1

Statistics and genetic toxicology – setting the scene

D.P. LOVELL

1.1 INTRODUCTION

The aim of this chapter is to explain in non-mathematical terms the strengths and limitations of aspects of the statistical analyses that can be applied to genotoxicity studies. First, to improve the understanding of subsequent chapters and statistical ideas in general; second, to try to remove the mystique surrounding aspects of statistics and thereby improve the dialogue between the individuals involved in conducting and assessing genetic toxicology tests.

There is a certain irony in the need for a book of this type at all. Much of current statistical thought developed from the work of researchers interested in the field of genetics. In particular, Sir Ronald Fisher made an enormous contribution to the field of what can now be termed biometrics. Many of the statistical techniques now in common use trace back to Fisher's studies on genetic variation and mutation in a range of species. In fact, one of Fisher's major contributions, the development of the analysis of variance, resulted from his unification in 1918 of the previously independent ideas and warring factions of Mendelian Geneticists and Biometricians (Fisher, 1918).

Fisher was, at the the same time, a geneticist, mathematician, statistician and experimentalist. Such expertise in the same person is rare. In genetic toxicology there is now a division of labour with specialists from different disciplines collaborating. This increased specialism has to some extent broken the statistical links with pioneers in the field.

1.2 WHAT IS STATISTICS FOR?

There are different schools of thought on whether statistical methods should be aimed at testing hypotheses (is the response of the treated group greater than that in the control group?) or estimation (how big is the difference and how confident are we of the difference?). Increasingly, the importance of estimation is being recognised and biomedical statisticians are insisting that statistical practices should reflect this (Gardner & Altman, 1986; Bulpitt, 1987).

An alternative approach is to use statistical methods for exploring and summarising the data. The move away from formal hypothesis testing and significance levels has developed over the last few years following on from the work of John Tukey (1977) in the area called Exploratory Data Analysis. Chatfield (1985) has suggested another somewhat similar pragmatic approach to data analysis called Introductory Data Analysis (IDA).

The increasing use of statistics as a method to explore data sets and to provide an improved summary of the results of an experiment whether by significance tests or exploratory data analysis is more relevant to how genotoxicity assays are actually assessed. The judicious use of statistics as an aid to determining the implications of results from genotoxicity assays is more important than an over-reliance on hypothesis testing methodology.

At times, discussions of statistical philosophy may confirm some biologists' prejudices about how statistics is irrelevant to the work they are doing. It is easy for the non-statistician to become confused and dispirited at the apparent intricacy of statistics and to dismiss its terminology and formality. However, statistics is a complex intellectual activity with roots in mathematics, logic and philosophy as well as the more mundane practice of data processing.

Statistical formulae therefore use mathematical symbols as a concise and precise method of specifying complex ideas. An example which is widely used in other chapters is the symbol $\sum$ (the Greek capital letter sigma) to designate the sum of a set of values.

The full symbol $\sum_{i=1}^{n} X_i$ is a succinct way of saying add all n values of a set of numbers designated $X_1, X_2, \ldots, X_n$. In fact, few of the practical statistical methods described in this book require mathematical skills beyond addition, subtraction, division and multiplication. Most methods are now either programmable for a computer or already exist as programs in software packages (see Appendix 2).

1.3 IS THERE A CORRECT STATISTICAL ANALYSIS?

Non-statisticians may well demand the 'best' approach or the 'most suitable' test. They become confused by the apparent competing claims for different approaches or the seemingly academic disputes over the various methods. Such a response is understandable. Concepts in statistics are still controversial issues raising fundamental questions. There is no single correct statistical method, but, instead, there are different schools of thought with alternative approaches.

An example of the continuing debate is how to analyse a 2 × 2 table such as the comparison between the proportion responding in control and treated groups – a common design in genetic toxicology. It is over 50 years since Frank Yates first described an analysis for this problem and yet the approach is still contentious. In 1984, on the fiftieth anniversary of Yates' original paper, the Royal Statistical Society held a meeting to commemorate its publication (Yates, 1984). Yates' lecture and the ensuing comments and correspondence takes up nearly 40 pages of the Royal Statistical Society's Journal!

There are certainly correct analyses of the various experimental designs used but there is probably no single correct analysis. In the case of the assays discussed in this book, it is possible to suggest statistical methods that work, i.e. give sensible answers. When more than one test gives sensible answers it may be difficult to decide which, if any, is 'correct'. Alternative analyses may well exist and the use of these methods should not necessarily be discouraged. One good reason is that often little is known of the comparative performance of different statistical tests on real data.

Another reason for not choosing a single 'correct' analysis is that many of the apparently different approaches are often the same statistical analysis in another guise. For instance, the tables of χ^2, t, F and the normal deviate (z) (see list of abbreviations for a description of these terms) found in the back of most introductory statistical text books are interrelated. All the other tables can, in fact, be shown to be special cases of the general F table. It is important, therefore, not to select one method and reject another simply because its terminology or formulation is different. For example, χ^2 tests on proportions can be expressed as tests based upon the normal deviate (z) while the two sample t-test is equivalent to a one-way analysis of variance with two groups.

Many statistical ideas can be related to an underlying general statistical model. An analogy is a continent covered by water with just the mountain tops appearing above the water as islands. Statistical tests like the chi-square and the analysis of variance which may appear independent are in fact inter-connected. The unification of statistical methods and models is an area of considerable intellectual effort which is now meeting with some success (McCullough & Nelder, 1983).

Statistical methods which are recommended in this book should not be taken to be '*the method*' to the exclusion of all other approaches. The methods included in this book may even be idiosynchratic or special cases of a more relevant general technique chosen because of familiarity with a particular approach. Computers and statistical packages are becoming more powerful and more widely accessible so that new statistical methods which will be more appropriate could replace today's optimum approaches.

It is appropriate to warn, however, that even with computers incorrect analyses can be carried out. There is no limit to people's ingenuity in the misuse of statistical packages. It is not possible to warn against all such types of errors in detail. However, some of the incorrect approaches based upon misunderstandings about statistical ideas are included in the discussion on each genotoxicity assay. The main way to guard against the use of incorrect methods is to maintain a familiarity with the data being analysed and by a healthy scepticism of computer printouts!

The aim of the individual chapters is to provide the statistical methods necessary for the genetic toxicologist to analyse the results of assays. Most of the methods are in fact straightforward and relatively easy to apply but difficulties may arise. If so, stop, and consult a statistician. The explanations in the chapters should provide sufficient information for a dialogue to take place between the genetic toxicologist and the statistician.

1.4 EXPERIMENTAL DESIGN

The experimental designs used in genotoxicity testing are not very different from those used in other areas of toxicology. The organisms may differ – bacteria, mice or mammalian cells; the end-points may vary – point mutations, chromosomal abberrations or abnormal foetuses – but the underlying experimental concepts are similar. A standard assay usually includes a negative control and a treated or experimental group as well as the inclusion of a positive control (where a treatment known to produce an effect is administered). There may also be a dose-response experiment to investigate the relationship of the end-point to variation in the doses. The experimental designs developed for genotoxicity assays do not differ appreciably from those developed by Finney and co-workers in the 1950s and 1960s for biological assays in general (Finney, 1978). Certainly, some of the statistical methods used in biological assay could be applied to the analysis of data from mutation experiments.

Statistical methods have been developed to try to detect effects of treatments above the 'noise' of biological variability. Variation can be present at any of a number of levels depending upon the mutagenicity assay; for instance, between cells, between plates or cultures, between animals, and between observers. Experimental designs are methods for taking this variation into account, estimating it and minimising its effect on the factors or the treatments which are the main purpose of the study. Choosing sample sizes, types of samples, doses, all form part of determining the experimental design. The presence or absence of variability at different levels of the design can influence how acceptable it is to pool data from different units such as plates or animals. The choice of the *experimental unit* in the analysis is important because variation between cultures in a chromosome aberration study or between males in a dominant lethal study may affect the significance testing of hypotheses.

In many of the more sophisticated statistical analyses the experimental design is a fundamental aspect of the analysis determining both the nature and the power of the statistical tests to be used. Identification and incorporation of potential sources of variation into the experimental design as a constituent part of the study can subsequently provide considerable insight into how successfully the objectives of the experiment were achieved.

1.5 STATISTICAL ISSUES RELEVANT TO ALL MUTAGENICITY ASSAYS

Some of the experimental designs permit a number of possible methods for analysis. Examples of the range of methods actually in use will be found in the subsequent chapters; however, a number of statistical issues (Table 1.1) need to be discussed which are relevant to an appreciation of statistical methods in general.

1.5.1 Probability and hypothesis testing

Probability is a very practical concept understandable to any gambler, whilst at the same time being a deeply philosophical statistical concept. An understanding of how often events occur in the 'long run' is a sufficient start for the interpretation of statistical tests. Some events are common although their actual occurrence at a particular point is uncertain; others are rare. Probability is measured on a scale from 0 to 1 where 0 means an event will never happen and 1 means that it is certain to.

One aspect of a statistical analysis is to provide some indication of whether the results obtained were compatible with some hypothesised effect. In fact a hypothesis, called the *null hypothesis*, is set up that suggests, for instance, that a particular treatment has no effect. The null hypothesis is deliberately created to be challenged and possibly rejected in favour of an *alternative hypothesis* that the treatment does have some effect. (Underlying this somewhat artificial formulation of the purpose of an experiment is

Table 1.1. *Statistical issues of relevance to all mutagenicity assays*

1.5.1	*Probability and hypothesis testing* Null hypothesis, alternative hypothesis, test statistics (F, t, χ^2, z), critical values, significance levels. Statistical significance is *not* a measure of the size of an effect.
1.5.2	*One- or two-sided significance tests* One- or two-sided test, one- or two-tailed statistical distributions, statistical tables of distributions.
1.5.3	*The power of a statistical test* Neyman-Pearson theory of hypothesis testing, false positives, false negatives. Prediction, sensitivity, specificity, accuracy, prevalence, Type I (α) error, Type II (β) error, power ($1-\beta$).
1.5.4	*Biological and statistical significance* Hypothesis testing, estimation, exploration of data. Statistical tests on positive controls. 'Statistical sanctification', high dose effects, thresholds, monotonic dose-response, statistical significance and/or biological importance, correlation.
1.5.5	*What to do with borderline significance levels* Weak effects or Type I errors, repetition of borderline experiments, pooling results, combination of probability values, sample size and power.
1.5.6	*Replication and randomisation* Replication of and randomisation of experimental units.
1.5.7	*Distribution of data* Parametric tests, normal distributions, 'robustness'.
1.5.8	*Transformations* Counts, Poisson, square root transformation; proportions, binomial, arcsine or angular transformation; logarithmic transformation, scales, back-transformation.
1.5.9	*Parametric v non-parametric tests* Parametric methods: Analysis of variance, t-test, linear regression, assume normal distributions, equal variances, independent data. Non-parametric methods, distribution free tests, relative rankings.
1.5.10	*Outliers* Artefacts, blunders, technical errors.
1.5.11	*Testing comparisons between means* Planned, *a priori*; unplanned, *a posteriori*; multiple comparisons, orthogonal comparisons. Dunnett's test, Bonferroni correction.

the concept that while a hypothesis can be disproved it can never be proved.)

The decision whether or not to reject a null hypothesis is based upon comparing the test statistic (e.g. F, t, χ^2, etc.) obtained from the statistical test with tables of *critical values* for that particular test. The critical values are those associated with particular *significance levels.*

These significance levels provide a probabilistic statement that the results observed occurred by chance alone when the null hypothesis is true. For historical reasons based upon the availability of suitable tables a number of critical values have become common. These are based upon probabilities of a by-chance-alone effect of one in 20, one in 100 or one in a 1000 experiments. These are sometimes expressed as percentages or as probabilities. Various ways of representing these effects are shown in Table 1.2.

There is no compulsion to be restricted to these critical values and increasingly statistical software packages are printing the actual probability levels associated with the statistical test. Many statisticians prefer to report these values arguing that is is misleading to draw arbitrary lines such that a probability level of 0.051 is non-significant while a level of 0.049 is significant.

It is important to appreciate that the significance level is not an absolute measure of the size of the difference between, say, two group means. It is a function of the experimental design and the statistical tests used.

Rare events which have small but finite probabilities do occur. Therefore an extremely small probability may represent a real effect of a treatment but it might also mean that this is a rare chance effect. This is an important consideration when many comparisons are being made. If, for instance, all the possible comparisons between 100 means are carried out, there would be 4950 comparisons. If many comparisons are made, or many experiments are carried out, rare chance effects will eventually be detected. In fact, the choice of specific significance levels actually defines the number of events that occur by chance. The problem of multiple comparisons will be discussed in more detail below.

1.5.2 One- or two-sided significance tests

The non-statistician is often confused by the problem of whether a particular statistical method involves a one- or two-sided significance test. The agonies of choosing which test can be avoided if certain key ideas are thought through. The core of a statistical test is the Null Hypothesis that a treatment has no effect. At the same time an alternative hypothesis is also proposed. This can either be that the experimenter is interested in a difference between the control and the treated (the difference can be in either direction and the statistical test is then two-sided) or that the experimenter is only interested in treatments which alter the incidence in one direction by increasing, for instance, the incidence of chromosomal aberrations (in this case the test is one-sided).

The actual choice of critical values for determining significance levels for testing the null hypothesis is determined from statistical tables and needs some care as the tables appear in many forms. The one-sided test is a less conservative test than the two-sided in that smaller effects will be

Table 1.2. *Different methods of showing statistically significant results*

Statistically significant			Expressed as				
percentages	by proportions	ratios	probability less than	or	probability between	Expressed as stars[a]	Expressed as words[b]
5%	0.05	1/20	$P < 0.05$		$0.01 < P < 0.05$	*	Significant
1%	0.01	1/100	$P < 0.01$		$0.001 < P < 0.01$	**	Highly significant
0.1%	0.001	1/1000	$P < 0.001$		$P < 0.001$	***	Very highly significant

Notes:
[a] Sokal & Rohlf (1969) p.169.
[b] Sprent (1981) p.44.

considered to be significant. A one-sided test is therefore appropriate when the direction of the effect is of interest: the treatment may or may not increase the damage or number of mutations compared with the control and whether or not it reduces the incidence below the control level is not of concern. (It is, however, important to check that incidences are not reduced in treated groups as such a finding might indicate problems with the experimental procedures such as high dose toxicity. This is an example of using statistics for exploring patterns in the data rather than for formal and precise significance testing.)

In general, one-sided statistical tests should be used in genotoxicity assays. (Conceivably, in a research project, an unknown compound might be either a mutagen or an anti-mutagen protecting against genetic damage, and experiments designed to determine which.) Care is needed to ensure that the correct sided test is in fact carried out especially when using statistical packages. Some tests such as the analysis of variance and testing for linear trends are routinely reported by many statistical packages as two-sided tests. Such tests will underestimate the statistical significance of any treatment related effects.

A further complication is that it is not always apparent from a statistical table whether the associated probability levels are for one- or two-sided tests. For instance, the table of critical values for the t distributions given by Snedecor & Cochran (1967, Table A4) are for two-sided tests while those given by Winer (1971, Table C2) are for a one-sided test. Also, tables such as those of the F and χ^2 distributions give the probabilities associated with *one tail* of the distribution but which are appropriate for a *two-sided* test. Figure 1.1 shows the appropriate tails of the distribution to use for either a one- or two-sided test. Mather (1964, Section 18, p. 46) provides a more technical but helpful description of the relationship between the various distributions.

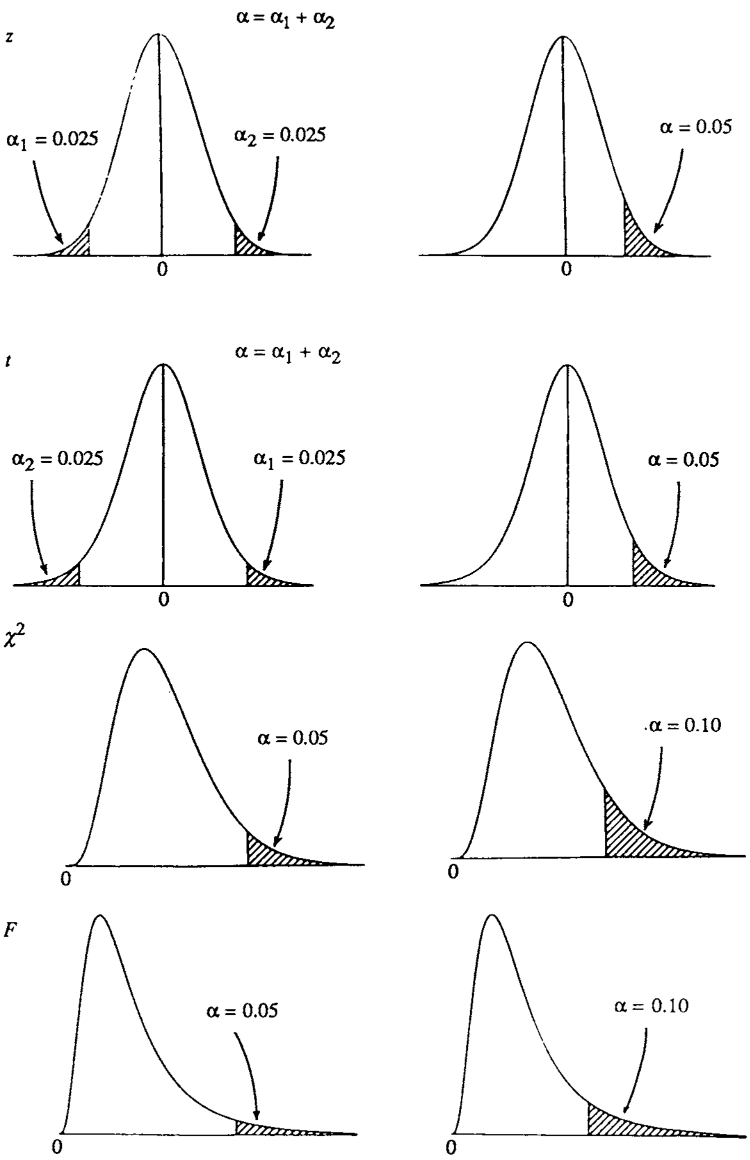
The symbol χ^2 can be confusing because it seems to appear in two forms: either the Greek letter chi (χ) or the Roman X. This is because the Greek and Roman letters are used in general to distinguish between a theoretical value and an observed value. The *theoretical* chi-square distribution is denoted by χ^2 while the *calculated* chi-square statistic is referred to as X^2. A similar distinction is made between the theoretical mean and standard deviation of a population called μ and σ (Greek letters mu and sigma) and the observed values obtained in an experiment designated $\overline{X}$ and s.

1.5.3 The power of a statistical test

The formulation of a statistical analysis in the form of a test of a hypothesis (called the Neyman-Pearson Theory of hypothesis testing after

Fig. 1.1. **Frequency curves of z, t, χ^2 and F distributions showing the respective tails of the distributions used for taking critical values for two- and one-sided tests at the $P < 0.05$ criterion.**

	Two-sided test	One-sided test
Null (H_0) hypothesis	$\mu_C = \mu_T$	$\mu_C = \mu_T$
Alternative (H_1) hypothesis	$\mu_C \neq \mu_T$	$\mu_T > \mu_C$


z
α = α₁ + α₂
α₁ = 0.025
α₂ = 0.025
α = 0.05
t
α = α₁ + α₂
α₂ = 0.025
α₁ = 0.025
α = 0.05
χ²
α = 0.05
α = 0.10
F
α = 0.05
α = 0.10