CHAPTER 1

# Preliminaries

## 1.1 COMPARATIVE EXPERIMENTS

This book is about the planning of experiments in which the effects under investigation tend to be masked by fluctuations outside the experimenter's control. Large uncontrolled variations are common in technological experiments and in many types of work in the biological sciences, and it is in these fields that the methods to be described are most used. It is likely, however, that acquaintance with the simpler methods is of some value in most branches of experimental science.

The following are some typical situations in which large erratic fluctuations occur.

*Example* 1.1. Most agricultural field trials have as their object the comparison of a number of varieties of some crop, or of a number of alternative manurial treatments, or of a number of systems of management, etc. The experimental area is divided into plots and the different varieties, or whatever is under comparison, are assigned one to each plot. The yield, or some other property, is then measured or estimated for each plot and from the observations a comparison of varieties is made. Experience shows that even if the same variety were to be sown on all plots there would still be substantial variations in yield from plot to plot, the main features of this variation being that

(*a*) neighboring plots tend to give yields more alike than distant plots;
(*b*) there may be systematic trends or locally periodic variations across a field;
(*c*) if the experiment is repeated in a different field or in a different year, there may be a substantial change in the mean yield.

It would be common for the yields on individual plots in a field to vary by as much as $\pm 30\%$ from their mean, and a systematic difference of $5\%$ between varieties might be of considerable practical importance. We shall be concerned with methods for arranging the experiment so that we may with confidence and accuracy separate the varietal differences, which interest us, from the uncontrolled variations, which do not.

The aim of such experiments is the comparison of varieties rather than the absolute determination of the yield per acre likely from a given variety under given conditions. There are two reasons for this. First, the differences between varieties determine any practical recommendations that may be based on the

experiment; i.e., a choice of which of the two varieties is to be preferred depends not on the absolute yields but on how much more one variety is likely to yield than another, and on differences between any other properties that are considered important. Second, it is common for the difference between varieties to remain relatively constant even when the substantial changes in mean yield mentioned in (*c*) occur. This implies that it is much more economical to make a direct comparison of varieties than to estimate, in separate experiments for each variety, the mean yield under representative conditions and then to compare the estimates.

To sum up the discussion of this example, we are concerned with an experiment in which

(*a*) the object is to compare a number of varieties (or treatments);
(*b*) in the absence of varietal differences there is a substantial variation in yield from plot to plot;
(*c*) differences between varieties are comparatively stable, even though the mean level of response may fluctuate.

It is convenient to introduce a standard terminology. We shall refer to the plots as *experimental units*, or more briefly as *units*, and to the varieties, fertilizers, etc. under comparison as *treatments*. The formal definition of an experimental unit is that it corresponds to the smallest division of the experimental material such that any two units may receive different treatments in the actual experiment. For example suppose that in order to estimate the yield from the plots, two sub-areas are taken on each plot and the crop on these harvested and weighed. These sub-areas are not the experimental units, because the two sub-areas on one plot always receive the same treatment.

*Example* 1.2. Many experiments in industrial technology have a similar form to Example 1.1. The object may be to compare a number of alternative methods of processing, or to assess the effect of a modification to a standard process. The experiment consists in dividing the raw material into batches and then processing one batch by one process in the first period (day, hour, etc.), another batch in the next period by, in general, a different process, and so on. Or there may be several sets of machinery in use simultaneously. An observation (mean strength, yield of product, etc.) is made for each batch. In the absence of process differences the observation will vary from batch to batch and in addition to apparently random variation there may be smooth trends following, for example, hour-to-hour and day-to-day variations in temperature and relative humidity, and also sudden discontinuities corresponding to the introduction of fresh consignments of raw material.

*Example* 1.3. When slates, on which are bases of *Balanus balanoides*, are exposed in sea-water, the setting of further barnacles of this type occurs rapidly. Knight-Jones (1953), in investigating the mechanism of setting, exposed untreated slates and slates that had been treated with a variety of chemical reagents. By finding which reagents produced a substantial decrease in the amount of setting he was able to infer something about the chemical processes involved.

This experiment has a feature additional to those of Examples 1.1 and 1.2 in that the comparison of treatments is of interest only insofar as it aids in revealing the nature of the phenomenon under investigation. The experiment is concerned with comparisons because it is advisable to include as a control a series of



untreated slates. This is to ensure that any observed decrease in the rate of setting after treatment is not due to a change in the natural rate of setting, which is subject to erratic fluctuations.

The experimental units are slates, the observation is the number of barnacles setting in a three-day period, and the treatments are the control and the various chemical reagents.

*Example* 1.4. One method of determining the potency of a drug is by direct comparison with an agreed standard in the following way: The drug is applied at a constant rate to an experimental animal and the dose at which death, or some other recognizable event, occurs is noted. This critical dose is called the tolerance or threshold. This is repeated for a number of animals using the drug under analysis and the standard. The tolerances vary from animal to animal but by comparing the mean log tolerances (see § 2.2) for the drug and for the standard a measure of potency is obtained. Here each animal is an experimental unit receiving one of two possible treatments, the drug and the standard.

An alternative procedure would be to measure the potency directly by, say, the mean log tolerance, without using a standard. This is usually unsatisfactory because the tolerance varies appreciably from group to group of animals, so that results in different laboratories and at different times would be only very roughly comparable. Experience shows that differences of log tolerance between a drug and a suitable standard are often little affected by systematic differences between groups of animals, so that the introduction of a standard into the experiment leads to a measure of potency that can be reproduced to within close limits at different times and places.

This simple form of comparative bioassay is discussed fully by Finney (1952).

*Example* 1.5. The clinical investigation of the use of new medical treatments raises similar problems of experimental design. It is almost always advisable to include a control treatment in the investigation, as well as the new treatment, because the effect of the new treatment may, except in dramatic cases, be shown by a comparatively small change in the proportion of cures. There are several cogent reasons, which will be discussed in detail later, why the determination of the proportion of cures for the control treatment should be part of the experiment and not just based on past experience. In this application each patient is an experimental unit, receiving one of two or more possible treatments.

In the treatment of serious diseases there is the complication that it will be considered unethical to withhold a treatment that is suspected to give increased chance of survival. This makes it imperative to conclude the experiment as soon as there is reasonable evidence that a particular treatment is in fact superior (Armitage, 1954).

An essential difference between the above experiments and many experiments in physics and chemistry is that in the latter, once the experimental technique is mastered and the apparatus working correctly, closely reproducible results are obtained. More precisely the uncontrolled variations are small compared with the effects to be expected when a change is imposed on the system. Therefore, if the system is altered and the observation changes, the imposed alterations may safely be assumed to be the cause of the change in the observation. In such cases the

methods described in this book are of little value, except as a safeguard against errors arising from defects in the apparatus. However, as soon as the effects under investigation become comparable with the uncontrolled variations, the problems we shall be concerned with become important.

Examples 1.1–1.5 are all of the same form. We have a number of experimental units and a number of alternative treatments. The experiment consists in applying one treatment to each unit and making one (or more) observations, the assignment of treatments to units being under the experimenter's control. When the object of such an experiment is the comparison of treatments rather than the determination of absolute values, the experiment will be called *comparative.**

The main planned investigations that are not comparative experiments are concerned with determining the properties of defined sets of things, such as the mean fiber diameter of a consignment of wool, the number of species of beetle in a particular area, or the characteristics of children in an area who watch television (*a*) frequently, (*b*) infrequently.

It is especially important to distinguish between the type of comparison that would be made in the last example and the type that would be made in a comparative experiment. The crucial distinction is that in the experiment the choice of treatment for each unit is made by the experimenter, whereas in the planned survey the observer has no control at all over what makes a particular individual fall in one group rather than another. Interesting conclusions can be drawn from planned surveys, particularly if comparisons are made within similar groups of individuals, for instance within groups of children of the same age, educational background, social class, etc. Nevertheless, much more cogent conclusions about causal effects can be drawn from experiments than from planned surveys. From this point onwards we restrict attention almost entirely to comparative experiments.

The discussion of the planning of such experiments falls into two almost distinct parts, dealing with the principles that should govern

(*a*) the choice of treatments to be compared, of observations to be made, and of experimental units to be used;

(*b*) the method of assigning treatments to the experimental units and the decision about how many units should be used.

Most of this book is about (*b*), but there is some attempt to discuss the first set of questions in Chapter 9.

It is convenient to discuss first the requirements for a good experiment.

* All measurements, including counting, are in a sense comparative, but this does not affect the distinction between comparative and other experiments, since within the framework of a particular experiment, measurements can usually be regarded as absolute.



## 1.2 REQUIREMENTS FOR A GOOD EXPERIMENT

We shall assume in this section that the treatments, the experimental units, and the nature of the observations have been decided on. The requirements for a good experiment are then that the treatment comparisons should as far as possible be free from systematic error, that they should be made sufficiently precisely, that the conclusions should have a wide range of validity, that the experimental arrangement should be as simple as possible, and finally that the uncertainty in the conclusions should be assessable.

These requirements will now be discussed in turn.

### (i) Absence of Systematic Error

This means that if an experiment of the given design were done using a large number of experimental units it would almost certainly give a correct estimate of each treatment comparison. Some examples should make the point clear.

*Example* 1.6. Consider an industrial experiment to compare two slightly different processes, $A$ and $B$, on the same machinery, in which $A$ is always used in the morning and $B$ in the afternoon. No matter how many lots are processed it is impossible, from the results of the experiment alone, to separate the difference between the processes from any systematic change in the performance of the machinery or operatives from morning to afternoon, unconnected with the difference between $A$ and $B$. Such systematic changes do sometimes exist. The difficulty is not met by a calculation of statistical significance; this may tell us that the apparent difference between $A$ and $B$ is unlikely to be a purely random one but cannot determine which of two or more possible explanations of the difference is the right one.

Of course it would be foolish to suggest that such an experiment is useless. Previous experimental work, or general knowledge of the process, or supplementary measurements on relevant variables (e.g., temperature, relative humidity) may suggest that any difference between conditions in the morning and afternoon is unimportant. Then, provided that it is clearly understood that the interpretation of the experiment rests on this extra assumption, no great harm may be done. But suppose that a surprising result is obtained, or a result that is in apparent contradiction with later work. Then unless the evidence for the absence of morning-afternoon differences is strong, the experiment may lose much of its cogency.

It is therefore a sound principle to plan an experiment so that such difficulties are as far as possible avoided, i.e., to ensure that experimental units receiving one treatment differ in no systematic way from those receiving another treatment.

Difficulties similar to those just discussed arise whenever the comparisons under test get completely mixed up with differences between batches of

experimental material, between observers, between different experimental methods, and so on. They are also liable to occur when all the units receiving one treatment are collected together in single groups and not left to respond independently.

*Example* 1.7. In animal feeding trials one possible plan is to have all animals receiving one treatment together in a single pen. This to some extent simulates practical conditions and also is very convenient in organizing the experimental work. If, however, we have one large pen of animals receiving the experimental ration, it is impossible to separate ration differences from systematic differences between pens or say from the presence in one pen of some disease wholly unconnected with the experimental treatments.

For example Yates (1934) has described an experiment on pigs in which the animals were divided into small groups housed separately, so that each treatment was tested on several entirely independent sets of pigs. It was found that pigs receiving no green food fell sick. Yates remarked that had the pigs receiving no green food been in a single pen it would probably have been concluded that the sickness was due to extraneous causes, particularly since previous experiments had suggested that green food was unnecessary. The fact that several independent sets of pigs receiving no green food fell sick and that no other pigs did so was, however, strong evidence that the treatment was responsible.

Another way of putting the difficulty is that in the experiment with single pens the experimental units are, in accordance with the definition of § 1.1, pens of animals, not single animals. Hence this is an experiment without replication for which further assumptions are needed before valid conclusions can be drawn.

The decision about what method of design to use in such experiments is not easy and the example is quoted primarily to illustrate the logical point involved. There are further discussions of animal feeding trials by Lucas (1948) and by Homeyer (1954).

A common type of experiment, of which Example 1.3 is an instance, involves applying a treatment, noting a change in the observation as compared with that expected in the absence of the treatment, and concluding that the treatment has caused the change. For such an experiment to be convincing by itself, the treated units must be compared with a control series of units, receiving no treatment, but included in the experiment under the same conditions as the treated units, and not being systematically different from them. To say that a certain observation has been obtained in the past, and that the treated units now give a different observation, is not by itself necessarily cogent evidence of a treatment effect, since there may be systematic differences among the experimental units or a systematic change in the external conditions. If past experience has shown that the observations on untreated units vary in a stable way, it may be in order to dispense with special control units, particularly in preliminary work. However this procedure is the same as allowing possible systematic differences between units in an experiment, such as in Example 1.6, and is best avoided in the great majority of cases.



A classical example of an experiment that was largely vitiated by the absence of controls is the following.

*Example* 1.8. McDougall (1927), to examine a possible Lamarckian effect in rats, taught some rats to choose between a lighted and an unlighted exit. He then bred from them and for each generation measured the speed with which the above task was learned. A Lamarckian effect would be shown by a steady increase in speed with generation number and this was in fact found. Certain other explanations, such as selection, were ruled out but there were no control units, i.e., no rats bred under the same conditions, but from untrained parents. Therefore it was possible that the effect was due to systematic uncontrolled variations in the experimental conditions.

Crew (1936) repeated the experiment with controls and found no apparent Lamarckian effect. Agar et al. (1954), in an experiment continued over a period of 20 years, found an initial increase in speed similar to McDougall's, but the same for the control as for the "treated" rats. They concluded that the effect was due to secular changes in the health of the colony of rats.

We can sum up as follows: experimental units receiving one treatment should show only random differences from units receiving any other treatment, including the control, and should be allowed to respond independently of one another. When it is impossible or impracticable to achieve this, any assumption about the absence of systematic differences should be explicitly recognized and as far as possible checked by supplementary measurements or by previous experience.

We shall see later how it is possible to ensure the absence of the main sources of systematic error by means of a randomization procedure.

**(ii) Precision**

If the absence of systematic errors is achieved by randomization (Chapter 5), the estimate of a treatment contrast obtained from the experiment will differ from its true value* only by random errors. It should be noted that the term random will be used throughout in its technical statistical sense. Roughly speaking this means that it refers to variations showing no reproducible pattern. For example, the variations of yield in a field described briefly in Example 1.1 are not random, because of the trends, correlation between yields on adjacent plots, etc.

The probable magnitude of the random errors in the estimate of the treatment contrast can usually be measured by the *standard error*. The precise definition and method of calculation of this is described in textbooks on statistical methods, for example in that of Goulden (1952, pp. 17–20), but for the present purpose a sufficiently good idea of its meaning can be grasped as follows:

In about one case out of three the estimate will be in error by more than plus or minus the standard error.

* The true value is defined more precisely in Chapter 2.

In about one case out of twenty the estimate will be in error by more than plus or minus twice the standard error.

In about one case out of a hundred the estimate will be in error by more than plus or minus two and one-half times the standard error.

These statements require some qualification depending on the form of the distribution of the errors and on the accuracy of the standard error, which itself has to be estimated; these points need not concern us at the moment.

The value of the standard error, and hence the precision of any particular experiment, will depend on

(*a*) the intrinsic variability of the experimental material and the accuracy of the experimental work;

(*b*) the number of experimental units (and on the number of repeat observations per experimental unit);

(*c*) the design of the experiment (and on the method of analysis if this is not fully efficient).

In most of the experiments where statistical design is useful, only a very limited increase in precision can be achieved by modifying the experimental material or by increasing the precision of measuring devices. This is partly because there is often an intrinsic variability that is very difficult to remove and partly because experiments under very controlled conditions, e.g., in greenhouses, in small-scale industrial plants, etc., cease to be representative of practical conditions. The point will be discussed again in Chapter 9.

If there is one observation per experimental unit, then, other things being equal, the standard error of the estimate of the difference between two treatments is inversely proportional to the square root of the number of units for each treatment. In fact, the standard error is

$$\text{standard deviation} \times \sqrt{\left(\frac{2}{\text{no. of units per treatment}}\right)}, \qquad (1)$$

or if there are differing numbers of observations on the two treatments $A$, $B$, it is

$$\text{standard deviation} \times \sqrt{\left(\frac{1}{\text{no. of units for } A} + \frac{1}{\text{no. of units for } B}\right)}. \qquad (2)$$

Here the standard deviation is a statistical measure of the random dispersion of the observations on experimental units treated alike (Goulden, 1952, p. 17).*

* Note that the standard deviation refers to the variation of the observations on individual units, whereas the standard error refers to the random variation of an estimate from a whole experiment.



From equation (1), the standard error is halved by a fourfold increase in the number of experimental units, but a hundredfold increase in the number of units is necessary to divide the standard error by ten. Although in theory the standard error can be made arbitrarily small by increasing the number of units, this is an expensive method of increasing precision.

The gain due to taking repeat observations on the experimental units is less than or equal to the gain from a corresponding increase in the number of units. It can be assessed from formulas similar to, but a little more complicated than, (1) and (2).

The third method of increasing precision is by improved design and it is with this that we shall be most concerned. The general idea is that whatever knowledge is available about the experimental units should be used to reduce the effective standard deviation in (1) and (2). It is sometimes possible to obtain an increase in precision equivalent to a substantial increase in the number of experimental units.

Our requirement about precision is, roughly speaking, that the standard error should be sufficiently small for us to be able to draw cogent conclusions, but not too small. If the standard error is large the experiment is, by itself, almost useless, whereas an unnecessarily small standard error implies a waste of experimental material. In the majority of cases the object is the estimation of treatment differences, and in these cases formulas (1) and (2) enable us to predict, when we are designing the experiment, the precision to be obtained with any given number of units or, alternatively, the number of units necessary for a given precision. For this we must know something about the standard deviation, i.e., the variability of the units, but approximate information from previous similar experiments is often available. Occasionally the object is not the estimation of treatment differences but is to reach an irreversible decision on, say, which of a number of treatments is the best. In this case if one treatment is much better than the rest and the units are tested in sequence, the experiment can be ended after a small number of observations, even though the precision of estimation is still low. This raises special problems. The whole question of the choice of number of units will be discussed in detail in Chapter 8.

### (iii) Range of Validity

When we estimate the difference between two treatments, we obtain conclusions referring to the particular set of units used in the experiment and to the conditions investigated in the experiment. If we wish to apply the conclusions to new conditions or units, some additional uncertainty is involved over and above the uncertainty measured by the standard error. The only exception to this statement is when the units

in the experiment are chosen from a well-defined population of units by a proper statistical sampling procedure.

The wider the range of conditions investigated in the experiment, the greater is the confidence we have in the extrapolation of the conclusions. Therefore if we can arrange, without decreasing the accuracy of the experiment, to examine a wide range of conditions, this is desirable. This is particularly important in experiments to decide some practical course of action and rather less so where the object is purely to gain insight into some phenomenon.

*Example* 1.9. "Student" (1931) mentions some experiments done by the Irish Department of Agriculture in connection with the introduction of Spratt–Archer barley. This was almost everywhere a great success; yet in one district the farmers refused to grow it, alleging that their own native race of barley was superior. After some time the Department, to demonstrate Spratt–Archer's superiority, produced a single-line culture of the native barley and tested it against the Spratt–Archer in the district in question. "Student" reports that to the Department's surprise the farmers were perfectly right: the native barley gave the higher yield. At the same time, the reason became clear: the barley in question grew more quickly and was able to smother the weeds, which flourished in that area; Spratt–Archer, growing less strongly to begin with, was, however, the victim of the weeds. Thus the original experiments, carried out on well-farmed land, were definitely misleading when their conclusions were applied elsewhere.

Similar points arise with other types of experiment. A new experimental technique that works very well when special attention is devoted to it may be quite unsuited to routine use. A new industrial process that works well under special supervision during an experiment may not be successful in routine production. Or, to take a more specific example, a modification to a textile process tested on a homogeneous batch of raw material may in fact be quite critically dependent on the oil content of the raw material. The difference between varieties of wheat may be dependent on soil and weather conditions, and so on.

There are several consequences of these remarks. First it is important, even in purely technological experiments, to have not just empirical knowledge about what the treatment differences are, but also some understanding of the reasons for the differences. Such knowledge will indicate what extrapolation of the conclusions is reasonable. Secondly we should, in designing the experiment, artificially vary conditions if we can do so without inflating the error. For example in comparing two methods of drawing wool, it may sometimes be expected that the difference between the methods is unaffected by the oil content of the wool. It would often be advantageous to include both lightly and heavily oiled wool in the experiment with the hope of providing a direct check on the independence



of the difference between methods and the oil content. The snag is, of course, that if several such supplementary factors are included the experiment may become difficult to organize, and also there is the possibility, if the system is a complicated one, that no clear-cut conclusions can be drawn, owing to no one set of conditions having been thoroughly investigated. This leads to the third point that it is important to recognize explicitly what are the restrictions on the conclusions of any particular experiment.

These considerations are rather less important in purely scientific work, where the best thing is usually to try to gain thorough insight into some very special situation rather than to obtain in one experiment a wide range of conclusions.

**(iv) Simplicity**

This is a very important matter which must be constantly borne in mind but about which it is difficult to make many general remarks. There are several considerations involved. If the experiment is to be done by relatively unskilled people, it may be difficult to ensure adherence to a complicated schedule of alterations. If an industrial experiment is to be run under production conditions, it will be important to disturb production as little as possible, i.e., to have a few long runs of the different processes rather than frequent changes. In scientific work, particularly in the preliminary stages of an investigation, it may be important to retain flexibility; the initial part of the experiment might suggest a much more promising line of enquiry, so that it will be a bad thing if a large experiment has to be completed before any worth-while results are obtained. Nevertheless there certainly are cases where a fairly complicated arrangement is advantageous and it is a matter of judgement and experience to decide how far it is safe to go in any particular application.

The above remarks apply to simplicity of design. It is also desirable to have simple methods of analysis. Fortunately the requirements of efficiency in design and simplicity in analysis are highly correlated and for nearly all the methods in this book, straightforward schemes of full statistical analysis are available, provided that certain assumptions to be described later are satisfied. If only estimates of the treatment differences are required, with no estimates of precision, few of the designs require more than simple averaging.

The use of electronic computers for the analysis of experimental results is an important recent development, particularly for those fields where either very large amounts of data are involved or where the time taken on the experimental work is comparable to or smaller than the time it would take to analyze the results by conventional methods. Once suitable

programs have been written, the time taken to make a statistical analysis on an electronic computer is likely to be very small in all ordinary circumstances.

### (v) The Calculation of Uncertainty

The previous requirements have not been statistical; this last one is. It is desirable that we should be able to calculate, if possible from the data themselves, the uncertainty in the estimates of the treatment differences. This usually means estimating the standard error of the differences, from which limits of error for the true differences can be calculated at any required level of probability, and from which the statistical significance of the differences between the treatments can be measured.

To be able to make this calculation rigorously we must have a set of experimental units responding independently to one treatment and differing only in a random way from the sets of units for the other treatments. A comparison, not necessarily straightforward, of the observations on units receiving the same treatment then gives a valid measure of error. The use of randomization, discussed in detail in Chapter 5, to eliminate systematic differences between units treated differently, automatically makes differences random and justifies the statistical analysis under weak assumptions. The distinction between such an analysis and that of Example 1.6 should be carefully noted.

In experiments with very small numbers of experimental units it may not be possible to obtain an effective estimate of the error standard deviation from the observations themselves. In such cases it will be necessary to use the results of previous experiments to estimate the standard deviation (see also § 8.3); the disadvantage of this is that we need to assume that the amount of random variation is unchanged.

As a general rule, methods of statistical analysis will not be described in this book. This is partly because there are a number of excellent accounts of such methods available, and partly because their inclusion would not only greatly increase the length of the book but would also tend to distract attention from considerations of design.

## SUMMARY

We deal mostly with experiments of the following form: there are a number of alternative *treatments* one of which is applied to each *experimental unit*, an *observation* (or several observations) then being made on each unit. The object is to be able to separate out differences between the treatments from the uncontrolled variation that is assumed to be present;



this may of course be only the first step towards understanding the phenomena under investigation.

Once the treatments, the experimental units, and the nature of the observations have been fixed, the main requirements are that

(*a*) experimental units receiving different treatments should differ in no systematic way from one another, i.e., that assumptions that certain sources of variation are absent or negligible should, as far as practicable, be avoided;

(*b*) random errors of estimation should be suitably small, and this should be achieved with as few experimental units as possible;

(*c*) the conclusions should have a wide range of validity;

(*d*) the experiment should be simple in design and analysis;

(*e*) a proper statistical analysis of the results should be possible without making artificial assumptions.

## REFERENCES*

Agar, W. E., F. H. Drummond, O. W. Tiegs, and M. M. Gunson. (1954). Fourth (final) report on a test of McDougall's Lamarckian experiment on the training of rats. *J. Exp. Biol.*, **31**, 307.

Armitage, P. (1954). Sequential tests in prophylactic and therapeutic trials. *Q. J. of Medicine*, **23**, 255.

Crew, F. A. E. (1936). A repetition of McDougall's Lamarckian experiment. *J. Genet.*, **33**, 61.

Finney, D. J. (1952). *Statistical method in biological assay*. London: Griffin.

Goulden, C. H. (1952). *Methods of statistical analysis*. 2nd ed. New York: Wiley.

Homeyer, P. G. (1954). Some problems of technique and design in animal feeding experiments. Chapter 31 of *Statistics and Mathematics in Biology*. Ames, Iowa: Iowa State College Press. Edited by O. Kempthorne et al.

Knight-Jones, E. W. (1953). Laboratory experiments on gregariousness during setting in *Balanus balanoides* and other barnacles. *J. Exp. Biol.*, **30**, 584.

Lucas, H. L. (1948). Designs in animal research. *Proc. Auburn Conference on Applied Statistics*, 77.

McDougall, W. (1927). An experiment for testing the hypothesis of Lamarck. *Brit. J. Psychol.*, **17**, 267.

"Student" (1931). Agricultural field experiments. *Nature*, **127**, 404. Reprinted in "*Student's*" *collected papers*. Cambridge, 1942.

Yates, F. (1934). A complex pig-feeding experiment. *J. Agric. Sci.*, **24**, 511.

* These are explicitly referred to in the text. There are some general references on p. 294.