Per Ottestad

METHODS OF EXPERIMENTAL RESEARCH

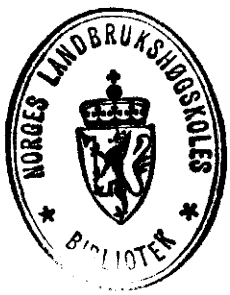
Summary of Lectures Given at
The Agricultural College of Norway

1967

Per Ottestad

METHODS OF EXPERIMENTAL RESEARCH

Summary of Lectures Given at
The Agricultural College of Norway



1967

Contents

1. Preliminaries.

About 40 years ago, important research work on the principles of experimentation was started at Rothamsted Experimental Station in England. The first general account of the results of this research work was given by R.A. Fisher in his book "The Design of Experiments", the first issue of which appeared in 1935. Ten years previously, the first issue of his "Statistical Methods for Research Workers" had been published. In this book the new statistical tool of analysis, known as the analysis of variance, was made known to research workers. A large number of papers and books, dealing with experimental design and statistical analysis, are inspired by these two important treatises.

It is probably well known that the results of the Rothamsted research work were not recognized and valued by the authorities on statistical methods at the time. Today the principles of the Rothamsted school are accepted by almost all statisticians, and it is interesting to notice that now these principles seem to be accepted "hook and line". On the other hand, the principles are not throughout accepted by all research workers. It is a fact that all over the world experimental research work is carried out according to other principles. Often the principle of randomization, perhaps the most important and a lasting contribution made by the Rothamsted school, is ignored. The consequence is that a large number of reports on experimental results are published, describing effects that are partially due to erroneous designings.

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

A research worker deals with questions. In planning and carrying out an experiment, he wants to obtain data upon which answers to his questions can be given. Then, he uses induction and this means that he discovers a rule or, merely presents statements, as answers to them. But surely, a rule or a 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. Usually, in practical situations, such a sample cannot be drawn. Drawing a random sample implies that it can be drawn from an existing population. If the population is an abstraction, no random sample can be drawn from it. Therefore, the only possibility left for the research worker, is to regard the sample as a random one, being the representative of

the population about which inferences are being drawn. This is, in fact, the population with which research workers in other fields of research most often have to be satisfied. But neither in experimental nor in other fields of research does this mean that the research worker has to be content with any sample.

In this treatise we shall throughout regard the experimental material as random in the sense that it consists of a number of replications, which are capable of being interpreted 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 the experimental material as non-random, are bound to encounter serious difficulties in their interpretation of the results of the experiment.

Turning next to the second point, it seems evident that the most commom 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" (6.ed., 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 up by most writers of papers and text-books dealing with experimental design and statistical analysis. But, of course, such extreme and unrealistic points of view are not shared by all. In some treatises the

problems concerning the estimation of treatment effects and differences in such effects are pointed out as just as important as those of testing null hypotheses. It may also be demonstrated that Fisher's point of view is not shared by independent research workers.

The function of an experiment is the production of data that can be used in order to find the answers to 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 questionsare asked in advance of the designing and the carrying out of the experiment. In order to answer the questions it is necessary to test statistical hypotheses and/or to estimate treatment effects and 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, particularly methods for testing purposes. But on the whole, it can hardly be maintained that the situation is quite satisfactory, i.e. satisfactory in the sense of meeting the requirements of the research workers.

Heterogeneity of the experimental material seems now to be commonly accepted. It has been known and discussed at considerable length by several writers, and it was 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 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, are disregarding the interactions between experimental factors and the heterogeneity factors. However, to proceed as if such interactions do not exist, would be to assume a too simple and unrealistic model of nature.

The model describing the null hypothesis can be written any way, provided it is capable of being tested. But, if it is unrealistic, the implication of the rejection of the null hypothesis may become very mixed. The usual models of null hypotheses presume additivity of treatment effects and the effects of the heterogeneity factors. Such models may give rise to strict mathematical treatment, but they are lacking realism. In dealing with the estimation of treatment effects and the differences between such effects, it is even more important that the model is realistic. Therefore, models that do not account for interactions between the treatments and the 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 certain description. Therefore, it is impossible to repeat a treatment two or more times, if 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. Furthermore, no two units of an experimental material are exactly alike. All kinds of experimental material are more or less heterogeneous. There are, therefore, always some variation in the effect of the same treatment among a number of experimental units. The most important factor causing this variation, is usually the 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 an 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 are identical in the populations represented by the two samples, it is necessary that the division is carried out by means of some technique of randomization. If such technique has not been used, we have no guarantee that the two samples are random representatives of the same population. Consequently, if a treatment T_1 is

applied to the units in the first sample and another treatment T_2 is applied to the units in the second sample, we have no guarantee that a comparison of the effects of the two treatments will turn out to be unbiassed. A 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 30 years, research workers still try to get around it, claiming that other ways of dividing the material lead to more precise comparisons and forgetting the bias. In the last section of this treatise we are returning to a particular aspect of the principle of randomization. Until then, we shall assume that the principle has consistently been applied.

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

1. The treatments are quantities, and the leading 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 between treatments chosen in advance.

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

In answering such questions, it is obviously important that the experimental material is such that the answers can be applied in a population of reasonable width. It is evident that the material can be chosen in such a way that small and unimportant differences 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 an 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 if the results are intended to be used for some practical purpose or, if the purpose is to supplement the insight and knowledge in some field. An experimental material which serves the latter purpose, might be largely unsatisfactory for the first. There are also possibilities for describing the population in which the inferences are intended to be applied, even if the description might turn out to be vague. Such a description is a description of the experimental material and the external circumstances under which the experiment has been carried out.

In the different treatises of the methodology of today there usually is a cry for efficiency. But, obviously, choosing a design that is more efficient than another, practically always implies a reduction of the width of the population and a reduction of the generality of the inferences. The consequence is that the same difference obtained with the more efficient design, does not

usually mean the same as if a less efficient design had been used. Therefore, the common and general recommendation to the effect that the most efficient design ought to be used, is liable to objection.

3. Complete Randomization.

Suppose that the 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 sample for treatment T_2, and the treatments are allocated 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 : an experimental material consisting of nk units, is divided randomly into k samples, and the k treatments are randomly allocated 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 outcome of the experiment. In the present case, with k treatments T_j (j=1,2,..k) and n experimental units for each treatment, the model is:

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

In this model x_{ji}* are the observations, μ is a general level,

*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.

and a_j are effects of the treatments. Without loss of generality we can let $\Sigma a_j = 0$ because, if $\Sigma a_j \neq 0$, 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 assumption would imply that all effects of the treatments are included in a_j, and this would be a too simple idea concerning 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, $Var_j(e)$. If the necessary caution is exercised during the planning and the administration of the experiment, the e's can be regarded as being stochastically independent both within and between the treatments, and $Var_j(e)$ 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

$$(3.2) \qquad \bar{x}_j = \mu + a_j + \bar{e}_j$$

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, e.g. the difference $(a_p - a_q)$. It will be seen that

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

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

$$\mathrm{Var}(\bar{x}_p - \bar{x}_q) = [\mathrm{Var}_p(e) + \mathrm{Var}_q(e)]/n$$

Therefore, except if $\mathrm{Var}_j(e)$ is a constant, the precision of the estimator of a contrast is not the same for all contrasts. Thus, the common practice to use the same error mean square for the computation of the confidence limits of all contrasts, should not be recommended. The research worker can never know that $\mathrm{Var}_j(e)$ 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 $V_j = \frac{1}{n-1}\Sigma(x_{ji} - \bar{x}_j)^2$ 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 research 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

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

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 6-7, this implies that, if we compute the confidence limits of the two contrasts in the described way, 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 is used for each contrast.

In the methodology as it is usually presented, much emphasize 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 $Var_j(e)$ 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 the 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 actual questions.

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 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.

In this case the replications must be regarded as a random sample. Thus, the population is the one 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 the 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. Then, he can divide the whole area into a number of blocks of the size he wants to use, and from this existing population of blocks he can draw at random a sample of blocks. After having drawn this sample, he might find that the

blocks, belonging to the sample, are scattered over the whole area of the farm. He may therefore find that this sample is too troublesome to use in practice, and for that reason choose one of the samples having the practical advantage that the blocks are lying side by side. It is evident that usually this latter sample represents, in the sense of a random sample, an abstract population of less width than the one that is 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 hesitatingly, it has been forwarded by several authors. However, it is a fact - in our opinion a regrettable one - that this way of thinking has not been found worthy of being 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,..k, i=1,2,..n. Then, the general model for the experimental data is

$$(4.1) \qquad x_{ji} = \mu + a_j + z_i + u_{ji} + e_{ji}$$

In this model μ and a_j are parameters, 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 from the different treatments implying e.g. that Var(e) is not the same for all treatments.

The variables z and u are both effects of the inter block heterogeneity factors : z 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. Of course, some of the u's might be 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, and 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. that 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 we for each replication use the difference $d_{pqi} = x_{pi} - x_{qi}$, it will be found that $\bar{d}_{pq} = \bar{x}_p - \bar{x}_q$ and the variance is estimated by V_{pq}/n, where

$$V_{pq} = \frac{1}{n-1} \Sigma (d_{pqi} - \bar{d}_{pq})^2$$

Owing to the robustness of Student's t the research worker can be

confident that the probability of the interval

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

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 we refer.

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 right, 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, it is so also if complete randomization has been used. However, if the number of experimental units for each treatment is the same as in case a randomized block design had

been used, the number of degrees of freedom is greater for the first than it is for the latter plan, i.e. 2(n-1) for the first and (n-1) for the latter. For small n this difference means 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 to 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 result in removing a very large part of the heterogeneity in order that the difference in 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 are to such an extent deviating from the rest of the observations that it is reasonable to doubt if they are correctly recorded. Such results may happen through failure to record, or to gross errors.

In order to restore the orthogonality of the observations, techniques known as missing plot techniques have been invented, presuming 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 research worker cannot know that it is realistic, we think that these techniques should not be recommended. 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 the research worker wants to estimate the contrast (a_p-a_q), he should be content with those observations that he has obtained and accepted.

If the research worker is interested in carrying out an analysis of variance and an F test, it might not do any damage if a few observations are replaced by means of a missing plot technique. But, not even then, the use of such a technique is necessary as there always is some part of the observations which is orthogonal. For this part an analysis of variance can be carried out and, if it matters much, the observations for the other treatments can be linked to the orthogonal part 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 to be evident that the use of a missing plot technique to any large extent might completely 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 cases the observation is to such an extent different from what should be expected, that there can be no doubt that a gross error in the recording has been made. In such a case it is reasonable to treat the observation as a missing datum. However, there are cases in which the research worker may be in doubt concerning the reliability of the record. Then, it may be very difficult to say what to do about it. The most unsatisfactory way of dealing with

the matter in such a case, is to use a missing plot technique. An apparent faultily recorded observation might be due to interaction between the treatment and the heterogeneity factors, and the use of a technique which is invented 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 applied mathematics. However, mathematics has played and still playsan important role in the development of statistics and research method. It 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 implies that usually the result obtained by mathematical deduction, if it holds any interest whatever, is merely a 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 the use of it is limited to cases satisfying the premises or, if it can safely be applied in a wider field. In general, the premises that are used, are too limited in scope to justify the 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, (15), for which a rigorous proof was given by R.A. Fisher (13). An important premise for the mathematical deduction was that the observed random variable is normally distributed. There are several grounds for doubling the realism of this premise. It is hardly possible that any random variable exists, which is so distributed. Certainly, a large number of actual random variables are found, the distributions of

which closely resemble the normal form, but there also are 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 test are satisfactory, and the t distribution is therefore now commonly accepted as a tool of research within a very wide field.

In the development of a statistical method there usually are two elements : mathematical deduction on 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 and regarded as a body, consists partly of a large bulk of techniques that are never tested satisfactorily, if at all. This may be the result because most people find mathematical deductions more interesting and entertaining than they find the very tedious work involved in the testing of techniques. With the development of the electronic computers the testing of techniques is much simplified, so that research workers may 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 it has been possible. But we have not had the facilities to use the electronic computer to the extent we would have wanted to. Therefore, results from new 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 upon 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 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 contradict 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 paramenters, are not stochastically independent. This fact raises the problem of how confidence limits of the contrasts ought to be computed. Several methods have been suggested. We refer to the summary given by Federer (8), to Mood and Graybill (24), to Miller (23), and to the literature cited in these treatises.

The solution has been sought in the experimentwise confidence coefficient, which is the confidence probability of the confidence intervals of all possible contrasts simultaneously. Mood and Graybill (24 ,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 .95." These attempts to find the 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. We think it is easy to see that this limit is (k-1), where k is the number of treatments.

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

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

for which $\Sigma A_{jp} = 0$. If a set of (k-1) contrasts is chosen in such a way, so that none of the contrasts can be derived from the other ones, all other contrasts are linear functions of sub-sets or the whole set of the chosen ones. This implies that the estimates of C_p for $p \geq k$ can be derived from the estimates of C_p for $p < k$. The confidence limits of C_p for $p \geq k$ 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.

Suppose that there are $m \leq k-1$ contrasts to be estimated, and the confidence limits of these contrasts are being computed. Then, the use of the experimentwise confidence techniques implies that the limits ought to be computed in such a way that the probability of all intervals covering the contrasts is equal to $(1-\alpha)$, e.g. 0.95. This means that the confidence probability of the confidence intervals simultaneously covering the contrasts is chosen independent of the number of contrasts.

We are not able to see the justification of this principle. In our opinion the limits of the intervals should be computed in such a way that the confidence probability of the intervals simultaneously covering the contrasts is equal to $(1-\alpha)^m$. This implies that the intervals computed by means of the observations obtained

in the same experiment, even if there are correlations among the estimators, will obey the same probability rule as the intervals obtained from independent experiments. The technique for the computations of such confidence limits is treated in the next section to which we refer.

If we are dealing with tests of significance, we are also faced with the problem of testing m null hypotheses in cases in which correlations are found between the different test variables. Then, in the same way, we should use such points of significance as will make the probability equal to α^m for simultaneous false rejections of all null hypotheses.

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, c.p. the preceeding section.

The usual methods for the 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 confidence probability of the confidence interval. He should therefore use the methods described in sections 3 and 4. Then, choosing the value of α in advance (e.g. $\alpha = 0.05$) and using these methods, the research worker 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 wants 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 in order

that the probability of the k-1 confidence intervals covering simultaneously the contrasts, is equal to $(1-\alpha)^{k-1}$, c.p. section 6. It will now be shown that, in spite of the correlations and to the extent our data can be relied upon, the methods described 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 $d_2 = \bar{x}_2-\bar{x}_3$, d_1 and d_2 being defined in section 4. Let V_1 and V_2 be the two relevant mean squares (c.p. 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 multiple distribution is

$$(7.1) \qquad F(t_1,t_2,V_1,V_2,r) = Q(V_1V_2)^{\frac{1}{2}(n-2)}(1-r^2)^{\frac{1}{2}(n-4)}\exp.\frac{-M}{2(1-\rho^2)}$$

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

$$M = \frac{t_1+n-1}{\sigma_1^2}V_1 + \frac{t_2+n-1}{\sigma_2^2}V_2 - 2\rho\left\{t_1t_2 + (n-1)r\right\}\frac{\sqrt{V_1V_2}}{\sigma_1\sigma_2}$$

The probability of the numerical values of t_1 and t_2 being simultaneously less than t_α, where t_α is the point of significance 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 : $-t_\alpha \leqq t \leqq +t_\alpha$, $0 \leqq V \leqq \ldots$ and $-1 \leqq r \leqq +1$.

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 $\alpha = 0.05$, $\sigma_1 = \sigma_2 = 1$ and some chosen values of n and ρ. The results for $A^{\frac{1}{2}}$ are shown in Table 7.1. It will be seen that the values are but slightly larger than $1-\alpha = 0.95$, indicating that the effect of n and ρ on the confidence probability is too small to be of practical significance.

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 treatment mean squares. Then, assuming that the observed random variable is normally distributed, the multiple distribution $F(t_1, t_2, V_1, V_2, V_3)$ can be derived. Then, let

$$A = \int \cdots \int F \cdot \, dt_1 dt_2 dV_1 dV_2 dV_3$$

the integration intervals being $-t_\alpha \leq t \leq t_\alpha$ and $o \leq V \leq \infty$, where t_α is the point of significance of Student's t for $2(n-1)$ degrees of freedom. Numerical computations of this integral have been carried out for $\sigma_1 = \sigma_2 = \sigma_3 = 1$, $\alpha = 0.05$, and for three chosen values of n. The results for $A^{\frac{1}{2}}$ are shown in Table 7.2. It will be seen that in this case also the values are but slightly larger than $1-\alpha = 0.95$.

Table 7.2.

n	2(n-1)	$A^{\frac{1}{2}}$
3	4	0.953
5	8	0.953
10	18	0.954

The implication of these results (Table 7.1 and 7.2) is: having chosen the value of α and computed the confidence limits of the two contrasts in the described way, i.e. by (4.3) and (3.4), the research worker can be satisfied that the confidence 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 have prevented the extension in these directions. As a substitute we have carried out tests by means of con-

structed examples.

Three examples of randomized block experiments were constructed by means of Wold's table of normal deviates, Wold (35). 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,..n=5, h=2,3,..(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), and (30)

for treatments T_1, T_2,...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 (c.p. section 6):

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

and the expected number of such intervals will be N.f(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.

Table 7.3. n = 5 blocks, k treatments, $\alpha = 0.05$.

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_o/N		0.855		0.610		0.760
0.95^{k-1}		0.857		0.630		0.774
$(n_o/N)^{1/k-1}$		0.949		0.947		0.947
$1-\bar{r}/k-1$		0.949		0.946		0.947

Let $1-\alpha'$ be the confidence probability of the confidence interval of a single contrast regarded alone. Then, if the before mentioned 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_o/N. Thus, the estimator of $1-\alpha'$ would be $(n_o/N)^{1/k-1}$. 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 (c.p. Table 7.3), also $1-\bar{r}/k-1$, where $\bar{r}$ is the arithmetic mean of r, is 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. We have therefore 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. 0.95. That this is so in other cases as well, is shown by the following examples.

In examples nos. 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 (31), column 8, which are sampled from the two-sided exponential. 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.

This part of the investigation was accomplished several years ago, at a time when we had no access to the use of an

electronic computer. Therefore, we were bound to use existing tables of random values, and small numbers (N) of samples. In our last six examples most of the work has been carried out on an electronic computer.

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

$$f(z) = R\, z^a\, (10-z)^b \qquad 0 \leqq z \leqq 10$$

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

" 7 " 10: $a=2$, $b=4$, $E(z) = 3.75$

" 8 " 11: $a=0$, $b=2$, $E(z) = 2.5$

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

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

The values of β_j were for $j = 1,2,..10$:

(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

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

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

(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

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

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. The confidence limits were computed by (3.4) and (4.3) and $\alpha = 0.05$. It will be seen that for all contrasts and examples the percentage number is very close to 95%. Since the departure of the distributions in these examples from the normal is considerable, and the variances are changed to a large extent among the treatments, the results are new verifications of the robustness of Student's t distribution.

Table. 7.4. Percentage number of confidence intervals which cover the contrast. $\alpha = 0.05$.

j	Example no.					
	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
Total	94.24	94.25	93.45	94.75	94.67	94.89

Among the k-1 = 9 estimators in examples 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 no. 6, 7 and 8 and k = 5.

Let n_o 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 no. 4 to 11 the values of n_o/N, $(1-\alpha)^{k-1} = 0.95^{k-1}$, $(n_o/N)^{1/k-1}$ and $1-\bar{r}/k-1$. It will be found that some of the frequencies (n_o) differ significantly from $N(1-\alpha)^{k-1}$, which is to be expected. Nevertheless, the values of both $(n_o/N)^{1/k-1}$ and $1-\bar{r}/k-1$ are very close to $1-\alpha = 0.95$ for all examples.

Table 7.5. $\alpha = 0.05$.

Example no.	Design	k	n	N	n_o/N	0.95^{k-1}	$(n_o/N)^{1/k-1}$	$1-\bar{r}/k-1$
4	Compl.Rand	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	Rand.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	0.945

Comments on these results are not necessary. It might be well to remember, however, that in practice there hardly exists a case 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. $\alpha = 0.05$, he should remember that the confidence probability of the resulting interval is hardly ever exactly equal to $1-\alpha = 0.95$. 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 mean squares, do not materially affect the confidence probability. Of course, a pertinent question is whether or not the included examples 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 on the results. We have tried to cover as much ground as it has been possible for us to do. 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,..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 (10) introduced the two mean squares

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

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

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

$$E(V_T) = \frac{1}{k} \Sigma Var_j(e) + \frac{n}{k-1} \Sigma a_j^2$$

and

$$E(V_R) = \frac{1}{k} \Sigma Var_j(e)$$

where $Var_j(e)$ is the mean square 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 problem still 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_o), 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_o 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 not so much to the form of the distribution of e. We confine ourselves to referring to Horsnell (19), the summaries given by Cochran and Cox (5) and Scheffé (32), 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 is bound to place reliance on the results of the different investigations, which indicate that the effect is not important. This is substanciated by the results of some new investigations to which we are

returning. Even so the F test should be used with some amount of reserve also in the present case.

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

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

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

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

The assumptions underlying the F test for this case is the simplified model, c.p. model (4.1),

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

in which 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 distribution.

As a side issue it must be pointed out that the two mean square ratios are not independent. Therefore, to enter the F table

with the two ratios simultaneously cannot be recommended. In practice 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 $Var_j(u)$ and $Var_j(e)$, the formulae are

$$E(V_T) = Var(e)+Var(u)-\frac{2}{k(k-1)}\Sigma\Sigma Covar(u_p,u_q)+\frac{n}{k-1}\Sigma a_j^2$$

where $p \neq q$, and

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

Therefore, also in this case $V_T \gg V_{TR}$ 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, the F test is merely an approximation. That this is so, has been recognized by several statisticians. But as far as we have been able to make out, sufficient information on the degree of approximation is wanting.

Lacking the necessary facilities, we have not been able to use large numbers 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 5 F-values less than 1/4.56 = 0.22 and the same number larger than 2.71. The number of F-values actually found in the different classes are

$F \leq 0.22$	16
$0.22 \leq F \leq 2.71$	68
$F \geq 2.71$	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 is totally lacking realism. The unrealistic part of these experiments is that intra block interactions are not included. It is likely that the influence of the latter interactions is to the effect of bringing the distribution of F into better agreement with the standard distribution of the normal theory. The results from some small experiments that have been carried out with normal deviates, seem to substanciate this belief.

We have also used examples 6 and 9 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 $F \geq F_\alpha$.

Table 8.1

Example no	α	N	r	r/N
6	0.1	375	43	0.115
	0.05	"	20	0.053
	0.01	"	4	0.011
9	0.01	300	49	0.163
	0.05	"	31	0.103
	0.01	"	11	0.037

It will be seen that in example 6 the relative frequencies (r/N) are approximately equal to the expected ones 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 experi-

ments 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 $a_j = 0$, ought to be regarded with considerable lack of confidence. But of course, the research worker can use the F test, choosing a lower level of significance than the one he would have used if he regarded the test as being fully reliable, e.g. 1 per cent instead of 5 per cent level of significance.

To be in doubt with regard to the reliability of the F test does not imply, however, to be in doubt regarding 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 use the F test.

9. The F test in Cases in which a Number of Mean Square Ratios are computed by Means of the same Residual Mean Square.

In some cases the research worker wants to test a number of null hypotheses by means of the F test and is bound 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_1V_1/σ^2, v_2V_2/σ^2 and v_oV_o/σ^2 are stochastically independent χ^2 with v_1, v_2 and v_o degrees of freedom, and let $F_1 = V_1/V_o$ and $F_2 = V_2/V_o$. Then, it can be shown that the regression of F_2 on F_1 is linear and that the coefficient of correlation is

$$\rho = \sqrt{\frac{v_1v_2}{(v_o+v_1-2)(v_o+v_2-2)}} \quad .$$

It will be seen that if v_o is large as compared to v_1 and v_2, the correlation is trivial and cannot make the F test invalid.

However, the effect of the correlation is better measured by means of the conditional probability $P(F_2 \geq F_\alpha | F_1 \geq F_\alpha)$, where F_α are the tabulated points of significance corresponding to the respective numbers of degrees of freedom: v_2 and v_o for F_2 and v_1 and v_o for F_1. Under the stated assumption, this probability can be computed. In Table 9.1 the values of P are shown for $\alpha = 0.05$, $v_1 = v_2 = 1$ and some values of v_o.

Table 9,1.

v_o	P
2	0.380
4	0.217
10	0.111
60	0.059
200	0.053

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

The usual way of dealing with this problem seems to be to ignore it. This attitude is rather surprising, since simultaneous tests of null hypotheses in such circumstances occur regularly in both experimental and non-experimental research work. It is also surprising, since solutions of the problem have been forwarded. 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 (17), Finney (9) and Nair (25), have resulted in the generalization due to Hartley (18). If there are m null hypotheses, Hartley has suggested the use of $f_{\alpha/m}(v_i, v_o)$, i=1,2,..m, as the points of significance. 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, c.p. the next section.

A different solution was suggested by the present author (28). For the simultaneous testing of m null hypotheses (H_{oi}, i=1,2,..m)

it was suggested that all H_{oi} should be rejected only if all $F_i \geq c_i$, where

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

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 $f_\alpha(v_i, v_o)$ simultaneously. Therefore, since the same technique is used for all F ratios, rejection of H_{oi} if $F_i \geq c_i$ 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 c_i with m substituted by m-1. If then, all F_i (number m-1) are larger than the new c_i, the corresponding H_{oi} 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 new c_i (number m-2) with m substituted 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 $F_\alpha(v_i, v_o)$, 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 necessary to make a start with all F ratios. On the first step only those ratios that are $\geq F_{\alpha}(v_i, v_o)$ should be included. Those ratios which are $< F_{\alpha}(v_i, v_o)$ can be judged not significant at once and removed.

The assumptions underlying this testing method 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. It is necessary that the research worker, using this method, does not forget that usually the test is too sensitive.

10. The Regression Method.

It will be assumed now that the treatments are quantities, $x_{1j}(j=1,2,..k)$, and that the purpose of the experiment is to produce the data upon which a response function can be estimated. The first question turning up is then: response on what? Since an exact repetition of a treatment is never possible, this is an appropriate question. Following Berkson (3) we can write $x_{1j} = h_j + v_j$ and assume that v_j are random errors for which it can be assumed that $E(v) = 0$ for each j. There are, therefore, two response functions. If $\bar{x}_{oj}$ are the means of the observed random variable, the response variable, the two functions are

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

As yet no recommendable method seems to have been found by means of which the latter function can be estimated except, perhaps, if the function is linear. We are therefore dealing with the first function only.

Since the formula of $f(x_1)$ is hardly ever known, the research worker is bound to assume that it can be substituted by the Taylor expansion, and the practical problem is the very common one: to estimate the coefficient $(\beta_o, \beta_1, \beta_2, \ldots)$ in the equation

$$\bar{x}_{oj} = \beta_o + \beta_1 x_{1j} + \beta_2 x_{1j}^2 + \ldots + e_j \tag{10.1}$$

It can be assumed that $E(e) = 0$ for each j and that e is stochastically independent of x_1, $x_1^2 \ldots$. But we cannot assume that Var(e)

is the same for all j.

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 invention of the electronic computers has simplified matters, as it is now possible, without too heavy cost, to include a rather large number of terms. Of course, in the final, estimated function the maximum number of terms is k (the constant term included), and the terms need not necessarily be a subset of the set r = 1,2,..(k-1 However, in practice the research worker has to compromise to avoid being involved in too heavy and expensive computations. If his experience from former investigations does not indicate that different terms ought to be used, it is, perhaps, sound practice to include initially the terms for r = 1,2,3 and 4. Having included these terms, the research worker will be able to decide, both if all these terms should be included in the final estimated response function and if it is adviceable to try to include additional terms.

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

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 is a number of sets of such functions. One of the sets is

$$u_1 = y_2$$

$$u_2 = y_1 - b_{12}y_2$$

$$u_3 = y_4 - b_{42.1}y_2 - b_{41.2}y_1$$

$$u_4 = y_3 - b_{34.12}y_4 - b_{32.14}y_2 - b_{31.24}y_1$$

where the coefficients (b) are least squares regression coefficients. However, this is one of the possible sets of such 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 square) due to the different u's are as presented in Table 10.1, where the R's are correlation coefficients (simple or multiple).

Table 10.1.

	Reduction	Degrees of freedom	Mean Square
u_1	$R_{0.2}^2\, n\Sigma(\bar{x}_{oj}-\bar{x}_o)^2$	1	V_1
u_2	$\{R_{o.12}^2-R_{o.2}^2\}n\Sigma(\bar{x}_{oj}-\bar{x}_o)^2$	1	V_2
u_3	$\{R_{o.124}^2-R_{o.12}^2\}n\Sigma(\bar{x}_{oj}-\bar{x}_o)^2$	1	V_3
u_4	$\{R_{o.1234}^2-R_{o.124}^2\}\, n\Sigma(\bar{x}_{oj}-\bar{x}_o)^2$	1	V_4
Residual	$\{1-R_{o.1234}^2\}n\Sigma(\bar{x}_{oj}-\bar{x}_o)^2$	k-5	V_5
Total	$n\Sigma(\bar{x}_{oj}-\bar{x}_o)^2$	k-1	

In model (10.1) x_1, x_1^2 ... can be substituted by u_1, u_2... and the result is

$$(10.2) \qquad \bar{x}_{oj} = \lambda_o + \lambda_1 u_{1j} + \lambda_2 u_{2j} + \ldots + e_j$$

Assuming 1) that e is a random variable, independent of u_1, u_2... 2) that Var(e) is the same for all j, and 3) that e is normally distributed, it can be shown that $V_r/\mathrm{Var}(e)$ $(r = 1,2,3\ldots)$ is a χ^2 if $\lambda_r = 0$. Therefore, if the assumptions are fulfilled, the null hypotheses $\lambda_r = 0$ can be tested by means of the mean square ratios $F_1 = V_1/V_R$, $F_2 = V_2/V_R$ ..., 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

$$V_R = \frac{1}{k(n-1)} \Sigma\Sigma(x_{oji} - \bar{x}_{oj})^2$$

and if randomized blocks have been used,

$$V_R = V_{TR} = \frac{1}{(k-1)(n-1)} \Sigma\Sigma(x_{oji} - \bar{x}_{oi} - \bar{x}_{oj} + \bar{x}_o)^2$$

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 we refer.

In our 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 declaired 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 being significant. But such 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 result should be taken as indicating that probably at least one of the terms $\lambda_r u_r$ (r=5,6.. 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.

Of course, the greater part of the residual reduction might be due to one of the variables u_5, u_6, ... u_{k-1}. For this reason the residual ought to be used with one degree of freedom, as this will imply that we are using a more efficient test. However, significance ~~does~~ merely indicates that a more satisfactory description of the response function might be obtained if at least one of the terms r=5,6 ...k-1 is included. It does not imply that it will be found that this is so. In the experience of the author such outcome of the testing will happen very rarely. The reason is, of course, that response functions

are not usually so complicated that a linear function of $x_1 \ldots x_1^4$ does not, when estimated, give a sufficiently accurate description of them.

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 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. Then, it is a very 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 $F_\alpha(1, v_R - 5)$, c.p. 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 described way. 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_o is a normal random variable , $E(x_o) = 0$ and $\sigma = 1$

x_1 is a normal random variable , $E(x_1) = 0$ and $\sigma = 1$

e is a normal random variable , $E(e) = 0$ and $\sigma = 1$

$x_2 = \beta x_1 + e$

The chosen values of β were $\beta = 1/3$ in example 1 and $\beta = 3/4$ in example 2. Thus, the coefficient of correlation is $\rho_{12} = 0.32$ in example 1 and $\rho_{12} = 0.6$ in example 2.

Suppose now, that the area covering both F ratios $\geqq 0$ is divided into three parts : A being the part for which both F ratios are $\leqq F_\alpha(1,v_R)$, C being the part for which both F ratios are $\geqq F_\alpha(1,v_R/2)$, 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 by the binomial terms $(1-\alpha)^2$, $2\alpha(1-\alpha)$ and α^2. Choosing $\alpha = 0.05$, the values of these terms (P) are as shown in Table 10.2.

The number of experiments that were constructed and analysed, is 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).

Table 10.2. $\alpha = 0.05$.

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

In these examples the variable 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 for the reason that 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). Then, it is 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 of the u-variables. These computations are carried out by standard technique described in a number of text-books and we do not, therefore, go into the matter here. The last step consists of substituting the u-variables in the response function by x-variables.

The described technique can hardly be recommended to research workers lacking the facilities to use an electronic computer. For these research workers this technique would be too time consuming. As an alternative approach it can be recommended to choose initially a particular set of u-variables, for instance

$$u_1 = y_1$$
$$u_2 = y_2 - b_{21}y_1$$
$$u_3 = y_3 - b_{32.1}y_2 - b_{31.2}y_1$$
$$\cdots\cdots\cdots\cdots$$

Choosing initially such a set, the u-variables can be included one at the time, and for each new u-variable the residual reduction in Table 10.1 can be computed. In this way the research worker can decide at each step if it seems worth while to continue adding new variables. Thus, the work can be reduced to such a minimum that the computations are easily carried out by means of a desk calculator. Using this technique, the research worker will be lacking the opportunity of trying the different combinations of the y-variables, and the final estimated response function cannot be claimed to be the "best" one 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 happen 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. None of these assumptions can be regarded as being realistic. We have discussed this point in section 8 and shall not repeat our arguments. 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. $\alpha = 0.05$, but usually an inflated one.

In a case in which the treatments are quantities, the research worker may want to estimate the response function and, at the same time, he may want 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 want 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 the response is a maximum. As 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 be able to estimate such values of x_1 it is necessary that the space of the selected values covers them. This means that the research worker must be in possession of advance information with regard to these values and use such information in the planning of the experiment.

An important question concerns the choice of the values of

x_1. For the computations it would be advantageous to choose equally spaced values or, equally spaced values of transforms such as log x_1 and $(x_1)^{\frac{1}{2}}$. This will enable the research worker to use the orthogonal polynomials introduced by Fisher (12), tabulated by Fisher and Yates (14) and by Pearson and Hartley (29). 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 means the use of a particular set of u-variables, implying that the research worker renounces from trying out the different sets of these variables.

11. The Problem of the Gaps and the Grouping of the Treatments.

In planning the experiment it is not always possible for the research worker to decide on particular contrasts that he wants 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, the gaps) between two and two neighbour means. Let r be the rank, r = 1,2,..k, where k is the number of treatments. Then, $u_r = \bar{x}_{r+1} - \bar{x}_r$ (r = 1,2,...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. that $f(\bar{x}_j) = 1/A$ ($0 \leq \bar{x}_j \leq A$), it can be shown that the distribution of u_r is

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

and that $E(u_r) = A/k+1$, 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 not able 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 $F = V_T/V_R$ (c.p. section 8) is significant on some chosen level of significanse, e.g. the 5 per cent level. Then, a strongly

* C.p. Tukey (33).

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. Then, the mean range can be used to advantage. The mean range $E(W_k)$ of the normal distribution has been tabulated by Pearson and Hartley (29) 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 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 wanted. 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 pick out among the k treatments those that, in a certain sense, are superior. In such cases the research 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, on 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 research worker has to be content with the group found on the first step. If, however, the mean square ratio is significant on 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 on a third step, a fourth step a.s.o., and will be terminated as soon as a non-significant mean square ratio is found.

It is necessary, perhaps, 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 the computation of the conditional range, merely implies the pointing out of the consequence of the null hypothesis having been rejected by means 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 (c.p. section 8) it will add to 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 information regarding the distribution of the range and the mean range has been given by Kendall and Stuart (21), to which we refer. Some new information could have been obtained from the examples used in section 7, if lack of funds had not prevented the 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 $0.5976\ \beta_j$ - 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 cheaply for selected,

mathematically simple distribution. It is well known for the rectangular distribution. For the exponential $f(x) = e^{-x}$ $(x \geq 0)$ it can be shown that the mean range is equal to

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

where k is the size of the sample. Some values are shown in Table 11.1.

Table 11.1.

k	Example no. 6		$E(W_k)$	
	N	$\hat{W}_k$	Normal	$f(x)=e^{-x}$
5	150	2.353	2.326	2.083
10	148	3.081	3.078	2.829
20	72	3.570	3.735	3.548
375	10	5.833	5.896	6.503

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

Federer (8), p. 122, presents the results of a randomized block experiment for the comparison of k = 7 varieties in n = 5 blocks. We choose this example because prior informations concerning the grouping of the varieties are 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 $F = V_T/T_R = 3.4$ 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 that $(V_R/n)^{\frac{1}{2}} = 1.179$, and since $E(W_7)=2.70436$ for the normal distribution, we find $V_7 = 3.19$ and that

$$\bar{x}_{min} + V_7 = 6.34 + 3.19 = 9.53$$

$$\bar{x}_{max} - V_7 = 13.32 - 3.19 = 10.13$$

The seven varieties are therefore divided into three groups:

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

Then, treating with group C only, it is found that $V_T = 6.49$ (with 4 degrees of freedom) and $V_R = 6.23$ (with 16 degrees of freedom), implying that no division of the group should be attempted. Hence, the grouping on 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 the statistical analysis, is directly observed, e.g. yield in an agricultural field plot experiment. 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 $100\,y$.

The statistical problem concerns the method of investigating the effect on the probability $P(A) = p$ of the different treatments and the comparison between the treatments. Both m and p might be dependent on the heterogeneity factors, implying that both of them must be regarded as being random variables. In any case, the research worker can never assume that m and p are independent on the heterogeneity factors. Therefore, it must also be assumed that m and p are correlated.

In this case we are therefore concerned with a situation in which we have to deal with three random variables (x, m and p), and only two of them (x and m) can be observed. The distribution is

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

Writing $w = 1/m$, it can be shown that

(12.2) $$E(y) = E(p) = p_o$$

(12.3) $$Var(y) = p_o(1-p_o)E(w) + [1-E(w)]\,Var(p) + Covar(p,w) - Covar(p^2,w)$$

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) $$y_{ji} = \mu + a_j + z_i + u_{ji} + e_{ji}$$

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

(12.5) $$\bar{y}_j = p_{oj} + \bar{z} + \bar{u}_j + \bar{e}_j$$

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

The difficulties met with in the statistical analysis, first and foremost originate in the interaction between the treatments and the heterogeneity factors, causing differences in Var(y) and correlations between contrast estimators among the treatments. Difficulties also partly spring from excessive

skewness of the distribution of y. We cannot see, however, that the difficulties are greater and our mistrust in the suggested statistical technique ought to be more serious, than is the case with other random variables. Our conclusion is, therefore, that the methods described in sections 7-11 are applicable in these cases also.

In statistical literature it will be found that certain transformations are recommended, c.p. Bartlett (2). In the present case, examples are log y, $\sqrt{y}$ and the inverse sine function, arcsine$\sqrt{y}$.

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 insignificant.

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 on p. But if the model is non-additive, the effect may be very small.

Some years ago the present author (27) recommended the use of co-variance 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 the variance among the treatments. We have found, however, that the effect is not sufficient to counterbalance the reduction of the width of the population in 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 we refer.

13. The Idea of the Non-Random Experimental Material.

In the proceeding sections the replications have been regarded as a sample representing an abstract population. We now go on to show in more detail what the difference is between a non-random sample 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 (the blocks) be numbered i=1,2,..n, and the treatments numbered j=1,2,..k. Also, let the experimental units within any replication be numbered r=1,2,..k. Then, the value (and, the observation) of the random variable under consideration can be symbolized by $x_{(j)ri}$, meaning that it stands for the value of the random variable which would have occurred for the r'th unit in replication no. i provided the treatment T_j were 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 no. r in replication no. 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, being the effect of the heterogeneity

of the experimental material.

As we understand it, this is the null hypothesis considered by Fisher (11). A modification is suggested by Neyman (26).

Suppose that N=nk values of a variable are randomly arranged in all K ways. Then, to 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 one 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 experimental materials 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. Most often a satisfactory compatibility has been found. We refer to papers by Eden and Yates (7), Welch (34), Pitman (30), Hack (16), 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 the 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 might sometimes be recognized and included in the term e. Usually e is regarded as a random variable belonging to the population consisting of the K arrangements. It is also found that it is regarded as being a normal random variable.

The pupulation consisting of the K arrangements, is certainly a lucid construction. It seems to be a fact that the majority of statisticians and, possibly also the majority of research workers, regard it as being fully adequate, For reasons explained in section 1, we do not regard 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. By some of them it is recommended to generalize to a broader population. For instance, Kempthorne (20), 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 consisting of the K arrangement, is the one the research worker is interested in. We have therefore come to the conclusion that this population is inadequate and that, if the research methodology is founded upon this construction, the research 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 open for avoiding the difficulty, is to regard the actual replications (blocks) as a random sample of replications representing the population. This population is always an abstraction and is the one the sample represents in the sense of a random sample. This is the answer in other fields of empirical research work and it is also so when we are dealing with experimental research.

It seems also to be a fact that most statisticians accept the assumption that the effect of treatments and that of replications are additive. Some writers even think that if this assumption is not satisfied, the randomized block design cannot be used. However, it is impossible to accept such an assumption because such acceptace would imply that the research

worker knows very much about the effects of the treatments in advance. To allow for interactions merely means that an unprejudiced point of view is taken. It does not mean that it is held generally that interactions always exist. The 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 if 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 in 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 the rs combinations of two factors P_p ($p=1,2,...r$) and Q_q ($q=1,2,...s$) are included in the experiment. Then, the experiment can be regarded as an experiment with the treatments T_j ($j=1,2,..k=rs$), and the analysis can be carried out as if just one factor is involved. Of course, in these cases some particular contrasts are planned to be estimated by means of linear functions of the treatment means.

The simplest method of analysis amounts to 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 of a_j in model (3.1) and model (4.1) by

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

The term u_{ji} in model (4.1) must be substituted

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

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

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

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

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

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 statement: 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 the testing. Using for the mean squares the symbols V_p, V_q, V_{PQ}, V_{PR}, V_{QR} and V_{PQR} (R symbolising replication), we have for example

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

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

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. Writing b=0, it will be found from model (14.2) that

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

and

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

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 wants 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_{PR}	
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 it is in experiments with one factor, large values of the mean square ratios are indications 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 $F \geq F_\alpha$ 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, c.p. section 9.

Findings of significant main effects and significant interaction are certainly interesting. 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 findings do not imply that the analysis is completed. Usually the research worker wants to know more about the details.

Then, the method of analysis will be different for the different types of treatments. If the alternatives of both factors are quantitative, e.g. quantitative levels of fertilizers, a careful problem analysis prior to and included in the designing of the experiment, may very often point out 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 is described in section 7. A careful problem analysis might also show that it is unnecessary to include all rs combinations of the alternatives of the factors. This will imply a reduction of cost of the experiment, a reduction that may be used to increase precision by increasing the number of replications.

Regression analysis is an alternative technique in such cases. Suppose that the quantitative alternatives 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} $(p=1,2,..r)$ for the P factor and x_{2q} $(q=1,2,..s)$ 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 ... $(x_{1p}x_{2q})$, $(x_{1p}^2 x_{2q})$ .... 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 we refer.

It is well known, however, that in some cases the 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, c.p. section 11. In a factorial experiment this can be carried out for each alternative 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 alternative of Q. It is reasonable to expect, however, that the alternatives of P belonging to the group of superior treatments, are different for the different alternatives 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 very well be found, and would be very important and useful.

We do not think, however, that a definite methodology for the analysis of data obtained in factorial experiments, should be recommended. The diversity of the questions that are wanted answered, is too great from one case to another. The important thing is that a careful problem analysis is

carried out in advance, and that the experiment is planned and carried out in such a way that answers can be expected to the questions which such an analysis has pointed out. If such a working rule is adopted, the methods outlined in sections 7-11, will serve the purpose.

It is well known that it is a disadvantage of factorial experiment that an increasing numbers of factors, even if the number of alternatives for each factor is small, may lead to large and sometimes to a prohibitive number of treatments. The difficulties that follow from such large numbers of treatments, have been tried overcome in various ways. In a forthcoming section we shall return to the problem, and shall for the present confine ourselves with 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 other treatments. Federer (8) has listed a number of cases in which the split-plot plan ought to be used.

In an agricultural field plot experiment with two factors P_p and Q_q (p=1,2,..r, q=1,2,..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,..m). Now e_{pqi} stands for the effect of the heterogeneity factors intra main-plots and interaction between these factors and P_p. In the same way e'_{qi} stands for the effect of the heterogeneity factors inter 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

$$\bar{x}_{pq} = \mu + b_p + c_q + d_{pq} + \bar{z} + \bar{u}_p + \bar{v}_q + \bar{w}_{pq} + \bar{e}'_q + \bar{e}_{pq}$$

$$\bar{x}_p = \mu + b_p + \bar{z} + \bar{u}_p + \bar{v} + \bar{w}_p + \bar{e}' + \bar{e}_p$$

$$\bar{x}_q = \mu + c_q + \bar{z} + \bar{u} + \bar{v}_q + \bar{w}_q + \bar{e}'_q + \bar{e}_q$$

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-alternatives 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. If the method described in section 7 is adopted, confidence limits of contrasts among Q alternatives, among P alternatives, as well as contrasts among PQ alternatives can easily be computed.

15. On Methods intended to yield Estimators of increased Precision.

It is but natural that both research workers and statisticians have been concerned with the development of experimental designs which are intended to yield increased precision of the estimators. Among these designs, that of confounding and the use of concommitant random variables are perhaps the most utilized in practice. Identical twins and the like are used in some more exceptional situations. If the replications are regarded as a sample, representing an abstract population, it is easy to see, however, that most designs invented with the point in view of increasing precision, are meeting the requirement at the sacrifice 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 is, a rule or merely a statement, it is not unrestrictedly universal. For instance, it 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 width of the population. In basic research the date produced by an experiment, may be satisfactory as evidence for some rule or statement even if the heterogeneity of the experimental material is very small. However, if the research worker wants to find a rule that can be used as a guidance for practical activity, it is necessary that it is inferred from data obtained in an experiment which is planned and

carried out in such a way, that the heterogeneity of the material is dependent on all those factors which are not controlled in the practical activity. If the rule is an inference based on data from an experiment on material of less heterogeneity, it might be overshadowed by the effect of these factors. Therefore, if such a rule is used as a guidance, the chance for the activity to achieve the desired end might 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 the comparison of the effects of two feeding alternatives, and the purpose is to learn which of the two alternatives 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 an experiment of such a kind may be important if the limitations of the validity of the results are not forgotten or ignored. Other examples of this sort are discussed by Linder (22), p.13, and Cox (6), p. 25.

The use of confounding means 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. Then, some of the treatment effects are confounded with the effects of heterogeneity among the main units. Thus, the split-plot design belongs to this class. The consequence 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 the nonconfounded effects are estimated with higher precision than the other ones.

The inference is different if observations of a concommitant 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 alternatives 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 to some extent the heterogeneity by means of the 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 reduce 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 to the validity of the result obtained without the use of the observations of x_1.

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

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

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

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

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

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

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

Writing $A_j = \Sigma(x_{1ji}-\bar{x}_{1j})^2$ and $B_j = \Sigma(x_{oji}-\bar{x}_{oj})^2$ 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_o and x_1 for the treatments T_p and T_q. Hence, approximately correct confidence limits for the contrast (a_p-a_q) are $d'_{pq} \mp t_\alpha s_{d'}$

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

$$s^2_{d'} = \frac{B_p(1-r_p^2) + B_q(1-r_q^2)}{n_p+n_q-4} \left\{ \frac{n_p+n_q}{n_p \cdot n_q} + \frac{(\bar{x}_{1p}-\bar{x}_1)^2}{A_p} + \frac{(x_{1q}-x_1)^2}{A_q} \right\}$$

the number of degrees of freedom being 2(n-2).

In the formula of s_d, both $(\bar{x}_{1p}-\bar{x}_1)^2$, $(\bar{x}_{1q}-\bar{x}_1)^2$, A_p, and A_q must be treated as fixed, non-random, quantities. This represents the sources of the loss of validity of the inference. Therefore, if d' is used instead of $d = \bar{x}_{op}-\bar{x}_{oq}$ as the estimator of the contrast, and s_d, is used for the computation of the confidence limits, the inference cannot be applied in the whole population represented by the experimental units. We are now bound to deal with a sub-population characterized by A_p, A_q, $(\bar{x}_{1p}-\bar{x}_1)^2$, and $(\bar{x}_{1q}-\bar{x}_1)^2$. This loss of validity of the inference should not be ignored, as it usually is.

Among the designs that are intended to increase precision we may also include the Latin Square design. In a randomized block experiment randomization is carried out according to the principle of complete randomization within each replication. The function of randomization is to prevent a bias from damaging the estimators of the treatment contrasts.

In field plot experimentation it has been tempting to affect a partial control over the heterogeneity, not only in one direction but in two orthogonal directions. This has to lead 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 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 concerning this design and the statistical analysis deals with a population that consists of all possible k^2 squares. It does not seem possible to look at this design in a different way. Therefore, if we are concerned with an abstract population represented by a sample of replications, the Latin Square design is lacking significance, except if 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, being concerned with designs to be used in order to produce data upon which inferences with regard to an abstract population can be drawn, our conclusion is that the Latin Square design should not be recommended.

16. Experiments with large Numbers of Treatments.

In some experimental situations the number of treatments is very large as, for instance, in some 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 large that the advantage of the randomized block design over complete randomization is illusory. Other examples are factorial experiments with large numbers of factors and/or large numbers of alternatives of the single factors. However, a large number of treatments sometimes means that the number is large in comparison to the number of easily accessible experimental units.

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

In the lattice designs for the comparison of k treatments, the replications are divided into a number of main units (usually called blocks), each consisting of m experimental units. This means that the number of main plots is k/m for each replication. In a quadratic lattice $k = m^2$, so that the number of main plots per replication is equal to m and the total number of main plots equal to $b = nm$. The treatments are also divided into m groups, each consisting of m treatments, and the groups are allocated

the plots in such a way that the comparison between two treatments, belonging to the same group, can be made between plots belonging to the same main plot. The grouping of the treatments is changed from one replication to another according to the rule that each pair of treatments shall occur together in the same main plot the same number of times.

It is usually thought that such arrangements of the treatments lead to increased precision of the estimators of the contrasts, as compared to randomized block experiments without grouping. However, disadvantages have also been recognized. Federer (9) 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 thinking, research workers should always regard missing data to be likely to happen, and equal error variances are practically never realized. Therefore, we find it very 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 the number of them is very large. For instance, in a field plot experiment such activities as the planting and the sowing take considerable time. The same is the case with operations during harvesting. Therefore, it would be advantageous if the area of land that represents a replication, were divided into main plots, so that the research worker can deal with these one at the time. The split-plot plan will meet this requirement.

Suppose that the treatments are divided into a number (s) of groups and the replications are divided into the same number of main units. Then, the s groups of treatments can be allocated the main units in a random way, and the treatments belonging to a group can be allocated randomly the experimental units within the main unit. Statisticians are familiar with the use of the split-plot design in this way, c.p. Cochran and Cox (6). The reason why incomplete block designs are preferred and recommended seems to be, first and foremost, that it is thought that these designs yield comparisons of equal precision. Research workers who do not believe in equal precision of the contrast estimators, 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 the split-plot design are the following: 1) It is unnecessary that the groups are of the same size, i.e. that they cover the same number of treatments, On the contrary, it is important that the treatments are 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 have been used.

Interactions between the treatments and the heterogeneity factors render any prior judgement of the relative precision of the different contrast estimators quite impossible. 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 the 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 no. 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, allocated the experimental units in a random way. If for practical reasons, the research worker deals with the main units one at the time, a time factor is introduced. In such a case it is particularly important that a check treatment is included, so that the bias caused by the time factor can be removed.

Let the observations of some random variable (e.g. yield) be x_{pqi}, where $p=1,2,..r_q$, $q=1,2,..s$, and $i=1,2,..n$, 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 thus

$$x_{pqi} = \mu+z_i+a_{pq}+v_{qi}+w_{pqi}+e'_{qi}+e_{pqi} \quad . \tag{16.1}$$

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 the treatments and the heterogeneity factors. The terms v and w stand for the interactions between groups and replications, and between treatments within groups and replications.

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 being 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 $\Sigma\bar{a}_q = o$.

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 ranging of the treatments. Consequently, a linear function of treatment means is an unbiassed estimator of the corresponding contrast. Interactions between treatments and the heterogeneity factors imply, in this case also, 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 has been included, T_o (say) x_{pqi} can be substituted in all the procedures by $y_{pqi} = x_{pqi} - x_{oqi}$, where x_{oqi} are the observations for the check treatment. 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_{oqi} stands for the mean value of two observations.

The preceeding discussion concerns cases in which the number of treatments is large, but where there is no shortage with regard to 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 wants to use litters as replications in a feeding experiment to pigs. In such a 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 the heterogeneity among the replications, and the split-plot plan ought to be combined with complete randomization. If the research worker wants to use the split-plot plan and the randomized block design, the replications ought to 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 to which we are returning in the next section.

17. Experiments which are intended to give Results for Practical Utilisation.

It has been touched upon previously that if an experiment is carried out for the express purpose of providing a base upon which advice to practitioners can be given, it must 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 usually fully known to the research worker, and the experiment must be planned in such a way, that the effect of them can be regarded as random effects. Also, the hard fact is that the inference, if any, can only be applied in the population represented by the actual experimental material in the sense of a random sample. This abstract population might not be broad enough to cover all cases that may occur in practical activity. Therefore, the research worker being consulted in a particular case, is well advised to show such modesty as to recommend a treatment only provided the case belongs to this population. If he does not make such a reservation he may take the chance of using extrapolation of his experimental result outside the sphere covered by the experiment. However, the research worker can do much to ensure that the population is broad enough to cover the great majority of cases occurring in practice.

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 under which a particular treatment among a number of treatments is the superior one. 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 field of research recommends a certain variety of wheat to all farmers in a geographical area, he should know from the results of his experiment that it is to be expected that the use of this variety will inflate the yield for the whole area as compared to 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 width of the population. Then, first of all he must decide what kind of experimental units should be used. In agricultural experimentations a unit must be a field suited for the growing of the plant in question, and which fields are suited is something that must be decided upon. In industrial research a unit may be an industrial plant, but it is not evident that all plants should be regarded as being suited. Therefore, in practically all cases there are a number of difficult decisions to be taken in advance of the planning of the experiment,

decisions to be taken for the purpose of marking the borders of the population about which knowledge is wanted.

When these decisions have been made, it might seem obvious what to do next: take one or more random samples from the accepted cases 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 often happen that a field selected for an agricultural experiment, is planned 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 centered around the experiments that can be characterised as "local". A local experiment in agricultural field plot experimentation, is one carried out in a chosen field in one season. A local experiment in experimentation on feeding pigs, is one carried out with pigs chosen from a single stock and at a chosen farm. A local experiment is also one carried out in a single industrial plant.

It is evident that for an experimental result to be used as a base for recommendations for practical activities, 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 narrow. In any case this is true if interactions exist between the treatments and the environmental factors that are not controlled during the practical acitvity.

Among research workers in the agricultural field of research it now seems to be generally recognized that both geographic heterogeneity factors and factors, the effects of which are varying 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. Therefore, research workers in this field of research are compelled to 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 importance of the interactions between the treatments and the heterogeneity factors may be quite different for the different cases.

Consulting the literature, dealing with such experimental situations, it is most often found that the experiment is regarded as consisting of a number of repeated local experiments. In our opinion it should not be considered thus. It ought to be regarded as an experiment of its own, planned and carried out for its own specific purpose.

Keeping to the example from the agricultural field of research, the 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 probably the best one, since it can be expected that the heterogeneity factors are better covered in an experiment according to this plan than they are if the same

sample of localities are used for all seasons. But it is evident that the use of the same sample of localities in all seasons is the simpler one 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 a sample of seasons and a population of seasons are too vague. The term "season" does not imply more than a kind of classification with regard to the variation in the effects of some environmental factors.

The conclusion is that whether a new sample of localities has been taken for each season or, the same sample is used for all seasons, the research worker has to be content with a sample of replications, and the population in which the inferences can be applied is the one this sample of replication represents in the sense of a random sample. Most often the width of the population is larger if a new sample of localities is taken for each season than it is if the same sample is used in all seasons. But the difference cannot ordinarily be very important. In both cases the question is whether the sample gives a satisfactory coverage for the geographical heterogeneity factors and for that period of time for which the inferences (or, forecasts) are intended. It is possible to ascertain to some extent whether there is a satisfactory coverage for the geographical factors. But for the time factors, i.e. climatic factors, the answer to the question regarding the coverage depends on what can be said about the changes of the climate in coming years, which for the present is very little. It is evident, however, that if relatively large

climatic changes have taken place in the seasons covered by the experimental replications, the research worker can be more confident in giving advice to practitioners than he can be if the replications cover less variation in the climatic factors. The interaction between a treatment and the climatic factors may be small and 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 can feel if the interaction is greater. The same is the case with regard to the interaction between the treatments and the geographic heterogeneity factors. Therefore, both kinds of interaction should be taken into account when the research worker is dealing with the ranging and classification of the treatments.

It has been mentioned above that there are, at least in theory, two ways that can be used for the sampling of replications. We can take a new sample of localities for each season included in the experiment or, we can use the same sample of localities in all seasons and merely change the site within the localities. In the first case, if there are n_1 seasons and n_2 localities, for each season, the sample consists of $n=n_1n_2$ replications. If only the site is changed from season to season and the number of seasons and localities are n_1 and n_2, the sample still consists of $n=n_1n_2$ replications. In both cases the population is the one 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 in all seasons has the advantage over the other plan 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, it may be 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,..k) and replication no. i (i=1,2,..m) is

$$\bar{x}_{ji} = \mu + a_j + z_i + u_{ji} + \bar{e}_{ji} \tag{17.1}$$

where the different terms stand for the same effects as they do 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). Therefore, 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 $\frac{1}{k}\Sigma Var_j(e)$, 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 according to the randomized block experiment, it must be expected that $Var(u) = \frac{1}{k}\Sigma Var_j(u)$ is much inflated as compared to 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 $F \geq F_\alpha$ being also inflated. However, the treatment mean $\bar{x}_j$ is an unbiassed estimator of the treatment effect, implying that an unbiassed ranging of the treatments is obtained by means of the treatment means.

Suppose, next, that the same sample of localities is used in all seasons, the mean of the observed random variable for treatment T_j (j=1,2,..k), in season no. i (i=1,2,..n_1), and locality no. h (h=1,2,..n_2) can be written $\bar{x}_{jih}$, and the model is

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

Also in this case 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 $(a_p - a_q)$.

It can also be shown, no matter which of the two plans is used, 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.

It has briefly been touched upon that, if the problem is to group the treatments and, particularly, if the problem is to isolate a group of superior treatments, the research worker should also consider the interactions, first and foremost the interaction between the treatments and the 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 in all seasons, we may consider the function

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

and the graphs of δ_{ji} against i, one for each treatment, as a probably useful aid. Probably also

$$\lambda_j = \Sigma_i \delta_{ji}^2 / \Sigma\Sigma\delta_{ji}^2 \qquad (17.3)$$

may prove to be useful for the 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

$$\lambda'_j = \Sigma\delta_{jh}^2 / \Sigma\Sigma\delta_{jh}^2 \qquad (17.4)$$

where

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

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 interaction.

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, to a large extent reflect 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 plants and effects of climatic factors are negligible. But it is most likely that such examples are exceptions rather than the rule.

To return to another example, we were in section 16 discussing the design for an experiment for the comparison of a number of feeding alternatives to pigs 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 according to the split-plot plan. 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 certainly is heterogeneity among stocks, particularly heterogeneity due to genetic factors, as there is

heterogeneity 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 is found 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 activity can be based, it is likely that it will 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 covers merely a part of the sphere of practical activity 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 repli-

cations. If we reflect on the best way of using the resources which are at the research worker's command, the conclusion will be to the effect that a very simple design should be used for the single replication. Merely one experimental unit for each treatment per replication will do, but in practice a couple of units ought to be used in order to guard against failures.

If m units (h=1,2,..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,..k; i=1,2,..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

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

It will be seen that the effect of increasing m merely is 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 of m will not strenghten the precision of the estimator very much. On the other hand, an increase of the number (n) of replications will always affect the precision favourably.

From the preceeding discussion it will be found that, as regards the method of analysis, the situation is equivalent to the one 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 7-12 are adequate for the statistical analysis.

18. Some Supplementary Matters.

A. Every research worker who is consistently using the principle of randomization, sooner or later will come across examples where the result of the randomization may seem unacceptable. The reason for this is, that most often some trend or regularity must be assumed to exist among the experimental units. For instance, this is so in a field plot experiment where there often is some regularity in more than one direction of the quality of the units. Then, if the 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 do something about it. He might, of course, stick to the randomization principle and accept the result, knowing that also such a result must be left a place in a long run procedure. However, often the research worker has to make a decision immediately, and therefore, it is natural for him to consider rejecting the result of the randomization and rerandomize.

Regarded from a principal point of view, any tampering with the result of the randomization ought to be refuted. But, we do not think that this would be the right 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 problem of experimental design, the question is usually ignored. However, Cox (6) has discussed the question at some length, referring also to relevant

literature. In his treatice some methods of dealing with the question, are being discussed. Two of the methods are founded upon the idea of the rejection of the more extreme arrangements of the treatments. The difficulty involved in this approach, is that it will be necessary to use a dichotomy, grouping the results of the randomization into acceptable and unacceptable arrangements.

Since it is necessary to submit to the fact that our statistical tools are merely approximate, the problem to consider is what effect, e.g. on the F test, rejection of some of the arrangements might have. Probably, the effect of such a restriction of the randomization is to bring the distribution of F into better harmony with the standard distribution of the normal theory. Regarding the confidence probability of the confidence interval of a contrast, it is reasonable to think that the effect is small and is directed towards an inflation of the confidence coefficient. But these statements are based on mere guessing. A statistician who has ample access to an electronic computer, might be able to obtain satisfactory evidence by using constructed examples. Then, of course, the examples must be constructed according to realistic models, and the rejection of extreme arrangements of the treatments must be carried out on an exaggerated scale. Only as soon as results from such investigations are presented, is it possible to make up ones mind what standpoint should be taken to the practice of curtailing the random arrangements.

B. At the time when the research work on experimental designs began, the common attitude among learned statisticians was that useful informations could hardly be obtained from small samples. The explanation of some of the criticism raised against the work by Fisher and his collaborators, may be found in this attitude. To-day it is generally recognized that even very small samples may yield data upon which important conclusions can be drawn. However, it is hardly questionable that the founders of the designs of experiments went to the other extreme, partly because they were too engaged in the problems of tests of significance. If the attention is turned to the problem of the estimation of contrasts, larger samples are usually required.

For instance, suppose that an industrial leader is contemplating to replace old mass manufacturing machinery by new machinery. Then, it is not enough for him to know that it has been shown by some test of significance that e.g. the new machinery is producing at a higher rate of speed than the old one. In his economic calculations he needs some measure of the difference of speed 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 the new machinery are the treatments and 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 is not too wide, which in fact implies that the size of the experiment or, the number of replications, cannot be very small.

We believe that this is a most common situation. If the purpose of an experiment is changed from that of significant test to the estimation of contrasts, an increase of the number of replications is usually required. But, of course, in some cases the replacement of one treatment by another does not imply new economic investments and, if so, 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 is 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 one is that the presence of interactions between the treatments and the 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 confidence intervals. Therefore, the use of a common error mean square will always mean 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 the computations of the confidence limits of the contrasts, implies that the number of replications cannot be too small. If this number is too small and, consequently, the confidence intervals of the contrasts are very wide, it is difficult to see what object the experimental data are capable of achieving. Therefore,

the research worker, in planning his experiment, should always try to estimate the number of replications that will be necessary in order that a chosen minimum precision can be expected to be obtained. Obviously, this is a very difficult task to be charged with, 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 the local and the non-local experiments described in section 17. In planning an experiment it is important to know if the results are intended to be used for some practical purpose or, if the purpose is to supplement the research workers knowledge in some 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 in the capacity of experimental factors. Therefore, a local experiment as described, will not meet the requirements in either case. However, an experiment of such a kind may furnish the necessary data upon which preliminary conclusions can be drawn, conclusions that may be used as a guide for the planning of a non-local experiment. For instance the data may show that some treatments are to such an extent inferior that they can be left out in the planning of 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 can be reduced to a minimum.

There are, of course, exceptions to this appraisal of the local experiments as, for instance, in our example in which it was

assumed that an industrial leader is interested in the comparison of two kinds of machinery. In this case the outcome of an experiment may be important to the particular industrial plant in question. Therefore, the experiment can 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, consisting of such experiments as are carried out in a laboratory or, under laboratory conditions, where a number of external factors can be controlled. A fourth category consists of such experiments, discussed in section 15, as 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. Then, one of the problems for the research leader is to decide how and to what extent the different categories ought to be utilized. In view of the fact that the research funds are usually very restricted, it is important that a balance is found in order to achieve a kind of optimum. In practice to find such a balance is certainly very difficult. If agricultural field plot experimentation is considered, it seems to be a fact that research workers are inclined to spend a too great part of the research fund on local experiments. This may partly be due to the history of the development of the experimental designs. In this field of research it is rather obvious that the outcome of a local experiment can be regarded as being merely preliminary. Therefore, it also seems somewhat confusing that so much emphasize is placed on the development of the designs of such experiments.

Bibliography.

1. Baker, F.B. and Collin Jr., R.O. (1966). J. Am. Stat. Ass. Vol. 61. No. 315.

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

3. Berkson, J. (1950). J. Am. Stat. Ass. Vol. 45, No. 250.

4. Box, G.E.P. (1953). Biometrika, Vol. 40.

5. Cochran, W.G. and Cox, G.M. (1950). Experimental Designs. John Wiley & Sons, Inc. New York.

6. Cox. D.R. (1958). Planning of Experiments. John Wiley & Sons, Inc. New York.

7. Eden, T. and Yates, F. (1933). J. Agric. Sci. Vol 23.

8. Federer, W.T. (1955). Experimental Designs. The Macmillan Comp. New York.

9. Finney, D.J. (1941). Ann. Eugen., Vol. 11, part 2.

10. Fisher, R.A. (1936). Statistical Methods for Research Workers. Oliver and Boyd, London.

11. Fisher, R.A. (1951). The Design of Experiments. 6 ed. Oliver and Boyd, London.

12. Fisher, R.A. (1921). J. Agric. Sci., Vol 11.

13. Fisher, R.A. (1923). Proc. Camb. Phil. Soc. Vol. XXI.

14. Fisher, R.A. and Yates, F. (1948). Statistical Tables for Biological, Agricultural and Medical Research. Oliver and Boyd, London.

15. Gosset, W.S., ("Student"), (1908). Biometrika, Vol. 6.

16. Hack, H.R.B. (1966). Biometrika, Vol. 45,

17. Hartley, H.O. (1938). Suppl.J. Roy. Stat. Soc., Vol. V, No. 1.

18. Hartley, H.O. (1955). Communication on Pure and Applied Mathematics. Vol. VIII.

19. Horsnell, G. (1953), Biometrika, Vol. 40.

20. Kempthorne, O. (1952). The Design and Analysis of Experiments. John Wiley & Sons. Inc. New York.

21. Kendall, M.G. and Stuart, A. (1958-1966). The Advanced Theory of Statistics. Charles Griffin & Comp. Limited, London.

22. Linder, A. (1953). Planen und Auswerten von Versuchen. Verlag Birkhäuser, Basel/Stuttgart.

23. Miller, R.G. Jr. (1966). Simultaneous Statistical Inference. McGraw-Hill Book Company, New York.

24. Mood. A.M. and Graybill, F.A. (1963). Introduction to the Theory of Statistics. McGraw-Hill Book Company, New York.

25. Nair, K.R. (1948). Biometrika, Vol. 35.

26. Neyman, J. (1935). J.Roy. Stat. Soc. Suppl. Vol II, No. 2.

27. Ottestad, P. (1953). Skandinavisk Aktuarietidskrift, 1952.

28. Ottestad, P. (1960). Sci. Reports from The Agric. Coll. of Norway, Vol. 39, No. 7.

29. Pearson, E.G. and Hartley, H.O. (1958). Biometrika Tables for Statisticians, Vol. I.

30. Pitman, E.J.G. (1937). Biometrika, Vol. 29.

31. Quenouille, M.H. (1959). New Statistical Tables Series No. XXVII (Biometrika). Vol. 46, parts 1 and 2.

32. Scheffé, H. (1959). The Analysis of Variance. John Wiley & Sons, Inc. New York.

33. Tukey, J.W. (1946). Biometrics, Vol. 5, No. 2.

34. Welch, B.L. (1937). Biometrika, Vol. 29.

35. Wold, H. (1954). Tracts for Computers, Vol. XXV. Cambridge Univ. Press.