods and Models of Statistical Inference**

Under classical statistical theory, random samples are taken from large, defined, populations. Statistical inferences refer to the populations that have been randomly sampled. This is the essence of the population model of inference [16].

But random samples, as defined above, are rarely taken in biomedical research. In practice, investigators take a non-random sample of convenience, then divide it into two or more groups by a process of randomization. Statistical inferences refer only to the sample of convenience. This is the essence of the randomization model of inference [16–18]. This is the process used in clinical trials, where the sample of convenience might constitute all patients with a certain diagnosis attending a hospital as outpatients or inpatients. In laboratory research, the sample of convenience might be mice, rats or rabbits available from a breeding program.

The statistical tests with which we are familiar are designed for use under the population model of inference and include t tests, analysis of variance and $\chi^2$ tests on categorical data. Statistical tests that are designed for use under the randomization model of inference are known variously as permutation, randomization or exact tests [16–18]. But there is no need to despair – in most cases, tests that are valid under the population model of inference behave satisfactorily when there has been randomization, not random sampling. The exceptions are when group sizes are small or if the values within the groups are asymmetrically distributed.

---

**Table 1.** An example of one-way ANOVA with post hoc contrasts

<table>
<thead>
<tr>
<th>Group</th>
<th>Treatment</th>
<th>Group size</th>
<th>Mean blood pressure ± SE, mm Hg</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>placebo (control)</td>
<td>4</td>
<td>110 ± 1.8</td>
</tr>
<tr>
<td>B</td>
<td>β-blocker</td>
<td>4</td>
<td>106 ± 1.3</td>
</tr>
<tr>
<td>C</td>
<td>diuretic</td>
<td>5</td>
<td>104 ± 1.5</td>
</tr>
<tr>
<td>D</td>
<td>β-blocker + diuretic</td>
<td>5</td>
<td>90 ± 1.5</td>
</tr>
</tbody>
</table>

**b Pairwise post hoc contrasts with placebo control**

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Uncorrected p</th>
<th>Dunnett-corrected p</th>
</tr>
</thead>
<tbody>
<tr>
<td>A vs. B</td>
<td>0.121</td>
<td>0.270</td>
</tr>
<tr>
<td>A vs. C</td>
<td>0.021</td>
<td>0.052</td>
</tr>
<tr>
<td>A vs. D</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

A study is done in which 4 groups of hypertensive patients are recruited, and assigned to treatment with placebo, β-blocker, diuretic, or β-blocker plus diuretic. Global outcome of one-way ANOVA: $p < 0.001$. Group D is obviously the odd man out.

---

**The General Linear Model**

This constitutes the theoretical basis for almost all the statistical tests on continuous (interval scale) data [19]. It embraces Student’s t test, all forms of ANOVA, and linear regression analysis. It presupposes that (a) there has been random sampling of defined populations, (b) such populations are normally distributed (or at least symmetrically distributed), and (c) the populations of interest have equal variance. Postulate (a) rarely, if ever, occurs in biomedical research – but, for the moment, this can be ignored. Postulate (b) is not very important – Monte Carlo simulations have shown that ANOVA is very robust towards moderate degrees of non-normality. But postulate (c) is very important – a variance ratio between or among groups greater than 2–3 is anathema to t tests and ANOVA. In this case, the equivalent distribution-free test should be considered (see later).

The following three equations are examples of general linear models:

\[ Y = a + bx \]  \hspace{1cm} (1)
\[ Y = a + b_1x_1 + b_2x_2 + \ldots + b_nx_n \]  \hspace{1cm} (2)
\[ Y = a + b_1x_1 + b_2x_2 + b_3x_1x_2 \]  \hspace{1cm} (3)

$a$ and $b$ are regression coefficients, $a$ indicating the Y intercept and $b$ the slope of the regression. $Y$ is defined as
the dependent (or predicted) variable. The terms on the right of the equations are independent (or predictive) variables. Equation 1 is the model for one-way ANOVA, or for simple linear regression. Equation 2 is the model for multiway ANOVA, or for multiple linear regression. The terms on the right are often known in ANOVA as main effects. Equation 3 is the model for two-way ANOVA with two main effects and a two-way interaction term (b3x1x2). I shall show later what interactions mean and how important they are.

Comparing Two Means
This is surely the most elementary form of statistical analysis. It is usually done by Student’s t test – or its exact equivalent, one-way ANOVA (in the case of two means, F = t2). The formula for the t statistic is:

\[ t = \frac{\text{Difference between means}}{\text{Standard error of the difference}} \]  

(4)

How to calculate the standard error (SE) of the difference in the cases of independent (unpaired) and related (paired) sample can be found in standard texts [20, 21]. But look at the data in figure 1, representing two independent groups. The raw values for chronic obstructive airways disease in figure 1a seem to have a positive skew, and the sample variance is 2.71 times that of normal (table 2). The skewness is at least partly corrected by log10 transformation, so that the variance ratio is reduced to 2.22 (fig. 1b; table 2). Now look at the outcomes of statistical testing (table 2). A conventional t test on the raw data gives two-sided p = 0.049 – statistically significant! A modified t test in which the group variances are treated separately gives p = 0.052 – not significant! Would we do better with the log-transformed data, where there is less inequality of variance? This seems to have been successful, in that the corresponding p values are now 0.048 and 0.050. Let us try another strategy. Nonparametric (rank-order) tests are sometimes popular. The Wilcoxon-Mann-Whitney procedure is the appropriate one. It gives p = 0.072 and 0.072 – decidedly not significant. But what does the Wilcoxon-Mann-Whitney procedure test? It does not refer to the difference between medians – a common fallacy. Instead, it indicates that there is no significant difference in group mean ranks (table 2); however, biomedical researchers and readers find it hard to understand

| n1, n2 = Group sizes; \( \bar{X}_1, \bar{X}_2 = \text{means for groups } x_1 \text{ and } x_2 \) (normal, chronic obstructive airways disease); \( \bar{R}_1, \bar{R}_2 = \text{mean ranks for groups } x_1 \text{ and } x_2 \); SE1, SE2 = standard errors for \( x_1 \text{ and } x_2 \); \( s_1^2, s_2^2 = \text{sample variances for } x_1 \text{ and } x_2 \). SYSTAT 12 (Systat Software Inc.) used for t tests. StatXact 7 (Cytel Software Corporation) used for exact tests. |
|---|---|---|---|---|---|---|---|---|---|
| Raw | 10 | 10 | 14.2 | 15.4 | 8.2 | 12.8 | 0.31 | 0.52 | 0.99 | 2.68 | 2.71 |
| log10 | 10 | 10 | 1.15 | 1.19 | 8.2 | 12.8 | 0.03 | 0.04 | 0.0009 | 0.0020 | 2.22 |

Table 2. Outcomes of analyzing the data of figure 1

<table>
<thead>
<tr>
<th>Test</th>
<th>t</th>
<th>d.f.</th>
<th>Two-sided p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( t_{\text{pooled variance}} )</td>
<td>2.114</td>
<td>18</td>
<td>0.049</td>
</tr>
<tr>
<td>( t_{\text{separate variance}} )</td>
<td>2.114</td>
<td>14.8</td>
<td>0.052</td>
</tr>
<tr>
<td>Exact Wilcoxon-Mann-Whitney</td>
<td></td>
<td></td>
<td>0.072</td>
</tr>
<tr>
<td>Exact permutation on means</td>
<td></td>
<td></td>
<td>0.049</td>
</tr>
<tr>
<td>log10 – transformed data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( t_{\text{pooled variance}} )</td>
<td>2.123</td>
<td>18</td>
<td>0.048</td>
</tr>
<tr>
<td>( t_{\text{separate variance}} )</td>
<td>2.123</td>
<td>15.8</td>
<td>0.050</td>
</tr>
<tr>
<td>Exact Wilcoxon-Mann-Whitney</td>
<td></td>
<td></td>
<td>0.072</td>
</tr>
<tr>
<td>Exact permutation on means</td>
<td></td>
<td></td>
<td>0.048</td>
</tr>
</tbody>
</table>
what this means [22]! For this reason, I recommend that
rank-order tests never be used – they are a sort of poor
man’s permutation test. If one is to go down the distribu-
tion-free pathway, as I believe one should in such cases
[23], then the procedure to use is a permutation test for
the difference between means. This is not easy to explain
and it is best to refer to published accounts [17, 18]. The
formula is as follows:

\[
p = \frac{\text{No. permutations in which the}}{\text{All possible permutations,}}
\]
\[
\text{difference between means } \geq \text{ observed}
\]

\[
\text{retaining the actual group sizes}
\]

Using a permutation test makes sense, both in terms of
the hypothesis being tested and because it operates under
the randomization model of inference (randomization,
not random sampling). It gives \( p = 0.049 \) and 0.048. It
would be quite improper to do all these tests and select
the one that gives the desired outcome (e.g. the smallest
\( p \) value). The father of modern statistics, R.A. Fisher,
wrote in 1935 that permutation tests give ‘the possibility
of an independent check on the more expeditious meth-
ods in common use’ (he meant \( t \) tests). He had to do per-
mutation tests by hand, which took days or weeks, but
they can now be executed in a few milliseconds by means
of specialized computer statistics software (see Appen-
dices 2 and 3). I and some others believe that \( t \) tests are out-
moded and should be replaced by exact permutation tests
on means [16, 17, 21].

It is scarcely necessary to make the distinction be-
tween the independent-sample (unpaired) \( t \) test and the
related-sample (paired) \( t \) test, though I have come across
instances in which the wrong form was used.

**Comparing Many Independent Means**

Traditionally, this is done by means of one-way
ANOVA (see equation 1). It can also be done by the per-
mutation equivalent, which is an extension of equation 5.
Note that, in either case, the means must be indepen-
dent – a different group of experimental units is exposed
to each of the several conditions. If the \( p \) value that results
from one-way ANOVA is \( \leq 0.05 \), we are entitled to infer
that there is some inequality of means. Just where the in-
equality resides can be inferred by inspection – usually,
there is an odd man out. If investigators want to go fur-
ther with their analysis and make post hoc paired con-
trasts, a multiple comparison procedure must be applied
[14, 20]. **Dunnett’s procedure** is designed for comparing a
control mean with all the others. The **Tukey-Kramer pro-
cedure** is designed for all possible pairwise contrasts.

Both of these are proof against excessive familywise type
I error. An example of one-way ANOVA followed by pair-
wise contrasts is given in table 1. Most statistical software
packages have these two procedures in their repertoire.

**Comparing Many Related Means**

This is where two-way (see equation 3), or multiway,
ANOVA comes into play. The same group of experimen-
tal units is exposed to two or more sets of conditions, in
random order (fig. 2). In two-way ANOVA there are three
terms in the analysis – two main effects and one interac-
tion (equation 3). The main effects are of little interest.
They consist of a \( p \) value for equality of the two group
means, across both sets of conditions, and a \( p \) value for
equality of the means of the two sets of conditions, across
both groups. The effect of interest is the two-way interac-
tion between groups and sets of conditions. The \( p \) value
for the interaction term tests for parallelism (fig. 2), which
is the only hypothesis of interest. Provided the hypothesis
of interest is stated in advance, no multiple comparison
procedure needs to be applied to the three \( p \) values that
result from two-way ANOVA. No satisfactory permuta-
tion test has been devised for two-way ANOVA [16].

**Comparing the Means of Serial Observations**

This experimental design is used more often in animal
than in human studies. In it, the group or groups of ex-
perimental units is/are exposed to different conditions.
that are presented in serial (as opposed to random) order. The conditions can be, for instance, a sequence of times or distances, ascending strengths of a stimulus, or ascending doses of a drug. RM-ANOVA caters for this very powerful experimental design [15]. As in two-way ANOVA, the effect of interest is an interaction term; p for this term tests for identity of profile according to group and along the serial observations (fig. 3). As in two-way ANOVA, no multiple comparison procedure is necessary provided the interaction term is nominated in advance as testing the hypothesis of interest. But the raw p value does need adjustment for a phenomenon known as multisample asphericity or, more simply, as correlation between successive observations. The two most popular adjustments are the Greenhouse-Geisser and the Huynh-Feldt [15]. Though RM-ANOVA is a very powerful and justifiably popular test, it has one drawback. As pointed out earlier, if just one value for one experimental unit is missing, then all observations on that experimental unit are deleted from the analysis.

The outcome of the RM-ANOVA on the data of figure 3 is given in table 3. For the interaction between dose and treatment, p = 0.0156 (Greenhouse-Geisser adjustment). This indicates nonparallelism between placebo and antagonist across the ascending doses. Note that multiple pairwise contrasts at each dose give the uninformative outcome that only at a dose of 50 µg is p < 0.05 (table 3).

**Regression and Correlation**

Before addressing the details of regression analysis, I introduce the two models of regression. In model I regression, the x values are fixed by the design of the experiment, and only the y values are free to vary. In model II regression, both the y and the x values are free to vary.

The best-known example of model I regression analysis is simple, or ordinary least squares, linear regression analysis. In ordinary least squares regression, the line of best fit is arrived at by minimizing the sum of the squared deviations of the y values from the regression line. We associate regression with correlation and, in particular, with Pearson’s product-moment correlation coefficient (r).

The formula for this is:

\[
\frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{(x - \bar{x})^2(y - \bar{y})^2}}
\]

Note that the value of Pearson’s r takes into account the deviations of both x and y from their respective means. It is thus consistent with model II rather than model I regression. Almost everyone recognizes that r is an index of association and does not imply cause and effect. But correlation can also act as a measure of goodness of fit of the data to the estimated line of best fit. Thus, r² is the proportion of the total variance of the x and y values which is accounted for by the regression line. This also applies to multiple linear regression (equation 2) and to many nonlinear regression equations.

**Table 3.** Global outcome, and multiple pairwise contrasts by dose, from RM-ANOVA (see fig. 3)

<table>
<thead>
<tr>
<th>Dose, µg</th>
<th>p</th>
<th>p’</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.25</td>
<td>0.041</td>
<td>0.164</td>
</tr>
<tr>
<td>12.5</td>
<td>0.469</td>
<td>0.938</td>
</tr>
<tr>
<td>25</td>
<td>0.027</td>
<td>0.135</td>
</tr>
<tr>
<td>50</td>
<td>0.0012</td>
<td>0.0072</td>
</tr>
<tr>
<td>100</td>
<td>0.064</td>
<td>0.192</td>
</tr>
<tr>
<td>200</td>
<td>0.728</td>
<td>0.938</td>
</tr>
</tbody>
</table>

For parallelism of fall in blood pressure according to dose of placebo and antagonist: p = 0.0156 (Greenhouse-Geisser correction). Thus the profiles of placebo and antagonist over increasing dose can be regarded as different. Data from Ludbrook [15]. p = Raw two-sided p values for contrast of placebo versus antagonist; p’ = raw p values adjusted for 6 comparisons by Ryan-Holm step-down Bonferroni procedure.
In a data set such as that for figure 4, both y and x values are free to vary, so that model II regression analysis should be applied. There are many forms of this, but my preference is for what I have called ordinary least products (OLP) regression analysis. In this, it is the sum of the products of the deviations of the y and the x values from the regression line that is minimized (equation 6). This technique has been used in figure 4b.

One of the common abuses of correlation is in comparing two methods of measurement or two measurers. It is not only wrong but dangerous to use r for this purpose [24–26]. What we should look for is not agreement but disagreement, or bias. The simplest and most popular way of demonstrating the latter is the Altman-Bland method of differences [24]. This consists in plotting the difference between the two measurements (y axis) against the mean of the two measurements (x axis) (fig. 4a). If there is no bias, the mean difference should be indistinguishable from 0, and the slope of the regression of y on x should also be indistinguishable from 0. A difficulty with the Altman-Bland method is as follows. A little work with a pencil and paper should convince you that if the slope of the regression of differences on means differs significantly from 0 (implying proportional bias), it is almost inevitable that the mean of the differences also differs from 0, even though there is no fixed bias. The exception to this rule is if there is genuine fixed bias, in which case the mean of the differences may equal 0, or even deviate from 0 in the opposite direction. It is for these reasons that I prefer the more complicated and computer-intensive method of OLP linear regression analysis (fig. 4b), because this identifies and distinguishes between fixed and proportional bias whereas the Altman-Bland method does not [25, 26]. OLP regression also allows one method of measurement to be calibrated against another.

There is a powerful and useful technique for comparing the slopes and elevations of two or more straight lines, called analysis of covariance. There are accounts of this in standard statistical texts [20, 21], but these are difficult to follow. Feldman [27] gives a much more user-friendly account, with informative examples.

**Nonlinear Regression Analysis**

This is an enormous and complex subject, to which I cannot possibly do justice. It refers to fitting exponential, hyperbolic, parabolic, sigmoidal or other nonlinear curves to experimental data. All the statistics software packages listed in Appendix 2 will execute such analyses, usually by iterative (successive approximation) techniques. Usually, the user is required to enter ‘starting values’ – that is, guessimates of the regression coefficients. This keeps the number of iterations to a minimum. More importantly, it prevents the fitting process from going crazy and giving ridiculous values for the coefficients.

---

**Fig. 4.** Comparing two methods of measurement of systolic blood pressure (BP). a The Altman-Bland method of differences [24]. The regression line is difference = −6.9 − 0.174 (mean). The slope differs from 0 (p = 0.024). The mean difference (●) (−30.50 ± 1.76) differs from 0 (p < 0.001). b The OLP regression method [25, 26]. The OLP line of best fit is method B = −7.0 + 0.844 (method A). The 95% CI for the Y intercept (−26.0 to 12.0) includes 0, whereas the 95% CI for slope (0.720–0.969) does not include 1. Method A suggests (falsely) both fixed and proportional bias. Method B indicates only proportional bias.
This is a good point at which to mention the two reasons for fitting regressions to experimental data. One is empirical – the regression line serves the strictly utilitarian function of allowing one variable to be calibrated against the other. The other reason is mechanistic, or explanatory. Here the goal is to explain the form of the fitted regression line in terms of underlying biological processes. A straight line can serve an empirical function when observations have been made over a limited range. But if, say, the straight line is extended beyond the limits of the observations, it may seem to result in absurd predictions of, for instance, a negative blood pressure or a body temperature of over 150°C. In such cases the investigators should get together with the statistical consultant and agree on a mathematical function that respects the biological limits of the observed variable, and can perhaps give an insight into the biological processes that are at work.

I shall mention only one example of nonlinear regression. It is the construction of sigmoidal curves to describe stimulus-response relationships. This technique is sometimes known as four-parameter logistic regression analysis, and is applicable to any situation in which the stimulus-response relationship is, in effect, a cumulative normal distribution. It has found a special place in characterizing baroreceptor reflexes, in terms of the estimated upper and lower limits of the response and the maximal slope (gain). Clear descriptions of the technique are hard to find, but I refer readers to two accounts of applying it [28, 29].

The Analysis of Rates, Proportions, and Frequencies

This topic is inevitably associated with tables of frequency, the $\chi^2$ distribution and the $\chi^2$ statistic. The problem is that the $\chi^2$ distribution is a continuous one, whereas the categorical data being analyzed are discrete. So an innate flaw in $\chi^2$ tests is that they are only approximate, and when tables of frequency are small the approximation is often not a good one. For this reason, I believe that the approximate $\chi^2$ test should be discarded from the statistical repertoire, and replaced by exact tests. Fisher’s exact test for $2 \times 2$ tables of frequency was described over 80 years ago, is well known, and for small $2 \times 2$ tables can even be done by hand. But for larger and more complex tables, specialized statistics software must be used. The most comprehensive is the free-standing program StatXact (see Appendix 3), but there are modules for exact tests, based on StatXact, in statistics software programs such as SAS, SPSS, Stata and a forthcoming version of SYSTAT (see Appendix 3). Easily understood accounts of

The tests that will be discussed can be found in elementary texts [18, 30], and Agresti [31] gives detailed theoretical descriptions.

Tables of Independent Frequencies

The convention I shall use is to refer to the number of rows in a given table as $r$, and the number of columns as $c$. In $2 \times 2$ and $2 \times c$ tables, the null hypothesis is (as an approximation) that the proportions in the table are equal. More specific exact tests can be constructed for $2 \times 2$ tables, in which the null hypothesis is that the odds ratio = 1, or that the relative risk = 1. In $r \times c$ tables, the null hypothesis is rather vague. It can be stated as a lack of interaction between rows and columns, or – as Karl Pearson, inventor of the $\chi^2$ test, put it – goodness of fit (of observed to expected frequencies).

Tables of Related Frequencies

In the case that a $2 \times 2$ table represents the proportions in a group, before and after some intervention, the exact McNemar test is appropriate.

Tests for Trends

Thus far I have considered only tables in which the variables are nominal – for instance, alive/dead, red/white/blue, and so forth. However, sometimes the variables can take the form of ordered categories – for instance, the achievements of schoolchildren as minimal, partial, satisfactory, good or excellent. In such instances,
a test for trends is appropriate. In table 4, there are two
2 × c tables of data constructed before and after an in-
tervention. The question is whether the trend in propor-
tions is altered by the intervention. The exact Cochran-
Armitage test for linear trends is the appropriate one to
use in these circumstances.

Binomial Logistic Regression Analysis
A detailed description of this is given by Hosmer and
Lemeshow [32]. The logistic regression equation takes the
following form:

\[ \text{Logit } Y = a + b_1x_1 + b_2x_2 + \ldots + b_nx_n \]  

(7)

where \( a \) and \( b \) are regression coefficients. \( Y \) is a binomial
categorical variable that indicates the occurrence or not
of an event such as dead/alive. Logit \( Y \) can be rewritten
\( \ln (\text{odds of } Y) \), where \( \ln \) is the natural logarithm. The
independent (x) variables can be dichotomous (male/
female), unordered polychotomous (specific hospitals),
or ordered polychotomous (age groups). The x variables
can even be continuous (age in years), and the logistic re-
gression can include interactions as well as main effects.

The usual approach is to include all independent, pre-
pdictive, variables of interest, then perform a stepwise re-
gression analysis. In this, the independent variables for
which \( p > 0.05 \) are successively eliminated, to arrive at a
final, best, model. Though a program for performing lo-
gistic regression analysis in an exact fashion is available
(LogXact, Appendix 3), for most purposes the routines
provided in general purpose software are adequate (see
Appendix 2). Logistic regression analysis is not a job for
amateurs – the help of a biostatistician is essential. What
is more, it has been described as more an art than a sci-
ence!

One of the really useful inferences one can make from
binomial logistic regression analysis is to be able to pre-
dict, for an individual patient, the risk (odds) of an event
such as death [32].

Survival Analysis
Analysis of survival is an essential component of most
prospective clinical trials, though it can be used in any
study in which a dichotomous event (yes/no) can occur at
some time after entry to the study. There is a good, all-
round, account of clinical trials which includes simple
descriptions of techniques of survival analysis [10], and
an in-depth account of survival analysis suitable for stat-
isticians [11]. I can do little more here than give a simple
description, and the nomenclature, of some of the tech-
niques.

Table 5. Steps in preparing a manuscript: statistical methods and
presentation of results of statistical analyses

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
</table>
| 1 | Write the statistical methods section giving the methods used,
|   | why they were chosen, and what to expect from them – and in
|   | language that biomedical readers will understand |
| 2 | State what statistics program you used, version, vendor, etc.: |
|   | for instance, SPSS v. 14 (SPSS Inc., Chicago Ill., USA) |
| 3 | Assemble your numerical results and outcomes of statistical
|   | analyses into provisional tables and figures |
| 4 | Define symbols – for instance: n (group size), p (probability of
|   | null hypothesis), 95% CI, SE, s (standard deviation), r (product-
|   | moment correlation coefficient) (see Appendix 1) |
| 5 | Decide on statistical notation and abbreviations: P or p (some
|   | UK journals prefer p)? Mean ± standard error (plus n), or
|   | mean ± standard deviation or 95% CI? Define the cut-off for
|   | statistical significance – conventionally two-sided \( p \leq 0.05 \) |
| 6 | Whenever possible, put your \( p \) values in the tables (or, some-
|   | times, figures or their legends), rather than cluttering up the
|   | text; give actual \( p \) values to a sensible number of decimal plac-
|   | es (for instance, \( p = 0.13 \), or \( p < 0.001 \), or \( p \) always
|   | >0.07); do not give \( p < 0.05 \) or NS (not significant) |
| 7 | Acknowledge the assistance of your biostatistician – this may
|   | ward off criticisms from reviewers! |

The Kaplan-Meier technique is not a form of statistical
analysis, but merely a method for constructing a survival
curve graphically. The Mantel-Haenszel log rank tech-
nique tests whether the survival experience under two
sets of conditions (for instance, placebo or active treat-
ment) is indistinguishable. The Cox proportional haz-
ards regression technique allows comparison of survival
experience for more than two treatments (for instance,
placebo or standard treatment or new treatment).

Preparing the Manuscript

I refer here only to statistical aspects of the manu-
script, because individual journals issue instructions to
contributors about the format in which manuscripts
should be submitted. But these instructions do not always
extend to statistical matters. There are two bodies that do
give this extension: the International Committee of Med-
ical Journal Editors (ICMJE), which is concerned chiefly
with general medical journals [33], and the Council of
Science Editors (CSE) [4]. The CSE has embraced the re-
commendations of the American Physiological Society [3]
regarding the presentation of statistics and statistical
analyses. With some exceptions to which I shall refer, I suggest that these should be followed.

Some of the important matters are summarized in table 5 and discussed below. Mean values should always be accompanied by n, the number of observations on which the mean is based, and with an indication of the uncertainty that surrounds the mean. This can be either the SE, or the 95% CI. Both indicate the uncertainty that the sample mean coincides with the mean of the parent population. Others favor the standard deviation (s), which is a measure of the scatter of values in the parent population. In figure 5, I suggest how best to present graphically the results of a study.

Finally, I return to the place of biostatisticians. It is wise to formally acknowledge their assistance, because it can preempt critical comments from reviewers. Furthermore, investigators are likely to learn far more about statistics from this than from reading articles such as this one.

**Acknowledgements**

I thank all those who have, over the past 50 years, helped me towards an understanding of biostatistics. The comments of the three reviewers of the manuscript have been of great assistance. Dr. Roger Evans helped me construct the figures.

**Appendix 1:**

**Glossary of Statistical Symbols and Terms**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>number of units in a group (sample)</td>
</tr>
<tr>
<td>$\bar{x}$, $\bar{y}$</td>
<td>group (sample) mean</td>
</tr>
<tr>
<td>$\hat{R}$</td>
<td>group mean rank</td>
</tr>
<tr>
<td>SE</td>
<td>standard error of the mean</td>
</tr>
<tr>
<td>s</td>
<td>group (sample) standard deviation</td>
</tr>
<tr>
<td>$s^2$</td>
<td>group (sample) variance</td>
</tr>
<tr>
<td>ln</td>
<td>natural logarithm</td>
</tr>
<tr>
<td>log10</td>
<td>logarithm to base 10</td>
</tr>
<tr>
<td>t</td>
<td>Student’s t statistic</td>
</tr>
<tr>
<td>F</td>
<td>F statistic (after R.A. Fisher)</td>
</tr>
<tr>
<td>a, b</td>
<td>regression coefficients</td>
</tr>
<tr>
<td>r</td>
<td>Pearson’s product-moment correlation coefficient</td>
</tr>
<tr>
<td>$\Sigma$</td>
<td>the sum of</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>chi-squared statistic</td>
</tr>
<tr>
<td>r, c</td>
<td>rows and columns in frequency tables</td>
</tr>
<tr>
<td>type I error</td>
<td>false rejection of the null hypothesis</td>
</tr>
<tr>
<td>type II error</td>
<td>false acceptance of the null hypothesis</td>
</tr>
<tr>
<td>p</td>
<td>probability of type I or type II error</td>
</tr>
<tr>
<td>$p'$</td>
<td>adjusted value of p</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>probability of type I error (falsely rejecting the null hypothesis)</td>
</tr>
<tr>
<td>$\beta$</td>
<td>probability of type II error (falsely accepting the null hypothesis)</td>
</tr>
<tr>
<td>$1 - \beta$</td>
<td>power to reject the null hypothesis</td>
</tr>
</tbody>
</table>
Appendix 2: General Purpose Statistical Software Packages

NCSS (NCSS, Kaysville, Utah, USA)
SAS (SAS Institute Inc., Cary, N.C., USA)
SigmaStat (Systat Software Inc., San Jose, Calif., USA)
SPSS (SPSS Inc., Chicago, Ill., USA)
Stata (Stata Corporation, College Station, Tex., USA)
SYSTAT (Systat Software Inc., San Jose, Calif., USA)

References