THE STATISTICIAN, THE COMPUTER AND EXPERIMENTATION*

D.J. FINNEY

The University, Edinburgh, Scotland

Introduction

To the mathematician, the interest of experimental design lies in the combinatorial problems presented by the search for and the enumeration of designs of different types. To the scientific investigator, the more important features of experimental design are those that concern the choice of design for a particular purpose and the integration of well-chosen designs into a good programme of investigation. The statistician needs to bridge the gap between the two.

In this paper, I discuss topics connected with the choice of design, giving special attention to the role of the statistician and to some of the ways in which computers can now contribute to the design and interpretation of experiments. I shall concern myself mostly with individual experments, though I recognize that what I say could be extended to series of experiments whether planned for simultaneous execution or as a sequence in time.

The problems I must discuss are no easier than those of combinatorial theory for design, but they require very different approaches. In most of them, neither questions nor answers can be exactly defined: their full consideration requires general experience of applied statistics, knowledge of the particular field of experimentation, and judgement. I shall indicate some general principles in designing real experiments; I hope also to make the reader aware of, even interested in, the difficulties of satisfactory definition and communication.

University courses on statistics, especially those intended for educating future professional statisticians, usually concentrate on formal mathematical theory, because this employs well-defined concepts and is well-documented for teaching and reference. Many of us who teach the courses and who write the text-books also spend much time in the practice of statistics, as consultants to academic colleagues or collaborators in investigations into various fields of science and technology. Yet we seldom transmit to students the benefits of the experience we have gained in years of applying statistical techniques to the real world. We instruct in the theory, and perhaps remember to tell our students that they too must have years of experience

*Based on a lecture delivered in the Medical University, Pécs, Hungary, in September 1970.

before they become competent practitioners! To some extent doubtless this is true. Nevertheless, I believe that one of the most urgent tasks for statisticians is to transform what we have had to learn by experience into a more systematic body of knowledge that will expedite learning by our students. We must avoid undue dogmatism, but we should where appropriate express forcefully our views on the responsibilities of the statistician and the manner of his involvement in research investigations; we should expose these views to critical discussion, seeking therefrom the emergence of an acceptable codification of practice.

2. Purposes of Quantitative Experiment

I shall restrict myself to experiments intended to produce comparative assessments, qualitative or quantitative; thus I exclude an experiment intended solely to demonstrate the feasibility of a chemical synthesis or the occurrence of a phenomenon. The obvious example is that of an experiment in which biological or physical entities are assigned to either of two treatments, a measurement (often conveniently termed a *yield*) is made on each, and the set of yields is used in assessing the relative merits of the treatments. The measurement is commonly a numerical value on a continuous scale, but can be a qualitative epithet ("survived", "dead", or "good", "damaged", "broken"). The intention to compare two or more treatments is essential.

At least three types of comparative experiment can be distinguished:

Pilot experiments,

Decision experiments,

Learning experiments.

A *pilot experiment* is conducted to provide preliminary information that aids the planning of definitive research. Its object may be to suggest which treatments can be omitted from the main experiment because they are useless, to indicate the form of a response curve so that levels of a quantitative factor can be well chosen, or to guide the judgement of replication from roughly estimated variance components. All these matters concern the statistician who advises on the definitive experiment. Moreover, if optimal utilization of limited resources is important, questions will arise about how much (if any) of these should be expended on a pilot experiment. The statistician may be asked whether improvements in design coming from expending 15% of resources on a pilot is likely to compensate for loss of replication later. The possible situations are so diverse that no well-defined theory exists: indeed, as far as I know even general principles have not been stated. Experiments for screeing a large number of treatments or materials, in order to select a few for a definitive experiment, may be included in this category.

Decision experiments are undertaken in order that a choice may be made between treatments for practical use. Obvious examples are the decision between three concentrations of a chemical reagent to be used in a manufacturing process and the decision whether or not to adopt a dietary supplement as a standard practice in pig rearing. For optimal planning, costs must be well specified, these including costs of experimentation in relation to experimental design as well as the value placed on increased production. The experiments are usually simple in treatment structure, because interest has been narrowed to a few possibilities. They can provide uncontroversial uses of subjective prior distributions, by which the initial uncertainty of the magnitude of an effect may be expressed, in order that an experiment may be planned to maximize the probability of a correct decision.

Learning experiments are those of greatest concern to general scientific research. Knowledge of any field of science advances by continued incorporation and interpretation of new information rather than by decisions that statements are or are not true. Interest usually lies in estimating parameters that measure properties of treatments with the maximum precision that resources allow rather than in making decisions or testing significance. For example, a decision experiment may seek the conditions that approximately maximize the ordinate to a response surface, but a learning experiment may be concerned with the whole form of that surface in a prescribed region.

In practice, these types are not always easily distinguished. Thus a programme of agronomic research may use a series of factorial experiments initially for the exploration of the dependence of production on the inter-relation of various conditions of crop growth; subsequently, results from the same experiments may decide what advise shall be given to farmers.

3. The Statistician

The statistician must be far more than the human instrument by which experimental yields are analyzed. He should be consulted on the selection of treatments and on the numbers of levels of different factors. His experience may help in the specification of the physical dimensions of units for treatment, the ages of animals, the judgement that the production of several different machines rather than only one be used as experimental material, and so on. He may assist in specifying what measurements are to be made: the purpose of an experiment determines one or two characteristics to be measured, but the desirability of making supplementary measurements often has statistical undertones. Additional measurements may improve precision (by covariance analysis or other form of standardization), may elucidate the developmental aspects of treatment effects, or may provide information of secondary importance at a small marginal cost.

4. The Computer

I asked a friend, who is experienced in animal experimentation, in what ways he thought computers ought to affect design in animal research. He replied "None", but subsequently admitted to some exaggeration. I am here interpreting the role of statistics in experimental design more broadly than in this brief conversation.

No badly designed experiment will be made good merely by analyzing its results on a computer. However, a computer makes possible calculations that otherwise are intolerably laborious. Hence, one may be able to adopt better designs because some aspects can be examined in advance. At the stage of analysis, more comprehensive calculations may improve interpretation of experimental results. I want to emphasize the discriminating use of computers, stressing the dangers that arise if they become masters instead of servants.

5. Requirements of Design

The internal structure of an experiment must permit inferences both valid and relevant to the circumstances of their use. The main requirement for internal validity of an experiment is the random allocation of plots* or units to treatments. Constraints on complete randomness are permissible, and often desirable, particularly those implied by the arrangement of an experiment in blocks ; within the residual freedom left by the constraints, random allocation is essential. Non-random allocation, seldom unavoidable or excusable, always introduces risks of biases that are the greater danger because they may pass unsuspected when results are examined. An experimenter who refuses to randomize must accept responsibilities for the untestable assertion that he is not distorting his results : yet the wish to depart from strict randomization is commonly based upon a belief that the nature of the allocation will affect the results.

The preparation and listing of the randomization required for an experiment is a task that can well be undertaken by a computer. When the treatments and structural constraints for the design have been specified, the computer should be able to refer to a store of random digits (or to a generator of pseudo-random digits) and to produce a listing that allocates treatments to plots. Moreover, the format should be planned to be most convenient for the experimenter who must implement the instructions. In an earlier version of this paper, I wrote that the programming required had not yet been systematically undertaken ; a few weeks later, I was delighted to learn that one of my own staff had now written such a program ! In the past, the time required for randomizing has sometimes been quoted as an excuse for not using independent randomizations for each of a series of experiments of the same simple design ; to-day, 500 random arrangements of one basic design can be produced with little effort, rapidly, and in a form suitable for immediate distribution to experimental sites.

*Hereafter, the word *plot* is used for the ultimate unit to which a treatment is applied, whether it be an area of land (the natural meaning), an animal, a bacterial culture, a piece of machinery, or even a complicated unit such as the two front wheels of a vehicle for one week of test of a pair of tyres.

When conclusions from an experiment are applied to practice, experimental conditions and the population of plots from which those in the experiment were taken may be vitally important. Despite insistence by statisticians that their science is objective, some subjective judgement is then inevitable. No one expects inferences from an experiment on the feeding of rats to be applicable at all exactly to the feeding of humans. To what extent are inferences from an experiment on one strain of rats applicable to another strain or at another age ? Many such questions can be dealt with by extending the scope of the experiment, using factorial design or repeating trials at different times and places. Nevertheless, an investigator is always limited in the number of different conditions that he can include, and he therefore judges which he thinks most important. Similarly, anyone using published results from many sources as a guide to action, say in the management of dairy cattle, is obliged to exercise judgement on what differences are important. There may be widespread agreement that breed and type of housing are important to the relevance of experiments, and that phase of the moon or colour of shirt worn by the man in charge of the stock are unimportant ; on many other factors, opinion will be far fiom unanimous.

In planning an experiment, the statistician must ensure clear definition of the population of conditions being studied, so as to avoid dangers from uncontrolled heterogeneities. Secondly, and a little less rigorously, he may help in judging whether experimental materials and conditions can reasonably be regarded as representative of a population to which conclusions are to be applied.

6. What the Computer Does Not Do

A computer does not absolve the investigator from the need to plan experiments carefully, nor does it remove the distinction between planned experiments and miscellaneous observations. Unfortunately, writers of computer programs, and even statisticians bewitched by computers, sometimes mislead on this.

Confusion arises because a computer can easily handle large matrices. Any analysis of variance can be computed as a form of multiple regression, although this outlook is not conducive to clear revelation of truth. Consequently, if an experiment has been "planned" without attention to balance, symmetry, and orthogonality, the computer can complete an analysis that would have been intolerably laborious with desk calculators. If such an experiment has been performed, it can be analyzed ; almost certainly, it should not have been performed. The belief that statisticians recommend symmetry in design primarily in order to facilitate analysis is false. The real reason is that, taking into account classifications inherent in the experimental material or constraints that must be imposed, symmetry usually leads to estimation of the important comparisons between treatments with a precision that is maximum for the available resources. On occasion, a special purpose underlying an experiment causes the best design not to be the most symmetrical. A first duty for an investigator is to define the purpose of his experiment ; if he then chooses his design haphazardly (or, as some have unwisely proposed, selects treatments at random from a large set), he is deliberately evading his responsibilities.

More extreme heresy is exhibited by those who think that planned experiment can be replaced by multiple regression on non-experimental records. This is often a strong temptation, especially in the medical and social sciences. Hospital records of patients suffering from a specified disease may include dosages of various drugs, diet, duration of stay, and other measures of the treatment given. A regression equation of some measure of subsequent health on variates representative of hospital care may be of interest descriptively, and may suggest problems for further study ; it cannot itself provide evidence for causal influences. For example, recovery may be less complete for patients who received the more intensive drug therapy and who stayed the longer in hospital, not because these factors were themselves harmful but because the patients judged to require more treatment and longer hospital care were those initially more severely diseased. The association may be statistically highly significant yet tell nothing about whether increased dosage of a drug aids or hinders recovery. Randomization is lacking : trustworthy inference is impossible. (This is not to condemn a cautious analysis of the record as a basis for arousing suspicions about the effects of various factors and suggesting hypotheses that may later be more critically examined.)

7. The Computer as Labour Saver

For many years past, knowledge of statistical method and availability of mechanical aids to statistical calculation have sufficed to make computational simplicity seldom a necessary limitation on the planning of experiments. Occasionally an experimenter wisely adopted a design simpler than might have seemed ideal, either because his computational facilities were inadequate or because a slight mistake in conduct of the experiment would vastly complicate the correct analysis. The computer should remove this difficulty.

Perfection is not yet achieved. Indeed, the number of entirely satisfactory programs for the statistical analysis of experiments is still too small. Anyone who states that his computer installation has a standard program package for analyzing experiments should be regarded with scepticism (but not with disbelief). Probably this will not go beyond the simplest of designs, and provision for producing summary tables may be quite indequate ; anything more ambitious, or modifications to take account of missing obeservations, transformations, and other complications will then require special programming. For example, a few years ago, I advised use of lattice designs for certain experiments. A year later, I received the results of three experiments and wanted analyses quickly as a basis for further research. No program existed. I was the only person involved who could have written one, and I did not have the time. I had assistants who could work desk calculators, and therefore the experiments were analyzed in the old-fashioned manner.

Such difficulties will continue until we statisticians have performed our duty of writing good general program packages. These must have clear instructions, be reasonably easy to use, and produce good output. They must be much more than transcriptions of method from standard text books into a computer language. For example, they should permit much fuller examination of the adequacy of standard linear models than has been practicable with desk calculators, and should contain facilities for dealing with some types of breakdown of linearity. They must permit input of data in several different forms, for scrutiny and monitoring checks on data that will draw attention to all abnormalities, and for standardizations and transformations needed before analysis. Output of results should be by tables that do not require rearrangement or further calculation before being interpreted by the investigator (section 14). They must take account of misfortunes such as randomly-occurring missing observations. They must permit the user to nominate special comparisons among treatments for examination and tabulation, to require output of individual plot-residuals, and to call for completion of other special laborious calculations. Although barely possible to-day, they will soon be expected to allow conversational interaction between computer and user, so that further steps in the analysis can be specified after inspection of results from preliminary stages. A few excellent programs already exist, but scope for further development to meet the demands I have stated and others will remain for some years.

8. The Choice of Design

If conditions permit, a design with maximum symmetry, chosen from the extensive range described in standard sources, will be the wisest choice. Occasionally other constraints and demands may prevent this. For example, within the frame work of limited resources, an experimenter might wish that certain comparisons between treatments be more precisely estimated than others: he might hope to specify approximately the ratios of variances and to determine the optimal design in a particular system of blocks. I believe that a computer program that would produce at least a good approximation to the optimal ought to be possible, but I do not know of one.

I shall illustrate another situation in which limitation on materials prevents adoption of an ideally symmetric design. Suppose that treatments A, B, C are to be compared with one another and with a control treatment, S, using at most 31 plots arranged as:

1 block of 6,
1 block of 5,
2 blocks of 4,
4 blocks of 3.

This might occur in work with pigs, where experience suggests that animals from the same litter be grouped as a block and the 8 available litters have the sizes stated. I assume that σ^2, the variance per unit within blocks, is constant over all blocks.

symmetric design could be achieved by discarding 2 units from the first block, 1 from the second, and then having a full replicate of the 4 treatments in blocks of 4 ; each of the possible sets of three treatments (SAB, SAC, SBC, ABC) would be assigned to one block of 3. The variance of the estimated difference between any pair of treatment means is then easily proved to be $0{\cdot}3\sigma^2$.

Table 1 shows ten asymmetric ways of allocating treatments to blocks ; many others exist, but these have been chosen to look moderately well balanced while having some excess of units assigned to treatment S. Table 2 shows, for each design, the variances for the six differences between pairs of treatments. Several designs improve appreciably on the symmetric design first mentioned. If the aim is to compare each of A, B, C, with S, and little interest attaches to comparisons among A, B, C, designs 1 and 10 are superior to all others examined. If comparisons between A, B and C are equally interesting, possibly designs 5 and 9 will be preferred. My purpose here is less to decide the best design for this problem than to illustrate the fact that any wise decision requires a summary such as that of Table 2. Designs such as 4 or 7 can be rejected unhesitatingly because at least one other (e.g. 5) is better in respect of *every* treatment difference. The choice between 5, 9, 10, however, must rest upon the relative importance attached to different features of precision. Some would argue for basing decision on the mean variance of the three or six treatment differences, some would prefer a minimax rule : in practice neither is satisfactory if used uncritically, though often both may lead to much the same choice.

The calculations for this small experiment were easily made without a computer. A more thorough examination of possible designs, or the study of a large and more complex experiment, would be practicable only with good assistance from a computer. Evidently the computer could be programmed to make an exhaustive search of all designs, or of all satisfying some conditions that exclude the obviously useless. More interesting would be a program that moves step by step towards a design optimal in respect of some criterion based upon variances. At the recent Biometric Conference in Hannover, Justesen described a systematic process which shows considerable promise for further development. A further interesting constraint is introduced if limitations of supplies for certain treatment limit the replication of these. For example, such a design may be required for early tests of new varieties of a crop when seed is scarce.

9. The Size of Experiment

If an expriment is not restricted by availability of materials, its size can satisfy a condition on the quality of the results. The commonest, but not the only relevant, condition would specify the maximum variance that will be tolerated. This requires consideration of the familiar formula for the variance of a mean, σ^2/n, which involves no computational difficulty ! Uncertainty about the value of σ^2 usually makes the inference untrustworthy ; although theory based upon estimates of σ^2 has been suggested, the frequency distributions obtained are complicated. Since any estimate of σ^2 must be based upon one or more previous experiments, I think it unlikely that formal distribution theory can contribute much without an empirical component. Only if experimental conditions are very tighty controlled is σ^2 in the n ew experiment likely to be identical with σ^2 in earlier experiments, so that distribution theory solely dependent upon sampling variation under fixed conditions may be seriously misleading. If progress is made here, almost certainly computer evaluation of distributions will be needed.

An alternative to deciding in advance the size required for an experiment is to employ a sequential stopping rule. In some fields of research, the time scale or other considerations make this inappropriate. When it is appropriate, the organization of a project is likely to require some kind of advance estimate of total size. Except for rather simple sequential plans, properties of the stopping rule (such as the average number of subjects that will have been tested when the experiment terminates or the probability that a specified number of subjects is exceeded) are mathematically intractable. A computer can be used to simulate experimental records (Section 13), by generation of suitable random elements, and the frequency distribution of the size of experiment can then be studied empirically.

10. Planning for Variates

An experiment will be intended to compare different categories of plot or other experimental unit, distinguished in respect of treatments, without bias, with maximum precision, and under conditions that ensure relevance to subsequent application of the results. The computer does not modify this. Usually, however, the aim will be understood in relation to one variate of greatest importance (wh'ch may itself be calculated from two or more distinct measurements), or to a compromise between optimal requirements of several variates.

Some experiments easily generate large numbers of variates. For example, any study of meat production can produce many carcase measurements, and an experiment on milk yields is likely to include measurements of many physical and chemical properties. If a suitable computer program for one variate exists, it can produce analyses and tabulations for each variate in turn. This not only allows fuller utilization of the information recorded : it encourages the investigator to plan broadly for the collection of observations. Thus the aims of the experiment may be extended, and planning in respect of precision may balance the needs of different variates.

Nevertheless, proliferation of variates must be kept within bounds. An experiment may easily generate so many variates, and so many statistical summaries, that the investigator, is smothered by computer output and his attention is diverted from matters of prime importance. I remember an instance of an agronomist seeking advice from a statistician about the analysis of his latest series of experiments. When he tried to discover what variates should be submitted to full analysis or given priority, the statistician found rational decision impeded by the fact that the experiment had not yet digested the computer output from earlier experiments analyzed two years previously ! Research might have progressed more effectively had effort been concentrated on three or four primary measures of yield.

Records of a variate are no use as long as they merely remain on file without any form of statistical examination. Neverthless, it is sometimes better to decide against analyzing a variate after a brief scrutiny, or even arbitrarily, than to increase output to an extent that may obscure rather than enlighten. The statistician must also remember that some of the variates directly measured may not be those that should be separately analyzed. Uncritical transfer from original records to a standard computer program should be discouraged. For example, if body temperatures of human patients were recorded twice daily for two weeks, 28 separate analysis probably would not be very helpful ; derived measures of average temperature for the period, rate of change, and daily range might be more informative. In a potato experiment, the produce of each plot may be divided into 6 size groups, and the numbers and weight of produce from each recorded. As separate variates, these have little interest, and anaylsis will be confused by complex intercorrelations. Consideration of the frequency distribution of size of individual tubers may give better variates for analysis, such as mean size, variance of size, percentage exceeding a minimal marketable size, and so on.

11. Multivariate Analysis

If for a particular design a program for the analysis of variance of one variate exists, it can readily be extended to simultaneous variance and covariance analysis for several variates.

Classical covariance analysis, involving adjustments in treatment means for one variate by reference to its error regression coefficients on other variates, can enable some extraneous sources of irrelevant variation to be eliminated. Proper randomization of the original design and use only of concomitant variates not themselves influenced by treatment are essential. The methods are well-known, but for complex designs the arithmetic becomes extensive. The computer removes this labour and also the practical restriction to having at most three simultaneous concomitants, though I doubt whether we shall often meet situations in which covariance on four or more concomitants is useful. Possibly more important is the practicability of trying alternative covariance schemes, including non-linear regressions if a curvature is suspected. Moreover, emergency devices for the rescue of experiments that have suffered disaster, through accidental losses or confused recordings of observations of unexpected environmental trends, can be constructed in terms of dummy covariates. Some study is needed of the consequences for an analysis of choosing the apparent best out of many covariance schemes.

I hope I have already (Section 6) sufficiently condemned use of multiple regression as a substitute for the proper designing of experiments. Nevertheless, regressions arising in covariance analysis may suggest relationships that can rightly form the subject of further experimentation.

When all variates are of intrinsic interest, the situation is very different and a general study of the effect of treatments on all variates may be wanted. Inspection of a full analysis of squares and products may be valuable to the experienced statistician. For example, two logically distinct variates may be seen to be so closely correlated that, under the conditions of the experiment, they are effectively equivalent. At a more sophisticated level, interest lies in mathematical techniques that determine a function of the variates with defined optimal properties. Factor analysis, canonical analysis, and principal component analysis are among the best known. As numerical devices for the exploration of data, and for stimulating the generatian of hypotheses that can subsequently be studied more deeply, they have their uses, but those who employ them naively are apt to be misled. A combination of variates determined from internal evidence of the data will have its own experimental error and may be more meaningful when refined by the investigator. For example, in a long-term experiment on dairy cattle, milk yields might be recorded for three successive lactations. A formal multivariate analysis of these variates might suggest

$$1.4x_1+1.7x_2+1.3x_3$$

and

$$1.1x_1-0.3x_2-0.9x_3$$

as having some optimal property in explanation of total variation. These functions have no natural meaning and are unlikely to be repeatable in other experiments. However, they point to total yield,

$$x_1+x_2+x_3,$$

and change over the period of study,

$$x_1-x_3,$$

as more meaningful alternatives so closely correlated with the first two as not to depart far from optimality. Further difficulties arise when variates differ in units of measurement.

I am not asserting that these forms of multivariate analysis are wrong or valueless; indeed, I believe that we should explore their potentialities thoroughly now that computers ease the labour, but their indiscriminate use is folly.

12. The Transformation of Variates

In discussing the transformation of data, elementary text-books of statistics commonly convey the impression that the subject is easy : look for warning signals in the data and, if these appear, analyze instead the square root, the logarithm, or other stated function of each observational value. On the other hand, some papers in statistical journals make immense mystery of transformations, suggesting adjustments and other special tactics the practical use of which is small because the information on which to base them is seldom available. When an experiment is intended for the study of a specified variate, to report results in terms of a mathematical transformation of that variate is irresponsible and possibly useless. Here textbook advice is rarely good. Data may be transformed in order to reconcile them with requirements of valid statistical analysis, but statistical theory has neglected the need to present conclusions on the original scale.

With a computer different analysis of the same data can be compared. Exploration of the stability of conclusions under various assumptions may help to justify an opinion that the exact choice of transformation matters rather little ; my own experience is that data for which theory indicates the desirability of transformation can often be safely analyzed and interpreted as they stand. I illustrate by a very simple experiment. Table 3 shows systolic blood pressures in two sets of 10 cats after exposure to alternative drugs that should reduce blood pressure. These would ordinarily be analyzed by a simple t-test. Although cats on drug B varied in blood pressure substantially more than those on drug A, I personally would not have judged the data to need transforming. However, I have studied the variability by employing the useful family of transformations.

$$y=(x+q)^p.$$

With $q=0$, this includes square root, cube root, reciprocal, and other simple power transformations; the inclusion of q gives extra flexibility. If p tends to zero, the transformed variate behaves like log $(x+q)$, as may be seen by considering the modified form

$$y^*=[(x+q)^p-1]/p.$$

If q becomes large, the transformation behaves like x itself, as of course it does for any q and $p=1$.

For many combinations of p and q, I looked at two quantities obtained in the analysis of y, the ratio of the mean square deviations for the two drugs and the value of the t statistic. Tables 4 and 5 contain some results. A variance ratio test of the homogeneity of the mean squares for A and B rejects the hypothesis of equal variances if the observed ratio exceeds 4.0 or is less than 0.25. Table 4 shows the ratio to vary widely, and to be not much above the lower limit when $p=1$. The value of t, on the other hand, is remarkably stable. By coincidence, it seems to achieve its minimum near to $p=1$. Nevertheless, even extreme transformations such as $p=4.0$, $q=0$ or $p=2.5$, $q=-4.0$ (which might be adopted if the criterion of making the ratio of mean squares nearly unity were enforced) alter the value of t only from 2.5 to 2.3. For significance at probability 0.05, t must exceed 2.1 : from Table 4, any combination of p and q that makes t less than 2.1 may be judged manifestly absurd.

To estimate the difference in means on the original scale is not easy. As a rough guide, I have taken the mean of y for each drug and used the inverse transformation

$$X=\bar{y}^{\frac{1}{p}}-q,$$

followed by subtracting one X from the other. This quantity, which is certainly not an unbiased estimate, is shown in Table 6. More interesting, perhaps, are the pseudo-t values in Table 7, the quotient of each entry in Table 6 by an approximate standard error ; the latter is calculated in the ordinary way from a pooled mean square and the differential coefficient dy/dx. Comparison of Tables 5 and 7 illustrates the robustness of t under approximations in calculation. Although values in Table 7 change rather more with changes in p and q, in the central region of the table the two tables are very much alike.

This study, practicable only by computer, reassures me that Table 3 is satisfactorily analyzed without transformation. I believe that statisticians will develop greater interest in this kind of approach. After a survey of a family of transformations, one would like to choose that most suitable to the data, to make inferences by its use without any bias consequent upon how it was chosen, and eventually to return to interpretation on the original scale of measurement. More rigorous theory is needed.

13. The Computer as Experimental Instrument

The computer can itself become an experimental instrument when it is used in simulation. So extensive are the possibilities that they seem likely to cause some statisticians to become in part experimental scientists. The heavy demands on computer time, however, indicate the desirability of seeking an analytical solution to a problem before recourse to simulation.

The computer cannot fulfil the experimenter's dream by permitting him to compare real treatments without the labour of a real experiment ! Simulation should enter earlier, to facilitate the comparison of alternative designs. Section 8 illustrated how relatively simple designs can be compared in terms of their variance patterns. A complex network or sequence of experiments may be less easily studied by direct evaluation of variances, either because general formulae would be exceedingly clumsy or because the relevant criterion for good experiment is not just a function of the variances. In animal breeding, questions arise about numbers of sires and of dams to be tested, numbers of progeny to be raised from a mating, length of time for which the performance of progeny is to be recorded before selection of parents for the next generation, and number of generations. How shall efforts and resources be divided between different stages of the program ? Assumptions may be made about the genetic and environmental determinants of economically important characters, including of course the variation attributable to random sampling. Although the investigator is constrained by limitations of resources and by requirements on the use of the results, he still has great freedom of action. He can have more sires at the price of fewer dams per sire, or more generations at the price of shorter test periods per generation. Within the constraints, what policy maximizes the gain from the whole programme of breeding and selection, or minimizes the risk that the gain will be too small to be economic ?

A computer cannot produce the final animals or even predict their quality exactly. It can be used to study alternative deployments of resources, perhaps simulating the whole selection programme for each and finding the average performance from 1000 different sets of initial animals. Thus not only may the optimal conditions be found in a problem too complicated for exact mathematical analysis, but the risks of failure to achieve precisely optimal conditions because of uncertainties about the values of parameters can be assessed. Thereafter, selection policy can be based upon reasoned judgement about desirable conditions, and upon knowledge of how robest the recommendations are. Much the same analysis arises in connexion with plant breeding. The problem is also closely paralleled by that of screening large numbers of chemical compounds by a preliminary test intended to detect which of them may have useful therapeutic properties.

14. Can the Computer Aid Interpretation ?

The interpreter of an experiment or set of experiments must relate results to other information in the same field and prepare a consistent statement of conclusions: this is usually a final duty for the experimenter. Good tabulations of means and other numerical or graphical summaries of data are essential. Manual construction of summary tables and diagrams is laborious, and is frustrating because only a few of them are eventually used; consequently the statistician is often tempted to decide which to prepare rather arbitrarily. Human reasoning and judgement are essential, but the computer can assist the rapid scanning of many different summaries.

Many years ago, I was consulted about research into the keeping quality of milk. Samples of milk stored in various ways were subjected to numerous tests, some objective and perhaps not closely related to what the consumer regards as good or bad milk, others based on consumer criteria and inevitably more subjective. The aims were less well defined than was desirable, but one clearly necessary step was to examine the relations between most pairs of variates; as these might be non-linear, many scatter-diagrams were needed. How much easier such a task is to-day! The records of all variates can be stored on tape or disk, and a very simple program can recover the information on a particular pair of variates. Even without special hardware for visual display or for graph plotting, a line printer can be used to produce diagrams adequate for the rapid inspection appropriate to this kind of problem. For extensive data, the diagrams are especially valuable in drawing attention to non-linearities or indicating where transformation may lead to a simpler representation; they are more immediately useful than correlation coefficients and regression equations, although these are easily calculated at the same time if requested.

A program for graphical representation, or for a systematic preparation of tables, helps the interpretation of data by suggesting on what the statistician should concentrate in a more detailed analysis. Indiscriminate use can be dangerous if it overwhelms the user with computer output. I once talked with a scientist who had administered a battery of 500 questions to several thousand human subjects. The answer to each question was necessarily "Yes" or "No"; he proposed to form all possible 2×2 tables between pairs of questions and to calculate X^2 for each as a first stage in his study. He was insistent that he needed to have the 125,000 tables and X^2 values in front of him before he decided what to do next. He could not be brought to see that inspection of 10 kilometres of line-printer output was likely to obstruct rather than to expedite understanding.

Can we learn what features of diagrams and tables distinguish those the investigator will use further from those he will discard? If so, the computer could be programmed to do its own preliminary scanning of tables and to be more discriminating in its output. Anything as uncritical as asking the computer only to print tables that showed close or statistically significant relations between variates could be thoroughly inadequate. The user of the program must always be allowed to override the standard rules, by insistence that tables of his own choice be produced.

Still greater flexibility of analysis will come with the development of multiple access to large computers from remote consoles. Data can be held in store and the console can be used to apply small segments of a standard program, or even *ad hoc* instructions written directly on the console, so as to build up a pattern of analysis most appropriate to the particular data. No longer will it be necessary to formulate a comprehensive plan of analysis at one time, and the interpretation of some types of data may be greatly aided when a full exploitation of these new facilities becomes possible. I do not think this will become of major importance for standard experiments, but it will certainly be valuable in some circumstances; for example, the questions discussed in Sections 8 and 12 may prove particularly amenable to this approach.

Not least important in the interpretation of statistical analyses is the form of output that must be inspected. General programs are often deplorably unsatisfactory in the quality of output. Until recently, storage capacity has been a serious limitation on programs, and programmers have been reluctant to use space for instruction on the format of tabulations. Moreover, preparation of these instructions is a tedious part of program writing, disliked by those more interested in analytical ingenuity. This dislike must be overcome! A good general program should produce summary tables that are well labelled, easily read, and after minimal modification suitable for transfer to reports for publication. One should not have to read an analysis of variance that gives only mean squares, suppressing the sums of squares and various sub-totals inherent in the structure of the analysis. Only as a rare exception should tables of means or entries in an analysis of variance be expressed in floating point notation: to program for location of the decimal point as a basis for constructing neater tables is no more than a minor nuisance to the programmer. The reader of the output should not have to refer to other records in order to discover units of measurement, or to identify the variates used in tables. Names of treatments also should be properly identified in the output, even though a simple coding be used in tables. In the analysis of factorial experiments, output that merely lists treatment combinations and mean yields, leaving the reader to construct multifactor tables, is thoroughly bad. Any program for a large experiment should permit the user to nominate the multifactor or other tables that he wants. All standard errors likely to be required, and any special adjustments of means, should be part of the standard output. Much time can be wasted in manual disentanglement of poor computer output in order to prepare for final presentation. The computer encourages analysis of more variates than was customary 20 years ago: it should be made to do all the tedious work needed for good presentation.

The existence of good computer facilities and reasonably good programs can cause the most important conclusions from an experiment to be buried beneath a large body of unimportant output. I believe that not the least of contributions to the interpretation of data is to incorporate in every program a good range of options on output. Quite rightly to-day we ask for more analyses and more details than when we depended upon desk calculators. The custom of asking for *everything*, in order to have a look at it and then to discard what is not informative, can do great harm by diversion of attention. If an experiment has produced a vast amount of data, decisions must be taken that certain variates, or certain aspects of the data, will not be reported in detail. Some decisions of this kind are best taken before analysis, others should be made conditional on criteria that are specified within the program and the printing or suppression of the relevant output should be controlled at that point, and certainly a few may have to be deferred until after an experienced scientific eye has scrutinized the computer output.

15. Summary

The paper begins by defining comparative experiments and distinguishing between different purposes for which these are conducted : as a pilot for definitive experiments, for technological decision, for scientific learning. The functions of the statistician and of the computer are then briefly stated.

The first objective of experimental design is that experiments be relevant to the problem studied, free from bias, and as precise as resources permit (Section 5). The computer can help in many ways, but those who employ it must always keep in mind its moronic character as well as its memory and its speed (Section 6). Important though computer speed is, one must not forget that the existence of a program is a pre-requisite of speedy analysis (Section 7). A computer cannot of itself determine a suitable design for an experiment; when relevance and unbiasedness have been secured, however, it can be used in various ways for comparing the merits of different designs and different allocations of resources (Sections 8, 9, 13).

A good design for an experiment is worthless if the aims of theex periment are ill-defined, or if the investigator lacks ideas on how to exploit the situation that he is studying. In Section 10, I discuss how availability of a computer should affect the aims of an experiment, by permitting the investigator to plan for a broader and deeper study that will depend upon much more arithmetic.

A well-designed experiment with well-conceived aims will fail to make any impact on science or technology unless good statistical techniques are applied. Sometimes little more than simple averaging or graphical presentation is needed but extensive bodies of records may deserve a far more complicated analysis (Section 7). Although in one sense the analysis is distinct from the design, the nature of the analysis must be greatly influenced by the design. I have therefore chosen not to discuss details of analysis, but to comment on general principles that are influenced by the computer revolution. I note particularly the opportunity for thorough study of the implications of transforming variates (Section 12) and the need for better use of multivariate analysis now that the computational burden is so much less serious (Section 11). I am particularly concerned by the need to organize computer programs so that statistician and all scientific investigators can obtain their own selections of tables and digrams, in clear and easily utilized format (Section 14). Above all else, the computer must not be allowed to become a burden instead of a help: it must not pour out "summaries" of the data in quantity far exceeding the capacity of its user to absorb and interpret.

TABLE 1

Structures of Ten Designs for an Experiment in Randomized Unequal Blocks

Design 1	*Design 2*	*Design 3*	*Design 4*	*Design 5*
SSABCC	SSAABB	SSSABC	SSAAAA	SAABBC
SSABC	SSABC	SSABC	SSAAA	SABCC
SABC	SABC	SABC	SBBB	SABC
SABC	SABC	SABC	SCCC	SABC
SAB	SCC	SAB	SBB	SAB
SAB	SAB	SAB	SBB	SAB
SAC	SAC	SBC	SCC	SAC
SBC	SBC	SAC	SCC	SBC

Design 6	*Design 7*	*Design 8*	*Design 9*	*Design 10*
SABBCC	SSSABC	SABBCC	SAABBC	SSSABC
SSABC	SAABB	SAABC	SABCC	SSABC
SABC	SACC	SABC	SABC	SABC
SABC	SBCC	SABC	SABC	SABC
SAA	SAB	SAB	SAB	SAB
SAB	SAA	SAC	SAC	SAC
SAC	SBB	SSB	SBC	SBC
SBC	SCC	SSC	ABC	ABC

TABLE 2

Variance of Estimated Treatment Differences for Designs of Table 1

(Table shows multiples of 0·001 σ^2)

Design	*S-A*	*S-B*	*S-C*	*A-B*	*A-C*	*B-C*
1	251	251	255	300	312	312
2	257	257	280	302	362	362
3	244	244	279	300	331	331
4	395	480	480	875	875	960
5	258	258	285	261	295	295
6	273	276	276	321	321	300
7	287	287	314	369	453	453
8	285	272	272	309	309	316
9	283	283	283	261	267	267
10	247	247	247	293	293	293

TABLE 3

Final Systolic Blood Pressures (mg. Hg.) of Cats

	Drug A	*Drug B*
	90	55
	80	80
	100	80
	80	70
	80	80
	95	95
	90	85
	75	70
	85	50
	85	70
Means	86.0	73.5

TABLE 4

Ratios of Mean Square Deviations for Transformations of Data in Table 3

Values of *q*

Values of p	—49	—45	—40	—30	—20	0	50	100	1000
—6·00	0·000	0·000	0 000	0·000	0·001	0·005	0·036	0·073	0·263
—5·00	0·000	0·000	0 000	0·000	0·002	0·010	0·051	0·093	0·272
—4·00	0 000	0·000	0·000	0·001	0·005	0·020	0·072	0·117	0·280
—3·50	0·000	0·000	0 000	0·002	0·008	0·027	0 086	0·131	0·285
—3·00	0·000	0·000	0 000	0·004	0·013	0·038	0·101	0·146	0·289
—2·50	0 000	0 000	0 001	0 008	0 021	0·052	0·119	0·162	0·294
—2·00	0·000	0 000	0·003	0·015	0·033	0·070	0·139	0 181	0·298
—1·50	0·000	0 002	0 008	0·029	0·052	0·093	0·162	0·201	0·303
—1·00	0·000	0·007	0 020	0·051	0·079	0·123	0·188	0·222	0·308
—0·50	0 004	0·024	0 048	0·087	0·118	0·161	0·218	0·246	0·312
0·00	0·033	0·071	0·101	0·142	1·170	0·207	0·250	0·271	0·317
0 25	0·071	0·111	0·141	0 179	0 203	0·233	0·268	0·284	0·319
0·50	0·132	0·167	0·192	0·221	0 240	0·262	0·287	0·298	0·322
0·75	0·219	0·239	0·254	0·271	0·281	0·293	0·306	0·312	0·324
1·00	0·327	0·327	0·327	0·327	0·327	0·327	0·327	0·327	0·327
1·50	0·589	0·543	0·505	0·459	0·432	0·402	0·371	0·358	0·332
2·00	0·885	0·796	0·716	0·616	0·555	0·487	0·418	0·391	0·337
2·50	1·192	1·069	0·950	0·792	0 694	0·583	0·470	0·426	0·342
3·00	1·502	1·350	1·197	0·984	0·846	0 687	0·526	0·463	0·347
3·50	1·813	1·634	1·451	1·186	1·009	0·799	0·585	0·502	0·352
4·00	2 126	1·921	1·709	1·395	1·180	0·919	0·648	0·543	0·357
5 00	2·775	2·509	2 236	1·827	1·540	1·176	0·784	0·631	0·368
6·00	3·487	3·139	2 790	2 277	1·917	1·452	0·932	0·727	0·378

TABLE 5

Values of t for Transformations of Data in Table 3

Values of *q*

Values of p	—49	—45	—40	—30	—20	0	50	100	1000
—6·00	1·000	1·017	1·098	1·300	1·477	1·756	2·146	2·313	2·526
—5·00	1·000	1·035	1·150	1·390	1·586	1·873	2·228	2·364	2·528
—4·00	1·001	1·072	1·233	1·515	1·725	2·005	2 306	2·411	2·529
—3·50	1·002	1·104	1·293	1·595	1·808	2·075	2·343	2·432	2·530
—3·00	1·005	1·151	1·370	1·689	1·899	2·147	2·379	2·452	2·530
—2·50	1·014	1·221	1·471	1·798	1·998	2·219	2·411	2·470	2·531
—2·00	1·035	1·326	1·600	1·921	2·101	2·288	2·441	2·486	2 531
—1·50	1·090	1·479	1·760	2·054	2·206	2·354	2·467	2·490	2·532
—1·00	1·225	1·691	1·950	2·191	2·307	2·412	2·400	2·511	2 532
—0·50	1·518	1·957	2 153	2·321	2·395	2·461	2·507	2·520	2·532
0·00	1·971	2-233	2·341	2·428	2 466	2·498	2·521	2·527	2·532
0·25	2·196	2·351	2·417	2·470	2·493	2·512	2·526	2·529	2·533
0·50	2·373	2 443	2·475	2·502	2·513	2·522	2·529	2·531	2·533
0 75	2·484	2·504	2·514	2·522	2·526	2·529	2·531	2·532	2·533
1·00	2·533	2·533	2·533	2 533	2·533	2·533	2·533	2·533	2·533
1·50	2·498	2·508	2·515	2 522	2 525	2·529	2·531	2·532	2·533
2·00	2·377	2·413	2·443	2 477	2·494	2·512	2 525	2·529	2·532
2·50	2·233	2·289	2·341	2·407	2·444	2·483	2·514	2·523	2·532
3 00	2·091	2·159	2·228	2 322	2·381	2·444	2·499	2·515	2·532
3·50	1·961	2·036	2·114	2·231	2·309	2·398	2·479	2·505	2·532
4·00	1·846	1·923	2·007	2·139	2·233	2·347	2 457	2·492	2·531
5·00	1·655	1·731	1·818	1·965	2·079	2·233	2 402	2 462	2·530
6·00	1·510	1·580	1·664	1·812	1·935	2·116	2 338	2·424	2·529

TABLE 6

Differences between Back-transforms of Means for Transformations of Data in Table 3

Values of *q*

Values of p	*—49*	*—45*	*—40*	*—30*	*—20*	*0*	*50*	*100*	*1000*
—6·00	31·27	29·78	28·01	25·02	22·71	19·66	16·37	15 11	12·87
—5·00	31·64	29·68	27·48	24·02	21·59	18·67	15 78	14·71	12·82
—4·00	31·97	29·30	26·53	22·65	20·24	17·61	15·20	14·32	12·76
—3·50	32·10	28·92	25·82	21·80	19·48	17·07	14 91	14·13	12·74
—3·00	32·17	28·32	24·88	20·85	18·68	16·52	14·62	13:94	12·71
—2·50	32·13	27·41	23·70	19·80	17 86	15·98	14·34	13·75	12·68
—2·00	31·84	26 04	22·24	18·69	17·02	15·43	14·06	13·56	12·66
—1·50	30·97	24·09	20·54	17·54	16·18	14 90	13·79	13·38	12·63
—1·00	28·73	21·58	18·70	16·39	15·36	14·38	13·52	13·20	12·61
—0·50	24 18	18·81	16·86	15·30	14·58	13·88	13·25	13·02	12·58
0·00	18·76	16 22	15·18	14 27	13·83	13·40	13·00	12·84	12·55
0·25	16·57	15 10	14·42	13·79	13·48	13·16	12·87	12·76	12·54
0·50	14·85	14·11	13·72	13 33	13 14	12·94	12·74	12·67	12·53
0 75	13·52	13·25	13·08	12 90	12·81	12·72	12·62	12·58	12·51
1·00	12·50	12·50	12·50	12·50	12·50	12 50	12·50	12·50	12·50
1·50	11·04	11·30	11·51	11·76	11·92	12·09	12 26	12·33	12·47
2·00	10·06	10·39	10·70	I1·12	11·38	11·70	12·03	12·17	12·45
2·50	9·33	9·68	10·04	10·56	10·90	11·33	11·81	12·01	12·42
3·00	8 76	9·12	9·49	10·06	10·47	10·99	11·59	11·85	12·40
3 50	8 30	8 65	9·03	9·63	10·08	10·67	11·38	11·70	12 37
4·00	7·92	8·25	8·63	9 25	9·72	10·37	11·18	11·55	12·34
5·00	7·29	7·61	7·98	8·60	9·10	9·82	10·79	11·26	12·29
6·00	6·79	7·10	7·46	8 07	8·58	9·35	10·43	10 98	12 24

TABLE 7

Pseudo-t Values from Table 6 and Approximate Standard Error

Values of *q*

Values of p	*−49*	*−45*	*−40*	*−30*	*—20*	*0*	*50*	*100*	*1000*
−6 00	0·000	0·000	0·010	0·115	0·337	0·872	1·745	2·104	2·522
—5·00	0·000	0·002	0·031	0·235	0·548	1·146	1·930	2·213	2·524
—4·00	0·000	0·013	0·099	0·456	0 ·856	1·459	2·101	2·308	2·527
—3·50	0·000	0·030	0·171	0·621	1·048	1·623	2·179	2·351	2·528
—3 00	0·001	0 069	0·290	0·829	1·263	1·786	2·251	2·389	2·529
—2·50	0·005	0·150	0·474	1 079	1·493	1·944	2·316	2·423	2·530
—2·00	0·021	0·315	0·745	1·364	1·727	2·091	2 374	2·452	2·531
—1·50	0·090	0·612	1·104	1·664	1·952	2·222	2·422	2·477	2·531
—1 00	0·334	1·070	1·523	1:952	2·153	2·333	2·462	2·497	2·532
—0·50	0·949	1·625	1·934	2·200	2·318	2·421	2·493	2·513	2·532
0·00	1·783	2·119	2·264	2·385	2·438	2·483	2·515	2·524	2·532
0·25	2·119	2·301	2·382	2·450	2·479	2·505	2·523	2·528	2·532
0·50	2·353	2:430	2·465	2·496	2·509	2·520	2·528	2·530	2·533
0·75	2·485	2·505	2·515	2·523	2·526	2·529	2·532	2·532	2·533
1 00	2·533	2·533	2·533	2·533	2:533	2·533	2·533	2 533	2·533
1·50	2·463	2·479	2·491	2·506	2·514	2·522	2 529	2·530	2·532
2·00	2·294	2·340	2·381	2·433	2·462	2 492	2·517	2·524	2 532
2·50	2 102	2·171	2·238	2·329	2·385	2·445	2·498	2·514	2·532
3·00	1·918	2·001	2·086	2·210	2·292	2·386	2·473	2·500	2·532
3 50	1·752	1·841	1·938	2 086	2·190	2·317	2·442	2·483	2·531
4·00	1·607	1·698	1·799	1·964	2·085	2·242	2·406	2·463	2·530
5·00	1·371	1·458	1·561	1·738	1·881	2 084	2·323	2·415	2·529
6·00	1·193	1 273	1·370	1·545	1·696	1·926	2·230	2·357	2·527