Review

STATISTICAL SCIENCE
AND QUANTITATIVE UNDERSTANDING

David J. FINNEY

"Some scientists use statistics in the way that a drunken man uses
a lamp post: he expects support but not illumination."
[A remark often heard from statisticians!]
13 Oswald Court, South Oswald Road, Edinburgh EH9 2HY, UK

(Received February 25, 1997; Accepted March 16, 1998)

ABSTRACT — Authors of papers in biological journals need to make their uses of statistical methods
and software explicit and maximally comprehensible. Editors and referees have important responsibilities
for this. Inexact use of technical terminology causes confusion. Computer software is invaluable
but not infallible. Graphical presentation should not conceal numerical results. Tests of statistical significance
should be reserved for specific needs, which will rarely include multiple comparison procedures.
Experiments that involve repeated measurements need special care in statistical analysis. Full attention
should be given to principles of statistical estimation as well as to choice of appropriate statistical technique.
At all times, ethical standards of scientific integrity must contribute to precision and clarity.
Clinical research that neglects well-established statistical principles may be intrinsically unethical.

KEY WORDS: Diagrams, Ethics, Probability, Significance, Software, Vocabulary

INTRODUCTION

This text was prepared as material on which I would speak at the Twenty-fifth Annual Meeting of the
Japanese Toxicological Society. The pre-eminence of English in modern scientific communication is convenient for me. Although I take pride in its quality for science, I make no claim that English alone is suited for this position: I deplore all impoverishment of human culture that may result from linguistic imperialism.

The paper is concerned with choosing statistical methods appropriate to a problem in biological research, to the data available, and to presentation of statistical argument in the eventual publication. I intend to be provocative, but to avoid giving offence. I write for discussion at a conference, with the hope that my audience will listen in critical spirit and will be stimulated to think further about my ideas, not necessarily always agreeing but attempting to formulate clear reasons before dismissing any recommendation as impracticable or undesirable. Effective publication requires language that flows readily from the author, unambiguously in meaning in respect of concepts, modes of thought, and technology. Often, good statistical argument is essential to clear presentation.

Statistical computations and statistical interpretation of data are not sauces to add to biological research solely in order to please personal taste. Commonly, they are essential to integrity, precision, clarity of understanding, and sound scientific conclusion. Even for good data, statistical inference should never begin without attention to detailed information on their acquisition. For an experiment, details of design, the nature of treatments (including quantitative values and logical structure), randomization, and units of measurement must be in front of whoever is to decide on a statistical method.

A LITTLE HISTORY

In 1939, I became responsible for statistical analysis and reporting of experiments on field crops at an agricultural research institute, for the designing of new experiments, and for statistical advice to various biologists. Much of my work was standard analysis of variance, laborious with the electromechanical calculators

Correspondence: David J. FINNEY at the above address.
of the time but teaching me always to organize and check arithmetic carefully. Help to other scientists could involve newer problems, possibly needing development of appropriate theory and technique for analysis. If the method of analysis that I recommended was too intricate and too time-consuming for a botanist or entomologist to undertake himself, I might need to advise on the adequacy of popular approximations. I soon became aware of the power of iterative techniques.

Most scientists now have access to a powerful computer. This may have software able to complete almost every statistical analysis that its user is likely to want. The entry of data may be automated, or may need to be done at a keyboard in an operation comparable with what my desk calculator once required. For most data, the arithmetic may take a fraction of a second or possibly, for a data set that a biologist would consider large, 5 to 10 seconds as compared with 1 to 18 hr in 1939.

This change has not always improved scientific understanding. Although the computer user, whether statistician or pharmacologist or toxicologist, may have a great range of software, to recognize what is best for his research and data may be difficult. The outcome can be that an inappropriate analysis is conducted. Recent years have seen proliferation of publications on new statistical theories and methods, most of which can easily be programmed for analyzing data. These may be incorporated into software packages; new methods can be mathematically and logically correct, yet dangers arise in the absence of critical advice on the circumstances that call for using them.

Unfortunately, not every software package is correct in its theory and logic. MINITAB is popular with teachers of statistical practice; it is powerful and easily used, but it can output nonsense from a simple significance test. Even in published papers, I have seen assertions that an analysis using this package led to $P \leq 0.0000!$. There is need for revision of some software, with a view to preventing absurd output that may pollute scientific journals. As I illustrate later, software producers are not always ready to correct mistakes.

**PUBLICATION AND EDITORS**

Whatever his own language, in all that he publishes about his research a good scientist will wish to make every sentence and every table or diagram present factual truth with maximal comprehensibility. Many who have had little statistical training employ powerful software packages that they do not fully understand. They may then unjustifiably assume that computer output is always expressed in standard terminology and notation. The reader of a published paper may consequently be unable to discover exactly what methods have been used and what is meant by symbols and technical terms.

In recent years, despite increasing sophistication of available software and methods, standards of presentation of scientific discussion relating to statistical analyses have declined. Descriptions of design for experiments and sample surveys often leave uncertain whether selection of units was truly random or thinly-disguised systematic. Computer graphics can encourage publication of elegant statistical diagrams ornamented with unexplained lines and arrows.

Some editors try to maintain statistical standards by advising authors on clarity of presentation. A few send many or most submitted papers to a statistical referee. An editor who asked me to be a referee stated that all he wanted from me was: "firm statements that a particular statistical treatment of data is right or wrong". No referee can make such statements unless he is told the method and the software used: many authors omit this vital information, and even fail to explain exactly how they obtained their data. Altman *et al.* (1983), Bailar & Mosteller (1988), Gore *et al.* (1992), Murray (1988), and Finney (1997) have commented on the problem. Altman (1998) has produced a valuable comprehensive survey.

A reader of a published paper about research in his own field (whether this be chemistry or botany or toxicology) should be able to discover, from the text and the bibliography, the methods and procedures (experimental, observational, chemical, physical, statistical and computational) that the writer used. If that reader has appropriate resources, he should then, without making guesses at meaning, be able himself to undertake comparable research. Listed references to sources in other books and papers need to be complete and unambiguous. Too many papers omit information that is essential to trustworthy inference, such as the number of observations, the number of degrees of freedom, or even clear identification of the statistical methods employed. A recent paper in the *British Medical Journal* mentioned use of "the Andersen-Gill extension of the Cox model", identifying this unfamiliar procedure only as a facility provided by a named software package!

Chemists have followed Lavoisier in adopting a complex symbolism that utilizes a large part of the
Roman alphabet. The scientific truths of chemistry are in no way dependent upon the universal use of C to symbolize carbon, yet so firmly established are many symbols that expressions such as H₂O and CO₂ are now widely understood in colloquial speech. Comparison with mathematics is interesting: despite affection for the constants π and e, every mathematician demands great freedom. He will on occasion use these letters with other meanings, but a strong tradition will ensure that, if he does, he first defines his choice of notation.

Statistical science has not adopted an internationally agreed nomenclature or symbolism. A paper submitted to a biological or a statistical journal may use symbols that follow the convention of 'Greek letters for parameters', but may include others that are peculiar to a software package or to an idiosyncratic textbook. An author who assumes these to be generally understood may mystify readers. The editor and his referees should endeavour to secure clarification before publication. This raises a general question: what statistical symbols and terminology can properly be used by an author without explicit definition, in the belief that all readers will understand? Too commonly, a reader is left to infer that a parenthetic comment "(n = 50)" is stating the number of animals per group for the treatments currently under discussion (or perhaps a number of degrees of freedom), and even to guess the implications of a statement: "Standard statistical methods were used".

For one biological journal, Finney & Harper (1993) proposed a code of statistical practice. This lists symbols and technical terms that, if used with the stated meanings, the editor will accept without explanation. For example, the code advises that the symbol 'r' should always be accepted as signifying the Product Moment Correlation Coefficient, modifiable to rₓᵧ if the variables involved need to be named, but that an author ought to insert explicit definition of any symbol he chooses to use for a rank correlation coefficient (perhaps rₛ for Spearman) or other measure of association. Similarly, 'b' should denote a simple linear regression coefficient of y on x, where x and y have already been defined, modifiable to bₓᵧ if the variables need to be specified. The word 'slope' should not be used as a synonym for 'regression coefficient'. An author who chooses to use symbols and terms not included in the code should define or explain them or give bibliographic references (Finney, 1995b).

No ban on other usages is implied. Still less is there any limitation on choice of method of analysis or discouragement of new methods. A code of statistical practice should not be designed as dogmatic restriction requiring that referees undertake a heavy task of censorship. It need be no more limiting to a scientist's freedom of expression than insistence that, in every chemical context, Fe shall represent iron. Possibly each major journal can benefit from having its own code appropriate to current interests in a particular field of science. In implementing such a code, an editor should intend both to simplify the task of referees and to aid readers of the journal.

Scientists should ensure that their units of measurement and recording are unambiguous. A careless writer may discuss "prevalence" without stating whether he expresses this as a proportion, a percentage, or even as a transformed value such as the arcsine. Many forget that a regression coefficient is a 'rate of change' involving a pair of dimensions of measurement: a numerical value may need to be stated as kg per week, or ml per °C, according to the units for the dependent and independent variables. Omission of units for a regression coefficient, unless these are clearly implied by the context, is a flaw that renders the numerical value meaningless.

USE OF WORDS

Every scientific discipline has special vocabulary. Even within a single language, the scientist must learn to distinguish differences of meaning for words that look and sound similar: as examples, I mention "ferric" and "ferrous", "genetic" and "generic", and again "typhoid" and "typhus". Statistical science has particular difficulties for the unwary, because developers of new ideas and methods regrettably often take words from colloquial language and, without change or after slight modification, give different special meanings to them.

In English, the word "error" can refer to any small mistake, such as using an incorrect telephone number; with this meaning, it may be wanted in an account of a scientific study. In a statistical context, if an author does not distinguish "standard error" from "standard deviation", he leaves his reader to guess whether or not there has been division by √n, a potentially misleading departure from exact statement. The symbols > and < should be used only in numerical or algebraic statements: they are not permissible in a sentence such as "In adults, the drug was < effective."

Many common verbal mistakes can contribute to misunderstanding and misinterpretation of results. A "parameter" is a quantitative property of a conceptual population, not a property of actual data. The word
"Random" should never be used except in relation to a strict process of random selection. "Likelihood" is a quantity calculated from data, never a synonym for probability. A "histogram" is a particular type of block diagram for graphical representation of frequencies, never a name for a Manhattan diagram. "Significant" in relation to a statistical test has no reference to the practical importance or scientific interest of the effect under discussion. Even "regression" and "variance" are common in colloquial English with meanings totally different from their uses in statistics. I shall not attempt to give a complete catalogue of potential verbal confusions!

MATERIALS AND METHODS

In any publication, a section with this title is surely the place for satisfying editors who have urged authors to: 'Describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to verify the reported results' (ICMJE,1988). This should minimize the need for a reader to make an inspired guess, or an ingenious inference, in order to discover what an author has done!

Information about use of unusual methods or special software should also be stated (Finney, 1993). This is important for journals in which statistical analyses and statistical modes of thought are commonly used. Some terms, such as 'multivariate analysis', 'cluster analysis', and the currently fashionable 'metanalysis', relate to types of statistical technique with many variants: a reader needs more detail in order to identify exactly the statistical methods used.

Within most biological disciplines, certain statistical terms and idioms can be assumed familiar to all serious readers. In a code of statistical practice, an editor might list those that he judges (necessarily subjectively) now to be part of the accepted vocabulary of the science to which his journal relates. Thereafter, no author need define or explain these, but authors should insert bibliographic references to identify use of any variants or other statistical terms. Even that well-known symbol $x^2$ does not uniquely identify a method, since it arises in connexion with many different statistical procedures. Similarly, a process of random selection may need detailed description, in order to ensure that its strict sense is distinguished from a possibly biased haphazard choice. A technique such as the 'Mantel-Haenszel method' may be popular with ecologists yet is perhaps almost unknown to pharmacologists. A new method or concept may have been named by its originator, but even a professional statistician may not recognize a name that is rarely seen outside a special field.

COMPUTERS AND SOFTWARE

The labour of statistical arithmetic has been banished by computers. No statistical analysis or inference should ever be attempted without specification of the origins and manner of acquiring the data. Application of sophisticated software to data structured differently from assumptions inherent in the logic of the software can bring disaster. Choice of software may have important consequences for publication of results. Even intrinsically good software cannot convert poor data into good science.

Unless editors act firmly, papers will continue to appear in which an author states: "All statistical analyses were performed with "STATMAGIC", yet fails to disclose what this package is or where it can be obtained. How can a reader comprehend the paper, especially if it displays diagrams showing fitted curves and limits of error without indicating what principles underlie the graphical presentation. Even when he has used a package of high repute, an author should not assume computer output to be intelligible without careful explanation. Every author has a duty to make the reporting of his results and the analysis of his data comprehensible to readers, so as to aid any who may wish to adopt similar methods.

In order to use a novel method of statistical analysis, a scientist may choose to write his own program. This potentially addictive activity can be relatively easy, but it requires experience. Failure to check program code and its performance may be disastrous. A few years ago, a statistician wrote a program for estimating a clinically important parameter from N pairs of values of X and Y (Finney, 1995a). This involved simple but unusual computations. Personal anxieties caused him accidentally to interchange X and Y, so obtaining a program that gave the reciprocal of the desired result. His omitted check led the World Health Organization to publish a report that included a misleadingly faulty numerical example!

A first step towards a new program should be to inquire whether any existing well-tested package contains what is wanted. Among the ever-growing number of general-purpose statistical packages, a few have been so extensively used that they are now recognized as standard accurate and tested products, with names assimilated to the general language of science; in unbiased alphabetic order, and with no implication that
Understanding Statistical Science.

Several standard packages, such as BMDP and SAS, include software that handles the probit and related methods by trustworthy computations for maximizing likelihood and producing parameter estimates to any desired numerical accuracy. The hr of iterative calculation required 60 years ago are today replaced by only a few minutes for preparing data for input and at most 5 seconds of PC time. My own programs BLISS and QUANTAL are specially designed for quantal response data. I believe that software known as Y, M, W is popular in Japan for the same purpose, but I have no experience with it. I advise against summarizing results solely in terms of ED50s. This quantity may depend upon uncontrolled and varying factors: assertions that the effective or safe dose of some manufactured product can be expressed solely by its ED50 in mg per kg body weight might have deplorable consequences.

The SAS user's manual illustrates dangers. I have no reason to doubt the correctness of the theory used in its probit computations, but the explanatory text has a wrong definition of the probit function and other confusing errors. The manual's advice on how to use the program runs contrary to my own experience and recommendations. In July 1997, I informed the company of these flaws, but I have not received any response.

DIAGRAMS

Fifty years ago, to draw a diagram for a publication or for a letter was a tedious task. Today, ingenious software will accept a file of data and rapidly output well-drawn diagrams. Authors find some possibilities so attractive that they and editors may express preference for a presentation of results that is not always best for the reader. I have tried elsewhere (Finney, 1986) to classify graphical representations of data systematically according to objectives. Every scientist can benefit from the superb book by Tufte (1983).

Some currently popular packages (Microsoft EXCEL is one) produce a block diagram that represents a set of mean values by rectangles of heights proportional to the means, a style that could almost be an artist's impression of the Manhattan skyline! On the roof of each skyscraper, the height of a radio mast may represent a standard error or a confidence interval; the software allows the user little control over details. The output may be uncritically inserted into a publication, often with no mention of the software used. It is NOT a "histogram"! Many an author who presents such dia-

these alone are good, I mention BMDP, Genstat, GLIM, MINITAB, SAS. A text reference to one such package may need no identification beyond statement of the choice of method within it. A wise statistician will discourage use of software of unknown quality, however attractively it may be advertised. If he fails in this, he may bring his profession into disrepute among those who discover too late that software used on their problems has produced absurdities.

In the 1920s, statisticians began to propagate ideas on optimal procedures for estimating unknown parameters. Chief among these was the "Principle of Maximum Likelihood"; this states that the best (according to various criteria) estimate of an unknown parameter is that which makes a certain function of the data (the LIKELIHOOD) as numerically large as possible. For many problems, the classical method of finding a maximum, namely differentiate the function with respect to the unknown and equate to zero, breaks down because the equations obtained cannot be solved analytically. This difficulty can be overcome by iterative computation: a value for the parameter is guessed and used in a sequence of simple but possibly laborious calculations to obtain an improved value, with which the process can be repeated as often as desired. In most statistical problems, the succession of values converges to the desired maximum likelihood estimate.

Toxicologists have attached particular importance to estimates of the ED50 from quantal response data. In the early 1930s, an ingenious system of computation was developed by C. I. Bliss and R. A. Fisher. This probit method was labour intensive, especially when used by biologists whose only mechanical help to arithmetic was a hand-operated calculator. A single cycle of iteration from one initial value to the next might easily take an hr or longer, and five cycles might be needed for one set of data. Many people proposed simplified approximate computations, or even ingenious graphical procedures, thereby earning gratitude for saving time but sacrificing some accuracy. Among these, the Spearman-Kärber, Reed-Muench, and Litchfield-Wilcoxon methods became popular.

These approximations, often termed 'quick and dirty methods', should now be abandoned. To insist on retaining one of them is analogous to declaring in 1998: "During my happy childhood, my home was lit by oil lamps and candles. Even now, I continue to use oil lamps instead of risking the dangers of electricity!". All modern computers are well suited to iterative computation, and so can quickly maximize an arbitrary mathematical function.
grams does not show any other quantitative results from his research.

Fig. 1 is typical of these 'Manhattan Diagrams'. It purports to compare blood concentration of a substance in subjects who have received one of four drug treatments. The diagram represents treatment means and their standard errors on the vertical scale.

The order of skyscrapers along the street (horizontal axis) is arbitrary and without quantitative meaning. The text may explain that the vertical masts, often termed "error bars", represent standard deviations, or standard errors, or confidence limits; some authors, erroneously regarding them as a recognized convention, leave them unexplained. I have seen instances in which no vertical scale is shown, so leaving no clue to numerical values. Despite my personal dislike, I concede that Manhattan diagrams can have valuable dramatic impact, especially for a presentation of general interest rather than formal science. They are of course valid representations of facts, but they can be rather silly when used to portray means from only two treatments, quantities that could easily be stated within a short sentence in the text.

A more serious fault is the concealment of important information that was in front of the author when he was writing his paper. A reader may wish to compare a result from the paper with information in other publications or from his own research. Even in order to make a simple statement such as: "Subjects on treatment D had a blood concentration 70% greater than those on treatment C", he must measure heights of rectangles on the printed page and may also need to look at conversions of scale. I question the ethics of a practice that obstructs the reader's access to a simple table of means such as:

![Graph showing blood concentration across treatments with standard errors indicated by error bars.](image)

Fig. 1. [based on invented data] Mean blood concentration of substance under study (µg per ml) for patients in the four treatment groups. Vertical bars show standard errors; asterisks indicate significant excess over the Control.

| Table 1. Values from which Fig. 1 was drawn: Mean blood concentration (µg per ml) for the four treatments, standard errors in (µg/l) |
|------------------------|--------|--------|--------|--------|
| Control               | B      | C      | D      |
| 7.7                   | 14.8   | 10.3   | 17.7   |
| (1.81)                | (2.52) | (1.94) | (2.75) |
A Manhattan diagram should not be an automatic first choice for m presenting summarized data. Even the Journal of Toxicological Sciences has too many, and thus often obstructs a reader who wishes to examine numerical values for means. At the present time, there is no general agreement on the conventions adopted for representing standard errors or standard deviations. Nor is it clear whether a reader should assume that the underlying analysis has taken account of non-orthogonalities arising from unequal numbers in groups.

**REPEATED MEASUREMENTS**

Often in toxicological research, the value of data may be augmented by making repeated measurements on each subject in an experiment, each treatment under study might be allocated to several subjects (human volunteers, animals, plants, cell cultures, etc.), and on each subject a measurement, such as weight or blood composition, is recorded at regular time intervals (perhaps every 10 days or every month). Treatments may remain constant throughout, or from time to time be repeated or modified. There may be a complex pattern of balanced changes in the treatments, so giving what is termed a cross-over or change-over design.

Experimental design and the time sequence of measurements need good planning, and the statistical analysis must receive critical attention. A software manual may seem to imply that: "If a procedure in this package will accept your data and produce an analysis, you can safely believe the results!". This uncritical attitude takes no account of whether assumptions underlying the program correspond with a model reasonable for the data. The ability to perform computations is no guarantee that a method is appropriate to the data.

Here lies a trap for a pharmacologist whose experience has not fitted him to assess available software, and who has no access to anyone who can advise him (Finney, 1996). Textbooks or software (notably STATVIEW) may encourage the idea that successive measurements on a subject can be regarded in the same way as data from the split-plot designs sometimes used in agricultural field trials. This outlook takes no account of the impossibility of randomizing in time, so making logically indefensible an analysis of variance appropriate to a factorial experiment with 'Time' as one factor. Such an analysis grossly mishandles the error structure of the experiment: it may lead to biased estimation of experimental error as well as to meaningless tests of significance. A correct analysis can be based either upon comparisons among measures of individual time trends, or upon a special form of multivariate analysis.

I do not know of any general software written to analyze data of this kind. I have stressed the plausibility but potential faults of particular packages. A toxicologist who undertakes such experiments might persuade a statistical colleague to write a good general program; the task should not prove unduly laborious.

**STATISTICAL SIGNIFICANCE**

I deprecate excessive emphasis on tests of statistical significance. Some editors, notably of medical journals, seem rarely to permit any statement of conclusions unaccompanied by a test of significance. They may even appear to judge the suitability of a paper for publication by its abundance of tests and associated asterisks.

A significance test has only one purpose: it assesses the strength of evidence against a stated, or clearly implied, NULL HYPOTHESIS. The hypothesis asserts some quantity to be zero, typically the difference between two different categories of subjects in respect of the 'expected' or population means for a measurement that can be made upon each subject. The test uses records of the measurement, from experiment or observation on individuals, computes empirical means for these individuals, and compares the difference in means with the variability among individuals. The conclusion has the form: "If the null hypothesis is true, then in a series of similar trials a difference at least as large as that found would occur with probability, or relative frequency, P". If P is small, say P < 0.05, this occurrence is by convention regarded as so infrequent as to justify rejection of the null hypothesis on which it is based! I believe that most thinking persons argue in similar manner in respect of everyday matters, but without the formal quantitative framework.

Suppose that a significance test has led to P = 0.048; some who do not fully understand may infer that 0.952, or (1-P), is the probability that the null hypothesis is true, but this is false logic. A significance test can help interpretation of data, especially in discouraging an investigator's wish to assert that even a very small apparent advantage for a new drug or treatment is important. No one should make a significance test without having clearly in mind his null hypothesis. Comprehension of the hypothesis under examination often suffices to convince that its trivial nature removes all real interest from testing it. Does one need a signifi-
cance test to demonstrate that the body weight of a mammalian species is NOT UNAFFECTED by increased dietary intake of carbohydrate?

Excessive and purposeless testing has been stimulated by numerous software packages and by many publications of new test procedures, especially those that are 'distribution-free'. Those who act as though all advances in science can be viewed in terms of significance tests and associated probabilities sometimes state a conclusion: 'This test failed to achieve statistical significance', almost as though such failure were analogous to loss of a football match.

The bad practice of authors omitting to identify their methods is particularly common in respect of the multitude of significance tests known only by the names of their developers. For example, in a major biological journal, the text of a recent paper on sexual selection in fish contained the following conclusion: "The probability of virgin females being inseminated decreased with male relative size (Wald = 6.26, r = -0.248, $P < 0.01$, n = 50)''. Possibly 'Wald' relates to a test devised by the distinguished statistician Abraham Wald, but I think it unlikely that his surname is in general use as the symbol for a test statistic. Only a somewhat obscure reference gave any clue to the source of the test, and the author failed to tell his readers whether $n$ denotes the number of virgin females examined or a number of degrees of freedom.

Widely available software can now compute tail probabilities for test statistics much more exactly than in the standard tables of fifty years ago, values with possibly many initial zeros. Does inclusion of these indicate the high quality of a journal? Only excessive faith in Normality, or in true randomness of allocation (as distinct from an ill-defined allocation adopted in the belief that it is 'effectively random'), can justify 8-digit accuracy in a probability computed from real experimental data. A value of $F$, the familiar variance ratio statistic, may be calculated to have an associated probability $0.04282906$; this may be mathematically correct from distribution theory, but circumstances will scarcely justify total belief in the assumptions underlying that theory. An editor can be misled by this apparent accuracy into thinking it indicative of advanced, or 'state of the art', statistical analysis! Such spurious numerical exactness is rarely desirable: commonly no more need be published than "$P = 0.043$'' or the formal "statistically significant at $P = 0.05$''

**PROBABILITIES**

Biologists are not alone in finding probability a difficult concept. Some use euphemisms, such as 'chance' or 'odds': unfortunately, these often introduce ambiguity. A statement that the chance of an accident is 25% ought to mean exactly the same as that the probability is 0.25: because advertisers and politicians so misleadingly exploit uses of percentages, I advise that probabilities be always stated as decimal quantities.

I think that the word 'odds' originated with gamblers. Careless writers, journalists and even scientists, often fail to distinguish between odds for an event and odds against that event! A newspaper report on an escape from drowning stated: "Doctors said that before rescue the man's chance of survival had been a million to one''. The search for phrasing that will not scare an innumerate reader can lead to similar ambiguity, a possibility exploited by publicists for such follies as lotteries. Who can fail to find attractive the assertion: "The odds or chance of each single ticket winning a prize is 15,000 to 1''? This type of mistake can be found in scientific journals, but there likely to have been a result of ignorance rather than intent to deceive.

A correct statement of odds is exactly related to a probability, but it is not the same thing. The simple relation is that an event with odds of 'f to 1 in its favour' has probability $f/(f+1)$, whereas if the odds are 'a to 1 against' the event, the probability of the event is $1/(a+1)$. Within the terms of these equivalences, it is legitimate to replace any probability by a statement of odds and vice versa doubt whether scientific writing is ever helped even by correct use of odds.

**MULTIPLE COMPARISON TESTS**

Some experiments compare many treatments simultaneously, each administered to its own small set of subjects: K related chemical compounds may be screened for therapeutic activity, or a plant breeder may make preliminary trial of the yielding potential of K newly produced genotypes. If the K distinct treatments can be classified into groups with the property that all belonging to the same group are identical in respect of the quality that the research is intended to improve, conclusions and practical recommendations will be simpler.

Statistical test procedures intended to discover such a group structure have been devised. Alternative specifications of the problem have demanded much ingenious mathematics. These so-called 'multiple com-
comparison tests' purport to classify the K treatments into, say, I groups with every difference between two groups being statistically significant. Procedures of steadily increasing complexity have been known for over 50 years. Superficial simplicity makes them attractive to authors and editors, but I have never encountered a practical situation in which this group structure seemed to have sufficient inherent plausibility to be worth trying to use. Greater attention to the source and nature of data, and to objectives consistent with these, might discourage uncritical presentation of results in a way that may derive from statistical imagination rather than possess external meaning. Especially undesirable is the use of multiple comparison tests on treatments that comprise a series of different doses of a drug. Almost certainly, the measurement being studied should show some continuous trend relative to dose, and a form of regression analysis is likely to be desirable. Chance variations may then cause multiple comparisons to produce nonsensical groupings, or to find no differences.

ESTIMATION

Research should usually be directed at finding widely applicable generalizations, rather than at making numerous tests of significance. For example, a taxonomist might be content simply to count features -limbs, teeth, colour patterns, etc. - and also to list purely descriptive properties. To a toxicologist, however, interest attaches to quantitative characters, such as tissue damage or survival time or ED50; even among individuals treated alike, these may vary substantially. Statistical science is essential to discussion of the effects of differential treatment on measures of them. A significance test does no more than examine the tenability of a hypothesis that on average the character of interest is unaffected by treatment. If that hypothesis is rejected, by formal test or otherwise, interest turns to the magnitudes of differences.

Quantitative expression of treatment effects and differences, perhaps by relative potencies or other parameters of the complex of organisms and materials under study, requires statistical estimation, supplemented by confidence intervals and standard errors to represent the uncertainty arising from inherent variability. I cannot today undertake general discussion of estimation techniques but, because it is insufficiently used, I wish to mention analysis of covariance. By taking account of variation in an otherwise unimportant concomitant measurement (such as initial weight of an animal), this technique can greatly increase the precision of estimating differences between means. If the procedure for obtaining the analysis of variance for an experiment is known, or is available from software, its elaboration into an analysis of covariance is easy. Corrupting its name to 'ANCOVA', or stating its purpose as 'Controlling for X' is a confusing misuse of words.

The classical approach to estimation for the drug Q is to hypothesize that, in respect of a desirable and measurable character, administration of Q causes a mean benefit equal to \( \theta \) units, where \( \theta \) is a totally unknown parameter. If Q is harmful, then \( \theta < 0 \); alternatively, it might be a fairly large positive number. An experiment that compares Q with untreated control subjects leads to measurements of the character, calculation of means, and estimates of variance; one then estimates \( \theta \) by \( d = \text{difference of means, and calculates its standard error. Unless I had strong reasons to the contrary, I might assume Normality of error distribution; I could then make a test of significance of the deviation of d from zero. I would conclude by finding limits, such that I could with confidence conclude } \langle d \rangle < \theta < d \rangle \text{, computed in such a way that d would be judged significantly different from any } \theta \text{ outside the interval (d1, d2).}

Popular with a few statisticians today are so-called Bayesian estimation procedures. If logically sound, these would have attractive features for the interpretation and presentation of results. I regard them as logically unacceptable, except occasionally as exploratory techniques and as leading to suggestions for new research plans. A Bayesian would begin by introducing a prior probability distribution for \( \theta \), a step that I think nonsense: a parameter, being a fixed quantity, does not have a probability distribution.

However, if he were convinced that no value for \( \theta \) outside the interval (0,20) is conceivable, he might assert an 'equal distribution of ignorance' over the interval and express this by asserting for \( \theta \) a rectangular prior distribution from 0 to 20. To me, this is folly; my unfamiliarity with subsequent details prevents me from showing you the mathematical development. However, if such a prior is adopted, the same Normality assumptions and orthodox mathematical operations can be compounded with the measurements in the experiment so as to express conclusions by a resultant posterior distribution of \( \theta \). This process employs a valid theorem, due to Thomas Bayes, but the argument depends vitally upon that initial assertion of 'equal distribution of ignorance' between 0 and 20.

This posterior distribution may look to be a simple
and useful summary of results; from it, one can easily evaluate particular probabilities, perhaps in such form as Prob.(θ > 2.5) = 0.062. Doubtless every statistician and every scientist would be delighted to have conclusions in this form. As long ago as 1925, most statisticians had discarded prior distributions of parameters as logically untenable. Although I admit that opinions are still divided, I regard the name 'Bayesian' as an insult to the memory of an eighteenth century philosopher of probability, who first explored these ideas and who wisely decided not to publish them! The typical mathematical operations may have uses for entirely different purposes. Moreover, there are circumstances, such as in statistical genetics under conditions of random mating, where the prior distribution of a parameter can be a sensible concept. I advise avoidance of Bayesian techniques, unless my reservations are clearly inapplicable to your problem.

The presentation of estimates from experiments is often best achieved by simple tables of means, each with its standard error; publications should not omit these tables. Editors should ensure that they are not replaced by lists of significance tests, collections of "***", statements of "n.s." or values of P, or Manhattan skyscrapers! Good experimental design is vital for the control of irrelevant variation. Statistical analysis for a complex design can be difficult for a novice: analysis of variance is usually the main technique, for which there is a vast literature that should ensure conformity to the rule that every constraint on the structure of a design shall correspond to one component of the analysis of variance. Most designs likely to be chosen can be handled by one of the many available software packages.

CHOICE OF STATISTICAL TECHNIQUE

The wealth of techniques of statistical analysis available today may complicate the choice of how to handle a new set of data. A biologist may be tempted to analyze his own data with help from a text book chosen as simple and attractively presented, scarcely adequate criteria for scientific reliability: such a book may contain absurd elementary errors! A typical fault is failure to distinguish between without replacement and 'with replacement' in simple probability problems involving selection from a finite population of possibilities. A pharmacologist in need of information on a chemical analysis or a physical property will not adopt a method solely because it looks simple!

The main principles of analysis and the broad style of eventual publication should have been known when the research study was being planned, but new factors may enter later. For an experiment, whoever will decide on a statistical method needs to be fully informed on the design, the nature of treatments (including quantitative values of doses and logical interrelations), randomization procedures, and units of measurement. He should also be acquainted with the research objectives and definitions of any parameters or other quantities that are to be estimated.

Attention to these matters should facilitate publication. Unless publication is to be an expository account of new methods, it need not include all arithmetical detail (tables of analyses of variance or lists of statistics and probabilities pertaining to significance tests) that may have formed a background to inference and the stated conclusions. For example, even if a significance test is desirable as part of a discussion of dependence of reaction times upon dose of a drug, a statement such as: "(F_{1,8} = 4.413, p = 0.05)" contains nothing of intrinsic interest; discussion might be helped much more by a regression coefficient or by a pair of means with an associated standard error. As explained in Section 11, emphasis on objectives consistent with the design of a project will often indicate the unhelpfulness of multiple comparison procedures.

Some modern software may tempt the user into a strange confusion. He may begin by making an analysis of variance, which automatically leads him to a pooled estimate of error variance. Then, in outputting a list of treatment means or in preparing a Manhattan diagram, he may receive a new estimated standard error for each mean, computed only from the replicates of the treatment without regard to the pooling or to constraints of design.

ETHICAL CONSIDERATIONS

A scientist should publish information on his research, either for the help of other scientists or because of potential benefits to mankind. In every sentence, and in every table or diagram, he should maintain factual truth and maximal comprehensibility. Whether he should suppress potentially harmful knowledge is a difficult moral issue.

I shall comment on statistical aspects of scientific ethics. For example, I question whether it is ever right to plan, conduct, and attempt interpretation of quantitative biological research without employing sound statistical practices, especially if experiments involve human subjects or other sentient beings. If I were to be
asked to analyze data that I judged to have been acquired by an unethical research programme, I believe that I ought to refuse.

A well-known medical scientist has written: "If you are scanning an article about therapy and it is not a randomised trial, why on earth are you wasting your time?". Should there be ethical doubts about the random allocation of human subjects to treatments in clinical trials? If a clinical trial is without a well-chosen experimental design, without proper randomization, without provision for an appropriate analysis, and without explicit rules for handling crises such as drop-outs or unavoidable changes of treatment, then patients with a disease may suffer discomforts and exposure to dangers without the reward of contributing to an important definitive study. Similar comments are relevant to ill-planned and poorly conducted animal experiments.

Scholarly integrity, truth and honesty, allied to fair dealing with the work and reputations of others, are vital to the ethical handling of data. Deliberately misleading presentation of research with a strong statistical component may be rare. Yet, if society permits the status of a biologist, or the continued funding of the institute in which he works, to depend upon the frequency with which he publishes evidence for statistically significant medical benefit from new treatments, temptations will exist. Software can be exploited so as to analyse one set of data in many ways; dishonesty enters if only the most dramatically successful conclusions are chosen for publication! To detect or prevent this source of bias is difficult, but alertness and care by editors and referees are essential.

Uninhibited use of Manhattan diagrams is potentially damaging to ethical standards. Author and editor may think that no reader will want to see the numbers lying behind a diagram. Yet should the published findings from a research project, possibly financed from public funds, ever conceal exact numerical values in a manner that makes excessively laborious the task of any reader who wishes to compare them with results of his own, or to compound them into a metanalytic review? I do not suggest that editors should ban these diagrams, but the present enthusiasm for them should be moderated.

A clever software writer might incorporate automatic procedures for adjustment of data before analysis. He might reject 'outliers', observations that fall outside arbitrary limits of plausibility, or he might apply uncritical procedures for handling missing values. These can amount to arbitrary interference with the data actually recorded: output from the software should report all such actions to its user, leaving him to assess the justification. Any resulting publication must take account of ethical aspects of alterations to the true data.

A scientist should report all relevant facts accurately. In clinical research, certain features of data may be highly confidential, but statistical collaboration should imply that the statistician has full access to all information that he needs. I once met a young statistician employed by a pharmaceutical company who was asked to analyze data consisting only of the percentage of successful responses for each treatment in a trial of six analgesics. I was told that a request to see the detailed records of subjects and responses might have been regarded as disrespectful from a junior to a senior clinician, although potentially relevant to her analysis.

A modern laboratory instrument can incorporate a microchip that will perform relevant computations while the instrument is in routine use. Auto-analyzers for hospital laboratories have been marketed pre-loaded with software for analyzing immunoassays and outputting the estimated hormone status of serum from each patient. A manufacturer may regard the algorithmic structure of that software as his commercial secret. He may have had expert advice on its writing, but this non-trivial task demands familiarity with aspects of bioassay. Secrecy conceals assumptions about the form of a response curve, rules for dealing with outliers, and all tests that data do not disagree with assumptions. Hence the diagnosis and subsequent treatment of patients may depend upon software that has never been checked by anyone other than its author, or subjected to critical appraisal by a statistician.

Does this remain true in 1998? We limit the sale or prescribing of drugs of unstated constitution. I ask: "Is it acceptable to society that care of patients may depend upon output from secret technology that is not open to scientific inspection and criticism?". I believe that to prevent disclosure of what statistical methods were employed in any statistical analysis is never justifiable.

**FINALLY**

I have discussed choice of statistical methods appropriate to a problem under study and to the data available, and also standards of presentation of statistics in biological publications. I have emphasized the value of good software and the dangers from uncritical use of unsuitable software. I urge that editors and their chosen referees should give greater attention to the elimination of common misuses of statistics, and should be insistent on authors making clear statements
of what they have done. I have advised reduced use of Manhattan Skyline Diagrams. I condemn excessive reliance upon significance tests and claim that understanding is rarely helped by multiple comparison tests. Data from good experiments in which measurements have been made repeatedly on the same subjects require special care if grossly misleading statistical analysis is to be avoided. I counsel great care in the choice of statistical techniques appropriate to the problem under study, allied to understanding of the essentials of statistical estimation; I unhesitatingly advise against the adoption of so-called Bayesian estimation. I conclude by surveying some of the ethical problems that may arise in analysis of data or omission of sound analysis of data from human subjects.

REFERENCES

FINNEY, D. J. (1993): On presenting the results of statistical analyses in scientific journals, Biometrical Journal, 4, 499-509./