P. OTTESTAD, Ph.D. (Oslo)

STATISTICAL MODELS AND THEIR EXPERIMENTAL APPLICATION

BEING
NUMBER TWENTY-FIVE
OF

RIFFIN'S STATISTICAL
MONOGRAPHS & COURSES

EDITED BY ALAN STUART, D.Sc.(Econ.)

GRIFFIN'S STATISTICAL MONOGRAPHS AND COURSES

No. 1: *The analysis of multiple time-series* M. H. QUENOUILLE

No. 2: *A course in multivariate analysis* M. G. KENDALL

No. 3: *The fundamentals of statistical reasoning* M. H. QUENOUILLE

No. 4: *Basic ideas of scientific sampling* A. STUART

No. 5: *Characteristic functions** E. LUKACS

No. 6: *An introduction to infinitely many variates* E. A. ROBINSON

No. 7: *Mathematical methods in the theory of queueing* A. Y. KHINTCHINE

No. 8: *A course in the geometry of* n *dimensions* M. G. KENDALL

No. 9: *Random wavelets and cybernetic systems* E. A. ROBINSON

No. 10: *Geometrical probability* M. G. KENDALL and P. A. P. MORAN

No. 11: *An introduction to symbolic programming* P. WEGNER

No. 12: *The method of paired comparisons* H. A. DAVID

No. 13: *Statistical assessment of the life characteristic: a bibliographic guide* W. R. BUCKLAND

No. 14: *Applications of characteristic functions* E. LUKACS and R. G. LAHA

No. 15: *Elements of linear programming with economic applications* R. C. GEARY and M. D. MCCARTHY

No. 16: *Inequalities on distribution functions* H. J. GODWIN

No. 17: *Green's function methods in probability theory* J. KEILSON

No. 18: *The analysis of variance* A. HUITSON

No. 19: *The linear hypothesis: a general theory* G. A. F. SEBER

No. 20: *Econometric techniques and problems* C. E. V. LESER

No. 21: *Stochastically dependent equations: an introductory text for econometricians* P. R. FISK

No. 22: *Patterns and configurations in finite spaces* S. VAJDA

No. 23: *The mathematics of experimental design: incomplete block designs and Latin squares* S. VAJDA

No. 24: *Cumulative sum tests: theory and practice* C. S. VAN DOBBEN DE BRUYN

No. 25: *Statistical models and their experimental application* P. OTTESTAD

No. 26: *Statistical tolerance regions: classical and Bayesian* I. GUTTMAN

No. 27: *Families of bivariate distributions* K. V. MARDIA

*Now published independently of the Series.

For a list of other statistical and mathematical books see back cover.

STATISTICAL MODELS AND THEIR EXPERIMENTAL APPLICATION

PER OTTESTAD
Professor of Statistics
The Agricultural College of Norway

BEING NUMBER TWENTY-FIVE OF
GRIFFIN'S STATISTICAL
MONOGRAPHS & COURSES
EDITED BY
ALAN STUART, D.Sc.(Econ.)

GRIFFIN 1820 LONDON

CHARLES GRIFFIN & COMPANY LIMITED
42 DRURY LANE, LONDON, WC2

First published 1970
SBN: 85264 166 4

Set by E W C Wilkins & Associates Ltd London N20
Printed in Great Britain by Latimer Trend & Co Ltd Whitstable

PREFACE

When, thirty years ago, I became acquainted with the then recently invented methods of experimental research, I accepted the ideas rather uncritically, save for two: the idea of non-random experimental material, and the now famous lady tea-taster introduced by R.A. Fisher in *The Design of Experiments*. Regarding the latter example, I could see its power as a means of explaining the principle of significance tests, and I have sometimes used it for this purpose in my teaching. But I could not and still cannot see why anyone would be interested in knowing that a certain lady is, or is not, able to discriminate between two kinds of tea. It was clear that tea-room proprietors would probably be interested in knowing whether their guests were usually able to tell the difference, but then the task would be to plan and carry out experiments with samples of tea-room guests as the experimental material.

Having been trained as a biologist and involved in biological research, I could easily see the tremendous progress signified by these methods. But what reliance should be placed on the idea of non-random experimental material? Soon also doubt crept in concerning the models underlying the methodology.

So for thirty years I have, on and off, returned to these problems. Years ago I found solutions which appeared to be satisfactory, although it would have been premature to publish anything about them at the time. First it was necessary to carry out certain investigations in order to justify the different methods. Until quite recently I had no means of doing this, but having now completed the necessary work, I find the results so encouraging that I am convinced of the soundness of the ideas which have guided this research.

It has been a great effort for me to write this treatise, particularly as I had to explain the underlying ideas in a language foreign to me. Another difficulty has been: how to deal at all adequately with the literature concerning the subject? The number of relevant monographs and articles in scientific journals is now very great, and I found that to cite the monographs and articles known to me would make my treatise almost unreadable. I have therefore cut down on citations and references to an extent that might be regarded as unjustifiable.

In closing, I wish to convey my thanks to a number of friends with whom I have had profitable talks about the problems. To Professor Alan Stuart, who has read my manuscript, I am most indepted for helpful comments. Thanks are also due to my wife for the assistance she has given me in writing the text in English.

PER OTTESTAD

The Agricultural College of Norway
October, 1969

CONTENTS

1 Preliminaries

About forty years ago, important research work on the principles of experimentation was begun at Rothamsted Experimental Station in England. The first general account of the results of this research was given by R.A. Fisher in his book *The Design of Experiments*, which originally appeared in 1935. Ten years previously, the same author's *Statistical Methods for Research Workers* had been published; in this book the new statistical tool of analysis, called the analysis of variance, was made known to research workers. A large number of papers and books, dealing with experimental design and statistical analysis, have since been inspired by these two important treatises.

It is well known that the results of the Rothamsted research work were not adequately recognized and valued by authorities on statistical methods at the time. Today, the principles of the Rothamsted school seem to be accepted unreservedly by almost all statisticians. On the other hand, these ideas are not accepted throughout by all research workers, and it is a fact that experimental research is very often carried out according to other rules. Often the principle of randomization, perhaps the most important and a lasting contribution made by the Rothamsted school, is ignored. The consequence is that many reports on experimental results are published, describing effects that are partly due to erroneous design.

The work of the Rothamsted school on design and statistical methods of analysis is certainly most important, but it is difficult to accept the principles in full. Briefly, criticism can be raised against the following elements: (1) the conception of experimental material as something fixed; (2) the purpose for which an experiment is carried out; and (3) the models upon which the theory rests.

(1) A research worker deals with questions. In planning and carrying out an experiment, he wishes to obtain data from which answers to these questions can be given. Then, by the use of induction, he discovers a rule, or merely presents statements

which provide the answers to these problems. But, surely, a rule or statement is always something that refers to a population. In experimental research this population is an abstraction. Therefore, the research worker cannot look upon his experimental material as fixed, because, if he does so, the population cannot be an abstraction.

In statistical theory we are taught that a generalization is justified only if some units or replications are – or can be regarded as – a random sample. However, if the population is an abstraction no random sample can be drawn from it, since the act of drawing a sample requires that the population be an existing one. Therefore the only possibility left to the research worker is to *regard* the sample as random, i.e. as being a random representation of the population about which inferences are drawn. This is, in fact, the population with which investigators in other fields of research most often have to be satisfied. But neither in experimental nor in other research does this mean that the scientist has to be content with any sample. To acquire satisfactory samples is most important in every research program.

Throughout this treatise we shall regard experimental material as random in the sense that it consists of a number of replications which can be considered as a random sample. We do not see that any serious objections can be raised against this point of view, even if there might be difficulties to overcome in some cases, e.g. in field plot experimentation. On the other hand, it is evident that research workers who regard their experimental material as non-random are certain to encounter serious difficulties in their interpretation of the results of the experiment.

(2) Turning next to the second point, it seems evident that the most common view among statisticians who accept the Rothamsted principles is that the testing of null hypotheses is the principal purpose for which an experiment is carried out. In *The Design of Experiments* (6th edn, p. 16) Fisher writes: "Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis." Even if this point of view is often regarded as extreme, it is, in the main, followed by most writers of papers and text-books dealing with experimental design and statistical analysis. But such extreme and unrealistic views are not shared by all writers. In some

treatises, problems concerning the estimation of treatment effects and differences between such effects are considered as no less important than those of testing null hypotheses. It may also be demonstrated that Fisher's point of view is not shared by some independent research workers.

The function of an experiment, as we have seen, is the production of data that can be used in order to find the answers to particular questions. What these questions are, is the concern of the research worker. In a discussion of the methodology of experimental research, it must be emphasized that the questions should be asked in advance of the designing and carrying out of the experiment, and in order to answer them it is necessary to test statistical hypotheses and/or to estimate treatment effects and the differences between such effects.

For the testing of statistical hypotheses and the estimation of treatment effects, a number of apparently satisfactory methods have been invented. But, on the whole, it can hardly be maintained that the situation is satisfactory in the sense of fully meeting the requirements of research workers.

Heterogeneity of experimental material now seems to be commonly accepted; it has been known and discussed at considerable length by several writers, and was in fact discovered before the work on experimental design was begun at Rothamsted. It is, of course, the combined effect of a number of factors which are not under control of the research worker. These factors affect the experimental units in the same way as do the experimental factors, and therefore interactions between the two groups of factors must be assumed to exist. It can be noted as a rather curious circumstance that writers who are much concerned with the possible interactions between experimental factors tend to disregard the interactions between experimental factors and heterogeneity factors. But to proceed as if such interactions did not exist would be to assume a too simple and unrealistic model of nature.

(3) The model describing the null hypothesis can be written in any way, provided it is capable of being tested. But if it is unrealistic, the implications of the rejection of the null hypothesis may become very involved. The usual models of null hypotheses presume additivity of treatment effects and effects of heterogeneity

factors. Such models may allow of strict mathematical treatment, but they are lacking in realism. In dealing with the estimation of treatment effects and the differences between such effects, it is even more important that the model should be realistic. Therefore, models that do not account for interactions between treatments and heterogeneity factors should never be accepted.

2 Treatments, questions and randomization

To apply a certain treatment to an experimental unit means, of course, that it is being applied according to a fixed procedure. Therefore, it is impossible to repeat a treatment if, by this, perfect repetition is understood. A treatment can only be repeated in the sense that a particular description of the treatment is fulfilled. Therefore, even if it were possible to find a number of experimental units that are exactly alike, the same treatment applied to these units would not produce exactly the same effect. However, no two units of experimental material are exactly alike; all kinds of experimental material are more or less heterogeneous. There is always, therefore, some variation in the effect of the same treatment among a number of experimental units. The most important factor causing this variation is usually heterogeneity of the material, but the failure of the treatment to be exactly repeated plays some part. There are also errors of observation.

Suppose now that the units of the experimental material are divided into two samples, and that the same treatment is applied to the units in both samples. Then, in order that the distributions of the observed random variable should be identical in the populations represented by the two samples, it is necessary that the division be carried out by some technique of randomization. If such a technique has not been used, and a treatment T_1 is applied to the units in the first sample and another treatment T_2 is applied to those in the second sample, we have no guarantee that a comparison of the effects of the two treatments will turn out to be unbiassed. Division of the material in a non-random way will therefore very often lead to false conclusions with regard to the relative effects of the two treatments. In spite of the fact that this consequence has been known for the last thirty years, research workers still try to get round it by claiming that other ways of dividing the material lead to more precise comparisons, and forgetting the bias. In the final section of this treatise we will return to a particular aspect of the principle of randomization. Until then, we shall assume that the principle has been consistently applied.

The purpose for which an experiment is planned and carried out is the concern of the research worker. But if our intention is to specify the method of statistical treatment of the experimental data, a general classification of the questions can be framed. The following three groupings should be satisfactory for all situations:-

(1) The treatments are qualities, and the principal question concerns the ranking of them on the outcome of the experiment.

(2) The treatments are qualities and/or quantities, and the question concerns the differences of the effects as between treatments chosen in advance.

(3) The treatments are quantities, and the question concerns the rule, if any, describing the way in which the effect depends on these quantities.

In answering such questions, it is obviously important that the experimental material should be such that the answers can be applied to a population of reasonable size. It is evident that the material can be chosen in such a way that small differences, unimportant in themselves, may turn out to be statistically significant. Moreover, there is probably always some difference between the effects of two treatments, so that the null hypothesis can be rejected only by choosing experimental material having sufficiently small heterogeneity.

The research worker should therefore always ask himself what he is going to do with the results of the experiment. It is important to know whether the results are intended to be used for some practical purpose,* or whether the intention is to supplement the investigator's insight and knowledge in some field. Experimental material which serves the latter purpose might be unsatisfactory for the first. There is also the possibility of describing the population to which the inferences are intended to be applied, even if the description turns out to be vague. Such a description

* This may merely mean that the experiment is part of a research program having a practical purpose. In Section 17 we return to more general problems concerning experiments of this kind.

would refer both to the experimental material and to the external circumstances under which the experiment has been carried out.

In treatises on methodology today there is usually a demand for efficiency. But, obviously, the choice of a design that is more efficient than another almost always implies a reduction in size of the relevant population and a reduction of the generality of the inferences. The consequence is that the interpretation of the result obtained with a more efficient design will not usually be the same as that obtained with a less efficient design. Therefore, the common and general recommendation that the most efficient design should be used is open to objection.

3 Complete randomization

Suppose that an experimental material consists of $2n$ units or replications and that the experimenter divides it, in a random way, into two samples, each sample consisting of n units. Then, if one of the samples is used for treatment T_1 and the other for treatment T_2 , and the treatments are allocated to the samples in a random way, the research worker can be confident that the difference between the effects of the two treatments (the "contrast") can be estimated without bias. Therefore, the most important requirement of estimation is fulfilled. Also, confidence limits of the contrasts can be computed.

The generalization to $k > 2$ treatments is simple and straightforward: the experimental material consisting of nk units is divided randomly into k samples, and the k treatments are randomly allocated to the samples. In this case also a contrast between treatments can be estimated without bias.

It is hardly possible to deal with any experimental situation without the aid of a model that gives a general description of the possible outcomes of the experiment. In the present case, with k treatments T_j $(j = 1, 2, \dots , k)$ and n experimental units for each treatment, the model is:

$$(3.1) \qquad x_{ji} = \mu + a_j + e_{ji} \qquad (i = 1, 2, \dots , n)$$

In this model, the x_{ji}* are the observations, μ is a general mean, and a_j are effects of the treatments. Without loss of generality we can let $\Sigma\, a_j = 0$ because, if $\Sigma\, a_j \neq 0$, the a_j contain a common element that can be included in μ.

The e's are ordinary random variables. Without loss of generality it can be assumed that $E(e_{ji}) = 0$, and we may also assume that the form of the distribution of e is the same for all treatments. But it cannot be assumed that the k distributions are identical. Such an assumption would imply that all effects of

* Here and in the following sections we shall use the same letter to denote a random variable and the observation of it. This simplification can hardly lead to confusion.

the treatments are included in a_j, and this would be too simple a view to take of the rather complicated mechanism that usually regulates the effect of a treatment.

The differences between the k distributions of e may be differences in skewness and differences in kurtosis. But the differences that are most important for the analysis of the experimental data are differences in the variance of e among the treatments. This means that the research worker, in his analysis of the data, has to deal with k variances, $\text{var}_j(e)$. If the necessary caution is exercised during the planning and administration of the experiment, the e's can be regarded as being stochastically independent both within and between the treatments, and $\text{var}_j(e)$ can therefore be estimated in the usual way.

It will be found that the mean of x_{ji} for treatment T_j is equal to

$$\bar{x}_j = \mu + a_j + \bar{e}_j \,. \tag{3.2}$$

Since $E(e_{ji}) = 0$, it will be seen that $E(\bar{x}_j) = \mu + a_j$, showing that $\bar{x}_j$ is an unbiassed estimator of the effect of T_j. Therefore, the means yield an unbiassed ranking of the treatments.

A contrast is by definition a linear function of a_j or a linear function of a sub-set of these parameters (cf. page 20). The difference $(a_p - a_q)$ is an example. It will be seen that

$$\bar{x}_p - \bar{x}_q = (a_p - a_q) + (\bar{e}_p - \bar{e}_q) \tag{3.3}$$

and hence that the difference between the means is an unbiassed estimator of the contrast. It will also be found that the variance of the difference is equal to

$$\text{var}(\bar{x}_p - \bar{x}_q) = [\text{var}_p(e) + \text{var}_q(e)]/n \,.$$

Therefore, unless $\text{var}_j(e)$ is a constant, the precision of the estimator of a contrast is not the same for all contrasts. Thus, the common practice of using the same error mean square for the computation of the confidence limits of all contrasts is not to be recommended. The research worker can never know that $\text{var}_j(e)$ is the same for all treatments. On the contrary, it is very unlikely that this variance is ever a constant.

If the distribution of e is normal and

$$V_j = \frac{1}{(n-1)} \sum (x_{ji} - \bar{x}_j)^2 ,$$

approximately correct confidence limits of the contrast $(a_p - a_q)$ are:

$$(3.4) \qquad (\bar{x}_p - \bar{x}_q) \mp t_\alpha \sqrt{\{(V_p + V_q)/n\}},$$

where t_α is the tabulated significance point of "Student's" t, the number of degrees of freedom being $2(n - 1)$. That the limits are approximately correct means, of course, that the probability of the interval covering the contrast is approximately equal to $(1 - \alpha)$.

Usually, however, the research worker wants to estimate more than one contrast. If two contrasts are $(a_p - a_q)$ and $(a_r - a_s)$, where $p \neq q \neq r \neq s$, no difficulty is involved. But the worker may want to deal with, e.g., the contrasts $(a_p - a_q)$ and $(a_p - a_r)$ simultaneously. In this case the two estimators $(\bar{x}_p - \bar{x}_q)$ and $(\bar{x}_p - \bar{x}_r)$ are correlated. The same is the case with $(V_p + V_q)$ and $(V_p + V_r)$. Nevertheless, the probability of the intervals

$$(\bar{x}_p - \bar{x}_q) \mp t_\alpha \sqrt{\{(V_p + V_q)/n\}} \text{ and } (\bar{x}_p - \bar{x}_r) \mp t_\alpha \sqrt{\{(V_p + V_r)/n\}}$$

simultaneously covering the contrasts $(a_p - a_q)$ and $(a_p - a_r)$ is approximately equal to $(1 - \alpha)^2$. As will be shown in Sections 6 and 7, this implies that, if we compute the confidence limits of the two contrasts in the way described, the confidence probability of each of the two intervals is but slightly different from $(1 - \alpha)$.

It will also be shown that this result can be generalized to cover k treatments and $(k - 1)$ contrasts, or that there is ample ground for such a generalization. It is very important, however, that a separate error mean square should be used for each contrast.

In the methodology as it is usually presented, much emphasis is placed on the so-called "orthogonal functions" of the treatment means. For instance,

$$y_1 = \bar{x}_1 - \bar{x}_2 \quad \text{and} \quad y_2 = \bar{x}_1 + \bar{x}_2 - 2\bar{x}_3$$

are regarded as being orthogonal, i.e. non-correlated. It is easy to show, however, that the two functions are orthogonal only if $\text{var}_j(e)$ is the same for $j = 1$, 2 and 3. In practice it would therefore be rather rash to regard them as being orthogonal. But, in defence of the use of such functions, it must be pointed out that it is reasonable to assume that the correlation between them is weaker than the correlation between other functions, and that they may be preferred for that very reason. The difficulty is that they very seldom correspond to the actual questions being asked by the research worker.

4 Randomized blocks

In a randomized block experiment a replication is a group of experimental units, and the number of units per replication is usually chosen to be equal to the number of treatments. For instance, in a feeding experiment in which a pig is an experimental unit, a litter can be used as a replication. In a field experiment, the experimental area is divided into a number of smaller areas of equal size, the "blocks" or "replications", and each of these into a number of "plots" (the units). In these cases randomization means complete randomization within each replication.

Here also the replications must be regarded as a random sample. Thus the population is the one which the sample of replications represents – in the sense of a random sample – and it is an abstraction. In our first example this idea is easily conceived, as the sample of litters might actually have been drawn at random from an existing population of litters, which in turn can be regarded as a random representation of an abstract population.

In our second example the idea might be more difficult to accept. However, suppose a research worker is planning a local field-plot experiment, and that the total cultivated area of a farm is placed at his disposal. He can then divide the whole area into a number of blocks of the size he wishes to use, and from this existing population of blocks he can draw at random a sample of blocks. Having drawn this sample, he might find that the blocks belonging to it are scattered over the whole area of the farm. He may then decide that this sample is too troublesome to use in practice, and for that reason choose another sample having the practical advantage that the blocks are adjacent. It is evident that this latter sample will usually represent, in the sense of a random sample, an abstract population less general than the one represented by the randomly drawn sample. Nevertheless, the chosen sample of blocks can be regarded as a random representation of *some* abstract population. Usually this population is rather narrow, and therefore the inferences (if any) that are drawn from the experimental data can be applied in a small range only.

This idea is not a new one. Somewhat hesitantly, it has been put forward by several authors. However, it is a fact – and in our opinion a regrettable one – that this way of thinking has not been found worthy to be followed up.

In this case there are always two components of heterogeneity of the experimental material: heterogeneity among the units within the replications, and heterogeneity among the replications. Therefore, we must deal with "intra-block" and "inter-block" heterogeneity factors. They are not necessarily different factors *per se*. In a field experiment they are usually the same factors. Nevertheless, it is necessary to distinguish between them because of the interactions between the treatments and these factors.

Suppose that the number of treatments is k, the number of replications is n, and let $j = 1, 2, \ldots, k$, $i = 1, 2, \ldots, n$. Then the general model for the experimental data is

$$x_{ji} = \mu + a_j + z_i + u_{ji} + e_{ji}. \tag{4.1}$$

In this model μ and a_j are parameters, and z, u, and e are random variables. Without loss of generality we can let $\Sigma a_j = 0$ and $E(e) = 0$ for each j and i. However, since e is an effect of the intra-block heterogeneity factors, and therefore also covers the interactions between the treatments and these factors, the distribution of e must be taken to be different for the different treatments, implying, for example, that $\text{var}(e)$ is not the same for all treatments.

The variables z and u are both effects of the inter-block heterogeneity factors: z being the effect common to all treatments, and u the interactions between the treatments and the heterogeneity factors. Without loss of generality we can let $E(z) = 0$ and $E(u) = 0$ for each j. But in other characteristics (e.g. the variance) the distribution of u must be assumed to be dependent on the treatments. It is important to notice that z and u cannot be taken to be independent variables, and that the u's cannot be regarded as being independent among themselves, although, of course, some of the u's might be mutually independent. In saying that correlations are present, we do not mean that such is the case for all comparisons and under all circumstances. It is evident, however, that the research worker can never know that such correlations do not exist; he must therefore use such statistical treatment of the experimental data as allows for these correlations.

It will be found that the mean of x for treatment T_j is equal to

$$(4.2) \qquad \bar{x}_j = \mu + a_j + \bar{z} + \bar{u}_j + \bar{e}_j ,$$

and – since $E(z) = E(u) = E(e) = 0$ – that $E(\bar{x}_j) = \mu + a_j$. This shows that the mean is an unbiassed estimator of the effect $(\mu + a_j)$, and hence that the means yield an unbiassed ranking of the treatments.

For $j = p$ and $j = q$ it will be found that

$$\bar{x}_p - \bar{x}_q = (a_p - a_q) + (\bar{u}_p - \bar{u}_q) + (\bar{e}_p - \bar{e}_q)$$

and hence that $E(\bar{x}_p - \bar{x}_q) = a_p - a_q$, i.e. the difference between the means is an unbiassed estimator of the contrast. On account of the interactions, the variance of the difference cannot be taken to be the same for all contrasts, and an individual estimate of the variance must therefore be used for each contrast. If for each replication we use the difference $d_i = x_{pi} - x_{qi}$, it will be found that $\bar{d} = \bar{x}_p - \bar{x}_q$, and the variance is estimated by V/n, where

$$V = \frac{1}{n-1} \sum (d_i - \bar{d})^2.$$

Owing to the robustness of "Student's" t, the research worker can be confident that the probability of the interval

$$(4.3) \qquad \bar{d} \mp t_\alpha \sqrt{(V/n)}$$

covering the contrast $(a_p - a_q)$ is approximately equal to $1 - \alpha$.

The method of computing the confidence limits can be used for any contrast. But in this case also, the research worker usually wants to estimate more than one contrast. On account of the interactions between the treatments and the inter-block heterogeneity factors, the estimators of the different contrasts are correlated, having different variances. Nevertheless, the confidence probability of each of the intervals, the limits of which are computed as described, is but slightly different from $(1 - \alpha)$. We return to this statement in Section 7, to which the reader is referred.

It is evident that if the number (n) of replications is small, the precision of the estimator of a contrast is usually very low. It is true, of course, that even if n is very small, interesting inferences might be drawn. But usually these inferences are such as are obtained through the rejection of the null hypothesis. If the research worker is interested in the estimation of contrasts, and the number of replications is very small, he cannot expect to find the estimators precise enough to serve any reasonable purpose.

Of course, this is also the case when there has been complete randomization. However, if the number of experimental units for each treatment is the same as if a randomized block design had been used, the number of degrees of freedom is greater for the first than for the latter plan, i.e. $2(n - 1)$ for the first and $(n - 1)$ for the latter. For small n this difference makes an important difference in the value of t_α. This difference may, however, be more than counterbalanced if the inter-block heterogeneity is materially greater than the intra-block heterogeneity. Therefore, the precision of randomized blocks, as compared with complete randomization, depends both on the value of n and on the difference between the inter- and intra-block heterogeneity. Thus, if n is small, the arrangement of the experimental units into blocks must remove a very large part of the heterogeneity in order that the difference in the value of t_α can be expected to be neutralized.

Having carried out a randomized block experiment, the research worker may find that some observations are missing, or that they deviate so widely from the rest of the observations that it is reasonable to doubt whether they are correctly recorded. Such results may happen through failure to record, or through gross errors of various kinds.

In order to restore the orthogonality of the observations, techniques known as “missing plot” techniques have been invented, which presume additivity of the effects of the treatments and the heterogeneity factors. Since we do not regard such a model as a realistic one, and the experimenter cannot know whether it is realistic, we do not recommend these techniques. It is obvious that if the research worker is engaged in the estimation of contrasts, the use of such techniques is unnecessary. If one or more observations are missing for two treatments T_p and T_q, and he wishes to estimate the contrast $(a_p - a_q)$, he should be content with the observations that he has obtained and accepted.

If the experimenter is interested in carrying out an analysis of variance and an F-test, it might do no harm to replace a few observations by means of a missing plot technique. But even then the use of such a technique is not essential as there are always some parts of the observations which are orthogonal. For these parts an analysis of variance can be carried out and, if necessary, the observations for the other treatments can be linked to the orthogonal parts by means of linear functions. Even if the number of degrees of freedom for the error mean square is reduced by one unit for each restored observation, it seems evident that the use of a missing plot technique to any large extent might falsify the result of the analysis.

The situation might be much more difficult to deal with if an observation seems to be faultily recorded. In some instances an observation may differ so greatly from what would be expected that there can be no doubt that a gross error in recording has been made; it would then be reasonable to treat the observation as a missing datum. However, there are cases in which the worker may be in doubt concerning the reliability of the record, and it may then be difficult to say what should be done. The most unsatisfactory procedure in such a case would be to use a missing plot technique. An apparently faulty observation may be due to interaction between the treatment and the heterogeneity factors, and the use of a technique which is worked out under the assumption of additivity might therefore lead to false conclusions.

5 The role of mathematics

If by "statistics" is meant "method of research", statistics is not merely applied mathematics. However, mathematics has played, and continues to play, an important role in the development of statistics and research method – and this must necessarily be so. But research workers should always remember that a mathematical deduction needs some premises. It should also be remembered that such premises as it has been necessary to use are rarely in keeping with the actual experimental situation.

This usually implies that the result obtained by mathematical deduction, if it holds any interest whatever, is merely part of the development of a research method. In one way or another the result has to be tested in order to find out whether its use is limited to cases satisfying the premises, or whether it can safely be applied in a wider field. In general, the premises that are used are too limited in scope to justify classification of the result of a mathematical deduction as a method of research.

For instance, consider the distribution of the statistic t developed by W.S. Gosset ("Student") [17], for which a rigorous proof was given by R.A. Fisher [15]. An important premise for the mathematical deduction was that the observed random variable is normally distributed. There are several grounds for doubting the realism of this premise. It is hardly possible that any random variable exists which is exactly so distributed. Certainly, a large number of actual random variables are found, the distributions of which closely resemble the normal form, but there are also actual distributions that deviate considerably from this model. In consequence, the distribution of t, as developed by Gosset, had to be tested. On the whole, the results of these tests are satisfactory, and the t-distribution is therefore now commonly accepted as a tool of research in a very wide field.

In the development of a statistical method there are usually two elements : mathematical deduction from chosen premises, and the testing of the result of the deduction in order to see whether or not the premises are important. Statistics, as it is presented today, consists of a mass of techniques that are never tested satisfactorily, if at all. This may be due to the fact that most people find mathematical deductions more interesting and entertaining than the very tedious work involved in the testing of techniques. However, with the development of electronic computers the testing of techniques has been greatly simplified, so that research workers can now look forward to interesting and useful developments.

In the present treatise some new techniques are suggested. We have tried to test them as elaborately as possible, but we have not had the use of computer facilities to the extent we would have wished. Therefore, results from further tests would be very welcome.

6 Simultaneous statistical inferences

Suppose that m independent experiments have been carried out by one or a number of research workers, for the specific purpose of producing data from which a certain parameter can be estimated. Moreover, suppose that the confidence limits of the parameter are computed for each of the m cases, and it is stated for each case that the value of the parameter is covered by the confidence interval. Then the probability of r correct statements is given by the binomial

$$(6.1) \qquad P_r = \binom{m}{r}(1 - \alpha)^r \alpha^{m-r}$$

where $(1 - \alpha)$ is the chosen confidence probability. Therefore, the expected number of correct statements is $m(1 - \alpha)$. It is also worth noticing that the probability of all the statements being true is $P_m = (1 - \alpha)^m$, and the probability of at least one false statement is $1 - (1 - \alpha)^m$. Consequently, in a very large number (m) of cases, the probability of all statements being true approaches zero, and the probability of at least one false statement approaches unity.

These results are consistent with the conclusion that, if the number of cases is large enough, at least two confidence intervals will be found that do not overlap, and hence that at least two statements refute each other. It is fairly easy to see that the results can be extended to cases in which different parameters are being estimated.

Now suppose that the research worker wants to estimate two parameters, θ_1 and θ_2. Then, in order to obtain two confidence intervals that are consistent with (6.1), he should carry out two independent experiments, one for the purpose of estimating θ_1 and one for the purpose of estimating θ_2. However, this would be too expensive. Therefore he has to be content with one experiment, the consequence being that the data which are used for the estimation of the parameters are not stochastically independent. This fact raises the question of how confidence limits of the contrasts ought to be computed. Several methods have been suggested. We refer to the summary given by Federer [10], to Mood and Graybill [26], to Miller [25], and to the literature cited in these treatises.

A solution has been sought in what is termed the experimentwise confidence coefficient, which is the probability of the confidence intervals of all possible contrasts simultaneously. Mood and Graybill [26, p. 268] write: "If in 95 per cent of the experiments each of the $t(t - 1)$ confidence intervals covers its respective difference $(\mu_i - \mu_j)$, we shall say that the experimentwise confidence coefficient is $\cdot 95$." These attempts to find a solution to an intricate problem give rise to the following questions and objections.

There must be an upper limit to the number of contrasts, less than the total number of possible contrasts, that can be immediately estimated. It is easy to see that this limit is $(k - 1)$, where k is the number of treatments.

As we have seen, a contrast is by definition a linear function of the parameters $\theta_j = \mu + a_j \quad (j = 1, 2, \dots, k)$, i.e.

$$C_p = \sum A_{jp}\theta_j = \sum A_{jp}a_j$$

for which $\Sigma A_{jp} = 0$. If a set of $(k - 1)$ contrasts is chosen in such a way that none of the contrasts can be derived from the others, all other contrasts are linear functions of sub-sets or of the whole set of the chosen contrasts. This implies that the estimates of C_p for $p \geqslant k$ can be derived from the estimates of C_p for $p < k$. The confidence limits of C_p for $p \geqslant k$ cannot be derived from the confidence limits of C_p for $p < k$, but the central values of the confidence intervals can be regarded as derived estimates. Therefore our argument also holds for the confidence intervals. This conclusion is consistent with the well-known fact that the treatment mean square in the analysis of variance can be divided into $(k - 1)$ components.

The use of the experimentwise confidence technique implies that the limits of the confidence intervals are computed in such a way that the probability of all intervals covering the contrasts is equal to $(1 - \alpha)$, e.g. $0{\cdot}95$. This means that the confidence probability of the intervals simultaneously covering the contrasts is chosen independently of the number of contrasts. It also implies that the confidence probability of the interval of any one individual contrast depends on the number of contrasts. For these reasons the principle can hardly be accepted.

The research worker is primarily concerned with the confidence probability of each individual interval, and this probability ought to be chosen by himself. But at the same time he must also consider the confidence probability of all intervals simultaneously. If the $m \leqslant k - 1$ contrasts are estimated by means of data from m independent experiments, the scientist can choose the confidence probability $(1 - \alpha)$ for the individual intervals, and he will know that the probability of $r = 0, 1, 2, \dots, m$ intervals covering the contrasts is given by (6.1). There is no doubt, however, that (6.1) also applies in cases in which the data are obtained in a single experiment. Consider, for example, a randomized block experiment for the comparison of three treatments (T_1, T_2, and T_3). Then, if the two contrasts are $(a_1 - a_2)$ and $(a_2 - a_3)$, the differences $(x_{1i} - x_{2i})$ and $(x_{2i} - x_{3i})$ should be used (cf. Section 4). It may happen that the two differences are observations of two independent random variables, and if it is known in advance that this is so, the differences can be used as if $(x_{1i} - x_{2i})$ and $(x_{2i} - x_{3i})$ were data obtained in two independent experiments.

If the confidence limits are correlated among the contrasts, as will usually be the case if the data come from the same experiment, (6.1) cannot be expected to apply exactly. But fortunately, in the cases dealt with in Sections 3 and 4, the effect of the correlations among the contrasts is small and has no practical significance.

7 The estimation of contrasts

It will now be assumed that in planning the experiment, the research worker has decided on the contrasts he wants to estimate. If the number of these contrasts is $k - 1$, the experiment must be carried out with k treatments, as in the preceding section.

The usual methods for computation of the confidence limits of a contrast rest on the assumption that the effects of the treatments and the heterogeneity factors are additive. The confidence limits of the contrast are therefore computed by means of the error mean square for the whole experiment. As the assumption of additivity is unrealistic, this method is lacking justification and, if it is used, the research worker cannot know the probability of the confidence interval. He should therefore adopt the methods described in Sections 3 and 4. Then, choosing the value of α in advance (e.g. $\alpha = 0{\cdot}05$) and using these methods, he can be reasonably certain that he is working on a confidence level that is very close to $(1 - \alpha)$.

However, in practice the research worker usually wishes to estimate more than one contrast. In fact, if k treatments have been included in the experiment and the principal purpose is to estimate contrasts, the reason for including k treatments must be that he has decided upon $k - 1$ contrasts. Then the problem is to decide which method should be used so that the probability of the $k - 1$ confidence intervals simultaneously covering the contrasts be equal to $(1 - \alpha)^{k-1}$ (cf. Section 6). It will now be shown that, in spite of the correlations, and to the extent to which our data can be relied upon, the methods given by (3.4) and (4.3) approximately satisfy this requirement.

Suppose that the experiment is a randomized block experiment with $k = 3$ treatments and n replications. Let the two contrasts be $C_1 = a_1 - a_2$ and $C_2 = a_2 - a_3$. The unbiassed estimators of these contrasts are $\bar{d}_1 = \bar{x}_1 - \bar{x}_2$ and $\bar{d}_2 = \bar{x}_2 - \bar{x}_3$, d_1 and d_2 being as defined in Section 4. Let V_1 and V_2 be the two relevant mean squares (cf. Section 4), σ_1^2 and σ_2^2 the corresponding population variances, r the sample correlation coefficient, and ρ the population correlation coefficient between d_1 and d_2. Then, assuming that d_1 and d_2 are both normally distributed, it will be found that the joint distribution is

$$(7.1) \qquad f(t_1, t_2, V_1, V_2, r)$$
$$= Q(V_1 V_2)^{1/2(n-2)}(1 - r^2)^{1/2(n-4)} \exp\left\{\frac{-M}{2(1-\rho^2)}\right\},$$

where Q is a known constant,

$$t_1 = \frac{\bar{d}_1 - C_1}{\sqrt{(V_1/n)}} \quad \text{and} \quad t_2 = \frac{\bar{d}_2 - C_2}{\sqrt{(V_2/n)}},$$

and

$$M = \frac{t_1^2 + n - 1}{\sigma_1^2} V_1 + \frac{t_2^2 + n - 1}{\sigma_2^2} V_2 - 2\rho\{t_1 t_2 + (n-1)r\} \frac{\sqrt{(V_1 V_2)}}{\sigma_1 \sigma_2}.$$

The probability that $|t_1|$ and $|t_2|$ are simultaneously less than $|t_\alpha|$, where t_α is the significance point of "Student's" t for $n - 1$ degrees of freedom, is then equal to the integral:

$$A = \int \ldots \int f \, . \, dt_1 \, dt_2 \, dV_1 \, dV_2 \, dr.$$

The integration intervals are: $-t_\alpha \leqslant t \leqslant + t_\alpha$, $0 \leqslant V \leqslant \infty$, and $-1 \leqslant r \leqslant + 1$.

For given values of σ_1 and σ_2, A depends on ρ and n. It can be shown that for any n, A is a minimum for $\rho = 0$, the minimum being equal to $(1 - \alpha)^2$. In order to find to what extent A depends on ρ and n, numerical integrations have been carried out for $\alpha = 0{\cdot}05$, $\sigma_1 = \sigma_2 = 1$ and some chosen values of n and ρ. The results for $A^{1/2}$ are shown in Table 7.1. It will be seen that the values are but slightly larger than $1 - \alpha = 0{\cdot}95$, indicating that the effect of n and ρ on the confidence probability is too small to be of practical importance.

Table 7.1 $\sqrt{A}$

n	ρ		
	0·3	0·6	0·9
4	0·950	0·953	0·959
8	0·951	0·955	–
15	0·951	0·955	–

Turning next to an experiment assumed to be carried out according to the principle of complete randomization, we shall consider the contrasts $C_1 = a_1 - a_2$ and $C_2 = a_2 - a_3$. The estimators are $\bar{x}_1 - \bar{x}_2$ and $\bar{x}_2 - \bar{x}_3$, which are both unbiassed. Let

$$t_1 = \frac{\bar{x}_1 - \bar{x}_2 - C_1}{\sqrt{\{(V_1 + V_2)/n\}}} \quad \text{and} \quad t_2 = \frac{\bar{x}_2 - \bar{x}_3 - C_2}{\sqrt{\{(V_2 + V_3)/n\}}},$$

where the V's are the usual mean squares. Then, assuming that the observed random variable is normally distributed, the joint distribution $f(t_1, t_2, V_1, V_2, V_3)$ can be derived. Let

$$A = \int \dots \int f . dt_1\, dt_2\, dV_1\, dV_2\, dV_3 ,$$

the integration intervals being $-t_\alpha \leqslant t \leqslant t_\alpha$ and $0 \leqslant V \leqslant \infty$, where t_α is the significance point of "Student's" t for $2(n-1)$ degrees of freedom. Numerical computations of this integral have been carried out for $\sigma_1 = \sigma_2 = \sigma_3 = 1$, $\alpha = 0{\cdot}05$, and for three chosen values of n. The results for $A^{1/2}$ are shown in Table 7.2. It will be seen that in this case also the values are but slightly larger than $1 - \alpha = 0{\cdot}95$.

Table 7.2

n	$2(n-1)$	$A^{1/2}$
3	4	0·953
5	8	0·953
10	18	0·954

The implication of these results (Table 7.1 and 7.2) is that having chosen the value of α and computed the confidence limits of the two contrasts in the way described, i.e. by (4.3) and (3.4), the research worker can be satisfied that the probability of the two confidence intervals simultaneously covering the contrasts is approximately equal to $(1-\alpha)^2$. This means that, in spite of the correlation, the confidence probability of each of the two intervals is approximately equal to $1 - \alpha$.

It is obvious that the scope of these results is rather limited. It has been assumed that the random variable is normally distributed, and that there are no interactions between the treatments and the heterogeneity factors. Furthermore, no more than

$k = 3$ treatments have been included. In order to widen the scope, such computations might have been extended to cases covering larger numbers of treatments and non-normal random variables. The computations should also have been carried out for different values of α. Lack of facilities, however, have prevented an extension in these directions. As a substitute, we have carried out tests by means of artificial examples.

Three examples of randomized block experiments were constructed by means of Wold's table of normal deviates – see Wold [37]. The rows in this table were then regarded as representatives of the replications. If h stands for the column number, the examples were constructed according to the model

$$x_{ji} = \mu + \beta_j z_{1i} + z_{hi},$$

where the z's are the normal deviates, $i = 1, 2, \ldots, n = 5$, $h = 2, 3, \ldots, (k + 1)$, and $j = h - 1$. In Examples 1 and 2, β_j was chosen equal to unity for all j. In Example 3 the chosen values of β_j were:

$$(-10), (-5), (10), (20), (25), \text{and } (30)$$

for treatments $T_1, T_2, \ldots, T_6$.

It will be seen that in the first two examples additivity is assumed, while in the third example interactions between the treatments and the inter-block heterogeneity factors are included. Confidence limits of the contrasts $a_j - a_{j+1} = 0$ were computed by (4.3), using the observed differences $d_{ji} = x_{ji} - x_{(j+1)i}$.

Let r stand for the number of confidence intervals that do not cover the contrast. Then, if the correlations between the d's among the contrasts do not affect the confidence probability, the probability of $(k - 1 - r)$ intervals covering the contrast will be the binomial (cf. Section 6)

$$f(r) = \binom{k-1}{r} \alpha^r (1 - \alpha)^{k-1-r},$$

and the expected number of such intervals will be $Nf(r)$, where N is the number of samples. In Table 7.3 the observed number (n_r) and the expected number of such samples are compared for each of the three examples.

Let $1 - \alpha'$ be the probability of the confidence interval of a single contrast regarded alone. Then, if the above-mentioned

Table 7.3 $n = 5$ **blocks,** k **treatments,** $\alpha = 0{\cdot}05$

r	Example No.					
	1 ($k = 4$)		2 ($k = 10$)		3 ($k = 6$)	
	n_r	$Nf(r)$	n_r	$Nf(r)$	n_r	$Nf(r)$
0	159	159·47	61	63·02	76	77·38
1	25	25·18	29	29·84	22	20·36
2	2	1·35	10	7·14	2	2·26
N	186		100		100	
n_0/N		0·855		0·610		0·760
$0{\cdot}95^{k-1}$		0·857		0·630		0·774
$(n_0/N)^{1/(k-1)}$		0·949		0·947		0·947
$1-\bar{r}/(k-1)$		0·949		0·946		0·947

correlations do not affect the confidence level, the confidence probability of all intervals simultaneously covering the contrasts is equal to $(1 - \alpha')^{k-1}$, the estimator of which is n_0/N. Thus the estimator of $1 - \alpha'$ would be $(n_0 /N)^{1/(k-1)}$. The latter estimator is not unbiassed, but if the number (N) of samples is large, it will give a fairly satisfactory approximation.

On the other hand, if the correlations do not affect the distribution of r (cf. Table 7.3), then $1 - \bar{r}/(k-1)$, where $\bar{r}$ is the arithmetic mean of r, is also an unbiassed estimator of $1 - \alpha'$. However, the correlations do, in fact, change the distribution of r to some extent, and therefore not even the latter estimator of $1 - \alpha'$ is quite satisfactory. Accordingly we have used both estimators in our examples. It will be seen from Table 7.3 that for the three cases considered, the values of both estimators are very close to the chosen value of $1 - \alpha$, i.e. $0{\cdot}95$. That this is so in other cases as well is shown by the following examples.

In Examples 4 and 5 the experiments were carried out according to the principle of complete randomization, and in both examples the additive model was used. For Example 4 ($n = 5$, $k = 3$), the observations were taken from Wold's table of normal deviates in the same way as in the first two examples, but now the values in a column were regarded as observations in a one-way classification. The estimated contrasts were $(a_1 - a_2)$

and $(a_2 - a_3)$. In our fifth example ($n = 5$, $k = 5$) the observations were taken in exactly the same way from the table presented by Quenouille [33], column 8, which are sampled from the two-sided exponential distribution. The estimated contrasts were $(a_1 - a_2)$, $(a_2 - a_3)$, $(a_3 - a_4)$, and $(a_4 - a_5)$. In both examples the confidence limits of the contrasts were computed by (3.4), using separate mean squares for the different contrasts. The results are recorded in Table 7.5, page 29.

This part of our investigation was carried out several years ago, when we had no access to an electronic computer. We were therefore compelled to use existing tables of random values, and small numbers (N) of samples. In our last six examples, however, most of the work has been carried out on an electronic computer.

In the last six examples the samples were drawn from the distribution

$$f(z) = Rz^a (10 - z)^b \qquad 0 \leqslant z \leqslant 10.$$

In Examples 6 and 9 : $a = b = 2$, $\quad E(z) = 5$

In Examples 7 and 10: $a = 2$, $b = 4$, $\quad E(z) = 3{\cdot}75$

In Examples 8 and 11: $a = 0$, $b = 2$, $\quad E(z) = 2{\cdot}5$.

In Examples 6, 7 and 8 the experiments were carried out according to the principle of complete randomization, and the model was

$$x_{ji} = \mu + \beta_j \, [z_{ji} - E(z)] \qquad \begin{Bmatrix} i = 1, 2, \dots, 10 \\ j = 1, 2, \dots, 10 \end{Bmatrix}.$$

The values of β_j were, for $j = 1, 2, \dots, 10$,

(4), (1·5), (1), (3), (3·75), (2·75), (3·5), (2·5), (3·25), (2).

In Examples 9, 10 and 11 the experiments were carried out according to the randomized block design, and the model was

$$x_{ji} = \mu + \beta_h \, [z_{1i} - E(z)] + \gamma_h \, [z_{hi} - E(z)],$$

$i = 1, 2, \dots, 10 = n$, $h = 2, 3, \dots, 11 = k + 1$, $j = h - 1$. The values of β_h were, for $j = 1, 2, \dots, 10$,

(4), (1·5), (1), (3), (3·8), (2·8), (3·5), (2·5), (3·2), (2)

and $\gamma_h = \frac{1}{2} \beta_h$.

In all six examples the estimated contrasts were

$$C_1 = a_1 - a_2, \quad C_2 = a_2 - a_3, \dots, C_9 = a_9 - a_{10} \, .$$

For each of the six examples, $N = 300$ experiments were sampled.

The percentage number of experiments for which the contrast ($= 0$) is covered by the confidence interval is shown in Table 7.4.

Table 7.4 **Percentage number of confidence intervals which cover the contrast. $\alpha = 0{\cdot}05$**

j	Example					
	6	7	8	9	10	11
1	93	95	92	94	94	94
2	94	95	94	96	93	94
3	92	96	93	93	95	94
4	96	94	93	97	96	96
5	95	94	94	96	96	96
6	96	93	93	92	96	97
7	94	95	94	94	96	94
8	93	94	94	95	93	94
9	94	93	93	96	93	95

The confidence limits were computed by (3.4) and (4.3) and $\alpha = 0{\cdot}05$. It will be seen that for all contrasts and examples the percentage number is very close to 95. Since in these examples the departure of the distributions from the normal is considerable, and the variances are changed to a large extent among the treatments, these results are further verifications of the robustness of "Student's" t distribution.

Among the $k - 1 = 9$ estimators in Fxamples 6, 7 and 8, independent sets can be selected. For instance, there are two sets of four estimators. The results for these are given in Table 7.5 under the notation: Examples 6, 7 and 8 and $k = 5$.

Let n_0 stand for the number of samples, or experiments, for which the contrast is covered by the confidence interval. In Table 7.5 are shown for Examples 4 to 11 the values of n_0/N, $(1 - \alpha)^{k-1} = 0{\cdot}95^{k-1}$, $(n_0/N)^{1/(k-1)}$ and $1 - \bar{r}/(k-1)$. It will be seen that the values of both $(n_c/N)^{1/(k-1)}$ and $1 - \bar{r}/(k-1)$ are very close to $1 - \alpha = 0{\cdot}95$ for all examples.

Comment on these results is unnecessary. It might be well to remember, however, that in practice there hardly exists a case

Table 7.5 $\alpha = 0{\cdot}05$

Example No.	Design	k	n	N	n_0/N	$0{\cdot}95^{k-1}$	$(n_0/N)^{1/(k-1)}$	$1-\bar{r}/(k-1)$
4	Complete randomization	3	5	100	0·920	0·903	0·959	0·955
5	"	5	5	40	0·875	0·815	0·967	0·963
6	"	10	10	300	0·617	0·630	0·948	0·942
7	"	10	10	300	0·600	0·630	0·945	0·943
8	"	10	10	300	0·573	0·630	0·940	0·937
6	"	5	10	600	0·815	0·815	0·950	0·944
7	"	5	10	600	0·803	0·815	0·946	0·944
8	"	5	10	600	0·783	0·815	0·941	0·935
9	Randomized blocks	10	10	300	0·680	0·630	0·958	0·947
10	"	10	10	300	0·663	0·630	0·955	0·947
11	"	10	10	300	0·687	0·630	0·959	0·949

which perfectly satisfies the assumptions underlying the use of "Student's" t for the computation of confidence limits of a parameter. Therefore, if the research worker computes the confidence limits of a contrast, using the tabulated value of t that corresponds to, e.g. $\alpha = 0{\cdot}05$, he should remember that the confidence probability of the resulting interval is hardly ever exactly equal to $1 - \alpha = 0{\cdot}95$. It is necessary for him to know, however, that the confidence probability is close to the chosen $1 - \alpha$.

The results obtained in our investigation indicate strongly that if the confidence limits of the contrast are computed by (3.4), or (4.3), the confidence probability of each contrast is simultaneously approximately equal to $1 - \alpha$. Non-normality, unequal variances, correlations between the estimators of the contrasts, and correlations between the estimators of the variances do not materially affect the confidence probability. Of course, a pertinent question is whether or not the examples included here cover so much ground that a general conclusion is justified. This is a question that may be raised in all situations of this kind. A general answer can hardly be given. However, the larger the number of examples is, the more confidence can be placed in the results. We have tried to cover as much ground as possible, but it is evident that results from new investigations are welcome.

8 The analysis of variance and the F-test

Consider an experiment according to the principle of complete randomization with k treatments (T_j , $j = 1, 2, \dots, k$) and n experimental units for each treatment. The general model for the observed random variable is given by (3.1). For this case R.A. Fisher [12] introduced the two mean squares

$$V_T = \frac{n}{k - 1} \sum (\bar{x}_j - \bar{x})^2$$

$$V_R = \frac{1}{k(n - 1)} \sum \sum (x_{ji} - \bar{x}_j)^2$$

and the statistic $z = \frac{1}{2} \log F$, where $F = V_T / V_R$.

It can be shown that if e_{ji} are stochastically independent values of a random variable e, the expectations of the two mean squares are

$$E(V_T) = \frac{1}{k} \sum \mathrm{var}_j(e) + \frac{n}{k - 1} \sum a_j^2$$

and

$$E(V_R) = \frac{1}{k} \sum \mathrm{var}_j(e)$$

where $\mathrm{var}_j(e)$ is the variance of e for treatment T_j . Therefore, $V_T \gg V_R$ indicates that $\Sigma a_j^2 > 0$, i.e. that the effect is not the same for all treatments. However, the question is how large $F = V_T/V_R$ must be in order to be taken as meaning, on some chosen level of significance, that $\Sigma a_j^2 > 0$. The answer to this question, given by Fisher, was his deduction of the distribution of F (or z), and the premise was

$$x_{ji} = \mu + e_{ji},$$

where e_{ji} is assumed to be $N = nk$ stochastically independent values of a normally distributed random variable. Regarding this as the null hypothesis (H_0), it can be tested by means of the tabulated significance points of F.

In practice it is usually taken for granted that rejection of H_0 implies that $\Sigma a_j^2 > 0$, but obviously this is not the only alternative. The null hypothesis also covers the statements that e is normally distributed and that var (e) is the same for all treatments. Therefore, statisticians have been concerned with the effect on the distribution of F of changing these two parts of the null hypothesis. The results of the different investigations are that the test seems to be too sensitive. This is chiefly due to differences in var (e) among the treatments and to lack of balance in the design, rather than to the form of the distribution of e. We confine ourselves to referring to Horsnell [21], the summaries given by Cochran and Cox [6] and Scheffé [34], and to the literature cited in these publications.

If proper randomization has been used, differences in var (e) among the treatments are due to interactions between the treatments and the heterogeneity factors. The research worker can hardly know to what extent the distribution of F is affected by such interactions in the actual case under consideration. He has to place reliance on the results of the different investigations, which indicate that the effect is not important. This is substantiated by the results of some new investigations to which we shall return. Even so, the F-test should be used with some reserve in the present case also.

Turning next to the randomized block experiment, we shall assume that there are k treatments $(T_j\ , j = 1, 2, \dots , k)$, n blocks or replications $(i = 1, 2, \dots , n)$ and m experimental units $(h = 1, 2, \dots , m)$ for each treatment and block. In this case the analysis of variance results in the following relevant mean squares:

$$V_T = \frac{nm}{k-1} \sum (\bar{x}_j - \bar{x})^2$$

$$V_{TR} = \frac{m}{(k-1)(n-1)} \sum \sum (\bar{x}_{ji} - \bar{x}_j - \bar{x}_i + \bar{x})^2$$

$$V_R = \frac{1}{nk(m-1)} \sum \sum \sum (x_{jih} - \bar{x}_{ji})^2 .$$

The assumption underlying the F-test for this case is the simplified model (cf. model (4.1)):

$$x_{jih} = \mu + z_i + e_{jih}$$

in which the e_{jih} are assumed to be $N = nkm$ stochastically independent values of a normally distributed random variable. It is also assumed that var (e) is the same for all treatments. Then it can be shown that $F_T = V_T/V_{TR}$ and $F_{TR} = V_{TR}/V_R$ are both distributed in the standard F form.

As a side issue it must be pointed out that the two mean square ratios are not independent. Therefore the two ratios should not be used simultaneously until the effect of the correlation between them is ascertained. In practice, however, m is usually chosen equal to unity, so that the problem of the effect of the correlation is not important.

From the general model (4.1) the expectations of the three mean squares can be developed easily. Letting $m = 1$, and writing, for short, var (u) and var (e) for the means of the k values of $\text{var}_j\,(u)$ and $\text{var}_j\,(e)$, the formulae are

$$E(V_T) = \text{var}(e) + \text{var}(u) - \frac{2}{k(k-1)} \sum\sum \text{covar}(u_p,\, u_q) + \frac{n}{k-1} \sum a_j^2$$

where $p \neq q$, and

$$E(V_{TR}) = E(V_T) - \frac{n}{k-1} \sum a_j^2 \,.$$

Therefore, in this case also $V_T \gg V_{TR}$ indicates that $\Sigma\, a_j^2 > 0$, i.e. that the effect is not the same for all treatments. However, if by the null hypothesis is meant $a_j = 0$ (or $\Sigma\, a_j^2 = 0$), the model to be tested is

$$x_{ji} = \mu + z_i + u_{ji} + e_{ji}.$$

Therefore, even if the e_{ji} are independent values of a normally distributed random variable, the F-test is merely an approximation. That this is so has been recognized by several statisticians, but at present there is insufficient information on the degree of approximation.

Lacking the necessary facilities, we have not been able to offer a large number of examples showing the effect of the interactions. The results obtained by using the examples from Section 7 might, however, throw some light upon the reliability of the test.

In Example 3, $k = 6$ and $n = 5$, so that the numbers of degrees of freedom are $k - 1 = 5$ and $(k - 1)(n - 1) = 20$ for V_T and V_{TR}. According to the standard F-distribution we should, therefore, expect in 100 experiments to find five F-values less than $1/4{\cdot}56 = 0{\cdot}22$ and the same number larger than $2{\cdot}71$. The numbers of F-values actually found in the different classes are

$F \leqslant 0{\cdot}22$	16
$0{\cdot}22 \leqslant F \leqslant 2{\cdot}71$	68
$F \geqslant 2{\cdot}71$	16
	100

This result indicates that the null hypothesis $a_j = 0$ might be falsely rejected about three times as often as is prescribed by the theory underlying the tabulated points of significance. Since interaction between treatments and replications would be expected to affect the F-distribution in this direction, the trend shown by the result is not surprising. The interaction effects are not exaggerated to such an extent that the experimental situation totally lacks realism. The unrealistic part of these experiments is that intra-block interactions are not included. It is likely that the effect of the latter interactions is to bring the distribution of F into better agreement with the standard distribution of the normal theory. The results from some small experiments carried out with normal deviates seem to substantiate this belief.

We have also used Examples 6 and 9, described in Section 7. In both cases the observations were drawn from a symmetrical Beta distribution. In Example 6 the experiments were constructed according to the principle of complete randomization with $n = 10$ replications and $k = 10$ treatments. In Example 9 we used the randomized block design with $n = 10$ replications and $k = 10$ treatments. The results are shown in Table 8.1, where N is the number of experiments and r the frequency of $F \geqslant F_{\alpha}$. It will be seen that in Example 6 the relative frequencies (r/N) are

Table 8.1

Example No.	α	N	r	r/N
6	0·1	375	43	0·115
	0·05	375	20	0·053
	0·01	375	4	0·011
9	0·1	300	49	0·163
	0·05	300	31	0·103
	0·01	300	11	0·037

approximately equal to those expected according to the standard F-distribution. However, in Example 9 the frequencies are larger, e.g. the estimated probability of F exceeding the 5 per cent level of significance is equal to 0·1.

These results are consistent with the results found with normal deviates. Together, the results show that if the randomized block design is used, the effect of interactions between the treatments and the inter-block heterogeneity factors is an inflation of the sensitivity of the F-test. This does not seem to be so for experiments carried out according to the principle of complete randomization.

In cases in which the experiment has been carried out according to the randomized block design, the F-test as a test of the null hypothesis $a_j = 0$ should be regarded with considerable reserve. But, of course, the research worker can use the F-test if he chooses a lower level of significance than that which he would have used if he had regarded the test as being fully reliable, e.g. the 2·5 per cent instead of the 5 per cent level of significance.

To be in doubt with regard to the reliability of the F-test does not, however, imply any lack of confidence in the usefulness of the analysis of variance. A research worker may very well be interested in the results of such an analysis, even if he does not resort to the F-test.

9 The F-test in cases in which a number of mean square ratios are computed by the same residual mean square

In some cases the research worker wishes to test a number of null hypotheses by means of the F-test and is compelled to use the same residual (or error) mean square for all F-ratios. It is evident that in such cases the mean square ratios are not stochastically independent. This implies that the ratios cannot be gauged against the tabulated points of significance of the standard F-distribution.

Suppose that $v_1 V_1/\sigma^2$, $v_2 V_2/\sigma^2$ and $v_0 V_0/\sigma^2$ are stochastically independent χ^2 with v_1, v_2 and v_0 degrees of freedom, and let $F_1 = V_1/V_0$ and $F_2 = V_2/V_0$. Then, it can be shown that the regression of F_2 on F_1 is linear and that the coefficient of correlation for $v_0 > 1$ is

$$\rho = \sqrt{\left\{\frac{v_1 v_2}{(v_0 + v_1 - 2)(v_0 + v_2 - 2)}\right\}}.$$

It will be seen that if v_0 is large compared with v_1 and v_2, the correlation is trivial and cannot invalidate the F-test.

However, the effect of the correlation is better measured by means of the conditional probability $P(F_2 \geqslant F_\alpha \mid F_1 \geqslant F_\alpha)$, where the F_α are the tabulated significance points corresponding to the respective numbers of degrees of freedom: v_2 and v_0 for F_2, and v_1 and v_0 for F_1. Under the stated assumption, this probability can be computed. In Table 9.1 the values of P are shown for $\alpha = 0{\cdot}05$, $v_1 = v_2 = 1$ and some values of v_0.

Table 9.1

v_0	P
2	0·380
4	0·217
10	0·111
60	0·059
200	0·053

It will be seen that if $v_0 > 60$, the correlation between the two ratios does not matter, but this is not so for smaller values of v_0. Furthermore, the effect of the correlation can be shown to be greater for larger values of v_1 and v_2.

The usual way of dealing with this problem seems to be to ignore it. This attitude is surprising, since simultaneous tests of null hypotheses in such circumstances occur regularly in both experimental and non-experimental research work; it is also unnecessary, because several solutions of the problem have already been put forward.

One of the solutions has been sought in the development and tabulation of the distribution of the largest ratio. Investigations along this line, by Hartley [19], Finney [11] and Nair [27], have resulted in the generalization due to Hartley [20].* If there are m null hypotheses, Hartley has suggested the use of $F_{\alpha/m}(v_i, v_0)$, $i = 1, 2, \ldots, m$, as approximative significance points. In our view the use of this technique implies that we take a too critical attitude, and it might in some cases result in inacceptable inferences; cf. the next section.

A different solution was suggested by the present author [30]. For the simultaneous testing of m null hypotheses (H_{0i}, $i = 1, 2, \ldots, m$) it was suggested that all H_{0i} should be rejected only if all $F_i \geqslant c_i$, where

$$(9.1) \qquad c_i = F_\alpha(v_i, v_0/m)\left[1 + \frac{(v_i - 2)(m - 1)}{v_0^2}\right].$$

It was shown that the probability (under the null hypotheses) of all F_i exceeding c_i is approximately equal to α^m. In the case in which the F-ratios are stochastically independent, this is the probability of all F_i exceeding $F_\alpha(v_i, v_0)$ simultaneously. Therefore, since the same technique is used for all F-ratios, rejection of H_{0i} if $F_i \geqslant c_i$ means rejection of any one null hypothesis on the α level of significance.

Most often some of the F-ratios are smaller than c_i. For such cases it was suggested that we should proceed sequentially:

Step 1: Remove the F-ratio with the smallest value of F/c. Then compute the c_i with m replaced by $m - 1$. If then, all the

* See also M.G. Kendall and A. Stuart [23], Vol. 3, pp. 40–43.

F_i (number $m - 1$) are larger than the new c_i, the corresponding H_{0i} are rejected. If at least one $F_i < c_i$, proceed to the second step.

Step 2: Remove the F-ratio (among the remaining $m - 1$ ratios) having the smallest value of F/c. Then compute the new c_i (number $m - 2$) with m replaced by $m - 2$, and proceed as under step 1.

It is evident that if at least one of the F-ratios is larger than the corresponding $F_\alpha(v_i, v_0)$, this step-wise procedure will eventually cease with at least one F-ratio judged significant.

It might be necessary to emphasize that to remove an F-ratio does not imply that the corresponding null hypothesis is accepted. It merely means that it is placed among those null hypotheses that are not rejected on the chosen level of significance by the experimental facts. This distinction is obviously very important.

Of course, it is not essential to begin with all the F-ratios. At the outset, only those ratios that are $\geqslant F_\alpha(v_i, v_0)$ should be included. Those ratios which are $< F_\alpha(v_i, v_0)$ can be judged not significant at once and removed.

The assumptions underlying this method of testing are (1) that $v_i V_i/\sigma^2$ are independent χ^2, and (2) that σ^2 is a constant variance. None of these assumptions are realized in actual experiments. Therefore, c_i by (9.1) is merely an approximation. When using this method, the research worker should not forget that the test is usually too sensitive.

Most often when using this technique of testing null hypotheses, the research worker is faced with the problem of how to interpolate in the F-table. If v_1 is the number of degrees of freedom for the numerator of the mean square ratio, and v_2 that of the denominator, such interpolation is rather difficult for small values of v_2. For this reason we have computed the points of significance for $v_1 = 1, 2, 3$, and 4 and $v_2 = 1{\cdot}1$ to $4{\cdot}9$. The values are given in the tables presented in the Appendix.

It will now be assumed that the treatments are quantities x_{1j} $(j = 1, 2, \dots, k)$, and that the purpose of the experiment is to produce data from which a response function can be estimated. The first question that presents itself is: response to what? Since an exact repetition of a treatment is never possible, this is an appropriate question. In this case the treatments (x_{1j}) are modified controlled observations as defined by Berkson [4]*, so we can write $x_{1j} = h_j + v_j$ and assume that the v_j are random errors for which $E(v_j) = 0$. There are, therefore, two response functions. If $\bar{x}_{0j}$ are the means of the observed response variable, the two functions are

$$E(\bar{x}_{0j}) = f(x_{1j}) \quad \text{and} \quad E(\bar{x}_{0j}) = g(h_j)$$

No satisfactory method of estimating the latter function seems to have been found, except, perhaps, when the function is linear. We therefore consider the first function only. This is the response function the research worker will be interested in if his purpose is to use it as a guide in some practical activity.

Since the formula of $f(x_1)$ is hardly ever known, the worker has to assume that it can be approximated by the Taylor expansion, and the practical problem is the common one of estimating the coefficient $(\beta_0, \beta_1, \beta_2, \dots)$ in the equation

$$(10.1) \qquad \bar{x}_{0j} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{1j}^2 + \dots + e_j .$$

It can be assumed that $E(e) = 0$ for each j and that e is stochastically independent of $x_1, x_1^2, \dots$ But we cannot assume that $\text{var}(e)$ is the same for all j.

It has been shown that curvature of the regression function can in some cases be removed by means of suitable transforms of the independent variable, such as, e.g., the logarithm and the square root. Most often, however, such transforms as have been invented do not remove the curvature completely, and the

* See also Kendall and Stuart [23], Vol. 2, pp 408–409.

research worker is therefore left with model (10.1), even if a transform has been used.

It is well known that if e is stochastically independent of x_1, the method of least squares yields unbiassed estimators of the regression coefficients (β_r). The difficulty is that the research worker must decide in advance which terms $\beta_r x_{1j}^r$ ought to be included initially. The advent of electronic computers has simplified matters, and it is now possible to include a considerable number of terms at not too great a cost. Of course, in the function as finally estimated, the maximum number of terms is k (the constant term included), and the terms need not necessarily be a sub-set of the set $r = 1, 2, \ldots, (k - 1)$. However, in practice the research worker has to compromise in order to avoid being involved in very heavy and expensive computations. Unless his experience from earlier investigations indicates that different terms ought to be used, it is, perhaps, sound practice to include initially the terms for $r = 1, 2, 3$ and 4. Working on these lines, he will be able to decide whether all these terms should be included in the final estimated response function, and whether it is advisable to bring in additional terms.

The advice to use initially the terms for $r = 1, 2, 3$ and 4 certainly lacks any logical justification; it is merely the author's inference from a rather limited field of experience. However, if the investigator, for one reason or another, wishes to start with a different set of terms, the technique is in principle exactly the same as with the choice of $r = 1, 2, 3$ and 4.

In order to simplify the formulae we shall introduce the deviations from the mean, $y_r = x_{1j}^r - \text{mean}\,(x_{1j}^r)$. Furthermore, we shall use orthogonal functions of these deviations. There are a number of sets of such functions. One of the sets is

$$\begin{aligned}
u_1 &= y_2 \\
u_2 &= y_1 - b_{12} y_2 \\
u_3 &= y_4 - b_{42.1}\, y_2 - b_{41.2}\, y_1 \\
u_4 &= y_3 - b_{34.12}\, y_4 - b_{32.14}\, y_2 - b_{31.24}\, y_1 ,
\end{aligned}$$

where the coefficients b are least-squares regression coefficients. However, this is only one of the possible such sets of functions, the total number of sets being $4! = 24$.

Suppose now that this particular set of such functions (u) has been chosen. Then it is possible to show that the reductions (mean squares) due to the different u's are as presented in Table 10.1, where the R's are correlation coefficients (simple or multiple), and $T = n\,\Sigma(\bar{x}_{0j} - \bar{x}_0)^2$.

Table 10.1

	Reduction	Degrees of freedom	Mean square
u_1	$R^2_{0.2}\,T$	1	V_1
u_2	$[R^2_{0.12} - R^2_{0.2}]\,T$	1	V_2
u_3	$[R^2_{0.124} - R^2_{0.12}]\,T$	1	V_3
u_4	$[R^2_{0.1234} - R^2_{0.124}]\,T$	1	V_4
Residual	$[1 - R^2_{0.1234}]\,T$	$k - 5$	V_5
Total	T	$k - 1$	

In model (10.1) x_1, $x_1^2, \ldots$ can be transformed to $u_1, u_2, \ldots$, and the result is

$$\bar{x}_{0j} = \lambda_0 + \lambda_1 u_{1j} + \lambda_2 u_{2j} + \ldots + e_j . \tag{10.2}$$

Assuming (1) that e is a random variable, independent of $u_1, u_2, \ldots$, (2) that $\text{var}(e)$ is the same for all j, and (3) that e is normally distributed, it can be shown that $V_r/\text{var}(e)$ ($r = 1, 2, 3, \ldots$) is a χ^2 if $\lambda_r = 0$. Therefore, if the assumptions are fulfilled, the null hypotheses $\lambda_r = 0$ can be tested by the mean square ratios $F_1 = V_1/V_R$, $F_2 = V_2/V_R, \ldots$, where V_R is the residual (or error) mean square in the analysis of variance. If the experiment has been carried out according to the principle of complete randomization,

$$V_R = \frac{1}{k(n-1)}\sum\sum(x_{0ji} - \bar{x}_{0j})^2$$

and if randomized blocks have been used,

$$V_R = V_{TR} = \frac{1}{(k-1)(n-1)}\sum\sum(x_{0ji} - \bar{x}_{0i} - \bar{x}_{0j} + \bar{x}_0)^2 .$$

In both cases n is the number of replications.

It will be seen that this is a case in which a number of null hypotheses are being tested simultaneously by mean square ratios, which are correlated because a common V_R is used. The problem has been treated in Section 9, to which the reader is referred.

In dealing with the problem of choosing a test method, it was stated that the use of the largest ratio might result in inacceptable inferences. Suppose now, for the sake of argument, that F_1 is the largest ratio and that it is greater than $F_\alpha(1, v_R)$, v_R being the number of degrees of freedom of V_R. Suppose, furthermore, that $F_1 < F_{\alpha/m}(1, v_R)$, where $m = 5$ in the present case. Then, if the technique based on the largest ratio is used, and if all ratios are declared not significant if the largest is less than $F_{\alpha/m}(1, v_R)$, none of the reductions in Table 10.1 should be regarded as significant. But such an inference is hardly acceptable because, since $F_1 > F_\alpha(1, v_R)$, the research worker would reasonably regard the reduction due to u_1 as being significant and include u_1 in the regression function.

Suppose now that the method described in Section 9 is used, and that it is found that V_5/V_R is significant on the chosen level of significance. This is a result that should be taken as indicating that probably at least one of the terms $\lambda_r u_r$ $(r = 5, 6, \dots, k - 1)$ ought to be included in the response function. However, it does not imply that we shall succeed if we try to do so. In the author's experience, such an outcome of the testing will happen very rarely. The reason is, of course, that response functions are usually not so complicated that a linear function of $x_1, \dots, x_1^4$ does not, when estimated, give a sufficiently accurate description of them.

The residual reduction $[1 - R^2_{0.1234}]\,T = (k - 5)\,V_5$, giving rise to the mean square V_5, can be divided into $(k - 5)$ reductions which are due to the variables $u_5, u_6, \dots, u_{k-1}$. If such a division is carried out, it may be found that the greater part of the reduction $(k - 5)\,V_5$ is due to only one of these variables. However, such a result must be accepted as a possibility in advance, the consequence being that $(k - 5)\,V_5$ should be used with one degree of freedom, leading to the mean square ratio $(k - 5)\,V_5/V_R$ and a more efficient test. It is evident, however, that in this case also a significant result merely indicates that a more satisfactory description of the response function might be obtained.

In our description of the statistical procedure we have assumed that the set $u_1, \ldots, u_4$ has been chosen. But we have pointed out that there are $4! = 24$ such sets. Furthermore, in order to compute the reductions in Table 10.1, the sample values of the coefficients of correlation must be known. Now, among 4 variables, taking account of their order, there are 4 sets consisting of one variable, 12 sets consisting of two variables, 24 sets consisting of three variables and 24 sets consisting of four variables. This makes a total of 64 sets. It is reasonable to suppose that for any electronic computer a program can be worked out for the selection of all these sets and at the same time for the computation of the corresponding coefficients of correlation. It is then a simple matter, at each step, to select among the variables (u) that are not included, the one that yields the largest reduction.

Testing null hypotheses concerning the reductions due to the different u-variables, the four F-ratios must be gauged against $F_\alpha(1, \nu_R/5)$ (cf. Section 9). Then, if the set of u-variables is chosen in advance, the null hypotheses will be rejected simultaneously on a level of significance that is approximately equal to α. It is not obvious, however, that this is so if the set of variables is chosen in the way described above. In order to estimate the effect on the level of significance of the selection of the u-variables, we have carried out experiments according to the following plan: x_0, x_1 and e are independent standardized normal variables, and

$$x_2 = \beta x_1 + e .$$

The chosen values of β were $\beta = \frac{1}{3}$ in Example 1 and $\beta = \frac{3}{4}$ in Example 2. Thus the coefficient of correlation is $\rho_{12} = 0{\cdot}32$ in Example 1 and $\rho_{12} = 0{\cdot}6$ in Example 2.

Suppose now that the area covering both F-ratios $\geqslant 0$ is divided into three parts : A being the part for which both F-ratios are $\leqslant F_\alpha(1, \nu_R)$, C being the part for which both F-ratios are $\geqslant F_\alpha(1, \nu_R/2)$, and B being the rest of the area. Then, if the two u-variables are chosen in advance, the approximate probabilities of the two F-ratios falling inside A, B and C are given respectively by the binomial terms $(1 - \alpha)^2$, $2\alpha(1 - \alpha)$ and α^2.*

* Cf. ref. [30].

Choosing α = 0·05, the values of these terms (P) are as shown in Table 10.2.

Table 10.2 $\alpha = 0{\cdot}05$

Area	P	Example 1		Example 2	
		n	NP	n	NP
A	0·9025	948	945·82	926	925·06
B	0·0950	97	99·56	98	97·38
C	0·0025	3	2·62	1	2·56
Total	1·0000	1048	1048·00	1025	1025·00

The number of experiments that were constructed and analysed was N = 1048 in Example 1 and N = 1025 in Example 2. In Table 10.2, n stands for the number of experiments for which the two F-ratios were found in the different areas (A, B and C). If it is assumed that the particular way in which the u-variables are selected does not affect the level of significance, NP will be the expected number of experiments. It will be seen that for both examples the values of n are consistent with those expected (NP).

In these examples the variables x_1 and x_2, and hence the u-variables, are random variables, while in the experimental case they are values chosen by the research worker. Our investigation was planned in this way because we were concerned with the problem as it is presented in multiple regression. It seems evident, however, that the results can be applied in the situation with which we are dealing in the present section.

Having chosen the set of u-variables and having decided which of the variables ought to be included in the estimated response function, it will be necessary to estimate the regression coefficients in model (10.2). It is then also necessary to compute the regression coefficients for the different regressions among the variables $x_1, \ldots , x_1^4$ which are included in the formulae for the u-variables. These computations are carried out by standard techniques described in a number of textbooks and we do not, therefore, go into the matter here. The last step consists of transforming the u-variables in the response function to x-variables.

The technique described in the foregoing can hardly be recommended to research workers who lack facilities to use an electronic computer; for such workers this technique would be too time-consuming. An alternative approach would be to choose initially a particular set of u-variables, for instance:

$$\begin{aligned} u_1 &= y_1 \\ u_2 &= y_2 - b_{21} y_1 \\ u_3 &= y_3 - b_{32.1} y_2 - b_{31.2} y_1 \\ &\cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \end{aligned}$$

With such a set, the u-variables can be included one at a time, and for each new u-variable the residual reduction in Table 10.1 can be computed. In this way the investigator can decide at each step whether it seems worth while to continue adding new terms. Thus the work can be reduced to a minimum, so that computations can easily be carried out by means of a desk calculator. Using this technique, the research worker will lack the opportunity of trying different combinations of the y-variables, and the final estimated response function cannot be claimed to be the "best" in the sense that it includes the minimum number of terms; even so, the estimated function may be quite acceptable as a description of the response function – it might even turn out to be the "best" one.

The assumptions underlying the F-tests used above are (1) that e in the models is a normally distributed random variable, and (2) that var(e) is the same for all treatments, i.e. the same for all chosen values of x_1. Neither of these assumptions can be regarded as being realistic. We have discussed this point in Section 8 and shall not repeat the arguments here. We shall confine ourselves to pointing out that the research worker ought to remember that the level of significance is not the one he has chosen, e.g. $\alpha = 0{\cdot}05$, but usually an inflated one.

In a case in which the treatments are quantities, the worker may need to estimate the response function and, at the same time, he may wish to estimate particular contrasts. A contrast can, of course, be estimated by means of the estimated response function. But in our opinion, the methods described in Section 7 are better suited for this purpose.

The research worker may also require to estimate particular

x_1-values as, for instance, the value for which the response is a maximum, or the value for which the increase of response is a maximum. So far as we can see, unbiassed estimates of these values of x_1 can never be obtained, but, even so, useful approximations can be found. It is evident, however, that in order to estimate such values of x_1 it is necessary that the range of the selected values should cover them. This means that the worker must be in possession of advance information with regard to these values and must use such information in the planning of the experiment.

An important question is that concerning the choice of values of x_1. For the computations it will be advantageous to choose equally spaced values, or equally spaced values of transforms such as the logarithm and the square root. This will enable the research worker to use the orthogonal polynomials introduced by Fisher [14].* These polynomials are proportional to our u-variables, and the use of the polynomials will therefore effectively simplify the computations. This is important, especially if the computations have to be carried out by desk calculator. It must be remembered, however, that the use of these polynomials entails the use of a particular set of u-variables, implying that the worker must abandon the idea of trying out different sets of these variables.

Except for the method of testing, the procedure described above is, in the main, identical with the modern stepwise regression programs now often used. As to these programs, we confine ourselves to referring to Draper and Smith [8], Kendall and Stuart [23], Vol. 2, and to Beale, Kendall and Mann [3].

* Tables of these polynomials will be found in Fisher and Yates [16] and in Pearson and Hartley [31].

11 The problem of gaps and the grouping of treatments

In planning an experiment it is not always possible for the research worker to decide on the particular contrasts that he wishes to estimate. In such cases a commonly used and acceptable procedure is to range the treatments according to the value of the treatment means ($\bar{x}_j$) and to compute the differences (or gaps) between two neighbouring means. Let r be the rank ($r = 1, 2, \dots, k$, where k is the number of treatments). Then $u_r = \bar{x}_{r+1} - \bar{x}_r$ ($r = 1, 2, \dots, k-1$) are the gaps.*

It is evident that the expectation of a gap is positive, i.e. that $E(u_r) > 0$, and that it is usually dependent on r. If the distribution of $\bar{x}_j$ is rectangular, i.e. if $f(\bar{x}_j) = 1/A$ $(0 \leqslant \bar{x}_j \leqslant A)$, it can be shown that the distribution of u_r is

$$f(u_r) = kA^{-k}(A - u_r)^{k-1}$$

and that $E(u_r) = A/k+\dot{1}$, i.e. that it is the same for all gaps. In other cases, e.g. the normal, the distribution of u_r depends on r and k, and the expectation is a function of r and k. Since the research worker cannot know this function, he is unable to utilize the differences $u_r - E(u_r)$. However, in practice the gaps might be used, even if $E(u_r)$ remains unknown.

Suppose that an analysis of variance has been carried out and that $F = V_T/V_R$ (cf. Section 8) is significant at some chosen level of significance, e.g. the 5 per cent level. Then a strongly marked gap in the series u_r can reasonably be taken to be an indication of a grouping of the treatments. There might also be indications of more than two groups.

In most cases, however, a more detailed analysis is needed. The mean range can then be used to advantage. The mean range $E(W_k)$ of the normal distribution has been tabulated by Pearson and Hartley [31] for sample sizes ranging from 2 to 1000. Using this mean range, the conditional expected range of the means ($\bar{x}_j$), i.e. conditioned by V_R regarded as a non-random quantity, is $V_k = E(W_k)\sqrt{(V_R/n)}$. In this formula, n is the number of

* Cf. Tukey [35]

replications, e.g. the number of blocks in a randomized block experiment. Then if $\bar{x}_{\min}$ and $\bar{x}_{\max}$ are the smallest and the largest treatment means, we can use $(\bar{x}_{\min} + V_k)$ and $(\bar{x}_{\max} - V_k)$ as borders between groups of treatments. The treatments, the means of which are included between $\bar{x}_{\min}$ and $(\bar{x}_{\min} + V_k)$, are regarded as one group. In the same way the treatments, the means of which are included between $(\bar{x}_{\max} - V_k)$ and $\bar{x}_{\max}$, are regarded as another group.

There are three possible outcomes of this preliminary grouping of the treatments:

(a) $(\bar{x}_{\min} + V_k) < (\bar{x}_{\max} - V_k)$, and no mean is found between the two borders.

(b) A number, at least one, of treatment means is found in the interval between the two borders.

(c) The two intervals are overlapping.

When trying to divide the treatments into groups, it is necessary to know the purpose for which the grouping is required. In basic research work it may be important to find the borders between all groups, and then it may be necessary to proceed with the two or three intervals found on the first step of the analysis. Most often, however, the purpose is to select among the k treatments those which, in a certain sense, are superior. In such cases the worker need not bother with more than one of the preliminary groups. Suppose that a treatment having a large value of $E(\bar{x}_j)$ is regarded as being superior, and that the interval bordered by $(\bar{x}_{\max} - V_k)$ and $\bar{x}_{\max}$ covers m treatments. Then, at the second step an analysis of variance should be carried out for this group alone. If the randomized block design has been used, this means the computation of a new treatment mean square V_T and a new residual mean square V_R, the numbers of degrees of freedom being now $(m - 1)$ and $(m - 1)(n - 1)$. Then if the new ratio $F = V_T / V_R$ is found to be non-significant, the worker has to be content with the group found at the first step. If, however, the mean square ratio is significant at some chosen level of significance, a new group border can be computed by $\bar{x}_{\max} - V_m$, where $V_m = E(W_m)\sqrt{(V_R/n)}$. This process can, of course, be repeated at a third step, fourth step, etc., and will be terminated as soon as a non-significant mean square ratio is found.

It is perhaps necessary to point out that the use of the mean range in this way must not be regarded as a test of significance. To use the mean range for computing the conditional range merely implies recognising the consequence of the null hypothesis having been rejected by the use of the mean square ratio.

A weak point in the suggested technique is that it is based upon the use of the mean range of the normal distribution. Because of the inflated sensitivity of the F-test (cf. Section 8) it will increase our confidence in the technique to know that this mean range is usually larger than it is in cases based on more realistic distributions. A summary of the present state of knowledge regarding the distribution of the range and the mean range has been given by Kendall and Stuart, to whose treatise [23] the reader is referred. New information could have been obtained from the examples used in Section 7 if lack of funds had not prevented utilization of the observations to this end. The treatment means obtained in Example 6 were, however, used for the purpose; in this example the observations were drawn from a symmetrical distribution. Since the standard deviation is known – for the treatment means it is equal to 0·5976 β_j – the mean range can be estimated in samples of experiments. The results are shown in Table 11.1, where k is the size of the sample and N the number of samples. For the sake of comparison the values of $E(W_k)$ for the normal distribution are included.

The mean range can also be computed economically for selected mathematically simple distributions. For the rectangular distribution it is equal to $\frac{(k-1)}{(k+1)}\sqrt{12}$ (cf. page 46), where k is the size of the sample. For the exponential $f(x) = e^{-x}$ $(x \geqslant 0)$ it can be shown that

$$E(W_k) = \sum_{1}^{k-1} \frac{1}{i}.$$

Some values of $E(W_k)$ for the exponential and the rectangular distribution are shown in Table 11.1.

It is obvious that such results do not justify the drawing of definite conclusions. But it will be seen that, for moderate values of k, the results appear to indicate that the mean range of the normal distribution is as large as that of the distributions which are being compared.

Table 11.1

k	Example 6		$E(W_k)$		
	N	W_k	Normal	$f(x) = e^{-x}$	Rectangular
5	150	2·353	2·326	2·083	2·309
10	148	3·081	3·078	2·829	2·834
20	72	3·570	3·735	3·548	3·134
375	10	5·833	5·896	6·503	3·446

Federer [10], p. 122, presents the results of a randomized block experiment for the comparison of $k = 7$ varieties in $n = 5$ blocks. We have chosen this example because prior information concerning the grouping of the varieties was available. The experiment was carried out with two units for each variety per block, and we have used the total for the two units.

In this case the mean square ratio $F = V_T/T_R = 3{\cdot}4$ which is significant, and the variety means are

Variety	$\bar{x}_j$
416	6·34
405	10·08
109	10·22
407	11·09
593	11·42
130	11·83
406	13·32

From the data given by Federer we have $(V_R/n)^{1/2} = 1{\cdot}179$, and since $E(W_7) = 2{\cdot}70436$ for the normal distribution, we find $V_7 = 3{\cdot}19$ and that

$$\bar{x}_{\min} + V_7 = 6{\cdot}34 + 3{\cdot}19 = 9{\cdot}53,$$

$$\bar{x}_{\max} - V_7 = 13{\cdot}32 - 3{\cdot}19 = 10{\cdot}13.$$

The seven varieties are therefore divided into three groups:

group A : variety 416
group B : variety 405
group C : the rest of the varieties.

Then, dealing with group C only, it is found that $V_T = 6{\cdot}49$ (with 4 degrees of freedom) and $V_R = 6{\cdot}23$ (with 16 degrees of freedom), implying that no division of the group should be attempted. Hence, the grouping at the first step is the final one.

Federer, using the prior information, concluded that there are significant differences (1) between group (130, 406, 593) versus group (405, 407, 416), and (2) among (405, 407, 416). It will be seen that these results are consistent with our findings. His second conclusion is consistent with our finding that 405, 407 and 416 belong to different groups. His first conclusion is consistent with our finding that 130, 406 and 593 belong to group C, while 405 belongs to group B and 416 to group A.

12 The statistical treatment of fractions

Most often the random variable that is the subject of a statistical analysis is directly observed, e.g. the yield in an agricultural field plot experiment; but it is not always so. For instance, in some cases the research worker observes the number (m) of units of a certain kind within each experimental unit, and at the same time he observes the number (x) of these units having a certain characteristic (A, say). In an agricultural field plot experiment, m may be the number of roots within plots and x the number of diseased roots. In such cases the research worker has to deal with $y = x/m$ or, in percentage, $100\,y$.

It must be assumed in general that both m and the probability $P(A) = p$ depend on the heterogeneity factors, and hence that they must be regarded as random variables. For the same reason it would be unrealistic to assume that they are independent. In this case, therefore, we are faced with a situation in which we have to deal with three correlated random variables (x, m, and p), but only two of them (x and m) can be observed.

The distribution of the three random variables is thus

$$(12.1) \qquad f(x,\, m,\, p) \;=\; \phi(m,\, p) \binom{m}{x} p^{x} \,(1 - p)^{m-x} ,$$

where $\phi(m, p)$ is the joint distribution of m and p, and the binomial part is the conditional distribution of x.

The formulae for $E(y)$ and $\mathrm{var}(y)$ can be derived easily; compare reference [29]. However, they do not show how the observations of $y = x/m$ should be treated statistically, and they will not be given here.

The model is evidently the same as it is for any other random variable. If the randomized block design has been used, the model is

$$(12.4) \qquad y_{ji} \;=\; \mu + a_j + z_i + u_{ji} + e_{ji}$$

($j = 1, 2, \ldots, k$, $i = 1, 2, \ldots, n$), where k is the number of treatments and n the number of replications. In this model, ($\mu + a_j$) can evidently be replaced by p_{0j}. Using this substitution, it will be seen that

$$\bar{y}_j = p_{0j} + \bar{z} + \bar{u}_j + \bar{e}_j . \tag{12.5}$$

Referring to Section 4 regarding the properties of z, u and e, it will also be found that $E(\bar{y}_j) = p_{0j}$ and that $E(\bar{y}_p - \bar{y}_q) = p_{0p} - p_{0q}$. A contrast in this case is, of course, a linear function of all, or a sub-set of, the p_{0j}. It will be seen that the same function of $\bar{y}_j$ is an unbiassed estimator.

The difficulties met with in the statistical analysis originate mainly in the interaction between the treatments and heterogeneity factors, causing differences in var(y) and correlations between contrast estimators among the treatments. Difficulties also spring partly from excessive skewness of the distribution of y. We cannot see, however, that the statistical treatment of the observations of x/m calls for methods different from those used for observations of other random variables. Our conclusion is, therefore, that the methods described in Section 7 to 11 are applicable in these cases also.

In the literature certain transformations are recommended; cf. Bartlett [2]. In the present case, examples are $\log y$, $\sqrt{y}$ and the inverse sine function, arc $\sin\sqrt{y}$.

It is evident that skewness of the distribution will be reduced by means of the logarithmic or the square root transformation, but the effect may be small and not worthwhile.

The purpose of some transformations, e.g. the inverse sine function, is to stabilize the variance. Assuming an additive model and that $p = P(A)$ is a constant, it can be shown that var(y) is approximately independent of p. But if the model is non-additive, the effect may be very small.

Some years ago the present author [29] recommended the use of covariance analysis, using $w = 1/m$ as the independent random variable in the regression function. No doubt the effect of the use of such regression is to reduce the differences in variance among the treatments. We have found, however, that the effect is not sufficient to counterbalance reduction of the generality of the population to which the conclusions are being applied. Problems arising as a consequence of the use of covariance technique will be dealt with in Section 15, to which the reader is referred.

13 Non-random experimental material

In the preceding sections, the replications have been regarded as a sample representing an abstract population. We now go on to show in more detail the difference between a non-random and a random sample of replications, and what this difference implies.

As an example we shall use an experiment carried out according to the randomized block design. Let the replications (blocks) be numbered $i = 1, 2, \ldots, n$, and the treatments numbered $j = 1, 2, \ldots, k$. Also, let the experimental units within any replication be numbered $r = 1, 2, \ldots, k$. Then the value (and the observation) of the random variable under consideration can be symbolized by $x_{(j)ri}$, which stands for the value of the random variable which would have been observed for the rth unit in replication i if the treatment T_j had been applied to this unit as a result of randomization. If the experimental material is regarded as non-random, we have to consider the number of possible allocations of the k treatments to the units. This number is $K = (k!)^n$.

The null hypothesis under consideration is now one stating that $x_{(j)ri}$ is the same for all j. This means that the result for unit r in replication i is the same, irrespective of the treatment actually placed on the unit by randomization. Even if this null hypothesis is true, some variation of the observations will occur, because of the heterogeneity of the experimental material. As we understand it, this is the null hypothesis considered by Fisher [13]. A modification is suggested by Neyman [28].

Suppose that $N = nk$ values of a variable are randomly arranged in all K ways. Then for any arrangement we have the usual two relevant mean squares: V_T with $k - 1$ degrees of freedom, and V_R with $(k - 1)(n - 1)$ degrees of freedom. Hence, for each arrangement there is a mean square ratio $F = V_T/V_R$. The distribution of these K values of F is thus the distribution of F in the population that consists of the K random arrangements. Since no two samples of experimental units are exactly alike, the distribution of F will change from one experiment to another. In any actual case the research worker knows the observations of the random variable under consideration for the actual arrangement. However, the distribution of F remains unknown to him.

For some cases in which Fisher's null hypothesis can be regarded as being satisfied, the distribution of F has been obtained by rearrangements of the values of the observed random variable. The distributions obtained in this way have been compared with the F-distribution derived under the normal theory. Usually a satisfactory compatibility has been found. We refer to papers by Eden and Yates [9], Welch [36], Pitman [32], Hack [18], Baker and Collier [1], and to other contributions cited in these papers. Some of these papers deal with experiments according to the principle of complete randomization.

The results obtained through such investigations are supposed to establish the necessary foundation for the use of the F-test of normal theory in the analysis of experimental data.

The model which is assumed to give a satisfactory description of the observed random variable (x) is

$$x_{ji} = \mu + b_i + a_j + e_{ji},$$

where μ, a_j and b_i are regarded as parameters, while e is regarded as a random variable. It will be seen that no interaction between the treatments and the inter-block heterogeneity factors is included. On the other hand, interaction between the treatments and the intra-block heterogeneity factors is sometimes recognized and included in the term e. Usually e is regarded as a random variable that belongs to the population consisting of the K arrangements. It is sometimes also regarded as a normal random variable.

This population of the K arrangements is certainly a simple construction, and the majority of statisticians – and possibly also of research workers – regard it as being fully adequate. For reasons explained in Section 1, we do not consider it as such. We also find that some writers who accept it seem to feel some uneasiness with regard to the interpretation of the experimental results associated with it. Some writers recommend generalizing to a broader population. For instance, Kempthorne [22], p. 152, writes: "We shall regard the inferences that we make as being inferences about the experimental units actually used, the extrapolation of these to a broader population being a matter of judgment in the present state of knowledge." This is important, and it shows the inadequacy of the commonly accepted idea of a population. We think nobody will insist that the population of the *K* arrangements is the one the research worker is actually interested in. We have therefore come to the conclusion that this population is inadequate and that if research methodology is founded upon this construction, the worker can make inferences about the treatment effects only by means of support from evidence obtained outside the experiment. This is certainly most unsatisfactory, and it seems to us that the only way of avoiding the difficulty is to regard the actual replications as a random sample of replications representing the population. This population is always an abstraction, and is the one the sample represents, when regarded as a random sample. This is the answer in other fields of empirical research work, and it also applies when we are dealing with experimental research.

It also appears that most statisticians accept the assumption that the effect of treatments and that of replications are additive; some writers even think that if the assumption is not satisfied, then randomized block designs cannot be used. However, we cannot accept this assumption because it would imply that the research worker has advance information about the effects of the treatments. To allow for interactions merely means that an unprejudiced point of view is taken; it does not mean assuming that interactions always exist. Our standpoint is that the research worker can never know in advance – and hardly by analysis of the experimental data – whether interactions exist; and, therefore, that he must treat his data as though interactions are present.

14 Factorial experiments and the split-plot design

For a number of reasons, investigations concerning the combined effect of two or more factors are important. When research is started in some new field, it is natural to begin with single-factor experiments and, by means of the data obtained from these, to learn something about the effects of several factors taken alone. But the effect of a factor may depend upon other factors, and therefore it will become necessary to carry out experiments with combinations.

Suppose that rs combinations of two factors P_p ($p = 1, 2, \dots, r$) and Q_q ($q = 1, 2, \dots, s$) are included in the experiment. Then the experiment can be regarded as concerned with the treatments T_j ($j = 1, 2, \dots, k = rs$), and the analysis can be carried out as if only one factor were involved. Of course, in these cases some particular contrasts are intended to be estimated by means of linear functions of the treatment means.

The simplest method of analysis consists in dividing the total treatment effect into a main effect (or sole effect) of each factor, and an interaction. The models that cover such divisions are obtained by the substitution for a_j in model (3.1) and model (4.1) of

$$a_j = b_p + c_q + d_{pq} .$$

For the term u_{ji} in model (4.1) we must substitute

$$u_{ji} = u_{pi} + v_{qi} + w_{pqi} .$$

The model for an experiment carried out according to the principle of complete randomization is thus

$$x_{pqi} = \mu + b_p + c_q + d_{pq} + e_{pqi} . \tag{14.1}$$

For randomized blocks with one experimental unit for each treatment per block, the model is

$$x_{pqi} = \mu + b_p + c_q + d_{pq} + z_i + u_{pi} + v_{qi} + w_{pqi} + e_{pqi} . \tag{14.2}$$

In both models e_{pqi} stands for the effect of the heterogeneity factors (for randomized blocks: the intra-block heterogeneity factors) and the interactions between these factors and the experimental factors. In model (14.2) u_{pi}, v_{qi}, and w_{pqi} stand for the interactions between the experimental factors and the inter-block heterogeneity factors. Without loss of generality we can let $\Sigma b_p = 0$, $\Sigma c_q = 0$, $\Sigma_p d_{pq} = \Sigma_q d_{pq} = 0$, and $E(z) = 0$. Referring to the discussion in Section 4, we confine ourselves to the following statements: (1) z_i stands for n independent values of a random variable z; (2) u_{pi} stands for n independent values of each of r random variables, one for each P_p; (3) v_{qi} stands for n independent values of each of s random variables, one for each Q_q; and (4) w_{pqi} stands for n independent values of each of rs random variables, one for each PQ combination. Since z, u, v and w are effects of inter-block heterogeneity factors, they cannot be assumed to be stochastically independent. Neither can it be assumed that var(u), var(v), var(w), and var(e) are constant among the treatments.

If the experiment has been carried out for testing purposes, there are three null hypotheses to be considered, i.e. $b = 0$, $c = 0$ and $d = 0$. In the case of randomized blocks, six mean squares and three F-ratios are available for testing. Using for the mean squares the symbols V_p, V_q, V_{PQ}, V_{PR}, V_{QR} and V_{PQR} (R symbolizing replication), we have, for example,

$$V_P = \frac{ns}{r-1}\sum(\bar{x}_p - \bar{x})^2$$

$$V_{PR} = \frac{s}{(n-1)(r-1)}\sum\sum(\bar{x}_{pi} - \bar{x}_i - \bar{x}_p + \bar{x})^2,$$

the numbers of degrees of freedom being $(r-1)$ and $(n-1)(r-1)$, and $F_P = V_P/V_{PR}$. This ratio is the only one, if any, that can be used for the testing of $b = 0$. Wŗiting $b = 0$, it will be seen from model (14.2) that

$$\bar{x}_p - \bar{x} = (\bar{u}_p - \bar{u}) + (\bar{w}_p - \bar{w}) + (\bar{e}_p - \bar{e})$$

and

$$\bar{x}_{pi} - \bar{x}_i - \bar{x}_p + \bar{x} = (u_{pi} - \bar{u}_i - \bar{u}_p + \bar{u}) + $$
$$+ (\bar{w}_{pi} - \bar{w}_i - \bar{w}_p + \bar{w}) + (\bar{e}_{pi} - \bar{e}_i - \bar{e}_p + \bar{e}).$$

It will be seen that the two differences both depend on u, w and e, and it can be shown that if $b = 0$, $E(V_P) = E(V_{PR})$. It can also be shown that if $c = 0$, $E(V_Q) = E(V_{QR})$, and if $d = 0$ that $E(V_{PQ}) = E(V_{PQR})$. Therefore, if the research worker wishes to test the three null hypotheses, using the F-test, an analysis of variance must be carried out according to the key shown in **Table 14.1.**

Table 14.1

Source of variation	Number of degrees of freedom	Mean square	F
Replication	$n - 1$		
P	$r - 1$	V_P	V_P / V_{PR}
PR	$(n - 1)(r - 1)$	V_{FR}	
Q	$s - 1$	V_Q	V_Q / V_{QR}
QR	$(n - 1)(s - 1)$	V_{QR}	
PQ	$(r - 1)(s - 1)$	V_{PQ}	V_{PQ} / V_{PQR}
PQR	$(n - 1)(r - 1)(s - 1)$	V_{PQR}	

As in experiments with one factor, large values of the mean square ratios are an indication of significant departures from the null hypotheses. However, in this case also it should always be remembered that the probability on the null hypotheses of $F \geqslant F_\alpha$ is larger than α, implying that the interactions tend to inflate the level of significance.

In cases in which complete randomization has been used, we have the same mean squares for P, Q and PQ and only one residual (or error) mean square with $rs(n - 1)$ degrees of freedom. Using this mean square, the research worker must choose a method of testing that is adapted for such correlated F-ratios; cf. Section 9.

The finding of significant main effects and significant interactions will certainly be of interest; and in cases in which the experiment has been planned for some practical purpose such knowledge might also be useful. It is evident, however, that such a finding does not imply that the analysis is completed. Usually the investigator wishes to know more about the details.

Then, the method of analysis will be different for the different types of treatments. For instance, if the categories of both factors are quantitative, e.g. quantitative levels of fertilizers, a careful problem analysis prior to and included in the design of the experiment will often indicate the treatment contrasts that should be estimated. The statistical problem will thus be reduced to the estimation (including computation of confidence limits) of these contrasts, the technique of which has been described in Section 7. Problem analysis may also show that it is unnecessary to include all rs combinations of the categories of the factors. This will enable a reduction in the cost of the experiment – a reduction that may be used to increase precision by increasing the number of replications.

An alternative technique in such cases is regression analysis. Suppose that the levels of the P factor are $x_{11}, x_{12}, \ldots, x_{1r}$, and those of the Q factor are $x_{21}, x_{22}, \ldots, x_{2s}$, or x_{1p} $(p = 1, 2, \ldots, r)$ for the P factor and x_{2q} $(q = 1, 2, \ldots, s)$ for the Q factor. Then a regression analysis can be carried out, using for independent variables $x_{1p}, x_{2q}, x_{1p}^2, x_{2q}^2, \ldots, (x_{1p}\ x_{2q})$, $(x_{1p}^2\, x_{2q}), \ldots$. The maximum number of independent variables is, of course, equal to $rs - 1$. If, for instance, $r = s = 2$, the independent variables that should be used are x_{1p}, x_{2q} and $(x_{1p}\, x_{2q})$. The regression technique is described in Section 10, to which the reader is referred.

As is well known, however, in some cases problem analysis pointing out the contrast to be estimated is very difficult. For instance, in agricultural experiments with varieties such an analysis might often be impossible. In such cases the statistical problem is reduced to the problem of ranking the treatments (P, say) according to the mean value of the observed random variable, and possibly to the grouping of the treatments; cf. Section 11. In a factorial experiment this can be carried out for each category of the other factor (Q). Then, if the interaction between the two factors is trivial, the ranking of P_p will be expected to be the same for the different categories of Q. It is reasonable to expect, however, that the categories of P belonging to the group of superior treatments are different for the different categories of Q. For instance, if P_p are varieties of wheat, Q_q different levels of nitrogen fertilizer, and the observed random variable is stiffness of the straw, such results may well be found, and would be very important and useful.

We do not, however, propose to recommend the use of a definite methodology for the analysis of data obtained in factorial experiments. From one case to another, there is too great a diversity of questions requiring an answer. The important things are that a careful problem analysis should be made in advance, and that the experiment should be planned and carried out in such a way that answers can be expected to the questions which the analysis has indicated. If such a working rule is adopted, the methods outlined in Sections 7–11 will serve the purpose.

It is a well-known disadvantage of factorial experiments that an increasing number of factors – even if the number of categories for each factor is small – may lead to a large and sometimes prohibitive number of treatments. Various attempts have been made to overcome the difficulties that ensue from such large numbers of treatments. In a later section we shall return to the problem; for the present we confine ourselves to the plan known as the *split-plot design.*

The necessity for the use of this design arises in two ways. It may arise because the number of possible experimental units belonging to the same replication is less than the number of treatment combinations – for instance, if a replication consists of animals (e.g. pigs) belonging to the same litter. It may also arise because some treatments need larger experimental units than others. Federer [10] has listed a number of cases in which the split-plot plan should be used.

In an agricultural field-plot experiment with two factors P_p and Q_q ($p = 1, 2, \ldots, r$, $q = 1, 2, \ldots, s$) the replications (blocks) are each divided into s (say) main plots, and each main plot is divided into r sub-plots. The main plots are treated with Q_q, randomly allocated. The sub-plots are treated with P_p, also randomly allocated.

The model for this case is a simple extension of (14.2), the extension being the inclusion of a term e'_{qi} ($i = 1, 2, \ldots, n$). Now e_{pqi} stands for the effect of the heterogeneity factors within main plots and the interaction between these factors and P_p. In the same way e'_{qi} stands for the effect of the heterogeneity factors between main plots and the interaction between these factors and Q_q. As in (14.2), u, v and w are the interactions between the experimental factors and the inter-block heterogeneity factors.

It will now be found that

$$\begin{aligned}
\bar{x}_{pq} &= \mu + b_p + c_q + d_{pq} + \bar{z} + \bar{u}_p + \bar{v}_q + \overline{w}_{pq} + \bar{e}'_q + \bar{e}_{pq}\,; \\
\bar{x}_p &= \mu + b_p + \bar{z} + \bar{u}_p + \bar{v} + \bar{w}_p + \bar{e}' + \bar{e}_p\,; \\
\bar{x}_q &= \mu + c_q + \bar{z} + \bar{u} + \bar{v}_q + \bar{w}_q + \bar{e}'_q + \bar{e}_q\,.
\end{aligned}$$

It will be seen that the difference between two $\bar{x}_q$, being an unbiassed estimator of the corresponding contrast, depends on v, w, e' and e, while the difference between two $\bar{x}_p$ is dependent on u, w and e. It is commonly thought that a contrast among P-altematives is estimated with higher precision than a contrast among Q-alternatives. This would certainly be true if the additive model were adequate. However, if we use a non-additive and realistic model, including all kinds of interactions, nothing can in general be known about the relative precisions of the two estimators. But if the method described in Section 7 is adopted, confidence limits of contrasts among Q categories and among P categories, as well as contrasts among PQ categories, can easily be computed.

15 Methods intended to yield estimators of increased precision

It is only natural that both research workers and statisticians should have been concerned with the development of experimental designs which are intended to yield increased precision of the estimators. Among these designs, those based on confounding and on the use of concomitant random variables are perhaps the most frequently adopted in practice. In exceptional situations, identical twins and the like are used as replications in experiments based on the randomized block principle. If the replications are regarded as a sample representing an abstract population, it is easy to see, however, that most designs invented for the purpose of increasing the precision meet this requirement at the expense of the generality of the inferences. Therefore it is important that the research worker should always bear in mind the purpose for which the experiment is planned.

It is a well-known fact that whatever the outcome of an experiment may be – whether a rule or merely a statement – this cannot have universal validity. For instance, such a statement is always restricted by the limited heterogeneity of the experimental material. With regard to the precision of an estimator, this implies that the extent of heterogeneity is directly related to the generality of the population. In basic research, the data produced by an experiment may be satisfactory as evidence for some rule or statement, even if the heterogeneity of the experimental material is small. However, if the investigator is seeking a rule that can be used as a guide in practical situations, it is necessary that it be inferred from data obtained in an experiment which is so planned and carried out that the heterogeneity of the material is dependent on all those factors which are not controlled in practice. If the rule is an inference based on data from an experiment on material of less heterogeneity, its practical significance may be lost because of the effects of these factors, and in such circumstances the likelihood of the project achieving its desired end may be very small.

For instance, this would be the case if identical twin calves were used as experimental units in a randomized block experiment for comparison of the effects of two feeding categories, the purpose being to learn which of the two should be used in practice for the feeding of calves. The use of identical twins as units implies that the research worker controls genetic factors which may be important sources of heterogeneity. However, the results of such an experiment may be important, provided the limitations on the validity of the results are not forgotten or ignored. Other examples of this sort are discussed by Linder [24], p. 13, and Cox [7], p. 25.

The use of confounding implies that each replication is divided into a number of main units (usually called blocks), and each main unit is divided into a number of sub-units. Some of the treatment effects are then confounded with the effects of heterogeneity among the main units. Thus the split-plot design belongs to this class. The effect of the use of confounding is that the contrasts corresponding to the differences between confounded effects must be estimated by means of differences between main units. The other contrasts can be estimated by means of differences between sub-units. However, all contrasts can be estimated by means of observations obtained in each single replication, and the inferences thus possess such validity or generality as the sample of replications permits. Often, but not always, contrasts corresponding to unconfounded effects are estimated with higher precision than the other contrasts.

The inference is different if observations of a concomitant random variable are used for the purpose of reducing the heterogeneity of the experimental material. Suppose, for instance, that an experiment for the comparison of the effects of $k = 2$ feeding categories to calves is carried out, and that the design is complete randomization. Let the principal random variable (x_0) be increase of weight during the feeding period. In this case it might be possible to reduce the heterogeneity to some extent by means of observations of the weight (x_1) of the animals at the start of the experiment. Then, assuming that the use of the observations of x_1 reduces the heterogeneity, it is evident that the validity of the inference with regard to the relative effects of the treatments is also reduced as compared with the validity of the result obtained without the use of observations of x_1.

Suppose that e in (3.1) is replaced by $\beta_j(x_{1ji} - \bar{x}_1) + e'_{ji}$, where x_1 is the concomitant random variable, the distribution of which is completely independent of the treatments. Thus the model is

$$x_{0ji} = \mu + a_j + \beta_j(x_{1ji} - \overline{x}_1) + e'_{ji}$$

$(j = 1, 2, \ldots, k,\ \ i = 1, 2, \ldots, n)$. It follows that

$$\bar{x}_{0j} = \mu + a_j + \beta_j(\bar{x}_{1j} - \bar{x}_1) + \bar{e}'_j\,.$$

It will be seen that it is not assumed that the regression coefficient of e on x_1 is the same for all treatments. In our opinion such an assumption would be unrealistic.

The so-called adjusted treatment means are

$$\bar{x}'_{0j} = \bar{x}_{0j} - b_j(\bar{x}_{1j} - \bar{x}_1).$$

It will be found that $E(\bar{x}'_{0j}) = \mu + a_j$, so that an unbiassed ranking of the treatments will be obtained by using the adjusted means. Therefore, the difference $d'_{pq} = \bar{x}'_{0p} - \bar{x}'_{0q}$ is an unbiassed estimator of the contrast $(a_p - a_q)$.

Writing $A_j = \Sigma(x_{1ji} - \bar{x}_{1j})^2$ and $B_j = \Sigma(x_{0ji} - \bar{x}_{0j})^2$, it will be found that the mean square of d'_{pq} is equal to*

$$s^2_{d'} = \frac{B_p(1 - r_p^2) + B_q(1 - r_q^2)}{2(n-2)}\left\{\frac{2}{n} + \frac{(\bar{x}_{1p} - \bar{x}_1)^2}{A_p} + \frac{(\bar{x}_{1q} - \bar{x}_1)^2}{A_q}\right\},$$

where r_p and r_q are the coefficients of correlation between x_0 and x_1 for the treatments T_p and T_q. Hence, approximately correct confidence limits for the contrast $(a_p - a_q)$ are $d'_{pq} \mp t_\alpha s_d$, the number of degrees of freedom being $2(n-2)$.

In the formula for $s_{d'}$ $(\bar{x}_{1p} - \bar{x}_1)^2$, $(\bar{x}_{1q} - \bar{x}_1)^2$, A_p and A_q must each be regarded as fixed, non-random quantities; this indicates the reason for the loss of validity of the inference.

* If the number (n) of replications is different for the different treatments, the formula is

$$s^2_{d'} = \frac{B_p(1 - r_p^2) + B_q(1 - r_q^2)}{n_p + n_q - 4}\left\{\frac{n_p + n_q}{n_p \cdot n_q} + \frac{(\bar{x}_{1p} - \bar{x}_1)^2}{A_p} + \frac{(\bar{x}_{1q} - \bar{x}_1)^2}{A_q}\right\}.$$

Therefore, if d' is used instead of $d = \bar{x}_{0p} - \bar{x}_{0q}$ as the estimator of the contrast, and $s_{d'}$ is used for the computation of the confidence limits, no inference can be applied to the whole population represented by the experimental units. We now have to deal with a sub-population characterized by $\bar{x}_{1p}$, $\bar{x}_{1q}$, $\bar{x}_1$, A_p and A_q. This loss of validity of the inferences should not be ignored.

Among designs that are intended to increase precision we may also include the *Latin square design.* In field-plot experimentation it has been tempting to try to obtain partial control over heterogeneity not merely in one direction, as in randomized blocks, but in two orthogonal directions, and this led to the invention of the Latin square design. If the number of treatments is k, the field is divided into k^2 units (plots) lying in k rows and k columns. The treatments are then allocated to these units in a random manner, but in such a way that each treatment occurs once in each row and once in each column.

The theory of this design andthe associated statistical analysis deal with a population that consists of all possible k^2 squares; it does not seem possible to regard this design in any other way. Thus, if we are concerned with an abstract population represented by a sample of replications, the Latin square design does not comply with our conditions, unless the whole square is regarded as a replication. In the latter case the experiment must be carried out by means of a sample of such squares. Such an experiment would be very expensive and would not necessarily yield significantly more precise estimators than a randomized block experiment. Therefore, concerned as we are with designs intended to produce data from which inferences about an abstract population can be drawn, we conclude that the Latin square design is not to be recommended.

16 Experiments with large numbers of treatments

In some experimental situations the number of treatments is very large, as, for instance, in field-plot experiments where the treatments are varieties. In such cases the number of experimental units necessary for a complete replication in a randomized block layout might become so great that the advantage of the randomized block design over complete randomization would be illusory. Other examples are factorial experiments with large numbers of factors and/or large numbers of categories of the individual factors. However, a large number of treatments sometimes means merely that the number is large in comparison with the number of easily accessible experimental units.

In order to counterbalance the loss of precision caused by large complete replications, a number of designs known as *incomplete blocks* and *lattices* have been invented. Much work and time is spent and much ingenuity shown in the construction of these designs. Most modern textbooks on experimental design give detailed descriptions of the different types.

It is hardly possible to give a general account of these designs. The main idea is, however, that a replication is divided into a number of main units (usually called "blocks") in the same way as in the split-plot design. If we compare this plan with the latter design, the difference is that the treatments are allocated to the experimental units in such a way that in certain circumstances it is reasonable to assume that the different contrasts are estimated with the same precision. The consequence is that the grouping of the treatments is changed from one replication to another.

It is usually thought that these arrangements of the treatments lead both to equal precision of the estimators of the different contrasts and to increased precision as compared with randomized block experiments without grouping of the treatments. However, certain disadvantages have also been recognized. Federer [10] writes: "Missing data or unequal error variances considerably complicate the analysis; if either situation is likely to occur, it is suggested that the experimenter improve the experimental technique and (or) use a randomized complete block design." In our view, research workers should always regard missing data as liable to occur, and equal error variances are practically never realized. We therefore find it difficult to recommend the use of these designs. Besides, we also think that the designs are impracticable because of their inflexibility.

There are, of course, practical advantages in grouping the treatments if their number is very large. For instance, in a field-plot experiment such activities as planting and sowing take considerable time, and the same applies to operations during harvesting. It would therefore be advantageous if the area of land that represents a replication were divided into main plots, so that the experimenter can deal with these one at a time. The split-plot plan will meet this requirement.

Suppose that the treatments are divided into a number (s) of groups and that the replications are divided into the same number of main units. Then, the s groups of treatments can be allocated to the main units in a random way, and the treatments belonging to a group can be allocated randomly to the experimental units within the main unit. Statisticians are familiar with this use of the split-plot design; of. Cochran and Cox [6]. The reason why balanced incomplete block designs are preferred and recommended seems to be, first and foremost, that such designs are thought to yield comparisons of equal precision. Research workers who do not believe in equal precision of estimators of contrasts will hardly find any advantage in the incomplete block designs over the split-plot plan. On the other hand, it is easy to point out several advantages of the latter.

The most important advantages of split-plot designs are the following. (1) It is unnecessary for the groups to be of the same size, i.e. to cover the same number of treatments; on the contrary, it is important that the treatments be divided, if possible, into "natural" groups. (2) There is no rule connecting the number of treatments, the number of groups, and the number of replications. (3) Missing data and interactions between the treatments and the heterogeneity factors do not complicate the statistical analysis any more than if randomized blocks without any grouping of the treatments had been used.

Interactions between the treatments and the heterogeneity factors make it impossible to form any prior judgement of the relative precision of the different contrast estimators. It is likely, however, that the precision is higher for contrasts among treatments belonging to the same group than it is among treatments that belong to different groups. To some extent the effect of heterogeneity among the main units can be reduced if a check treatment is included in all treatment groups. Then if r_q is the number of treatments in group q, the main unit used for this group must cover at least $r_q + 1$ experimental units. The check treatment must be regarded as belonging to the group and, along with the other treatments, must be allocated to the experimental units in a random way. If, for practical reasons, the experimenter deals with the main units one at a time, a time-factor is introduced. In such a case it is particularly important to include a check treatment, so that bias caused by the time-factor can be removed.

Let the observations of some random variable (e.g. yield) be x_{pqi}, where $p = 1, 2, \ldots, r_q$, $q = 1, 2, \ldots, s$, and $i = 1, 2, \ldots, n$, n being the number of replications (blocks). The model describing x_{pqi} is now a simplification of the model for a two-factor experiment according to the split-plot design, and can be written

$$(16.1) \qquad x_{pqi} = \mu + z_i + a_{pq} + v_{qi} + w_{pqi} + e'_{qi} + e_{pqi} \, .$$

In this model e' stands for the effect of heterogeneity among main units and e for the effect of heterogeneity among the experimental units within main units; but, of course, e' and e also cover the interaction between treatments and heterogeneity factors. The terms v and w stand for the interactions between groups and replications, and between treatments within groups and replications, respectively. Since the replications are regarded as a random sample, representing an abstract population, all terms in the model except μ and a must be regarded as random variables. The term a_{pq} can be written $a_{pq} = \bar{a}_q + (a_{pq} - \bar{a}_q)$, and we can without loss of generality let $\Sigma\bar{a}_q = 0$.

It will be found in this case also that the treatment mean $\bar{x}_{pq}$ is an unbiassed estimator of the effect $(\mu + a_{pq})$, and hence that the treatment means yield an unbiassed ranking of the treatments. Consequently, a linear function of treatment means is an unbiassed estimator of the corresponding contrast. Here, also, interactions between treatments and the heterogeneity factors imply that correlations exist between x_{pqi} among the treatments and that var(x) is different for the different treatments. However, the method described in Section 7 can be used for the computation of the confidence limits of the contrasts. The method described in Section 11 can be used for the grouping of the treatments, e.g. for the isolation of a group of superior treatments. It is likely, but not obvious, that the confidence intervals of some of the contrasts can be shortened by means of the observations for the check treatment.

If a check treatment T_0 (say) has been included, x_{pqi} can be replaced in all the procedures by $y_{pqi} = x_{pqi} - x_{0qi}$, where x_{0qi} are the observations for the T_0 . In practice the research worker will hardly use more than one, or perhaps two experimental units per main unit and replication for the check treatment. If two units have been used, x_{0qi} stands for the mean value of two observations.

The preceding discussion concerns cases in which the number of treatments is large, but where there is no shortage of experimental units per replication. There are cases, however, where there may be such a shortage; for example, this may be so if the research worker wishes to use litters as replications in a pig-feeding experiment; in this case the number of treatments that can be included is much restricted.

In such cases a number of samples of experimental units can be used as the main units in an experiment according to the split-plot design. Then, a replication would consist of a sample of such main units. In our example, the research worker can use litters to represent the main units, and a replication would then consist of a number of litters. Thus, if the total number of litters for the whole experiment is sampled from the same stock, the heterogeneity among the main units is equal to that among the replications, and the split-plot plan ought to be combined with complete randomization. If the investigator wishes to use the split-plot plan and the randomized block design, the replications should be sampled in a different way. For instance, he can use a sample of stocks to represent the sample of replications. There are other reasons for such selection of the replications which we consider in the next section.

17 Experiments intended to give results for practical utilization

We have seen in a previous section that if an experiment is carried out for the express purpose of providing a basis upon which advice to practitioners can be given, it should be designed so that none of those factors are controlled that are not under control in practice. To design an experiment that satisfies this requirement is certainly a difficult task. These factors are not as a rule fully known to the research worker, and the experiment must be planned in such a way that their effects can be regarded as random effects. Also, the hard fact is that inferences, if any, can only be applied to the population represented by the actual experimental material regarded as a random sample. This abstract population may not be broad enough to cover all cases that may occur in practical situations. Therefore, the experimenter, when consulted on a particular case, would be well advised to show such modesty as to recommend a treatment only if it is known that the case belongs to this population. If he does not make such a reservation, he may take the risk of extrapolating his experimental result outside the sphere covered by the experiment. However, the investigator can usually ensure that the population is broad enough to cover the great majority of cases that actually occur.

Another fact is that if the research worker's recommendation is acted upon by all practitioners and the choice of treatment involves economic consequences, very often some practitioners would be better off by using another treatment. This is so because all populations consist of sub-populations differing in one characteristic or another, and the most successful treatment might not be the same in all sub-populations. In practice there are always limits to what a research worker can know about the circumstances that make one particular treatment superior to others that might be used. Therefore, he can only recommend a particular treatment for cases belonging to a certain population, which is the one that is represented by his experimental replications.

For instance, if a research worker in the agricultural sphere recommends a certain variety of wheat to all farmers in a geographical area, he should know from the results of his experiment that the use of this variety is expected to inflate the yield for the whole area, as compared with the use of another variety included in the experiment. But possibly he also knows, or can guess, that the yield might be even larger if some of the farmers do not act upon his recommendations.

Before he starts the detailed planning of the experiment, it is necessary for the research worker to make some difficult decisions, however vague, with regard to the generality of the population. This implies that he must decide what kind of experimental units should be used. In agricultural experimentation a unit must be a field suitable for growing the plant in question; and which fields are suitable is something that must be definitely known. In industrial research a unit may be an industrial plant, but it is not evident that all plants should be regarded as being suitable. Thus in practically all cases there are a number of difficult questions that must be dealt with in advance of the planning of the experiment – decisions to be taken for the purpose of delimiting the borders of the population about which knowledge is required.

When these decisions have been made, the next step might seem obvious, i.e. to take from the accepted cases one or more random samples to be used as the experimental material. But it may not be so simple as that. The research worker may encounter many obstacles; for instance, it may happen that a field selected for an agricultural experiment is already intended to be used for some other purpose.

Consulting the literature dealing with experimental design, it will be found that most scientific work and discussion have centred on experiments that can be characterized as *local*. In agricultural field-plot experimentation, a "local" experiment is one that is carried out in a chosen field in one season; in experimentation on feeding pigs, it is one carried out with pigs chosen from a single stock and at a chosen farm; or it can be an experiment carried out on a single industrial plant.

It is evident that for an experimental result to be used as a basis for recommendations for practical work, a local experiment does not suffice. The reason is, of course, that most often the population in which such results can be applied is too restricted. In any case this is true if interactions exist between the treatments and the environmental factors which are not controlled in actual practice.

Among research workers in the agricultural sphere it now seems generally recognized that both geographic heterogeneity factors and factors the effects of which vary from season to season make themselves felt, and also that there are interactions between these heterogeneity factors and the treatments. If it were not so, the results from local experiments would be sufficient. Research workers in this field must therefore plan and carry out experiments in such a way that the replications cover a geographical area and a number of seasons. In principle the situation is hardly different in other fields of research, even if the interactions between treatments and heterogeneity factors may be of varying importance in the different cases.

In the literature concerning such experimental situations, an experiment is usually regarded as consisting of a number of repeated local experiments. In our opinion it should not be considered in this sense, but as a single experiment on its own, planned and carried out for its own specific purpose.

Keeping to our agricultural example, extension both geographically and in time can be achieved in two different ways. A sample of localities must be chosen, and within each locality a site for the replication. Then the research worker can use the same sample of localities for all seasons, only changing the site for the replication from season to season. He can also choose a new sample of localities for each season included in the experiment. The latter plan is perhaps preferable, since one may expect that the heterogeneity factors would be better covered in an experiment planned in this way than if the same sample of localities were used for all seasons. But it is evident that the use of the same sample of localities at all seasons is the simpler of the two plans to manage in practice. Now the research worker may succeed in choosing localities and sites so as to have a sample of replications which closely resembles an ordinary random sample, representing the agricultural area. But both the concepts of a "sample of seasons" and a "population of seasons" are too vague. The term "season" implies no more than a kind of classification with regard to the variation in the effects of some environmental factors.

Whether or not a new sample of localities is taken for each season, the question is : will the sample give a satisfactory coverage for the geographical heterogeneity factors and for the period of time for which the inferences (or forecasts) are intended ? It is possible to be fairly certain that there is satisfactory coverage for the geographical factors. But for the time factors – i.e. climatic factors – the answer to the question of coverage depends on what can be said about climatic changes in coming years, which at present is very little. It is evident, however, that if relatively large climatic changes have taken place during the seasons covered by the experimental replications, the research worker can be more confident in giving advice to practitioners than he could be if the replications covered less climatic variation. The interaction between a treatment and the climatic factors may be insignificant, and it is obvious that the research worker should feel more confident in recommending a treatment showing small interaction with these factors than he could if the interaction were greater. The same is the case with regard to the interaction between treatments and geographic heterogeneity factors. Therefore both kinds of interaction should be taken into account when the research worker is dealing with the ranking and classification of the treatments.

If a new sample of localities is taken for each season, and if there are n_1 seasons and n_2 localities for each season, the sample consists of $n = n_1 n_2$ replications. If only the site is changed from season to season and the numbers of seasons and localities are n_1 and n_2, the sample still consists of $n = n_1 n_2$ replications. In both cases the population is the one which the sample of replications represents in the sense of a random sample. The two populations are not identical, but the difference cannot be important. On the other hand, the use of the same sample of localities at all seasons has an advantage over the other plan in that it makes a more detailed analysis possible.

No matter which of the two plans is used, a replication does not usually consist of k experimental units, k being the number of treatments. Most often a number (m) of units is used for each treatment, and the design may be complete randomization, or perhaps randomized blocks.

Suppose first that a new sample of localities is taken for each season, and that the design for each replication is complete randomization. Then the model for the mean of the observed random variable for treatment T_j $(j = 1, 2, \dots, k)$ and replication i $(i = 1, 2, \dots, n)$ is

$$(17.1) \qquad \bar{x}_{ji} = \mu + a_j + z_i + u_{ji} + \bar{e}_{ji} ,$$

where the different terms stand for the same effects as in the model for a randomized block experiment. It will be found that the model for the treatment mean is

$$\bar{x}_j = \mu + a_j + \bar{z} + \bar{u}_j + \bar{e}_j .$$

Without loss of generality, we can let $\Sigma a_j = 0$, $E(z) = 0$, $E(u) = 0$ for each j, and $E(\bar{e}) = 0$ for each combination (j, i). Hence it will be found that

$$E(\bar{x}_j) = \mu + a_j \quad \text{and} \quad E(\bar{x}_p - \bar{x}_q) = a_p - a_q ,$$

i.e. that $\bar{x}_j$ is an unbiassed estimator of the treatment effect $(\mu + a_j)$, and $(\bar{x}_p - \bar{x}_q)$ is an unbiassed estimator of the contrast $(a_p - a_q)$. If the experimental units are of the same size as those used in a local experiment, it will usually be found that var(e), or $\frac{1}{k}\Sigma\,\text{var}_j(e)$, is approximately the same as is found in a local experiment. However, since the heterogeneity among the replications is usually much greater than it is in a local experiment based on the randomized block principle, it must be expected that $\text{var}(u) = \frac{1}{k}\Sigma\,\text{var}_j(u)$ is much inflated as compared with a local experiment. Therefore the reliability of the F-test for the testing of the null hypothesis $a_j = 0$ is questionable, the probability of $F \geqslant F_\alpha$ being also inflated. However, the treatment mean $\bar{x}_j$ is an unbiassed estimator of the treatment effect, implying that an unbiassed ranking of the treatments is obtained by means of the treatment means.

Suppose, next, that the same sample of localities is used at all seasons, then the mean of the observed random variable for treatment T_j $(j = 1, 2, \dots, k)$, in season i $(i = 1, 2, \dots, n_1)$ and locality h $(h = 1, 2, \dots, n_2)$, can be written $\bar{x}_{jih}$, and the model is

$$(17.2) \quad \bar{x}_{jih} = \mu + a_j + z_i + y_h + u_{ih} + v_{ji} + w_{jh} + e'_{jih} + \bar{e}_{jih} .$$

In this case also it can be shown that the treatment mean $\bar{x}_j$ is an unbiassed estimator of the effect $(\mu + a_j)$, and that $(\bar{x}_p - \bar{x}_q)$ is an unbiassed estimator of the contrast $(a_p - a_q)$.

In addition it can be shown, using either plan, that a linear function of the treatment means is an unbiassed estimator of the corresponding contrast. However, the precision of the estimator is different for the different contrasts. Therefore the confidence limits of the contrasts must be computed by means of individual mean squares, as described in Section 7.

We have seen that if the problem is to group the treatments, and particularly if it is required to isolate a group of superior treatments, the research worker should also consider the interactions, first and foremost the interaction between treatments and seasons. Now if a new sample of localities is chosen for each season, it is impossible to separate the treatment–locality interaction and the treatment–season interaction. However, if the same sample of localities has been used at all seasons, we may consider that the function

$$\delta_{ji} = \bar{x}_{ji} - \bar{x}_j$$

and the graphs of δ_{ji} against i, one for each treatment, will probably be useful aids. Probably also

$$\lambda_j = \sum_i \delta_{ji}^2 \Big/ \sum\sum \delta_{ji}^2 \tag{17.3}$$

may prove to be useful for characterization of the treatments. For example, if the treatments are varieties, the research worker would prefer for recommendation a high-yielding variety for which the value of λ_j is small.

With regard to the treatment–locality interaction the equivalent statistic is

$$\lambda'_j = \sum_h \delta_{jh}^2 / \sum\sum \delta_{jh}^2 \tag{17.4}$$

where

$$\delta_{jh} = \bar{x}_{jh} - \bar{x}_j .$$

It is evident, however, that of the two λ, (17.3) is the more useful in practice.

If a new sample of localities is taken for each season, the research worker can use (17.3), but in this case λ_j is dependent upon the confounded treatment–locality and the treatment–season interactions.

So far we have been concerned with problems in agricultural field experimentation. It seems likely, however, that the difficulties encountered in this sphere of research reflect to a large extent the problems with which research workers have to deal generally. It may be possible, of course, that industrial plants are so far advanced technically that heterogeneity among the plants and the effects of climatic factors are negligible; but it is more likely that such examples are exceptions rather than the rule.

To return to another example, we discussed in Section 16 the design of an experiment for comparing a number of alternative pig-feeds in a situation where the number of treatments is too large to be covered by a litter. It was suggested that a litter should be used in the same way as a main unit in a field experiment based on the split-plot plan, and in this case a number of litters must be sampled to constitute a replication. It was further suggested that the replications should be sampled from different stocks, one replication from each stock. There is certainly heterogeneity among stocks, particularly heterogeneity due to genetic factors, as there is among localities in a field experiment. But there might also exist heterogeneity due to differences in the environmental conditions under which the pigs are living, indicating that climatic factors may be important in this case also. In fact the experimental situation is essentially the same as the one described above. The difference exists merely in the relative heterogeneity due to the different sources. Our conclusion is, therefore, that such experiments should be planned and carried out according to the same principles as those used in field experimentation. The experimental material (the replications) should be sampled in such a way that a reasonable amount of the heterogeneity among stocks as well as heterogeneity due to differences in living conditions are covered. In order to satisfy the latter requirement the replications must cover a number of years.

If the purpose is to obtain data upon which rules for practical work can be based, it will probably be found that there are numerous factors causing heterogeneity that cannot be or are not controlled in practice. The experiment must therefore be planned so that the sample of replications covers the heterogeneity due to these factors. If great care is not taken to ensure that this requirement is satisfied, the population represented by the sample of replications will cover merely a part of the actual situations for which the research worker's recommendations are intended.

It is evident that an experiment of this kind and for this purpose should cover the largest possible number of replications. If we reflect on the best way of using the resources which are at the research worker's command, our conclusion will be that a very simple design should be used for the single replication. One experimental unit for each treatment per replication will suffice, but in practice a couple of units should be used in order to guard against failures.

If m units $(h = 1, 2, \ldots, m)$ are used for each treatment per replication and the design is complete randomization, the model for the observed random variable is

$$x_{jih} = \mu + a_j + z_i + u_{ji} + e_{jih}$$

$(j = 1, 2, \ldots, k,\ \ i = 1, 2, \ldots, n)$. For the contrast $(a_p - a_q)$ the estimator is $(\bar{x}_p - \bar{x}_q)$, the variance of which can be shown to be

$$\text{var}(\bar{x}_p - \bar{x}_q) = \frac{\text{var}(u_p - u_q)}{n} + \frac{\text{var}(e_p - e_q)}{nm}.$$

It will be seen that the effect of increasing m is merely to reduce the last term. Therefore, except in cases in which the interaction (u) between the treatments and the heterogeneity factors is very small, an increase in m will not greatly strengthen the precision of the estimator. On the other hand, an increase in the number (n) of replications will always affect the precision favourably.

From the preceding discussion it will be seen that, as regards the method of analysis, the situation is equivalent to the one which is met with in randomized block experiments. The difference is that, most often, the interaction between the treatments and the heterogeneity factors is more important than it is in randomized block experiments. Even so, it is thought that the methods described in Sections 7 to 12 are adequate for statistical analysis.

18 Some supplementary matters

(A) Every research worker who consistently uses the principle of randomization will sooner or later come across examples where the result of randomization may appear unacceptable. The reason for this is that most often some trend or regularity must be assumed to exist among the experimental units. This is so, for instance, in a field-plot experiment where there is frequently some regularity in more than one direction of the quality of the units. Then, if randomization leads to a result showing congruity between the allocation of the treatments to the experimental units and the regularity among the units, the research worker is probably tempted to take some corrective action. He might, of course, stick to the randomization principle and accept the result, knowing that such a result must take its place in a long-run procedure. However, the worker must often come to a decision immediately, and in such a case it is natural for him to consider rejecting the result of the randomization and to re-randomize.

In principle, any tampering with the result of randomization should be excluded. But we do not think that this would be a reasonable attitude to take, and it is known that highly qualified research workers do, in fact, reject some arrangements of the treatments.

In the literature dealing with the problems of experimental design, the question is usually ignored. However, Cox [7] has discussed the question at some length, referring also to relevant literature. In his treatise, methods of dealing with the question are discussed, e.g. methods based upon the idea of rejecting the more extreme arrangements of the treatments. The difficulty involved in this approach is that it entails using a dichotomy, grouping the results of randomization into acceptable and unacceptable arrangements.

Since we have to recognize the fact that our statistical tools are merely approximations, the problem to consider is what effect, e.g. on the F-test, rejection of some of the arrangements is likely to have. Probably the effect of such a restriction of randomization will be to bring the distribution of F into better harmony with the standard distribution of normal theory. In so far as concerns the probability of the confidence interval of a contrast, it is reasonable to think that the effect is small and will tend to inflate the confidence coefficient. But these statements are no more than guesswork. A statistician who has ample access to an electronic computer might be able to obtain satisfactory evidence by using artificial examples. Here, of course, the examples must be constructed according to realistic models, and the rejection of extreme arrangements of the treatments must be carried to excess. Only when the results of such investigations are presented will it be possible to make up one's mind concerning the practice of curtailing random arrangements.

(B) In the early days of research on experimental design, the common attitude among statisticians was that useful information could hardly be obtained from small samples. The explanation of some of the criticisms raised against this work by Fisher and his collaborators may be found in this attitude. Today it is generally recognized that even very small samples may yield data from which important conclusions can be drawn. However, it is hardly questionable that the founders of experimental design went to the other extreme, partly because they laid too much stress on tests of significance. If one is dealing with the estimation of contrasts, larger samples are usually required.

For instance, suppose that an industrialist contemplates replacing old mass-production machinery by new plant. Then it is not enough for him to know that a test of significance shows that, e.g. the new machinery produces at a higher rate than the old. In his calculations he needs some measure of this difference of rate, and also a value showing the lower margin for the difference. This means that he must utilize the outcome of an experiment in which the old and new machinery are the treatments, and must base his calculations upon the resulting estimate of the contrast and the confidence limits of the contrast. Also it is important to him that the confidence interval of the contrast should not be too wide, which in fact implies that the size of the experiment, or the number of replications, must not be very small.

We believe that this is quite a common situation. If the purpose of an experiment is changed from that of a significance test to the estimation of contrasts, an increase in the number of replications is usually required. But, of course, in some cases the replacement of one treatment by another does not imply any difference of cost, and if it does not, it is enough to know that at least one of the treatments can be classified as the superior one.

Suppose now that k treatments are included in an experiment carried out according to the randomized block design, n being the number of replications. In Section 7 it was explained why the research worker in this case should use "Student's" t with $(n-1)$ degrees of freedom in his computations of the confidence limits of a contrast. There are two principal reasons for this point of view. The first is that the presence of interactions between the treatments and heterogeneity factors implies that there are differences in precision among the contrasts. The second reason is that if a common error mean square is used for all contrasts, the research worker cannot possibly know the level of confidence of the intervals. The use of a common error mean square will always imply that the confidence limits of the contrasts are biassed, and hence that they may be misleading.

It is evident that the use of individual mean squares in computing the confidence limits of the contrasts implies that the number of replications cannot be too small. If this number is in fact too small and, consequently, the confidence intervals of the contrasts are very wide, it is difficult to see what object the experimental data can achieve. Therefore the research worker, in planning his experiment, should always try to estimate the number of replications that will be necessary so that he may expect to obtain a chosen minimum precision. Obviously, this is a very difficult task, and it is evident that the research worker has to utilize experience from previously conducted experiments of a similar kind.

(C) The last question to be considered concerns the relative importance of local and non-local experiments, as described in Section 17. In planning an experiment, it is important to know whether the results are intended to be used for some practical purpose, or whether they are to supplement the investigator's knowledge in his field of research. In the first case it is evident that a non-local experiment is needed. In the second case, what is needed is either a non-local experiment or an experiment in which a very large number of external factors are included as the experimental factors. Thus a local experiment, as described earlier, will not meet the requirements in either case, although such an experiment may furnish the necessary data for drawing preliminary conclusions which can be used as a guide in planning a non-local experiment. For instance, the data may show that some treatments are so inferior that they can be omitted in planning the non-local experiment. This is important, because a non-local experiment is usually very expensive, and it is therefore important that the number of treatments be reduced to a minimum.

There are, of course, exceptions to this appraisal of local experiments, as, for instance, in our example of the industrialist who is interested in comparing two kinds of machinery. In this case the outcome of an experiment may be important merely for the particular industrial plant in question. The experiment can therefore be carried out as a local experiment, even if the outcome is intended to be used as a guide for some practical decision.

There is also a third category of experiments, namely those carried out in a laboratory or under laboratory conditions, where a number of external factors can be controlled*. A fourth category consists of experiments – discussed in Section 15 – which are planned to yield high precision of the contrast estimators. In a comprehensive research program it may be possible to make advantageous use of all these categories of experimental plans. Thus, one of the problems for the research leader is to decide how, and to what extent, the different categories should be utilized. In view of the fact that research funds are usually very restricted, it is important that a balance be found in order to achieve a kind of optimum. In practice, to strike such a balance is certainly difficult.

In most fields of research the investigators seem inclined to spend too great a part of the research funds on local experiments. This may be due to the way in which research methodology is

* The usual design of such experiments may not be satisfactory for biological research work. For such work, therefore, the design should be subjected to critical examination in order to make the data and inferences more realistic.

presented today. To the present author it is apparent that a change is necessary. In some cases – for instance in agricultural research as now practised by some scientists – this change is on its way.

Table I Significance points of F : $\alpha = 0{\cdot}05$

v_1 v_2	1	2	3	4
1·0	161·4	199·5	215·7	224·6
1·1	105·9	127·4	135·9	140·8
1·2	74·7	87·6	92·6	95·6
1·3	56·1	64·4	67·5	69·4
1·4	44·4	50·0	52·1	53·4
1·5	36·2	40·0	41·4	42·3
1·6	30·4	33·0	33·9	34·5
1·7	26·2	28·0	28·6	29·0
1·8	23·0	24·2	24·7	25·0
1·9	20·5	21·3	21·6	21·8
2·0	18·5	19·0	19·2	19·3
2·1	16·9	17·2	17·2	17·3
2·2	15·6	15·7	15·7	15·7
2·3	14·5	14·4	14·4	14·3
2·4	13·6	13·4	13·2	13·1
2·5	12·8	12·5	12·3	12·2
2·6	12·1	11·7	11·5	11·4
2·7	11·5	11·1	10·9	10·7
2·8	11·0	10·5	10·3	10·1
2·9	10·5	10·0	9·7	9·6
3·0	10·1	9·6	9·3	9·1
3·1	9·7	9·2	8·9	8·7
3·2	9·4	8·8	8·5	8·3
3·3	9·1	8·5	8·2	8·0
3·4	8·9	8·2	7·9	7·7
3·5	8·7	7·9	7·6	7·4
3·6	8·5	7·7	7·4	7·2
3·7	8·3	7·5	7·2	7·0
3·8	8·1	7·3	7·0	6·8
3·9	7·9	7·1	6·8	6·6
4·0	7·7	6·9	6·6	6·4
4·1	7·5	6·7	6·4	6·2
4·2	7·4	6·6	6·2	6·0
4·3	7·3	6·5	6·1	5·9
4·4	7·2	6·4	6·0	5·8
4·5	7·1	6·3	5·9	5·7
4·6	7·0	6·2	5·8	5·6
4·7	6·9	6·1	5·7	5·5
4·8	6·8	6·0	5·6	5·4
4·9	6·7	5·9	5·5	5·3

Table II Significance points of F: α = 0·025

ν_2 \ ν_1	1	2	3	4
1·0	648	800	864	900
1·1	375	450	479	494
1·2	240	280	296	304
1·3	166	189	198	202
1·4	121	135	141	144
1·5	92·9	102	105	107
1·6	74·0	79·7	81·9	83·0
1·7	60·8	64·3	65·7	66·4
1·8	51·2	53·3	54·1	54·6
1·9	44·0	45·2	45·6	45·8
2·0	38·5	39·0	39·2	39·3
2·1	34·2	34·2	34·2	34·1
2·2	30·7	30·4	30·2	30·1
2·3	27·9	27·3	27·0	26·9
2·4	25·6	24·8	24·4	24·2
2·5	23·7	22·7	22·2	22·0
2·6	22·0	20·9	20·4	20·1
2·7	20·6	19·4	18·9	18·6
2·8	19·4	18·1	17·6	17·3
2·9	18·4	17·0	16·4	16·1
3·0	17·4	16·0	15·4	15·1
3·1	16·6	15·2	14·6	14·3
3·2	15·9	14·5	13·8	13·5
3·3	15·3	13·8	13·1	12·8
3·4	14·7	13·2	12·5	12·2
3·5	14·2	12·7	12·0	11·6
3·6	13·7	12·2	11·5	11·1
3·7	13·3	11·8	11·1	10·7
3·8	12·9	11·4	10·7	10·3
3·9	12·5	11·0	10·3	9·9
4·0	12·2	10·7	10·0	9·6
4·1	11·9	10·4	9·7	9·3
4·2	11·6	10·1	9·4	9·0
4·3	11·3	9·8	9·1	8·8
4·4	11·0	9·5	8·9	8·5
4·5	10·8	9·3	8·7	8·3
4·6	10·6	9·1	8·5	8·1
4·7	10·4	8·9	8·3	7·9
4·8	10·3	8·7	8·1	7·7
4·9	10·2	8·6	7·9	7·6

Table III Significance points of *F*: $\alpha = 0{\cdot}01$

v_1 v_2	1	2	3	4
1·0	4052	5000	5403	5625
1·1	1992	2380	2534	2618
1·2	1110	1293	1362	1401
1·3	682	776	812	831
1·4	452	503	523	533
1·5	308	347	359	365
1·6	235	252	259	262
1·7	181	191	195	197
1·8	144	149	151	152
1·9	118	120	121	121
2·0	98·5	99·0	99·2	99·3
2·1	84·0	83·2	82·9	82·7
2·2	72·8	71·3	70·6	70·3
2·3	64·1	61·9	61·0	60·5
2·4	57·0	54·5	53·4	52·9
2·5	51·3	48·5	47·4	46·7
2·6	46·6	43·6	42·4	41·7
2·7	42·7	39·5	38·2	37·5
2·8	39·4	36·2	34·8	34·1
2·9	36·6	33·3	31·9	31·2
3·0	34·1	30·8	29·5	28·7
3·1	32·0	28·7	27·3	26·6
3·2	30·2	26·9	25·5	24·7
3·3	28·6	25·3	23·9	23·1
3·4	27·2	23·8	22·5	21·7
3·5	25·9	22·6	21·2	20·5
3·6	24·8	21·5	20·1	19·4
3·7	23·7	20·4	19·1	18·4
3·8	22·8	19·6	18·2	17·5
3·9	22·0	18·7	17·4	16·7
4·0	21·2	18·0	16·7	16·0
4·1	20·5	17·3	16·0	15·3
4·2	19·9	16·7	15·4	14·7
4·3	19·3	16·2	14·9	14·2
4·4	18·8	15·7	14·4	13·7
4·5	18·3	15·2	13·9	13·2
4·6	17·8	14·7	13·5	12·8
4·7	17·4	14·3	13·1	12·4
4·8	17·0	14·0	12·7	12·0
4·9	16·6	13·6	12·4	11·7

BIBLIOGRAPHY

[1] Baker, F.B. and Collier, R.O., Jr. (1966). *J. Amer. Statist. Ass.*, Vol. 61, No. 315.

[2] Bartlett, M.S. (1947). *Biometrics*, Vol. 3, No. 1.

[3] Beale, E.M.L., Kendall, M.G. and Mann, D.W. (1967). *Biometrika*, Vol. 54, pp. 357–65.

[4] Berkson, J. (1950). *J. Amer. Statist. Ass.*, Vol. 45, No. 250.

[5] Box, G.E.P. (1953). *Biometrika*, Vol. 40.

[6] Cochran, W.G. and Cox, G.M. (1950). *Experimental Designs*. Wiley, New York.

[7] Cox D.R. (1958). *Planning of Experiments*. Wiley, New York.

[8] Draper, N.R. and Smith, H. (1966). *Applied Regression Analysis*. Wiley, New York.

[9] Eden, T. and Yates, F. (1933). *J. Agric. Sci.*, Vol. 23.

[10] Federer, W.T. (1955). *Experimental Designs*. Macmillan, New York.

[11] Finney, D.J. (1941). *Ann. Eugen.*, Vol. 11, Part 2.

[12] Fisher, R.A. (1936). *Statistical Methods for Research Workers* (6th edn), Oliver and Boyd, Edinburgh and London.

[13] Fisher, R.A. (1951). *The Design of Experiments* (6th edn), Oliver and Boyd, Edinburgh and London.

[14] Fisher, R.A. (1921). *J. Agric. Sci.*, Vol. 11.

[15] Fisher, R.A. (1923). *Proc. Camb. Phil. Soc.*, Vol. 21.

[16] Fisher, R.A. and Yates, F. (1948). *Statistical Tables for Biological, Agricultural and Medical Research.* Oliver and Boyd, Edinburgh and London.

[17] Gosset, W.S. ("Student"), (1908). *Biometrika*, Vol. 6.

[18] Hack, H.R.B. (1966). *Biometrika*, Vol. 45.

[19] Hartley, H.O. (1938). *Suppl. J. Roy. Statist. Soc.*, Vol. 5, No. 1.

[20] Hartley, H.O. (1955). *Communications on Pure and Applied Mathematics*, Vol. 8.

[21] Horsnell, G. (1953). *Biometrika*, Vol. 40.

[22] Kempthorne, O. (1952). *The Design and Analysis of Experiments.* Wiley, New York.

[23] Kendall, M.G. and Stuart, A. (1958–1966). *The Advanced Theory of Statistics*. Griffin, London.

[24] Linder, A. (1953). *Planen und Auswerten von Versuchen.* Verlag Birkhäuser, Basel/Stuttgart.

[25] Miller, R.G. Jr. (1966). *Simultaneous Statistical Inference.* McGraw-Hill, New York.

[26] Mood, A.M. and Graybill, F.A. (1963). *Introduction to the Theory of Statistics*. McGraw-Hill, New York.

[27] Nair, K.R. (1948). *Biometrika*, Vol. 35.

[28] Neyman, J. (1935). *J. Roy. Statist. Soc., Suppl.*, Vol. II, No. 2.

[29] Ottestad, P. (1953). *Skandinavisk Aktuarietidskrift*, 1952.

[30] Ottestad, P. (1960). *Sci. Reports from the Agric. Coll. of Norway*, Vol. 39, No. 7.

[31] Pearson, E.S. and Hartley, H.O. (1958). *Biometrika Tables for Statisticians*, Vol. 1.

[32] Pitman, E.J.G. (1937). *Biometrika*, Vol. 29.

[33] Quenouille, M.H. (1959). *New Statistical Tables*, Series No. XXVII (Biometrika), Vol. 46, Parts 1 and 2.

[34] Scheffé, H. (1959). *The Analysis of Variance*. Wiley, New York.

[35] Tukey, J.W. (1946). *Biometrics*, Vol. 5, No. 2.

[36] Welch, B.L. (1937). *Biometrika*. Vol. 29.

[37] Wold, H. (1954). *Tracts for Computers*, Vol. XXV, Cambridge Univ. Press.

OTHER GRIFFIN BOOKS ON STATISTICS AND MATHEMATICS

A statistical primer	F. N. DAVID
A course in theoretical statistics	N. A. RAHMAN
An introduction to the theory of statistics	G. U. YULE & M. G. KENDALL
The advanced theory of statistics	M. G. KENDALL & A. STUART
Rank correlation methods	M. G. KENDALL
Exercises in theoretical statistics	M. G. KENDALL
Exercises in probability and statistics	N. A. RAHMAN
Characteristic functions (2nd edn)	E. LUKACS
Rapid statistical calculations	M. H. QUENOUILLE
Statistical communication and detection: with special reference to digital data processing of radar and seismic signals	E. A. ROBINSON
Biomathematics	C. A. B. SMITH
Random processes and the growth of firms	J. STEINDL
The design and analysis of experiment	M. H. QUENOUILLE
Mathematical model building in economics and industry	C-E-I-R LTD.
Computer simulation models	JOHN SMITH
Sampling methods for censuses and surveys	F. YATES
Statistical method in biological assay	D. J. FINNEY
The mathematical theory of epidemics	N. T. J. BAILEY
Probability and the weighing of evidence	I. J. GOOD
Combinatorial chance	F. N. DAVID & D. E. BARTON
Games, gods and gambling: the origins and history of probability and statistical ideas	F. N. DAVID
Mathematical puzzles for the connoisseur	P. M. H. KENDALL & G. M. THOMAS

Descriptive brochure available from the Publishers

26s
£1.30
net

SBN: 85264 166 4