UNSOLVED PROBLEMS OF EXPERIMENTAL STATISTICS*

JOHN W. TUKEY
Princeton University

IT WOULD not be misleading to suggest that there is really only one unsolved problem of experimental statistics: "How can we recognize the problems of experimental statistics?" We can recognize a good many unsolved problems by accident, but we probably miss many important ones for far too many years. Difficulties in identifying problems have delayed statistics far more than difficulties in solving problems. This seems likely to be the case in the future, too.

Thus it is appropriate to be as systematic as we can about unsolved problems. Any system may be a start toward, or even a partial solution of, this problem of recognition. I shall try to do this by stating first some principles and then some consequences. I shall strive to phrase all these principles as generally as possible, in the hope of prolonging their useful life.

A discussion of examples of these 18 general principles will set forth a certain number of unsolved problems, while a list of 51 provocative questions poses many more. (This list is admittedly and intentionally incomplete.) The account closes with a discussion of the possibility of orienting experimental statistics toward problems rather than techniques.

SOME GENERAL PRINCIPLES

If we feel that the detailed problems of experimental statistics arise from the interaction of certain general principles among themselves and with classes of experiments, it is reasonable to try to state and illustrate some of these principles. Before stating the hypergeneral principles on which these general principles hang, we need to explain the sense in which three terms, *ends*, *areas* and *considerations* will be used there and in the sequel.

By an *end* we refer to real purposes of the user of the statistical technique. These purposes are often unformulated, and their partial formulation often requires the statistician to "psychoanalyze" his client (in the writer's view this is one of the most important functions of the statistical consultant!). An *immediate end* is a formalized (and almost certainly partial) end such as to describe an appearance (e.g., by a point estimate), to make a test of significance, to make a decision, or to reach a confidence statement.

* Prepared in connection with research sponsored by the Office of Naval Research. Presented to the American Statistical Association and the Biometric Society 28 December 1953.

An *area* is a class of situations with qualitatively similar data, such, for example, as the class where two sets of observations are presented for the comparison of the "typical" values of the corresponding populations (means, medians, and the like serve as "typical" values). Within an area, different techniques are competitive. Within an area, the historical, evolutionary, and logical relations of different techniques are relatively clear.

A *consideration* is a recognition that the world may very well be more complex, annoying, and difficult than our earlier techniques had supposed. Thus we might admit—nay, even take into consideration—the possibility that we did not know the variance, that the distribution might not be normal, that a certain fraction of the observations are affected by blunders, etc.

The four hypergeneral principles, which may seem harmless until we come to their consequences, run as follows:

(*A*) Different ends require different means and different logical structures.

(*B*) In each area, statistical method must and does evolve, mainly by adding *both* immediate ends *and* considerations.

(*C*) While techniques are important in experimental statistics, knowing when to use them and why to use them are more important.

(*D*) In the long run, it does not pay a statistician to fool either himself or his clients.

We have one hypergeneral principle about logical structure, two about statistical method, and one about statisticians. The last may seem to be of smallest scope, but when we consider matters carefully, we see that (*A*), (*B*), and (*C*) all follow from (*D*). To insist on one means or one logical structure for different ends, or to feel that there is *a* solution to the problems of method, are obvious attempts of the statistician to fool himself.

Clearly, one very general consequence is this: "This complexity of experimental statistics will clearly increase."

Reducing the generality somewhat, we list some consequences of (*A*), (*B*), (*C*), and (*D*) which are themselves general principles:

(*A1*) Statistics needs constantly to recognize new ends for which it should try to furnish new means *and* new logical structures.

(*A2*) Statistics needs to avoid over-unification, while encouraging coordination.

(*A3*) Statistical methods should be tailored to the real needs of the user.

(*A4*) Statistics needs continually to compare its own logical structures with the logical structures currently used or being put into use by science, engineering, business, and military administration, and other fields.

(*B1*) In any area of statistical method, analysis cannot be usefully considered alone for more than a limited time; after a time appropriate to the area, design must be brought in.

(*B2*) There are normal sequences (patterns) of growth in immediate ends.

(*B3*) There are normal sequences (patterns) of growth in considerations.

(*B4*) Growth in immediate ends can sometimes be neglected, but growth in considerations is almost never to be neglected.

(*B5*) At any one time, different areas of statistical methodology will be in different states of evolution, both in immediate ends and in considerations.

(*C1*) Competitive statistical techniques indicate a need for manuals of "when to choose which" and not just selection of "the best" technique.

(*C2*) Statisticians owe their clients help in choosing wisely between high confidence in a short inference and low confidence in a long inference.

(*C3*) Techniques of evaluating both the isolated experiment and history down to date will continue to be useful.

(*C4*) "What should be done" is almost always more important than "what can be done exactly." Hence new developments in experimental statistics are more likely to come in the form of approximate methods than in the form of exact ones.

(*D1*) Statisticians must face up to the existence and varying importance of systematic errors.

(*D2*) Statisticians have an obligation to clarify the foundations of their techniques for their clients.

(*D3*) Statisticians should be honest and expository about the relation of precise "assumptions" and exactly "optimum" solutions to real situations.

(*D4*) In every statistical area, we almost certainly need methods admitting one more nuisance parameter, methods of one higher level of robustness and de-parametrization, methods with both of these desiderata.

(*D5*) Statistics must continually study the behavior of its techniques when their conventional assumptions are *not* true.

ILLUSTRATIVE EXAMPLES

I will try to illustrate these principles by discussing particular problems of experimental statistics which show their impact. These examples are not intended to be an exhaustive list. In the light of general principle (*C*), a problem in experimental statistics is not solved by the existence of a mathematical statistical paper showing how to find a solution, or even by the existence of a technique with tables. There is needed an understanding of when and why to use the technique, and this understanding must be spread through a certain minimum number, sometimes small and sometimes large, of experimental statisticians. Thus we may, and should, discuss as unsolved problems some which others may consider as already solved.

(*A1*) *Statistics needs constantly to recognize new ends for which it should try to furnish new means and new logical structures.* A very good illustration of this principle is provided by recent developments in connection with the problem of multiple comparisons. Where one immediate end grew a few years ago, three immediate ends flourish today and promise to flourish for a long time. These three are:

(*1*) The immediate end of providing increments to the store of established knowledge. This to be done by the analysis of existent data with control of the error rate. The analysis to be formulated in confidence or significance statements (*cf.* Tukey [35, 36, 37], Duncan [11, 12, 13] and others).

(*2*) The immediate end of providing protection against too bad a selection among candidates. This to be done by a sequential design of measurement. The result to be selection of the apparently leading candidate when the "stop rule" takes effect. (*cf.* Bechhofer, Dunnett, Sobel [1, 2, 14]).

(*3*) The immediate end of minimizing, in some sense, the sum of the costs of experimentation and the costs of poor choice. This is to be done by a sequential design of measurement. The result to be selection of the apparently leading candidate when the "stop rule" takes effect (*cf.* Grundy, Healy, and Yates [40, 41], Sommerville [31]).

In my judgment, there will be a continuing place for all three immediate ends. To a reasonable extent these places correspond to the terms "basic research," "developmental research," and "operations research," [cp. 22].

This problem of multiple comparisons is still unsolved as a problem of multiple comparisons, because the necessary minimum numbers of experimental statisticians have not yet acquired a working understanding of the new immediate ends involved, or of when which technique is appropriate. Analogous problems, involving immediate ends which differ in analogous ways, are to be expected in more areas of statistics.

(*A2*) *Statistics needs to avoid over-unification, while encouraging coordination.* It is now known to mathematical statisticians that all the currently routine modes of statistical technique—significance statements, point estimates, confidence statements, etc.—can be formulated as decision problems. There is a tendency in the air to do so to an increasing degree. This *may* be good mathematical statistics, because it *may* encourage the interchange of useful mathematical techniques among the modes. (We are likely to see in due course whether or not this is true.) But it would surely be very *bad* experimental statistics to treat all these modes in too unified a way. For then some experimental statisticians might be led to forget whether their clients wanted (explicitly or implicitly) a decision or a confidence statement, whether they had done the experiment as a basis for immediate action or as a contribution to knowledge. What more important matter could be forgotten by any experimental statistician?

In almost every area of experimental statistics, there is a problem of providing enough *different* methods to meet the user's needs.

(*A3*) *Statistical methods should be tailored to the real needs of the user.* In a number of cases, statisticians have led themselves astray by choosing a problem which they could solve exactly but which was far from the needs of their clients. They could have chosen a problem closer to their client's needs at the price of an approximate solution. In most of these cases, tailoring the statistical method to the real needs of the client would have meant, and still means, giving up exactness for the sake of usefulness. Realistic assessment of value must urge us to make such "deals" freely and frequently.

The broadest class of such cases comes from the choice of significance procedures rather than confidence procedures. It is often much easier to be "exact" about significance procedures than about confidence procedures. By considering only the most null "null hypothesis" many inconvenient possibilities can be avoided. If the varieties are not different they cannot interact with fertilizers or blocks. If the treatment has no effect, we do not have to be concerned with how its effect varies with the weight or health of the animal or child. And so on—and on. In these examples, it will be clear to many that we are dodging substantial issues.

But throughout experimental statistics there are many areas with significance procedures but without confidence procedures. Almost every one of these areas needs one or more *rough* confidence procedures. Rough procedures will be adequate because the assumptions are not likely to be closely true, so that the probability statements need not follow precisely from the assumptions either. One or more, because techniques based on alternative assumptions give both greater freedom of action and greater confidence in results to the analytical statistician. Here are many unsolved problems in experimental statistics!

At another level of unsolution are the problems where the approximate mathematical statistics has been done, but no use has been made of the results. One outstanding example is the computation by Haldane [19] of the effect of non-normality on the variance of the estimated correlation coefficient. Who has put this to use? Yet it surely is enough to support an empirical robustification procedure involving an effective number of pairs of observations. There must be many more examples like this, where the results have not been carried through to practical usability.

(A4) Statistics needs continually to compare its own logical structures with the logical structures currently used or being put into use by science, engineering, business, and military administration, and other fields. We can indicate an unsolved problem here which is not likely to be solved in the near future. This is the problem of formalizing some further part of the process of developing new scientific concepts and new scientific theories. Only the most elementary steps in this process have been formalized (in terms of the analysis of conventional types of experiments, of the testing of goodness of fit, and the like). Undoubtedly some, at least, of the less elementary steps can be formalized, but how? And which ones?

This is a vague and diffuse problem, but it is a very important problem indeed. Some would construe it as a problem for philosophers, but I feel that it will require quantitative philosophers (that is, experimental statisticians).

(B1) In any area of statistical method, analysis cannot be usefully considered alone for more than a limited time; after a time appropriate to the area, design must be brought in. The second and third types of multiple comparison procedures cited above (*A1*) furnish an excellent example of the need for design. For the immediate ends involved the only action, once the measurements are made, is to take the seemingly best candidate. That this is reasonable is, and has been, clear to all. Even a very moderate degree of sophistication was barred from these

situations until the question of when to stop taking measurements was introduced. There must now be many similar cases in other areas today where design considerations have not yet been properly introduced.

(*B2*) *There are normal sequences of growth in immediate ends.* One natural sequence of immediate ends follows the sequence:

(1) Description
(2) Significance statements
(3) Estimation
(4) Confidence statement
(5) Evaluation

In the case of a double binomial the successive levels are illustrative by the sequence of statements.

(*1*) The percentage of success observed among A's was higher than among B's.
(*2*) The percentage of success among A's was significantly greater than among B's.
(*3*) The observed percentage of success among A's exceeded that among B's by a difference of 0.28 in logits. (Or, perhaps, by 15 per cent.)
(*4*) The difference in logits corresponding to the increased percentage of success in A's as against B's is between 0.18 and 0.43 with 95 per cent confidence. (Between 10 per cent and 22 per cent with 95 per cent confidence, perhaps.)
(*5*) Considering both this experiment, and all the observations reported by Smith, Jones, Brown, Robinson, and their coworkers, the indicated difference in logits lies between 0.32 and 0.36 with 5 per cent diffidence (the difference in per cent lies between 17 and 19, perhaps).

The order of (*2*) and (*3*) is not nearly so well defined as that of any other pair. In some areas, and to some experimental statisticians either order would be wrong. We have chosen this order for definiteness and not with sureness.

In the actual case of the double binomial, almost every experimental statistician can handle (*1*), (*2*), and (*3*) easily. Some are not perturbed by (*4*) and of these most but not all can handle (*4*) correctly. No one, so far as the writer knows can treat (*5*) adequately. In other areas we may stop at level (*1*), at level (*2*), at level (*3*), or at level (*4*), but in almost every case there is a next level which represents an unsolved problem.

How to operate at level (*5*) seems to represent an unsolved problem

in many areas. It is a real and important problem, and one whose solution should not be approached flippantly or lightly. Either the classical example of the charge on the electron (as of 1938) or the current example of the heat of sublimation of carbon (which has not improved during the last 25 years) shows that the proper evaluatory answer may be: "The available determinations fall into two systematically different groups, which correspond to values between A and B and between C and D, respectively, and which we are confident cannot be brought into agreement without the introduction of a new systematic adjustment." How many other unusual (from the point of view of formal statistics as found in the books) kinds of conclusions are reasonable in evaluation of all available data? This is not an easy question, but its solution (at least its partial solution) is a prerequisite to that of any problem of evaluation.

There are, of course, other normal sequences of immediate ends, leading mainly through various decision procedures, which are appropriate to development research and to operations research, just as the sequence we have just discussed is appropriate to basic research. (Here "There are, of course" means "There must be! We are sure they exist, but we cannot specify them today.")

(*B3*) *There are normal sequences of growth in considerations.* The area of comparing the typical values of two populations with aid of a sample drawn from each illustrates a customary sequence of evolution in considerations quite nicely. The sequence runs:

(*1*) Normal populations of equal and known variance.
(*2*) Normal populations of general (i.e., probably unequal) and known variances.
(*3*) Normal populations of identical but unknown (but estimated) variance.
(*4*) Normal populations of general and unknown (but estimated) variances.
(*5*) Symmetrical populations of unknown shape and unknown but equal variance.
(*6*) Symmetrical populations of the same unknown shape but general and unknown variances.
(*7*) Symmetrical populations of unknown shapes and variances.
(*8*) Populations of unknown but equal shape and variance.
(*9*) Populations of the same unknown shape and unknown and general variances.
(*10*) Populations of general and unknown shapes and variances.

Here we have exemplified the growth in considerations like these:

(*a*) The scale of the populations might be different.
(*b*) The variance might not be known.
(*c*) The symmetrical populations might not be normal.
(*d*) The populations might not have the same shape.
(*e*) The populations might not be symmetrical.

It is by considering such unpleasant possibilities that we sharpen our techniques and strengthen our understanding.

The normal distribution suffices for levels (*1*) and (*2*), while level (*3*) requires Student's *t*. The next level, (*4*), provides the Fisher-Behrens problem, while (*5*) seems to be the likely end of the direct application of Wilcoxon-Walsh [38–39] procedures (so far only applied to the matched observation case). Beyond this point the *terra* is rather *incognita*, but we may note that through level (*7*) we need to make no distinction between medians and means, while simple rank order procedures are exact through level (*8*).

Not only does this area—and remember that it is one of the most carefully worked over of all areas—provide a good example of a normal sequence of growth in considerations, but it also provides many examples of unsolved problems. The Fisher-Behrens problem arises quite early, at only level (*4*) in the list, yet today the Fisher solution is known not to be unique [33], even in the domain of fiducial probability, while the Aspin-Welch solution may or may not correspond to an exact solution as well as an asymptotic one. What should a poor experimental statistician do?

Who has good-looking solutions for the problems posed by (*5*), (*7*), (*9*), or (*10*)? Who knows how the solutions for level (*4*) just mentioned behave as to error rate when (*5*), (*6*), or (*7*) represents the facts? How do the solutions for level (*4*) behave as to power when either (*4*) or (*3*) represents the facts? And the reader can add many more.

The foreseeable, normal growth in considerations will provide unsolved problems for a long time to come in almost *every area* of statistics.

(*B4*) *Growth in immediate ends can sometmes be neglected, but growth in considerations is almost never to be neglected.* We can use the two-sample area to illustrate this principle also. If we had a clear and reasonable solution to the Fisher-Behrens problem, very few experimental statisticians would dare ignore it. But many are content to teach significance testing without confidence procedures. (The young chemist who can analyze the variance of Latin squares and snatch out single degrees of freedom with zest and ease, but who cannot use Student's *t*

to set confidence limits on $A-B$, because no one ever mentioned it to him, is a poor witness to the teaching of chemists by statisticians!)

(B5) At any one time, different areas of statistical methodology will be in different states of evolution, both in immediate ends and in considerations. We have only to contrast the two-sample area with the $m \times n$-contingency-table area or the correlation-coefficient area with the measures-of-nonnormality-for-time-series area to find application of this general principle.

(C1) Competitive statistical techniques indicate a need for manuals of "when to choose which" and not just selection of "the best" technique. Our discussion of the two-sample area should have made it clear that what is needed here is a guide to the various techniques explaining why and when to use them. No selection of a single "best" technique is going to be satisfactory.

Another widely separated area which illustrates the principle nicely is the response maximization area. Here we have a spectrum of suggestions from the carefully thought-out "circle and bee-line (possibly repeated) and then survey" technique of Box and Wilson [5] to the creeping technique of Friedman and Savage [16] and the sophisticated but so far one-dimensional technique of Robbins and Monro [30]. I am sure that all of those named have their place, as do, no doubt, some of the intermediate points in the spectrum. I have, indeed, some idea of where these places are. But I would like to know far more precisely where these places are and why. (You couldn't possibly sell me a *single* best method!)

(C2) Statisticians owe their clients help in choosing wisely between high confidence in a short inference and low confidence in a long inference. In the analysis of three and more way analyses of variance, there arises the problem of choosing the correct error term (e.g. Goulden [17]). This is the first big problem in the analysis of variance, and one that is still very effective in separating the statisticians from the children. If one classification is years, one choice can be put into words as follows: Will you have differences in average performance averaged over *these particular* years, with narrow confidence limits, or will you have differences in average performance, averaged over a *population of years* of which these years are a sample, with much broader confidence limits. With regard to this particular example, most experimental statisticians are clear and effective. Thus, it may be a solved problem. But in many other areas the corresponding problem is not only unsolved but unposed!

Some have queried the use of "short" and "long" in this context, and

have tried to relate this choice to that of the proper "breadth" of foundation (the advantages of sufficiently broad basis of inference have, of course, been ably discussed by Fisher [15, Section 39]). It is important to avoid possible confusion in this regard. Considerations of breadth arise during the design of an experiment, while considerations of length arise in its interpretation. Thus an experiment to compare certain psychological characteristics within brother-sister pairs would be broadened as to foundation if changed from 50 pairs drawn from Indiana to 5 subgroups of ten pairs each from 5 geographically and culturally separated areas. For *either* experiment, there will be a problem of length of inference! Will we make statements about the average over the 50 pairs of perfectly measured differences, or shall we make statements concerning the average differences in larger populations of which these 50 pairs, or these 5 sets of 10 pairs are a sample or samples? The two questions are quite separate.

(*C3*) *Techniques of evaluating both the isolated experiment and history down to date will continue to be useful.* There are many experimental procedures that involve either the regular measurements of control specimens or the regular use of special calibration procedures. After a new calibration, should we use the old calibration? Should we use only the new calibration? Or should we combine old and new values? With what relative weights? This is a recurrent problem, one whose solution might improve measurement accuracies per dollar in a wide variety of applications. But who has the solution? or better "the solutions," because the path is long from the isolated group of occasional measurements to the production line producing measurements steadily. Different locations along this path will require different solutions. Work on this problem has undoubtedly been hampered by the tradition of the self-contained experiment. But many measurement procedures are far from self-contained experiments.

Like unto this first example is a second. Most procedures of statistical analysis today include a measure of spread in this particular experiment, be it an estimated variance, a total or mean range, or the mean square in a certain line of the analysis of variance. Usually there is past evidence as to the variability in question. In assessing the results of a particular experiment shall we use only the estimate from within the experiment? Only past history? Some combination of the two? Which combination?

This problem of how far to look back is widespread and unsolved. A solution might allow us to narrow the wide confidence limits that go with wide apparent variation and to widen the falsely narrow ones

which go with narrow apparent variation. This would equalize our exposure to error, and tend to let us make sharper statements on the average. Again the philosophy of "each experiment to itself" has stood in the way. But why should we allow this to go on? (Of course the philosophy of "each experiment to itself" is important, of course it must be widely used, but neither always or everywhere! Just another example of (*A2*) and (*C1*).)

(*C4*) *"What should be done" is almost always more important than "what can be done exactly." Hence new developments in experimental statistics are more likely to come in the form of approximate methods than in the form of exact ones.* Once upon a time the calculation of the first four moments was an honorable art in statistics. Then came those who could calculate the exact distributions of simple expressions. And because their results were "exact" they took over the place of honor. (Partly too, perhaps, because the moment calculators failed on occasion to transform their expressions wisely before calculating the moments.) And it came to be *infra dig* to find moments. In seminars one heard *A*'s achievement of calculating the first four moments for n's up to 12 belittled in comparison with *B*'s proof that the distribution tended to normality as n tended to infinity. Yet which result was more useful to the experimental statistician with experimental data for n equal to 5, 10, 20 or even 50—? Probably the first four moments.

If the moments had been on MacArthur's staff, their parting statement would have read "we shall return!" But when? I think that it is high time to bring the calculation of moments back to that high estate which it deserves. We shall always have to deal with messy expressions, whose exact distribution will be found by no one, at least for a long time. Moments may allow us to get on with the work. If they do allow us to do this, let us use them.

The variability of estimates of spectra of time series provides a case in point. Even with the normality assumption, the exact distribution is not going to be easily manageable. Yet the first *two* moments can be found, and found with very useful results. Considerable recent progress in the analysis of physical time series rests on those two moments [e.g. 27, 29].

(*D1*) *Statisticians must face up to the existence and varying importance of systematic errors.* The failure of the statistician to take sufficient cognizance of systematic errors has been in part an escape phenomenon. To a man looking hopefully for a way to shorten a confidence interval by 7 per cent of its length by ingenious devices, the thought of systematic errors which might make it twice as long comes as a severe shock,

and all men try to avoid shocks. Perhaps, too, the recent development of statistics in connection with the uncomfortable sciences like agriculture and biology—uncomfortable because *un*systematic errors tend to be so large—may have much to do with this. Only the sampling survey statisticians, with their recent treatment of "non-sampling errors" seem to be facing up to the existence of systematic errors.

What should experimental statistics as a whole do about systematic errors? Should we change from "95 per cent confidence" to "5 per cent diffidence" and impress on our clients that more diffidence has to be added because of systematic errors? Have we been overselling our clients on the confidence with which they should accept the results of our analyses? Is this why physics is the most-resistant of all the sciences to the penetration of statistics?

Some there will be who will claim that the old ways are good enough, since in comparative experiments the systematic errors tend to be very much smaller than in absolute experiments. Very much smaller, but not zero, is the answer. (The experimental statistician dare not shrink from the war cry of the analyst "Only a fool would use it, but it's better than we used to use!," but on the other hand, he dare not take the motto as a permanent excuse for sloppy methods). Here is a real unsolved problem of experimental statistics; What about systematic errors?

(*D2*) *Statisticians have an obligation to clarify the foundations of their techniques for their clients.* I have the impression that, at the time the analysis of variance was introduced, the practice of adjusting yields for the apparent fertility of blocks was, or would have been, regarded with suspicion—"cooking the observations." Yet the analysis of variance which is quite equivalent in its results, seems to have spread without opposition of this sort. Was this because the arithmetic was so complicated that the poor client didn't understand what was going on? I am sorely afraid that this was the case.

At the beginning, it may have paid the statisticians to fool their clients about the analysis of variance, but does it today? I give vent to a hearty "no!", feeling that many clients get far less out of such analyses than they should, because they don't understand what is going on. How many of your clients really understand what sorts of additive decompositions of the observations underlie the analyses of variance you proudly return to them?

How to explain to the client what the analysis of variance is about? This is surely a problem of experimental statistics. Even if I should know a large part of the answer, as I hope I do, it is an unsolved prob-

lem, since the answer is not at the finger tips of enough experimental statisticians.

In how many other areas are we losing by fooling our clients?

(*D3*) *Statisticians should be honest and expository about the relation of precise "assumptions" and exactly "optimum" solutions to real situations.* As an example here, let us take a field currently under development. Box and his coworkers have been, and continue to be, active in the development of designs for the estimation of all the zeroth, first, and second degree coefficients in a second degree response surface, where the response is a function of 1, 2, 3, 4, 5, etc., variables. In the process he is resting heavily on such "exact" concepts as "orthogonailty" and "estimating all coefficients with the same variance." He is well aware that, because of the way the designs are to be used, these "exact" mathematical properties are not likely to correspond to any physical realities, that, in any particular situation, there is no reason to believe that the "exactly optimum" design is appreciably better than any nearby design. But even if "exactly optimum" does not mean what it says, it may well mean "likely to be quite useful,", as in this case it does.

How many of the potential *users* of such designs will understand that "exactly optimum" doesn't mean what it says? All too few, and for the others we statisticians are likely to be to blame. We have pushed "optimum" procedures for one reason or another, without adequate warning about idealizations and the real world. As a psychologist once said when Mosteller discussed "inefficient statistics" before the Eastern Psychological Association, "inefficient statistics, but efficient statisticians"! How often do we miss the chance to have "non-optimal techniques, but optimal statisticians " apply to us?

Another example of the same sort looms large on the horizon. It concerns all of bioassay and much of the transformation of counted data (a subject about which there are whispers of new discussion). Little attention has been paid to gains or losses from "exact" maximum likelihood, minimum chi-square, or unbiased solutions of bioassay problems. Much attention has been spent in getting these "exact" solutions. Does it matter whether we use logits, probits, or anglits? How much does it matter? (On this there is some information.) What happens if a little non-binomial fluctuation creeps in? Have we been realistic about *anything* in this whole area? Clearly there are many unsolved problems of experimental statistics here.

(*D4*) *In every statistical area, we almost certainly need methods admitting one more nuisance parameter, methods of one higher level of robustness and de-parametrization, methods with both of these desiderata.* Here

we may turn the carpet back to see the dirt—it is a large carpet trying to cover much dirt. We have a reasonably wide variety of procedures for analyzing counted data which assume pure binomial variation. Contingency tables, chi-square, and ω^2 goodness of fit tests, Kolmogoroff-Smirnoff bounds on the population distribution, all-or-none bioassay, and so on. The list is long. Many of the techniques are important. *All of them* need procedures admitting the possibility of additional nonbinomial variation. We gave up long ago assuming that we knew the variance of yield of soy bean plots of given size—even though we had empirical data on it. We blithely assume that we know the variance of preparing a dilution and the variance of death among guinea pigs injected with a single dilution—we assume one to be zero and the other to be binomial! We would criticize the varietal trial without an internal estimate of error, yet we look silently on the bioassay without one.

Perhaps in part we have not attacked these problems because of their resemblance to those cited under (*C3*). Perhaps we have not attacked them because their consideration would disturb our clients' techniques or bring to light new sources of variation. But whatever the reasons, they do not seem valid to me today.

Here are many unsolved problems in experimental statistics.

(*D5*) *Statistics must continually study the behavior of its techniques when their conventional assumptions are not true.* I have touched on some minor examples of this principle. Let me cite a few major ones.

Many statistical techniques assume homogeneity of variance, each of them needs a related technique assuming inhomogeneity of variance. How do the present techniques stand up under homogeneity?

Many statistical techniques utilize a normality assumption almost exclusively as a means for predicting the stability of estimated variances. Each needs a related robustified technique which allows for the effects of non-normality on this stability. How do the present techniques stand up under non-normality?

Many discussions of efficiency of estimation assume an underlying normal distribution. Each needs related studies assuming suitably varied nonnormal distributions.

How many unsolved problems do we need?

SOME PROVOCATIVE QUESTIONS

In providing examples of the various general principles, I have indicated a number of unsolved problems of experimental statistics, but there are a few more at the tip of the tongue. In this section I shall seek

to provide a few more, mostly indirectly, by trying to ask some provocative questions.

(*1*) *What are we trying to do with goodness of fit tests?* (Surely not to test whether the model fits exactly, since we know that no model fits *exactly!*) What then? Does it make sense to lump the effects of systematic deviations and over-binomial variation? How should we express the answers of such a test?

(*2*) *Why isn't someone writing a book on one- and two-sample techniques?* (After all, there is a book being written on the straight line!) Why does everyone write another general book? (Even 800 pages is now insufficient for a complete coverage of standard techniques.) How many other areas need independent monograph or book treatment?

(*3*) *Does anyone know when the correlation coefficient is useful, as opposed to when it is used?* If so, why not tell us? What substitutes are better for which purposes?

(*4*) *Why do we test normality?* What do we learn? What should we learn?

(*5*) *How soon are we going to develop a well-informed and consistent body of opinion on the multiple comparison problem?* Can we start soon with the immediate end of adding to knowledge? And even agree on the place of short cuts?

(*6*) *How soon are we going to separate regression situations from comparison situations in the analysis of variance?* When will we clearly distinguish between temperatures and brands, for example, as classifications?

(*7*) *What about regression problems?* Do we help our clients to use regression techniques blindly or wisely? What are the natural areas in regression? What techniques are appropriate in each? How many have considered the "analyses of variance" corresponding to taking out the regression coefficients in *all* possible orders?

(*8*) *What about significance vs. confidence?* How many experimental statisticians are feeding their clients significance procedures when available confidence procedures would be more useful? How many are doing the reverse?

(*9*) *Who has clarified, or can clarify, the problem of nonorthogonal (disproportionate) analysis of variance?* What should we be trying to do in such a situation? What do the available techniques do? Have we allowed the superstition that the individual sums of squares should add up to the total sum of squares to mislead us? Do we need to find new techniques, or to use old ones better?

(*10*) *What of the analysis of covariance?* (There are a few—at least

one [10]—discussions which have been thought about.) How many experimental statisticians know more than one technique of interpretation? How many of these know when to use each? What are all the reasonable immediate aims of using a covariable or covariables? What techniques correspond to each?

(*11*) *What of the analysis of variance for vectors?* Should we use overt multivariate procedures, or the simpler ones, ones that more closely resemble single variable techniques, which depend on the largest determinantal root? Who has a clear idea of the strength or scope of such methods?

(*12*) *What of the counting problems of nuclear physics?* (For some of these the physicists have sound asymptotic theory, for others repairs are needed—cf. Link [21].) What happens less asymptotically? What about the use of transformations? What sort of nuisance parameter is appropriate to allow for non-Poisson fluctuations? What about the more complex problems?

(*13*) *What about the use of transformations?* Have the pros and cons been assembled? Will the swing from significance to confidence increase the use of transformations? How accurate does a transformation need to be? Accurate in doing what?

(*14*) *Who has consolidated our knowledge about truncated and censored (cf. [18], p. 149) normal distributions so that it is available?* Why not a monograph here that really tells the story? Presumably the techniques and insight here are relatively useful, but how and for what?

(*15*) *What about range-based methods for more complex situations?* (We have methods for the analysis of single and double classifications based on ranges.) What about methods for more complex designs like balanced incomplete blocks, higher and fractional factorials, lattices, etc.? In which areas would they be quicker and easier? In which areas would they lead to deeper insight?

(*16*) *Do the recent active discussions about bioassay indicate the solution or impending solution of any problems?* What about logits vs. probits? Minimum chi-square vs. maximum likelihood? Less sophisticated methods vs. all these? Which methods are safe in the hands of an expert? Which in the hands of a novice? Does a prescribed routine with a precise "correct answer" have any value as such?

(*17*) *What about life testing?* What models should be considered between the exponential distribution and the arbitrary distribution? What about accelerated testing? (Clearly we must use it for long-lived items.) To what extent must we rely on actual service use to teach us about life performance?

(*18*) *How widely should we use angular randomization* [*4*]*?* What are its psychological handicaps and advantages? Dare we use it in exploratory experimentation? What will be its repercussions on the selection of spacings?

(*19*) *How should we seek specified sorts of inhomogeneity of variance about a regression?* What about simple procedures? Can we merely regress the squared deviations from the fitted line on a suitable function? (Let us not depend on normality of distribution in any case!) What other approaches are helpful?

(*20*) *How soon can we begin to integrate selection theory?* How does the classical theory for an infinite population (as reviewed by Cochran [8]) fit together with the second immediate aim of multiple comparisons (Bechhofer *et al.* [1, 2, 14]) and with the *a priori* views of Berkson [3] and Brown [6]? What are the essential parameters for the characterization of a specific selection problem?

(*21*) *What are appropriate logical formulations for item analysis (as used in the construction of psychological tests)?* (Surely simple significance tests are inappropriate!) Should we use the method introduced by Eddington [32, pp. 101–4] to estimate the true distribution of selectivity? Should we then calculate the optimum cut off point for this estimated true distribution? Or what?

(*22*) *What should we do when the items are large and correlated?* (If, for example, we start with 150 measures of personality, and seek to find the few most thoroughly related to a given response or attitude.) What kind of sequential procedure? How much can we rely on routine item analysis techniques? How does experiment for insight differ from experiment for prediction?

(*23*) *How many experimental statisticians are aware of the problems of astronomy?* What is there in Trumpler and Weaver's book [32] that is new to most experimental statisticians? What in other observational problems like the distribution of nebulae (e.g. [23, 26])?

(*24*) *How many experimental statisticians are aware of the problems of geology?* What is there in the papers on statistics in geology in the *Journal of Geology* for November 1953 and January 1954 that is new to most experimental statisticians? What untreated problems are suggested there?

(*25*) *How many experimental statisticians are aware of the problems of meteorology?* What is there in the books of Conrad and Pollak [9] and of Carruthers and Brooks [7] that is new to most experimental statisticians? What untreated problems are suggested there?

(*26*) *How many experimental statisticians are aware of the problems*

of particle size distributions? What is there in Herdan's book [21] on small particle statistics that is new to most experimental statisticians? What untreated problems are suggested there?

(*27*) *What is the real situation concerning the efficiency of designs with self-adjustable analyses*—lattices, self-weighted means, etc.—*as compared with their apparent efficiency?* Meier [25] has attacked this problem for some of standard cases, but what are the repercussions? What will happen in other cases? Is there any generally applicable rule of thumb which will make approximate allowance for the biases of unsophisticated procedures?

(*28*) *How can we bring the common principles of design of experiments into psychometric work?* How can we make allowance for order, practice, transfer of training, and the like through specific designs? Are environmental variations large enough so that factorial studies should always be done simultaneously in a number of geographically separated locations? Don't we really want to factor variance components? If so, why not design psychometric experiments to measure variance components?

(*29*) *How soon will we appreciate that the columns (or rows) of a contingency table usually have an order?* When there is an order, shouldn't we take this in account in our analyses? How can they be efficient otherwise? Should we test *only* against ordered alternatives? If not, what is a good rule of thumb for allocating error rates? Yates [40] has proposed one technique. What of some others and a comparison of their effectivenesses?

We come now to a set of questions which belong in the list, but which we shall treat only briefly since substantial work is known to be in progress:

(*30*) What usefully can be done with $m \times n$ contingency tables?

(*31*) What of a very general treatment of variance components?

(*32*) What should we really do with complex analyses of variance?

(*33*) How can we modify means and variances to provide good efficiency for underlying distributions which may or may not be normal?

(*34*) What about statistical techniques for data about queues, telephone traffic, and other similar stochastic processes?

(*35*) What are the possibilities of very simple methods of spectral analysis of time series?

(*36*) What are the variances of cospectral and quadrature spectral estimates in the Gaussian case?

(*37*) What are useful general representations for higher moments of stationary time series?

Next we revert to open questions:

(*38*) *How should we measure and analyze data where several coordinates replace the time?* What determines the efficiency of a design? Should we use numerical filtering followed by conventional analysis? How much can we do inside the crater?

(*39*) *What of an iterative approach to discrimination?* Can Penrose's technique [28] be usefully applied in a multistage or iterative way or both? Does selecting two composites from each of several subgroups and then selecting supercomposities from all these composites pay? If we remove regression on the first two composites from all variables, can we usefully select two new composites from among the residuals?

(*40*) *Can the Penrose idea be applied usefully to other multiple regression situations?* Can we use either the simple Penrose or the special methods suggested above?

(*41*) *Is there any sense in seeking a method of "internal discriminant analysis"?* Such a method would resemble factor analysis in resting on no external criterion, but *might* use discriminant-function-like techniques.

(*42*) *Why is there not a clearer discussion of higher fractionation?* Fractionation (by which we include both fractional factorials and confounding) is reasonably well expounded for the 2^m case. But who can make 3^m, 4^m, 5^m etc. relatively intelligible?

(*43*) *How many useful fractional factorial designs escape the present group theoretical techniques?* After all, Latin Squares are $k$ths of a k^3, and most transformation sets do not correspond to simple group theory.

(*44*) *In many applications of higher fractionals, the factors are scaled—why don't we know more about the confounding of the various orthogonal polynomials and their interactions (products)?* Even a little inquiry shows that some particular fractionals are much better than others of the same type.

(*45*) *What about redundant fractions of mixed factorials?* We know perfectly well that there is no useful simple (nonredundant) fraction of a $2^2 3^3 4^1$, but there may be a redundant one, where we omit some observations in estimating each effect. What would it be like?

A number of further provocative questions have been suggested by others as a result of the distribution of advance copies of this paper and its oral presentation. I indicate some of them in my own words and attitude:

(*46*) *To what extent should we emphasize the practical power of a test?* Here the practical power is defined as the product of the probability

of reaching a definite decision given that a certain technique is used by the probability of using the technique. (C. Eisenhart)

(*47*) *What of regression with error in x?* Are the existing techniques satisfactory in the linear case? What of the nonlinear case? (K. A Brownlee)

(*48*) *What of regression when the errors suffer from unknown autocorrelations?* What techniques can be used? How often is it wise to use them? (K. A. Brownlee)

(*49*) *How can we make it easier for the statistician to "psychoanalyze" his client?* What are his needs? How can the statistician uncover them? What sort of a book or seminar would help him? (W. H. Kruskal)

(*50*) *How can statisticians be successful without fooling their clients to some degree?* Isn't their professional-to-client relation like that of a medical man? Must they not follow *some* of the principles? Do statisticians need a paraphrase of the Hippocratic Oath? (W. H. Kruskal)

(*51*) *How far dare a consultant go when invited?* Once a consultant is trusted in statistical analysis and design, then his opinion is asked on a wider and wider variety of questions. Should he express his opinion on the general direction that a project should follow? Where should he draw the line? (R. L. Anderson)

In closing these questions, it should not be necessary to remind the reader that neither in the last section of examples or in this section of provocative questions have we tried to suggest an order of importance for the unsolved questions suggested. We leave that to the reader.

TOOL BUILDING VS. PROBLEM SOLVING

To judge from published books and articles, experimental statistics has grown by finding tools somehow, and then running around using them. (This impression is undoubtedly *somewhat* inaccurate.) Why has experimental statistics not been more obviously concerned with problems? Partly, perhaps, because it is just beginning to get its growth. Partly, perhaps, because dealing with problems is difficult and likely to lead to approximate solutions. These are valid reasons, but not valid excuses.

As experimental statistics grows toward maturity, it surely should orient more toward areas rather than toward techniques. How much more may be a question. But an essential prerequisite to such reorientation is some picture of what are the areas. This picture will not spring forth full armed, but will come from much work and discussion. As an attempted trigger for this work and discussion, the next section presents a feeble first attempt at classification. Reader, can you do better?

A FEEBLE GUIDE TO AREAS

We shall set up with a digital classification, but without prejudice as to whether the classification provided by one digit is crossed with or nested inside that provided by another. The digits provided will usually not specify an area completely, but they will usually narrow the situation down to a small number of areas.

The first digit classification refers to the general end of the analysis as follows:

(The assessment of, or determination of a wise action in view of)

(*1*) Typical response
(*2*) Variability of response
(*3*) Distribution of response
(*4*) Concealed structures and their coefficients
(*5*) Control charts and other "spotting" procedures
(*9*) Miscellaneous

(If answers are expressible in simple or mixed cumulants, then the degree of these cumulants with respect to response variables is controlling. (*1*) contains cases of degree 1; (*2*) contains cases of degree 2; (*3*) contains cases of higher degree.) Under (*1*) are included regression coefficients as well as means, while correlation analysis considered as a study of predictability comes under (*2*). Contingency tables fall under (*1*), except when the issue is homogeneity, when they fall under (*2*). Factor analysis seems better placed under (*4*) than under (*2*), but structural regression, as practiced in econometrics, seems to fall most naturally under (*1*).

The second digit classification refers to the situation of measurement, and, in description at least, has to be subordinated to the first digit. It runs

(*-1*) Isolated (one or a few) responses, isolated (one or a few) variabilities, isolated (one or a few) distributions, etc.
(*-2*) Response curves or surfaces, variabilities as functions of environmental variables, etc.
(*-3*) Inverse responses (what environment(s) produces a given response), inverse variabilities, etc.
(*-4*) Response to nonenvironmental variable (e.g. time shape of pulses, distribution of grain sizes, power spectrum of time series.)
(*-9*) Miscellaneous

All of bioassay and sensitivity testing will of course be found in (*-3*).

Problems of maximization of response by altering quantitative variables fall best into (-*2*), since attempts to put them into (-*3*) as the search for that environment where the derivatives vanish seem unwise.

The third digit classification refers to the nature of the measurement, and is easy to apply, namely

(--*1*) Absolute measurements without calibration problems
(--*2*) Intermediate cases
(--*3*) Absolute measurements by comparison with a standard
(--*4*) Comparative measurements among a family without calibration problems
(--*5*) Intermediate cases
(--*6*) Comparative measurements among a family with the aid of standards
(--*9*) Miscellaneous

The conventional problems of bioassay fall in (--*3*), while sensitivity to explosion or breakage problems based on falling weights may fall in (--*1*). Conventional comparisons of varieties and fertilizers are usually thought to fall in (--*4*), but must, in many cases, fall in (--*5*).

The fourth digit expresses the kind of response considered, and is again easy to apply. The classes are:

(---, *1*) Directly measured responses
(---, *2*) Responses measured as slopes or regression coefficients
(---, *3*) Adjusted responses (as by covariance)

No examples seem to be needed.

The fifth digit specifies the nature of the response, as follows:

(---, -*1*) Measured response (on reproducible scale)
(---, -*2*) Scored or rated response (by judge or panel)
(---, -*3*) Counted (all-or-none) response
(---, -*9*) Miscellaneous

At the present, the impact of this digit on statistical technique is very noticeable. Should it remain so?

The sixth digit specifies the complexity of the response, as follows:

(---, --*1*) Single variate response
(---, --*2*) Bivariate response
(and so on)
(---, --*8*) Many variate response
(---, --*9*) Miscellaneous

Examples here are not needed.

The seventh digit describes the complexity of the environments considered, as follows:

(---, ---, *1*) Environment varied only randomly
(---, ---, *2*) Environment varied in one measured way
(---, ---, *3*) Environment varied randomly and in one measured way
(---, ---, *4*) Environment varied in two measured ways
(---, ---, *5*) Environment varied in a more complex manner
(---, ---, *9*) Miscellaneous

REFERENCES

[1] Bechhofer, Robert E., Dunnett, Charles W., and Sobel, Milton, "A two-sample multiple decision procedure for ranking means of normal populations with unknown variances," *Annals of Mathematical Statistics*, 24 (1953), 136 (abstract).

[2] Bechhofer, Robert E., and Sobel, Milton, "A sequential multiple decision procedure for ranking means of normal populations with known variances," *Annals of Mathematical Statistics*, 24 (1953), 136 (abstract).

[3] Berkson, Joseph, "'Cost-Utility' as a measure of the efficiency of a test," *Journal American Statistical Association*, 42 (1947), 246–55.

[4] Box, G. E. P., "Multifactor designs of first order," *Biometrika*, 39 (1952), 49–57.

[5] Box, G. E. P., and Wilson, K. B., "On the experimental attainment of optimum conditions," *Journal of the Royal Statistical Society*, B13 (1951), 1–45.

[6] Brown, George W., "Basic principles for construction and application of discriminators," *Journal of Clinical Psychology*, 6 (1950), 58–60.

[7] Brooks, C. E. P. and Carruthers, N., *Handbook of statistical methods in meteorology*. London, H. M. Stationery Office (1953).

[8] Cochran, W. G., "Improvement by means of selection," *Proceedings of 2nd Berkeley Symposium on Mathematical Statistics and Probability* (1951), 449–70.

[9] Conrad, Victor, and Pollak, L. W., *Methods in Climatology*, 2nd Edition, Cambridge, Mass., Harvard University Press (1950), pp. 459.

[10] DeLury, D. B., "The analysis of covariance," *Biometrics*, 4 (1946), 153–70.

[11] Duncan, David B., "A significance test for differences between ranked treatments in an analysis of variance" *Virginia Journal of Science*, 3 (1951), 172–89 (abstract in *Annals of Mathematical Statistics*, 22 (1951), 142).

[12] Duncan, David B., "On the properties of the multiple comparisons test," *Virginia Journal of Science*, 3 (1952), 49–67.

[13] Duncan, David B., "Multiple range and multiple F tests." Presented to meeting of American Chemical Society, 1953, also Technical Report 6a, Department of Statistics and Statistical Laboratory, Virginia Polytechnic Institute.

[14] Dunnett, Charles W. and Sobel, Milton, "On a multivariate analogue ot Student's *t*-distribution, with some tables for the bivariate case." *Annals of Mathematical Statistics*, 24 (1953), 492 (abstract).

[15] Fisher, R. A., *The Design of Experiments*, First Edition, Edinburgh, Oliver and Boyd (1936).
[16] Friedman, Milton, and Savage, L. J., "Planning experiments seeking maxima," Chap. 13 of Statistical Research Group, Columbia University, *Techniques of Statistical Analysis* (edited by Churchill Eisenhart, Millard W. Hastay, and W. Allen Wallis), New York, McGraw-Hill (1947).
[17] Goulden, C. H., *Methods of Statistical Analysis*, 1st Edition, New York, John Wiley and Sons (1939), especially Chapter XI, Section 6 (pp. 122 ff.); 2nd Edition 1952, especially Chapter 5, Section 13 (pp. 90 ff.).
[18] Hald, Anders, *Statistical Theory with Engineering Applications*, New York, John Wiley and Sons (1952).
[19] Haldane, J. B. S., "A note on non-normal correlation," *Biometrika*, 36 (1949), 467–68.
[20] Healy, M. J. R., "Decision between two alternatives—how many experiments," Paper at the third international biometric conference, Bellagio, September, 1953.
[21] Herdan, G., *Small particle statistics*, Amsterdam-Houston-New York-Paris, Elsevier (1953)
[22] Hitchman, Norman, "What is the mission of operations research?" *Journal of the Operations Research Society of America*, 1 (1953), 241–42.
[23] Limber, D. Nelson, "The analysis of counts of the extragalactic nebulae in terms of a fluctuating density field," Submitted to *Astrophysical Journal*.
[24] Link, Richard F., "Some Statistical Techniques Useful for Estimating the Mean Life of a Radioactive Source," Doctoral thesis, Princeton University, 1953.
[25a] Meier, Paul R., "Weighted means and lattice designs," Doctoral thesis, Princeton University.
[25b] Meier, Paul R., "Variance of a weighted mean," *Biometrics*, 9 (1953), 59–73.
[26] Newman, J., Scott, E. L., and Shane, C. D., "On the spatial distribution of the galaxies. A specific model," *Astrophysica Journal*, 117 (1953), 92–138.
[27] Panofsky, H. A., and McCormick, R. A., "The vertical momentum flux at Brookhaven at 109 meters," *Geophysical Research Papers*, 19 (International Symposium Atmos. Turbulence Boundary Layer) (1952), 219–30.
[28] Penrose, L. S., "Some notes on discrimination," *Annals of Eugenics*, 13 (1946–7), 228–37.
[29] Pierson, Willard J., Jr., *A unified mathematical theory for the analysis, propagation and refraction of storm generated ocean surface waves*. Department of Meteorology, New York University, 1952.
[30] Robbins, Herbert, and Monro, Sutton, "A stochastic approximation method," *Annals of Mathematical Statistics*, 22 (1951), 400–7.
[31] Somerville, Paul N., "Optimum sample size for choosing the largest of $k+1$ parameters," Paper at the Institute of Mathematical Statistics meeting Kingston, Ontario, 3 September 1953.
[32] Trumpler, Robert J., and Weaver, Harold F., *Statistical Astronomy*, Berkeley, University of California Press (1953).
[33] Tukey, John W., "Purposes of fiducial inference," Paper before the Institute of Mathematical Statistics, Minneapolis, 6 September 1951.

[34] Tukey, John W., "Allowances for various types of error rates," Paper before Institute of Mathematical Statistics, Blacksburg, 19 March 1952.
[35] Tukey, John W., "Multiple Comparisons," Paper before American Statistical Association and Biometric Society, Chicago, 28 December 1952.
[36] Tukey, John W., "Various methods from a unified point of view," Paper before Institute of Mathematical Statistics, Chicago, 29 December 1952.
[37] Tukey, John W., *The problem of multiple comparisons,* In preparation.
[38] Walsh, J. E., "Some significance tests for the median which are valid under very general conditions," *Annals of Mathematical Statistics,* 20 (1949), 64–81.
[39] Wilcoxon, Frank, *Some rapid approximate statistical procedures,* Insecticide and Fungicide Section, Stanford Research Laboratories, American Cyanimide Co., 1948.
[40] Yates, Frank, "The analysis of contingency tables with groupings based on quantitative characters," *Biometrika* 35 (1948), 176–81.
[41] Yates, F., "Principles governing the amount of experimentation in developmental work." *Nature,* 170 (1952), 138–40.