Environmental Health Perspectives Supplements
Vol. 102 Suppl 1:43–47 1994

Statistical Evaluation of Mutagenicity Test Data: Recommendations of the U.K. Environmental Mutagen Society

by David J. Kirkland

Most of the many published guidelines on how to conduct mutagenicity tests do not give advice or references on statistical analysis of data. The U.K. Environmental Mutagen Society decided to address this omission, and in 1985 established 8 working groups comprising genetic toxicologists from all sectors of the science, plus at least 2 statisticians per group, to produce statistics guidelines on 10 different test systems. Each group gave advice on how to determine the suitability of data for distribution fitting, when data are unsuitable, when and how data should be transformed, which statistical tests are suitable for a given set of data, which factors govern the choice of statistical test, an order of preference, and some worked examples using real data. In addition, groups gave advice on statistical issues in the design of experiments. Strong recommendations were made that for *in vitro* tests, sufficient cells be treated and sampled to provide meaningful values of spontaneous mutant/aberration frequencies, for genuine, independent replicate treatments to be used, and that the acceptability of an experiment should be based on homogeneity between replicates as well as comparison of negative and positive control responses with historical ranges. It was recommended that most *in vitro* studies should include independent repeat experiments, and advice was given on how to check for consistency between experiments and then combine data for further significance testing. For *in vivo* tests, it was generally believed that increasing the number of dose levels and reducing the number of animals per dose impoves statistical sensitivity; there was some uncertainty about how to handle data when heterogeneity was found within a group of animals, but there was a consensus that statistical tests and interpretation of the biological findings should proceed.

Introduction

Many guidelines for mutagenicity testing [e.g., Organisation for Economic Cooperation and Development (*1–3*). European Economic Community (EEC) (*4*), and U.K. Environmental Mutagen Society (UKEMS) (*5,6*)] that have made recommendations on methods for generation of data have not made similar recommendations on analysis of data. Statements such as "data should be analysed using appropriate statistical methods" are common, without referring the reader to any useful publication. Interestingly, it appears the Japanese Ministry of Health and Welfare has not requested in their guideline (*7*) that mutagenicity data be analysed statistically. UKEMS decided that, having published two guideline books on how to generate data (*5,6*), it should attempt to prepare a similar guideline on statistical analysis of data.

Organization

As in the past, UKEMS decided recommendations should be achieved by consensus, rather than being those of an individual, and for each topic a working group consisting of five to nine members was convened. Each group was chaired by a genetic toxicologist with recognized experience in the topic area, and each group included at least two statisticians. Working group members were selected to represent all sectors of the science, namely, academic, industrial, and contract laboratory genetic toxicologists and statisticians. A steering group (effectively a subcommittee of UKEMS) was established to oversee the exercise, and comprised seven genetic toxicologists and three statisticians representing the same scientific sectors as above, but also including representatives of U.K. regulatory authorities.

Aims and objectives

Ten different mutagenicity test systems were selected for assessment and grouped into eight topics. For each topic, the working group was required to consider: *a*) how to determine the suitability of data obtained from an assay for fitting a distribution, when the data are unsuitable, when and how data should be transformed; *b*) the types of statistical analyses that can be used with the assay data under consideration, which, if any, factors govern the choice of analysis, an order of preference if several types of analysis may be used; and *c*) some examples using real data to help the reader understand the above. In addition, working groups were asked to consider the statistical implications of experimental design, and to make recommendations where appropriate.

Finally, with the exception of general principles that would be presented in an introductory chapter, and could be referred to in any of the individual reports, each report was to be written

Hazleton Microtest, Otley Road, Harrogate, HG3, 1 PY, UK.
This paper was presented at the International Biostatistics Conference on the Study of Toxicology that was held May 13–25, 1991, in Tokyo, Japan.

to stand alone. The reason for this dates back to the original UKEMS reports, which were rather like handbooks, and an experimenter interested in one topic would find all the relevant information in a single chapter. It was recognized that this approach with a statistics book could lead to some repetition or even some contradictions. UKEMS was prepared to accept repetition in return for the benefits of providing integral chapters. Contradictions would be avoided by following a previously established procedure, namely, *a*) working groups discuss objectives, chairman (or statistician) drafts manuscript, other group members comment; *b*) corrected manuscript reviewed by steering group members individually, comments collated on one copy, returned to working group chairman; *c*) steering group and all working group chairmen meet to see if all comments can be accommodated, particularly aiming to remove inconsistencies between different manuscripts.

Recommendations

All of the working groups associated with *in vitro* mutagenicity techniques made strong recommendations for treating and sampling sufficient cells to provide meaningful values for spontaneous mutant/aberration frequencies. All recommended the use of genuinely independent replicate treatments in all cases (minimum of three for bacterial colony assays, mimimum of two for all other *in vitro* assays). They also recommended that judgment of the acceptability of an experiment should be based on two factors: comparison of negative and positive control values with some appropriate historical range, and a measure of heterogeneity/dispersion between replicate cultures. For all *in vitro* tests except chromosomal aberration, strong recommendations were made for experiments to be repeated at least once. In many cases, advice was given on how to check for consistency between experiments, how to combine data from separate but consistent experiments, and how to perform further significance tests. If consistency was not obtained, additional experiments were recommended.

The working groups concerned with *in vivo* tests in mammals also made some recommendations regarding study design. In general it was felt that increasing the number of doses and reducing the number of animals per dose improved the statistical sensitivity, but checks for heterogeneity between animals should be made. Other specific recommendations of the working groups are summarized below.

Microbial Colony Assays (Ames Test)

In the most widely used assays, 10^7–10^8 *Salmonella* or *E. coli* bacteria with a nutritional mutation are treated with test chemical, with and without exogenous metabolism, plated in agar with a trace amount of the required amino acid, and incubated for 2–3 days. Only bacteria that have fully reverted to independence can grow after the trace of amino acid has been exhausted, and they produce discrete colonies. Spontaneous mutation rates lead to 5–200 colonies/plate (depending on strains), and increases in numbers of colonies are indicative of a mutagenic effect. Some authors have reported Ames colony counts to be distributed according to Poisson statistics (*8*) and others have reported them to be more variable than would be expected from the Poisson (*9,10*). Our authors therefore recommended sample variation be first determined by dividing the χ^2 value for the data set by its degrees of freedom to give the m statistic. If the m-statistic is ≤ 1, then significance tests which assume the Poisson distribution can be used.

If m lies between 1 and, say, twice the average historical value in the laboratory, then a method allowing variation greater than Poisson, should be used. If m exceeds twice the laboratory average, then the experiment should be discarded.

As Ames test data are often not Poisson, statistical significance methods based on observed variance are perhaps the most logical. Of the parametric methods that allow for multiple comparisons, Dunnett's t-test (*11,12*) is preferred. Various regression methods were recommended for looking at dose response, the type of regression depending on the choice of transformation or weighting, and whether any downturn in the response curve is excluded or modeled. Of the nonparametric methods, Wahrendorf's ranking method (*13*) was preferred.

Computer simulations were used to compare the sensitivity of these three methods, using untransformed data and data transformed by various methods. The following conclusions were reached: *a*) untransformed data yielded more significant effects than data transformed by square root, inverse hyperbolic sine, or log transformations; *b*) linear regression and Wahrendorf's method were more powerful than Dunnett's test, particularly when colony counts were small and highly variable; *c*) toxicity-induced reduction in colony counts at high doses affected the power of linear regression and Wahrendorf's methods much more than it affected Dunnett's test.

Mammalian Cell Gene Mutation Colony Assays

In gene mutation colony tests, cells that are normally sensitive to a poison are examined for resistance to its toxic effects after treatment. Colonies may be selected in agar (similar to the Ames test) or in liquid medium, in which case they grow as discrete colonies on a plastic surface and are usually stained to visualize them. Spontaneous mutant frequencies depend on the cell type and genetic locus examined, but are generally higher than in Ames bacteria and range from 4×10^{-7} to 1×10^{-4}. To avoid zero mutant counts on control or treated plates, large numbers of cells must be plated, and technical restrictions can make this impractical for some systems. The working group therefore recommended suspension rather than monolayer cultures, and genetic loci with high spontaneous mutant frequencies. Thus, *TK* mutation in mouse lymphoma cells becomes the method of choice on statistical grounds.

Even with this system, the assays are so large that there are usually insufficient genuinely independent replicate observations to permit the use of nonparametric methods. Of the more powerful parametric methods, the authors preferred weighted regression to transformation of the data because it allows a test for a direct relationship between mutant frequency and dose. Various forms of weighting may be used, but Poisson-derived weights are simpler to calculate, and may give more realistic weighting when plates have been lost; and there was no firm evidence to suggest the counts were not Poisson distributed.

The group recommended the first stage of analysis should be analysis of variance on weighted mutant frequencies to determine if differences between groups were greater than between

replicates. The second stage should then be a t-test to examine significance at a test dose compared with control. The recommended procedure uses estimates of between-replicates residual error mean square and weights to obtain estimates of variance and standard error of mean mutant frequency at each dose.

Finally, dose response could be examined by performing a test for linear trend within the analysis of variance table, the slope of the straight line being

$$\frac{\text{cross product (mutant frequency} \times \text{dose).}}{\text{doses sum of squares}}$$

Bacterial/Mammalian Cell Fluctuation Tests

Bacterial and mammalian cell fluctuation tests are similar in principle to the colony tests discussed above, except the selection of mutants does not take place in agar or on large plastic dishes, but the population of cells/bacteria is divided up into many small wells, and numbers of empty wells are counted instead of colonies.

If one looks at replicate 96-well plates from the same culture, then the proportion of empty wells varies binomially. However, extra variability is found when one observes plates from different cultures, particularly after prolonged subculturing (*14*). The authors compared observed variances from several experiments with theoretical binomial variances and found the ratio to be fairly constant. The groups believed this was a good measure, therefore, of variability within an experiment and decided to call the ratio the heterogeneity factor after its comparable use in biological assays (*15*). Although the ratio can be estimated from a single experiment, the authors recommended each new ratio from a new experiment be used to update (say, in the proportions of 19:1) a historical value and then be compared with it. According to the F-distribution, heterogenicity factors exceeding the updated historical value more than 10.8-fold would be extremely rare (0.1% one-sided, with 1 and infinite degrees of freedom), and cultures with such values should be excluded.

For mammalian cell tests, analysis using two types of statistical test was recommended, first to compare the weighted-mean log mutant frequency from each treatment with control, and second to check for linear trend by weighted regression. Both methods use the heterogeneity factor described above to obtain a modified estimate of variance.

For bacterial tests, there is continuous incubation with the test compound and no subculturing to introduce additional error as in the mammalian cell version. Variation between replicate trays appears to be binomially distributed, and therefore direct comparisons between treated and control cultures using 2×2 tables can be made. Values of χ^2 can then be compared with Dunnett's values for multiple comparisons. There is no separate measure of viability after treatment in the bacterial fluctuation test, and so dose response should be assessed using a test for isotonic trend, which is much more robust in the presence of toxicity (*16*).

In Vitro Cytogenetic Assays

Cytogenetic tests examine induction of gross chromosomal damage in metaphase preparations of rodent or human cells at appropriate times after treatment. After much discussion, and perhaps controversy, the authors decided the cell, and not the chromosome, was the experimental unit.

They thus recommended the data be classified into two basic groups: normal or aberrant cells. As the biological consequences of small discontinuities (gaps) in the chromosome are uncertain, the aberrant cells are usually classified as including or excluding gaps, but most conclusions are drawn on proportions of aberrant cells excluding gaps.

In an acceptable assay, it is assumed that the variability between cells sampled from different cultures is no greater than that between cells sampled from the same culture. It is therefore recommended that acceptable homogeneity be checked using the binomial dispersion test.

Assuming acceptable homogeneity, recommended data from replicates are combined, giving an overall proportion of aberrant cells for each treatment or negative control; the proportions at each treatment are compared with the control using Fisher's exact test.

Sister Chromatid Exchange Tests

Sister chromatid exchanges (SCEs) are reciprocal exchanges between the sister chromatids of a chromosome and represent a consequence of genetic damage, which is, as yet, poorly understood. The experimental unit is either the culture (*in vitro*) or the animal (*in vivo*).

Sister chromatid exchanges in Chinese hamster ovary (CHO) cells are Poisson distributed (*17*), but in human lymphocytes and in animals they are not; and there is no single family of distributions that can be used for all data sets (*18*). Although a square-root transformation can therefore be used for SCE in CHO cells, appropriate transformations for other systems need to be empirically determined.

The transformation is to ensure that the data to be analyzed are of approximately constant variance, and then it is recommended that analysis of variance (ANOVA) be carried out. The form of ANOVA can, however, be chosen when differences between replicate cultures or animals have been checked. Between-cells or between-cultures estimate errors will then be chosen as appropriate.

Finally, dose response can be checked by performing a trend test of treatment totals, and this can be performed on untransformed data when a Poisson model is satisfied, as in the case of CHO cells.

Micronucleus Test *In Vivo*

In this assay, chromosome fragments or whole chromosomes that do not segregate correctly become detached from the main nucleus of bone marrow cells and, when the main nucleus is expelled to form an erythrocyte, they are left behind and appear like micronuclei. Cells either have micronuclei or they do not, and the relative rarity of micronucleated cells in control animals ($<4/1000$) means their distribution approximates to Poisson.

A whole raft of different significance tests (analysis of variance, likelihood ratio tests, generalized linear models, and 2×2 contingency tables) have all produced similar conclusions with sample micronucleus data and can be equally recommended. The Kruskal-Wallis, Mann-Whitney U and Jonckheere's nonparametric tests are all feasible alternatives.

The authors did recommend, however, that a t-test should not be used without transformation of the data, and Kastenbaum and Bowman tables should not be used without first checking for heterogeneity between animals.

Chromosomal Aberrations *In Vivo*

The end point of the chromosomal aberrations assay is similar to that *in vitro*, and it is assumed that variation between cells in an animal is binomially distributed. However, variability between animals *in vivo* is generally greater than between replicate cultures *in vitro*.

A χ^2 test or analysis of variance after arcsin transformation of the data is recommended to test significance, but there is some debate as to whether these tests should be modified if heterogeneity within groups of animals is found. A compromise was suggested such that a test for heterogeneity is performed, and the findings reported, but then χ^2 tests carried out regardless.

Other Tests

The dominant lethal and Drosophila sex-linked recessive lethal tests are infrequently used, and will be dealt with only briefly. A variety of approaches such as nonparametric, normal distribution, and β-binomial methods can be used with dominant lethal data, but the design of the studies must include deliberate randomization of animals if any of these approaches is to be valid.

For Drosophila assays, a test based on the normal approximation to the binomial distribution is favored, unless sample sizes are small, and then the conditional binomial should be used. The data are not suitable for analysis by Fisher's exact test.

Conclusions

It has been possible here to make only a very cursory overview of the lengthy discussions presented in the UKEMS statistical guidelines (*19*), but it is hoped that the benefits of a joint approach between statisticians and biologists will be clear to all and that it may prompt others to use a similar collaborative approach in the future.

Appendix

Discussion of Lecture

Q. Some guidelines recommend three plates/dose for the Ames test and others (e.g., Japan) recommend two. What does UKEMS recommend?

A. As many bacteria as possible should be plated. I think the working group actually recommended four replicates/dose as a minimum but knew this was out of line with OECD/EEC guidelines. What was most important was to increase the number of plates in negative control groups to twice those in treated groups (e.g., 6 plates/control) to improve sensitivity.

Q. A referee of a paper was asking for transformation of *in vivo* SCE data because heterogeneity between animals was seen at the doses producing a positive response. Did UKEMS recommend all SCE data be transformed?

A. No. UKEMS recognized SCE in CHO cells were Poisson distributed and so could easily be transformed by taking square roots. It was noted that in other cell types and *in vivo*, SCE were not Poisson distributed. The recommendation was to look at the data and see what transformation, if any, was appropriate. Furthermore, we would usually exclude positively responding animals from heterogeneity checks because we expect greater variation in those animals, so I do not agree with the referee's logic or conclusion.

Q. Why did UKEMS not use the nonparametric recursive ranking method of Simpson and Margolin for the Ames test?

A. Probably because the author of this chapter had worked with Wahrendorf.

Q. Sometimes Ames data from some strains in the negative control situation, and other times with positive responses, show non-Poisson distribution. How does this affect UKEMS recommendation for tests?

A. We recognized that overdispersion occurs on some occasions, particularly where there are positive responses, and so the simulation study used a positive response with data distributed according to negative binomial. As I mentioned earlier, untransformed data gave higher significance values than transformed data and, depending on whether there was a downturn through toxicity, you could choose Dunnett's test, linear regression, or Wahrendorf's test. I think we tried in the book to encourage people to look at their data and see the best way to handle it, not being too rigid. I think we also made one very clear statement: whatever result the statistical analysis gives,the biological conclusion is most important. Statistical analysis is an aid to that biological conclusion; it is not the conlusion itself.

REFERENCES

1. OECD. Guideline for Testing of Chemicals. Genetic Toxicology, No. 471-474. Organisation for Economic Cooperation and Development, Paris, 1983.
2. OECD. Guideline for Testing of Chemicals. Genetic Toxicology, No. 475-478, Organisation for Economic Cooperation and Development, Paris, 1984.
3. OECD. Guideline for Testing of Chemicals. Genetic Toxicology, No. 479-485, Organisation for Economic Cooperation and Development, Paris, 1986.
4. EEC. Methods for the determination of physico-chemical properties, toxicity and ecotoxicity; Annex V to Directive 79/831/EEC. In: Official Journal of the European Communities, No. L251, European Economic Community, Brussels, 1984, pp. 131-145.
5. UKEMS. (UKEMS Sub-committee on Guidelines for Mutagenicity Testing. Report.) Part 1, Basic Test Battery (B. J. Dean, Ed.), United Kingdom Environmental Mutagen Society, Swansea, UK, 1983.
6. UKEMS. (UKEMS Sub-committee on Guidelines for Mutagenicity Testing. Report) Part 2, Supplementary Tests, (B. J. Dean, Ed.), United Kingdom Environmental Mutagen Society, Swansea, UK, 1984.
7. Mutagenicity Studies. In: Drug Approval and Licensing Procedures in Japan 1986. Yakugyo Jiho Co. Ltd., Tokyo, 1986, pp. 173-178.
8. Stead, A. G., Hasselblad, V., Creason, J. P., and Caxton, L. Modelling the Ames test. Mutat. Res. 85: 13-27 (1981).
9. Margolin, B. H., Kaplan, N., and Zeiger, E. Statistical analysis of the Ames Salmonella/microsome test. Proc. Natl. Acad. Sci. U.S.A. 78: 3779-3783 (1981).
10. Bernstein, L., Kaldor, J., McCann, J., and Pike, M. C. An empirical approach to the statistical analysis of mutagenesis data from the Salmonella test. Mutat. Res. 97: 267-281 (1982).
11. Dunnett, C. W. A multiple comparison procedure for comparing several treatments with a control. J. Am. Stat. Assoc. 50: 1096-1121 (1955).
12. Dunnett, C. W. New tables for multiple comparisons with a control. Biometrics 20: 482-491 (1964).
13. Wahrendorf, J., Mahon, G. A. T., and Schumacher, M. A non-parametric approach to the statistical analysis of mutagenicity data. Mutat. Res. 147: 5-13 (1985).
14. Leong, P.-M., Thilly, W. G., and Morgenthaler, S. Variance estimation in single-cell mutation assays: comparison to experimental observations in human lymphoblasts at 4 gene loci. Mutat. Res. 150: 403-410 (1985).
15. Finney, D. J. Statistical Method in Biological Assay, 3rd ed. Griffin, London, 1978, p. 373.

16. Collings, B. J., Margolin, B. H., and Oehlert, G. W. Analyses for binomial data, with application to the fluctuation test for mutagenicity. Biometrics 37: 775–794 (1981).
17. Margolin, B. H., Resnick, M. A., Rimpo, J. Y., Archer, P., Galloway, S. M., Bloom, A. D., and Zeiger, E. Statistical analyses for *in vitro* cytogenetic assays using Chinese hamster ovary cells. Environ. Mutagen. 8: 183–204 (1986).
18. Carrano, A. V., and Moore, D. H., II. The rationale and methodology for quantifying sister chromatid exchange in humans. In: Mutagenicity: New Horizons in Genetic Toxicology (J. Heddle, Ed.), Academic Press, New York, 1982, pp. 267–304.
19. Kirkland, D. J, Ed. Statistical Evaluation of Mutagenicity Test Data. Cambridge University Press, Cambridge, 1989.