Experimental Designs

WILLIAM G. COCHRAN
Professor of Biostatistics
The Johns Hopkins University

and

GERTRUDE M. COX
Director, Institute of Statistics
University of North Carolina

SECOND EDITION

New York · John Wiley & Sons, Inc.
London · Chapman & Hall, Limited

Library of Congress Catalog Card Number: 57-5908

PRINTED IN THE UNITED STATES OF AMERICA

Experimental Designs

Preface to the Second Edition

DURING THE SIX YEARS SINCE THE PUBLICATION OF THE FIRST edition, there has been a substantial increase in the attention paid by research workers to the principles of experimental design. Evidence of this change can be seen in the medical and social sciences, in physics and chemistry, and in industrial research. Another encouraging trend is that workers in these areas, although still willing to utilize appropriate designs taken from agricultural experimentation, have begun to examine their own problems of experimentation in a fresh light and to produce new designs better suited to their particular conditions. The recent literature also reflects a move toward greater depth and comprehensiveness in experimental work, as instanced by numerous papers devoted to experimentation with more than one factor.

In preparing a second edition we decided to leave the framework of the book unchanged. New material is included wherever necessary to fulfill our original objective of making available the types of design most likely to be useful in practice.

Two chapters have been added. Chapter 6A deals with the fractional replication of factorial experiments. This technique, initially applied in agricultural field experiments, was described briefly in the first edition. In the intervening period, fractional replication has proved so productive in exploratory research, particularly in industry, that a more complete account is justified.

Chapter 8A is also concerned with experiments of the factorial type. It presents methods and new designs for experiments in which the factors represent quantitative variables, measured on a continuous scale. In this situation it is often natural to think of the yield or response as related to the levels of the factors by some mathematical function which, if it could be found, would summarize compactly all the results of the experiment. The new methods are applicable when this function can be approximated, within the limits of the experimental region, by a polynomial of the first or second degree. The second half of this chapter gives an introductory account of experimental strategies for finding, as quickly as possible, the levels at which the factors must be set in order to obtain the maximum response. These techniques, produced for industrial research and development, have many potential applications in other areas.

The appearance of several series of new incomplete block designs during the last six years necessitated expansions in some of the later chapters. The investigator who has found these designs useful now has a considerably wider selection. In fact, the large number of new designs forced us to revise one of our original objectives. When it appeared, the first edition contained the plans of substantially all the designs in common use at that time. To include the new plans, however, would involve devoting over 200 extra pages to plans alone. Even if this were done, our set of plans would soon become incomplete as further research is published. We have tried to handle this problem by giving, at the end of Chapter 9, an index to the incomplete block designs now available. For designs for which no plan is presented, the purpose of the design and the method of analysis are explained at the appropriate place in a later chapter, with references to the sources of the plans. The principal series of designs represented in this way are some new confounded factorial designs in single replication, partially balanced incomplete block designs (with elimination of either a single or a double grouping), the chain block and generalized chain block designs and some further designs related to the latin square.

To cover other developments, a number of sections have been added. Section 2.22a presents a table for estimating the numbers of replications needed for comparing two treatments when the data are arranged in two classes, as dead or alive, sound or defective. Sections 4.15a and 4.27a give the methods of analysis with data of this type for completely randomized and randomized block designs. These sections fill a gap in the first edition, which gave methods of analysis only for continuous data.

Sections 4.61a to 4.66a contain a discussion of the use of latin squares in adjusting for residual effects which may be present when the treatments are applied in sequence to the same subject.

Topics that are presented more briefly are sequential experimentation (section 2.23a), the testing of effects suggested by the data (section 3.53), the problem of making several tests of significance in the same experiment (section 3.54a), Yates' automatic method of computing factorial effect totals (section 5.24a), additional standard error formulae for split-plot experiments (section 7.22), the effects of errors in the weights on the recovery of inter-block information (section 10.12a), and the use of balanced incomplete block designs in taste and preference testing (section 11.1a). Some of these topics deserve more space than we have given them, but the number of pages added in this edition is already greater than we should have liked.

A number of users of the book, particularly those who use it for teaching, asked us to prepare this edition so that the new material is clearly identifiable. To accomplish this, we have numbered the new chapters as 6A and 8A. Any new section, table, or plan carries the letter *a*, e.g., section 2.23a, table 11.1a. Where a minor revision or addition has been made within a section, the old section number remains unchanged.

For permission to use data as examples we are indebted to the General Foods Corporation and to Doctors J. Doull and J. N. Tolentino. Considerable help has been received from staff members at our respective institutions and from colleagues elsewhere.

The large sale of the first edition of a book that is not elementary has been highly gratifying. It testifies to the alertness of workers in the experimental sciences in seeking out any new ideas and methods that may make their investigations more efficient and productive.

W. G. Cochran
G. M. Cox

May, 1957

Preface to the First Edition

WORK ON THIS BOOK WAS STARTED WHEN BOTH OF US WERE MEMbers of the staff of Iowa State College. At that time requests were received rather frequently from research workers. Some wanted advice on the conduct of a specific experiment: others, who had decided to use one of the more complex designs that have been discovered in recent years, asked for a plan or layout that could be followed during the experimental operations. Although the logical principles governing the subject of experimentation are admirably expounded in Fisher's book *The Design of Experiments*, these requests indicated a need for a different type of book, one which would describe in some detail the most useful of the designs that have been developed, with accompanying plans and an account of the experimental situations for which each design is most suitable. Such a book is directed at the experimenter and is intended to serve as a handbook which is consulted when a new experiment is under consideration.

Mainly on account of the war, slow progress was made. In 1944 we completed a mimeographed draft of which several hundred copies were distributed. Many helpful suggestions were made by readers. Of these, the most frequent was to the effect that we should include more material dealing with the statistical analysis of the results. In the mimeographed draft, our practice was to give references to worked examples of the analysis in cases where they could be found in the literature, and to present examples for only those designs whose analysis was not available. To this it was objected that many research workers did not have easy access to our references.

This suggestion raised a difficult issue. To present a self-contained account of the analysis of variance in all its ramifications would make the book, it seemed to us, unwarrantably long and expensive. Consequently we have continued to assume that the reader has some knowledge of the principles of analysis of variance and of the computational methods involved. We have included a brief review of the basic theory and an extensive set of worked examples of the analysis for both common and less common types of design. Although strenuous efforts were made to obtain a selection of examples from diverse fields of research, a preponderance from biology, and more particularly from agriculture, was dictated by our own experience in those areas. On several occasions we rejected an example which would have made an attractive addition to the scope because we did not feel sufficiently familiar with the conditions of experimentation to give a realistic account of the problems encountered.

Since courses of lectures on the design of experiments are being introduced in many colleges and universities, some teachers may be interested in the potentialities of the book as a textbook. Several comments are prompted by our own experience in this connection. First, it will often be necessary for the teacher to provide a more systematic development of the analysis of variance than is given here. Second, interest in such a course is greatly enhanced by examples from an environment with which the listeners are familiar, and especially by examples of experiments that have been conducted by some of the listeners. Thus the teacher is well advised to build the course around his own examples, using those in this book mainly as supplementary material for the students. Finally, a selective use of the book is in order, because it contains much more material than can be covered in a typical one-quarter course, and because some difficult topics have been dealt with in the early parts of the book.

We wish to express our gratitude to the many staff members at Iowa State College and North Carolina State College who helped us by putting experimental data at our disposal, by providing painstaking descriptions of their experimental techniques, or by lending their assistance in the preparation of the manuscript and plans. For similar kindnesses in connection with certain of the examples we are indebted to G. F. Potter, D. Y. Solandt, D. B. DeLury and J. Hunter, F. M. Wadley and C. F. Rainwater, and the Wailuku Sugar Company, Honolulu. Some theoretical results used in Chapter 14 were developed from research conducted at Raleigh under a contract with the Office of Naval Research. Finally, our thanks go to George W. Snedecor, who participated in the original plans for this book and made a careful reading of the first draft.

W. G. COCHRAN
G. M. COX

January, 1950

Contents

CHAPTER 1

Introduction

1.1 The Contribution of Statistics to Experimentation

1.11 The Problem of Interpretation. Since statisticians do not usually perform experiments, their claim to attention when they write on this subject requires some explanation. It is true that on many important aspects of experimentation the statistician has no expert knowledge. Nevertheless, in recent years, research workers have turned increasingly to statisticians for help both in planning their experiments and in drawing conclusions from the results. That this has happened is convincing evidence that statistics has something to contribute.

At first, requests for assistance were nearly always concerned with the interpretation of the results. It is a common characteristic of experiments in widely diverse fields of research that, when they are repeated, the effects of the experimental treatments vary from trial to trial. This variation introduces a degree of uncertainty into any conclusions that are drawn from the results. Even after a number of repetitions, or *replications* as they are called, the investigator still does not know by how much his results would be changed if the experiment were repeated further under the same conditions. Successive trials may be so discrepant in their results that it is doubtful which of two treatments would turn out better in the long run.

As an illustration of this variation, data are given in table 1.1 from a simple experiment. The data are the times (minus 2 minutes) required to perform a routine statistical calculation, that of finding the sum of squares of 27 observations. Each sum of squares was computed separately by the same person on both of two standard computing machines. In all, 10 different sums of squares were worked, making 10 trials or replications of the experiment. It will be noted that the differences in speed range from 17 seconds in favor of machine B to 2 seconds in favor of machine A. Some experimenters may comment that the results of this experiment are remarkably well behaved, and exhibit nothing like the variation with which they have to contend. The results will, however, serve our purpose.

TABLE 1.1 TIME IN SECONDS (MINUS 2 MINUTES) REQUIRED FOR COMPUTING SUMS OF SQUARES

Replication	Machine A	Machine B	Difference (A − B)
1	30	14	16
2	21	21	0
3	22	5	17
4	22	13	9
5	18	13	5
6	29	17	12
7	16	7	9
8	12	14	−2
9	23	8	15
10	23	24	−1
Means	21.6	13.6	8.0

The object of the experiment is, of course, to compare the speeds of the two machines for this calculation. More specifically, two objects might be stated. The first is to answer the question: is there any difference in speed? or, to put it another way, to test the hypothesis that there is no difference in speed. The second object, which is related to the first, is to estimate the size of the difference in speed. Almost all experiments are carried out for one or both of these purposes—the testing of hypotheses and the estimation of differences in the effects of different treatments.

As regards the test of the hypothesis that there is no difference in speed, we might report, as relevant evidence, that B proved faster 7 times out of 10, A twice, while once there was a tie. For the problem of estimation we might report that the average difference in speed in the experiment was 8 seconds in favor of B. These purely descriptive statements do not carry us very far. Their weakness is that they supply no information about the reliability of the figures presented. For example, have we any confidence that, if the experiment were continued for another set of 10 trials, the advantage at the end would still be close to 8 seconds in favor of B?

Because of the deficiencies in the descriptive approach, experimenters have adopted a different point of view in the summarization of their results. They tend to reason as follows. Suppose that it were feasible to continue the experiment indefinitely under the same conditions. The average difference in speed between the two machines would presumably settle down to some fixed value. This value, which will be independent of the size of the experiment that was actually carried out, may reasonably be called the *true* difference between A and B. From this point of view, the problem of summarizing the results may be restated in the question: what can we say about the true difference between A and B? This is a problem in *induction* from the part to the whole, or in statistical language, from the sample to the population. A solution to this problem has been developed by means of the theory of statistics. It is this solution that constitutes the principal contribution of statistics to the interpretation of the results.

1.12 Statistical Inference. Obviously, it cannot be expected that the solution will provide the exact value of the unknown true difference. As a less ambitious goal we might hope to be able to find 2 limits within which the exact value is certain to lie, but even this cannot quite be attained. What can be done is that for any chosen probability, say .95, two limits are found such that the probability that they enclose the true difference is .95. In other words, limits can be found that are almost certain to enclose the true difference, where the degree of certainty, as measured by the probability, can be chosen by the experimenter. Since we wish to focus attention on the type of inferential statement that can be made rather than on the method of calculating the limits, the computations will not be discussed at present. For the example in table 1.1 they will be found in section 4.42. When the probability is .95, the limits for the true difference in speed between the 2 machines turn out to be 3.3 and 12.7 seconds in favor of machine B. A statement that B is faster by an amount that lies between 3.3 and 12.7 seconds has a 1 in 20 chance of being wrong. If the degree of certainty is decreased by lowering the probability to .8, the limits are narrowed to 5.1 and 10.9 seconds. If the probability is raised to .99, the limits become 1.1 and 14.9 seconds, and as the probability is brought closer to certainty, the limits steadily become farther apart. The limits are called confidence limits, and the probabilities are called confidence probabilities.*

These probabilities are not merely academic abstractions: they can be subjected to at least a rough experimental verification. For verification, we need a situation where the true difference between the effects of two treatments is known. In toxicology, for instance, this situation can sometimes be obtained by diluting a standard poison to a known extent. The dilution is sent to the laboratory, labelled as an "unknown" poison. By experiments on animals, confidence limits are found for the

* Fisher (1.2) has developed statistical inference in terms of *fiducial* limits. The two concepts, fiducial and confidence limits, have different logical backgrounds, although in all simple applications to controlled experiments the actual values of the fiducial and confidence limits are identical. For a discussion, see Kendall (1.1).

amount of the unknown that has the same toxicity as a given amount of the standard. Since the persons originating the experiment know the true value for this amount, they can verify whether the statement that the true amount lies within the limits is correct. The practical difficulty with this type of check is the labor required. A large number of experiments would be needed to verify whether about 95% of the statements made with confidence probability .95 were in fact correct, and about 5% were wrong.

As we have seen, the statistical solution to the problem of estimation consists of a statement that the true difference lies between certain limits, plus a probability that the statement is correct. It is of interest to consider whether this type of information is sufficiently precise to permit decisions of practical importance to be made. Although a thorough discussion of this question would be rather lengthy, inferences of this type often permit definite action to be taken with confidence that the action will be fruitful. When they fail to decide the point at issue, the reason is nearly always that the data obtained are insufficient. For illustration, suppose that it is desired to discover whether the application of a dressing of some fertilizer to a crop will be profitable. The cost of the fertilizer is such that its application will be profitable only if it increases yields on the average by 2 or more bushels. A series of experiments is carried out in order to estimate the true average response to the fertilizer. If the 95% confidence limits for the increase due to the fertilizer are 4 and 11 bushels, its use can be adopted with a good deal of assurance that it will be profitable. Similarly, if the 95% confidence limits are −5 and 1 bushels, a decision not to use the fertilizer follows. A case where there is uncertainty occurs when the limits are 0 and 5 bushels. Here it is likely that there will be either a small gain or a small loss, but no recommendation can be made without considerable risk of its being wrong. If it is important to make the correct decision, further experiments must be conducted in order to narrow the distance between the confidence limits.

Thus far we have considered the problem of estimating the true difference between the effects of two treatments. In testing hypotheses, we are interested in the supposition that the true difference has some specified value, most commonly zero. As in the case of estimation, difficulty arises because of the variability that is typical of experimental data. As a result of this variability, the data are never exactly in agreement with the hypothesis, and the problem is to decide whether the discrepancy between the data and the hypothesis is to be ascribed to these variations or to the fact that the hypothesis is not true. The contribution of statistics is the operation known as a *test of significance*.