Jour. Ind. Soc. Ag. Statistics
Vol. XXXVIII, No. 1 (1986), pp. 60-70

TABLES AND DIAGRAMS : WILL THEY HELP OR HINDER ?*

D. J. FINNEY
Department of Statistics, University of Edinburgh, Scotland

1. Introduction

Much has been written about technical aspects of presenting statistical information in tables and diagrams—the need for numbering, self-explanatory titles, logical placement in text, consistency of style within a book or article, and so on. I know of little discussion of what should be presented, what should be omitted, and why. A table or a diagram is part of the communication between an author and his readers; the author should plan it at least as carefully as he chooses words and sentence structure in his text.

I have recently had occasion to think carefully about the functions that tables and diagrams must perform in relation to scientific research. A systematic account of some aspects of these may be thought-provoking, and should be a useful guide to the production of tables well-suited to particular classes of presentation. I emphasize the need for standard errors as aids to future users of published tables, and condemn vigorously alternatives that restrict freedom of use.

2. Scientific Writing

Scientific endeavour is incomplete unless its results are communicated—to other scientists as increasing the whole body of knowledge, to those who will use them for technological advance, and to the public that ultimately provides the financial and other resources. We have three main modes of communication, words, tables, and pictures. Most good scien-

*'Dr V. G. Panse Memorial Lecture' (10th in the Series) delivered at IASRI, New Delhi on 19th November, 1985.

tists recognize the need for care in the language in which they report their work, even though not all succeed in writing well. This paper relates to the other two modes, and more particularly to tables and diagrams that present results of a statistical character. A scientist will take great care to produce a good photograph of an insect, or to ensuring the accuracy of a botanical key to the identification of species; yet, too often he will assume that his first draft of a table summarizing an experiment on insect control, or of a diagram showing plant growth under various conditions, will serve his purpose. Good tables and good diagrams need as much care in drafting as does good prose. There may be many satisfactory ways of presenting one table, none obviously the best, but avoidance of the even more numerous bad ways requires thought. The first step is to believe that good presentation matters.

First I list several classes of scientific writing. They may not always be sharply distinguishable, but an author is usually clear which is chiefly concerning him.

A. Research Communication

Written as a book, a journal paper, or a thesis to present new scientific knowledge (other than statistical theory).

B. Science Teaching

Exposition of established knowledge at any level from secondary school to postgraduate training.

C. Lectures and Seminars

Distinguished from A by the fact that material displayed is seen for a very short time.

D. Popularization of Science

Although some scientists may disparage this activity, in an age when most research is supported by public funds care must be taken to present science in a manner that makes some impact on public thinking. In this class I include such newspapers and other general publications as are prepared to report science seriously, good magazines of the types of *The New Scientist* and *The Scientific American*, annual reports of some research organizations and institutes, and some more purposeful writings such as grant applications.

E. Archives

Major reports in which the writing is primarily explanation of tables that are to be put on record for future reference. Reports on census and survey projects are typical, as are collections of information on large numbers of cultivars of a crop or on case histories of hospital patients.

F. Statistical Methodology

Exposition of new or established statistical techniques designed to help others to use them.

3. Types of Table and Diagram

I distinguish five types of table according to function. Again, of course, distinctions are not unambiguous but I believe an author should have some such classification in mind at the time he prepares a table.

Type I (Support)

Tables intended to provide quantitative support for text statements or to clarify a complex argument.

Type II (Reference)

Tables wanted solely for archive and reference purposes, and not for consecutive reading; associated text will explain what the tables contain, but the tables are not conceived as illuminating the text.

TYPE III (Mixed)

Tables having substantial elements of both these functions simultaneously. For examples, yields of sugarbeet under various fertilizer treatments may be tabulated to illustrate a discussion of new methods of application; the author may also recognize that a subsequent compilation of fertilizer responses may need to include his results, and possibly to combine them in ways not at present evident to him. Similar comments could be made on responses recorded in three mammalian species to various doses of several beta-blocker drugs.

TYPE IV (Lecture)

Tables for use in a lecture or other oral presentation.

TYPE V (Expository)

Tables used to aid the exposition of statistical methods or other scientific techniques.

Diagrams can be similarly classified, but for them Types I, IV, V are the most important. A diagram of Type II is unlikely to be wanted. The nature of Type III is that, in addition to its immediate effect, a reader may wish to extract numerical values for combination with other information; reconstructing numerical values from a printed diagram is a tedious (and often inaccurate) task, and consideration for this need will commonly suggest to the original author the wisdom of producing a table rather than a diagram. Tufte (1983) gives excellent advice, with horrifying examples to the contrary, on the construction of statistical diagrams. His insistence on a high ratio between useful information and total consumption of ink is equally applicable to tables.

4. Uses and Functions

Tables of Type I are the most universally useful, being appropriate to all classes of scientific writing except *E* (and even then they may be sometimes wanted for illustration). They should talk to the reader as fluently as does the surrounding text, with which they are closely integrated, so that every entry contributes to the impact. They must be short and to the point, rarely having more than 10 entries and preferably no more than 6. When the story to be told is complicated, two or three small tables will be more effective than one large one. Broad indications are more important than exact numerical information : this does not excuse distortion of fact, but accuracy of two (or at most three) digits suffices and coarse grouping may convey a message more dramatically than detailed values. For example, if data relate to insect counts, the reader will be concerned about a difference between 167 and 34 for two categories, but not by the difference between 167 and 182 : a grading as few (0-10), moderate (11-50), many (51-250), very many (251-) may inform more rapidly than do the exact numbers. Diagrams are similarly wanted for all classes of writing and also should be kept simple. If they can be well designed and produced, they are likely to be more effective than tables, but they are less adaptable. One rule, too often neglected, is that any diagram representing a relation between observations *must* have data points (or treatment means) indicated; a picture of lines and curves is little use unless the reader can himself appraise the fit to the data.

Type II tables are the other extreme. Their chief duty is to form the backbone of *E*. They may sometimes be wanted for *A*, but should usually appear as appendices to the main text (where they can be as long and as

detailed as their compilers deem advisable). A well-written article is not improved by the need for frequent page-turns to bypass lengthy reference tables, in the manner that advertisements must be bypassed in some commercial publications. Tables of mathematical functions and catalogues of designs perhaps should be judged of Type II in relation to *F*, and again an appendix is the right place. With a few exceptions, diagrams are an extravagant and unsatisfactory way of storing reference material; at one time, nomograms for statistical procedures would have seemed appropriate to *F*, but the ease of computing today must have greately reduced their use.

Type III tables are the most difficult to plan. The material should be well digested and reduced so that irrelevancies are removed, differences that are clearly trivial given minimal emphasis, and the immediately important facts brought into sharp focus. As one extreme among readers, I would sometimes welcome full tabulation of individual plot yields from an experiment because they could be useful as examples for teaching. At the other extreme, some readers might be content with seeing solely the treatment differences that are statistically and scientifically "significant". The one is impracticable, the other improvident. A particular experiment may show no indication that responses of potatoes to fertilizer vary with date of application; presentation of a two-way table eases the way for a subsequent averaging of experiments, which may disclose an appreciable advantage for application at the time of planting. There is therefore a need for summary tables to show information on factors that are not statistically significant in individual experiments. A further consideration is avoidance of bias. If tests of sowing cabbages at full moon or of causing rain by gunfire are reported only when a measured difference is statistically significant, a compilation of results from many sources may lead someone to conclude that research has shown positive effects. My remark may seem trivial, but this type of bias has almost certainly encouraged support for wild ideas in human medicine. Although sometimes a diagram may serve as an alternative to a Type III table, it will have serious disadvantages for an inquisitive reader who wishes to reconstruct numerical values as a basis for further calculations. If an editor will not accept both table and diagram, the author may have to make a difficult decision between visual impact for the message of the current work and information conveniently presented for future use.

Even more than those of Type I, Type IV tables must be kept simple. Figures displayed on a screen will do little to help an audience unless their message is so clear that it can be observed in 30 seconds. All unnecessary digits should be pruned, the number of different values shown should be very small, the layout and labelling should be most carefully planned. "These next few tables will give you an idea of our recent

results", followed by presentation of 6 tightly packed slides in 40 seconds, will seldom do much to help a lecturer's case. Here diagrams are ideal for giving a general impression very quickly. If the presentation requires that extensive numerical values be seen, and possibly referred to at intervals, a lecturer will be wise to consider the merits of a handout.

Type V tables obey no rules beyond that of clarity, for their nature must be determined by the particular point to be demonstrated. They may show aspects of computation (such as analyses of variance) that rarely need explicit presentation as part of a publication in an applied field but that require full explanation in an account of statistical methods. They may have a number of digits that even for Type II would be excessive. They can rarely be replaced by diagrams.

The five types may not always be unambiguously distinguishable, but the writer who keeps in mind their separate functions, as summarized in Table 1, will avoid inflicting on his readers tables and diagrams totally unsuited to his purpose.

TABLE 1—THE SUITABILITY OF TYPES OF TABLE TO CLASSES OF SCIENTIFIC WRITING

Nature of Presentation	*Type of Table*				
	I (*Support*)	*II* (*Reference*)	*III* (*Mixed*)	*IV* (*Lecture*)	*V* (*Didactic*)
A (Research)	++	?	++	−	?
B (Teaching)	+	−	+	−	++
C (Oral)	++	−	?	++	−
D (Popularization)	++	−	?	+	−
E (Archives)	?	++	+	−	−
F (Methodology)	+	+	+	−	++

Presentation of tables (and diagrams) is to some extent a matter of taste. Two statisticians or two biologists will not necessarily agree on the most effective way of arranging a particular table. Three rules should be generally agreed :

(i) The facts must be correct;
(ii) The arrangement should be designed with a view to easing comprehension by a reader rather than to pleasing the author;
(iii) The information content should be high.

By (iii), I mean that excessive digits should be removed, that factors common to all or most treatments should be stated at the head or in a footnote and not repeated on every line, that the order of entries should facilitate logical comparisons, and that unnecessary distractions to eye and brain should be avoided.

5. Measures of Precision

Almost every statistical table summarizing research in an applied biological science needs to indicate the precision of the results stated. Omission may sometimes be permissible for Type IV or even for Type II, but not for Types I and III. The critical reader, or the user of reference material, is entitled both himself to confirm statements made in the text and to construct for himself types of comparison other than those a table presents directly. There are five distinct types of indicator :

(i) Standard deviation per plot or experimental unit (SD);
(ii) Standard error of means (SE);
(iii) Standard error of differences of pairs of means (SEdiff);
(iv) Least significant, or critical, difference (LSD);
(v) Results of a multiple range test (MRT);

Of these, the SD is valuable as a record of the intrinsic precision of an experiment, for comparing with other experiments. In some respects, its square (the variance) is preferable. Some statisticians prefer to quote the coefficient of variation, often I suspect without much thought for its implications. None of these is directly usable in interpreting the current results. The SE, SEdiff, and LSD are to some extent interchangeable, since knowledge of any one permits the other two to be easily calculated. Although the SEdiff and the LSD might seem to be the most immediately applicable, they have two disadvantages : examination of a table may call for a greater number of different values of these than of SEs, and comparisons of combinations of means may require first a backward step to reconstruct the SE and then a forward step to form a new SEdiff. The LSD, of course, will relate to the author's choice of level of significance. There is an understandable reluctance to state the LSD when no differences exceed it, a further reason for preferring the unobjectionable SE or SEdiff.

I believe that, for the vast number of tables that relate to patient health, crop yields, animal growth or other characteristics, the most appropriate measures of precision, or of statistical variability, are the standard errrors of means. Whatever other indications of statistical variablity may be shown in a table, the relevant values of SE, SEdiff, or

LSD should *always* be there, although my peference is strongly for the SE, I accept that a case can be made for either of the others. Inclusion of one of these indicators is especially important for tables of Types I-III, uses of which are not confined to the current text. Standard errors are essential to any user of a table who wants to look at groupings of treatments (Any indication that the tricyclics as a group differ from the other compounds under test?), to compare a treatment difference with a cost (Is the estimated gain in crop yield from an extra herbicide application enough to pay for itself), or to combine evidence on the same comparison from different experiments (How strong is the evidence that responses to anti-hypertensive drugs differ in Britain, Italy, and Sweden? Does an advantage of one wheat variety over another appear statistically significant as an average over 10 experiments in 2 seasons, even though in no one experiment was the difference significant?). In rapid reading of a table, the approximate relations

$$\text{SEdiff} = 1.5 \times \text{SE}$$
$$\text{LSD (0.05)} = 3 \times \text{SE}$$

commonly suffice for immediate assessments, with more exact calculation following where proper estimation seems necessary. Of course, the number of degrees of freedom for error should be stated whenever it is small, say 15 or fewer; the probability level used for any LSD should be stated.

Indicators of precision based on MRT have increased in popularity in recent years. They have the attraction of giving immediate answers on tests of significance; very approximate SE values can be guessed from them, but the process is tedious and ought not to be needed. In many, perhaps most, tables chief interest attaches to estimation of differences among treatment means, with tests of significance being little more than statisticians' jargon to aid the drawing of a broad distinction between those that are large enough to be worth follow-up and those that on present evidence can be neglected. Does anyone really believe that the difference between population means for two logically distinct treatments is zero rather than perhaps 0.001? Certainly many authors (and editors) give far too much attention to tests of significance. Portrayal of their results by asterisks or by alphabetic codes tends to clutter a table with symbols that distract attention from the quantities tabulated. The responsibility for this pollution is shared among statisticians who have come to regard testing significance as their primary task, other scientists who feel nakedly exposed to criticism unless they can quote the conclusions from a ritual set of tests, editors who do not consider a paper fit for publication unless it is decorated with asterisks or other codes, and some software writers who exercise their ingenuity in outputting tables already

equipped with these evidences of completeness. A column of means that all receive the code "a" as an indicator from a multiple range test conveys almost no information on precision. Possibly there are situations in which a test such as the popular Duncan's multiple range test or DMRT (Duncan 1955) is harmless; undoubtedly there are many situations where it is bad. If it is to be used, let it be restricted to types of comparison for which a good argument of appropriateness can be made.

6. Multiple Range Tests

I have made a collection of examples to illustrate small and large misuses of range tests. Although I would like to include a selection here, there are two difficulties. First, the most striking examples are large tables; each would require much space for the table itself, background information, and my own suggested presentation. Secondly, my collecting has concentrated on small regions of Alphabet Jungle, and I am reluctant to exhibit specimens from only two or three publications. Bryan-Jones and Finney (1983) have discussed a few specific examples from one source, but have missed the real prizes. Here, instead, I shall try to establish some general principles on when multiple range tests are to be avoided :

(i) *Variety trials.* One feature of different cultivars of a crop plant is that they differ. They may differ in practically every measurable characteristic, although for many the reality of a difference would be apparent only from averages of immense numbers. A non-significant difference is commonly a comment on the replication and intrinsic precision of the experiment rather than on the true characters of the cultivars. If each entry in a table of mean yields has letters purporting to indicate which pairs are different, no useful information is added. A single standard error at the foot of the table, say 0.36, informs immediately that the true difference between any pair may be wrong by as much as 1.0 in either direction, as well as being valuable if results from eight of the cultivars are to be averaged with those from four other experiments.

(ii) *Regression.* If a regression has been shown as significant, tests of significance on differences between successive levels usually are pointless. For example, if analysis of an experiment on several concentrations of a drug shows clear evidence of linear regression of a response on log concentration, there is little sense in also claiming absence of difference between 0.5% and 1.0%. The only reason for lack of significance is insufficient precision of individual means: on the evidence available, I would confidently assert that a difference exists between 0.65% and 0.67% but that only a very large

and precise experiment would demonstrate this. There can be exceptions. A scientist familiar with the topic may be prepared to consider that at certain levels the response curve becomes perfectly flat, or that it rises to a maximum and then declines.

(iii) *Factorial design.* If an experiment has factorial structure, a listing of all treatment combinations and presentation of significance tests among their means is unlikely to be useful, and is certainly not the first step to be tried. Yet this is whats ome standard programs produce. Comparison between yields of a crop for levels of fertilizer corresponding to levels 140, 30, 0 and 0, 0, 30 of nitrogen, phosphate, potash is not a very profitable line of enquiry. If an experiment has been "planned" so that only this kind of comparison is possible, its design is bad; if these are two treatments from a $3 \times 2 \times 2$ factorial experiment, the difference between them tells almost nothing that can be usefully interpreted.

(iv) *Main effect and interactions.* The most useful interpretation of a factorial experiment is likely to be in terms of main effects and selected components of interaction (sometimes main effects of one factor, and effects of a second at each level of the first). All tests made will be based on the same error mean square. There must surely be no wish that a conclusion on factor *A* shall be influenced by the form of test on factor *B*. This introduces some complication into the question of whether error rates per experiment, per factor, or per comparison are being used, a topic on which much has been written without great profit for the statistician or experimenter whose aim is to make scientific inferences from data.

(v) There are other situations analogous to (i), for example comparisons among many different pesticides or herbicides.

Of course, there can be situations where, initially at least, interest lies equally with every comparison among treatments because no pattern among them is known in advance. This is unlikely to be absolutely true, but it may rspresent a reasonable approximation to the state of mind of an experimenter. A multiple range test may then be useful in drawing attention to major differences. However, in a well-replicated experiment, this step may be immediately followed by a recognition of a relevant treatment structure or by an admission that differences among "good" treatments are far more worth reporting than differences among "poor" treatments. Consequently, the interpretative emphasis may change and even in this situation the formal multiple range test may be of transient interest. A procedure that may occasionally be worth having at the stage of statistical analysis does not necessarily have to be exhibited at the stage of publication.

Possibly other classes of tabulation in which MRT is definitely bad can be identified. I am chiefly concerned to emphasize the positive recommendation that standard errors are almost always needed; once they are inserted, the case for any additional indications of precision or statistical significance will disappear from many tables. My own experience leads me to a very firm outlook. I like to have the SE. I can live with the LSD or the SEdiff. Despite frequently thinking about the matter during the past 30 years, I have never yet been able to imagine a situation in which the DMRT, or any other multiple range test, is the slightest help. If we can stop using them, we shall remove from our tables a lot of unnecessary ink that distracts the eye and obscures the meaning.

Scientists have always been reluctant, in my view rightly, to permit statisticians to dictate how they shall present their results, although I think they are often grateful for our advice. Unfortunately, some popular packages for statistical analysis of experiments have incorporated Duncan's multiple range test as a standard part of the output. Not only do the software writers produce this as an ornament on a listing of means for all treatment combinations, whatever the factorial structure : they seem also to be under the illusion that the questions answered by the test bear some relation to those that should interest the scientist who conducted the experiment, and who carefully chose a factorially structured set of treatments. The consequences are at best confusing, at worst disastrous. In some circles, the procedure has become so popular that "to DMRT" has almost become a verb to be used by a biologist in his request to a statistician, or by a statistician in his instructions to a technician. Yet it amounts to insidious censorship of the data, for the reader is prevented from answering any questions that were not in the mind of the author of a table. Today there is a danger that scientists will permit standard computer packages to dictate the form of tables for publication—yet surely a computer package is even less desirable than a statistician as a dictator of style in scientific writing !

REFERENCES

Bryan-Jones, J. and Finney, D. J. (1983) : On an error in "Instructions to Authors", *Hortscience*, **18** : 279-282.

Duncan, D. B. (1955) : Multiple range and multiple *F* tests, *Biometrics*, **11** : 1-42.

Tufte, E. R. (1983) : *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, Connecticut, USA.