UNIVERSAL
LIBRARY
OU_148800
UNIVERSAL
LIBRARY

OSMANIA UNIVERSITY LIBRARY

Call No. 311.9 P195 Accession No. 13705

Author Pawse + Sukhatme

Title Statistical methods for Agriculture

This book should be returned on or before the date last marked below.

STATISTICAL METHODS
FOR
AGRICULTURAL WORKERS

STATISTICAL METHODS
FOR
AGRICULTURAL WORKERS

BY

V. G. PANSE, B.Sc., Ph.D. (Lond.), F.N.I.

Statistical Adviser
Indian Council of Agricultural Research, New Delhi

AND

P. V. SUKHATME, Ph.D., D.Sc. (Lond.), F.N.I.

Chief, Statistics Branch
Food and Agriculture Organization of the United Nations, Rome



Post Graduate Library
College of Arts & Commerce

Published by
INDIAN COUNCIL OF AGRICULTURAL RESEARCH
NEW DELHI

Published *1954*
Reprinted *1957*

PREFACE

THIS book had its origin in the proposal to revise and reprint the 'Handbook of Statistics for Use in Plant Breeding and Agricultural Problems' by F. J. F. Shaw published by the Indian (then Imperial) Council of Agricultural Research in 1936 but which has been out of print for some years. The present authors were invited by the Council to revise and bring it up to date in order that it should meet the normal requirements of the present-day agricultural research workers. In the course of the revision it became clear that not only would there have to be a change in the order and manner of the presentation of the subject-matter but also a considerable expansion in the scope of the material included in the book, even while keeping within the limits set out by the objective of the book, namely, to provide an elementary text-book for the use of agricultural research workers whose mathematical attainments are modest. As it turned out finally the revision amounted to an almost complete rewriting of the book and addition of several new chapters.

The book now consists of sixteen chapters divided into two parts, the first with six chapters dealing with statistical methods and the second with the design of experiments. The elaboration of the second part is in keeping with the greater emphasis now rightly laid on the proper planning of experiments in order that they may supply efficiently and economically the information sought. The subject-matter has been illustrated throughout with appropriate examples and the appendices would provide material helpful in practical work. With the new arrangement the book commences with a chapter on frequency distribution, graphic representation, averages and measures of dispersion, etc., which were dealt with in the second and third chapters of the previous publication. The material relating to normal curve and probability integral is presented in the second chapter in a more concise and systematic manner than was done in the fourth and fifth chapters of the original. A note on binomial distribution is also included in this chapter on account of its utility in dealing with frequency data, proportions, etc. The third chapter introduces

the idea of sampling and sampling errors and is practically new. Opportunity is taken for an early introduction of the analysis of variance by explaining the technique in an elementary form in the fourth chapter. The fifth chapter is devoted to methods of handling frequency data. Besides giving the various uses of the χ^2 test, the more important methods of estimating linkage including the one based on maximum likelihood are discussed. In presenting methods dealing with the relations between two variables considerable changes have been made and the technique of regression has been given due prominence. It is presented in the sixth chapter. The topics of partial and multiple regression and correlation and intraclass correlation which were not treated in the previous publication are incorporated in the same chapter.

The remaining chapters of the book deal with design of experiments. Apart from randomized block and latin square designs and some ideas relating to soil heterogeneity and analysis of covariance, the entire material in this part is new. Chapter VII introduces the principles of replication, randomization and local control underlying the design of experiments as well as the idea of efficiency of an experiment. Randomized block and latin square designs are discussed in the eighth chapter with the addition of material on size and shape of blocks, number of replications, the process of randomization, etc. Three new chapters, IX to XI, have been added in order to deal with factorial experimentation, confounding and split-plot and strip-plot designs. Analysis of covariance and methods of dealing with incomplete data are presented in Chapter XII. Chapter XIII should be of special interest to plant breeders as it discusses the designs and analysis of replicated progeny row and compact family block trials and introduces incomplete block designs. The important problem of the analysis of groups of experiments has been discussed in Chapter XIV. A special chapter has been devoted to the consideration of practical aspects of field experimentation in order to emphasize their importance in the efficient prosecution of experimental studies. The last chapter, XVI, deals with the problem of conducting experiments on farmers' own land, which is assuming a rapidly growing importance and in which interesting developments are expected.

In view of the restriction on the scope of the book mentioned earlier and in due regard to its size it was not considered desirable to attempt a more exhaustive treatment of the subject-matter or to include more explanatory material concerning the principles discussed. For the benefit of the more enterprising student, who wishes to pursue the study of the subject in greater detail, references to further sources of information concerning various aspects of statistical methods and design of experiments have been given. It is to be hoped that in its present form the book will meet the day to day needs of the agricultural and other biological research workers and form a suitable text for teaching the subject to the students of these sciences.

The authors wish to record their keen appreciation of the careful and painstaking work done by Mr. S. D. Bokil, who was associated throughout with the project of preparing the present text. Their thanks are also due to Mr. V. N. Amble for his help in preparing the drafts of certain new sections and in making a critical scrutiny of several chapters of the manuscript and to Messrs. K. S. Krishnan and K. S. Avadhani for checking the numerical calculations. Lastly, they wish to make a special mention of the help they received from Dr. P. N. Saxena in reading the proofs.

The authors are indebted to Professor Sir Ronald A. Fisher, Cambridge, and to Messrs. Oliver and Boyd Ltd., Edinburgh, for their permission to reprint Tables I and III from their book 'Statistical Methods for Research Workers'.

V. G. PANSE.

New Delhi, January 1954. P. V. SUKHATME.

CONTENTS

PAGE

PREFACE v

PART I. STATISTICAL METHODS

CHAPTER

I. FREQUENCY DISTRIBUTIONS

1a.1 Introduction 1
1b.1 Frequency Distribution 2
1b.2 Graphic Representation 9
1b.3 Frequency Curve and its Characteristics 10
1c.1 Measures of Central Value or Location 10
1c.2 Measures of Dispersion 12
1c.3 The Coefficient of Variation 14
1d.1 Methods of Computation 15

II. THE NORMAL AND BINOMIAL DISTRIBUTIONS

2a.1 The Probability Integral 22
2a.2 The Normal Curve 25
2a.3 Use of the Normal Probability Integral 27
2b.1 Binomial Distribution 32

III. SAMPLING METHOD AND STANDARD ERRORS

3a.1 Introduction 36
3b.1 Random Sampling 36
3b.2 Use of Random Numbers 38
3b.3 Expectation of Sample Estimates . . . 40
3b.4 Requisites of Randomization 40
3c.1 Distribution of Sample Means . . . 42
3c.2 Confidence Limits 44

PAGE

3*d* 1 Standard Error of Sum and Difference of Means 45

3*d*.2 Standard Error of Linear Function of Independent Variates 47

3*d*.3 Sample Estimate of Variance 48

3*d*.4 Standard Error of Estimate of Proportion 48

IV. TESTS OF SIGNIFICANCE OF MEANS AND THEIR DIFFERENCES

4*a*.1 Significance of Means in Large Samples 51

4*a*.2 Significance of Difference of Means in Large Samples 54

4*b*.1 Small Samples: '*t*' Test 57

4*b*.2 '*t*' Test in Paired Samples 62

4*c*.1 Analysis of Variance 63

4*c*.2 The '*F*' Test 66

4*d*.1 Fisher and Behren's '*d*' Test 67

V. THE χ^2 TEST AND ESTIMATION OF LINKAGE

5*a*.1 Introduction 70

5*a*.2 Testing the Significance of Proportion . 70

5*b*.1 The χ^2 Test of Goodness of Fit . . . 72

5*b*.2 Test of Independence: 2×2 Table . . 75

5*b*.3 Test of Independence: $2\times r$ Table . . 79

5*b*.4 Testing Heterogeneity 81

5*b*.5 Test of Independence: $r\times c$ Table . . 83

5*c*.1 Use of χ^2 Test in Genetic Experiments . 83

5*c*.2 Size of Experiments 85

5*c*.3 The Partition of χ^2 86

5*c*.4 Detection of Linkage 88

		PAGE
5*c*.5	Heterogeneity: Linkage Ratio . . .	93
5*d*.1	Estimation of Linkage	93
5*d*.2	Emerson's Method	95
5*d*.3	Method of Maximum Likelihood . . .	96
5*d*.4	Product-Ratio Method	98
5*d*.5	Comparison of the Methods	99

VI. CORRELATION AND REGRESSION

6*a*.1	Introduction	100
6*a*.2	Correlation Table	101
6*a*.3	Coefficient of Correlation	103
6*a*.4	Significance of the Correlation Coefficient	108
6*a*.5	'*z*' Transformation	109
6*b*.1	Regression	113
6*b*.2	Relation between Regression and Correlation	115
6*b*.3	Significance of Regression	116
6*c*.1	Partial Correlation	119
6*c*.2	Significance of Partial Correlation . .	121
6*d*.1	Partial Regression	121
6*d*.2	Significance of Partial Regression . . .	123
6*e*.1	Curvilinear Regression	127
6*f*.1	Intraclass Correlation	131

PART II. DESIGN OF EXPERIMENTS

VII. PRINCIPLES OF FIELD EXPERIMENTATION

7*a*.1	The Problem	137
7*a*.2	A Uniformity Trial	137
7*a*.3	Replication	142

			PAGE
	7*a*.4	Randomization	143
	7*a*.5	Local Control	148
	7*a*.6	Efficiency of Designs	150
VIII.	RANDOMIZED BLOCKS AND LATIN SQUARE		
	8*a*.1	Randomized Blocks	152
	8*a*.2	Critical Difference	155
	8*a*.3	Number of Replications	157
	8*a*.4	Shape and Size of Blocks and Plots: Efficiency	159
	8*b*.1	Latin Square	161
	8*c*.1	Randomized Blocks *Versus* Latin Square	165
IX.	FACTORIAL EXPERIMENTS		
	9*a*.1	The Factorial Concept	166
	9*a*.2	Advantages of a Factorial Experiment .	174
X.	CONFOUNDING		
	10*a*.1	Control of Soil Heterogeneity . . .	176
	10*b*.1	A 2^4 Confounded Experiment . . .	177
	10*b*.2	Symbolical Representation of Treatment Combinations	179
	10*b*.3	Symbolical Representation of Main Effects and First Order Interactions .	180
	10*b*.4	Symbolical Representation of Second and Higher Order Interactions	183
	10*c*.1	Principles and Technique of Confounding	184
	10*d*.1	Analysis of a Confounded Design of 2^n Type	185
	10*d*.2	Efficiency	187
	10*e*.1	Confounding with Factors at 3 Levels .	189

PAGE

10*e*.2 Technique of Confounding in a 3^3 Experiment 189

10*e*.3 Statistical Analysis of a Confounded 3^3 Experiment 193

10*e*.4 Efficiency 197

XI. SPLIT-PLOT AND STRIP-PLOT DESIGNS

11*a*.1 The Split-Plot Design 199

11*a*.2 Advantages of the Design 202

11*a*.3 Intraclsss Correlation between Sub-Plots 202

11*a*.4 Analysis of Variance 203

11*a*.5 Efficiency 206

11*a*.6 Standard Errors of Comparisons . . 207

11*a*.7 Extension of the Split-Plot Design . . 210

11*b*.1 The Strip-Plot Design 211

11*b*.2 Statistical Analysis 212

11*b*.3 Standard Errors of Comparisons . . 216

XII. ANALYSIS OF COVARIANCE AND MISSING PLOT TECHNIQUE

12*a*.1 Use of Ancillary Information . . . 217

12*a*.2 Computations for the Analysis of Covariance 221

12*a*.3 Adjustment for Regression and Test of Significance 222

12*a*.4 Individual Comparisons 224

12*a*.5 Efficiency of Covariance Procedure . . 227

12*a*.6 Extension of the Technique of Analysis of Covariance 228

12*b*.1 Analysis of Incomplete Observations . 229

12*b*.2 Bartlett's Technique for Missing Plots . 230

PAGE

12*b*.3 Adjustment of Averages 232

12*b*.4 Standard Errors of Comparisons . . 234

12*c*.1 Method of Substitution for the Missing Value 234

12*c*.2 Extension of the Covariance Technique to More than One Missing Value . . 236

12*c*.3 The Iterative Method 237

12*d*.1 Missing Plot in the Latin Square . . 239

XIII. Designs for Plant Breeding Trials

13*a*.1 Introduction 241

13*a*.2 Replicated Progeny Rows 241

13*a*.3 Compact Family Blocks 245

13*a*.4 Efficient Utilisation of Seed 250

13*b*.1 Incomplete Block Designs 253

13*b*.2 The Simple Lattice Design 254

13*b*.3 The Statistical Analysis of Incomplete Block Designs 255

13*b*.4 Analysis of a Simple Lattice Design. . 257

13*b*.5 Concept of Balancing 267

13*b*.6 The Balanced Lattice 268

13*b*.7 Analysis of a Balanced Lattice Design . 272

13*b*.8 The Use of Incomplete Block Designs . 278

XIV. Groups of Experiments

14*a*.1 The Problem 280

14*b*.1 Test of Homogeneity of Experimental Errors and Treatment Comparisons . 281

14*b*.2 Heterogeneity of Interactions 286

PAGE

14c.1 Test of Interaction and Method of Combining Results 288

14c.2 Test of Single Degree of Freedom in the Absence of Interaction 293

XV. PRACTICAL CONSIDERATIONS IN FIELD EXPERIMENTATION

15a.1 Introduction 298

15a.2 Objectives and Scope of Experiment . . 298

15a.3 Observation Plots 301

15a.4 Choice of Experimental Design . . . 301

15a.5 Uniformity of Site 302

15a.6 Size, Shape and Arrangement of Plots . 303

15a.7 Arrangement of Blocks 306

15a.8 Modifications of the Usual Layouts . . 308

15a.9 Conducting the Experiment 309

15a.10 Ancillary Observations and Sampling . 310

15a.11 Analysis and Interpretation of Results . 311

15a.12 Transformations 315

XVI. EXPERIMENTS IN CULTIVATORS' FIELDS

16a.1 The Place of Experiments in Cultivators' Fields in Research 317

16a.2 Practical Difficulties 317

16a.3 Need for using the Principle of Random Sampling in Selecting Sites 319

16a.4 The Design of Experiment 321

16a.5 The Number of Experimental Fields and its Distribution Between and Within Places 323

16a.6 Choice of Treatments 326

PAGE

16*a*.7 Auxiliary Observations 330

16*a*.8 The Analysis of Data 331

REFERENCES 337

APPENDICES

I. Interpolation 339

II. Random Numbers 343

III. Values of Normal Deviate (x) for Chosen Levels of P 348

IV. Values of t for $P = \cdot 05$ and $P = \cdot 01$. 349

V. Values of F (Variance Ratio) 350

VI. Values of χ^2 352

VII. Notes on Computation 354

SUBJECT INDEX 357

PART I

STATISTICAL METHODS

CHAPTER 1

FREQUENCY DISTRIBUTIONS

1*a*.1 INTRODUCTION

STATISTICS is a science of collecting numerical observations and interpreting them. When the observations deal with organisms or living things the science is called *biometry*. The part dealing with the collection of observations is known as the *design of experiments*. That dealing with interpretation is termed *statistical methods*. The principal aim of the latter is to summarize data which by their very bulk are difficult to comprehend and which, therefore, need to be analyzed by appropriate methods before they can be understood or made to supply useful information. We shall deal with the latter first.

To start with, we have to distinguish between two kinds of data, concerning *qualitative* and *quantitative* characters. In the case of qualitative characters, the individuals comprising the material under consideration are distinguished by some quality or attribute. Colour in flowers or smooth or wrinkled nature of surface in peas are instances of qualitative characters. In the case of quantitative characters, the individuals are distinguished by a measurement or count. Thus we have the quantitative characters, height of a plant, yield of a crop, number of petals in a flower, etc. In quantitative characters the variation may be *continuous* or *discrete*. A quantity that varies from individual to individual, is called a *variate* and the aggregate of individual values or the individuals themselves from which these are obtained is called a *population.*

In continuous variation the variate can take any value in the range of variation while in discrete variation the variate can assume only certain values in the range, usually integral values. Height or yield are instances of the former kind of variation while number of petals in a flower or number of abdominal hair in drosophila are instances of the latter kind. However, methods of treatment in the case of discrete variation are the same as those

for continuous variation if the different values that the variate can take are sufficiently numerous. This is generally the case; hence the distinction between continuous and discontinuous variates is of no importance at this stage.

1*b*.1 Frequency Distribution

With either kind of data, whether qualitative or quantitative, we generally have records consisting of observations corresponding to each individual in the material under study. By merely inspecting these individual records we cannot hope to gain any definite knowledge for the material as a whole concerning the character measured. It is necessary to put some order in the data before we can derive useful information from them and the first step to take is to classify the observations and study their resulting distribution. Thus, if it is desired to study the occurrence of flower colour in a population of linseed plants, the attribute to be observed would be the flower colour of individual plants. The observer would be confronted at the conclusion of his observations with a long list of records of flower colour of different plants and his first task would be to classify the plants according to the colour of flower and count the number in each class. The number occurring in each class is termed the *frequency* of that class. The manner in which the frequencies are distributed over the different classes is called the *frequency distribution* of the character under study. Table 1.1 gives such a classification of an F_2 population of a linseed cross.

Table 1.1

Distribution of petal colour in F_2 *population of linseed*

Class (1)	Frequency (2)
Blue	169
Lilac	61
White	62
Pink	22
Total	314

In this case the population of 314 plants which comprise the material under study has been divided into four classes as shown in column 1 of the table. Column 2 shows the distribution of the 314 plants among the classes. The division of the population into classes is natural, since each class is characterized by an attribute clear and distinct from that characterizing any other class. The frequency distribution of a quantitative character, on the other hand, necessitates an arbitrary classification of the observations under study. If, for instance, it is desired to measure the length of earhead of a variety of wheat, the observer will obtain a large number of measurements from a field of wheat and these measurements may take any value within a certain range. His first task will be to classify these data with the object of reducing them to a form in which they can be conveniently handled. The classification of such a mass of data involves the division of the range of variation into a number of class intervals and recording the number of individual observations falling in each class. Table 1.2 includes the data on length of earhead to the nearest tenth of a centimetre recorded for 400 earheads of Pusa 12 wheat and Table 1.3 shows the frequency distribution grouped into 17 classes.

TABLE 1.2

Data of length of earhead and number of grains per ear in Pusa 12 wheat, 1930–31

Serial number	Length of earhead	Number of grains per earhead	Serial number	Length of earhead	Number of grains per earhead
1	10·0	25	16	10·0	30
2	10·7	36	17	9·6	26
3	9·8	27	18	8·8	27
4	10·6	32	19	11·8	42
5	11·2	32	20	9·5	26
6	9·2	27	21	9·5	26
7	9·2	31	22	9·4	27
8	11·2	36	23	10·0	28
9	8·5	26	24	10·3	28
10	9·7	32	25	7·5	14
11	12·5	42	26	8·4	26
12	8·3	22	27	11·6	39
13	9·8	32	28	11·1	34
14	10·4	28	29	9·5	26
15	10·8	33	30	11·3	44

TABLE 1.2—*Contd.*

Serial number	Length of earhead	Number of grains per earhead	Serial number	Length of earhead	Number of grains per earhead
31	8·9	24	84	10·8	28
32	9·4	27	85	10·7	30
33	10·2	35	86	10·5	32
34	11·4	32	87	9·5	31
35	9·9	36	88	9·1	37
36	10·2	30	89	9·1	27
37	10·1	28	90	12·1	42
38	10·4	32	91	12·5	43
39	9·9	36	92	9·0	32
40	9·4	28	93	11·2	39
41	8·9	26	94	9·7	30
42	12·0	38	95	11·0	32
43	13·0	41	96	11·1	32
44	9·7	28	97	9·7	26
45	8·9	25	98	8·9	25
46	7·9	19	99	10·1	29
47	8·9	24	100	11·4	42
48	9·5	30	101	10·4	31
49	11·0	35	102	9·7	40
50	9·8	27	103	9·7	28
51	10·2	38	104	10·4	27
52	8·9	20	105	11·0	40
53	9·8	26	106	11·2	31
54	8·4	24	107	11·5	33
55	8·8	23	108	10·9	31
56	9·0	28	109	13·6	41
57	9·5	29	110	11·0	40
58	13·4	48	111	9·5	23
59	10·5	33	112	11·1	36
60	7·7	17	113	10·5	33
61	10·1	30	114	9·8	31
62	8·2	29	115	12·2	36
63	11·4	38	116	8·5	28
64	7·5	20	117	8·4	33
65	9·6	34	118	9·3	29
66	6·6	17	119	9·1	28
67	9·9	29	120	10·6	29
68	12·7	42	121	11·5	30
69	10·2	38	122	10·5	33
70	9·6	27	123	10·4	24
71	10·0	28	124	10·4	34
72	11·7	38	125	9·4	23
73	8·7	34	126	9·0	22
74	8·4	30	127	9·4	22
75	10·0	32	128	11·3	35
76	8·3	21	129	11·8	34
77	7·9	20	130	9·6	29
78	10·0	32	131	8·4	23
79	10·9	26	132	8·7	26
80	9·7	32	133	8·5	25
81	12·5	34	134	11·4	40
82	10·5	40	135	10·8	34
83	11·3	40	136	10·6	30

TABLE 1.2—*Contd.*

Serial number	Length of earhead	Number of grains per earhead	Serial number	Length of earhead	Number of grains per earhead
137	10·5	35	190	10·6	38
138	9·9	23	191	8·8	26
139	6·7	17	192	9·3	29
140	9·5	26	193	13·3	38
141	7·0	17	194	11·5	37
142	10·6	28	195	11·3	37
143	11·2	39	196	10·0	30
144	11·5	45	197	10·7	30
145	8·3	21	198	9·8	26
146	8·4	25	199	6·7	15
147	7·0	15	200	8·5	24
148	9·6	25	201	9·7	31
149	11·7	33	202	11·7	44
150	8·9	28	203	8·9	23
151	10·9	32	204	11·6	30
152	7·8	18	205	11·7	36
153	10·0	29	206	9·2	31
154	11·9	37	207	10·2	31
155	8·7	27	208	10·6	32
156	9·4	28	209	10·4	29
157	11·7	32	210	11·8	41
158	13·1	45	211	10·6	39
159	9·6	32	212	10·3	30
160	10·2	40	213	10·7	26
161	10·4	37	214	11·8	32
162	9·5	23	215	11·8	33
163	12·0	40	216	9·8	30
164	10·6	30	217	9·9	35
165	8·7	25	218	9·6	25
166	10·9	31	219	9·6	31
167	12·2	39	220	9·6	24
168	8·0	14	221	13·7	51
169	12·8	50	222	9·9	26
170	9·7	25	223	5·7	15
171	6·5	17	224	9·2	18
172	11·5	35	225	9·4	32
173	9·8	31	226	10·7	45
174	10·0	27	227	8·8	27
175	9·8	32	228	8·8	25
176	9·4	21	229	11·1	32
177	6·3	17	230	10·4	27
178	10·9	29	231	11·5	32
179	7·6	21	232	11·9	44
180	7·6	23	233	10·5	40
181	10·1	33	234	7·1	19
182	8·9	24	235	10·8	30
183	10·9	35	236	9·5	30
184	8·9	31	237	8·6	35
185	9·4	31	238	12·2	49
186	5·5	13	239	11·6	33
187	6·3	16	240	10·7	42
188	10·1	29	241	9·0	29
189	8·6	33	242	8·6	26

TABLE 1.2—*Contd.*

Serial number	Length of earhead	Number of grains per earhead	Serial number	Length of earhead	Number of grains per earhead
243	10·1	32	296	10·0	31
244	11·2	23	297	9·3	30
245	9·3	25	298	13·5	54
246	11·2	38	299	11·0	31
247	7·9	19	300	10·2	30
248	9·0	26	301	8·7	23
249	9·4	29	302	10·2	30
250	7·4	17	303	12·0	41
251	10·0	31	304	10·4	35
252	8·4	23	305	12·0	42
253	10·6	29	306	11·9	33
254	10·6	39	307	8·5	23
255	9·2	34	308	6·4	14
256	10·5	34	309	11·4	41
257	9·4	26	310	9·8	28
258	10·0	27	311	10·2	28
259	9·3	32	312	10·9	38
260	11·2	38	313	10·1	33
261	8·7	27	314	12·1	41
262	9·6	34	315	9·5	34
263	9·2	33	316	10·4	36
264	9·3	27	317	9·9	39
265	9·5	34	318	9·2	27
266	9·8	29	319	11·3	38
267	10·0	27	320	8·8	26
268	11·2	31	321	8·4	36
269	11·0	34	322	8·7	22
270	11·6	38	323	9·9	26
271	10·0	37	324	12·1	42
272	13·2	39	325	10·5	33
273	7·6	18	326	7·0	21
274	10·1	34	327	11·5	35
275	11·4	36	328	9·7	28
276	12·0	35	329	8·2	21
277	10·0	40	330	9·6	29
278	10·3	34	331	11·6	35
279	10·5	40	332	8·4	23
280	10·9	43	333	9·5	27
281	10·8	34	334	8·6	33
282	9·0	31	335	12·9	40
283	10·4	28	336	9·2	26
284	12·3	43	337	10·2	29
285	9·7	30	338	10·3	32
286	9·1	20	339	12·4	33
287	10·0	33	340	9·4	30
288	8·8	28	341	8·9	25
289	9·8	28	342	9·9	24
290	9·7	34	343	9·8	26
291	5·4	10	344	6·9	21
292	10·1	29	345	11·8	36
293	8·3	28	346	13·1	32
294	9·0	25	347	9·4	25
295	10·7	38	348	10·5	31

TABLE 1.2—*Contd.*

Serial number	Length of earhead	Number of grains per earhead	Serial number	Length of earhead	Number of grains per earhead
349	13·6	39	375	10·8	36
350	8·7	21	376	10·5	35
351	10·2	31	377	11·8	35
352	12·1	41	378	10·4	31
353	6·2	16	379	10·4	39
354	10·8	35	380	9·5	27
355	10·9	32	381	11·8	42
356	11·0	34	382	9·8	30
357	10·5	34	383	8·5	23
358	10·0	29	384	10·7	35
359	11·5	35	385	10·0	30
360	9·2	28	386	10·7	30
361	8·4	24	387	9·2	24
362	11·8	38	388	10·8	32
363	9·7	28	389	10·3	36
364	9·0	27	390	8·2	22
365	6·3	13	391	9·5	34
366	10·3	41	392	6·8	19
367	8·6	27	393	10·6	36
368	10·2	32	394	10·0	31
369	8·0	23	395	12·4	39
370	10·0	31	396	8·1	24
371	10·4	33	397	7·5	21
372	11·4	33	398	10·3	32
373	8·9	29	399	8·8	27
374	11·5	38	400	10·2	31

In this example, the shortest earhead measures 5·4 cm. and the longest measures 13·7 cm. We, therefore, have a range of 5·3 to 13·7 cm. which can be divided into 17 classes by taking class intervals of 0·5 cm. The mid-point of each class is taken as the *class value* and the frequency of the class consists of the number of observations falling within the limits of that class.

The student must exercise caution in classifying those observations whose values fall on the limits of each class-range. For instance, the first class includes observations of 5·3 cm. or above upto and including observations of 5·7 cm., values of 5·8 cm. fall into the second class and similarly values of 6·3 cm. fall into the third class.

It will be seen that the distribution is marked by low frequencies in the extreme classes. The frequencies increase

TABLE 1.3

Frequency distribution of length of earhead in Pusa 12 wheat

Class (1)	Class value (2)	Frequency (3)	Frequency × Class value (4)
5·3– 5·7	5·5	3	16·5
5·8– 6·2	6·0	1	6·0
6·3– 6·7	6·5	8	52·0
6·8– 7·2	7·0	6	42·0
7·3– 7·7	7·5	8	60·0
7·8– 8·2	8·0	11	88·0
8·3– 8·7	8·5	32	272·0
8·8– 9·2	9·0	42	378·0
9·3– 9·7	9·5	58	551·0
9·8–10·2	10·0	65	650·0
10·3–10·7	10·5	55	577·5
10·8–11·2	11·0	37	407·0
11·3–11·7	11·5	31	356·5
11·8–12·2	12·0	24	288·0
12·3–12·7	12·5	7	87·5
12·8–13·2	13·0	6	78·0
13·3–13·7	13·5	6	81·0
Total	..	400	3991·0

gradually as one approaches the middle of the distribution giving the distribution a symmetrical appearance.

In choosing a class interval certain important considerations should be borne in mind.

(1) The class interval should be of uniform width and of such size that the characteristic features of the distribution are displayed. Thus, the class interval must not be so large that a considerable error would be involved in assuming that the mid-point of the interval is the average value of the class. It must not be so small as to give too many classes with zero or very small frequencies.

(2) The range of the classes should cover the entire range of the data and the classes must be continuous.

(3) As a general rule, the number of classes should be about 15 and never more than 30 nor less than 6.

(4) It is convenient to make the mid-point of a class a whole number.

1*b*.2 Graphic Representation

The information contained in Table 1.3 can also be expressed as a graph and indeed this method permits of a ready grasp of certain important features which are common to some types of frequency distributions. The graph is obtained by plotting class values as abscissae and class frequencies as ordinates. The curve obtained from the data in Table 1.3 is shown in Fig. 1.1 and depicts clearly the features of the distribution, namely, that it has

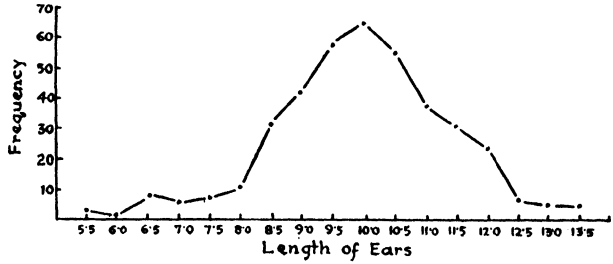


Fig. 1.1. Frequency polygon of length of 400 earheads of Pusa 12 wheat.

the maximum frequencies at the middle of the range and the class frequencies diminish more or less symmetrically in the direction of the two extremes. This curve is called a *frequency polygon.*

Another graphical method of depicting frequency distributions, illustrated in Fig. 1.2, is to measure along the horizontal axis

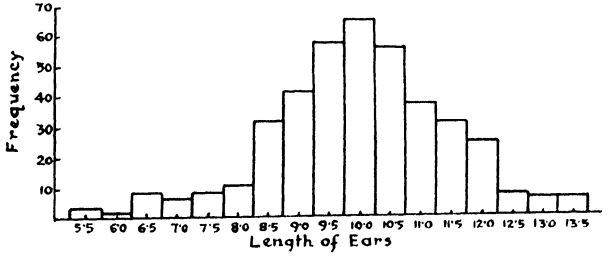


Fig. 1.2. Histogram showing the distribution of the length of 400 earheads of Pusa 12 wheat.

distances proportional to the class intervals and to raise on each of these distances, columns or rectangles proportional in

height to the number of individuals falling within the class. The resulting figure is called a *histogram.*

1*b*.3 Frequency Curve and its Characteristics

The above frequency graph for any data, whether it is a frequency polygon or a histogram, approaches more and more the form of a smooth curve as the number of observations increases and finer class intervals are used. For an hypothetically infinite body of data, the graph should preserve its regularity however narrow the class interval may be and, therefore, with an infinitely narrow class interval we expect to get a perfectly smooth curve. Frequency curves can be distinguished from one another by means of four characteristics: the central value, the spread of the curve around the central value, the symmetry or the departure from it termed *skewness* and the excess or deficiency of frequencies in the centre and the two extremes compared with the flanks termed the *kurtosis.* These characteristics can be measured from the data and the measures are sometimes called the constants of the distribution. For the majority of biological characters the frequency distributions approximate to a symmetrical bell-shaped curve known as the *normal curve.* For such distributions only the first two characteristics, namely, the central value and the spread, also called *dispersion*, are important. Methods of measuring these are described in the following sections.

1*c*.1 Measures of Central Value or Location

There are three well-known measures of the central tendency of a frequency distribution; these are the *mean*, the *mode* and the *median.*

(1) *Mean*

The mean is the arithmetic average and is the result obtained when the sum of the values of the individuals in the data is divided by the number of individuals in the data.

The mean is usually denoted by the symbol μ and given by

$$\mu = \frac{\Sigma x}{N} \qquad (1.1)$$

where Σ represents the summation which is taken over all the N observations in the population. When the data are expressed

in a frequency distribution, the mean is calculated from the formula

$$\mu = \frac{\Sigma f.x}{N} \qquad (1.2)$$

where

f = class frequency,

x = the class value,

N = the size of the population

and

Σ indicates summation of the products f with x over all classes.

This formula assumes that all the observations in any class are concentrated at the middle of the class interval. This assumption is of course not strictly true and the formula may therefore give results different from those obtained directly from the individuals without grouping, but the difference is generally negligible.

Substituting in formula (1.2) the values in the example in Table 1.3, we obtain

$$\mu = \frac{3991 \cdot 0}{400} = 9 \cdot 9775 \text{ cm.}^*$$

The mean in this case, therefore, is 9·9775 cm. Calculating the mean directly from the ungrouped data of Table 1.2, we find that the mean is

$$\mu = \frac{3990 \cdot 9}{400} = 9 \cdot 97725 \text{ cm.}^*$$

The difference is obviously negligible.

(2) *Mode*

The mode is that value of the variate which occurs most frequently. In a frequency table the modal class is the class which has the greatest frequency. This class can be determined at once from inspection, but the actual value of the mode will be located somewhere in that class interval, not necessarily at the mid-point of the class. In Table 1.3 the class with mid-value 10·0 is the modal class.

---

* More decimal places have been retained in this calculation than is warranted by the data for the purpose of illustration.

(3) *Median*

The median is the value which is located in the middle of a series when the observations are arranged in order of magnitude and it divides the series into two equal halves, half the number of the observations lying above it and half below. The determination of the median is a simple matter when there are an odd number of observations in the series. Thus, if 101 observations are placed in the order of their magnitude, the 51st observation will be the value of the median. If there are an even number of observations in a series, the average of the two central values may be taken as the median. In a frequency distribution, such as the one we have described for length of earhead of wheat, the median will be the value of the abscissa at which the ordinate divides the area under the curve into two equal halves.

Of these measures of central tendency the arithmetic mean is by far the most important and commonly used. The mode and the median are easier to find than the mean; but the arithmetic mean is preferred partly because of the ease with which it can be employed in the various statistical procedures to which data are commonly subjected and partly because unlike the mode and the median, all the observations enter into its calculation.

1*c*.2 MEASURES OF DISPERSION

The mean gives us an idea of the central value around which the individual observations are distributed; but it tells us nothing of how they are distributed. Thus each of the following five series:

(1)	7	7	7	7	7		
(2)	5	6	7	8	9		
(3)	4	5	6	8	9	10	
(4)	1	1	2	2	3	5	35
(5)	2	12					

has 7 as the mean, although the pattern of individual observations is different in the different series. The following measures are available for the measurement of spread or dispersion of the individual observations in the population.

(1) *Range*

The *range* of a distribution is the difference between the largest and the smallest of observations and gives us some idea of the amount of variability present. In our wheat example the range is from 5·4 cm. to 13·7 cm. for the length of earhead, that is, 8·3 cm.

(2) *Mean deviation*

Another measure of dispersion is provided by the *mean deviation*. This is calculated by adding the deviations of individual observations from their arithmetic mean without regard to the sign and dividing the sum by the number of observations. The formula for calculating the mean deviation for data classified in a frequency distribution is

$$\text{mean deviation} = \frac{\Sigma f.d}{N} \tag{1.3}$$

where

f = class frequency,

d = deviation of the mid-value of the class from the population mean, always taken positive,

N = total number of observations

and the summation Σ is taken over all the classes of the distribution.

(3) *Standard deviation*

This measure of dispersion is calculated by squaring the deviation of each observation from the mean, adding the squares, dividing by the number of observations and extracting the square root according to the formula

$$\sigma = \sqrt{\frac{\Sigma d^2}{N}} \tag{1.4}$$

where σ stands for the *standard deviation* and where the summation is taken over all the N deviations. When the data are grouped in the form of a frequency distribution the formula takes the form

$$\sigma = \sqrt{\frac{\Sigma f.d^2}{N}} \tag{1.5}$$

Since the deviations, both positive and negative, are squared, the products $f.d^2$ are always positive.

Unlike the mean, the difference between the two formulæ is not always negligible and a correction is required in computing the standard deviation from grouped data with the help of formula (1.5). This correction is dealt with in Section 1*d*.1.

It should be remembered that the measures of dispersion are expressed in the same units of measurement such as inches, grammes, pounds, etc., in which the observations themselves are measured.

The term *variance* is used to denote the square of the standard deviation, *i.e.*, σ^2. We have

$$\sigma^2 = \frac{\Sigma f.d^2}{N} \qquad (1.6)$$

Of these measures of dispersion, σ is the one most commonly used. The range, though it provides some indication of the spread and is the simplest to work out, suffers from the defect that it is based on only two out of the whole bulk of observations. For this reason it does not adequately reflect the information regarding spread contained in the data unless we are dealing with small sets of observations.

The mean deviation is, perhaps, a simpler measure of variability than the standard deviation, but is not easily amenable to algebraic treatment in the way the standard deviation is. In fact, most methods of statistical analysis have been evolved round the square of the standard deviation or the variance.

1*c*.3 The Coefficient of Variation

The above measures of dispersion are absolute, being expressed in units in which the observations are recorded. A measure of variation which is independent of the unit of measurement and is therefore useful for comparison between different populations is provided by the standard deviation expressed as a percentage of the mean. This measure is known as the *coefficient of variation* and is given by

$$C.V. = \frac{\sigma}{\mu} \times 100 \qquad (1.7)$$

Applying this formula to our example of the length of earheads in Pusa 12 wheat, we have

$$\mu = 9{\cdot}978 \text{ cm.}$$

$$\sigma = 1{\cdot}441 \text{ cm.}$$

$$\therefore \quad C.V. = \frac{1{\cdot}441 \times 100}{9{\cdot}978} = 14{\cdot}4 \text{ per cent.}$$

We say that the coefficient of variation is about 14 per cent.

The usefulness of the coefficient of variation can be illustrated on the following figures showing the mean yield of paddy determined by harvesting and weighing the produce of a large number of plots of uniform size, the standard deviation and the coefficient of variation for two districts.

District	Mean yield in lb. per acre	Standard deviation in lb. per acre	Coefficient of variation
Tanjore (Madras)	1324·5	140·5	10·61
Raipur (Madhya Pradesh)	1069·6	124·2	11·61

The table brings out the importance of the coefficient of variation as a relative measure of variation. Thus, though the actual standard deviation is greater for the Tanjore district, the coefficient of variation and hence the relative variation of yield of paddy in this district is smaller than in Raipur district.

1*d*.1 Methods of Computation

The ordinary method

This method is simply the straightforward application of the formulæ given in the preceding pages. Table 1.4 gives the details of the calculations for the number of grains per ear in Pusa 12 wheat. The data from Table 1.2 are first arranged in a frequency distribution as shown in columns 1, 2 and 3 of Table 1.4. For the character the number of grains per ear, we are dealing with a discrete variate taking only integral values going from 8 to 57. This range has been divided into 10 classes with

a class interval of 5 with the limits 8–12, 13–17, 18–22, ..., etc. Column 4 gives the products of the frequency with the class value and columns 5 to 7 deal with the deviation, d, of each class value from the mean and the products $f.d^2$. The standard deviation is calculated from the formula (1·5); but a correction factor has to be applied to compensate for the error introduced by grouping the observations into classes and basing the calculations on the values of class centres. This correction is called *Sheppard's Correction.* The Sheppard's correction is 1/12 of the square of the class interval and is to be deducted from $\Sigma f.d^2/N$ as is shown in the following example.

TABLE 1.4

Calculation of mean and standard deviation of number of grains per earhead in Pusa 12 wheat by ordinary method

Class	Class value	Frequency	Frequency × Class value	Deviation from the mean	Deviation squared	Frequency × Deviation squared
	x	f	$f.x$	d	d^2	$f.d^2$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
8–12	10	1	10	−20·55	422·3	422
13·17	15	17	255	−15·55	241·8	4111
18–22	20	25	500	−10·55	111·3	2782
23–27	25	86	2150	− 5·55	30·8	2649
28–32	30	125	3750	− 0·55	0·3	38
33–37	35	77	2695	4·45	19·8	1525
38–42	40	55	2200	9·45	89·3	4912
43–47	45	9	405	14·45	208·8	1879
48–52	50	4	200	19·45	378·3	1513
53–57	55	1	55	24·45	597·8	598
Total	..	N=400	12220	..	..	20429

$$\mu = \frac{\Sigma f.x}{N} = \frac{12220}{400} = 30{\cdot}55$$

$$\sigma = \sqrt{\frac{\Sigma f.d^2}{N} - \frac{1}{12}(I)^2}, \; I \text{ being the class interval}$$

$$= \sqrt{\frac{20429}{400} - 2{\cdot}0833}$$

$$= 6{\cdot}999$$

Without Sheppard's correction, we would have obtained

$$\sigma = \sqrt{\frac{20429}{400}} = 7 \cdot 146$$

By direct calculation from deviations of individual observations without grouping we get

$$\sigma = \sqrt{\frac{19198 \cdot 76}{400}} = 6 \cdot 928$$

It will be seen that the correction has considerably reduced the error due to grouping.

The frequency curve of the distribution of the number of grains per earhead is shown in Fig. 1.3.

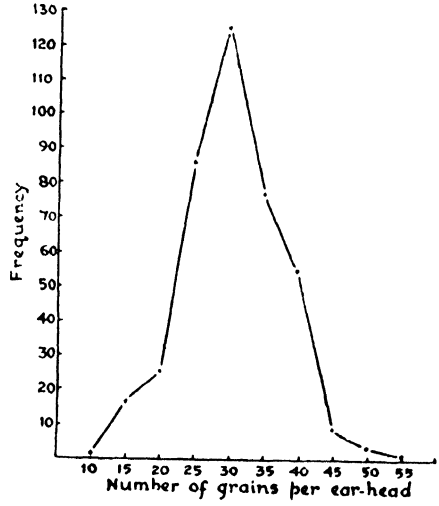


FIG. 1.3. Frequency polygon of the number of grains per earhead in Pusa 12 wheat.

Sheppard's correction is considered to be chiefly useful in distributions which tail off slowly.

Short Method

The heavy calculations of the preceding example may be considerably reduced by the use of the short method illustrated

in Tables 1.5 and 1.6. In this method a convenient assumed value, which must agree with one of the class values, is taken as the origin and is referred to by the letters *A.O.* meaning arbitrary origin. Thus in Table 1.5 the class value of the first class is taken as the arbitrary origin. The deviations, d', from this arbitrary origin are taken as shown in column 4 in Table 1.5 and are expressed in terms of the number of class intervals without reference to actual class values. This is known as *coding*. Thus the deviation d' of the last class is given as 9 because this class is 9 class intervals removed from the class which contains the arbitrary origin. For this last class the frequency is 1 and therefore, $f.d' = 1 \times 9$ and $f.d'^2 = 1 \times 9^2$. The products $f.d'$ and $f.d'^2$ are given in columns 5 and 6. Since the calculations have been made on deviations from the arbitrary origin and not the true mean and since the deviations are expressed as numbers of class intervals, corrections have to be applied to the arbitrary origin to determine the mean and to the usual formula for the standard deviation to determine the latter in the proper units. This is *decoding*. Thus

$$\mu = A.O. + I\left(\frac{\Sigma f.d'}{N}\right) \tag{1.8}$$

and

$$\sigma = \sqrt{\left\{\frac{\Sigma f.d'^2}{N} - \left(\frac{\Sigma f.d'}{N}\right)^2\right\}(I)^2 - \frac{1}{12}(I)^2} \tag{1.9}$$

A word of caution seems necessary in applying Sheppard's correction. This correction is 1/12th the value of the square of the class interval and must invariably be deducted from the variance as shown in formula (1.9).

In the above example the class value of the first class has been taken as the arbitrary origin and all deviations, d', are positive; this, however, is not necessary and a class value in the middle or in any other part of the distribution may be taken as the arbitrary origin, in which case the lower classes will give negative deviations and the higher classes will give positive deviations. It is always convenient to take the arbitrary origin at or near the middle of the distribution, as will be seen from Table 1.6 which gives the calculations of the mean and standard

deviation for the length of earhead in Pusa 12 wheat by the short method.

TABLE 1.5

Calculation of mean and standard deviation of number of grains per earhead in Pusa 12 wheat by short method

Class	Class value	Frequency	Deviation from arbitrary origin	Frequency × Deviation	Frequency × Deviation squared
	x	f	d'	$f.d'$	$f.d'^2$
(1)	(2)	(3)	(4)	(5)	(6)
8–12	10	1	0	0	0
13–17	15	17	1	17	17
18–22	20	25	2	50	100
23–27	25	86	3	258	774
28–32	30	125	4	500	2000
33–37	35	77	5	385	1925
38–42	40	55	6	330	1980
43–47	45	9	7	63	441
48–52	50	4	8	32	256
53–57	55	1	9	9	81
Total	..	400	..	1644	7574

$I = 5$

$A.O. = 10$

$$\mu = 10 + 5\left(\frac{1644}{400}\right)$$

$$= 30{\cdot}55$$

$$\sigma = \sqrt{\left\{\frac{7574}{400} - \left(\frac{1644}{400}\right)^2\right\}(5)^2 - \frac{1}{12}(5)^2}$$

$$= 6{\cdot}999$$

$$C.V. = \frac{6{\cdot}999 \times 100}{30{\cdot}55}$$

$$= 22{\cdot}9 \text{ per cent.}$$

TABLE 1.6

Calculation of mean and standard deviation of length of earhead in Pusa 12 wheat by short method

Class	Class value	Frequency	Deviation from arbitrary origin	Frequency × Deviation	Frequency × Deviation squared
	x	f	d'	$f.d'$	$f.d'^2$
(1)	(2)	(3)	(4)	(5)	(6)
5·3– 5·7	5·5	3	−8	−24	192
5·8– 6·2	6·0	1	−7	− 7	49
6·3– 6·7	6·5	8	−6	−48	288
6·8– 7·2	7·0	6	−5	−30	150
7·3– 7·7	7·5	8	−4	−32	128
7·8– 8·2	8·0	11	−3	−33	99
8·3– 8·7	8·5	32	−2	−64	128
8·8– 9·2	9·0	42	−1	−42	42
9·3– 9·7	9·5	58	0	0	0
9·8–10·2	10·0	65	1	65	65
10·3–10·7	10·5	55	2	110	220
10·8–11·2	11·0	37	3	111	333
11·3–11·7	11·5	31	4	124	496
11·8–12·2	12·0	24	5	120	600
12·3–12·7	12·5	7	6	42	252
12·8–13·2	13·0	6	7	42	294
13·3–13·7	13·5	6	8	48	384
Total	..	400	..	−280 +662 +382	3720

I = 0·5 cm.

$A.O.$ = 9·5 cm.

$$\mu = 9{\cdot}5 + 0{\cdot}5\left(\frac{382}{400}\right)$$

$$= 9{\cdot}978 \text{ cm.}$$

$$\sigma = \sqrt{\left\{\frac{3720}{400} - \left(\frac{382}{400}\right)^2\right\}(0{\cdot}5)^2 - \frac{1}{12}(0{\cdot}5)^2}$$

$$= 1{\cdot}441 \text{ cm.}$$

$$C.V. = \frac{1{\cdot}441 \times 100}{9{\cdot}978}$$

$$= 14{\cdot}4 \text{ per cent.}$$

The short-cut methods give the same values for the constants as the ordinary method and not approximations to the results achieved by the latter.

In the preceding examples we were dealing with large populations in which the data are grouped. It happens, however, that the mean and the standard deviations have generally to be calculated from a few observations, in which case the calculation is made directly from the observed data. In that case, the class interval as well as the correction for grouping drop out from the formula (1.9), and $f = 1$ for each observation. Hence the formula reduces to

$$\sigma = \sqrt{\frac{\Sigma d'^2}{N} - \left(\frac{\Sigma d'}{N}\right)^2} \qquad (1.10)$$

If a calculating machine is available it becomes convenient to use zero as the arbitrary origin. Then d' stands for the individual observation x and so the formula becomes

$$\sigma = \sqrt{\frac{\Sigma x^2 - T \cdot \mu}{N}} \qquad (1.11)$$

where T is the total of all the N observations.

This is the formula in general use for machine calculations.

CHAPTER II

THE NORMAL AND BINOMIAL DISTRIBUTIONS

2*a*.1 The Probability Integral

In Chapter I it has been mentioned that as the number of observations used for the construction of the frequency polygon increases and the class interval is reduced the graph approaches more and more the form of a smooth curve known as the frequency curve. The concept of a frequency curve is of great value in statistics, for a frequency curve provides an excellent summary of the data and reflects their characteristics. A little consideration would enable the reader to grasp the fact that the area cut off by any two ordinates and the curve is proportional to the frequency of observations lying between the values of the variate at which the ordinates are erected. For, in the limiting case, the curve is constructed by erecting at the mid-points of narrow class intervals ordinates proportional to the frequencies of observations within the respective classes. Consider the graph drawn in Fig. 2.1 which shows a normal frequency curve having the same mean and variance as the population of earheads, superposed on the histogram in Fig. 1.2.

The curve does not fit the data quite closely, but this is the curve which we may expect to obtain as a limit, when the number of observations is increased and the class interval narrowed down. The fit is seen to be good in certain regions like rectangles *a*, *b*, *g*, *h*, etc., but not in others. If we consider the figure formed by the two ordinates erected at A and A_1 and the curve intercepted by them, that is, the figure $AA''A_1''A_1$, it will be seen that the area of the figure is practically equal to that of the rectangle $AA'A_1'A_1$ which is obtained by erecting at C, the mid-point of AA_1 an ordinate CC' proportional to the frequency in the class interval and drawing at its upper end $A'A_1'$ parallel to AA_1. This can be verified by drawing the figure on graph paper and counting the number of squares in the rectangle $AA'A_1'A_1$ and the figure $AA''A_1''A_1$ reproduced on an enlarged scale in Fig. 2.2.

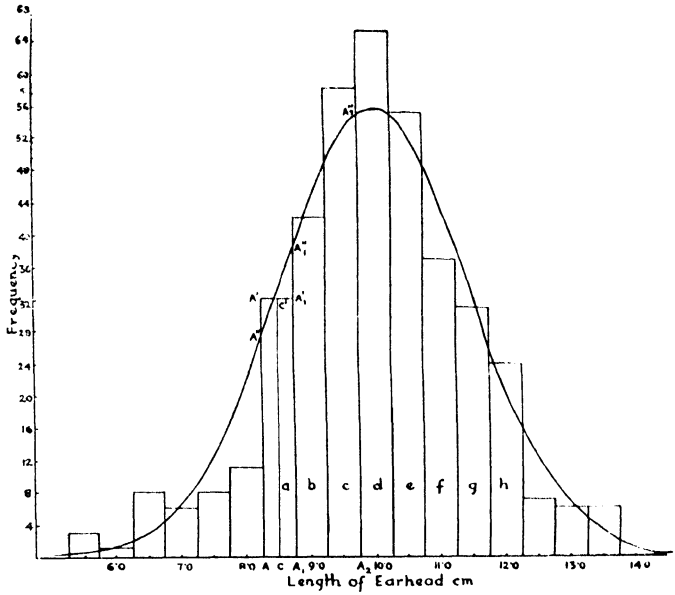


FIG. 2.1. Histogram and the normal curve for population of wheat earheads.

In our present example, this does not hold good of all rectangles; but in the limiting case where the smooth curve is obtained by increasing the number of observations and joining the ends of ordinates raised at mid-points of class intervals which have been made as small as possible, the area under the curve for each class interval will be equal to that of the rectangle erected over the class interval and equal in height to the ordinate raised at its mid-point.

Referring again to Fig. 2.1, a segment such as $AA''A_2''A_2$ can be regarded as made up of several rectangles *a*, *b*, etc., each proportional to the frequency in the class interval, so that the area of the segment would be proportional to the frequency of observations lying in the interval AA_2. This is true of any segment of the curve. This fact is of value, since, given the curve, it enables us to calculate the proportion of observations or what is the same thing the probability of an observation lying between

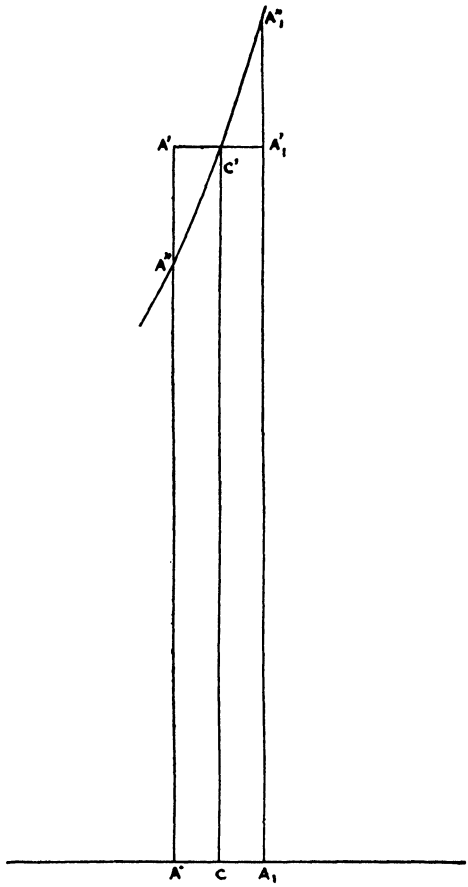


FIG. 2.2. Enlarged drawing of fig. $AA''A_1''A_1$ in Fig 2.1.

any two values or that of its being less than or exceeding a given value of the variate. When a table is constructed showing the proportions of observations lying below the successive values of the variate, it is known as the *cumulative frequency* or *probability integral* table. This in fact is an alternative method of expressing the distribution of a variate. For the normal frequency curve such a table has been calculated and is known as the normal

probability integral table. The use of this table will be explained later.

2*a*.2 THE NORMAL CURVE

The Normal Curve is determined by two constants representing the population, *viz.*, the mean and the standard deviation. In other words, knowing the mean and the standard deviation of a population, the curve can be drawn. Figure 2.3 gives the curves for three normal distributions with different values of μ and σ.

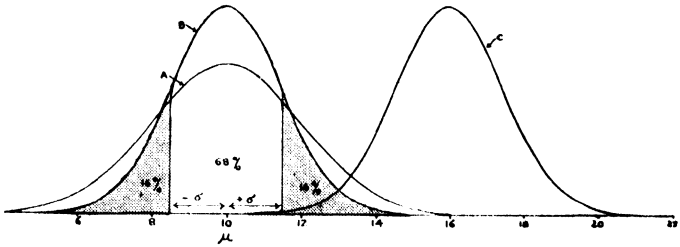


FIG. 2.3. Normal distributions.

The curve *B* is the same as in Fig. 2.1, namely, the one for the data for earhead lengths of wheat, the curve *A* has the same mean value μ as the curve *B* but different σ and the curve *C* has the same standard deviation σ as the curve *B* but different μ. It will be seen that all the three curves are perfectly symmetrical about the means and, therefore, the mean divides the area under the curve into two equal halves. The mean is, therefore, also the median. The maximum ordinate, that is, the peak, is also at the mean. Thus the mean, mode and median coincide in the case of the normal curve. The figure also shows that as we move away from the mean, whether to the left or right, the ordinate and hence the frequency to which it is proportional, diminishes rapidly. This means, the greater the deviation from the mean, the fewer the observations exceeding that value. It will be observed that curves *B* and *C*, which have the same σ but different μ, have the same form and differ only in location, whereas curves *A* and *B*, which have the same μ but different σ, have the same location but *A* has a greater spread. The curves *A* and *C* differ in μ as well as σ and they consequently differ in their location as well as spread.

If we erect two ordinates at a distance σ on both sides of the mean in any of the three curves, the area of the central portion so cut off is about 68 per cent. of the total area (unshaded portion of curve *B*). The area of the end portions or 'tails' taken together (shaded portion of curve *B*) amounts to about $(100 - 68) = 32$ per cent. As the curve is symmetrical, the tails are equal in area, each enclosing nearly 16 per cent. of the total area.

In terms of frequency this means that nearly $\frac{2}{3}$ of the observations lie within the range $\mu \pm \sigma$, about $\frac{1}{6}$ of the total number of observations lie beyond $\mu + \sigma$ and an equal number have values less than $\mu - \sigma$. In other words, only about $\frac{1}{3}$ of the total number of observations deviate from the mean by an amount equal to σ or more. Similarly, if we were to erect ordinates on both sides of the mean at a distance equal to 2σ, the central area so cut off would amount to about 95 per cent. of the total. In other words, only 5 per cent. of the total number of observations deviate from the mean by 2σ or more. In fact, for a normal curve the fraction of the total area thus cut off depends only on the ratio $(x - \mu)/\sigma$, where x stands for a given observation and $x - \mu$ its deviation from the mean. This ratio is called the *normal deviate*. Table 2.1 of the normal probability integral gives for different positive values of the normal deviate the fraction corresponding to the area lying to the left of the ordinate at x. This area represents the probability of a normally distributed variate being less than x and consists of the central portion and one of the tails and is represented by $\frac{1}{2}(1 + \alpha)$, where α is the area of the central portion. For negative values of $x - \mu$, the area to the left of the ordinate at x is given by $\frac{1}{2}(1 - \alpha)$, being obtained by subtracting from unity the area $\frac{1}{2}(1 + \alpha)$ corresponding to the positive value of $x - \mu$. The area or the frequency of observations lying outside the $\pm$ value of the normal deviate is $1 - \alpha$. Table 2.1 gives values of $\frac{1}{2}(1 + \alpha)$ for positive values of the normal deviate differing by 0·1 and shows the normal frequency distribution in a cumulative form.

Values of the fraction for intermediate values of the normal deviate may be obtained by interpolation (Appendix I) or by use of more extensive tables (*Tables for Statisticians and Biometricians*

TABLE 2.1

Normal probability integral showing area of normal curve for different values of normal deviate

Normal deviate $\frac{x-\mu}{\sigma}$	Area to left of ordinate $\frac{1}{2}(1+\alpha)$	Normal deviate $\frac{x-\mu}{\sigma}$	Area to left of ordinate $\frac{1}{2}(1+\alpha)$
0·0	0·50000	1·6	0·94520
0·1	0·53983	1·7	0·95543
0·2	0·57926	1·8	0·96407
0·3	0·61791	1·9	0·97128
0·4	0·65542	2·0	0·97725
0·5	0·69146	2·1	0·98214
0·6	0·72575	2·2	0·98610
0·7	0·75804	2·3	0·98928
0·8	0·78814	2·4	0·99180
0·9	0·81594	2·5	0·99379
1·0	0·84134	2·6	0·99534
1·1	0·86433	2·7	0·99653
1·2	0·88493	2·8	0·99744
1·3	0·90320	2·9	0·99813
1·4	0·91924	3·0	0·99865
1·5	0·93319		

by Karl Pearson). Before illustrating the use of this table, we shall explain in some detail the concept of probability.

2*a*.3 USE OF THE NORMAL PROBABILITY INTEGRAL

We shall first explain the concept of *probability* and its relationship with the frequency distribution. Suppose we have a large number of observations distributed in a distribution of the type we have considered and we note the observations, one on each of a number of tickets similar in all respects. If we mix the tickets well, draw a ticket, record the observation, replace it and repeat the process a fairly large number of times, we will have an aggregate of a number of observations. At the time we draw a ticket we do not know what the observation on the ticket would be, whether it will be greater or smaller than the mean; but since we know that half of the population lies below and half above the mean, we may reasonably expect that in the aggregate of tickets drawn nearly half would lie above and half below the mean. In a sufficiently large number of draws it actually happens that the proportion is close to half provided the

tickets are mixed well, do not stick together and are similar. Similarly, if we know that $\frac{1}{4}$ the members of the population exceed a certain value a then in a large number of draws of the tickets we would expect nearly $\frac{1}{4}$ of the draws to exceed that value and we would find our expectation fulfilled to a fairly close degree over a large number of draws provided the conditions mentioned above are satisfied. The drawing of a ticket is a chance event and for such events the proportions in which we expect the alternative events to happen signify the chance or probability that they will happen. For the type of draws we have just considered, the expected proportions would be the same as the proportions of the population lying in given regions. Hence these latter proportions represent the probability of a variate lying in those regions.

Following examples illustrate the use of the probability integral table.

Example 2.1

The mean length of earhead in a population of earheads of Pusa 12 wheat is 9·978 cm., and the standard deviation is 1·441 cm. What is the probability of occurrence of

(i) a head having a length of 12·128 cm. or more,

(ii) a head of 6·536 cm. or less

and

(iii) a head deviating from the mean by $\pm$ 2·581 cm. or more ?

(i) The first step is to calculate the normal deviate corresponding to the observation of 12·128 cm. We have

$$\frac{x-\mu}{\sigma} = \frac{12\cdot128 - 9\cdot978}{1\cdot441}$$

$$= \frac{2\cdot150}{1\cdot441}$$

$$= 1\cdot492$$

From Table 2.1 we find that the area $\frac{1}{2}(1+\alpha)$ corresponding to the value 1·4 of the normal deviate is 0·91924 and that corresponding to the value 1·5 is 0·93319, that is, a difference of 0·1

in the normal deviate corresponds to a difference of 0·01395 in the area $\frac{1}{2}(1 + \alpha)$.

A difference of 0·092 therefore corresponds to a difference of

$$\frac{0{\cdot}092 \times 0{\cdot}01395}{0{\cdot}1} \text{ or } 0{\cdot}01283$$

We may thus take the area corresponding to the normal deviate 1·492 to be (0·91924 + 0·01283), that is, 0·93207. The area of the tail to the right of 12·128 is therefore equal to (1 − 0·93207) or 0·06793. The probability of the occurrence of an earhead length equal to or greater than 12·128 cm. is therefore about 7 per cent.

In the above method of calculating the area $\frac{1}{2}(1 + \alpha)$ corresponding to any given value of the normal deviate intermediate between the tabulated values, we have assumed that a change in $\frac{1}{2}(1 + \alpha)$ is proportional to the change in the normal deviate over the whole interval between the tabulated values of the normal deviate. In other words, we have assumed a linear relationship between the area and the normal deviate. The method is, therefore, called the method of proportional parts or linear interpolation. The assumption of linearity is not strictly true and so the value obtained is approximate. A more exact method of calculating the intermediate values from such tables is described in Appendix I. For practical purposes linear interpolation is generally sufficient. The student may verify that the result obtained above is correct to four decimal places.

(ii) The value of the normal deviate corresponding to the given observation is

$$\frac{x - \mu}{\sigma} = \frac{6{\cdot}536 - 9{\cdot}978}{1{\cdot}441}$$

$$= -\frac{3{\cdot}442}{1{\cdot}441}$$

$$= -2{\cdot}389$$

The values of $\frac{1}{2}(1 + \alpha)$ corresponding to the values 2·3 and 2·4 of the normal deviate are 0·98928 and 0·99180 respectively. Interpolating between these values we obtain the area

$\frac{1}{2}(1 + \alpha)$ corresponding to 2·389. We have

$$\begin{aligned} \tfrac{1}{2}(1 + \alpha) &= 0{\cdot}98928 + 0{\cdot}00224 \\ &= 0{\cdot}99152 \\ \therefore \quad \alpha &= 2\,(0{\cdot}99152 - 0{\cdot}50000) \\ &= 0{\cdot}98304 \end{aligned}$$

and

$$\tfrac{1}{2}(1 - \alpha) = \tfrac{1}{2}(1 - 0{\cdot}98304) = 0{\cdot}00848$$

The probability of occurrence of a head length less than or equal to 6·536 cm. is thus approximately 0·85 per cent., which is very much less than that in (i).

(iii)
$$\begin{aligned} \frac{x - \mu}{\sigma} &= \frac{2{\cdot}581}{1{\cdot}441} \\ &= 1{\cdot}791 \end{aligned}$$

The corresponding area $\frac{1}{2}(1 + \alpha)$ is obtained by interpolating into Table 2.1 between the values 1·7 and 1·8 and is found to be 0·96329.

$$\begin{aligned} \text{Area in the two tails} &= (1 - \alpha) \\ &= 2\,(1 - 0{\cdot}96329) \\ &= 0{\cdot}07342 \end{aligned}$$

Thus the probability of occurrence of a head length deviating from the mean by more than $\pm$ 2·581 is 7·3 per cent. Also the area of the central portion is 1 — 0·07342 or 0·92658. In other words, the odds against the occurrence of a head deviating from the mean by 2·581 cm. or more are 0·927 : 0·073, that is 13 : 1 approximately.

Example 2.2

For comparison of an observed distribution such as the one in Table 1.3 with the hypothetical normal distribution having the same mean and variance we need to calculate the expected frequencies in the various classes. This is done with the help of the probability integral table as follows.

Suppose we wish to calculate for the data of wheat earheads, the frequency expected in the class with class value 10·5 cm. The

class limits are 10·25 and 10·75 cm. We can find the areas $\frac{1}{2}(1+\alpha)$ cut off by the ordinates at 10·25 and 10·75 cm. from the probability integral table after calculating the corresponding normal deviates.

Thus the normal deviate corresponding to 10·25 cm. is

$$\frac{10\cdot25 - 9\cdot98}{1\cdot441} = 0\cdot19$$

and that corresponding to 10·75 cm. is

$$\frac{10\cdot75 - 9\cdot98}{1\cdot441} = 0\cdot54$$

The corresponding areas, which may be obtained approximately as in Example 2.1, are found more exactly to be 0·5753 and 0·7054 respectively from Table 2, *Tables for Statisticians and Biometricians* by K. Pearson.

The fraction of the total area, within this class is therefore $0\cdot7054 - 0\cdot5753 = 0\cdot1301$. This is the relative frequency of this class in the population, the total frequency being taken to be unity. The actual total frequency being 400, the absolute frequency expected in this class is $400 \times 0\cdot1301 = 52\cdot04$ (52 to the nearest integer).

For calculating the frequency in the next class we need to know the areas cut off by the ordinates at 10·75 and 11·25 cm. The former we have already found. The latter is found thus:

The normal deviate corresponding to 11·25 cm. is

$$\frac{11\cdot25 - 9\cdot98}{1\cdot441} = 0\cdot88$$

and the corresponding area is 0·8106.

Hence the fraction of area in the class 10·75 to 11·25 cm. is $0\cdot8106 - 0\cdot7054 = 0\cdot1052$ and the absolute expected frequency is $400 \times 0\cdot1052 = 42\cdot08$.

In this way we can calculate the expected frequencies in all classes.

Another method often used for this purpose is the one based on the value of the ordinate of the normal curve. Pearson's table referred to above also gives values of the ordinate for various values of the normal deviate. If Z is the value of the ordinate at the middle of a class interval I_x, then the frequency in the class is

$$\frac{N.Z.I_x}{\sigma}$$

where N is the total number of observations (400 in the above case).

Thus for the class with the value 10·5 cm., we find the normal deviate to be

$$\frac{10{\cdot}50 - 9{\cdot}98}{1{\cdot}441} = 0{\cdot}36$$

and the corresponding ordinate from Pearson's table is 0·3739. Therefore, the frequency of this class is

$$\frac{400 \times 0{\cdot}3739 \times 0{\cdot}5}{1{\cdot}441} = 51{\cdot}9$$

which is very close to the one found by the previous method. Owing to the approximate nature of this calculation the two values might differ slightly. Provided the areas $\frac{1}{2}(1 + \alpha)$ are found accurately by interpolation, the former method should be regarded as giving a better approximation to the true value.

2*b*.1 Binomial Distribution

In the previous paragraphs we have dealt with the normal curve, which is by far the most important curve in the application of statistical theory to a large variety of biological data. Another distribution only next in importance to the normal curve is the *binomial*. Here, however, we are concerned with a discrete variate taking only certain values in the range.

Suppose we throw a die after thorough shaking and count the turning up of a particular face, say 1, as success and of any other face, as failure. For a die which is a perfect cube of homogeneous material, all the six faces would in the long run turn up with the same relative frequency. Hence the probability

of the face 1 turning up, that is of a success, is $\frac{1}{6}$ and of a failure $\frac{5}{6}$, these probabilities remaining constant so long as there is no change in the die. If we were to consider a certain number of independent throws, say 5, to constitute a trial the number of successes in a trial may have any value ranging from 0 to 5. If such trials are repeated a number of times the different number of successes 0, 1, 2, ..., 5 will occur with certain relative frequencies. If the number of trials is increased indefinitely, the relative frequencies will approach values given by the coefficients of terms in the expansion of the binomial

$$\left(\frac{5}{6}b + \frac{1}{6}a\right)^5$$

the coefficient of the term $a^r \cdot b^{5-r}$ giving the expected relative frequency (probability) of r successes. The variate r showing the number of successes in 5 throws or the proportion $r/5$ is said to be distributed in a binomial distribution. In general, if p is the constant probability of a success and q $[= (1 - p)]$ the complementary probability of a failure, the number of successes r in n independent events would occur with a probability given by the coefficient of the term $a^r b^{n-r}$ in the expansion of the binomial

$$(qb + pa)^n$$

and hence given by

$$\frac{n(n-1)(n-2)\ldots(n-r+1)}{1.2.3.\ldots r} p^r q^{n-r}$$

In a large number, N, of such trials with n events we may expect that trials with r successes in n throws will be obtained nearly

$$N.\frac{n(n-1)(n-2)\ldots(n-r+1)}{1.2.3.\ldots r} p^r q^{n-r}$$

times.

The probability that the number of successes in n events is r or less would be the sum of the probabilities that the number of successes would be r, $r - 1$, ..., 0 and is therefore obtained by adding the coefficients of appropriate terms in the binomial. This will be clear from the following example.

Example 2.3

What is the probability of obtaining 4 or more heads in 6 independent tosses with a perfect coin?

With a perfect coin the probabilities of obtaining a head or a tail are equal, in other words, $p = q = \frac{1}{2}$. The probabilities of 0, 1, 2, . . ., 6 heads are given by the coefficients of terms containing successive powers of a in the expansion of

$$(\tfrac{1}{2}b + \tfrac{1}{2}a)^6$$

$$i.e.,\ (\tfrac{1}{2}b)^6 + 6(\tfrac{1}{2}b)^5(\tfrac{1}{2}a) + 15(\tfrac{1}{2}b)^4(\tfrac{1}{2}a)^2 + 20(\tfrac{1}{2}b)^3(\tfrac{1}{2}a)^3 + 15(\tfrac{1}{2}b)^2(\tfrac{1}{2}a)^4 + 6(\tfrac{1}{2}b)(\tfrac{1}{2}a)^5 + (\tfrac{1}{2}a)^6$$

The coefficients of terms containing 4th, 5th and 6th powers of a are $15(\frac{1}{2})^6$, $6(\frac{1}{2})^6$ and $(\frac{1}{2})^6$ respectively.

Therefore the probability of 4 or more heads is given by

$$(15 + 6 + 1)(\tfrac{1}{2})^6 = \frac{11}{32}$$

With p and q each equal to $\frac{1}{2}$ the distribution is symmetrical. When p is not equal to q, we get a skew binomial distribution. For larger and larger values of n, the binomial distribution, whether skew or symmetrical, approaches the normal distribution more and more closely and its probability integral is given by the normal probability integral with sufficient accuracy when p lies between ·1 and ·9 and n is 50 or larger. This is a fortunate circumstance since ordinarily the computations involved in adding up the coefficients of the binomial are heavy especially when n is large.

It can be shown that the binomial distribution has the mean value np, which is the expected number of successes in n trials, and a variance equal to $np(1 - p)$ or a standard deviation equal to $\sqrt{np(1-p)}$. To obtain the probability of getting r or more successes by approximating the binomial distribution to the normal, we calculate the ratio

$$\frac{x - \mu}{\sigma} = \frac{r - \frac{1}{2} - np}{\sqrt{np(1-p)}}$$

and refer the value to the normal probability integral table. The fraction $\frac{1}{2}$ is to be subtracted from r in calculating the ratio,

as the value of the class r is supposed to be at the mid-point between the values r and $r - 1$, *i.e.*, at $r - \frac{1}{2}$.

When an event can occur in one of the two possible ways it is said to be a *dichotomous* classification. The binomial distribution finds considerable use in problems relating to such situations, as for example, in dealing with two-class segregations in genetics. In a two-class segregation, the two classes are characterized by certain constant probabilities and families of size n are distributed in the binomial form, p and q referring to the probabilities of the two classes. This fact is made use of in the following chapters.

For events which can occur in more than two alternative ways, the probabilities of different results that might be obtained in sets of n such independent events, are given by a more general distribution that goes by the name of *multinomial*. The binomial distribution can be regarded as a particular case of this distribution. It is beyond the scope of this book to deal with the multinomial distribution in any detail.

CHAPTER III

SAMPLING METHOD AND STANDARD ERRORS

3*a*.1 INTRODUCTION

So far we have been dealing with populations as a whole. In practice the populations are usually too large to permit our taking observations on all the individuals comprising them; often they are hypothetical. Consider, for instance, the problem of estimating the total production from a crop in a given season. It is out of question to harvest and weigh the produce from all the fields growing the crop, which constitute the population under study. Again, a person may wish to ascertain the probability of certain faces turning up in the throw of a die. Here the probability refers to the proportion in all the throws that can be obtained with such a die, which obviously constitute an infinite hypothetical population. Similar is the case of a segregating progeny. When a geneticist wishes to know whether the progeny of a backcross or an F_2 segregates in a particular ratio, he has in mind all the progeny that may be obtained under the same mating conditions. It is obvious that in all such cases, where the population is too large or hypothetical, the investigator has to obtain information about the population from a part thereof. The selection of a part of the population to represent the whole is known as *sampling* and the part selected is known as the *sample*. The scientific methods of selecting samples and the relations between the sample and the population form the subject-matter of the theory of sampling.

3*b*.1 RANDOM SAMPLING

As stated above the object of sampling is to get information regarding the population from which the sample is obtained. However, some limitations are inherent in a sampling procedure. The sample being a part of the population cannot completely represent the whole and it does not seem that a sample, whatever be the method of selection adopted, can ever give us a complete knowledge of the population. For a sample to be useful, therefore,

it is necessary to have an idea of the degree of reliance which can be placed on the information provided by the sample in respect of the population. The simplest of the methods of selection of a sample, which enables us to estimate the population characters from the sample as also provides a measure of the degree of uncertainty attached to these estimates, is the method of *random sampling*.

The process of random sampling consists in basing selection on a chance event such as drawing of lots, throw of a die, etc. It should be remembered that though the result of a toss of a coin or a throw of a die cannot be individually predicted, the results of a large number of such trials can be anticipated with considerable accuracy and confidence. This is a remarkable property of chance events; they are not so erratic as is commonly believed. And as will be seen later, the results of selection based on such events can also be similarly anticipated.

Let us start with a simple though artificial problem. Suppose we have a miniature population of 6 individuals of magnitudes 1, 2, 3, 4, 5 and 6 units, and that we wish to take samples of 2 individuals from this population. For choosing samples of 2 individuals randomly from the population of 6, we might use 6 uniform cards, numbered from 1 to 6, shuffle them thoroughly and take two topmost cards. The different pairs that can be drawn from the population have been enumerated in Table 3.1. It can be shown that all these pairs are equally likely to turn up in the first two cards. The table gives the 15 different samples, their means and values of a certain other quantity calculated therefrom. It should be noted that the mean of the sample means equals 3·5, which is identical with the mean of the population, that is, the arithmetic mean of numbers 1, 2, 3, 4, 5 and 6.

This method of selection which gives each of the possible samples an equal chance constitutes the method of random sampling. It might be observed that when we draw two cards in succession, every number has a chance of $\frac{1}{6}$ of being selected at the first draw, while at the second, each of the remaining numbers has a chance of $\frac{1}{6}$ of being selected. Even so, all the possible samples get an equal chance of selection, since the chance of selecting any given pair of cards in a specific order is the product

TABLE 3.1

Means and proportions in samples from finite population

Units in the sample	Sample mean	Proportion of occurrence of 1
(1)	(2)	(3)
1, 2	1·5	0·5
1, 3	2·0	0·5
1, 4	2·5	0·5
1, 5	3·0	0·5
1, 6	3·5	0·5
2, 3	2·5	0
2, 4	3·0	0
2, 5	3·5	0
2, 6	4·0	0
3, 4	3·5	0
3, 5	4·0	0
3, 6	4·5	0
4, 5	4·5	0
4, 6	5·0	0
5, 6	5·5	0
Total	52·5	2·5
Mean	3·5	1/6

of the two chances of selecting the specified cards at the first and second draw respectively and is therefore equal to $\frac{1}{6} \times \frac{1}{5}$ or $\frac{1}{30}$. Since the order of selection is immaterial the chance of selecting a pair of cards is equal to $\frac{2}{30}$, or $\frac{1}{15}$.

3*b*.2 USE OF RANDOM NUMBERS

It is a general property of randomly selected samples that if we give at the time of selection of the successive units of the sample each of the remaining units an equal chance of being selected, each of the possible samples automatically gets an equal chance of being selected. Therefore, random sampling is often described as one which gives each member of the population an equal chance of being selected. Thus, if we have to select samples of 5 earheads randomly from our population of 400 earheads of Pusa 12 wheat (Table 1.2) all we have to do is to give at the selection of each earhead, each of the remaining earheads an equal chance of being selected and in consequence all the possible samples of 5 which can be obtained from the population get an equal chance. Giving equal chance to each unit can be done

in various ways. It is obvious that a throw of a die gives one of 6 possible results which are equally likely. If we consider trials of two throws we find that corresponding to each of the results of the first throw there are six results of the second throw so that trials of two throws can give $6 \times 6 = 6^2 = 36$ different results which are equally likely. Following the same reasoning it can be easily seen that trials of four throws can give 6^4 or 1,296 results which are equally likely. If we fix any 400 of these results to correspond to the 400 earheads, five trials would give us a random sample of the required size out of 400 earheads. More simply, we may note 400 numbers from 1 to 400 on 400 different tickets, which are quite similar, and take out of them 5 tickets after thorough mixing. This would also give us a random sample of 5. More processes of this type can be thought of.

However, there are many difficulties associated with the use of tickets, dice, etc., for randomization, such as the difficulty of ensuring the exact similarity of the tickets, uniformity of dice, etc., apart from the tediousness of the procedures. The practical difficulties in the way of random selection have been largely overcome by the publication of tables of random numbers and random selection is most conveniently done with the help of these tables. Appendix II gives a page from a table of random numbers. These tables usually consist of columns of 1, 2, 3 or 4 digit numbers randomly drawn and tested for their randomness. To use these tables we have to number the units constituting the population, start anywhere in these tables and move up or down, across or diagonally and choose the numbers in the population as they occur. If we are using a three-digit column, then the total number in the population should not exceed 1,000 (using the occurrence of 000 to indicate the selection of the 1,000th unit in the population). Thus, if we wish to draw samples of 5 earheads from our population of 400 earheads in Table 1.2, we may start at 029 in the fifth three-digit column (say). Running down the column we shall come across in succession the numbers 029, 265, 689, 905, 531, 526, 700, 469, 226, 407, 047, 325, 748, 782, 352, etc., of which we choose the first five underlined numbers, which are less than 400. Thus we can get our sample of 5 earheads, and more such samples. More methods

of using tables of random numbers for various purposes are described in *Statistical Tables for Biological, Agricultural and Medical Research by* Fisher and Yates.

3*b*.3 Expectation of Sample Estimates

It is a property of random sampling that the mean of means derived from random samples approaches the population value as the number of sample means averaged increases. The same is true of proportions calculated from samples, that is, the mean of sample estimates of proportion is equal to the proportion in the population. Where the number of possible samples is itself finite the mean of sample means as well as the mean of estimates of proportion equals their population values. In the example of sampling from a population of 6 units we saw that the mean of sample means was actually equal to the population mean. We may consider the proportion of occurrence of any number, say 1, in various samples. Column 3 of Table 3.1 gives these proportions for the various samples, and from the mean of these values calculated at the end of the column we see that the mean of these estimates actually equals the population value of the proportion, namely, $\frac{1}{6}$. In mathematical language this property is expressed by the statement that the *expected value* of the sample mean or proportion is equal to the population mean and proportion respectively, and denoted by

$$E(m) = \mu$$
$$E(\hat{p}) = p$$

E in the above equation stands for ' expectation of ' and

m = the sample mean,
$\hat{p}$ = the sample proportion,
μ = the population mean,
p = the population proportion.

Where this property does not hold, *i.e.*, where the expectation of a sample estimate is different from the population value, the estimate is said to be *biased.*

3*b*.4 Requisites of Randomization

It is often thought that a random sample is obtained when the selection is made without deliberate discrimination. This is,

however, not true; for, in such sampling bias may be introduced on account of psychophysical causes. Consider, for instance, the problem of selecting a number of cotton plants randomly from a field of cotton for taking observations on, say, plant height. An experimenter might feel that he can go to the field and select a number of plants randomly by selecting them just anyhow, while, as a matter of fact, his selection is likely to include a greater proportion of the more conspicuous members than what occurs in the population. The result would be that the sample mean would systematically tend to exceed the mean plant height of the population and thus by repeated sampling he would not obtain a value approaching the population mean; that is the relation $E(m) = \mu$ will not hold good. In other words, the sampling would be biased. If he tries to guard against selection of too many tall plants he may err in the opposite direction and choose too many stunted ones. Moreover, estimation of the mean is not always the only object of sampling; variability is also important as in genetic studies. Even if therefore a worker succeeds in selecting plants whose mean is a reliable estimate of the population value the sample may still fail to reflect the true variability of the population. A random sample alone can reflect the different characters of the population within, of course, the limitation of its size. Personal selection can never be trusted to give such samples.

It may be imagined that if all plants are numbered and a set of numbers written down without any design, one would get a random sample of plants. In actual fact, this process is also vitiated by the experimenter's fancies for and prejudices against various numbers and it is found that a set of numbers thus written down often contains too many of certain numbers and too few of some others, as may be verified by a trial. The research worker should, therefore, trust nothing short of an objective process that has been tested for giving equal chance to all members of a population. The method in general use is, of course, selection with the help of tables of random numbers. When constructing such tables, tests are carried out to ensure that the tables give a truly random selection.

The application of the method of random sampling presumes the population under study to be divisible into a number of

distinct identifiable units from which selection is to be done directly by numbering them. Thus, in the above example of selecting cotton plants, each of the plants in the field can be recognized separately and it is possible, at least theoretically, to number all the plants and select a number of them with the help of random numbers. In practice, it will be a very laborious project requiring the numbering of thousands of plants. Sometimes it is not possible to recognize separately and number the individual plants which form the population. Thus, if we wish to select a sample of wheat plants in a field for taking observations it will not be possible to recognize separate wheat plants, as the tillers from neighbouring plants run into one another. Under these conditions, alternative procedures of random selection are available. Instead of numbering the plants a research worker may number the rows and find out a random row with the help of tables of random numbers and, as he knows the length of the row, he can choose in this row a random yard length or a foot length of the crop with the help of random numbers for taking observations. If it is possible to recognize the plants separately, he can further choose one or more plants randomly out of the plants occurring in the chosen yard or foot length of the row and study them. If not, all the plants in the entire selected length may be selected for observation. This procedure of selecting a random sample in successive stages is called *sub-sampling*. The device is extensively used in sampling on account of the ease in selection and economy of labour.

3c.1 Distribution of Sample Means

The behaviour of means derived from random samples can be best studied with the help of an actual example of sampling. It is recommended that the student should obtain a number of samples (say 100 or more) of sizes 5, 10 and 20 with the help of tables of random numbers from the population of 400 earheads of wheat in Table 1.2 and find the means from these samples. The results of one such sampling experiment are shown in Fig. 3.1 which gives the frequency polygons of the sample means for 400 samples of each size. As might be expected the various sample means differ amongst themselves. By an inspection of the diagram we find that the range of the means decreases with increasing size

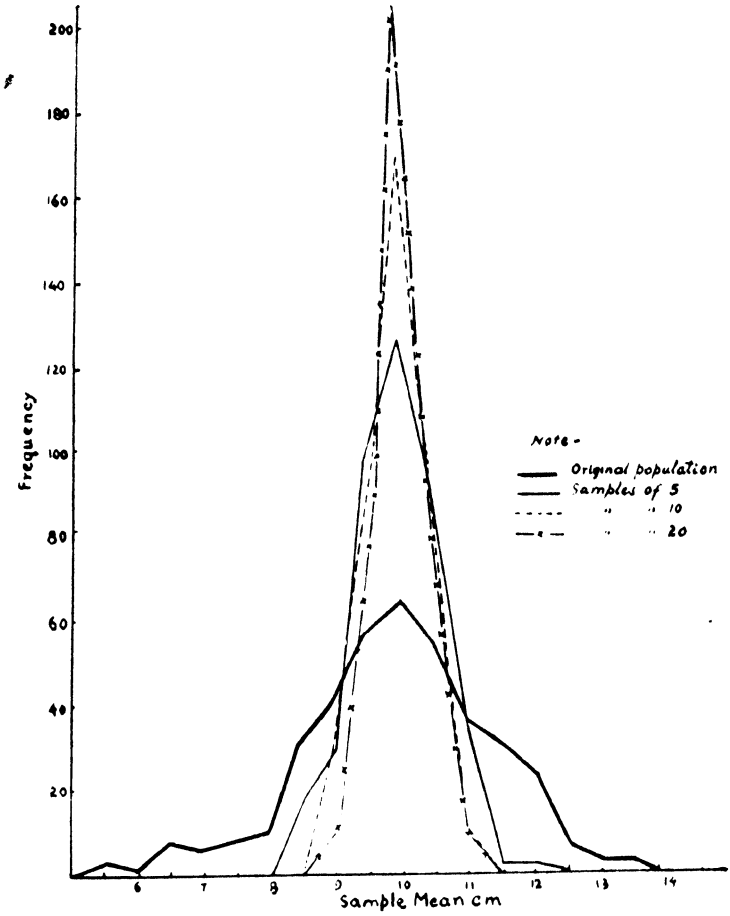


FIG. 3.1. Frequency polygons of sample means of 5, 10 and 20 earheads.

of the sample. In other words, the means of samples of the same size cluster more and more near the mean of the population as the sample size increases. If we calculate the means and standard deviations for each of the three distributions we find that the means practically coincide with the population mean while the standard deviations decrease with increase in sample size.

TABLE 3.2

Means and standard deviations of samples of different sizes

Size of sample	Mean cm.	Standard deviation cm.	Expected standard deviation cm.
Population	9·978	1·441	
5	9·934	0·630	0·644
10	9·921	0·473	0·456
20	10·008	0·357	0·322

As a matter of fact, the variability among means of random samples of a given size is related to the variability of the parent population by a definite mathematical relation. If $\sigma_{\bar{x}}$ stands for the standard deviation of means of samples of n individuals and σ for the standard deviation of the parent population, then

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{3.1}$$

The standard deviation of the mean is commonly termed the *standard error*. The square of the standard error is called the *sampling variance* of the mean. The standard deviations calculated from the above relationship are shown in the last column of Table 3.2. It will be seen that the actual and the expected values agree closely.

3*c*.2 CONFIDENCE LIMITS

An important property of random sampling is that if the original population is distributed normally or even approximately so, the distribution of the means of random samples is also normal. This fact enables us to find out with the help of the normal probability integral table the limits within which any given proportion of the sample means would lie. Thus, for instance, we can say that 95 per cent. of the means in samples of n lie between the limits $\mu \pm 2\sigma/\sqrt{n}$, μ being the population mean, or that there is 95 per cent. probability that the sample mean obtained lies within these limits. Conversely, if m is the sample mean, the limits $m \pm 2\sigma/\sqrt{n}$ would contain the population mean μ on an

average in 95 out of 100 cases. In other words, we may expect the inequality

$$m - \frac{2\sigma}{\sqrt{n}} \leqslant \mu \leqslant m + \frac{2\sigma}{\sqrt{n}} \tag{3.2}$$

to hold good on the average in 95 per cent. of the samples. The two limits in (3.2) are known as the *confidence limits* and the probability that the inequality (3.2) will hold good, in this case 0·95, is known as the *confidence coefficient*. The range between the two confidence limits is known as the *confidence interval*. It will be noted that the standard error of the mean and consequently the confidence interval becomes smaller as the size of the sample is increased. Using the inequality (3.2) we can calculate for $P = 0{\cdot}95$ the size of sample required to estimate the population mean within a given confidence interval.

Example 3.1

Calculate the size of sample required for estimating the mean length of earhead in the population (Table 1.2) with a margin of error not exceeding 5 per cent. of the population mean. The degree of assurance desired is $P = 0{\cdot}95$.

The mean μ is 9·978 cm. and the standard deviation σ is 1·441 cm. We want to determine n such that the probability that $m - \mu$ will not exceed $0{\cdot}05\mu$ is 0·95.

The value of the normal deviate corresponding to $P = 0{\cdot}95$ is approximately 2. Hence, substituting for m, μ, and σ in

$$\frac{m - \mu}{\frac{\sigma}{\sqrt{n}}} = 2$$

we get

$$\frac{\sqrt{n}\,(0{\cdot}05)\,(9{\cdot}978)}{1{\cdot}441} = 2$$

giving us

$$n = \left\{\frac{2\,(1{\cdot}441)}{9{\cdot}978 \times 0{\cdot}05}\right\}^2$$

$$= 33{\cdot}4$$

3*d*.1 Standard Error of Sum and Difference of Means

We shall now consider the standard error of sums and differences of means of two or more samples. Suppose we have two

samples of sizes n_1 and n_2 drawn from a population with mean μ and variance σ^2 and that m_1 and m_2 are the means of those samples. If m_1 deviates from μ by d_1 and m_2 by d_2, signs being taken into account, the deviation of $m_1 + m_2$ from 2μ will be $d_1 + d_2$. By definition the variance of a sample mean is the average value of the square of the deviation of sample mean from the population mean over all samples. The variance of $(m_1 + m_2)$ is therefore the average value of the quantity $(d_1 + d_2)^2$, that is of $d_1^2 + d_2^2 + 2d_1d_2$. Now by definition the average value of d_1^2 is the variance of m_1, namely σ^2/n_1. Similarly the average value of d_2^2 is the variance of m_2, namely σ^2/n_2. For independent samples the average value of d_1d_2 can be seen to be zero, since d_1 and d_2 will each occur with positive and negative signs equally frequently independent of each other, so that their products cancel one another. Hence

$$V(m_1 + m_2) = V(m_1) + V(m_2)$$

$$= \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} \tag{3.3}$$

and

$$S.E.(m_1 + m_2) = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$$

$$= \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{3.4}$$

Similarly the mean value of $(d_1 - d_2)^2$, that is, the variance of the difference $(m_1 - m_2)$, in independent samples, can be shown to be given by

$$V(m_1 - m_2) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} \tag{3.5}$$

and hence

$$S.E.(m_1 - m_2) = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{3.6}$$

The standard error of $m_1 - m_2$ is thus seen to be the same as that of the sum $m_1 + m_2$.

This result may appear somewhat curious. A little reflection would resolve the difficulty. It should be remembered that

both the deviations d_1 and d_2 take positive as well as negative values equally frequently. In case of $(d_1 + d_2)^2$ the quantity tends to be larger when both deviations are of the same sign than when they are of opposite sign. In case of the quantity $(d_1 - d_2)^2$ the opposite happens. On the whole, the deviations are of the same and opposite signs in equal number of cases, and so the pairs which contribute more to the first quantity contribute less to the second and *vice versa*; on the aggregate the two effects are equal.

3*d*.2 Standard Error of Linear Function of Independent Variates

It can be seen that what applies to m_1 and m_2 with variances σ^2/n_1 and σ^2/n_2 also applies to any two quantities x_1 and x_2 distributed independently with variances σ_1^2 and σ_2^2 respectively so that

$$V(x_1 \pm x_2) = \sigma_1^2 + \sigma_2^2 \tag{3.7}$$

If x_3 is another quantity distributed independently with a variance σ_3^2, then the variance of any of the quantities $x_1 \pm x_2 \pm x_3$ is given by

$$V(x_1 \pm x_2 \pm x_3) = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 \tag{3.8}$$

The result follows by considering $(x_1 \pm x_2)$ as the first and x_3 the second quantity. Similarly the argument can be extended to any number of independent quantities, so that

$$V(x_1 \pm x_2 \pm x_3 \pm x_4 \pm \ldots) = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2 + \ldots \tag{3.9}$$

If the variate x_1 deviates from its mean by d_1, a quantity kx_1 where k is a constant would deviate from the mean of kx_1 by kd_1 so that the square of the deviation would be $k^2d_1^2$, and consequently the variance of kx_1 will be $k^2\sigma_1^2$. If in equation (3.9) the quantities x_1, x_2, ... are replaced by k_1x_1, k_2x_2, ..., and σ_1^2, σ_2^2, ..., by $k_1^2\sigma_1^2$, $k_2^2\sigma_2^2$, ..., which are variances of k_1x_1, k_2x_2, etc., we get

$$V(k_1x_1 \pm k_2x_2 \pm k_3x_3 \pm \ldots) = k_1^2\sigma_1^2 + k_2^2\sigma_2^2 + k_3^2\sigma_3^2 + \ldots \tag{3.10}$$

This equation enables us to calculate the variance of a linear function of any number of independent variates. Its application will be illustrated in later chapters.

3*d*.3 Sample Estimate of Variance

It might be noted that for knowing the sampling variances of the various quantities derived from samples we need to know σ, the standard deviation for the population. This has to be estimated from the sample by the following formula appropriate for samples, namely,

$$s^2 = \frac{\Sigma d^2}{n-1} \tag{3.11}$$

where s^2 denotes an estimate of σ^2, d represents deviation of individual value from the sample mean and Σ is the summation taken over the n units in the sample.

The divisor $n-1$ for the sum of squares of deviations is termed the *degrees of freedom*. The degrees of freedom correspond to the number of independent deviations that are available from the data and can be calculated by deducting from the number of values available the number of constants that are calculated from the data. In the present example, the mean is such a constant and hence the number of degrees of freedom is one less than n, the number of observations.

3*d*.4 Standard Error of Estimate of Proportion

Just as the standard error of the mean is given by $\sigma/\sqrt{n}$ the standard error of an estimate of proportion p is given by

$$S.E.\,(\hat{p}) = \sqrt{\frac{p\,(1-p)}{n}} \tag{3.12}$$

This formula for the standard error of the proportion can be derived directly from formula (3.1) by scoring the presence of the attribute with 1 and its absence with 0. Then the mean of the score for the population will be $p\times 1 + q\times 0$ which is simply p. The variability of the score in the population is given by

$$\begin{aligned}\sigma^2 &= \Sigma\, f.d^2 = p\,(1-p)^2 + q\,(0-p)^2 \\ &= p\,(1-p)\,\{(1-p)+p\} \\ &= p\,(1-p)\end{aligned}$$

Hence if n is the sample size the variance of the sample mean is $p\,(1-p)/n$ according to the rule stated previously and hence the standard error of $\hat{p}_{\hat{p}}$, σ is given by

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{3.13}$$

It follows that the standard error of the number n_1 observed in any class with proportion p will be

$$S.E.(n_1) = n \cdot \sqrt{\frac{p(1-p)}{n}} \tag{3.14}$$

For calculating the standard error from the sample we must substitute for p the value actually observed. The correct divisor however is then $(n-1)$ and not n. Thus

$$S.E.(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} \tag{3.15}$$

and

$$S.E.(n_1) = n\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} \tag{3.16}$$

The variance of the sum or difference of proportions from samples can also be similarly obtained. We write

$$S.E.(\hat{p}_1 \pm \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1-1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2-1}} \tag{3.17}$$

where n_1 and n_2 stand for the sizes of the two samples.

Example 3.2

The following table gives the number of plants in the two classes of seed colour, green and yellow, for two families of peas. Calculate the joint estimate of the proportion of plants with green peas from the two families and estimate its error.

Family	Green	Yellow	Family total
1	110	40	150
2	120	36	156
Total	230	76	306

The joint estimate calculated from the total of two families is given by

$$\hat{p} = \frac{230}{306} = 0{\cdot}7516$$

This is obtained by pooling the families and amounts to taking a weighted mean of the proportions p_1 and p_2 estimated from individual families, weighting by the family sizes, n_1 and n_2.

Thus

$$\bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

The standard error of $\bar{p}$ is therefore given by

$$\begin{aligned} S.E.(\bar{p}) &= \frac{1}{n_1 + n_2} S.E.(n_1\hat{p}_1 + n_2\hat{p}_2) \\ &= \frac{1}{n_1+n_2}\sqrt{\frac{n_1^2 p(1-p)}{n_1} + \frac{n_2^2 p(1-p)}{n_2}} \\ &= \sqrt{\frac{p(1-p)}{n_1 + n_2}} \end{aligned}$$

where p stands for the population value of the proportion. Substituting for p its estimate $\bar{p}$ based on the joint data from the two families and reducing the denominator by 1, we have

$$\begin{aligned} S.E.(\bar{p}) &= \sqrt{\frac{0{\cdot}7516 \times 0{\cdot}2484}{305}} \\ &= 0{\cdot}0247 \end{aligned}$$

The formula for the standard error of proportion also enables us to calculate the family size required for estimating the population value of p within a given error provided some reasonable assumption can be made in advance about the value of p. Thus, suppose we wish to estimate the proportion p of plants with green peas in the above case with a standard error of 5 per cent. and that previous experience indicates that the value of p is around $\frac{3}{4}$. For this value of p the standard error would be $\sqrt{\frac{3}{4}.\frac{1}{4}.1/n}$. Equating this with $0{\cdot}05 \times \frac{3}{4}$, we get $n = 133$ to the nearest integer.

CHAPTER IV

TESTS OF SIGNIFICANCE OF MEANS AND THEIR DIFFERENCES

4*a*.1 Significance of Means in Large Samples

In Chapter III we saw that means of samples drawn randomly from a population differ amongst themselves and that it is possible to obtain a measure of variability of these means called the standard error. We obtained the standard errors not only of the means but also of the sums and differences of means of two or more independent samples. In this chapter we shall show with the help of examples how these standard errors are used in testing the *significance* of the results obtained from experiments or observational programmes in agricultural research. The meaning of the term significance in this connection will be clear from the examples themselves.

Example 4.1

Suppose we take a random sample of 100 earheads from a field of wheat and find that the sample gives a mean of 29·47 grains per earhead. The standard deviation of the number of grains per earhead is known to be 5·64 grains. If a mean of 30 grains per earhead has been postulated for the field, can we regard our sample as having been drawn from a population with this mean ?

Since the individual earheads have a standard deviation of 5·64 grains, the standard error of the mean of a random sample of 100 earheads will be

$$\frac{\sigma}{\sqrt{n}} \quad \frac{5\cdot64}{\sqrt{100}} \quad 0\cdot564$$

The means of samples of size 100 will therefore be distributed about the population mean with a standard error of 0·564. The deviation of the observed sample mean from the hypothetical population mean is $30\cdot00 - 29\cdot47 = 0\cdot53$. Hence, the ratio,

$$\frac{\text{deviation}}{\text{standard error}} \quad \frac{0\cdot53}{0\cdot564}$$

$$= 0\cdot94$$

Assuming that the sample means are distributed normally, which assumption is usually reasonable even when the original population deviates somewhat from the normal, the above ratio may be regarded as a normal deviate. By reference to the table of normal probability integral (Table 2.1) we find that the probability $(1 - \alpha)$, that the deviate is numerically as large as or larger than 0·94, is between 0·3681 and 0·3173, these values of probability corresponding to values 0·9 and 1·0 of the normal deviate. By linear interpolation, the probability corresponding to 0·94 is found to be 0·3478. This is the fraction of the infinitely large number of samples of this size that may be expected in repeated sampling to have the mean differing from the population mean by as much as or more than the observed amount. In other words, in sampling from a population with a mean of 30 grains and a standard deviation of 5·64 grains, we would obtain a deviation as big as the one observed or bigger, in nearly 35 per cent. of cases, for random samples of this size. Consequently, we have no reason to consider this sample mean to be in contradiction to the hypothetical population mean, for differences as large as the one observed or larger would arise as a result of sampling fluctuation quite frequently. A deviation of this magnitude is therefore said to be not significant. On the other hand, if the value of the probability of obtaining a deviation as large or larger than the one observed is low, we take it to be indicative of a real difference between the observed mean and the hypothetical value of the population mean; for, differences of such a magnitude or greater will rarely arise as a result of sampling fluctuations only.

When the probability of obtaining a deviation as large as or larger than the one observed is low, the deviation is said to be significant, that is, indicative of a real difference between the sample and the population means. What value of probability should be regarded as sufficiently low for indicating significance is to some extent a matter of choice. It is customary to consider 0·05 to be sufficiently low for this purpose, which means that if the probability of obtaining a deviation from the hypothetical value as large as or larger than that observed is 0·05 or less, we consider the deviation to be significant. The

probability chosen is known as the *level of significance* and the statistical procedure of deciding whether the hypothesis, termed the *null-hypothesis*—which in this case is that the sample comes from a population with a true mean of 30 grains per earhead—should be rejected or not is called the *test of significance*.

Obviously, the inferences about significance would go wrong a certain number of times depending upon the level of significance chosen. Since sampling fluctuation alone gives deviations significant at $P = 0{\cdot}05$ in 5 per cent. of cases, we would be wrongly interpreting a chance deviation as a real one once in twenty times on an average. It is not possible to get round this difficulty entirely. As stated in the previous chapter, inferences regarding the population, based on samples, are subject to some degree of uncertainty. The best that we can do is to reduce this uncertainty by choosing a more stringent level of probability for indicating significance. Thus we may choose the probability level to be $P = 0{\cdot}01$, that is, the 1 per cent. level of significance in which case a deviation arising from chance would be interpreted as a real one on an average only once in hundred times and we can place more confidence in the conclusion when a deviation is significant at $P = 0{\cdot}01$ level of significance than when it is significant at $P = 0{\cdot}05$ level. In either case, it is important to realize that statistical inference by its nature is uncertain but nevertheless rigorous in the sense that the degree of uncertainty is itself known or can be predetermined. The research worker should use a test of significance with discretion fully considering the situation at hand. He should not, on the one hand, set the level of significance so low as to be often misled by chance deviations. He should not, on the other, miss the indications of the results of experiments or observations when non-significant results are found, but should repeat and if necessary amplify the scope of his experiments with a view to obtaining more decisive results.

It can be seen from the table of normal probability integral that the deviation corresponding to a probability $P = 0{\cdot}05$ is about twice the magnitude of its standard error ($1{\cdot}96$ times the standard error, to be more exact), so that an observed deviation equal to twice the standard error is significant at the 5 per cent. level.

Similarly a deviation 2·58 times its standard error is significant at 1 per cent. level. In practice what we often require is not so much the probability corresponding to a particular value of the normal deviate but rather the value of the normal deviate corresponding to chosen levels of probability. Such a table of the normal deviate is given in Appendix III. If the observed value of the normal deviate is larger than the tabulated value at the chosen level of significance the deviation is considered to be significant.

The above discussion should make clear the idea underlying the testing of significance. The test of significance is a method of making due allowance for the sampling fluctuation affecting the results of experiments or observations. This is done in the present case by recognizing differences which are twice the standard error of the result, $2\sigma/\sqrt{n}$, or larger as significant of a real deviation. The fact that the results of biological experiments are affected by a considerable amount of uncontrolled variation makes such tests necessary. In sciences like Physics and Chemistry the uncontrolled variation is generally so small as compared to the magnitude of the observation that it is usually sufficient to consider the mean of two or three repetitions of an observation as accurately determined. In the interpretation of the results in experimental work in these sciences the need of statistical techniques is therefore not felt to the same extent as in the biological sciences.

The purpose of Example 4.1 is, chiefly, to explain the ideas connected with tests of significance. The test is applied there to a difference between an observed mean and a hypothetical value when the standard deviation of the population under study, σ, is known. Usually however the value of σ is not known. In that case the estimate of σ derived from the sample itself may be used without serious error provided the sample is sufficiently large, that is, contains over 30 individuals.

4*a*.2 Significance of Difference of Means in Large Samples

The comparison of an observed mean with its hypothetical value is not a problem of frequent occurrence. A problem more commonly met with in agricultural and other biological research

is the comparison of two sample means. Thus we may wish to compare the mean yields of two varieties of wheat, the staple length of two varieties of cotton, the sucrose percentage of two varieties of cane, etc., that is to say that we wish to test whether the two samples can be regarded as having been drawn from the same normal population. It may be recalled that the standard error of the difference of means of two samples of sizes n_1 and n_2 drawn from a population with standard deviation σ is

$$\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

and that its estimate is provided by

$$s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

where

$$s=\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} \qquad (4.1)$$

s_1 and s_2 standing for the standard deviations of the two samples respectively. Provided the samples are large enough for us to take the standard deviation σ to be accurately estimated from the samples, we can use formula (4.1) to test the significance of the difference between the two sample means with the help of the normal deviate. For this purpose we calculate the ratio

$$\frac{\text{difference of means}}{S.E.\text{ of difference}},$$

consider it to be a normal deviate and refer the value obtained to the table of normal probability integral. Example 4.2 will illustrate the procedure.

Example 4.2

Table 4·1 gives the data for the length of earheads of Pusa 4 wheat in two samples of 400 each taken in two seasons.

An estimate of the standard deviation of the population from which the two samples may be regarded as having been drawn is obtained by substituting for s_1 and s_2 in (4.1) the values

TABLE 4.1

Comparison of length of earhead in Pusa 4 wheat in two successive years

Year	Number of earheads in the sample	Mean length of earhead in cm.	Standard deviation of lengtn of earhead in cm.
1931–32	400	7·88	1·09
1932–33	400	7·82	0·90

1·09 and 0·90 respectively. We obtain

$$s = \sqrt{\frac{399\,(1\cdot09)^2 + 399\,(0\cdot90)^2}{798}} = 1\cdot00$$

In the absence of knowledge of the population value of σ, this may be taken as its value for the purpose of the test, since the two samples are large. Hence the standard error of the difference of means is

$$\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 1\cdot00\sqrt{\frac{2}{400}} = 0\cdot0707$$

The difference in the means of the two samples is 0·06 cm. Therefore, the ratio

$$\frac{\text{difference of means}}{S.E.\text{ of difference}} = \frac{0\cdot06^*}{0\cdot071} = 0\cdot8$$

It can be seen from the table of normal probability integral that the probability corresponding to this ratio is 0·424. The difference cannot, therefore, be regarded as significant.

It should be remembered while carrying out the test that the formula used for calculating the standard error of difference presumes the independence of the two samples and the significance of the difference should be tested in the manner of Example 4.2 only when the two samples concerned are independent. By

* The reader will note that since the difference between the means is given correct to one significant figure it is unnecessary to retain more than two significant figures in the denominator.

independence of the two samples we mean that there is no relationship or correspondence between the individuals constituting the two samples. Thus samples drawn from different populations or different parts of the same population will be independent. Samples drawn at random from the same population can be considered as independent when the population is large. If there is one to one correspondence between the individuals in the two samples and they can be paired through the existence of some common factor linking the members of a pair together, then we cannot regard the two samples as independent. In the present case, for instance, we cannot pair the 400 earheads in the first sample and the 400 earheads in the second sample according to any natural relationship or correspondence and we can consider the two samples to be independent. If, on the other hand, the 400 earheads in the second season had been obtained from the respective progenies of the 400 earheads selected in the previous season there would have been a parent-progeny correspondence between the individuals in the two samples and we could have paired them on the basis of this relationship. The above test would have been inappropriate in that case. The correct test for data in which pairing is possible is described later in the chapter. When, however, the seed is mixed and the sample in the second season obtained from the bulk crop, the correspondence is lost and the two samples can be considered to be independent, as in the above example.

4*b*.1 Small Samples: *t* Test

In Example 4.2 the standard deviation for each season was obtained from 400 values or 399 independent deviations which, as mentioned in the last chapter, are called degrees of freedom. The final estimate of the variance of the length of earhead is thus based on $399 + 399 = 798$ degrees of freedom. When a variance is based on such a large number of degrees of freedom it can be taken to provide an accurate estimate of the population variance and the test of significance of difference may be carried out safely on the basis of the normal deviate. However, it frequently happens in biological research that we have to compare means obtained from samples consisting of only a few observations. In such cases the variance obtained from the sample is

not a sufficiently accurate estimate of the population variance and it is not correct to test the significance of the difference with the help of the normal deviate. The distribution of the quantity

$$\frac{\text{difference}}{\text{S.E. of difference}},$$

when the population variance is estimated from a small sample, deviates considerably from normality. On the null-hypothesis the numerator of the ratio is as likely to be positive as negative so that the distribution of the ratio is symmetrical. The distribution basic to the testing of differences in small samples was first worked out by W. S. Gosset writing under the pseudonym *Student* and tables are available giving the values of the above ratio which are exceeded in sampling with a known probability, for samples of different sizes, that is, for different numbers of degrees of freedom (*Statistical Tables*, Fisher and Yates). The ratio

$$\frac{\text{difference}}{\text{S.E. of difference}},$$

when the standard error of difference is estimated from the sample, is denoted by t and the table giving the values of the ratio required for significance at various levels of probability and for different degrees of freedom is spoken of as the table of t. The table of t is to be entered with the number of degrees of freedom on which the estimate of the standard error is based. The values of t required for significance at the 5 and 1 per cent. levels of significance, for degrees of freedom ranging from 1 to 30, are given in Appendix IV. For brevity these values are written as $t_{5\%}$ and $t_{1\%}$. Thus the value $t_{5\%}$ is that value which is exceeded with a probability of 0·025 in the negative direction and 0·025 in the positive direction, making the total probability $P = 0{\cdot}05$. Figure 4.1 reproduces in a graphical form the $t_{5\%}$ and $t_{1\%}$ values for different numbers of degrees of freedom. The number of degrees of freedom n' is given along the abscissæ and the values of t required for significance at 5 and 1 per cent. levels are plotted along the y-axis and the points corresponding to each level are joined to give the two curves. These curves show how the value of t required for significance at either level

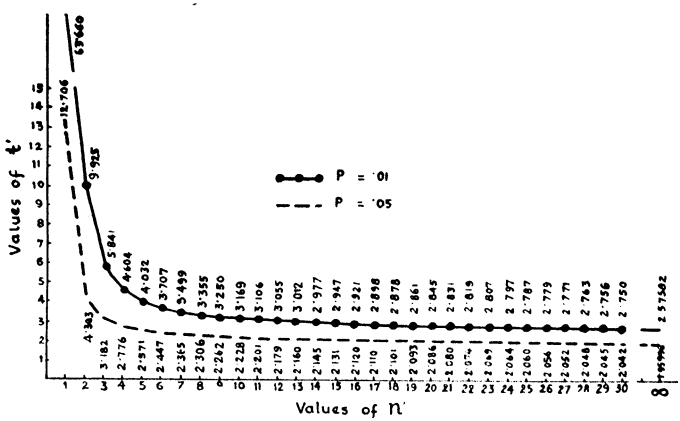


FIG. 4.1. Values of t at P = ·01 and P = ·05 levels of significance for different numbers of degrees of freedom.

depends upon the number of degrees of freedom, n'. From $n' = 1$ to $n' = 5$ or 6 the value decreases rapidly after which the decrease is much more gradual and practically negligible after $n' = 30$. It is sufficiently accurate to take $t = 2$ for significance at the 5 per cent. level for $n' = 30$ and above. For $n' = \infty$, the values of t are the same as those of the normal deviate and this part of the table is, in fact, the table of the normal deviate.

If m is the mean of a sample of size n and s is the standard deviation as estimated from the sample and if we are testing the deviation of m from a hypothetical value μ then

$$t = \frac{(m - \mu)\sqrt{n}}{s} \tag{4.2}$$

with $n - 1$ degrees of freedom.

If we want to test the significance of difference between two sample means m_1 and m_2 based on independent samples of sizes n_1 and n_2 and s_1 and s_2 are the estimated standard deviations, we first calculate an estimate of the common standard deviation applicable to both samples by pooling together the sums of squares corresponding to the two samples, dividing by the total number of degrees of freedom and taking the square root as in formula (4.1). Then

$$t = \frac{m_1 - m_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{m_1 - m_2}{s}\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \tag{4.3}$$

This value of t will be naturally entered in the table of t with $n' = n_1 + n_2 - 2$ degrees of freedom. In the special case when the two samples are of equal size, say n, the test reduces to

$$t = \frac{m_1 - m_2}{s}\sqrt{\frac{n}{2}} \tag{4.4}$$

and the t table will be entered with $n' = 2(n - 1)$ degrees of freedom.

It should be noted that the t test is appropriate when the variance of the population concerned is estimated from the sample, while the test based on the normal deviate is strictly applicable only when the variance is known from hypothesis. When the sample is large the two tests are identical for all practical purposes and therefore the t test is generally used only with small samples. It may be supposed that the t test is only an approximate test, influenced by errors in the estimation of the population standard deviation due to the small size of the sample. In fact, it differs from the test based on normal deviate in that it takes into account these errors. It is, therefore, a rigorous test of the hypothesis that the two samples are drawn from a population having the same mean and variance. The following example illustrates the application of the test.

Example 4.3

Table 4.2 gives the rainfall at two places A and B in 24 seasons 1924 to 1947, both years inclusive. Assuming that the 24 seasons constitute a representative sample of the rainfall at the two places, can we consider the two places to have the same mean annual rainfall ?

The 24 values of the rainfall at each of the two places cannot be considered to be independent samples of the rainfall at the two places since there is seasonwise correspondence between the values. Pairing of these values according to seasons is thus essential. However, we shall treat the data first as if the two samples are independent. This will incidentally serve to show

TABLE 4.2

Comparison of rainfall at two places A *and* B

Season	Rainfall in inches at *A*	Rainfall in inches at *B*	*A* + *B*	*A* − *B*
1	2	3	4	5
1924	39·59	39·48	79·07	0·11
1925	19·93	17·81	37·74	2·12
1926	23·91	24·47	48·38	−0·56
1927	29·38	24·32	53·70	5·06
1928	43·09	41·18	84·27	1·91
1929	25·34	23·41	48·75	1·93
1930	49·35	45·13	94·48	4·22
1931	39·62	42·83	82·45	−3·21
1932	42·90	46·94	89·84	−4·04
1933	53·35	51·51	104·86	1·84
1934	57·66	57·50	115·16	0·16
1935	37·05	34·35	71·40	2·70
1936	34·14	34·29	68·43	−0·15
1937	38·01	38·65	76·66	−0·64
1938	52·40	50·32	102·72	2·08
1939	32·20	29·94	62·14	2·26
1940	47·81	45·24	93·05	2·57
1941	33·98	34·13	68·11	−0·15
1942	39·46	40·68	80·14	−1·22
1943	37·78	35·54	73·32	2·24
1944	63·24	57·24	120·48	6·00
1945	39·04	42·05	81·09	−3·01
1946	60·51	55·33	115·84	5·18
1947	38·08	37·45	75·53	0·63
Total	977·82	949·79	1927·61	28·03
$\bar{x}$	40·74	39·57		1·17
$\sum x^2$	42740·21	40260·00		188·22
$\bar{x}\sum x$	39838·83	37587·54		32·74
$\sum(x-\bar{x})^2$	2901·38	2672·46		155·48
$\frac{1}{n-1}\sum(x-\bar{x})^2$	126·15	116·19		6·76
S.D.	11·23	10·78		2·60
Pooled *S.D.*	11·01			

how an erroneous conclusion may be obtained by an inappropriate application of the test.

The details of the calculation of the means and standard deviations of the two samples are shown under the corresponding columns of the table. Substituting these values in formula (4.4), we have

$$t = \frac{40{\cdot}74 - 39{\cdot}57}{11{\cdot}01}\sqrt{\frac{24}{2}}$$

$$= 0{\cdot}37$$

This value of t with 46 degrees of freedom will be seen to be much lower than the 5 per cent. value of t and on the basis of this evidence we would conclude that there is no significant difference in the rainfall at the two places.

4*b*.2 *t* Test in Paired Samples

The treatment of the rainfall data in the previous section as if the two sets of values are independent ignores the interdependence of the rainfall values at the two centres in the same season. The appropriate method is to take the differences in the corresponding values of the variate (the differences being taken in the same direction with due regard to sign) and analyze these as values of a separate variate. If there is no difference in the mean rainfall at the two places, the expected value of the difference would be zero. We can calculate directly the standard deviation of the difference and test the significance of the deviation of the mean difference from the hypothetical value, zero. The differences $(A - B)$ are given in the fifth column of the table and the calculations are done as in the previous case. The total stands for the algebraic total. The standard deviation of the difference is 2·60. Hence

$$t = \frac{(1 \cdot 17)\sqrt{24}}{2 \cdot 60} = 2 \cdot 20$$

Since there are 24 differences, the estimate of their standard deviation is based on 23 degrees of freedom. The table of t is therefore entered with $n' = 23$. We find that for this value of n', $t_{5\%} = 2 \cdot 07$. Since the value of t we have calculated exceeds this value we conclude that the difference between the mean rainfall at the two places is significant. Thus we have obtained a result different from the previous one. The reason is that on account of the parallel changes in the rainfall at the two places the differences are far less variable than would have been the case if they had been independent. Consequently by assuming their independence we have overestimated the error of the difference and underestimated its significance. While applying the t test the worker must, therefore, bear in mind this point about the independence of the two samples and guard himself against an erroneous application of the test. If there is any kind

of correspondence between the individual values in the two samples, they should be paired and differences taken and analyzed directly as is done above.

4c.1 Analysis of Variance

The test of significance of the difference in paired samples in the above example can also be put in another and a more instructive form. From the 48 values of the rainfall we can calculate the total sum of squares of deviations (briefly, S.S.) from the general mean, that is $\sum_{i=1}^{24}\sum_{j=1}^{2}(x_{ij}-\bar{x})^2$, where x_{ij} is the rainfall at the j-th place in the i-th season and $\bar{x}$ is the general mean. The corresponding mean square (briefly, M.S.) will be obtained by dividing the sum of squares by 47, the number of degrees of freedom on which it is based. We can look upon the total variation represented by the mean square as arising from three causes, namely variation due to places, variation due to seasons and variation due to all other indistinguishable causes, and we can estimate the components of variation due to these three factors in the following manner:

The quantity $(x_{ij}-\bar{x})$ can be expressed as

$$(x_{ij}-x_{i.}-x_{.j}+\bar{x})+(x_{i.}-\bar{x})+(x_{.j}-\bar{x})$$

where $x_{i.}$ and $x_{.j}$ are the respective season and place means.

Hence the quantity $\sum_{1}^{24}\sum_{1}^{2}(x_{ij}-\bar{x})^2$ can be put as

$$\sum_{1}^{24}\sum_{1}^{2}\{(x_{ij}-x_{i.}-x_{.j}+\bar{x})+(x_{i.}-\bar{x})+(x_{.j}-\bar{x})\}^2$$

This quantity, it is obvious, is equal to

$$\sum_{1}^{24}\sum_{1}^{2}(x_{ij}-x_{i.}-x_{.j}+\bar{x})^2+24\sum_{1}^{2}(x_{.j}-\bar{x})^2+2\sum_{1}^{24}(x_{i.}-\bar{x})^2$$

$$+2\sum_{1}^{24}\sum_{1}^{2}(x_{ij}-x_{i.}-x_{.j}+\bar{x})(x_{i.}-\bar{x})$$

$$+2\sum_{1}^{24}\sum_{1}^{2}(x_{ij}-x_{i.}-x_{.j}+\bar{x})(x_{.j}-\bar{x})$$

$$+2\sum_{1}^{24}\sum_{1}^{2}(x_{.j}-\bar{x})(x_{i.}-\bar{x})$$

The last three terms can be shown to be equal to zero. Thus taking the fifth term we have

$$\sum_{1}^{24}\sum_{1}^{2}(x_{ij} - x_{i.} - x_{.j} + \bar{x})(x_{.j} - \bar{x})$$

$$= \sum_{1}^{2}(x_{.j} - \bar{x})\sum_{1}^{24}(x_{ij} - x_{i.} - x_{.j} + \bar{x})$$

$$= \sum_{1}^{2}(x_{.j} - \bar{x})(24x_{.j} - 24x_{.j} - 24\bar{x} + 24\bar{x})$$

$$= 0$$

Similarly, the other two terms can be shown to be zero. Hence

$$\sum_{1}^{24}\sum_{1}^{2}(x_{ij} - \bar{x})^2 = \sum_{1}^{24}\sum_{1}^{2}(x_{ij} - x_{i.} - x_{.j} + \bar{x})^2 + 24\sum_{1}^{2}(x_{.j} - \bar{x})^2 + 2\sum_{1}^{24}(x_{i.} - \bar{x})^2$$

The quantity $2\sum_{1}^{24}(x_{i.} - \bar{x})^2$, which is derived from the seasonal means, gives the sum of squares due to seasonal variation and the mean square corresponding to seasonal variation is obtained by dividing the sum of squares by 23 degrees of freedom between seasons. Similarly the quantity $24\sum_{1}^{2}(x_{.j} - \bar{x})^2$ represents the sum of squares due to place variation and the corresponding mean square is obtained by dividing this sum of squares by 1 degree of freedom between places. The quantity $\sum_{1}^{24}\sum_{1}^{2}(x_{ij} - x_{i.} - x_{.j} + \bar{x})^2$ represents the sum of squares due to residual causes and the mean square obtained by dividing this sum of squares by the remaining 23 degrees of freedom provides a measure of variation due to residual causes. This last component of variation is spoken of as the *interaction* between seasons and places and represents the variation in place differences between different seasons not accounted for by pure seasonal differences or, which is the same thing, represents the variation in seasonal differences between the places not accounted for by the pure place differences. Since it represents the variation due to uncontrolled causes it is called the error mean square.

The different sums of squares are calculated in practice as explained below. It can be shown that

$$\sum_{1}^{24}\sum_{1}^{2}(x_{ij}-\bar{x})^2=\sum_{1}^{24}\sum_{1}^{2}x_{ij}^2-48\bar{x}^2$$

the term $48\bar{x}^2$ being termed the correction factor (*C.F.*). Similarly

$$2\sum_{1}^{24}(x_{i.}-\bar{x})^2=2\sum_{1}^{24}x_{i.}^2-48\bar{x}^2, \textit{ i.e., } 2\sum_{1}^{24}x_{i.}^2-C.F.$$

and

$$24\sum_{1}^{2}(x_{.j}-\bar{x})^2=24\sum_{1}^{2}x_{.j}^2-48\bar{x}^2, \textit{ i.e., } 24\sum_{1}^{2}x_{.j}^2-C.F.$$

In practice it is convenient to calculate the sum of squares from the class totals rather than from class means. If $T_{i.}$ is the i-th seasonal total, $T_{i.}=2x_{i.}$.

Hence

$$2\sum_{1}^{24}x_{i.}^2=\tfrac{1}{2}\sum_{1}^{24}T_{i.}^2$$

The sum of squares between seasons is therefore equal to

$$\tfrac{1}{2}\sum_{1}^{24}T_{i.}^2-C.F.$$

Similarly the sum of squares due to places is equal to

$$\frac{1}{24}\sum_{1}^{2}T_{.j}^2-C.F.$$

where $T_{.j}$ is the j-th place total, the divisor representing the number of values on which the class total is based. The correction factor (*C.F.*) is best calculated from the grand total G of all observations, being $G^2/48$.

The sum of squares due to the interaction is usually calculated by subtracting the two latter sums of squares from the total sum of squares. When the data consist of pairs of observations as in the above case, it might be obtained directly from the formula

$$\text{Interaction } S.S.=\tfrac{1}{2}\left[\sum_{1}^{24}d_i^2-24\bar{d}^2\right] \qquad (4.5)$$

where d_i stands for the i-th difference (taken with due regard to sign) and $\bar{d}$ is the mean difference. This result serves as a useful check on the accuracy of the calculations. The results of the calculation of the different mean squares are generally presented in a tabular form as shown in Table 4.3.

The above method of partitioning the total variation into components assignable to different causes is known as the *analysis of variance* and the table showing the various mean squares together with the corresponding degrees of freedom is called the analysis of variance table.

TABLE 4.3

Analysis of variance of rainfall data

Source of variation	Degrees of freedom	Sum of squares	Mean square
Places	1	16·37	16·37
Seasons	23	5496·10	238·96
Places × Seasons (error) ..	23	77·74	3·38
Total ..	47	5590·21	118·94

The reader should check the calculations as an exercise.

4c.2 THE *F* TEST

The analysis of variance table provides a ready means of testing the significance of differences between class means. Thus in the rainfall example it provides the means of testing differences between the two places or between the 24 seasons. Had the observed differences between the places or the seasons been entirely due to random fluctuations we would expect the various mean squares to be of the same order. Hence a comparison of the mean square due to any cause (such as seasons or places) with the error mean square provides a test of significance of differences arising from that particular cause. The comparison is done by finding the ratio of the mean square concerned (mean square due to places or seasons in the present example) to the error mean square. This ratio is known as the *variance ratio* and is denoted by *F*. Tables are available giving the values

of F required for significance at different levels of probability and for different degrees of freedom for the numerator and the denominator of the ratio. The table of F in Appendix V gives these values for the two levels of significance commonly used, namely, the 5 and 1 per cent. levels. The difference we wish to test here is the difference between places. Hence we have

$$F = \frac{16 \cdot 37}{3 \cdot 38} = 4 \cdot 84$$

From the table of F we find that with 1 and 23 degrees of freedom respectively for the numerator and the denominator the values required for significance at 5 and 1 per cent. levels are 4·28 and 7·88 respectively. Thus, this test leads to the same result as the one previously obtained, namely, that the place differences are just significant at the 5 per cent. level.

These two tests are, in fact, identical here, since for a single degree of freedom for the numerator the ratio F is identically equal to t^2 as can be verified from the values of t and F observed. The F test is actually of wider application in that it also provides an overall test of several differences whereas the t test provides a test of a single difference. We shall have frequent occasion to use the F test in succeeding chapters where its use will be explained in greater detail.

4*d*.1 Fisher and Behren's *d* Test

As mentioned earlier, the t test is a test of the hypothesis that the samples compared have been obtained from populations with the same mean and variance. Sometimes we may wish to test whether the two samples belong to populations having the same means though with different variances. Thus if the variances estimated from the samples, s_1^2 and s_2^2, seem to differ and we doubt the equality of population variances, the t test may appear inadequate and we may wish to test the hypothesis that the samples have been obtained from populations with the same mean but possibly differing in variances. A test of this hypothesis is provided by Fisher and Behren's d-test. For carrying out the test we first calculate s_1 and s_2 (based on samples of sizes n_1 and n_2 respectively) and find an angle θ such that

$$\tan \theta = \frac{s_1/\sqrt{n_1}}{s_2/\sqrt{n_2}}$$

or

$$\theta = \tan^{-1}\left\{\frac{s_1\sqrt{n_2}}{s_2\sqrt{n_1}}\right\} \tag{4.6}$$

Then we calculate the ratio of the difference between the two means m_1 and m_2 to its standard error given by

$$d = \frac{m_1 - m_2}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}} \tag{4.7}$$

where m_1 and m_2 are the two sample means and refer this ratio to the tables of d by Sukhatme (*Statistical Tables*, Fisher and Yates, Table V_1). The tables give the values of d required for significance at 5 and 1 per cent. levels for different combinations of n_1', n_2' the numbers of degrees of freedom and the angle θ where $n_1' = n_1 - 1$ and $n_2' = n_2 - 1$. The values of the ratio for intermediate values of n_1', n_2' and θ are obtained by interpolation (Appendix I). The application of the test is illustrated in the following example:

Example 4.4

Two samples of sizes 9 and 20 give the following values of m_1, m_2 and s_1^2 and s_2^2:

Sample 1: $m_1 = 10{\cdot}25$ cm. $s_1^2 = 0{\cdot}1316$ cm.^{2} and $n_1' = 8$

Sample 2: $m_2 = 9{\cdot}55$ cm. $s_2^2 = 3{\cdot}710$ cm.^{2} and $n_2' = 19$

We have

$$\tan \theta = \frac{\sqrt{0{\cdot}1316/9}}{\sqrt{3{\cdot}710/20}}$$

$$= \sqrt{\frac{0{\cdot}0146}{0{\cdot}1855}}$$

$$= 0{\cdot}281$$

From the table of natural tangents (*Statistical Tables*, Fisher and Yates, Table XXXII) we find that θ has a value of about 15° 42′. Substituting the values in (4.7) we get

$$d = \frac{10{\cdot}25 - 9{\cdot}55}{\sqrt{0{\cdot}0146 + 0{\cdot}1855}} = \frac{0{\cdot}70}{0{\cdot}447} = 1{\cdot}57$$

From the table of d (*Statistical Tables*, Fisher and Yates, Table V_1), it is found by interpolation that the value of d required for significance at the 5 per cent. level with $n_1' = 8$, $n_2' = 19$ and $\theta = 15°$ is 2·10. The observed value of d being smaller than 2·10 is thus non-significant.

If instead we calculate t, disregarding the inequality of variance, we have

$$s = \sqrt{\frac{n_1' s_1^2 + n_2' s_2^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{8 \times 0{\cdot}1316 + 19 \times 3{\cdot}710}{9 + 20 - 2}}$$

$$= \sqrt{2{\cdot}649}$$

$$= 1{\cdot}628$$

and

$$t = \frac{10{\cdot}25 - 9{\cdot}55}{1{\cdot}628\sqrt{\frac{1}{9} + \frac{1}{20}}}$$

$$= \frac{0{\cdot}70}{0{\cdot}653}$$

$$= 1{\cdot}07$$

The value of $t_{5\%}$ with 27 (*i.e.*, 8 + 19) degrees of freedom is 2·052. The observed value of t is, therefore, also non-significant. It should be observed, however, that the difference between the values of t calculated from the data and required for significance at the 5 per cent. level is greater than that between the corresponding values of d. The t test has under-estimated the significance of the deviation. It follows that it is important to use the d test when the variances of the two populations from which the samples are drawn are likely to be unequal.

CHAPTER V

THE χ^2 TEST AND ESTIMATION OF LINKAGE

5*a*.1 INTRODUCTION

IN the last chapter we considered the application of tests of significance to quantitative characters such as length of earhead, amount of rainfall, etc. Frequently the biological research worker has to deal with data of the qualitative type in which the observations are classified into several mutually exclusive classes or groups and the frequencies in these classes are given. The principles underlying the statistical treatment of such data are essentially similar to those dealt with in the previous chapter, but problems encountered in connection with these data are somewhat different and consequently the methods of treatment of such data also differ from those in the last chapter. Thus, we may want to know whether the frequency in any particular class is in agreement with a hypothetical value, or whether the entire distribution into a number of classes is in agreement with that on some hypothesis, or whether two such distributions are in agreement with one another, or whether sets of classifications are independent and so on. It is the purpose of the present chapter to set out the methods of dealing with such problems.

5*a*.2 TESTING THE SIGNIFICANCE OF PROPORTION

A problem of frequent occurrence in connection with data on frequencies is to test the agreement of a proportion of occurrence (or absence) of a character in a given number of observations with a hypothetical value. The reader may remember that at the end of Chapter III we gave the formula for the standard error of an estimate of proportion p. We can use it to test the significance of deviation of an observed value of the proportion from the hypothetical, by calculating the normal deviate as shown in the following example:

Example 5.1

In an F_2 population of chillies 831 plants with purple and 269 with non-purple chillies were observed. Is this ratio

consistent with a single factor ratio of 3 : 1 ? The observed ratio is

$$\frac{831}{1100} : \frac{269}{1100}, \text{ i.e., } 0{\cdot}755 : 0{\cdot}245$$

while the hypothetical ratio is

$$0{\cdot}750 : 0{\cdot}250.$$

The standard error of the observed proportion in either class will be

$$S.E.(\hat{p}) = \sqrt{\frac{3}{4} \cdot \frac{1}{4} \cdot \frac{1}{1100}} = 0{\cdot}0130$$

The ratio

$$\frac{\text{deviation}}{\text{standard error}} = \frac{0{\cdot}755 - 0{\cdot}750}{0{\cdot}0130}$$

$$= 0{\cdot}4 \text{ approximately}$$

This is a low value of the normal deviate the probability of exceeding it being nearly 69 per cent. The deviation is, therefore, non-significant and there is no reason to believe that the observed segregation is not consistent with the hypothetical ratio.

It is generally more convenient to make the test of significance on the observed number in a class than on the proportion. For this purpose we use the standard error of the observed number in any class given by

$$S.E.(n_1) = \sqrt{npq}$$

and

$$S.E.(n_2) = \sqrt{npq} \qquad (5.1)$$

where n_1 is the number observed in a sample of n, in class 1 and n_2 in class 2, so that $n_1 + n_2 = n$ and p is the population proportion in class 1 and q that in class 2, $p + q$ being equal to unity. Thus the standard error of the number in either class in the above example is

$$\sqrt{\frac{3}{4} \cdot \frac{1}{4} \cdot 1100} = 14{\cdot}4$$

The number expected in the purple class is $1100 \times \frac{3}{4}$ or 825. The deviation of the observed number from the expected in the purple class is therefore $831 - 825 = 6$. Hence, the ratio

$$\frac{\text{deviation}}{\text{standard error}} = \frac{6}{14 \cdot 4}$$

$$= 0 \cdot 4$$

same as in previous case as expected.

The testing of significance of an observed value of proportion against its standard error by reference to the normal probability integral table is not quite accurate, for as we saw in Section 2*b*.1 this procedure rests on the assumption that the observed value of proportion is normally distributed about its theoretical value p with the standard error $\sqrt{pq/n}$. For practical purposes, the approximation is sufficiently close when p does not lie outside the limits 0·1 to 0·9 and n is 50 or greater. Outside these limits it is advisable to use the test derived from the exact binomial distribution, by calculating the probability of obtaining a deviation as large as the one observed or larger.

5*b*.1 The χ^2 Test of Goodness of Fit

A test of wide applicability to numerous problems of significance in frequency data is the χ^2 test of *goodness of fit*. As its name indicates it is primarily used for testing the agreement of observed frequencies with those expected upon a given hypothesis as, for instance, in comparing an observed frequency distribution with a theoretical one like the normal. For carrying out this test we calculate, from the data, the quantity

$$\chi^2 = \sum \frac{(O - E)^2}{E} \tag{5.2}$$

where O stands for the observed and E for the expected frequency in any particular class of the distribution and Σ for the summation over all classes. It can be seen that for a complete agreement with the hypothetical distribution the value of the χ^2 score will be zero; but chance deviations are bound to occur and we shall obtain positive values of the score due to sampling fluctuations. The distribution of the χ^2 score in sampling is due to Karl Pearson

and has been tabulated in Appendix VI. The table shows the value of χ^2 that will be exceeded in sampling with given probabilities. This value depends on the number of classes of the observed distribution that can be filled up arbitrarily. The latter is also called the number of degrees of freedon on which the χ^2 is based. The table accordingly gives, corresponding to each value of the number of degrees of freedom, n', from 1 to 30, the value of χ^2 that will be exceeded with various probabilities. The values of χ^2 for values of n' beyond the range tabulated will be rarely required but for greater values of n', the significance of χ^2 can be tested by calculating the quantity $\sqrt{2\chi^2} - \sqrt{2n' - 1}$, and treating it as a normal deviate and judging the χ^2 value as significant on the 5 per cent. level whenever the calculated normal deviate exceeds 1·65.

A couple of points ought to be noted about the application of the test. The application of the χ^2 test to frequency data involves an approximation of the type we have noted in the application of the standard error for testing the significance of proportion p. The use of χ^2 test requires that the frequency expected in any class is not too small, that is, 5 or less. It is generally possible to pool the frequencies in the adjacent classes so as to obtain the expectation of pooled frequencies greater than 5. The pooled frequency will have to be treated as belonging to a single class and the degrees of freedom on which the χ^2 is based would consequently be reduced. This is a disadvantage. The pooling of frequencies in the adjacent classes should not therefore be carried on indiscriminately but only when it is essential and justifiable.

Another point to be remembered about this test is that workers are sometimes misled by its name into supposing that it can be applied for testing goodness of fit for all kinds of data. This is not so. The test is applicable only to comparisons of observed and expected values of absolute frequencies; as such, it should not be used for comparing the observed and expected values either of relative frequencies (proportions) or of measurements, as for instance, in comparing observed values of a variate with those calculated from some prediction formula. The gross inapplicability of the test for the latter purpose would be obvious

when it is noted that by a change in the scale of measurement, for example, from pounds to ounces, a value calculated according to formula (5.2) will be changed, altering very often the verdict about significance. Statistics appropriate to tests with measurement data, such as t or F on the other hand, remain unaffected by change of scale.

The following example illustrates the application of the χ^2 test to the comparison of an observed distribution with the normal distribution.

Example 5.2

Table 5.1 gives for plot yields of paddy in a sample survey, the class limits, frequencies observed and the frequencies expected on the basis of the normal distribution respectively. The last two columns give the quantities $O - E$ and $(O - E)^2/E$. The frequencies in the first two classes and in the last two classes are pooled to give expected frequencies more than 5. The expected frequencies were calculated by the method given in Chapter II.

TABLE 5.1

Distribution of plot yields of paddy

Class limits (plot yield in md. per acre)	Observed frequency (O)	Expected frequency (E)	$O - E$	$\frac{(O-E)^2}{E}$
1	2	3	4	5
0— 1·9 ..	1 } 5	3·3 } 6·2	−1·2	0·232
2·0— 3·9 ..	4 }	2·9 }		
4·0— 5·9 ..	8	5·0	3·0	1·800
6·0— 7·9 ..	8	7·8	0·2	0·005
8·0— 9·9 ..	11	11·2	−0·2	0·004
10·0—11·9 ..	17	15·2	1·8	0·213
12·0—13·9 ..	21	19·0	2·0	0·211
14·0—15·9 ..	22	22·1	−0·1	0·000
16·0—17·9 ..	17	23·9	−6·9	1·992
18·0—19·9 ..	23	23·9	−0·9	0·034
20·0—21·9 ..	17	22·2	−5·2	1·218
22·0—23·9 ..	23	19·2	3·8	0·752
24·0—25·9 ..	15	15·5	−0·5	0·161
26·0—27·9 ..	13	11·4	1·6	0·225
28·0—29·9 ..	10	7·9	2·1	0·558
30·0—31·9 ..	7	5·1	1·9	0·708
32·0—33·9 ..	5 }	3·1 }		0·306
34·0 and above ..	0 } 5	3·3 } 6·4	−1·4	
Total ..	222	222		$\chi^2 = 8·419$

We have 16 classes. Keeping the total frequency fixed and assuming two more constants namely, the mean and standard deviation, of the normal distribution to be fitted being the same as those of the observed distribution, the χ^2 calculated is based on 13 degrees of freedom. By reference to the table of χ^2 (Appendix VI) we find that for 13 degrees of freedom this value of χ^2, namely, 8·419 could be exceeded in over 80 per cent. of samples on account of sampling fluctuation. Hence we cannot regard the observed value of χ^2 as indicating a significant departure of the observed class frequencies from a normal distribution.

5*b*.2 Test of Independence: 2 × 2 Table

Another common use of the χ^2 test is in testing independence of classifications in what are known as *contingency* tables. When a group of individuals can be classified in two ways the results of the classification can be set out as in Table 5.2.

TABLE 5.2

Contingency table

Class	A_1	A_2	A_3, etc.
B_1	n_{11}	n_{21}	n_{31}
B_2	n_{12}	n_{22}	n_{32}
B_3	n_{13}	n_{23}	n_{33}
etc.			

Such a table giving the simultaneous classification of a body of data in two different ways is called a contingency table. If there are r rows and c columns, the table is said to be an $r \times c$ table. The simplest table of the kind is 2×2, also known as the fourfold table. The method of testing independence of classifications in a contingency table can be followed very easily with reference to the fourfold table. Consider the following example.

Example 5.3

Table 5.3 gives the classification of 1,282 cotton plants from an F_2 according to corolla colour and leaf shape (Hutchinson, 1934). Are the two classifications independent?

Consider the proportions of plants with yellow corolla and white corolla in each of the two classes, narrow leaf and broad

TABLE 5.3

Observed frequencies of cotton plants in F_2

Leaf shape \ Corolla colour	Yellow	White	Total
Narrow	717 (0·742)	249 (0·258)	966
	a	*b*	
Broad	236 (0·747)	80 (0·253)	316
	c	*d*	
Total ..	953 (0·7434)	329 (0·2566)	1282

leaf. These proportions are shown in brackets against actual frequencies in the table. If the classification for corolla colour bears no relation to the classification for leaf shape, the proportion of yellow and white corolla would be the same in both the leaf shape classes apart from sampling fluctuations. We want to test whether the magnitude of differences between the observed proportions is such as would arise frequently from sampling fluctuations.

If the proportions of yellow and white corolla are, in fact, consistent in the two leaf shape classes, these proportions are best estimated from the totals of the corolla colour classes, that is, 953 and 329. These are shown against the corresponding total frequencies and are seen to be 0·7434 and 0·2566. The expected frequencies in each class are then calculated on the basis of these proportions; thus the expected frequency in the class yellow corolla and narrow leaf is given by $966 \times 0{\cdot}7434$ or 718·1, that in white corolla and narrow leaf is given by 247·9 and so on. The expected frequencies thus calculated are set out in Table 5.4.

TABLE 5.4

Expected frequencies of cotton plants in F_2

Leaf shape \ Corolla colour	Yellow	White	Total
Narrow	718·1	247·9	966
Broad	234·9	81·1	316
Total ..	953	329	1282

Actually it is necessary to calculate the expected frequency in only one cell, the other frequencies being obtained by subtraction from the marginal totals. This shows that the calculated χ^2 is based on only 1 degree of freedom. Thus we have

$$\chi^2 = \frac{(717 - 718 \cdot 1)^2}{718 \cdot 1} + \frac{(249 - 247 \cdot 9)^2}{247 \cdot 9} + \frac{(236 - 234 \cdot 9)^2}{234 \cdot 9} + \frac{(80 - 81 \cdot 1)^2}{81 \cdot 1}$$

$$= \frac{(-1 \cdot 1)^2}{718 \cdot 1} + \frac{(+1 \cdot 1)^2}{247 \cdot 9} + \frac{(+1 \cdot 1)^2}{234 \cdot 9} + \frac{(-1 \cdot 1)^2}{81 \cdot 1}$$

$$= 0 \cdot 026$$

However, it is not necessary to calculate the expected frequencies for the calculation of the χ^2 from the fourfold table. The χ^2 can be directly calculated from the observed frequencies with the help of the following formula. If a, b, c and d denote the frequencies in the four cells of the 2×2 table (*see* Table 5.3), it can be shown that

$$\chi^2 = \frac{(ad - bc)^2 (a + b + c + d)}{(a + b)(a + c)(b + d)(c + d)} \tag{5.3}$$

This formula involves, apart from the factor $(ad - bc)^2$ the four marginal totals and the grand total. Making use of the formula we have

$$\chi^2 = \frac{(1404)^2 \times 1282}{953 \times 329 \times 966 \times 316}$$

which on simplification gives the same value as before, namely, 0·026.

This value of χ^2 is quite small and with 1 degree of freedom has a high probability of being exceeded in sampling, P being between 0·8 and 0·9; hence we have no reason to doubt the independence of the two characters. The probability in this case is high but not too high. It ought to be noted that though a low value of probability leads us to doubt the agreement with the hypothesis it does not mean that the higher the value of the probability the more firmly the hypothesis is established. Thus when the calculated value of χ^2 is very close to zero, it means that agreement between hypothesis and observation is so good that an agreement as good or better would be very rarely obtained. Such

agreement should raise our suspicion as we would be justified in regarding it too good to be believed.

The computation of χ^2 with the help of formula (5.3) is particularly convenient when the numbers in various classes are small. Consider the following example.

Example 5.4

Two batches each of 12 experimental animals, one inoculated and the other not inoculated, were exposed to the infection of a disease. The following frequencies of dead and surviving animals were noted in the two cases. Can the inoculation be regarded effective against the disease?

TABLE 5.5

Contingency table (small frequencies)

	Dead	Survived	Total
Inoculated ..	2	10	12
Not inoculated ..	8	4	12
Total ..	10	14	24

By formula (5.3) we get

$$\chi^2 = \frac{(72)^2 \times 24}{10 \times 14 \times 12 \times 12}$$

$$= 6{\cdot}171$$

This value of χ^2 is highly significant for 1 degree of freedom and shows that inoculation has been effective against disease. It should be noted, however, that the test is not sufficiently accurate when very low frequencies (5 or less) are involved as in the present case. In such cases a correction, which is due to Yates, is useful for application in a 2×2 table. This correction consists in increasing by $\frac{1}{2}$ the frequencies in the two cells along the diagonal whose product is less than the other diagonal product, changing the other frequencies so as to keep the marginal totals fixed, and calculating the χ^2 from the adjusted frequencies. This results in a reduction of the factor $(ad - bc)$ by $n/2$ and consequently the χ^2 value is reduced except when the value of $(ad - bc)$ is less than

$n/4$. This new value of χ^2 provides a more accurate test of the significance of the results.

Applying the correction in the above example we get

$$\chi^2 = \frac{[72 - \frac{1}{2}(24)]^2 \times 24}{10 \times 14 \times 12 \times 12} = 4{\cdot}286$$

Though this value of χ^2 is still significant at the 5 per cent. level of significance, the χ^2 is appreciably reduced. Even Yates' correction does not give a completely accurate test of significance with small frequencies and the exact treatment of such tables can be dealt with by following the method given by Fisher (1950). A discussion of the method is, however, beyond the scope of thi book.

5*b*.3 Test of Independence: $2 \times r$ Table

The test of independence in a $2 \times r$ table is carried out on the same lines as in a 2×2 table as illustrated in Example 5.5 below:

Example 5.5

In Table 5.6 is given the distribution of 886 fields selected in a yield survey on paddy classified according to (i) the type of manuring given and (ii) irrigation. We wish to test whether the type of manuring is independent of the supply of irrigation.

TABLE 5.6

Distribution of paddy fields according to irrigation and manuring

Manure	Irrigated a	Unirrigated a'	$a + a'$	$an_2 - a'n_1$	$\frac{(an_2 - a'n_1)^2}{a + a'}$
No Manure ..	123	413	536	−12158	275778
Farmyard manure ..	81	223	304	3062	30842
Oilcakes ..	8	6	14	3924	1099841
Other manures ..	14	18	32	5172	835924
Total ..	$n_1 = 226$	$n_0 = 660$	886		2242385
					$n_1 n_2 = 149160$

To test whether manuring is independent of the supply of irrigation is to test whether the proportion of irrigated and unirrigated fields is the same irrespective of the manure applied in those fields. We therefore calculate the proportions of irrigated and unirrigated fields in the entire number of fields in the survey, that is 0·2551 and 0·7449 calculated from the respective frequencies, 226 and 660 and the total frequency 886. The χ^2 can then be calculated from the actual and expected frequency in each class by using the formula (5.2). However, it is more convenient to use the following formula:

$$\chi^2 = \frac{1}{n_1 n_2} \sum \frac{(an_2 - a'n_1)^2}{a + a'} \tag{5.4}$$

where a and a' stand for the frequencies in the two-fold classification in a class belonging to the other classification, n_1 and n_2 the corresponding totals in the two-fold classification and the summation extends over all classes. $a + a'$ are class totals in the other classification. The formula is derived as explained below.

The expected frequencies in the two classes of the two-fold classification within each class of the other classification will be

$$\frac{n_1 (a + a')}{n_1 + n_2}$$

and

$$\frac{n_2 (a + a')}{n_1 + n_2}$$

Hence the two classes will contribute

$$\frac{\left\{a - \frac{n_1 (a + a')}{n_1 + n_2}\right\}^2}{\frac{n_1 (a + a')}{n_1 + n_2}} \text{ and } \frac{\left\{a' - \frac{n_2 (a + a')}{n_1 + n_2}\right\}^2}{\frac{n_2 (a + a')}{n_1 + n_2}}$$

to the total χ^2. The sum of these two quantities is

$$\frac{1}{(a + a')(n_1 + n_2)} \left\{\frac{(an_2 - a'n_1)^2}{n_1} + \frac{(a'n_1 - an_2)^2}{n_2}\right\}$$

which reduces on simplification to

$$\frac{(an_2 - a'n_1)^2}{n_1n_2(a + a')}$$

Hence, the sum of contributions of all classes is

$$\frac{1}{n_1n_2}\sum\frac{(an_2 - a'n_1)^2}{a + a'}$$

There are two columns in the above table. If there are r classes in each column, it is obvious that we can fill up $(r - 1)$ frequencies arbitrarily keeping the marginal totals fixed. Therefore the χ^2 table would be referred to with $r - 1$ degrees of freedom. The quantities $(a + a')$, $(an_2 - a'n_1)$ and $(an_2 - a'n_1)^2/(a + a')$ are set out in the last three columns of Table 5.6. The sum of quantities in the last column divided by n_1n_2 gives us the value of χ^2. We have

$$\chi^2 = \frac{2242385}{149160} = 15{\cdot}033 \text{ for } 3\ d.f.$$

which is a highly significant value and shows that irrigation and manuring are not independent; but that the type of manuring applied depends on the supply of irrigation.

5*b*.4 Testing Heterogeneity

The problem of testing heterogeneity is similar to that of testing independence. Thus if there are several sets of data, say r, each divisible into two classes, their frequencies would be set down as a $2 \times r$ table and the χ^2 calculated as in Example 5.5. The technique of testing heterogeneity of data in this manner is very useful in the treatment of data from genetic experiments where we frequently wish to consider the consistency or otherwise of several groups of data, generally families segregating into Mendelian classes. Such an example is given below. In this case we can use another useful modification of the formula, namely

$$\chi^2 = \frac{1}{\bar{p}\bar{q}}\{\Sigma ap - n_1\bar{p}\} \qquad (5.5)$$

where

$$p = \frac{a}{a + a'},\ \bar{p} = \frac{n_1}{n_1 + n_2} \text{ and } \bar{q} = (1 - \bar{p})$$

the symbols a, a', n_1, n_2 having the same meaning as in the previous formula (5.4). For application of the formula we need the marginal totals and the proportions p which should be calculated to 5 or 6 decimal places. The advantage of this method is that it shows actual proportions observed in each class. If the proportions are of no interest we can use formula (5.4), or another form of the formula (5.5), namely

$$\chi^2 = \frac{(n_1 + n_2)^2}{n_1 n_2} \left\{ \sum \frac{a^2}{a + a'} - \frac{n_1^2}{n_1 + n_2} \right\} \qquad (5.6)$$

In place of p we can tabulate the quantities $a^{2\prime}/(a + a')$, and the quantity $n_1^{2\prime}/(n_1 + n_2)$ and thence calculate the value of the curled bracket. Multiplying it by $(n_1 + n_2)^2/n_1 n_2$ we get the χ^2 value. This formula is a little more convenient to use than formula (5.4).

Example 5.6

Five F_2 families of a cotton cross (Baroda lintless × Dharwar glabrous lintless) segregating for the character lintlessness had the number of plants in the linted and lintless classes as shown in the table below (Govande, 1947). Are the families in agreement regarding the ratio linted : lintless that they indicate?

TABLE 5.7

Distribution of lintless and linted cotton plants in F_2 *families*

Family number	Lintless	Linted	Total	p (linted)
1	72	53	125	0·424000
2	37	24	61	0·393443
3	58	27	85	0·317647
4	36	30	66	0·454545
5	40	23	63	0·365079
Total ..	243	157	400	$\bar{p} = 0{\cdot}392500$

The last two columns give respectively the totals $a + a'$, and the observed proportions p. We have

$$\Sigma ap - n_1\bar{p} = 0{\cdot}901768$$

$$\bar{p}\bar{q} = 0{\cdot}3925\,(1 - 0{\cdot}3925) = 0{\cdot}238444$$

Hence using (5.5)

$$\chi^2 = 3{\cdot}782 \text{ with } 4\ d.f.$$

It will be seen from the table that this value of χ^2 is not significant at the 5 per cent. level of significance and the families can be considered to be in agreement with one another regarding the ratio of segregation indicated.

5*b*.5 Test of Independence: $r \times c$ Table

For testing independence or heterogeneity in a general $r \times c$ table (for example, Table 5.2) the expected frequencies in the cells are calculated in the same manner as for the n_{11} class of a 2×2 table, assuming independence of the classifications and marginal totals as known. Thus the frequency expected in the class n_{11} is $\Sigma A_1 . \Sigma B_1 / T$, where ΣA_1 and ΣB_1 are the marginal totals of A_1 and B_1 classes and T is the grand total. With the help of actual and expected frequencies in each class we then calculate the χ^2 value by application of the formula (5.2). If there are r rows and c columns, this χ^2 will be based on $(r-1)(c-1)$ degrees of freedom. For we can fill in a block of $(r-1)$ rows and $(c-1)$ columns of cells arbitrarily and yet have one row and one column left to adjust to make the total frequencies of rows and columns agree with the marginal totals.

5*c*.1 Use of χ^2 Test in Genetic Experiments

The χ^2 test is used perhaps most widely in connection with genetic experiments. Not only can it be employed to test the significance of deviation of an observed segregation from a theoretical one, but it can also be adapted to the simultaneous testing of a number of questions, such as, single factor ratios, linkage and heterogeneity. The method for the last two problems would be explained in the next section. In this section we shall describe and illustrate the use of the test in connection with testing the agreement of observed ratios of segregation with the hypothetical.

Example 5.7

The following table gives the observed frequencies of plants in an F_2 population of chillies, the expected frequencies on the basis of the ratio 1 : 3 : 8 : 4 and the calculation of χ^2 (Deshpande, 1933).

TABLE 5.8

Calculation of χ^2 for classification for colour in F_2 population of chillies

Phenotype	Frequency		$O - E$	$\frac{(O-E)^2}{E}$
	Observed	Expected		
Purple, deep ..	65	68·75	−3·75	0·2045
Purple, medium ..	203	206·25	−3·25	0·0512
Purple, light ..	563	550·00	13·00	0·3073
Green ..	269	275·00	−6·00	0·1309
Total	1100	1100		$\chi^2 = 0·6939$

From the table of χ^2 (Appendix VI) we find that for 3 degrees of freedom the probability of exceeding the calculated value of χ^2 lies between 0·8 and 0·9. The agreement with the theoretical ratio is clearly satisfactory.

The testing of a two class segregation is particularly simple. Only a single degree of freedom is involved and χ^2 can be most easily calculated by the use of the formula

$$\chi^2 = \frac{(l_2 a_1 - l_1 a_2)^2}{l_1 l_2 n} \qquad (5·7)$$

where a_1 and a_2 are the observed frequencies in classes expected to be in the ratio $l_1 : l_2$ and $n = a_1 + a_2$. The formula takes the following forms for the commonly occurring ratios of 1 : 1, 3 : 1, 9 : 7 and 15 : 1.

Ratio	χ^2
1 : 1	$\frac{(a_1 - a_2)^2}{n}$
3 : 1	$\frac{(a_1 - 3a_2)^2}{3n}$
9 : 7	$\frac{(7a_1 - 9a_2)^2}{63n}$
15 : 1	$\frac{(a_1 - 15a_2)^2}{15n}$

5c.2 Size of Experiments

One point often overlooked in connection with tests of agreement with genetic ratios is that the observed segregation may agree with more than one ratio and in consequence no decisive conclusion can be reached unless the size of the experiment is adequate. Suppose, for instance, we have a family of 16 individuals segregating into two classes, with 10 individuals of one kind and 6 of the other. Such a family can well arise either from a 1 : 1 or a 3 : 1 ratio of segregation and we cannot confidently assert on the basis of such a small family whether it agrees with one or the other of the two ratios. To be able to distinguish between the two ratios we need a family of adequate size. What constitutes this adequate size is explained below.

Suppose we have data for n plants segregating into two classes and we wish to ascertain whether these are in agreement with a 1 : 1 or a 3 : 1 ratio. Such a question can obviously arise when the number of plants in the smaller class is greater than $n/4$. If a_1 and a_2 are the numbers in the two classes the values of χ^2 on the basis of the two ratios will be

$$\frac{(a_1 - a_2)^2}{n}$$

and

$$\frac{(a_1 - 3a_2)^2}{3n}$$

respectively. To be able to distinguish between the two, the value of n has to be so large that on either hypothesis the χ^2 should have a low probability. If we choose the 5 per cent. level of probability as sufficiently low, then each of the above values should be equal to 3·841 at least, the value of χ^2 for $P = 0{\cdot}05$ and 1 degree of freedom. We thus have two equations in two unknowns, n and a_1, a_2 being equal to $n - a_1$. Solving these we get

$$\begin{aligned} n &= 3{\cdot}841\,(2 + \sqrt{3})^2 \\ &= 53{\cdot}5 \end{aligned}$$

In general it can be shown that for distinguishing a segregation l_1 : 1 from l_2 : 1 with the 5 per cent. level of probability

$$n = \frac{3{\cdot}841\,(1 + \sqrt{l_1 l_2})^2}{(\sqrt{l_1} - \sqrt{l_2})^2} \qquad (5{\cdot}8)$$

We can also calculate the size of family on the basis of the standard error test. Clearly the ratio $d/S.E.\ (d)$ should be so large in each case as to be significant at $P = 0{\cdot}05$ and therefore must be equal to $1{\cdot}96$. Hence we have the following equations:

$$a_1 - \frac{n}{2} = 1{\cdot}96\sqrt{\frac{n}{4}}$$

$$\frac{3n}{4} - a_1 = 1{\cdot}96\sqrt{\frac{3n}{16}}$$

By solving these, we get

$$n = \{1{\cdot}96\,(2 + \sqrt{3})\}^2$$
$$= 53{\cdot}5$$

The value is in agreement with the previous one in virtue of the fact that a χ^2 for a single degree of freedom is the square of a normal deviate, although this will not be so in other instances.

5c.3 The Partition of χ^2

The property of χ^2 which makes it so useful for a number of purposes and especially in the analysis of genetic experiments is that the sum of a number of χ^2 values derived from independent sources is distributed as a χ^2 with number of degrees of freedom equal to the total of the numbers of degrees of freedom on which the individual values are based. This permits us to partition χ^2 into components assignable to different factors. Thus, when we have a number of families of individuals segregating for the same ratio, we can total up the χ^2 values obtained from each family and partition this total into components assignable to deviation from the theoretical ratio and heterogeneity between families. The component assignable to deviation from the theoretical ratio can be obtained by pooling the classes over all the families and calculating the χ^2 due to deviation from these totals. The χ^2 due to heterogeneity is obtained by subtraction from the total χ^2. The following example would make the procedure clear.

Example 5.8

Test the 5 families in Example 5.6 for agreement with a 9 : 7 ratio and heterogeneity among families.

The following table sets out the values of χ^2 calculated from each of the 5 families and from their totals for the two classes by the use of the formula

$$\chi^2 = \frac{(7a_1 - 9a_2)^2}{63\,n}$$

a_1 and a_2 being the number of individuals in the lintless and linted classes and $n = a_1 + a_2$.

Family number	$7a_1 - 9a_2$	$63n$	χ^2	*D.F.*
1	27	7875	·093	1
2	43	3843	·481	1
3	163	5355	4·962	1
4	−18	4158	·078	1
5	73	3969	1·313	1
Due to deviation of total	288	25200	3·291	1

The sum of χ^2 values derived from each of the families gives us $\chi^2 = 6{\cdot}957$ for 5 degrees of freedom. The χ^2 for deviation calculated from the totals of classes, gives $\chi^2 = 3{\cdot}291$ for 1 degree of freedom. Subtracting this from the total χ^2 for 5 degrees of freedom we obtain $\chi^2 = 3{\cdot}666$ for 4 degrees of freedom due to heterogeneity among families. Since there are 5 families there should be 4 degrees of freedom for differences between them, that is, for heterogeneity. We may set down the results in the following form:

Source	χ^2	*D.F.*
Deviation from 9 : 7 ratio ..	3·291	1
Heterogeneity	3·666	4
Total ..	6·957	5

The analogy of this procedure with the analysis of variance can be easily seen.

By reference to the table of χ^2 we find that a value of 3·291 is not significant for 1 degree of freedom at $P = 0{\cdot}05$ though it

has a probability less than 0·1. Hence we cannot regard the deviation as significant but the low value of probability gives some cause for suspicion. A χ^2 value of 3·666 for 4 degrees of freedom has a fairly high probability, above 0·3, and we have no reason to doubt the agreement of the families with one another in respect of the segregation ratio.

It might be noted that χ^2 calculated in this example differs slightly from that calculated in Example 5.6, namely 3·782. This difference is due to the difference in the hypotheses tested in the two cases. While in Example 5.6 merely the agreement of families with one another in respect of their ratio of segregation was tested, in Example 5.8 the agreement of families in conforming to the hypothetical ratio 9: 7 is tested.

If the deviation from the hypothetical ratio is found to be significant, then the method of Example 5.6 provides the appropriate test of heterogeneity. When the deviation from the hypothetical ratio is non-significant, the two methods would give χ^2 values close to each other. However, if significant heterogeneity is revealed by the data, the agreement of the data as a whole with the hypothetical ratio will carry little meaning, since we cannot conclude that the segregation ratio is the same for all families. In such a contingency we shall have to consider the ratio of segregation for each family separately.

5*c*.4 Detection of Linkage

The method of partitioning χ^2 is also of great use in the detection of linkage. A geneticist studying the inheritance of characters concerns himself with not only one but several characters and studies populations segregating simultaneously for two or more characters. When the material under study is segregating simultaneously for two characters, he wants to know whether the two characters are inherited independently or tend to be associated. In doing this he has first to test the segregation of two characters separately for their agreement with the respective expected ratios. If he is satisfied about the segregation of individual characters he can proceed to test their independence. If two characters are segregating independently in the ratio l_1: 1 and l_2: 1 respectively, their joint segregation into 4 classes would be in the ratio $l_1 l_2$: l_1: l_2: 1. Thus we obtain the double backcross

ratio of 1 : 1 : 1 : 1 and by selfing a double heterozygote showing dominance for both characters, a ratio of 9 : 3 : 3 : 1, provided the two characters are segregating independently. We can calculate χ^2 by taking deviations of observed frequencies in the 4 classes from their expectations on the assumption of independence. This would represent the χ^2 due to linkage if the segregation of individual characters is strictly according to expectation. This is hardly ever the case and deviations from theoretical ratios also contribute to the value of χ^2 calculated above. We have to remove these contributions from the gross value of χ^2 before we get the χ^2 for linkage alone, which will have 1 degree of freedom.

This object is achieved in a systematic manner by partitioning the total χ^2 for 3 degrees of freedom into components by the use of *orthogonal functions*. By orthogonal functions we mean functions which give independent comparisons. Suppose the observed frequencies in the four classes are a_1, a_2, a_3, a_4 respectively. Then we can choose various linear functions of these frequencies such as

$$U = k_1a_1 + k_2a_2 + k_3a_3 + k_4a_4 \tag{5.9}$$

where k's are arbitrary constants.

If p_1, p_2, p_3 and p_4 are the expected frequencies in the classes corresponding to observed frequencies a_1, a_2, a_3, a_4 and if $\Sigma pk = 0$, the expected value of U will be zero and the variance given by

$$V(U) = n\Sigma pk^2 \tag{5.10}$$

The ratio of the square of the function to its variance is a χ^2 with 1 degree of freedom. Many such functions may be chosen, but they will not be all orthogonal or independent. The independence of any two such functions is tested by finding whether

$$\Sigma pkk' = 0$$

where k' stands for the corresponding coefficients in the second function. If it is, the two functions are independent.

Suppose we are considering the joint segregation of two characters each segregating into two classes A and a, and B and b, the ratio of each segregation being 3 : 1. If the two are segre-

gating independently the ratio of the four classes AB, Ab, aB and ab will be a 9:3:3:1 ratio. For testing the segregation of factor A alone we calculate

$$\chi^2 = \frac{\{(a_1 + a_2) - 3(a_3 + a_4)\}^2}{3n} \tag{5.11}$$

with 1 degree of freedom. The fraction written within the curled brackets in the numerator is a linear function of the observed frequencies of the type U with

$$k_1 = k_2 = 1, \; k_3 = k_4 = -3$$

By hypothesis we know that

$$p_1 = \frac{9}{16}, \; p_2 = p_3 = \frac{3}{16} \text{ and } p_4 = \frac{1}{16}$$

giving

$$\Sigma pk = \frac{9}{16} \times 1 + \frac{3}{16} \times 1 - \frac{3}{16} \times 3 - \frac{1}{16} \times 3 = 0$$

Hence the variance of the function under consideration

$$\begin{aligned} V(U) &= n\Sigma pk^2 \\ &= n\left\{\frac{9}{16}(1)^2 + \frac{3}{16}(1)^2 + \frac{3}{16}(-3)^2 + \frac{1}{16}(-3)^2\right\} \\ &= 3n \end{aligned}$$

which is the divisor in the formula for χ^2.

Similarly for the segregation of factor B we use the formula

$$\chi^2 = \frac{\{(a_1 + a_3 - 3(a_2 + a_4)\}^2}{3n} \tag{5.12}$$

It is easy to show that in this case also $\Sigma pk' = 0$ and $n\Sigma pk'^2 = 3n$. We further notice that the two functions are independent since

$$\Sigma pkk' = \frac{9}{16} \cdot 1 - \frac{3}{16} \cdot 3 - \frac{3}{16} \cdot 3 + \frac{1}{16} \cdot 9 = 0$$

The χ^2 due to linkage of the two factors is given by a linear function of the type U which is independent of both of the two previous functions. Such a function is given by $\Sigma kk'a$ where

it will be noticed that the coefficients of the observed frequencies a_1, a_2, $\cdots$ are the products of corresponding coefficients of the two previous functions. Thus if the two previous functions are

$$U_1 = a_1 + a_2 - 3a_3 - 3a_4$$

and

$$U_2 = a_1 - 3a_2 + a_3 - 3a_4$$

the new function will be

$$U_3 = a_1 - 3a_2 - 3a_3 + 9a_4$$

The independence of the new function and the previous functions can be easily proved by calculating the sum $\Sigma pkk'$. Similarly Σpk can be seen to be zero, while the sampling variance of the function will be given by

$$n\Sigma pk^2 = n\left\{\frac{9}{16}(1)^2 + \frac{3}{16}(-3)^2 + \frac{3}{16}(-3)^2 + \frac{1}{16}(9)^2\right\}$$

$$= 9n$$

Hence the χ^2 due to linkage will be given by

$$\chi^2 = \frac{(a_1 - 3a_2 - 3a_3 + 9a_4)^2}{9n} \qquad (5.13)$$

With 4 class frequencies there will be only 3 independent comparisons possible giving 3 degrees of freedom for the χ^2. Since we have obtained the χ^2 values for 3 independent comparisons their total must check up with the χ^2 for 3 degrees of freedom calculated directly. This is always so within the limits of accuracy of the computations and serves as a useful check on the correctness of the calculations. The following example would show the practical working of the method.

Example 5.9

Linseed type 11 has a deep lilac petal and a deep purple stigma while linseed type 121 has a lilac petal and a white stigma (Shaw *et al.*, 1931). In F_1 type 121 phenotype was dominant and the frequencies observed in F_2 are shown in Table 5.9.

TABLE 5.9

F_2 *frequencies for petal and stigma colours in linseed*

	Lilac petal		Deep lilac petal	
	White stigma	Purple stigma	White stigma	Purple stigma
	AB	*Ab*	*aB*	*ab*
Observed	357	37	33	94
Expected on the 9 : 3 : 3 : 1 ratio ..	293·06	97·69	97·69	32·56

Total χ^2 (3 *d.f.*) $= 13\cdot949 + 37\cdot702 + 42\cdot835 + 115\cdot918 = 210\cdot404$

$$\chi^2_{\text{petal colour}} = \frac{\{357+37-3\,(33+94)\}^2}{3\times521} = \frac{(13)^2}{1563} = 0\cdot108$$

$$\chi^2_{\text{stigma colour}} = \frac{\{357+33-3\,(37+94)\}^2}{3\times521} = \frac{(-3)^2}{1563} = 0\cdot006$$

$$\chi^2_{\text{linkage}} = \frac{\{357-3\,(33+373)+9\,(94)\}^2}{9\times521} = \frac{(993)^2}{4689} = 210\cdot290$$

Thus we have the following results:

Source	χ^2	*D.F.*
Petal colour factor	0·108	1
Stigma colour factor	0·006	1
Linkage	210·290	1
Total ..	210·404	3

The total is in agreement with the χ^2 calculated directly.

An inspection of the results shows that the data provide a very strong evidence of linkage, giving a χ^2 value over 200 for a single degree of freedom. This inference is now possible since we have partitioned χ^2 into components due to different causes. We might have inferred from the total χ^2 for 3 degrees of freedom that the segregation was not in agreement with the 9 : 3 : 3 : 1 ratio, but whether the discrepancy was due to failure of single factor ratios to conform with their theoretical expectations or whether the discrepancy was due to linkage could not have been settled without the partition of χ^2. In this lies the merit of the method; it enables us to locate precisely the source

of the discrepancy. The partitioning of χ^2 for other ratios of segregation can be done in a similar manner by deriving the appropriate functions.

5*c*.5 HETEROGENEITY: LINKAGE RATIO

When we have several groups of data of the same type we can test them individually for evidence of linkage. We can also test them for heterogeneity by pooling the χ^2 due to the linkage degree of freedom from each group and subtracting from the total the χ^2 due to linkage calculated from the pooled data. Thus we can test several groups of data for linkage as well as heterogeneity in the same way as we did in the case of single factor ratios earlier.

5*d*.1 ESTIMATION OF LINKAGE

When the existence of linkage has been conclusively proved, it becomes necessary to obtain a measure of the intensity of linkage. While the testing of linkage involves no hypothesis as to the nature of linkage, being merely the demonstration of the fact that the two characters do not segregate independently, estimation of its intensity can be done only in the light of some hypothesis about its nature. This is provided by the modern genetic theory which regards these deviations from independence as resulting from the genes controlling the characters being situated on the same chromosome. Thus the fact that lilac petal and white stigma in Example 5.9 appear together much more frequently than expected, finds explanation in the hypothesis that the genes producing these characters in the progeny are located on the same chromosome. The fact that we have plants in which purple stigma and lilac petal or white stigma and deep lilac petal are associated is explained by the phenomenon of crossing over, that is, exchange of parts between the chromosomes of a pair. The intensity of linkage is measured in inverse sense as the fraction of the total number of chromosome pairs in which the change-over takes place at gametogenesis. This is known as the *recombination fraction*. The smaller this fraction, the more intense the linkage.

Several methods of estimating this fraction from the observed data have been proposed from time to time and which method to choose for estimation is a problem which naturally puzzles the

research worker. Fortunately, as a result of the development of the statistical theory of estimation, we have now criteria by which to judge the various methods. From the point of view of the problem at hand the following two are important. One is that the estimate obtained should tend to the theoretical value as the sample is enlarged and the other is that the estimate should have the lowest possible variance for the type of data. The first criterion, known as the criterion of *consistency*, ensures that any bias in the estimate decreases to a negligible magnitude as the sample size becomes large, while the second, the criterion of *efficiency*, ensures that the estimate will be as precise as possible, under the particular conditions. We propose to give in the following pages methods in common use which are satisfactory from the point of view of these criteria and compare their advantages and disadvantages.

For making an estimate of the recombination fraction from observed frequencies we need to know the relation between these frequencies and the recombination fraction. In general, this is not difficult. Suppose we obtain the progeny of a double heterozygote crossed with a double recessive (that is, progeny of a double backcross) with dominance of both factors. In this case the double recessive parent will produce gametes of one kind only, of the *ab* type, while the heterozygous parent will produce 4 types of gametes, the *AB*, *Ab*, *aB* and *ab* types. If the assortment of the genes is independent these four types are produced in equal numbers, while if the *A*, *B* and *a*, *b* genes are linked initially, that is, are on the same chromosome, the 4 types would occur with the relative frequencies,

$$\tfrac{1}{2}(1-p),\ \tfrac{1}{2}p,\ \tfrac{1}{2}p,\ \tfrac{1}{2}(1-p)$$

respectively, where p is the recombination fraction. Since these gametes fuse with the same type of gamete, the different kinds of progeny produced will also be in the above proportions and p can be directly estimated as the relative frequency of the two middle classes taken together.

Similarly we can work out the expected frequencies in case of the progeny obtained by selfing a double heterozygote as in the F_2 of a cross. Here we have to consider the possibility of the

recombination fraction having different values at the male and female gametogenesis. If p and p' are the values of the fraction for the two parents, both assumed to be in the coupling phase, the frequencies of gametes of the four kinds produced by the parents would be as follows:

AB	Ab	aB	ab
$\frac{1}{2}(1-p)$	$\frac{1}{2}p$	$\frac{1}{2}p$	$\frac{1}{2}(1-p)$
$\frac{1}{2}(1-p')$	$\frac{1}{2}p'$	$\frac{1}{2}p'$	$\frac{1}{2}(1-p')$

Since the double recessive class can be obtained only by the fusion of ab gametes from both parents, its frequency will be given by $\frac{1}{4}(1-p)(1-p')$. Since the total frequency of the recessive class for any factor is $\frac{1}{4}$, the two classes with one factor in the recessive condition would each have the relative frequency $\frac{1}{4}\{1-(1-p)(1-p')\}$. The frequency of the double dominant class can be obtained by subtracting these three frequencies from unity and would be found to be $\frac{1}{4}\{2+(1-p)(1-p')\}$. It will be seen that all the four frequencies depend upon the single quantity $(1-p)(1-p')$. If we denote this quantity by θ the frequencies in the four phenotypic classes would be

AB	Ab	aB	ab
$\frac{1}{4}(2+\theta)$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$

Since the frequencies depend upon θ only, this is the quantity we can estimate from the observed frequencies for this type of data.

5*d*.2 EMERSON'S METHOD

Emerson's method of estimating θ from F_2 data is to put

$$\theta = \frac{E-M}{n} \tag{5.14}$$

where

E = sum of observed frequencies of AB and ab classes,

M = sum of observed frequencies of Ab and aB classes,

n = $E+M$, *i.e.*, total number of individuals in all classes.

If we substitute the expected values of the frequencies in the expression on the right of (5.14) we get θ on simplification which shows that formula (5.14) will provide an unbiased estimate of

θ. The standard error of the estimate of θ can be obtained from the formula

$$V_{\theta} = \frac{1 - \theta^2}{n} \tag{5.15}$$

Applying this method to the data of Example 5.9, we get

$$\hat{\theta} = \frac{\{357 + 94 - (33 + 37)\}}{521} = 0{\cdot}7313$$

Hence

$$V_{\hat{\theta}} = \frac{1 - (0{\cdot}7313)^2}{521} = 0{\cdot}0008929$$

If we assume $p = p'$, then

$$\hat{\theta} = (1 - \hat{p})^2 = 0{\cdot}7313$$

and

$$(1 - \hat{p}) = \sqrt{0{\cdot}7313} = 0{\cdot}854$$

or

$$\hat{p} = 1 - 0{\cdot}854 = 0{\cdot}146$$

Thus the recombination fraction comes out to be 14·6 per cent. The standard error of this estimate can be found from V_{θ} by making use of the formula

$$V_p = \frac{V_{\theta}}{4\theta} \tag{5·16}$$

where V_p is the variance of p.

Hence in the above case

$$V_{\hat{p}} = \frac{0{\cdot}0008929}{4 \times 0{\cdot}7313} = 0{\cdot}0003053$$

$$\therefore \quad S.E.(\hat{p}) = \sqrt{0{\cdot}0003053}$$

$$= 0{\cdot}0175$$

5*d*.3 Method of Maximum Likelihood

This is a method of unique importance, for it has been shown that in large samples no other method will give an estimate with a smaller sampling variance than the one given by this method (Fisher, 1921). This is, therefore, the most efficient method of estimation and the efficiency of other methods is judged by a comparison with this method. The method can be applied

to any type of data for which we can correctly work out the expected frequencies in terms of the quantity to be estimated (θ and p, in our case). The method is, therefore, applicable to data of various types such as data from a double backcross, F_2, etc. If p_1, p_2, p_3 and p_4 are the frequencies expected in the four classes (p_1, p_2, ... being expressions involving p or θ) and a_1, a_2, a_3 and a_4 are the observed frequencies, the method consists in obtaining that value of θ and hence of p which will maximize the value of the expression

$$a_1 \log p_1 + a_2 \log p_2 + a_3 \log p_3 + a_4 \log p_4$$

This is done by differentiating this expression with respect to θ and putting the result equal to zero. Thus in case of F_2, for which we have worked out the expectations, it can be shown that

$$\frac{a_1}{2+\theta} - \frac{a_2}{1-\theta} - \frac{a_3}{1-\theta} + \frac{a_4}{\theta} = 0$$

On simplification, this gives us the equation

$$2a_4 + \{a_1 - 2(a_2 + a_3) - a_4\}\theta - n\theta^2 = 0 \qquad (5.17)$$

which is a second degree equation in θ. By solving it we get the value of θ required. Substituting the values of a_1, a_2, ... from Example 5.9, we get the equation

$$188 + 123\theta - 521\theta^2 = 0$$

which gives

$$\hat{\theta} = 0{\cdot}7302$$

By assuming $p = p'$, we get as in the previous calculation

$$\hat{p} = 1 - \sqrt{0{\cdot}7302} = 0{\cdot}146$$

The results obtained by the two methods differ only slightly. The variance of the estimate of θ obtained by the method of maximum likelihood is given by

$$V_\theta = \frac{2\theta(1-\theta)(2+\theta)}{n(1+2\theta)} \qquad (5.18)$$

This gives us

$$V_{\hat{\theta}} = 0{\cdot}0008397$$

From this we obtain as in the previous case

$$V_{\hat{p}} = \frac{0{\cdot}0008392}{4 \times 0{\cdot}7302} = 0{\cdot}0002873$$

which gives

$$S.E.(\hat{p}) = \sqrt{0{\cdot}0002873} = 0{\cdot}0170$$

5*d*.4 Product-Ratio Method

Another useful method of estimating linkage from a four class segregation is the *product-ratio* method. In this method we equate the ratio of the product of extreme classes to the product of middle classes, with its theoretical value. This gives us the required equation for obtaining θ. Thus for the case of F_2 considered,

$$Q = \frac{a_1 a_4}{a_2 a_3} = \frac{(2 + \theta)\theta}{1 - 2\theta + \theta^2} \tag{5.19}$$

This is a quadratic in θ. Solving it we get

$$\theta = \frac{1 + Q - \sqrt{1 + 3Q}}{Q - 1} \tag{5.20}$$

From Example 5.9 we have

$$Q = \frac{a_1 a_4}{a_2 a_3}$$

$$= \frac{357 \times 94}{37 \times 33}$$

$$= 27{\cdot}484$$

Substituting this in (5.20) we get

$$\hat{\theta} = 0{\cdot}728$$

The formula for the variance of the product ratio estimate of θ is the same as that in the case of the maximum likelihood estimate given by (5.18). Hence we can calculate $\hat{p}$ and $V_{\hat{p}}$ as in the previous two cases and it will be found that the values differ only slightly from those calculated by the maximum likelihood method.

5*d*.5 COMPARISON OF THE METHODS

The three methods would be found adequate for the simpler types of experiments generally conducted. Their merits and demerits and their suitability in particular cases should be noted carefully. Emerson's method is generally less efficient than the other two methods, as it gives an estimate with a larger standard error. When the linkage is tight, however, the standard error of this estimate is nearly the same as those of the other two estimates and this method may then be preferred on account of its simplicity. The other two methods are equally efficient as can be seen from the fact that they have the same variance for the estimate obtained. The product ratio method has the advantage of being less influenced by viability disturbances which cannot be taken into account by the maximum likelihood method. With tight linkage, however, the recombination classes may have very few members and one class may at times be absent. In such a case the product ratio method gives zero as the estimate of the recombination fraction, while the other class shows that recombination has occurred. In this situation the maximum likelihood estimate is definitely superior and even the estimate obtained by Emerson's method is preferable. Wherever the frequencies expected in the different classes can be calculated exactly, the maximum likelihood method is superior. This method, besides being applicable to diverse types of data, is easily adapted to obtaining the joint estimate of the recombination fraction from different types of data as also to estimating heterogeneity between groups of data. It is beyond the scope of this book to give the use of the method for this purpose, for which reference may be made to *Measurement of Linkage in Heredity* by K. Mather.

CHAPTER VI

CORRELATION AND REGRESSION

6*a*.1 INTRODUCTION

WE have upto now considered the measurement of variation in only a single variable. We shall now consider the simultaneous variation of two variables. It often happens that changes in one variable are accompanied by changes in another and that a definite relation exists between the two. In other words, there is a *correlation* between the two variables. Thus in our example of variation in the length of earhead and number of grains per head in wheat we find that heads of greater length, in general, possess a higher number of grains.

When two variables change together in such a way that an increase in one variable is accompanied by an increase in the other, the variables are said to be *positively correlated.* An example of perfect positive correlation is the relationship between the changes in temperature and the length of an iron bar. For every single degree rise in temperature, the length of the bar increases by a fixed amount and the temperature and the length of the bar are positively correlated. In biological measurements, the relationship between the two variables is not likely to be so complete as this, but it is obvious that certain characters may be expected to show a strong correlation. For instance, in general we shall expect a strong positive correlation between the heights of human, beings and their weights and this has, in fact, been found to exist. Similarly, in plants a number of characters show such correlation, as tiller number and yield of wheat plants, number of bolls and yield in cotton plants and so on. Other well-known examples of correlation are provided by heights of related persons such as heights of father and son, heights of brothers, etc. Should an increase in one variable go hand in hand with a decrease in the other, these two variables are said to be *negatively correlated.* If there is no relationship between two variables, they are said to be independent or uncorrelated.

6*a*.2 CORRELATION TABLE

Analogous to the frequency table (Table 1.3) for a single variable we have the correlation table for the simultaneous distribution of two variables. The correlation table is a two-way classification similar to the one represented by Table 5.2 with the difference that the classifications along rows and columns represent intervals of two quantitative variables. This will be clear from Table 6.1 which gives the correlation table for the data of earheads of wheat given in Table 1.2. It will be seen that the contingency table and the correlation table are related to each other as Tables 1.1 and 1.3. When one or both variables in a two-way table are qualitative it is a contingency table. When both are quantitative it becomes a correlation table. In the previous chapter we have dealt with the significance of association in a contingency table. In this chapter we shall deal with the measurement and testing of the significance of association in the correlation table.

In Table 6.1, the classification for length of earhead is done along the rows and classification according to number of grains, along the columns. Consider any row or column of cells. In the row of cells characterized by 15 grains to the ear, for instance, there are two ears with an average length of 5·5 cm., 1 with 6 cm., 8 with 6·5 cm., 2 with 7 cm., 3 with 7·5 cm. and 1 with 8 cm. This shows how in a total of 17 ears with an average number of 15 grains to the ear the variation in length is distributed. In this way, all the 400 earheads are grouped together according to their respective lengths and the number of grains per ear. The shape of the distribution indicates the extent and the nature of the correlation; the more elliptical the distribution the stronger the correlation. If the long axis of the ellipse slopes from left to right the correlation is positive; negative correlation is indicated when the long axis of the ellipse slopes from right to left, and if the distribution in the correlation table is not markedly elliptical the characters are not correlated (Fig. 6.1).

From the nature of the ellipse enclosing the area covered by the data in Table 6.1, it is obvious that there is a strong positive correlation between the length of head and the number of grains per head in Pusa 12 wheat.

TABLE 6.1

Correlation table for length of earhead and number of grains per ear in Pusa 12 wheat, 1930–31

Number of grains		Length in cm. x																				
		5·5	6·0	6·5	7·0	7·5	8·0	8·5	9·0	9·5	10·0	10·5	11·0	11·5	12·0	12·5	13·0	13·5			
y	Dev. from *A.O.*	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16		f_y	$\Sigma f_y . d_y$
10	0	(0) 1																		1	0
15	1	(0) 2	(1) 1	(2) 8	(3) 2	(4) 3	(5) 1													17	40
20	2				(6) 4	(8) 4	(10) 6	(12) 5	(14) 4	(16) 2										25	264
25	3					(12) 1	(15) 3	(18) 18	(21) 23	(24) 22	(27) 13	(30) 4	(33) 2							86	1929
30	4						(20) 1	(24) 3	(28) 12	(32) 27	(36) 38	(40) 22	(44) 15	(48) 5	(52) 1		(60) 1			125	4552
35	5							(30) 6	(35) 3	(40) 6	(45) 9	(50) 18	(55) 12	(60) 13	(65) 8	(70) 2				77	3930
40	6									(48) 1	(54) 5	(60) 10	(66) 7	(72) 10	(78) 13	(84) 3	(90) 3	(96) 3		55	3924
45	7											(70) 1	(77) 1	(84) 3	(91) 1	(98) 2	(105) 1			9	791
50	8														(104) 1		(120) 1	(128) 2		4	480
55	9																	(144) 1		1	144
f_x		3	1	8	6	8	11	32	42	58	65	55	37	31	24	7	6	6		400	16054

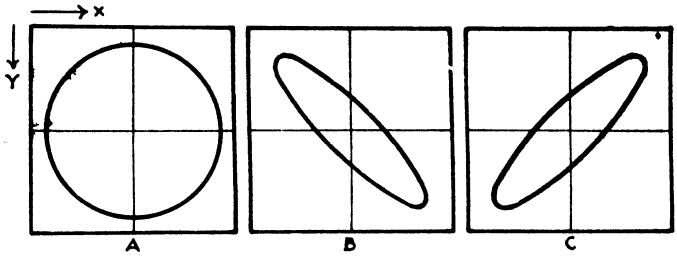


FIG. 6.1. Dispersion of data in correlation table.
(A) Uncorrelated, (B) Positive, and (C) Negative.

6*a*.3 COEFFICIENT OF CORRELATION

The intensity of correlation is measured by a coefficient, usually indicated by the symbol ρ which is computed according to the formula

$$\rho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \tag{6.1}$$

where

$$\sigma_{xy} = \frac{\Sigma f(x - \bar{x})(y - \bar{y})}{N} = \frac{\Sigma f.d_x.d_y}{N}$$

and denotes, what is known as the *mean product moment* or the *covariance* between x and y, and σ_x and σ_y are the standard deviations of x and y respectively. The estimate of ρ calculated from a sample is denoted by r. Since the deviations d_x and d_y can be both positive and negative the quantity $\Sigma f.d_x.d_y$ and hence the covariance σ_{xy} can assume both positive and negative values. In either case, however, it cannot exceed the product $\sigma_x\sigma_y$ and consequently the correlation coefficient ρ necessarily varies between $+1$ and -1, a perfect positive correlation being indicated by $+1$ and a perfect negative correlation by -1. The direct application of formula (6.1) is very laborious but a good short-cut method is described below.

In the short-cut method the lowest class value in each variable is taken as the arbitrary origin and higher class values are considered in serial order as deviations from this. Thus for the length of earheads the smallest class value is 5·5 cm. and taking this as the arbitrary origin equal to 0, the class value of 6 cm. is

a deviate of 1, the class value of 6·5 cm. is a deviate of 2, the class value of 8·5 cm. is a deviate of 6 and the largest class value of 13·5 cm. is a deviate of 16. A similar codification or transformation is carried out for the class values for the number of grains. In the correlation Table 6.1, in each cell containing a number expressing the frequency of that particular class, we insert a number which is the product of the deviations from the two arbitrary origins. For example, in the cell containing earheads with class value 10 cm. and number of grains 30 the deviations from the arbitrary origins are 9 and 4 respectively and their product, 36, is shown in brackets. The sum of the products of the frequencies and the numbers in brackets over all the cells of each row are denoted by $\Sigma f.d_x'.d_y'$ and are shown in the last column to the right of the table. The total of this column divided by the total number of individuals is called the mean product moment about the arbitrary origin. A correction has to be applied to this moment (the mean product moment about the arbitrary origin) to reduce it to the mean product moment about the true means of x and y and to convert it into proper units, that is, to make allowance for the fact that we take the deviations from an arbitrary origin and that we measure the deviation in arbitrary units and not in terms of the actual class intervals. The correction for the former factor is the product of the two corrections which are to be applied for calculating the means from the arbitrary origin. The result obtained after applying this correction is multiplied by the product of the class intervals (I_x and I_y) to arrive at the mean product moment in actual units. The corrected mean product moment divided by the product of the two standard deviations gives the coefficient of correlation. In the present example the calculations are shown as follows.

Example 6.1

Calculate the coefficient of correlation between the length of earhead and the number of grains per ear from the data given in Table 6.1.

The mean product moment about arbitrary origin

$$= \frac{\Sigma f.d_x'.d_y'}{N} = \frac{16054}{400} = 40{\cdot}135$$

The product of the two correction factors

$$(C.F)_x \times (C.F)_y = 8{\cdot}955 \times 4{\cdot}110 = 36{\cdot}805$$

The actual mean product moment

$$\left\{\frac{\Sigma f.d_x'.d_y'}{N} - (C.F)_x(C.F)_y\right\} I_x.I_y$$

$$= (40{\cdot}135 - 36{\cdot}805)(5 \times 0{\cdot}5) = 8{\cdot}325$$

The product of the two standard deviations is

$$1{\cdot}441 \times 6{\cdot}999 = 10{\cdot}085$$

$$\therefore \quad \rho = \frac{8{\cdot}325}{10{\cdot}085} = 0{\cdot}8255$$

It should be noted in calculating the correlation coefficient that while Sheppard's correction is applied in calculating the two variances, no such correction for grouping is required for the covariance term. Sheppard's correction, however, introduces some disturbance in the estimates of correlation coefficient calculated from small samples, giving at times values exceeding unity. In such cases the correlation coefficient should be calculated directly without grouping, thus dispensing with the correction. For tests of significance of the correlation the estimate of the correlation coefficient r obtained without the use of the correction should be employed.

In the above example we calculated the correlation coefficient from the mean product moment and the standard deviations. Ordinarily it is not necessary to calculate separately the values of the covariance and the two standard deviations. It will be recalled that

$$\sigma_x = \sqrt{\frac{\Sigma f.d_x^2}{N}}$$

and

$$\sigma_y = \sqrt{\frac{\Sigma f.d_y^2}{N}}$$

and the correlation coefficient can therefore be expressed as

$$\rho = \frac{\Sigma f.d_x.d_y}{N\sqrt{\frac{\Sigma f.d_x^2}{N}}\sqrt{\frac{\Sigma f.d_y^2}{N}}}$$

$$= \frac{\Sigma f.d_x.d_y}{\sqrt{(\Sigma f.d_x^2)\,(\Sigma f.d_y^2)}}$$

$$= \frac{\text{sum of products } (x-\bar{x})\,(y-\bar{y})}{\sqrt{S.S \text{ of } (x-\bar{x}) \times S.S \text{ of } (y-\bar{y})}} \qquad (6.2)$$

In other words we can calculate the correlation coefficient by dividing the sum of products of deviations from the means by the square-root of the product of the sums of squares of deviations from the respective means of the two variables.

If the number of individuals for which the x and y measurements are available is small, then the data need not be grouped in classes in the form of a correlation table. The following example shows the computation of the coefficient of correlation between the breakage of rice grains in milling and the temperature of unhusked rice (Rhind and Tin, 1933). In this problem the respective values of the temperature (x) and of the breakage percentage (y) are arranged in parallel columns. For calculating the sums of squares and products zero is taken as the arbitrary origin, which is particularly suitable for calculations on the machine. The sum of products is calculated from the formula

$$\Sigma\,(x-\bar{x})\,(y-\bar{y}) = \Sigma\,xy - \frac{1}{N}\Sigma\,x.\Sigma\,y \qquad (6.3)$$

where $\Sigma\,xy$ is the sum of products of x and y values (not their deviations from the means $\bar{x}$ and $\bar{y}$), $\Sigma\,x$ is the total of x values and $\Sigma\,y$ the total of y values.

TABLE 6.2

Correlation between temperature of unhusked rice and percentage breakage of rice grains in milling

Serial number of observation	Temperature x	Percentage breakage y	x^2	y^2	xy
1	33·9	27·3	1149·21	745·29	925·47
2	34·6	29·5	1197·16	870·25	1020·70
3	34·5	26·8	1190·25	718·24	924·60
4	36·9	29·5	1361·61	870·25	1088·55
5	37·1	30·5	1376·41	930·25	1131·55
6	37·3	29·7	1391·29	882·09	1107·81
7	28·8	25·6	829·44	655·36	737·28
8	29·6	25·4	876·16	645·16	751·84
9	30·7	24·6	942·49	605·16	755·22
10	31·2	23·6	973·44	556·96	736·32
11	31·6	26·1	998·56	681·21	824·76
12	32·2	24·9	1036·84	620·01	801·78
13	33·4	27·0	1115·56	729·00	901·80
14	33·6	25·6	1128·96	655·36	860·16
15	33·6	26·4	1128·96	696·96	887·04
16	33·9	27·2	1149·21	739·84	922·08
Total ..	532·9	429·7	17845·55	11601·39	14376·96

Example 6.2

Calculate the correlation coefficient between the temperature of unhusked rice and the percentage breakage of grains in milling from the data shown in Table 6.2. We have

$$\Sigma(x-\bar{x})^2 = 17845{\cdot}55 - \frac{(532{\cdot}9)^2}{16} = 96{\cdot}65$$

$$\Sigma(y-\bar{y})^2 = 11601{\cdot}39 - \frac{(429{\cdot}7)^2}{16} = 61{\cdot}26$$

$$\Sigma(x-\bar{x})(y-\bar{y}) = 14376{\cdot}96 - \frac{(532{\cdot}9)\ (429{\cdot}7)}{16} = 65{\cdot}26$$

Hence

$$r = \frac{65{\cdot}26}{\sqrt{96{\cdot}65 \times 61{\cdot}26}} = 0{\cdot}8482$$

It is seen that there is a strong positive correlation between the temperature of unhusked rice and the breakage of rice grains in milling.

As a descriptive measure of association the correlation coefficient is of considerable use in the study of observational data. It should be noted, however, that the coefficient only expresses association and by itself tells us nothing of the causal relationships of the variates. Thus, purely from the knowledge that two variates x and y are correlated we cannot say whether variation in x is the cause or the result of variation in y or whether the association results from mutual dependence of the two variates or from common causes affecting both of them. Similarly the mere existence of a high value of the correlation coefficient is not necessarily indicative of an underlying relationship between the two variates. Such a value can at times be purely accidental, the two variates having no connection whatever; for instance, a correlation of $-0{\cdot}98$ was observed between the birth rates in Great Britain and annual production of pig iron in United States over the period 1875 to 1920. Such correlations are known as *spurious* or *nonsense* correlations.

6*a*.4 Significance of the Correlation Coefficient

The correlation coefficient like the mean value and the standard deviation is a numerical character of the population and is estimated from a sample of the bivariate population. This sample estimate denoted by r is susceptible to sampling fluctuations. In interpreting any sample value or the difference between two values of r we have therefore to take into account the sampling errors of the estimates. A case of particular importance is testing the significance of deviation of the estimate of correlation from the hypothetical value zero. This amounts to testing whether the sample indicates a real correlation between the variates or whether the observed value can easily arise as a result of sampling fluctuations. If ρ is the correlation in the population, then the estimate r based on n pairs has a standard error given by

$$\sigma_r = \frac{1 - \rho^2}{\sqrt{n - 1}} \tag{6.4}$$

The standard error of the estimate r is, however, of little value in tests of significance, for the distribution of r is far from normal except for small or moderate values of r and large n, say greater

than 500. For testing the significance of r, when it is estimated from a small number of pairs, we calculate the ratio

$$t = \frac{r}{\sqrt{1 - r^2}}\sqrt{n-2} \tag{6.5}$$

where r is the estimate obtained from n pairs. It has been shown that this ratio is distributed in sampling as t (Chapter IV) with $n - 2$ degrees of freedom. The significance of an observed correlation can, therefore, be tested by reference to the t table. If this ratio exceeds the t value for $P = 0{\cdot}05$, we have reason to believe that the sample indicates a real correlation between the two variates. Calculating this ratio for Example 6.2 of rice grain, we have

$$t = \frac{0{\cdot}8482\sqrt{16-2}}{\sqrt{1-(0{\cdot}8482)^2}} = 5{\cdot}987$$

The result is highly significant and establishes the existence of correlation between the temperature and breakage percentage beyond doubt.

The test can be more easily applied with the help of a table (Table VI, *Statistical Tables* by Fisher and Yates), which gives the values of r required for significance at levels usually required, for degrees of freedom $(n - 2)$ ranging from 1 to 100, by steps of 5 after 20 and steps of 10 after 50. For $n - 2 = 14$, the table shows that the value of r required for significance at $P = 0{\cdot}01$ is $0{\cdot}6226$. Our value of r exceeds this and is therefore significant at this level.

6*a*.5 *z* TRANSFORMATION

For comparison of two estimates of the correlation coefficient or for comparison of the estimates with hypothetical values, the standard errors of the estimates are of little value for reasons given above. The tests of significance for such comparisons involve what is known as the z transformation. The transformation consists in calculating from the value of r the quantity z, specified by the relation

$$z = \tfrac{1}{2}\log_e \frac{1+r}{1-r}$$

$$= \tfrac{1}{2}\{\log_e(1+r) - \log_e(1-r)\} \tag{6.6}$$

This quantity z is very nearly normally distributed for all values of the number of pairs, n, with a standard error

$$\sigma_z = \frac{1}{\sqrt{n-3}} \tag{6.7}$$

Tests of significance of the transformed values can therefore be done with the help of normal probability integral table. The transformation of r to z and z to r can be done with the help of tables of natural logarithms or more conveniently with the help of tables prepared for the purpose (Table VII, *Statistical Tables* by Fisher and Yates), wherein are given the values of r corresponding to values of z from 0 to 3. For transformations beyond the range of the table we must consult the tables of natural logarithms. The transformation to z allows us not only to test the significance of the observed values of r or their differences but also enables us to combine a number of estimates to give a joint estimate of correlation. The procedures involved can be best understood with the help of actual examples.

Example 6.3

(*a*) Test the significance of the values of r in Examples 6.1 and 6.2. (*b*) Test the significance of the deviation of the same values of r from 0·8.

(*a*) The following table sets out the values of r in the two cases, the corresponding z values and their standard errors.

TABLE 6.3

Transformation of correlation coefficient into z

Correlation	r	z	S.E.(z)
Length and number of wheat grains	0·826	1·1754	0·0502
Temperature and breakage per cent. of rice grains..	0·848	1·2490	0·2773

The z values are calculated thus

$$\log_e(1+0{\cdot}826) = \log 1{\cdot}826 = 0{\cdot}60212$$

$$\log_e(1-0{\cdot}826) = \log 0{\cdot}174 = -1{\cdot}74870$$

Hence

$$z = \tfrac{1}{2}(0{\cdot}60212 + 1{\cdot}74870) = 1{\cdot}1754$$

Similarly

$$\log_e(1{\cdot}848) = 0{\cdot}61412$$

$$\log_e(0{\cdot}152) = -1{\cdot}88387$$

and

$$z = \tfrac{1}{2}(0{\cdot}61412 + 1{\cdot}88387) = 1{\cdot}2490$$

The value of z corresponding to $r = 0$ is zero. The ratios of the deviations of the z values from zero to their standard errors are therefore $1{\cdot}1754/0{\cdot}0502$ and $1{\cdot}2490/0{\cdot}2773$ or $23{\cdot}4$ and $4{\cdot}5$ respectively which are highly significant.

(*b*) The value of z corresponding to $r = 0{\cdot}8$ is $1{\cdot}0986$. The significance of deviation of the correlation coefficients from $0{\cdot}8$ can be tested by comparing the deviation of the z values from this quantity with the standard errors of z values. Thus the ratio of the deviation to the standard error in the two cases would be

$$\frac{0{\cdot}0768}{0{\cdot}0502} = 1{\cdot}53$$

and

$$\frac{0{\cdot}1504}{0{\cdot}2773} = 0{\cdot}54$$

respectively.

By reference to the table of probability integral (Table 2.1) we find that the probabilities of exceeding the observed deviations in the two cases are respectively, $0{\cdot}126$ and $0{\cdot}589$. Neither of the deviations is thus significant at $P = 0{\cdot}05$.

Example 6.4

The correlation coefficients between temperature of unhusked rice and breakage percentage calculated from two samples of 12 and 16 are $0{\cdot}8912$ and $0{\cdot}8482$ respectively.

(*a*) Do the two estimates differ significantly?

(*b*) Combine the two estimates to give a joint estimate of the correlation between the two characters.

The following table gives the two estimates of r, the corresponding z values and the quantities $(n-3)$, $1/(n-3)$ and $(n-3)z$:

TABLE 6.4

Comparison and combination of correlation coefficients by z transformation

(1) Sample	(2) r	(3) z	(4) $n-3$	(5) $\frac{1}{n-3}$	(6) $(n-3)z$
First ..	0·8912	1·4276	9	0·1111	12·8484
Second ..	0·8482	1·2496	13	0·0769	16·2448
Total ..	..	..	22	0·1880	29·0932

(*a*) For testing the significance of difference between the two values of r, we take the difference of the corresponding values of z, divide it by its standard error and test the ratio as a normal deviate. The variance of the z value is given by $1/(n-3)$ and is shown in column 5 of Table 6.4. The variance of the difference of the two z values is the sum of the variances and is given in the last row.

The standard error of difference is therefore $\sqrt{0{\cdot}1880}$ or 0·4336.

The difference of z values is $1{\cdot}4276 - 1{\cdot}2496 = 0{\cdot}1780$. The ratio,

$$\frac{\text{difference}}{S.E.\text{ of difference}} = \frac{0{\cdot}1780}{0{\cdot}4336} = 0{\cdot}41$$

From Table 2.1, we find that the probability of exceeding this ratio in either direction is over 0·68 and therefore the two estimates cannot be considered to differ significantly. It is therefore reasonable to combine these into a single estimate.

(*b*) To combine two (or more) estimates of r we take the total of the quantities $(n-3)\,z$ and divide it by the sum $\Sigma(n-3)$. This is said to be the *weighted mean* of the z values where the quantities $(n-3)$ representing the reciprocals of the respective variances are the weights. This weighted mean is the value of z corresponding to the joint estimate of the correlation coefficient and has a variance given by $1/\Sigma(n-3)$. The corresponding

value of r is found by reference to the table of r and z (Table VII, *Statistical Tables* by Fisher and Yates) or with the help of a table of natural logarithms.

In the present example, the weighted mean of the z values is

$$\frac{29 \cdot 0932}{22} \text{ or } 1 \cdot 3224$$

From Table VII (Fisher and Yates), we find that the value of r corresponding to $z = 1 \cdot 32$ is $0 \cdot 8668$, and that of r corresponding to $z = 1 \cdot 33$ is $0 \cdot 8692$. We notice that the difference of $0 \cdot 01$ in z corresponds to a difference of $0 \cdot 0024$ in r. Hence the value of r corresponding to $z = 1 \cdot 3224$ is

$$0 \cdot 8668 + \frac{(0 \cdot 0024)\ (0 \cdot 0024)}{0 \cdot 01} = 0 \cdot 8674$$

Thus $0 \cdot 8674$ is the joint estimate of r.

6*b*.1 Regression

A simpler and more useful approach to the study of simultaneous variation of two (or more) characters is the study of *regression*. If we are studying two characters x and y and have drawn up a correlation table like Table 6.1, we can find out for each of the columns (or rows) the mean of the character y (or x). It will be generally seen that these means bear some relationship with the values of the variate characterising the columns (or the rows). We may, for example, plot the means of the columns of Table 6.1 against the values of x corresponding to these columns. This is done in Fig. 6.2.

It can be observed from the figure that the dots lie more or less on a straight line which rises with increasing x. We cannot, however, hope to get a regular straight line by joining the points, for even if there is a linear relation between the mean y value and x in the population, the means obtained from an observed body of data, which must be necessarily finite, would deviate from the ideal line owing to sampling fluctuations. The underlying relation between y and x in a bivariate population can be expressed in the form of a mathematical equation known as the regression equation and said to represent the regression of the variate y on

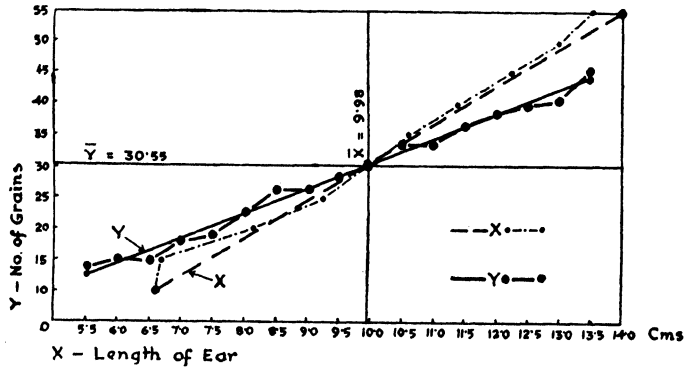


FIG. 6.2. Regressions of number of grains (y) on length of ear (x) and *vice-versa* in Pusa 12 wheat.

the variate x. y is here called the dependent and x the independent variate. For the simplest case of a straight line graph, the equation is of the form

$$Y_p = \alpha + \beta x \tag{6.8}$$

where α and β denote the population constants and Y_p the hypothetical population mean corresponding to any x.

The regression represented by a straight line is spoken of as the *linear regression*. In practice the values of α and β would be obtained from a sample. If a and b denote these estimates the regression line can be written as

$$Y = a + bx \tag{6.9}$$

where Y is the value one would expect to obtain corresponding to any given x on the basis of the observed linear relation in (6.9). a and b are so chosen that the quantity $\Sigma(y - Y)^2$ is minimum. This is known as the principle of *least squares*. In this quantity y stands for the observed value of the dependent variate. Y for the value obtained by substituting in equation (6.9) the corresponding x value and Σ indicates summation over all pairs of observations. This quantity, therefore, represents the sum of squares of the deviations of observed values of y from those predicted from the regression equation and forms the

residual variation after a regression line is fitted to the data. The regression equation is thus the relation between y and x which minimizes the residual variation of y. It can be shown that the residual variation is minimum when a and b are calculated from the formulæ

$$a = \bar{y} - b\bar{x} \tag{6.10}$$

and

$$b = \frac{\Sigma (y - \bar{y})(x - \bar{x})}{\Sigma (x - \bar{x})^2} \tag{6.11}$$

The quantity $\Sigma (y - \bar{y})(x - \bar{x})$, which is obviously the sum of products of deviations of x and y from the respective means, can also be expressed as

$$\Sigma (y - \bar{y})(x - \bar{x}) = \Sigma y (x - \bar{x}) = \Sigma x (y - \bar{y})$$

Thus b can be alternatively expressed as

$$b = \frac{\Sigma y (x - \bar{x})}{\Sigma (x - \bar{x})^2}$$

or

$$= \frac{\Sigma x (y - \bar{y})}{\Sigma (x - \bar{x})^2} \tag{6.12}$$

For the data in the Table 6.1 the values of a and b are given by

$$a = -9{\cdot}476$$

and

$$b = 4{\cdot}011$$

hence

$$Y = -9{\cdot}476 + 4{\cdot}011x$$

The values of Y calculated from the regression line are entered in Fig. 6.2.

6*b*.2 Relation between Regression and Correlation

If in the same manner as explained in the previous section we obtain the mean values of x for the rows and plot them on the graph, as is also shown in Fig. 6.2, it will be seen that these lie in general along a different line. If we try to express x in terms of y by means of an equation like

$$X = a' + b'y \tag{6.13}$$

and choose a' and b' so as to minimize the quantity $\Sigma(x - X)^2$, it would be found that this equation will be different from the one obtained previously. The constants a' and b' will be given by the formulæ

$$a' = \bar{x} - b'\bar{y} \tag{6.14}$$

and

$$b' = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(y - \bar{y})^2} \tag{6.15}$$

For the data in Table 6.1 these are found to be 4·788 and 0·170 respectively and are seen to be very different from the values a and b obtained previously. The resulting regression line is entered in Fig. 6.2. It should be noted that the regression of x on y is not the same as the regression of y on x and no attempt should be made to obtain the value of x corresponding to any given value of y from the regression of y on x and *vice versa*. The two regression lines are seen to intersect at the point $(\bar{x}, \bar{y})$.

If we multiply b and b', we obtain

$$bb' = \frac{\Sigma(y - \bar{y})(x - \bar{x})}{\Sigma(x - \bar{x})^2} \cdot \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(y - \bar{y})^2}$$

$$= \frac{\{\Sigma(y - \bar{y})(x - \bar{x})\}^2}{\Sigma(x - \bar{x})^2 \cdot \Sigma(y - \bar{y})^2}$$

which is seen to be the square of the correlation coefficient Hence

$$bb' = r^2$$

or

$$r = \sqrt{bb'} \tag{6.16}$$

The correlation coefficient is thus the geometric mean of the two regression coefficients and is alternatively defined as such.

6*b*.3 Significance of Regression

By substituting the value of Y obtained from the equation

$$Y = \bar{y} + b(x - \bar{x})$$

in $\Sigma(y - Y)^2$, we get the value of the residual variation. Thus

$$\begin{aligned}\Sigma(y - Y)^2 &= \Sigma[y - \{\bar{y} + b(x - \bar{x})\}]^2 \\ &= \Sigma\{(y - \bar{y}) - b(x - \bar{x})\}^2 \\ &= \Sigma(y - \bar{y})^2 + b^2\Sigma(x - \bar{x})^2 - 2b\Sigma(y - \bar{y})(x - \bar{x}) \\ &= \Sigma(y - \bar{y})^2 - b\Sigma(y - \bar{y})(x - \bar{x}) \qquad (6.17)\end{aligned}$$

since

$$\begin{aligned}b^2\Sigma(x - \bar{x})^2 &= b\{b\Sigma(x - \bar{x})^2\} \\ &= b\frac{\Sigma(y - \bar{y})(x - \bar{x})}{\Sigma(x - \bar{x})^2}\Sigma(x - x)^2 \\ &= b\Sigma(y - \bar{y})(x - \bar{x})\end{aligned}$$

The residual variation is, therefore, obtained by subtracting from the sum of squares of y the quantity $b\Sigma(y - \bar{y})(x - \bar{x})$. The latter quantity thus represents the variation accounted for by the regression. We can set down this result in the form of an analysis of variance of y values as follows:

Source	*D.F.*	*S.S.*
Regression ..	1	$b\Sigma(y-\bar{y})(x-\bar{x})$
Residual ..	$n-2$	$\Sigma(y-\bar{y})^2 - bS(y-\bar{y})(x-\bar{x})$
Total ..	$n-1$	$\Sigma(y-\bar{y})^2$

The total sum of squares, $\Sigma(y - \bar{y})^2$, is calculated by obtaining one constant $\bar{y}$ from the data and has $n - 1$ degrees of freedom. The sum of squares due to regression depends upon one other constant calculated from the data, namely b, apart from $\bar{x}$. In repeated samples, however, $x_1, x_2, \ldots x_n$ are assumed to be fixed. The sum of squares has therefore 1 degree of freedom. The residual variation after the removal of the sum of squares due to two constants, $\bar{y}$ and b, has naturally $n - 2$ degrees of freedom. By dividing the sum of squares due to residual variation by $n - 2$ we get the residual mean square also known as the *residual variance* which we may denote by s'^2. The residual variance s'^2 represents the variation which is not accounted for by the hypothesis of linear regression and therefore measures

the uncontrolled variation affecting the y values. By comparing the variation due to regression with s'^2 we can test the significance of the coefficient b. For this purpose, we calculate the F ratio. If this ratio is non-significant, we cannot assert with reasonable confidence that there is any relation between the values of y and x and we shall have to regard the apparent regression to be of such magnitude as can easily arise as a result of sampling fluctuations. The F ratio has 1 degree of freedom for the numerator and $n - 2$ for the denominator. Alternatively, we may test the significance of b against its standard error by calculating the ratio

$$t = \frac{b}{s'/\sqrt{\Sigma (x - \bar{x})^2}} \tag{6.18}$$

where the denominator is the standard error of b and refer the calculated value of t in the table with $n - 2$ degrees of freedom. The tests are identical as it might be recalled from Chapter IV that with 1 degree of freedom for the numerator, $F = t^2$.

Example 6.5

Calculate the regression of breakage percentage of rice (y) on temperature (x), from the data in Example 6.2 and test its significance. We have

$$\Sigma (x - \bar{x})^2 = 96{\cdot}65, \Sigma y (x - \bar{x}) = 65{\cdot}26$$

and

$$\Sigma (y - \bar{y})^2 = 61{\cdot}26$$

Hence

$$b = \frac{65{\cdot}26}{96{\cdot}65} = 0{\cdot}6752$$

$$\text{Regression } S.S. = 65{\cdot}26 \times 0{\cdot}6752 = 44{\cdot}06$$

$$\text{Residual } S.S. = 17{\cdot}20$$

$$\bar{y} = 26{\cdot}86$$

and

$$\bar{x} = 33{\cdot}31$$

Hence the regression equation is

$$Y = 26{\cdot}86 + 0{\cdot}6752 (x - 33{\cdot}31)$$

$$4{\cdot}37 + 0{\cdot}6752 x$$

We have the following analysis of variance of y:

Source	*D.F.*	*S.S.*	*M.S.*
Regression	1	44·06	44·06
Residual variation ..	14	17·20	1·23
Total ..	15	61·26	4·08

$$\text{The } F \text{ ratio} = \frac{44 \cdot 06}{1 \cdot 23} = 35 \cdot 82$$

The observed value of F is seen to be much larger than $F_{1\%}$ for 1 and 14 degrees of freedom and therefore the regression must be considered to remove a significant portion of the total variation. In other words the regression coefficient is significantly different from zero.

6*c*.1 Partial Correlation

The correlation coefficient defined and discussed earlier measures the association between two characters. We may wish to study the simultaneous variation of more than two characters also. A common instance is provided by the height, weight and age of children of school-going age. Another instance is provided by the characters, height, node number, number of bolls, average boll weight, yield, etc., for a cotton plant. In such cases we can, of course, obtain the correlation coefficients for various pairs of characters. The measures of association thus obtained will not however be independent. Since, for instance, weight and height are each correlated with age, weight and height would also be found to be correlated between themselves. In other words, if two characters are each associated with a third, they are bound to be correlated. We might in such cases wish to know whether a pair of characters show any correlation even when their correlation with the third is allowed for and to obtain a measure of correlation after making such allowance. This can be done in the following way. Consider the three variates x_1, x_2, x_3. For any individual set of three observations, their deviations from the respective means are given by $(x_1 - \bar{x}_1)$, $(x_2 - \bar{x}_2)$ and $(x_3 - \bar{x}_3)$. If b_{13} is the coefficient of regression of x_1 and x_3, a deviation of $b_{13}(x_3 - \bar{x}_3)$ is predicted for x_1 on account

of its association with x_3. If we remove this from the actual deviation, the residual

$$(x_1 - \bar{x}_1) - b_{13}(x_3 - \bar{x}_3)$$

is free from the effect of association of x_1 and x_3.

Similarly the residual

$$(x_2 - \bar{x}_2) - b_{23}(x_3 - \bar{x}_3)$$

is free from the effect of association of x_2 with x_3. If we consider the association of such pairs of residuals we get a measure of association of two variates x_1 and x_2 which is independent of their association with the third, namely x_3. If we denote by $r_{12\cdot3}$ the correlation coefficient of the residuals, it can be shown that

$$r_{12\cdot3} = \frac{r_{12} - r_{13}\cdot r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \tag{6.19}$$

where r_{12}, r_{13} and r_{23} are the correlation coefficients between the pairs of variates, x_1 and x_2; x_1 and x_3; and x_2 and x_3, respectively. The correlation coefficient $r_{12\cdot3}$ is known as a *partial correlation* coefficient, since it measures the association of two variates after making allowance for their association with the third, as against the correlation coefficients r_{12}, r_{13} and r_{23} which are known as *total correlation* coefficients.

Thus suppose the total correlation coefficients between yield and boll number, yield and height of plant and boll number and height of plant obtained from observations on 30 cotton plants are 0·86, 0·65 and 0·72 respectively. From these we can calculate the partial correlation between yield and boll number. Denoting the characters yield, boll number and height by 1, 2 and 3, we have

$$r_{12} = 0{\cdot}86$$

$$r_{13} = 0{\cdot}65$$

and

$$r_{23} = 0{\cdot}72$$

Hence

$$r_{12\cdot3} = \frac{0{\cdot}86 - (0{\cdot}65)(0{\cdot}72)}{\sqrt{\{1 - (0{\cdot}65)^2\}\{1 - (0{\cdot}72)^2\}}}$$

$$= \frac{0{\cdot}3920}{\sqrt{0{\cdot}4816 \times 0{\cdot}5775}}$$

$$= 0{\cdot}743$$

The variate whose influence is allowed for in the calculation of the partial correlation coefficient is spoken of as the *eliminated* variate. If we have partial correlation coefficients between three pairs formed by three variates in each of which a fourth variate is eliminated, we can further obtain from these three coefficients a partial correlation coefficient in which two variates are eliminated, by the application of the same formula. We can thus obtain by successive stages partial correlation coefficients in which three or more variates are eliminated.

6*c*.2 Significance of Partial Correlation

The testing of significance of partial correlation coefficients is done in the same way as the testing of total correlation coefficients with this difference that for each variate eliminated 1 degree of freedom is subtracted from the degrees of freedom for t or the transformed z value. Thus for the partial correlation coefficients obtained by eliminating one variate the t would be tested with $(n-2)-1=n-3$ degrees of freedom where the coefficient is calculated from n trios. Alternatively, the transformed z may be tested against its standard error, namely, $1/\sqrt{n-4}$. Thus in the example of 30 cotton plants in the previous section the z value will have a standard error of

$$\frac{1}{\sqrt{30-4}}=\frac{1}{\sqrt{26}}=0\cdot196$$

Similarly for coefficients with more eliminated variates we shall subtract 1 additional degree of freedom for each variate eliminated.

6*d*.1 Partial Regression

With only two variates, x and y, we expressed y as a linear function of x and chose a and b in the equation $Y=a+bx$ so as to minimize the residual variation.

With more than two variates we can express a dependent variate y in terms of a number of independent variates such as $x_1, x_2, x_3, \ldots$, by means of a linear equation of the type

$$Y=a+b_1x_1+b_2x_2+\ldots \qquad (6.20)$$

The coefficients a, b_1, b_2, ..., are estimated by minimizing the quantity $\Sigma(y - Y)^2$ representing the residual variation. This leads to the following equations:—

$$\begin{aligned}\Sigma y &= \Sigma Y \\ \Sigma x_1 y &= \Sigma x_1 Y \\ \Sigma x_2 y &= \Sigma x_2 Y \\ &\cdots\cdots\cdots\end{aligned}$$

From the first of these equations we write

$$\bar{y} = \bar{Y} = a + b_1\bar{x}_1 + b_2\bar{x}_2 + \ldots$$

Subtracting this from each of the other equations above we obtain

$$\begin{aligned}b_1\Sigma x_1'^2 + b_2\Sigma x_1'x_2' + b_3\Sigma x_1'x_3' + \ldots. &= \Sigma x_1'y' \\ b_1\Sigma x_1'x_2' + b_2\Sigma x_2'^2 + b_3\Sigma x_2'x_3' + \ldots. &= \Sigma x_2'y' \\ b_1\Sigma x_1'x_3' + b_2\Sigma x_2'x_3' + b_3\Sigma x_3'^2 + \ldots. &= \Sigma x_3'y' \\ &\cdots\cdots\cdots\end{aligned}$$

where $x_i' = x_i - \bar{x}_i$ and $y' = y - \bar{y}$ and Σ stands for the summation, there being as many equations as there are b's in the equation. The equation (6.20) is known as the *partial regression* equation and the coefficients b_1, b_2, etc., are known as the partial regression coefficients.

The method of solving these equations is given by Fisher (*Statistical Methods for Research Workers*, 1950) and consists in calculating from the sets of equations

$$\begin{aligned}b_1\Sigma x_1'^2 + b_2\Sigma x_1'x_2' + b_3\Sigma x_1'x_3' + \ldots &= 1\ 0\ 0\ 0 \ldots 0 \\ b_1\Sigma x_1'x_2' + b_2\Sigma x_2'^2 + b_3\Sigma x_2'x_3' + \ldots &= 0\ 1\ 0\ 0 \ldots 0 \\ b_1\Sigma x_1'x_3' + b_2\Sigma x_2'x_3'^2 + b_3\Sigma x_3'^2 + \ldots &= 0\ 0\ 1\ 0 \ldots 0 \\ &\cdots\cdots\cdots \\ &\cdots\cdots\cdots\end{aligned} \qquad (6.21)$$

obtained by giving the right-hand side in succession the values 1 and 0, the roots of each set of equations. With p independent

variates there will be p such sets. The p solutions of the successive equations can be denoted by

$$\begin{array}{l} c_{11},\ c_{12},\ c_{13}\ldots.c_{1p} \\ c_{21},\ c_{22},\ c_{23}\ldots.c_{2p} \\ c_{31},\ c_{32},\ c_{33}\ldots.c_{3p} \\ \ldots\ldots\ldots\ldots\ldots \\ c_{p1},\ c_{p2},\ c_{p3}\ldots.c_{pp} \end{array} \qquad (6.22)$$

It would be found that $c_{mn} = c_{nm}$. From these c coefficients, $b_1, b_2, \ldots$, can be obtained by means of the relations

$$\begin{aligned} b_1 &= c_{11}\Sigma x_1'y' + c_{12}\Sigma x_2'y' + c_{13}\Sigma x_3'y' + \ldots. c_{1p}\Sigma x_p'y' \\ b_2 &= c_{21}\Sigma x_1'y' + c_{22}\Sigma x_2'y' + c_{23}\Sigma x_3'y' + \ldots. c_{2p}\Sigma x_p'y' \\ b_3 &= c_{31}\Sigma x_1'y' + c_{32}\Sigma x_2'y' + c_{33}\Sigma x_3'y' + \ldots. c_{3p}\Sigma x_p'y' \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ b_p &= c_{p1}\Sigma x_1'y' + c_{p2}\Sigma x_2'y' + c_{p3}\Sigma x_3'y' + \ldots. c_{pp}\Sigma x_p'y' \end{aligned} \qquad (6.23)$$

The c values arranged in the manner shown in (6.22) are known as the *covariance matrix* and would be the same for the same set of independent variates. If we want to calculate the regression of some other variate z on this very set of independent variates all we have to do is to calculate the sums of products $\Sigma x_1'z', \Sigma x_2'z', \ldots$ and substitute in the equations for b and we would get the partial regression coefficients for the regression of the variate z on the variates $x_1, x_2, \ldots$

6*d*.2 Significance of Partial Regression

The c values also enable us to calculate the standard errors of the coefficients, $b_1, b_2, \ldots$ This is done as follows. We first obtain the residual sum of squares $\Sigma(y - Y)^2$ which is given by

$$\Sigma y'^2 - b_1\Sigma x_1'y' - b_2\Sigma x_2'y' - b_3\Sigma x_3'y' - \ldots. \qquad (6.24)$$

and divide it by $n - p - 1$ to obtain the residual mean square. Denoting this mean square by s'^2, we have

$$s'^2 = \frac{\Sigma(y - Y)^2}{n - p - 1} \qquad (6.25)$$

The standard errors of $b_1, b_2, \ldots$ are then given by $S.E.\ (b_1) = s'\sqrt{c_{11}}$, $S.E.\ (b_2) = s'\sqrt{c_{22}}, \ldots$ The whole procedure would be clear from a simple example of regression of y on two variates x_1 and x_2. Regression of a dependent variate on two independent variates is the simplest case of a partial regression equation and partial regressions with more variates can be obtained without any difficulty once the procedure in the simple case is understood.

Example 6.6

Table 6.5 gives for 25 progenies of cotton the data for mean fibre length of each progeny, the corresponding parent plant value and the mean value of the plot in which the parent plant was grown. It is found that both the parent value as well

TABLE 6.5

Fibre length of cotton progenies and parent plants

Number of progeny	Progeny mean (mm.) y	Parental plant value (mm.) x_1	Parental plot mean (mm.) x_2
1	24·30	26·0	25·50
2	24·48	28·8	25·50
3	23·41	25·2	25·50
4	21·60	23·4	25·00
5	22·49	26·6	25·00
6	23·62	25·4	24·60
7	22·75	23·4	24·60
8	24·40	27·6	23·60
9	22·60	24·4	23·60
10	25·36	24·0	24·42
11	23·21	24·2	24·42
12	24·76	26·0	24·42
13	21·53	22·8	22·56
14	21·32	20·8	22·56
15	22·81	24·8	22·56
16	25·41	26·2	24·90
17	24·30	27·2	24·90
18	23·65	26·6	24·91
19	24·31	25·0	24·91
20	21·88	23·4	24·05
21	24·10	25·6	24·05
22	21·91	23·0	24·05
23	22·24	25·4	24·57
24	23·45	23·4	24·57
25	22·10	24·2	24·57
Total	581·99	623·4	609·32

Note.—The values of y and x_2 being means are carried to one more decimal place than those for x_1.

as the plot mean bear some relationship with the progeny mean. Express this relation in the form of a partial regression equation with progeny mean as the dependent variate.

From the table the following sums of squares and products can be easily calculated:

$$\Sigma x_1'^2 = 74{\cdot}62 \qquad \Sigma x_1'y' = 34{\cdot}09$$
$$\Sigma x_2'^2 = 17{\cdot}33 \qquad \Sigma x_2'y' = 11{\cdot}09$$
$$\Sigma y'^2 = 35{\cdot}50 \qquad \Sigma x_1'x_2' = 18{\cdot}97$$

We then get the following two sets of equations

$$74{\cdot}62\, c_{11} + 18{\cdot}97\, c_{12} = 1$$
$$18{\cdot}97\, c_{11} + 17{\cdot}33\, c_{12} = 0$$

and

$$74{\cdot}62\, c_{21} + 18{\cdot}97\, c_{22} = 0$$
$$18{\cdot}97\, c_{21} + 17{\cdot}33\, c_{22} = 1$$

Taking the first set, namely,

$$74{\cdot}62\, c_{11} + 18{\cdot}97\, c_{12} = 1$$
$$18{\cdot}97\, c_{11} + 17{\cdot}33\, c_{12} = 0$$

and eliminating c_{12}, by multiplying the first equation by 17·33 and the second by 18·97 and subtracting the latter, we get

$$933{\cdot}30\, c_{11} = 17{\cdot}33$$

or

$$c_{11} = 0{\cdot}01857$$

Substituting this in the second equation, we have

$$17{\cdot}33\, c_{12} = -0{\cdot}35227$$

or

$$c_{12} = -0{\cdot}02033$$

Taking the second set we similarly obtain

$$c_{21} = -0{\cdot}02033$$
$$c_{22} = 0{\cdot}07997$$

Hence we have the covariance matrix

$$0{\cdot}01857, \quad -0{\cdot}02033$$
$$-0{\cdot}02033, \quad 0{\cdot}07997$$

and

$$b_1 = (0{\cdot}01857)(34{\cdot}09) + (-0{\cdot}02033)(11{\cdot}09) = 0{\cdot}4076$$

and

$$b_2 = (-0{\cdot}02033)(34{\cdot}09) + (0{\cdot}07997)(11{\cdot}09) = 0{\cdot}1938$$

$$\text{Residual } S.S. = 35{\cdot}50 - (0{\cdot}4076)(34{\cdot}09) - (0{\cdot}1938)(11{\cdot}09)$$
$$= 19{\cdot}46$$

$$\therefore \quad s'^2 = \frac{1}{22}(19{\cdot}46) = 0{\cdot}8845$$

and

$$s' = 0{\cdot}9405$$

Now

$$\sqrt{c_{11}} = 0{\cdot}1363$$

and

$$\sqrt{c_{22}} = 0{\cdot}2828$$

$$\therefore \quad S.E.\ (b_1) = 0{\cdot}1282$$

and

$$S.E.\ (b_2) = 0{\cdot}2660$$

The significance of the coefficients b_1 and b_2 can be tested by reference to the table of t with $n - p - 1$ degrees of freedom, which is 22 in this case.

For b_1

$$t = \frac{0{\cdot}4076}{0{\cdot}1282} = 3{\cdot}18$$

and for b_2

$$t = \frac{0{\cdot}1938}{0{\cdot}2660} = 0{\cdot}73$$

It is seen that b_1 is highly significant while b_2 is not. We can therefore conclude that while the parental value has definitely an influence on the progeny mean as evidenced by the data, the influence of plot mean is not significant. As in the case of the linear regression the total variation of y can be regarded as being divisible into components due to regression and residual variation

and the results set down in the table of analysis of variance as given below. The sum of squares due to regression is based on 2 degrees of freedom in this case, 2 being the number of independent variates.

Source		*D.F.*	*S.S.*	*M.S.*
Regression	..	2	16·04	8·02
Residual variation	..	22	19·46	0·8845
Total	..	24	35·50	

The significance of the joint regression will be tested by calculating the F ratio. Thus in the above case we have

$$F = \frac{8 \cdot 02}{0 \cdot 8845} = 9 \cdot 06$$

a highly significant value.

The square-root of the ratio of the regression sum of squares to the total sum of squares is known as the *multiple correlation* coefficient and denoted by R. It is always positive and obviously less than 1. It might be noted that if this ratio is calculated in the case of simple linear regression we get the square of the correlation coefficient r between x and y. Hence the square of the correlation coefficient, r^2, represents the fraction of the total variation of y that is accounted for by its association with x. Similarly, in this case, the square of the multiple correlation coefficient represents the fraction of the variation in y accounted for by its joint association with the variates x_1, x_2 ... The coefficient is therefore a measure of the joint association of all these variates with the dependent variate y and tells us how much of the variation in y could be accounted for by reference to these variates.

6*e*. 1 Curvilinear Regression

So far we have considered the regression of y on x represented by a linear relation of the type

$$Y = a + bx$$

There are however many variates the re'ation between which is not so simple. Consider, for example, the relation between the manure

applied and the yield of a crop, say, wheat. If by applying 100 lb. of ammonium sulphate to a field we get an increase in yield of 200 lb. per acre, it is not necessarily likely that 500 lb. of it would give an increase of 1000 lb. per acre or that 10 lb. of it would give an additional yield of 20 lb. per acre. The relationship between increase of yield and manure applied is not proportional or linear. It is found by experiment that additional doses of manure increase the yield only upto a certain limit, the rate of this increase for every additional unit of manure decreasing gradually. Further application of manure is not effective and may even decrease the yield. The student can think of more such instances, as for example, the relation between rainfall and yield of crops, between quantity of feed and milk yield of a cow and so on. These relations are found to be expressible in the form of more complex equations of the type:

$$Y = a + bx + cx^2$$

or

$$Y = a + bx + cx^2 + dx^3 + \ldots \qquad (6.26)$$

containing terms in higher powers of the independent variate x. By plotting the graphs of such equations it will be found that they give curves of various types. Consequently, regressions represented by these equations are known as *curvilinear regressions* and the expression on the right-hand side involving higher powers of x is known as the *polynomial* in x.

The coefficients $a, b, c, d, \ldots$ in the polynomial regression equation are estimated by minimizing as before $\Sigma(y - Y)^2$ leading to the equations

$$\begin{aligned} \Sigma y &= \Sigma Y \\ \Sigma xy &= \Sigma xY \\ \Sigma x^2y &= \Sigma x^2Y \\ &\cdots\cdots\cdots\cdots \end{aligned}$$

It is thus seen that a polynomial regression equation can be fitted by following a procedure exactly similar to that for fitting a multiple regression equation. For a second degree equation, for example, the coefficients b and c would be given by

$$\begin{aligned} &b\Sigma(x - \bar{x})^2 + c\Sigma(x - \bar{x})(x^2 - \overline{x^2}) = \Sigma(x - \bar{x})(y - \bar{y}) \\ &b\Sigma(x - \bar{x})(x^3 - \overline{x^2}) + c\Sigma(x^2 - \overline{x^2})^2 = \Sigma(x^2 - \overline{x^2})(y - \bar{y}) \end{aligned} \qquad (6.27)$$

where $\overline{x^2}$ stands for

$$\frac{1}{n}\Sigma x^2$$

and a is given by

$$a = \bar{y} - b\bar{x} - c\overline{x^2} \qquad (6{\cdot}28)$$

The covariance matrix of these coefficients may also be worked out and used for testing the significance of b, c, etc., in exactly the same manner as described for the multiple regression.

The degree of the polynomial curve that would be appropriate for any given body of data can often be judged from their correlation table. But even in the absence of such knowledge we can fit in succession 1st, 2nd, 3rd, ... degree curves and stop the process when the improvement obtained in the closeness of the fit of the regression curve to the data ceases to be significant, that is, when the coefficient of the highest power of x in the regression equation obtained comes out to be non-significant.

We shall fit the 2nd degree curve to the data of temperature and breakage percentage of rice in Example 6.2 by the above method. We have already found that

$$\Sigma(x - \bar{x})^2 = 96{\cdot}65,\ \Sigma(x - \bar{x})(y - \bar{y}) = 65{\cdot}26$$

and

$$\Sigma(y - \bar{y})^2 = 61{\cdot}26$$

We can similarly find that

$$\Sigma(x - \bar{x})(x^2 - \overline{x^2}) = 6552{\cdot}05,\ \Sigma(y - \bar{y})(x^2 - \overline{x^2}) = 4404{\cdot}11$$

and

$$\Sigma(x^2 - \overline{x^2})^2 = 428688{\cdot}58$$

For the covariance matrix we have the sets of equations

$$96{\cdot}65\, b_1 + 6552{\cdot}05\, b_2 = 1,\ 0$$

$$6552{\cdot}05\, b_1 + 428688{\cdot}58\, b_2 = 0,\ 1$$

which give

$$c_{11} = 8{\cdot}7709$$

$$c_{12} = -0{\cdot}1316$$

and

$$c_{22} = 0{\cdot}001977$$

Hence we get

$$b_1 = -7{\cdot}2931,\ b_2 = 0{\cdot}1194,\ a = 136{\cdot}62$$

and

$$s' = \sqrt{\frac{61{\cdot}26 - 49{\cdot}90}{13}} = \sqrt{0{\cdot}8738} = 0{\cdot}9348$$

$$S.E.\ (b_1) = s'\sqrt{c_{11}} = 2{\cdot}7684$$

$$S.E.\ (b_2) = s'\sqrt{c_{22}} = 0{\cdot}0416$$

The equation obtained is

$$Y = 136{\cdot}62 - 7{\cdot}2931x + 0{\cdot}1194x^2$$

The straight line and the second degree curve obtained from the data are shown together in Fig. 6.3.

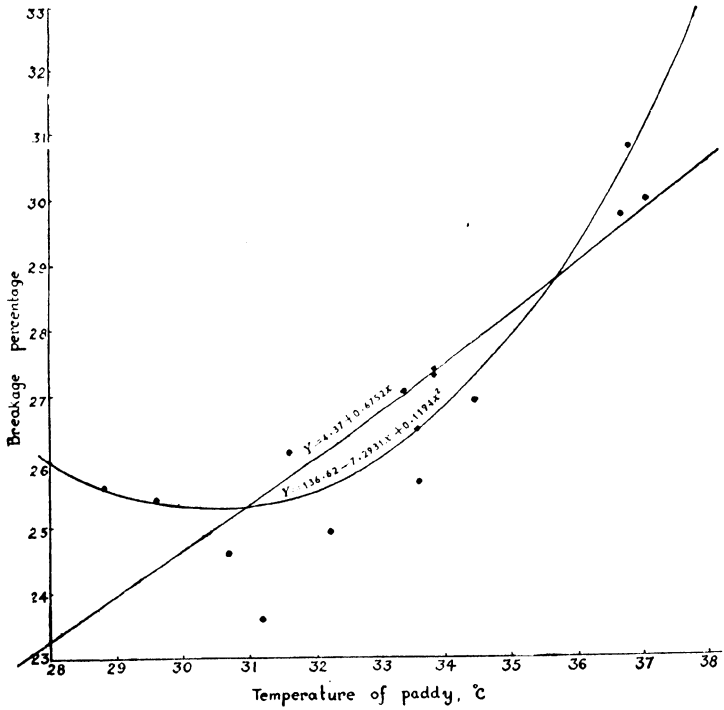


FIG. 6.3. Linear and quadratic regressions of breakage percentage of rice on temperature.

6*f*.1 Intraclass Correlation

The correlation coefficient was introduced as a measure of association between two variates. In the particular cases considered in that connection, the variates referred to measurements in different units and to characters that were quite distinct such as length of earhead and number of grains per earhead or temperature and breakage percentage of rice. It is, perhaps, easy to see that the study of association need not and cannot always be confined to such variates. We may, for instance, wish to consider whether the length of earheads on the same wheat plant or the fibre length of cotton from different bolls of the same plant tend to be alike. For studies of this type we should need data giving lengths of two or more earheads or the fibre lengths from two or more bolls from each of a number of plants. Such data however present a difficulty in classifying them in the form of a correlation table. Even for the simplest case of a pair of observations from each of a number of plants, we cannot directly calculate the correlation coefficient by the method described at the beginning of the chapter, for we cannot classify the individual observations in various pairs under two different heads as two distinct variates. One method of measuring association among such observations is to enter the observations in each pair twice, interchanging the classes arbitrarily marked as *A* and *B* or 1 and 2, when the observations are entered a second time. It is obvious that the distribution of values in both classes would be identical since all the observations would appear in each class. By treating this table by the method described earlier we shall obtain a correlation coefficient

$$\rho' = \frac{2\Sigma(x - \bar{x})(x' - \bar{x})}{\Sigma(x - \bar{x})^2 + \Sigma(x' - \bar{x})^2} \tag{6.29}$$

where x and x' are the corresponding measurements in pairs and $\bar{x}$ is the mean of all the observations, that is, $\Sigma(x + x')/2n$, there being n pairs of observations. Since the observations are not naturally divisible into two classes we call the coefficient ρ' an *intraclass* (within class) *correlation* coefficient to distinguish it from ρ the *interclass* (between class) *correlation* coefficient.

We have often to consider in similar circumstances association of individuals in groups of more than two. We may, for instance, have three earheads per plant or even more. The method is adapted to cover such cases by entering in a table, as in case of pairs, as many pairs as can be formed in each group and obtaining therefrom the intraclass correlation coefficient using an extension of formula (6.29). It is obvious, however, that this method would be very laborious and become impossible in practice if each group contained a large number of individuals. On the other hand the problem of measuring association between individuals within groups arises frequently. Thus we may be interested in calculating the association between the crop yields of a number of small plots within blocks into which land might be divided. Or we may wish to consider the association between rainfall recorded at a number of stations in a district from season to season or, as in the classical example of Harris (1913), between leaves of the same tree. Harris himself studied the resemblance of leaves from a tree by taking observations on 26 leaves from each of a number of trees and introduced a simple method of calculating the intraclass correlation. We shall describe this method in the following paragraph.

Suppose there are n groups each with k observations. Then, as in Example 4.3, Chapter IV, we can partition the total variability of the kn values into that between groups and within groups. The analysis of variance would take the form.

Source		D.F.	S.S.	M.S.
Between groups	..	$n-1$	$k\sum(\bar{x}_p-\bar{x})^2$	ks_b^2
Within groups	..	$n(k-1)$	$\sum(x-\bar{x})^2-k\sum(\bar{x}_p-\bar{x})^2$	s_w^2
	Total ..	$nk-1$	$\sum(x-\bar{x})^2$	s^2

where $\bar{x}_p$ stands for a group mean and ks_b^2 and s_w^2 stand for the mean squares between and within groups. Then the estimate of the intraclass correlation coefficient, r', can be shown to be given by

$$\frac{M.S.\text{ between groups} - M.S.\text{ within groups}}{M.S.\text{ between groups} + (k-1)\,M.S.\text{ within groups}}$$

i.e.,

$$r' = \frac{ks_b^2 - s_w^2}{ks_b^2 + (k-1)s_w^2} \tag{6.30}$$

It follows that

$$\frac{ks_b^2}{s_w^2} = \frac{(k-1)r' + 1}{1 - r'} \tag{6.31}$$

Since ks_b^2 and s_w^2 are both mean squares, they cannot be negative. Hence the ratio

$$\frac{(k-1)r' + 1}{1 - r'}$$

must be $\geqslant 0$. Consequently r' cannot have a value greater than $+1$ or less than $-1/(k-1)$.

From the analysis of variance table, we have

$$(nk-1)s^2 = (n-1)ks_b^2 + n(k-1)s_w^2 \tag{6.32}$$

Total	*S.S.* between	*S.S.* within
S.S.	groups	groups

For large values of n the formula (6.32) can be written as

$$s^2 = s_b^2 + \frac{k-1}{k} s_w^2 \tag{6.33}$$

Whence using (6.30) we can express s_b^2 and s_w^2 in terms of s^2 and r' as

$$s_b^2 = \frac{s^2}{k}\{1 + (k-1)r'\} \tag{6.34}$$

and

$$s_w^2 = s^2(1 - r') \tag{6.35}$$

If there are any causes of variation from group to group other than those producing variation within groups it is obvious that ks_b^2 would reflect their effect besides containing the component s_w^2. The quantity in the numerator of r' can therefore be considered to estimate the true variance between groups, that is, variance due to causes other than those producing the within group variation. The quantity in the denominator of r' as we saw in

(6.33) measures the total variation between all individuals, resulting from variation between as well as within groups. The intraclass correlation coefficient is thus seen to be the ratio of two variances, measuring the relative importance of factors producing variation from group to group compared to the total variation in the entire material. If the total variation produced were solely due to group differences there would be no differences within groups, s_w^2 would vanish and r' would equal 1 expressing perfect association within groups.

To illustrate the application of the formula we shall calculate the intraclass correlation coefficient from the analysis of variance set out in Example 4.3. We shall first consider whether rainfall at the two places tends to be alike in each season. For this purpose the pairs of values of rainfall at each place form the groups and for calculating the value of within group variation the sum of squares for 1 degree of freedom due to places must be pooled with the sum of squares for 23 degrees of freedom due to interaction. This gives us the following mean squares:

Between groups	$M.S. = 238{\cdot}96$
Within groups	$M.S. = 3{\cdot}92$
	$k = 2$

Hence

$$r' = \frac{238{\cdot}96 - 3{\cdot}92}{238{\cdot}96 + (2-1)\,3{\cdot}92} = 0{\cdot}968$$

The high value of the correlation coefficient indicates that the rainfall at the two places (which were quite near each other) tends to be alike in the different seasons. If, on the other hand, we wish to consider whether the seasons at the two places are alike, that is, tend to preserve constancy we would have to consider the 24 seasons at each place to form the two groups. The within group variation would be obtained by pooling the sum of squares due to 23 degrees of freedom for interaction with the sum of squares for seasons. We then get the following mean squares:

Between groups	$M.S. = 16{\cdot}37$
Within groups	$M.S. = 121{\cdot}17$
	$k = 24$

Hence

$$r' = \frac{16 \cdot 37 - 121 \cdot 17}{16 \cdot 37 + (24 - 1)(121 \cdot 17)} = -\frac{104 \cdot 80}{2803 \cdot 28}$$

$$= -0 \cdot 0374$$

The low value of the coefficient shows clearly that the rainfall tends to vary from season to season.

The example serves to illustrate the utility of the intraclass correlation. A positive value of the intraclass correlation means that within group variation is smaller than that between groups, while a negative value of the coefficient means that the group means tend to be alike in spite of variation within groups. A value near zero would show that the within and between group variations are of the same order, that is, there is no real variation between groups over and above that within groups. The intraclass correlation is not much used as a measure of association but the idea underlying the coefficient, namely, that of comparing the different components of variation such as variation between and within groups is important and forms the basis for designing of experiments in agriculture. The test of significance of the correlation coefficient follows from the analysis of variance and is, therefore, not required to be made separately.

PART II

DESIGN OF EXPERIMENTS

CHAPTER VII

PRINCIPLES OF FIELD EXPERIMENTATION

7*a*.1 The Problem

In agriculture a research worker is required to experiment mainly in the field. Whether it is new varieties, cultivation practices or methods of seed treatment, he has to try them out in the field before he can assess their value. These objects of comparison in his trials may be termed *treatments*. The simple procedure of trying these treatments each in a different field or plot does not seem adequate to ascertain their relative worth with reasonable confidence. For even after discovering from such a trial that some treatments have given a better performance than others the experimenter is left wondering whether the differences observed are due to treatments, to inherent fertility differences in the soil or some other accidental factors. Ideally, the research worker would like to try the treatments under identical conditions but even with the most uniform land that he can select, he finds that the inherent variation in the soil is quite considerable and the simple procedure of trying out different treatments on single plots side by side in the same field does not suffice for assessing the intrinsic worth of the treatments. A good idea of the nature and extent of fertility variation in land can be obtained from the results of what are known as *uniformity trials*.

7*a*.2 A Uniformity Trial

A uniformity trial consists in growing in a field or piece of land a particular crop with a uniform treatment, dividing the field into small units and harvesting and recording the produce from each of these units separately. The results of one uniformity trial on cotton carried out at the Institute of Plant Industry, Indore, are given in Table 7.1. The trial consisted of 128 rows of cotton with a spacing of 14″ between rows, their length being 186′ 8″. Each unit for harvesting was 4 rows wide and 4′ 8″ long, thus measuring 1/2000 acre. From the results of such a trial we can

TABLE

Uniformity trial on Malvi cotton, 1933–34,

Yield of seed cotton

93	95	111	116	97	82	80	131	80	102	97	95	87	136	118	89	57
49	63	50	115	83	78	69	93	85	84	64	69	51	44	115	61	54
89	90	95	129	97	133	72	115	97	76	91	106	40	147	119	66	122
85	91	115	107	99	77	76	63	70	71	59	78	58	85	61	104	84
63	96	68	69	101	120	96	100	70	110	88	75	101	61	106	113	93
86	82	113	74	117	28	83	81	109	91	53	108	83	74	129	89	79
62	86	119	80	79	92	72	103	89	77	67	88	97	50	101	111	100
94	76	119	75	125	96	95	74	90	63	68	66	96	66	85	74	160
45	96	115	130	99	77	104	116	64	77	32	88	67	76	72	87	59
51	79	111	79	84	84	67	63	66	40	56	80	58	68	52	76	87
81	80	72	102	104	81	42	76	85	80	56	74	93	88	60	111	128
74	51	66	86	71	73	98	87	61	111	81	61	73	61	63	85	115
85	89	112	88	105	76	76	78	72	92	59	81	90	52	82	68	88
84	64	101	90	69	101	87	70	63	75	63	88	74	81	68	87	82
85	55	88	87	73	119	60	76	60	73	77	44	83	82	74	52	97
82	83	102	80	88	57	81	55	75	69	70	84	59	73	77	81	98
46	54	83	60	57	75	73	50	66	31	56	66	43	70	59	35	90
71	68	39	80	73	86	64	89	101	88	63	83	74	78	74	50	126
58	52	85	99	72	96	67	86	37	71	62	48	63	64	106	90	91
65	45	77	78	93	60	94	82	72	81	84	64	75	97	114	109	111
57	52	60	71	55	89	71	53	64	74	95	59	68	94	85	90	98
60	90	65	66	59	55	51	79	80	62	95	94	95	96	114	105	113
59	69	39	79	74	68	54	72	69	67	81	72	56	63	92	62	103
63	78	60	83	51	60	60	72	75	45	75	69	93	89	75	87	114
23	67	47	83	74	58	95	108	81	86	88	99	82	90	133	97	85
39	57	83	70	77	89	79	113	77	102	101	74	79	97	100	41	77
79	29	57	77	75	82	87	77	61	85	108	65	64	95	56	81	153
51	54	74	87	71	74	62	62	90	80	83	60	68	53	82	107	152
58	47	86	69	81	60	73	96	73	59	83	83	89	130	74	76	133
56	78	62	55	60	58	111	54	57	76	82	103	114	92	70	39	87
42	70	68	87	79	89	87	91	85	74	67	98	89	69	50	57	56
57	41	81	80	90	78	61	76	78	87	106	41	49	61	70	78	145
44	68	44	71	75	79	64	68	62	118	60	78	63	53	67	51	84
99	33	65	85	64	74	96	74	95	101	56	59	61	77	57	44	88
58	63	68	71	54	59	84	71	84	87	82	73	73	73	48	42	28
79	66	95	74	81	77	94	106	132	80	142	30	83	46	56	29	82
89	51	66	48	65	82	119	101	90	53	69	66	45	54	61	68	47
62	69	58	66	45	77	94	84	98	44	57	58	17	68	45	39	82
81	106	43	84	46	74	56	109	115	61	60	39	66	61	57	37	109
31	87	69	64	76	82	61	64	87	59	99	31	81	45	50	31	52
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17

7.1

Institute of Plant Industry, Indore

per plot in gm.

106	118	97	72	120	117	117	66	123	116	86	87	79	78	56	1
54	36	88	62	75	52	91	74	66	76	79	54	54	109	87	2
63	72	34	47	106	132	97	70	78	56	81	59	81	78	61	3
104	59	56	70	66	51	83	77	64	87	87	62	80	87	107	4
62	92	109	84	117	101	75	100	104	129	96	107	116	82	101	5
89	75	114	106	94	93	71	103	70	63	114	55	92	67	126	6
96	66	113	98	86	97	69	105	76	87	73	108	104	116	73	7
148	84	110	110	110	83	154	101	103	89	137	79	111	120	82	8
123	67	59	113	131	55	101	105	98	98	57	69	78	62	99	9
72	91	81	153	82	70	133	96	108	68	92	127	91	110	88	10
124	117	118	77	84	104	105	119	101	104	129	53	62	99	75	11
80	82	76	86	97	65	75	68	98	67	69	75	89	72	88	12
94	68	106	87	106	80	88	71	85	82	108	104	98	95	90	13
80	64	105	67	70	58	79	88	58	73	53	85	64	96	86	14
65	59	75	110	79	101	74	66	84	87	100	67	83	61	70	15
74	95	110	119	93	95	81	102	82	74	82	80	75	85	99	16
62	89	96	87	72	80	88	67	67	107	69	79	48	89	59	17
98	104	109	45	106	107	100	137	110	102	79	134	45	60	78	18
86	83	104	55	70	73	67	67	50	91	88	89	70	56	66	19
74	107	86	53	52	90	53	72	70	43	53	106	49	67	62	20
94	84	62	60	54	50	35	64	70	54	62	85	77	78	76	21
49	123	104	60	71	83	81	79	50	69	76	111	79	108	119	22
89	98	64	115	41	73	99	78	72	104	56	121	80	128	157	23
147	138	118	106	106	86	85	95	154	78	98	133	125	133	101	24
113	137	82	130	82	106	96	53	102	103	82	124	108	116	108	25
98	71	98	102	75	77	83	112	106	95	106	93	68	94	103	26
116	45	132	103	88	94	123	76	145	103	107	122	94	108	121	27
40	80	124	130	105	103	60	75	115	138	129	165	135	143	138	28
30	43	81	96	85	86	140	103	92	88	93	97	38	82	103	29
35	46	39	52	157	80	63	59	69	63	72	64	68	102	62	30
54	71	70	69	92	47	41	29	52	46	79	87	74	73	65	31
40	79	103	71	29	51	28	30	37	37	53	49	81	28	36	32
40	64	46	34	15	33	56	32	35	42	37	35	97	56	66	33
91	29	50	40	32	61	47	8	27	54	72	76	67	89	99	34
36	31	13	17	35	18	23	14	22	33	42	52	82	60	72	35
25	41	30	23	46	47	35	40	58	63	61	63	92	57	78	36
44	32	32	50	37	32	53	17	30	43	5	42	36	62	61	37
61	33	72	65	86	37	57	53	44	47	56	76	56	75	87	38
71	81	52	79	92	45	62	49	87	111	73	93	58	56	51	39
40	50	53	57	97	52	81	82	76	54	55	91	12	50	80	40
18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	

prepare what is known as a *fertility contour map* showing lines passing through areas of equal fertility. A fertility contour map for the uniformity trial given in Table 7.1 is shown in Fig. 7.1.

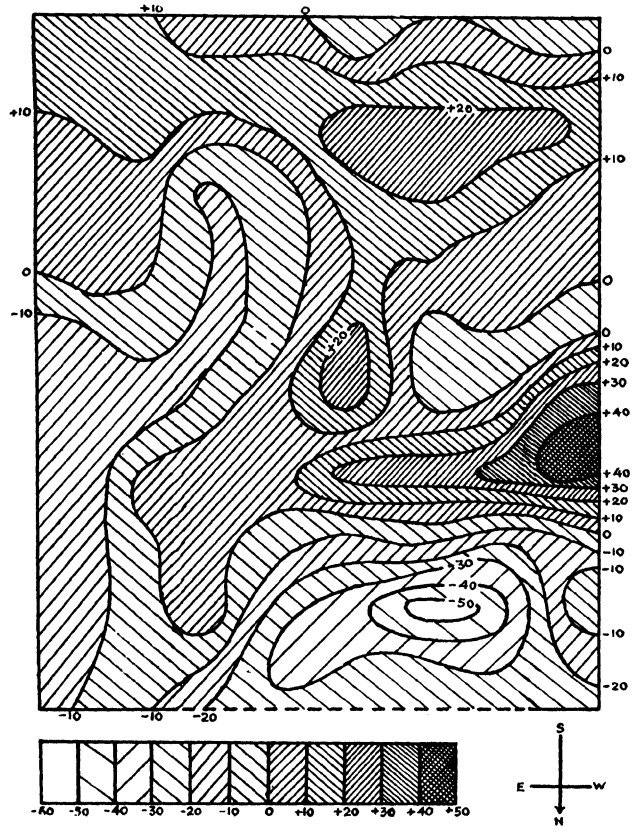


FIG 7.1. Fertility contour map from data of cotton uniformity trial.

An inspection of the fertility contour map shows that not only is there appreciable variation in fertility but that this variation does not follow any systematic pattern, that is, the fertility does not increase or decrease uniformly in any direction but that fertility variations are distributed over the field in an erratic fashion. However, it can be seen that relatively small areas are homogeneous. From the figures in Table 7.1 we can find the standard

deviation of the unit plot yields. This standard deviation gives an index of the inherent variability of the field. We can combine the yields of neighbouring units to give us plots of sizes 2, 4, 8 or more units and calculate from these the coefficient of variation for the plots of each size. The results thus obtained from data in Table 7.1 for plots of various sizes and shapes are given in Table 7.2.

TABLE 7.2

Coefficient of variation for plots of various sizes and shapes

Number of units across rows of cotton	Number of units along rows of cotton			
	1	2	4	8
1	32·0	26·4	22·4	18·3
2	26·4	22·8	20·0	16·8
4	23·4	20·9	18·7	15·8
8	20·6	18·6	17·2	14·5

We find from this table that the yields from plots, say, 2 units wide and 8 units long have a coefficient of variation of 17 per cent. This means that, notwithstanding the uniformity aimed at in respect of seed, sowing, cultivation, etc., there are other factors beyond the control of the experimenter reflected in natural differences in fertility which bring about a variation of this magnitude. It follows that apparent differences of the order of 20 to 25 per cent. can easily arise between the yields of plots under different treatments, without any real differences between the treatments. Such variation from plot to plot, caused by uncontrolled factors, is spoken of as *experimental error*. In order to allow for fluctuations due to experimental error, the research worker intuitively thinks of repeating the treatments. He feels that if the differences observed are obtained consistently in the repetitions, he may accept them as real and not due to fertility variations alone. When treatments are repeated on a number of plots, the observed variation between treatments may be partly due to real treatment differences if there are any and partly to experimental error which influences yield even in the absence of

any real treatment differences. It is necessary, therefore, to evaluate the magnitude of variation due to experimental error and compare with it the observed variation between treatments in order to see if the experiment indicates any real differences in the effects of the treatments.

7*a*.3 REPLICATION

The repetition of the treatments under investigation is known as *replication.* Since the variation in fertility cannot be allowed for directly owing to its unpredictable nature, the experimenter seeks to average out its influence over the different treatments by replication. The procedure, it should be noted, amounts to sampling. If we repeat a single treatment r times the mean of these repetitions will be subject to a standard error of $\sigma/\sqrt{r}$ where σ, the standard deviation of individual plots, is estimated from the experiment. The percentage standard error to which this mean will be subject $= C/\sqrt{r}$ where C is the per cent. coefficient of variation. The standard error of a difference of two treatment means expressed as percentage of the general mean will be $C\sqrt{2/r}$. On this basis we can calculate the number of replications required to enable us to infer the difference between two treatments as significant at a given level when the observed difference exceeds a given per cent. of the mean.

Example 7.1

Calculate the minimum number of replications required so that an observed difference of 10 per cent. of the mean will be regarded as significant at 5 per cent. level, the coefficient of variation of plot values being 12 per cent.

The standard error of the difference of treatment means based on r replications expressed as percentage of the common mean $= 12\sqrt{2/r}$.

When, the ratio,

$$\frac{\text{difference}}{\text{S.E. of difference}} = \frac{10}{12\sqrt{\frac{2}{r}}}$$

exceeds the value 1·96 the difference will be detected as significant at the 0·05 level of significance. Hence

$$\frac{10}{12\sqrt{\frac{2}{r}}} = 1{\cdot}96$$

$$\therefore \quad \sqrt{r} = \frac{1{\cdot}96 \times \sqrt{2} \times 12}{10}$$

i.e.,

$$r \geqslant 11{\cdot}1$$

Thus the minimum number of replications required for detecting the given difference at the 5 per cent. level of significance is the next integer greater than 11·1, that is 12.

However, it may be noted that here the true value of the coefficient of variation is assumed known. In practice this is not often the case and an estimate of the coefficient has to be used. An application of the above formula would then give only an approximate idea of the number of replications required.

It is not merely to provide stability to means that replication is to be regarded a necessity in scientific experimentation. Replication is, indeed, essential for rigorous comparison of treatment effects for a more fundamental reason, namely, that it is only by replication that we have—by considering the differences between plots under the same treatment in different replications—the means of estimating the experimental error.

7*a*.4 Randomization

It is necessary for an objective comparison between treatments to fulfil one more condition, namely, the random allocation of the treatments to various plots. It can be shown that the statistical procedures employed in making comparisons between treatments will hold good only provided the treatments are allocated randomly to various plots. Regarded from a different angle also the necessity of randomization would be clear. By replication the experimenter wishes to average out as far as possible the effects of environmental differences so as to give the various treatments equal scope to show their merit. This brings him to the question of arrangement of plots. How should he arrange the

plots so as to give his treatments equal scope ? In the absence of a prior and exact knowledge of fertility variation in the field he cannot attain this object by a deliberate allotment of the treatments to specific plots. By randomization he can ensure that the various treatments will, in the long run by repetition of the experiment, be subject to equal environmental effects. If there are two treatments *A* and *B* it is obvious that in any particular replication the two will not be affected by the soil variation to an exactly equal extent; but if *A* and *B* were randomly allocated within each replication they will tend to be more and more equally influenced as the number of replications or trials increases. If instead the experimenter allots treatments *A* and *B* to replications in the manner, say, *AB*, *AB*, *AB*, . . . this will not give the treatments equal environment if there is a gradient of fertility across the series of plots. Since the treatment *A* is always to the left it will be consistently favoured or be at a disadvantage depending on the direction of the gradient. The alternative arrangement *AB*, *BA*, *AB*, . . . will not give equal environment to both the treatments either, if there is a cyclic variation of fertility across the series of plots. This applies to other systematic arrangements of plots that have been suggested from time to time for eliminating variations of fertility. A common example of such systematic designs is the chess-board arrangement of plots such as

A	*B*	*C*	*D*
D	*A*	*B*	*C*
C	*D*	*A*	*B*
B	*C*	*D*	*A*

when 4 treatments, *A*, *B*, *C*, *D* are under test.

Since all the treatments appear in each row as well as column the influence of any fertility gradients along the sides of the rectangle will be eliminated by such an arrangement. But then it can be easily seen that if the fertility is decreasing on both sides of the diagonal *AA* it will consistently favour the treatment *A* as against the other treatments. It has to be made clear that even randomization does not remove the difficulty of securing an exactly equal environment for all treatments in any given experiment. Indeed, actual randomization in any particular experiment may result in one of the very systematic arrangements discussed above.

But this will happen in the long run only as frequently as is taken into account by the test of significance applied to treatment differences at a particular level of probability. The merit of randomization lies therefore in providing a rigorous basis for the test of significance in comparing an observed difference between treatment means against the difference brou ht about by unequal environm nt.

Incidentally it might be noted that in Example 7.1 we have tacitly assumed that the treatments were allocated randomly to various plots. This assumption is necessary since the law that the standard error of the mean is equal to $\sigma/\sqrt{r}$, is known to hold for random sampling only.

When the treatments are replicated and the plots allocated randomly to various treatments, we are in a position to test the significance of observed treatment differences by the use of procedures dealt with in Chapter IV. Thus if two treatments A and B are randomly allocated to a number of plots and there are no intrinsic differences between the two treatments the average of A will tend to equal the average of B. We can calculate the significance of the difference $A-B$ by comparing its observed value with its standard error derived from the within treatment variation. The ratio

$$\frac{\text{difference}}{S.E.\text{ of difference}}$$

will then be distributed as t. If the ratio comes out to be significant the hypothesis of the equality of the two treatments, termed the *null hypothesis*, will be suspect.

Example 7.2

Suppose we form twenty plots of size 2×8 units each from the results of uniformity trial shown in Table 7.1. The first plot would consist of first eight units in the first and second rows each, the second plot, of first eight units in the third and fourth rows each and so on. The yields of the twenty plots would then be as follows:

Serial No. of Plot	1	2	3	4	5	6	7	8	9	10
Yield in gm.	1405	1533	1382	1447	1400	1244	1375	1271	1068	1209
Serial No. of Plot	11	12	13	14	15	16	17	18	19	20
Yield in gm.	1033	1042	1155	1128	1098	1179	1067	1191	1178	1165

Ten of these plots may be allocated randomly to each of two dummy treatments labelled *A* and *B*. This is equivalent to an hypothetical experiment conducted on these twenty plots with two treatments between the effects of which there is no intrinsic difference.

The following ten random numbers were found for the treatment *A*:

15, 19, 13, 3, 6, 1, 8, 20, 10, 11.

Treating the yields of these plots as obtained for treatment *A* and of the remaining ten plots as those for treatment *B*, test the significance of the difference between *A* and *B*. We obtain the following table of yields (Table 7.3).

TABLE 7.3

Yields of cotton in gm. per plot in completely randomized experiment with two treatments A *and* B

	Treatment	
	A	*B*
	1405	1533
	1382	1447
	1244	1400
	1271	1375
	1209	1068
	1033	1042
	1155	1128
	1098	1179
	1178	1067
	1165	1191
Total ..	12140	12430
Mean ..	1214·0	1243·0
Grand Total ..	24570	
General Mean, $\bar{x}$ ..	1228·5	

From the figures in Table 7.3 we calculate the following:

	D.F.	*S.S.*
Variation between plots under treatment *A*	.. 9	122274
Variation between plots under treatment *B*	.. 9	289816

Hence we get the standard deviation also called the standard error per plot.

$$s = \sqrt{\frac{122274 + 289816}{18}} = \sqrt{22893 \cdot 9}$$

$$\therefore \text{ Standard error of difference of means} = \sqrt{22893 \cdot 9}\sqrt{\frac{2}{10}}$$

$$= 67 \cdot 6$$

$$\text{Difference } (A - B) = -29 \cdot 0$$

and

$$t = -\frac{29 \cdot 0}{67 \cdot 6} = -0 \cdot 429$$

a non-significant value as expected, there being no intrinsic difference between A and B.

Alternatively, we can analyse the total variation among the twenty plots into the two components:

Between A and B (1 d.f.)

Within A and B (18 d.f.)

and test the significance of the difference $A-B$ by the F test as in Chapter IV. Thus we obtain the following analysis of variance (Table 7.4):

TABLE 7.4

Analysis of variance of plot yields (gm./plot) in Example 7.2

Source	D.F.	S.S.	M.S.
Between treatments ..	1	4205·0	4205·0
Within treatments ..	18	412090·0	22893·9
Total ..	19	416295·0	

From the mean squares in Table 7.4 we obtain

$$F = \frac{4205 \cdot 0}{22893 \cdot 9}$$

$$= 0 \cdot 184$$

$$= (0 \cdot 429)^2$$

as expected since the value of F for 1 degree of freedom in the numerator is the square of t.

The alternative procedure has the advantage that it is also suited to making an overall test of several differences simultaneously, that is, for the comparison of more than two treatments. Thus any number of t eatments A, B, C, etc., might be replicated an equal number of times for simplicity and convenience and allocated randomly to various plots. The analysis of variance of plot yields obtained from these, splitting the total variation among all plots into components between treatments and within treatments, would give the mean squares required for making the F test, for testing the null hypothesis of the equality of all treatments.

7*a*.5 Local Control

The random allocation of treatments to plots while giving an estimate of treatment difference freed from any systematic influence of environment or bias as well as providing a correct test of significance of the difference is not quite efficient. It would be obviously desirable to reduce the experimental error as far as practicable without interfering with the statistical requirement of randomness, because a lower experimental error means that a smaller real difference between treatments can be detected to be significant. The reduction of experimental error can be achieved by making use of the fact observed earlier that adjacent areas in a field are relatively more homogeneous than those widely separated. If instead of randomizing the two treatments in our earlier example all over the field we divide the twenty plots into ten blocks of two plots each and allot the treatments A and B randomly within the plots of a block, the difference between A and B would be subject to the fertility variation within each block alone. Generally this variation would be less than that over the whole field. The following example illustrates such an arrangement and the analysis of data obtained therefrom and would serve to bring out the advantage that might be derived from this plan.

Example 7.3

Suppose we construct another dummy experiment with the uniformity trial data in Example 7.2 by dividing the twenty

plots into ten blocks of two contiguous plots each and allocating randomly the treatments *A* and *B* within each block. Suppose further that by randomization we obtain the following arrangement within blocks:

AB, BA, AB, BA, BA, AB, AB, BA, AB, BA.

Test the significance of the difference between the two treatment means. We have the following table of yields of *A* and *B* plots:

TABLE 7.5

Yield of cotton in gm. per plot in randomized block experiment with treatments A *and* B

Block	*A*	*B*	*A* + *B*
1	1405	1533	2938
2	1447	1382	2829
3	1400	1244	2644
4	1271	1375	2646
5	1209	1068	2277
6	1033	1042	2075
7	1155	1128	2283
8	1179	1098	2277
9	1067	1191	2258
10	1165	1178	2343
Total	12331	12239	24570

From the above table we obtain the following analysis of variance (Table 7.6):

TABLE 7.6

Analysis of variance of plot yields in Example 7.3

Source	*D.F.*	*S.S.*	*M.S.*
Blocks	9	367016·0	40779·6
Treatments ..	1	423·2	423·2
Error	9	48855·8	5428·4
Total ..	19	416295·0	

It can be seen from the table that the error variation is reduced by this arrangement to 5428·4 in comparison to the value

of 22893·9 which we would have obtained from a simple randomized layout without blocks (Example 7.2). This is due to the fact that the treatment difference is now subject only to the between-plot variation within blocks. This variation is generally likely to be lower than the plot-to-plot variation over the whole field regardless of blocks, owing to the positive intraclass correlation between plots of the same block.

Such arrangements in blocks can be extended easily to cases where more than two treatments are tried. Each group of contiguous plots forming a block would contain as many plots as there are treatments and the treatments would be randomly allocated to plots within each block. This arrangement is known as *randomized blocks* and is dealt with at greater length in the next chapter. The principle which this arrangement is used to illustrate here, namely, making use of greater homogeneity of groups of experimental units (due to physical contiguity as far as field experiments are concerned) in order to reduce experimental error, is known as *local control.* Various forms of plot arrangements to suit the requirements of particular problems have been evolved and are known as *experimental designs.* The underlying principle of all these designs is the same, namely, that they seek to provide by means of randomization and replication an unbiased comparison of treatments against their standard errors and aim at reducing these errors with the help of replication and local control.

7*a*.6 Efficiency of Designs

If σ^2 is the variation between plots regardless of blocks then it can be seen from equation (6.35) that the expected value of the error mean square in the analysis of variance in Example 7.3 is $\sigma^2(1-\rho')$ and that of the mean square between blocks of two plots each is $\sigma^2(1+\rho')$ where ρ' is the intraclass correlation between plots within blocks. Equating these two expressions to the respective mean squares we can estimate the values of σ^2 and ρ'. Thus we have from the mean squares in Table 7.6,

$$\sigma^2(1-\rho') = 5428{\cdot}4$$

and

$$\sigma^2(1+\rho') = 40779{\cdot}6$$

Hence

$$\sigma^2 = 23104 \cdot 0$$

and

$$\rho' = 0 \cdot 76504$$

The relative efficiency of various plot arrangements is measured by the ratio of the inverse of error mean squares which they give. The inverse of error mean square is termed the *amount of information* this being larger for an experiment with a lower error and *vice versa*. Since σ^2 denotes the variance per plot in the case of unrestricted randomization, $1/\sigma^2$ will be the corresponding amount of information and since $\sigma^2(1 - \rho')$ is the variance in the case of the paired plot arrangement $1/\sigma^2(1 - \rho')$ will represent the corresponding amount of information. The relative efficiency of the latter arrangement in comparison to the former is thus given by the ratio $\sigma^2/\sigma^2(1 - \rho')$ of the respective amounts of information or in percentage by $100/(1 - \rho')$. Thus in the present case the relative efficiency of the arrangement in blocks is

$$\frac{100}{1 - 0 \cdot 76504} = 425 \cdot 6 \text{ per cent.}$$

or a gain in efficiency of 325·6 per cent. over the simple random arrangement.

CHAPTER VIII

RANDOMIZED BLOCKS AND LATIN SQUARE

8*a*.1 RANDOMIZED BLOCKS

A SIMPLE application of the principles discussed in the last chapter and one of very common use in field trials is the design known as randomized blocks. The idea has been introduced already in the last chapter in connection with the comparison of two treatments. The design is, however, of wider applicability and several treatments can be tried together in the same way. For this purpose, the land on which the trial is to be carried out is divided into as many blocks of the same size and shape as there are to be replications and each of the blocks into as many plots of the same size and shape as there are treatments. If there are t treatments and r replications, there will be r blocks with t plots in each block, giving a total of tr plots altogether. The treatments are allocated randomly to the t plots in the block, the randomization being carried out by either drawing lots or with the help of a table of random numbers. The tr plot yields obtained from the tr plots furnish the data for the comparison of treatments. The comparisons are made using the analysis of variance technique.

The total sum of squares of deviations of the plot yields from the general mean, based on $tr - 1$ degrees of freedom, can be partitioned into the following components: (1) The sums of plot yields for each treatment from all blocks give t treatment totals from which we obtain the sum of squares based on $t - 1$ degrees of freedom due to treatment differences. (2) From r block totals we similarly obtain a sum of squares based on $r - 1$ degrees of freedom due to block differences. (3) The remaining sum of squares based on $(t - 1)(r - 1)$ degrees of freedom due to variation between similarly treated plots in different blocks represents the uncontrolled variation affecting the treatments. It is, therefore, known as the error sum of squares. By comparing the error mean square obtained by dividing the error sum of squares by the number of degrees of freedom with

the treatment mean square similarly obtained, we can test the significance of treatment differences. The procedure will be clear from the following example of a varietal trial carried out in randomized blocks.

Example 8.1

Seven selections of *mung* (*Phaseolous mungo*) numbered 1 to 7 were put under a yield trial in randomized blocks against the local type which formed treatment number 8. There were six replications. Each plot consisted of five rows, 63 feet long. The row spacing being $1\frac{1}{2}$ foot the plot dimensions were $7\frac{1}{2}$ feet $\times 63$ feet. One row on each side of a plot as well as $1\frac{1}{2}$ foot length of plants at either end of each row were discarded from the experimental plot yield to allow for the borders. The net plot thus measured $4\frac{1}{2}$ feet $\times$ 60 feet, that is, 1/160 acre (*cf.* Chapter XV for a discussion of the significance of these experimental details).

Fig. 8.1 gives the plan of the experiment, the strains sown and yield obtained from each experimental plot.

Block I		Block II		Block III	
4	19·5	6	27·0	3	13·5
6	25·0	2	15·0	8	13·0
3	22·0	3	21·0	2	20·5
8	11·5	7	15·0	4	19·0
7	14·5	5	22·0	7	16·0
5	21·0	8	18·0	1	15·0
1	17·5	1	20·0	5	5·5
2	27·0	4	11·0	6	12·0
2	19·5	4	24·0	2	20·5
3	24·5	7	14·0	7	14·0
7	17·5	6	31·5	6	22·0
6	29·0	2	25·0	5	24·5
5	24·5	8	16·5	3	20·5
1	31·5	3	26·0	1	21·0
8	11·0	1	20·0	8	15·5
4	19·0	5	26·0	4	16·5
Block IV		Block V		Block VI	

FIG. 8.1. Plan of randomized block trial with strains of *mung* (*Phaseolous mungo*).

For the purpose of analysis it is necessary to tabulate the yield figures according to strains and blocks in the following manner:

TABLE 8.1

Plot yields in a randomized block trial of mung (Phaseolous mungo) *strains (in oz. per plot)*

Strain \ Block	I	II	III	IV	V	VI	Strain totals	Strain means
1	17·5	20·0	15·0	31·5	20·0	21·0	125·0	20·83
2	27·0	15·0	20·5	19·5	25·0	20·5	127·5	21·25
3	22·0	21·0	13·5	24·5	26·0	20·5	127·5	21·25
4	19·5	11·0	19·0	19·0	24·0	16·5	109·0	18·17
5	21·0	22·0	5·5	24·5	26·0	24·5	123·5	20·58
6	25·0	27·0	12·0	29·0	31·5	22·0	146·5	24·42
7	14·5	15·0	16·0	17·5	14·0	14·0	91·0	15·17
8	11·5	18·0	13·0	11·0	16·5	15·5	85·5	14·25
Block totals	158·0	149·0	114·5	176·5	183·0	154·5	935·5	..

The last two columns of Table 8.1 give strain totals and strain means respectively, while the last row gives the block totals. The grand total of all the 48 plots is obtained from both the strain as well as the block totals—the double procedure serving as a check—and is found to be 935·5 oz.

From these figures we get the various sums of squares thus:

$$\text{Correction factor} = \frac{(935\cdot5)^2}{48} = 18232\cdot5$$

$$\text{The total } S.S. = \{(17\cdot5)^2 + (27\cdot0)^2 + \ldots.\} - C.F.$$
$$= 19734\cdot8 - 18232\cdot5$$
$$= 1502\cdot3$$

$$S.S. \text{ due to Blocks} = \tfrac{1}{8}\{(158\cdot0)^2 + (149\cdot0)^2 + \ldots.\} - C.F.$$
$$= 18598\cdot3 - 18232\cdot5$$
$$= 365\cdot8$$

$$S.S. \text{ due to strains} = \tfrac{1}{6}\{(125\cdot0)^2 + (127\cdot5)^2 + \ldots.\} - C.F.$$
$$= 18720\cdot7 - 18232\cdot5$$
$$= 488\cdot2$$

Hence we get the sum of squares due to error by difference. Thus

$$S.S. \text{ due to error} = 1502\cdot3 - (365\cdot8 + 488\cdot2)$$
$$= 648\cdot3$$

We set down these results in a Table of Analysis of Variance of plot yields as shown below:

TABLE 8.2

Analysis of variance of plot yields in a randomized block trial with mung *strains (oz./plot)*

Source of variation	*D.F.*	*S.S.*	*M.S.*	*F* ratio observed	*F* for $P=0\cdot05$	*F* for $P=0\cdot01$
Blocks ..	5	365·8	73·17	3·95**	2·49	3·60
Strains ..	7	488·2	69·74	3·77**	2·30	3·23
Error ..	35	648·3	18·52	..	..	..
Total ..	47	1502·3	..	..	..	..

Note.—As the tabulated yields are correct only to the nearest half ounce, it is not necessary to carry more than one figure after the decimal point in our calculated sums of squares. The mean squares are however calculated to one more decimal place as they are averages.

Table 8.2 also gives the 5 and 1 per cent. *F* values. It can be seen from the table that the *F* ratio is significant at 1 per cent. for blocks as well as strains. It is customary to indicate significance at 5 per cent. by an asterisk and at 1 per cent. by a double asterisk on the respective *F* values.

8*a*.2. CRITICAL DIFFERENCE

The square root of the error mean square measures the standard error per plot due to uncontrolled environmental effects. The strain means which are means of six plots will be subject, as a

result of environmental fluctuations, to a standard error of $s/\sqrt{6}$ where s is the square root of the error mean square. Hence the standard error of strain mean in our case will be

$$\sqrt{\frac{18 \cdot 52}{6}} = 1 \cdot 76 \text{ oz.}$$

The standard error of difference of means of two strains will be, therefore,

$$1 \cdot 76 \times \sqrt{2} = 2 \cdot 49 \text{ oz.}$$

From the value of the standard error of difference we can calculate the value of the difference which will be just significant at a chosen level of significance. This difference is known as the *critical difference* for the particular level of significance. It is usual to calculate this difference for $P = 0 \cdot 05$ as this level of significance is regarded adequate for most purposes. For calculating the critical difference at the significance level $P = 0 \cdot 05$ we equate the ratio

$$\frac{\text{difference}}{S.E. \text{ of difference}}$$

to $t_{5\%}$ value. We get

$$\frac{\text{difference}}{S.E. \text{ of difference}} = t_{5\%} \text{ for 35 d.f.}$$

$$= 2 \cdot 03$$

Hence critical difference which is $t_{5\%} \times$ standard error of difference

$$= 2 \cdot 03 \times 2 \cdot 49$$

$$= 5 \cdot 05 \text{ oz.}$$

The value of the critical difference is thus 5·05 oz. and strains differing by this amount or more will be recognised as differing significantly. It ought to be clearly understood, however, that the testing of significance of differences between individual treatments should not be done in a trial unless overall significant differences between treatments are indicated by F test in the analysis of variance; for when a large number of treatments are tried,

differences between pairs of treatments may occasionally exceed the critical difference purely through environmental factors. The F test which is an overall test of differences between treatments compares the average variation of the treatments with the uncontrolled variation and if the former is not significantly greater than the latter we have no reason to suspect that any observed differences are real, that is, due to intrinsic differences between treatments. In our example significant differences between strains are indicated from the analysis of variance and so we are justified in comparing the individual strains with the help of the critical difference. Setting down the strains in the order of performance we compare their means in the following manner:

Strain number ..	6	2	3	1	5	4	7	8 (local)
Mean yield oz. per plot	24·42	21·25	21·25	20·83	20·58	18·17	15·17	14·25

Standard error of strain mean, 1·76 oz.

Critical difference at 5 per cent., 5·05 oz.

Strains which do not differ significantly have been underlined by a bar. This method of using the critical difference value for underlining or covering by bars sets of treatments which do not differ significantly among themselves is sometimes adopted as being a concise way of indicating the significance of individual comparisons.

8*a*.3 Number of Replications

An important point to remember when planning a trial is that the estimate of error variation should be based on an adequate number of degrees of freedom. If the degrees of freedom for the error are few, the F ratio required for significance is large and it will not be possible to detect relatively small differences between treatments; in other words the test will be insensitive. We may, for example, consider the data for only two blocks out of the six in the experiment with strains in Example 8.10; say, the first two. We shall obtain the following analysis of variance:

TABLE 8.3

Analysis of variance of plot yields in blocks I and II in Table 8.1 (oz./plot)

Source	*D.F.*	*S.S.*	*M.S.*	*F* ratio observed	*F* 5 per cent.
Blocks ..	1	5·07	5·07	..	..
Strains ..	7	230·94	32·99	1·77	3·80
Error ..	7	130·43	18·63	..	..
Total ..	15	366·44	..	..	..

The significance of strain comparisons is greatly altered, the F ratio for strains being non-significant now. The F ratio for strains has decreased; but the test of significance has been also affected by the increase of F value required for significance. With 7 degrees of freedom for both the numerator and the denominator, the F value required for significance at the 5 per cent. level is 3·80 as compared to the previous value of 2·30. Similarly the t value required for significance of individual strain comparisons will also be greater. This insensitiveness only reflects the fact that the error mean square which is an estimate of the instrinsic variation from plot to plot is estimated with poor precision and consequently a larger difference has to be observed between treatments for the same degree of reliance to be placed on it as indicative of a real difference. The number of degrees of freedom for error is so to speak the effective number of observations on which the estimate of error variation per plot is based. However, while an adequate number of degrees of freedom for error may render the overall test for the differences among treatments sensitive, that alone is not sufficient to make the test for the significance of individual differences precise. When a large number of treatments is tried, we can have an adequate number of degrees of freedom for the error even with two or three replications; but since the standard error of treatment means is obtained by taking the square root of the ratio

$$\frac{\text{error mean square}}{\text{number of replications}}$$

the standard error naturally depends on the number of replications also, the standard error being smaller the greater the number of replications. With a smaller number of replications the standard error of the difference between treatment means as also the value of t required for significance are increased. This means that the critical difference between two treatment means becomes larger for the prescribed level of significance, namely, 5 per cent.

It is difficult to generalise on the number of replications which may be considered adequate when two, three, four, ... treatments are to be tried, since this depends also on the inherent variation from plot to plot for the character to be measured. In the absence of any knowledge regarding the magnitude of variability, the number of replications provided should be sufficient to ensure at least about 12 degrees of freedom for error. This is inferred from the fact that the tabulated values of F for the 5 and 1 per cent. levels of significance cease to fall off rapidly for values of n_2 beyond 12 or so. We have seen that when t treatments are replicated r times in randomized blocks the error is based on $(t-1)(r-1)$ degrees of freedom. From this consideration we find the minimum number of replications for various numbers of treatments to be as follows:

Number of treatments ..	2	3	4	5	6	7	8	9	10
Number of replications ..	13	7	5	4	4	3	3	3	3

If the magnitude of variability per plot were known accurately, the number of replications required could be calculated in the manner of Example 7.1. This, however, is rarely so and in cases where only an estimate of variability per plot is available, it is a much more complicated problem to estimate the number of replications. Reference might be made for more details on this point to a text-book such as *Experimental Designs* by Cox and Cochran.

8*a*.4 Shape and Size of Blocks and Plots: Efficiency

It has been stated already that the error mean square measures the soil fertility variation within blocks. This mean square and hence the standard error of treatment comparisons, therefore,

depends upon the uniformity of land chosen for experimentation as well as the shape and size of blocks and plots. These should be such as to result in the greatest homogeneity within blocks. This object is achieved by making the block shape such that block differences will remove maximum amount of variation. For this purpose it has been found that the blocks should be compact, that is, as nearly square in shape as possible. In dividing the blocks into plots the object is the reverse of this, namely, to have the plots as much alike one another as possible. This object is achieved by having the plots long and narrow and arranged in a row across the block. If there is a large number of treatments, it will not be possible to lay out the plots in a single row without making the blocks appreciably deviate from the compact shape. In such cases, the plots may be laid out in two or more rows. There is however one exception to this procedure. If as in the case of tea gardens, the site of experiment is sloping, there is likely to be a soil fertility gradient in the direction of the slope. In such cases the plots of the same block should all lie side by side with their longer sides parallel to the slope, even if the compactness of the blocks is disturbed by this arrangement. Different blocks may be located at different levels along the slope. As might be guessed, the size of the block also has an effect on the efficiency of the experiment since by increasing the size of a block the heterogeneity within a block is increased and so is the error variation. This increase of heterogeneity within blocks is evidenced from the results of uniformity trials where it has been found that the intraclass correlation between plots in a block decreases with increasing size of the block until the variation within blocks tends to be the same as that between blocks and no very useful purpose is then served by dividing the land into blocks and replicating the treatments in different blocks.

The efficiency of the randomized block design as compared with the simple random arrangement with the same number of replications may be measured by the ratio of the mean square obtained by pooling the blocks and error sums of squares and dividing by the total of the corresponding degrees of freedom, to the error mean square for the randomized block design. In the present case this turns out to be 310·4 per cent. corresponding to an intraclass correlation of + 0·68 obtained as we saw in the

last chapter by equating 310·4 with $100/(1-\rho')$ where ρ' is the intraclass correlation between plots within blocks.

8*b*.1 Latin Square

It may happen that fields chosen for laying out experiments sometimes exhibit fertility variations in strips, that is, there may be alternate strips of high or low fertility caused by cultivation. A layout in randomized blocks will be effective in this situation if the blocks happen to lie parallel to such strips; on the other hand the layout will be extremely inefficient if the blocks happen to be placed across the strips. The direction of alignment of such strips is rarely known initially. An ingenious method of eliminating such variation in fertility consists in an experimental arrangement which simultaneously controls variation in two directions at right angles. Such a layout is known as a *latin square*. In this design there have to be as many replications as there are treatments. The experimental area is divided into plots arranged in a square in such a manner that there are as many plots in each row as there are in each column, this number being also equal to the number of treatments. The plots are then assigned to the various treatments such that every treatment occurs only once in each row and once in each column. This can be done in a large number of ways and the way it is to be done in any particular layout must be determined randomly.

The different ways in which an experiment could be arranged in a latin square have been completely enumerated for squares of sizes 7×7 and smaller. In the *Statistical Tables* by Fisher and Yates sets of 4×4 and 5×5 latin squares are given from which all the possible arrangements could be obtained by permuting, that is, by obtaining all possible orders of rows keeping the first row unchanged and then permuting all columns. For size 6×6 also a set of squares is given from which all possible latin squares can be obtained by permuting rows, columns and letters, the permutation of letters meaning that various letters are assigned to various treatments in all possible ways. For 7×7 square, only four squares are given in the *Statistical Tables* by Fisher and Yates; but Saxena (1950) has enumerated all the different squares which would be found in the reference cited. For squares 8×8 to 12×12 Fisher and Yates give one square

each from which a large number of squares could be obtained by permutations of rows, columns and letters. The procedure of randomization is explained in the *Introduction to the Statistical Tables* and briefly consists in choosing for sizes 6×6 and smaller one of the given squares randomly with the help of key numbers noted under the squares and then permuting rows, columns and letters according to directions given in the Introduction. For squares of bigger size it is enough to take the square given (one of the four given in the case of a 7×7 square) and permute rows, columns and letters randomly. We shall illustrate the procedure with reference to a 5×5 square.

The key numbers given under the squares range upto 56. Choosing one of these numbers randomly with the help of the two-digit table in Appendix II, column three, say, we get the number 20. Being the second key number under the square we have to choose a square conjugate of the one given, that is, one obtained by treating the rows as columns and the columns as rows. We thus get the following square:

	A	*B*	*C*	*D*	*E*
(1)	*B*	*C*	*E*	*A*	*D*
(2)	*C*	*A*	*D*	*E*	*B*
(3)	*D*	*E*	*A*	*B*	*C*
(4)	*E*	*D*	*B*	*C*	*A*

We have to permute randomly all the rows of this square except the first. For this purpose we number the rows to be permuted as shown above and refer to a one-digit column in Appendix II, say fourth. In this column we might start anywhere and move up or down. We start at the top, say, and proceeding downwards find the numbers 1 to 4 in the order 2, 1, 4, 3. By changing the order of rows accordingly we get the square

(1)	(2)	(3)	(4)	(5)
A	*B*	*C*	*D*	*E*
C	*A*	*D*	*E*	*B*
B	*C*	*E*	*A*	*D*
E	*D*	*B*	*C*	*A*
D	*E*	*A*	*B*	*C*

For columns we similarly find a random permutation or order and it comes out to be, 1, 4, 3, 2, 5. Arranging the columns in that order we get the square

A	*D*	*C*	*B*	*E*
C	*E*	*D*	*A*	*B*
B	*A*	*E*	*C*	*D*
E	*C*	*B*	*D*	*A*
D	*B*	*A*	*E*	*C*

This is the selected square and the experiment will be laid out by applying treatment *A* to plots corresponding to positions of *A*'s in the above square, treatment *B*, to plots corresponding to positions of *B*'s and so on. It might be noted that the permutation of rows and columns has not affected the characteristic feature of a latin square, namely that each treatment occurs once and only once in each row and in each column. The layout and analysis of the experiment would be clear from the following example of a varietal trial with wheat carried out in the form of a latin square.

Example 8.2

Five varieties of wheat, *A*, *B*, *C*, *D* and *E* were tried. The gross size of the plot was 18 feet × 22 feet, the net plot being 14 feet × 18 feet. Thus the whole experiment occupied an area 90 feet × 110 feet. The plan, the varieties sown in each plot and yields obtained in oz. have been given in Fig. 8.2.

B 90	E 80	C 134	A 112	D 92
E 85	D 84	B 70	C 141	A 82
C 111	A 90	D 87	B 84	E 69
A 81	C 125	E 85	D 76	B 72
D 82	B 60	A 94	E 85	C 88

FIG. 8.2. Plan of wheat varietal trial in a latin square.

For the analysis we need the row and column totals as well as the treatment totals obtained from plot yields. From the 25 plot yields we obtain the sum of squares for 24 degrees of freedom. From this we subtract the sums of squares due to row, column and treatment differences each based on 4 degrees of freedom calculated from the respective totals, leaving 12 degrees of freedom for error. The results of analysis of the above experiment are given in Table 8.4:

TABLE 8.4

The analysis of variance of plot yields in oz.

Source	D.F.	S.S.	M.S.	F ratio observed	$F_{\cdot 05}$	$F_{\cdot 01}$
Rows ..	4	1075·0	268·8	2·65	3·26	5·41
Columns ..	4	1003·8	251·0	2·48	3·26	5·41
Varieties ..	4	6139·8	1535·0	15·16	3·26	5·41
Error ..	12	1215·2	101·3	..	..	..
Total ..	24	9433·8	..	..	..	..

The F ratios for rows and columns are not significant at $P = 0{\cdot}05$ while that for varieties is very highly significant. The fact that there are no significant differences between rows and columns shows that the latin square arrangement has not been advantageous. The trends in fertility along rows and columns are not very strong and it is possible that a randomized block trial with compact blocks and long and narrow plots laid out on the same land might have given as small an error. In fact, the latin square arrangement is suitable only in the special cases where the land exhibits marked trends in fertility. This design, since it requires as many replications as there are treatments, is suitable chiefly for 4 to 8 treatments. For comparison of a smaller number of treatments the number of replications is found to be inadequate while for a larger number of treatments the number has to be unduly increased. For the small number of treatments, however, more than one latin square may be laid out to secure adequate replication.

8c.1 Randomized Blocks *Versus* Latin Square

The randomized block design is superior to the latin square in many ways. The randomized block design is available for a wide range of treatments and there is no restriction on the number of replications. The analysis of variance is also more flexible. If there is an attack of some pest or disease in one or two of the blocks the data for these blocks can be easily omitted without any complication in the analysis, while results from a latin square experiment necessitate a much more complicated analysis under similar circumstances. In the field also the randomized block trial is easier to manage. It can be accommodated equally well in a rectangular or square field or a field of any other shape, while for a latin square trial it is necessary that the shape of the field should be approximately square or rectangular. When there are simultaneous trends of fertility variations in two directions at right angles (or what amounts to a diagonal trend in fertility), the latin square design is likely to be efficient. The latin square arrangement is also of considerable value in some other branches of agricultural research such as in animal experiments where it might be employed to control simultaneously two factors contributing to the experimental error. For example, litter and body-weight differences may be controlled in trials with guinea-pigs by assigning them to the rows and columns of a latin square.

CHAPTER IX

FACTORIAL EXPERIMENTS

9*a*.1 The Factorial Concept

We have considered in the previous chapters suitable designs for simple experiments involving the testing of variation in a single factor, be it different varieties of a crop, different kinds of manures, different doses of the same manure or methods of cultural treatment. In practice the experimenter has to deal quite often with simultaneous variation in more than one factor. He may be required for instance to find the most suitable level of irrigation and at the same time the optimum dose of a nitrogenous top dressing. Following the traditional procedure, he may investigate the problems one by one, varying a single factor at a time in simple experiments. He may first vary the irrigation levels and, adopting the optimum level as indicated by this experiment, proceed to find the optimum dose of nitrogen supply. The soundness of this approach rests on the implication that the response to different levels of water-supply is independent of the amount of manure given. To make such an assumption under all situations is obviously unwarranted. In the particular instance cited here, it is, in fact, known for most crops that upto a limit a higher level of irrigation is required to secure an adequate response from a higher dose of manure. The two factors, in other words, do not have independent or, what is technically termed, additive effects, but interact with each other. It will be seen that in such a situation the fixing of the supply of manure at a particular level in the first experiment in which the levels of irrigation are varied has to be done quite arbitrarily. Consequently the results are of limited application being strictly appropriate to the particular level of manuring used. What is much more serious, if the two factors do interact or depend upon each other, the optimum combination of the two cannot be discovered by conducting two simple experiments in the above manner. In fact, it is not possible to find from such experiments whether the two factors actually interact with each other or are independent in

their effects. The only effective approach for answering these questions lies in investigating the effect of the two factors together by comparing in one and the same experiment all the possible combinations of the levels of both the factors. This approach is known as the *factorial concept* of experimentation. We will illustrate this concept with the help of an example.

Example 9.1

An experiment was laid out on irrigated wheat to compare different varieties and to study the response to the application of ammonium sulphate. There were three varieties, *C.* 518, *C.* 591 and *C.* 520 which may be regarded as three levels of the first factor and conveniently designated as v_1, v_2, v_3 and four levels of the second factor, or dose of ammonium sulphate, *viz.*, $n_0 = 0$ lb. N per acre (unmanured control), $n_1 = 20$ lb. N per acre, $n_2 = 40$ lb. N per acre and $n_3 = 60$ lb. N per acre. There were thus 3×4 or 12 treatment combinations. These were tested in a randomized block design with six replications. Table 9.1 gives the plan and yields.

TABLE 9.1

*Layout and yields of wheat in varietal-*cum*-manurial trial*

(Yield in lb. per plot; plot size 1/80 acre)

REPLICATION I		REPLICATION II	
v_1n_1	17·4	v_3n_1	16·2
v_1n_2	14·8	v_1n_3	14·0
v_2n_3	17·1	v_2n_1	15·9
v_2n_2	18·6	v_2n_2	15·1
v_1n_0	15·2	v_3n_2	14·6
v_3n_1	17·5	v_1n_2	12·6
v_2n_1	17·1	v_3n_0	16·3
v_2n_3	19·3	v_1n_1	16·0
v_3n_3	19·7	v_2n_3	18·0
v_3n_0	13·3	v_1n_0	13·1
v_1n_3	15·8	v_2n_0	13·1
v_3n_0	14·4	v_3n_3	15·5

TABLE 9.1—(*Continued*)

REPLICATION III		REPLICATION V	
v_3n_0	15·3	v_1n_1	15·2
v_2n_3	19·0	v_2n_3	17·7
v_2n_2	20·3	v_1n_2	17·2
v_1n_1	16·0	v_1n_0	14·6
v_2n_0	17·2	v_3n_2	20·5
v_3n_3	19·5	v_3n_1	16·5
v_1n_2	16·9	v_2n_1	16·0
v_3n_2	21·6	v_3n_0	15·4
v_2n_1	20·4	v_3n_3	18·2
v_3n_1	19·9	v_1n_3	13·3
v_1n_3	17·1	v_2n_2	17·0
v_1n_0	16·4	v_2n_0	15·4
REPLICATION IV		REPLICATION VI	
v_2n_2	17·4	v_3n_0	13·7
v_1n_3	14·8	v_2n_0	14·8
v_3n_0	14·2	v_1n_2	17·4
v_3n_3	16·3	v_2n_2	16·5
v_3n_2	19·2	v_1n_1	17·8
v_2n_1	15·0	v_2n_1	12·3
v_1n_0	14·9	v_2n_3	18·8
v_1n_2	18·3	v_3n_3	20·3
v_2n_3	18·0	v_3n_2	14·5
v_3n_1	15·8	v_1n_3	18·6
v_2n_0	13·2	v_3n_1	14·1
v_1n_1	15·9	v_2n_0	15·4

Considering the 12 treatment combinations together, we can partition the total variation between plot yields into components for blocks, treatments and experimental error in the manner

dealt with already. The resulting analysis of variance is given in Table 9.2.

TABLE 9.2

Simple analysis of variance of plot yields of wheat trial

(Unit: lb. per plot of 1/80 acre)

Source of variation	*D.F.*	*S.S.*	*M.S.*	*F*
Blocks ..	5	68·29	13·66	5·40**
Treatments ..	11	120·03	10·91	4·31**
Error ..	55	139·18	2·53	..
Total ..	71	327·50	..	..

It is seen from Table 9.2 that the treatment combinations differ significantly at 1 per cent. level. Setting out the treatment means in lb. per acre in the descending order of their magnitude and noting that the critical difference at 5 per cent. level will be

$$t_{5\%} \times \sqrt{2} \times \sqrt{\frac{2 \cdot 53}{6}} \times 80 = 147 \cdot 2$$

the result will be as follows:

Treatment	v_3n_2	v_2n_3	v_3n_3	v_2n_2	v_3n_1	v_1n_1	v_1n_2	v_2n_1	v_1n_3	v_1n_0	v_3n_0	v_2n_0
Average yield in lb./acre	1480	1448	1443	1399	1333	1311	1296	1289	1248	1195	1191	1160

It will be observed that varieties v_2 and v_3 have given higher yields than v_1 in the presence of nitrogen; there is little to distinguish between the doses n_2 and n_3 in respect of any of the varieties, but there is an increased response to these as compared to n_1, in the case of variety v_2.

The above analysis does not, however, bring out effectively all the advantages of trying the varieties and the doses of ammonium sulphate together in a factorial scheme and does not elicit all the information that the experiment was planned to give. Let us consider, first of all, the yields (for the sake of convenience taken as totals of six plots each) under the various treatment combinations arranged in a two-way table as in Table 9.3.

TABLE 9.3

Treatment totals of six plots each in lb.

Varieties	Doses of nitrogen				*Total*
	n_0	n_1	n_2	n_3	
v_1 ..	89·6	98·3	97·2	93·6	378·7
v_2 ..	87·0	96·7	104·9	108·6	397·2
v_3 ..	89·3	100·0	111·0	108·2	408·5
Total ..	265·9	295·0	313·1	310·4	1184·4

A cursory examination of the table indicates that there is a response to the first dose n_1 of nitrogen. It is not possible however to decide without further statistical examination whether there is a further response to n_2 over n_1 and whether the varietal differences can be regarded as real. It will also be seen that while the yield (given as total of six plots) of v_3 in the absence of nitrogen is just less than that of v_1 by 0·3 lb., v_3 outweighs v_1 by 1·7 lb. under dose n_1 of nitrogen, by 13·8 lb. under n_2 and by 14·6 lb. under n_3. It would appear therefore that the differences between varieties are not independent of the application of nitrogen. We would like to know if such an inference may be drawn reasonably from the differences quoted above. We may enquire specifically whether the difference between v_3 and v_1 under dose n_1 can be regarded as significantly different from the difference between v_3 and v_1 under dose n_2. The difference between these two differences can be expressed as a linear combination of individual plot yields and carries, therefore, a single degree of freedom, which belongs to the 11 degrees of freedom for treatment differences. We may symbolically express the difference as

$$[(v_3n_2 - v_1n_2) - (v_3n_1 - v_1n_1)] = (v_3 - v_1)(n_2 - n_1)$$

and test its significance by the t-test, its standard error being estimated as $s\sqrt{24}$ (where s is the square root of the error mean square) since 24 plots are in all involved in the linear expression for the difference, each term with a coefficient ± 1. Hence

$$t = \frac{13{\cdot}8 - 1{\cdot}7}{6{\cdot}3} = 1{\cdot}9$$

which for 55 degrees of freedom is just short of the significance level. In the same manner other comparisons which are of interest could be made, but this procedure of merely treating the data as in the case of a simple randomized block design by testing the overall mean square for treatments against the error mean square and comparing the individual treatment combinations with one another will not permit us to elicit all the information that the experiment was planned to provide. The information we seek from the experiment can be set out as answers to three explicit questions: (1) whether the experimental data indicate significant differences among varieties on the one hand, (2) among the effects of the doses of nitrogen on the other, and (3) whether there is any interaction between these two factors. These points can be brought out by a simple extension of the technique of analysis of variance by further subdividing the treatment sum of squares and the corresponding degrees of freedom into components corresponding to (i) the average differences among varieties (technically called the *main effect* of varieties and denoted by the symbol V); (ii) the average differences among the effects of different doses of nitrogen (main effect N); and (iii) the interaction between varieties and doses of nitrogen (interaction VN). Among the three varieties two independent comparisons are possible. The main effect of varieties, V, has therefore 2 degrees of freedom. Similarly the main effect of nitrogen N has 3 degrees of freedom. The remaining 6 degrees of freedom out of the total of 11 for treatments are attributable to the interaction VN. The sums of squares corresponding to V, N and VN can be obtained by treating the two-way table of varieties and doses (Table 9.3) in the same manner as we would treat a two-way table of blocks and treatments in a simple randomized block design. The sum of squares for rows with a divisor of 24, since each row total is a total of 24 plots, gives the sum of squares for main effect V. Thus sum of squares for V is given by:

$$\frac{(378{\cdot}7)^2 + (397{\cdot}2)^2 + (408{\cdot}5)^2}{24} - \frac{(1184{\cdot}4)^2}{72} = 18{\cdot}86$$

The sum of squares for columns, with a divisor of 18, gives the sum of squares for the main effect N, namely,

$$\frac{(265\cdot9)^2 + (295\cdot0)^2 + (313\cdot1)^2 + (310\cdot4)^2}{18} - \frac{(1184\cdot4)^2}{72}$$
$$= 78\cdot15$$

The sum of squares for individual values of the two-way table is the treatment sum of squares in Table 9.2. We can therefore derive the sum of squares for interaction *VN* by subtracting the sums of squares due to *V* and *N* from the treatment sum of squares. The interaction sum of squares is thus equal to

$$120\cdot03 - 18\cdot86 - 78\cdot15 = 23\cdot02$$

The treatment sum of squares is in this manner partitioned as shown in Table 9.4.

TABLE 9.4

Partitioning of treatment sum of squares

Source of variation		*D.F.*	*S.S.*
Main effect of varieties *V*	..	2	18·86
Main effect of doses *N*	..	3	78·15
Interaction *VN*	..	6	23·02
Treatments	..	11	120·03

The mean square for any treatment effect *V*, *N* or *VN* can be tested against the error mean square in the usual manner for significance of each of these effects. Inserting the components of the treatment sum of squares in the analysis of variance (Table 9.2) we have finally the following table of the complete analysis of variance for the factorial experiment.

TABLE 9.5

Complete analysis of variance

(Unit: lb. per plot of 1/80 acre)

Source of variation		*D.F.*	*S.S.*	*M.S.*	*F*
Blocks	..	5	68·29	13·66	5·40**
Varieties *V*	..	2	18·86	9·43	3·72*
Doses *N*	..	3	78·15	26·05	10·30**
Interaction *VN*	..	6	23·02	3·83	1·51
Error	..	55	139·18	2·53	..
Total	..	71	327·50	..	..

The *F*-tests shown in the table indicate that the main effects for varieties as well as for doses of nitrogen are significant, the latter at 1 per cent. level, but the interaction is not significant indicating that the data do not reveal definite evidence of differential response of varieties to the different doses of nitrogen. This unambiguous information could not have been obtained without making the critical analysis of the results in the way indicated above. It may be pointed out that the suggestive indication obtained earlier that there is a differential response to v_3 relatively to v_1 at doses n_1 and n_2 has to be rejected in the light of the verdict of non-significance of the interaction effect.

The results are summarized in the two-way table as shown in Table 9.6. The summary gives the average yields in standard units, namely, pounds per acre, to facilitate comparison with previous experience. It may be noticed that the critical difference is given for both the sets of marginal differences corresponding to the two main effects which have proved significant. Had the interaction effect also proved significant a third critical difference for the difference between any two items in the body of the two-way table would have been necessary.

TABLE 9.6

Summary of average yields in lb. per acre

Varieties	Doses of nitrogen				*Average*
	n_0	n_1	n_2	n_3	
v_1 ..	1195	1311	1296	1248	1262
v_2 ..	1160	1289	1399	1448	1324
v_3 ..	1191	1333	1480	1443	1362
Average ..	1182	1311	1392	1380	1316 General mean

C.D. (5 per cent.) for varietal means = 73·6 lb. per acre.

C.D. (5 per cent.) for nitrogen means = 85·0 lb. per acre.

In the present example the treatments have been laid out in simple randomized blocks. They could as well have been tested in the latin square layout. With the large number of

treatment combinations usually required in a factorial experiment, however, the latin square arrangement would not be preferred normally for reasons already mentioned in the previous chapter.

9*a*.2. Advantages of a Factorial Experiment

We are now in a position to make an appraisal of the advantages of adopting the factorial approach. One great advantage has been discussed already, namely, that it is the factorial experiment alone which can furnish information regarding the interactions between the various factors under study. Another attractive feature of the factorial scheme is the comprehensiveness of the conclusions drawn from it. If the interaction between the factors is significant, we cannot consider the effect of any level of one factor without taking into account the levels of the other factor also and the optimum combination of the two can be discovered only through a factorial experiment. On the other hand, if the two factors do not interact, we may speak of the response to one factor irrespective of the level of the other factor. Whereas these advantages are obvious, it might appear at first that the increase in comprehensiveness of the conclusions from the factorial design is being achieved only at the expense of precision of the comparisons relating to the response to individual factors. It is a remarkable fact, however, that far from there being any loss of accuracy in adopting the factorial scheme, each such comparison may be made with as great a precision as though the entire experiment had been devoted to that comparison alone. In the illustrative experiment discussed above, for instance, the comparisons between the 4 doses of nitrogen are based on averages of 18 plots each as would be the case even if the entire experiment were a simple manurial trial. Similar is the case with the main response to varieties. Had separate experiments been conducted for the two factors, nearly twice the number of plots would have been required to attain the same level of accuracy and even then we would have to fix arbitrarily the level of the factor which is not varied. We might yet find in the end that the results are not applicable to the range of conditions met with in actual practice. It may be contended that both these shortcomings can be got over by conducting three simple manurial experiments, one with

each of the three varieties and by combining their results. Such a procedure would amount, however, to a crude approximation to the factorial approach but with the serious drawback that the results of the three different experiments, being subject to differences in experimental conditions, may not be at all comparable. Even if they are, their averages will be subject to much greater experimental variation than the corresponding averages of the factorial experiment. It will thus be evident that when more than one factor are under study, the factorial approach is the most informative and the most efficient method of experimentation.

CHAPTER X

CONFOUNDING

10*a*.1 Control of Soil Heterogeneity

The advantages of the factorial scheme of experimentation compared to simple experiments with one factor at a time have been point d out in the previous chapter. The factorial scheme has, however, certain drawbacks which tend to affect its efficiency adversely. We have been considering in the previous chapter a varietal-*cum*-manurial trial on wheat with 3 and 4 levels of the two factors respectively making 12 treatment combinations altogether. Suppose we wish to enlarge the scope of the enquiry by introducing yet another factor, say irrigation, at 3 levels. The treatment combinations would now number 36. With plots of the same size as in the previous experiment, it is obvious that each block, having to contain 36 plots instead of 12, would be enlarged three times in size and consequently would become more heterogeneous. The plot to plot variation within blocks, on which is based the estimate of experimental error, would thus increase and lower the precision of the experiment. One possibility of reintroducing homogeneity within blocks, would lie in reducing the size of the block by reducing the size of the plot. Convenience of agricultural operations and the need for allowing non-experimental border rows however place a lower limit on the size of the plot. Consequently wherever a large number of factors or a large number of levels of each factor or both are included in a trial the difficulty of controlling the size of the block will invariably arise. The only way of overcoming this difficulty and of retaining simultaneously the comprehensiveness of the factorial concept and the advantage of increased precision through local control by having homogeneous blocks, is to adopt the principle of *confounding*, that is, to give up the restriction of the randomized block design that each block should form a complete replication and, instead, divide the replication into two or more compact blocks of suitable size. Such a subdivision of a replication into two or more blocks, however, disturbs the precision with which

different treatment comparisons are made. While it permits some treatment comparisons to be made with the highest possible accuracy, others are affected by differences between blocks within a replication in addition to plot to plot variation within blocks and in the limiting case the effect of block differences may be so overwhelming that certain comparisons may have to be sacrificed altogether. The treatment combinations to be allotted to different blocks therefore require to be grouped in such a manner as to ensure that the important comparisons in which we are interested are kept free from block differences. This important device of confounding certain relatively unimportant treatment contrasts with blocks plays a very useful role in many exploratory experiments, in which a number of factors are required to be tried simultaneously in the absence of any previous knowledge regarding their interactions. Any detailed treatment of the subject is beyond the scope of this book. We shall only attempt to illustrate the concept with simpler examples, one with 4 factors each at 2 levels and the other with 3 factors each at 3 levels.

10*b*.1 A 2^4 CONFOUNDED EXPERIMENT

Example 10.1

At the Nagina Experimental Station, Uttar Pradesh, an experiment on paddy was conducted with four factors at two levels each, namely, (i) two varieties of paddy T.21 and T. 22, denoted as v_1 and v_2 for brevity; (ii) two seed rates of 25 sr. and 50 sr. per acre, r_1 and r_2 (1 seer = 2·06 lb); (iii) ammonium sulphate applied at the two levels of nitrogen, 0 lb. and 50 lb. per acre, n_0 and n_1 respectively, and (iv) superphosphate applied at the two levels of P_2O_2, 0 lb. and 50 lb. per acre, p_0 and p_1 respectively. The experiment thus included 16 treatment combinations. We designate it as a $2\times2\times2\times2$ or 2^4 experiment. The layout and yields are given in Table 10.1.

It will be seen that the design adopted is not a simple randomized block design but that each replication of 16 treatment combinations is spread over two blocks, with the expectation that a much better control over plot to plot variation could be had with blocks of 8 plots only instead of 16. Blocks 1 and 2

TABLE 10.1

Plan and yields of 2^4 experiment on paddy with partial confounding of second order interactions

(Yields in lb. per plot: plot size 1/68 acre)

Block 1		Block 4	
$v_1r_2n_1p_1$	37·5	$v_1r_2n_0p_0$	27·5
$v_2r_2n_0p_0$	18·8	$v_1r_1n_0p_1$	15·5
$v_1r_1n_0p_1$	17·8	$v_1r_1n_1p_1$	14·0
$v_2r_2n_0p_1$	26·9	$v_2r_2n_1p_1$	28·3
$v_2r_1n_1p_1$	34·0	$v_1r_2n_1p_0$	19·5
$v_1r_1n_0p_0$	15·5	$v_1r_2n_0p_1$	32·8
$v_1r_2n_1p_0$	37·8	$v_2r_1n_0p_0$	31·0
$v_2r_1n_1p_0$	35·3	$v_2r_1n_1p_0$	27·5
Block 2		Block 5	
$v_1r_2n_0p_1$	33·8	$v_1r_1n_0p_1$	16·0
$v_2r_2n_1p_0$	26·3	$v_1r_1n_1p_0$	22·8
$v_2r_1n_0p_0$	29·3	$v_2r_1n_0p_0$	29·8
$v_1r_1n_1p_0$	17·8	$v_1r_2n_1p_0$	30·3
$v_2r_1n_0p_1$	29·0	$v_2r_2n_1p_1$	45·0
$v_1r_2n_0p_0$	34·3	$v_2r_1n_1p_1$	38·0
$v_2r_2n_1p_1$	30·8	$v_1r_2n_0p_1$	21·5
$v_1r_1n_1p_1$	23·8	$v_2r_2n_0p_0$	28·0
Block 3		Block 6	
$v_1r_1n_1p_0$	19·0	$v_2r_1n_1p_0$	44·8
$v_2r_1n_0p_1$	29·5	$v_1r^2n_1p_1$	30·3
$v_1r_2n_1p_1$	25·5	$v_2r_2n^1p_0$	37·0
$v_2r_2n_1p_0$	32·0	$v_1r_1n_0p_0$	21·8
$v_2r_2n_0p_0$	31·3	$v_1r_2n_0p_0$	23·8
$v_2r_1n_1p_1$	37·0	$v_2r_1n_0p_1$	33·0
$v_1r_2n_0p_1$	13·5	$v_1r_1n_1p_1$	24·8
$v_1r_1n_0p_0$	16·0	$v_2r_2n_0p_1$	31·5

TABLE 10·1—*Contd.*

Block 7		Block 8	
$v_2r_2n_0p_0$	33·0	$v_1r_1n_1p_1$	28·3
$v_2r_1n_1p_0$	36·3	$v_1r_2n_0p_1$	25·5
$v_1r_1n_1p_0$	24·3	$v_1r_2n_1p_0$	24·0
$v_1r_1n_0p_1$	17·5	$v_2r_1n_0p_0$	36·3
$v_2r_1n_0p_1$	35·0	$v_2r_1n_1p_1$	42·0
$v_1r_2n_1p_1$	25·8	$v_2r_2n_0p_1$	27·5
$v_2r_2n_1p_1$	30·0	$v_2r_2n_1p_0$	30·8
$v_1r_2n_0p_0$	23·0	$v_1r_1n_0p_0$	18·5

constitute a replication, blocks 3 and 4 form another, blocks 5 and 6 a third and blocks 7 and 8, fourth. The treatment comparison comprising of the contrast between the group of treatments allotted to two blocks of the same replication is inseparable from the natural soil fertility difference between the two blocks and is therefore said to be confounded with blocks in that replication. Thus in replication I the treatment contrast

$$\begin{aligned}&\{v_2r_2n_1p_1 + v_1r_1n_1p_1 + v_1r_2n_0p_1 + v_2r_1n_0p_1 + v_2r_2n_1p_0 + v_1r_1n_1p_0\\&\quad + v_1r_2n_0p_0 + v_2r_1n_0p_0\}\\&\quad - \{v_1r_2n_1p_1 + v_2r_1n_1p_1 + v_2r_2n_0p_1 + v_1r_1n_0p_1 + v_1r_2n_1p_0\\&\quad + v_2r_1n_1p_0 + v_2r_2n_0p_0 + v_1r_1n_0p_0\}\end{aligned}$$

is confounded with blocks. In order to see what effect precisely the contrast represents physically, we shall have to extend the concept of interaction to more than two factors. In doing so we shall also derive expressions for the main effects and the various interactions in the following sections.

10*b*.2 SYMBOLICAL REPRESENTATION OF TREATMENT COMBINATIONS

At the outset we shall add a word regarding symbols. It will have been noticed that in describing the varietal-*cum*-manurial trial in the last chapter, the treatment combinations have been denoted by the combinations of small letters with

numerical suffixes, the former representing the factors and the latter the levels of each factor. The corresponding effects have been denoted by capital letters. This is the convention. In the present experiment there are four factors each at two levels which we have denoted as follows:

(i) Variety : v_1 and v_2;

(ii) Seed rate : r_1 and r_2;

(iii) Nitrogen : n_0 (no N) and n_1; and

(iv) Phosphorus: p_0 (no P) and p_1.

Thus $v_1r_2n_1p_0$ stands for the treatment combination consisting of variety T. 21 sown at 50 sr./acre, with a dose of nitrogen at 50 lb./acre but with no phosphorus. It is, however, convenient in the case of factors at two levels to adopt the further simplification of replacing the lower level of each factor by 1 and allowing the small letter denoting the factor itself to stand for the other level. In writing out the treatment combinations, the symbols are further treated as if they are algebraic quantities which can be multiplied out. Thus the combination $v_1r_2n_1p_0$ would be replaced by (1) (r) (n) (1) and shortened merely into rn. It is easily verified that if the convention is strictly followed it causes no ambiguity once the factors involved in an investigation and their levels are known. Thus knowing that the factors involved are v, r, n and p the symbol rn can be interpreted without ambiguity as standing for the combination of the lower levels of factors v and p with the higher levels of r and n.

10*b*.3 Symbolical Representation of Main Effects and First Order Interactions

We shall next consider the symbolic representation of treatment effects. Consider the response of one variety compared with another. In each replication we have 8 plots under each variety and corresponding to a plot under v_2 we have exactly one plot under v_1 treated precisely similarly as the former in all respects save that it is under v_1; for example, the plot with the treatment $v_1r_2n_1p_0$, or rn according to the shortened symbol, corresponds to the plot to which is assigned the treatment $v_2r_2n_1p_0$ or vrn. The difference in their yields is

therefore attributable to the varietal difference alone apart from experimental error. In each replication we thus have estimates of the difference between the varieties under all the 8 possible combinations of the remaining three factors r, n and p. The average of these taken over all the replications, gives the estimate of the main response to varieties which we denote by V. Thus V is proportional* to the sum over all the replications of the difference which we may symbolically express by the following:

$$\begin{aligned} & v_2r_2n_1p_1 + v_2r_2n_1p_0 + v_2r_2n_0p_1 + v_2r_2n_0p_0 + v_2r_1n_1p_1 + v_2r_1n_1p_0 \\ & \quad + v_2r_1n_0p_1 + v_2r_1n_0p_0 \\ & \quad - v_1r_2n_1p_1 - v_1r_2n_1p_0 - v_1r_2n_0p_1 - v_1r_2n_0p_0 - v_1r_1n_1p_1 \\ & \quad - v_1r_1n_1p_0 - v_1r_1n_0p_1 - v_1r_1n_0p_0 \end{aligned}$$

or, using the shortened symbols, by

$$\begin{aligned} & vrnp + vrn + vrp + vr + vnp + vn + vp + v \\ & \quad - rnp - rn - rp - r - np - n - p - 1 \end{aligned}$$

By treating the symbols as algebraic quantities this may be written as

$$(v - 1)(r + 1)(n + 1)(p + 1)$$

Conversely, if we expand the brackets in the latter expression by multiplication we get the difference of treatment combinations corresponding to the main effect V. It will be noticed that each bracket in the above expression consists of the symbols for the two levels of a single factor. These are connected by the *minus* sign only in the bracket corresponding to the factor whose main response is to be estimated while the two levels of all other factors, over which the difference between the levels of the factor under estimation is averaged, are connected by the *plus* sign. This furnishes a clue for obtaining expressions for the other main responses. The main effect for seed-rate, symbolised by R, would, for instance, be obtained by expanding the expression $(v + 1)(r - 1)(n + 1)(p + 1)$ and carrying out over all the replications the algebraical summation of the yields

* We devide the total difference by half the total number of plots, namely, 32 to get the response per plot ; since the divisor remains the same for all treatment effects and does not alter the sums of squares, it is more convenient to deal with the totals themselves.

of the various treatment combinations in the manner indicated by the expanded expression. Similarly for the main effects N and P.

We next turn to interactions. The response to nitrogen may itself depend on the application of phosphorus. This differential response to nitrogen at the different levels of the other factor phosphorus is what we term the interaction between the two factors and denote by the symbol NP. The difference $v_2r_2n_1p_1 - v_2r_2n_0p_1$ or $vr(np - p)$, for instance, gives the response to nitrogen for the combination $v_2r_2p_1$ of the other factors; now the corresponding response at the lower level of phosphorus, given by $v_2r_2n_1p_0 - v_2r_2n_0p_0$ or $vr(n - 1)$ may not be the same, even allowing for experimental errors. The difference between them, namely,

$$[(v_2r_2n_1p_1 - v_2r_2n_0p_1) - (v_2r_2n_1p_0 - v_2r_2n_0p_0)] = v_2r_2(n_1 - n_0)(p_1 - p_0)$$

or $vr(n - 1)(p - 1)$ contributes therefore to the estimate of the differential response NP. This is the contribution at the levels v_2 and r_2 of the first two factors. We have in all four such differences per replication corresponding to the four combinations of v and r. The complete estimate of the interaction NP is obtained by averaging these four over all the replications. NP is therefore proportional (in shortened symbols) to

$$\begin{aligned}&(vrnp - vrp - vrn + vr)\\ &+ (vnp - vp - vn + v)\\ &+ (rnp - rp - rn + r)\\ &+ (np - p - n + 1)\end{aligned}$$

which may be summarised into

$$(v + 1)(r + 1)(n - 1)(p - 1)$$

It will be noticed that the symbolical expression contains *minus* sign in the two brackets corresponding to the factors whose interaction it stands for and the positive sign in the other brackets. This provides the method for obtaining the other interactions as well. The interaction between variety and seed-rate VR, for instance, may be obtained from the expression

$$(v - 1)(r - 1)(n + 1)(p + 1)$$

10*b*.4 Symbolical Representation of Second and Higher Order Interactions

So far we have dealt with only pairs of factors. The concept of interaction may also be extended to more than two factors. The magnitude of an interaction between a pair of factors, say, N and P, may itself depend on the level of a third factor such as v, and we may be interested in estimating this differential response. It is obviously proportional to the difference between the two rectangular brackets

$$[(vrnp - vrp - vrn + vr) + (vnp - vp - vn + v)]$$

and

$$[(rnp - rp - rn + r) + (np - p - n + 1)]$$

The first of these stands for the interaction NP at the level v_2 of the factor V, that is, for $v(r + 1)(n - 1)(p - 1)$ and the second for the same interaction at the level v_1 and may be expressed as $(r + 1)(n - 1)(p - 1)$. This difference reduces to

$$(v - 1)(r + 1)(n - 1)(p - 1)$$

This expression suggests what may be established by argument also, that the differential response of NP with respect to effect V is the same as the differential response of NV with respect to P or of VP with respect to N. This common effect is termed the interaction VNP. Such interactions are spoken of as three-factor or *second order interactions*; the two-factor interactions, the simplest kind, being termed as of *first order*. With the four factors in the experiment we have six first order interactions, namely, VR, VN, VP, RN, RP and NP and four second order interactions, namely, VRN, VRP, VNP and RNP. Lastly, any of these three-factor interactions may itself be conceived of as varying with the level of the remaining factor and the resulting differential response is called the four-factor or *third order interaction*. Obviously there is only one such interaction in the present experiment, namely, $VRNP$, which is proportional to

$$(v - 1)(r - 1)(n - 1)(p - 1)$$

We may extend the concept of interactions in the same manner to cover a larger number of factors. The concept is

quite a general one applicable to factors at more than two levels also. In the case of factors each at two levels, however, it is possible to express every effect, whether a main effect or an interaction involving any number of factors, explicitly as a single difference between the treatment combinations with one degree of freedom.

10c.1 Principles and Technique of Confounding

It may now be seen from the treatment combinations allocated to the two blocks of the first replicate of our experiment (Table 10.1) that the contrast between the two is represented as $(v_2 - v_1)(r_2 - r_1)(n_1 - n_0)(p_1 + p_0)$ which is clearly the three-factor interaction *VRN*. From the first replication, therefore, this interaction cannot be estimated, as it is inseparably tied or confounded with the natural soil fertility differences between the two blocks of the replicate. Similarly it may be verified that interactions *VRP*, *VNP* and *RNP* are confounded with blocks in the second, third and fourth replication respectively. If the same second order interaction, say *VRN*, were confounded with blocks in all the replications, it would not be possible to estimate it at all, as it cannot be distinguished from the natural soil fertility differences between the blocks of each replicate. It is for this reason that each second order interaction is confounded in only one replicate and it can be estimated from the remaining three. The design is said to be a *partially confounded design* and permits the estimation of the confounded components of treatment contrasts though with less precision than the unconfounded components.

The confounding of these particular effects in the layout we are considering has not been an accidental outcome, but has been deliberately arranged. It may be verified that in every replication the estimates of all the main effects and the two-factor interactions are composed of differences within blocks and are consequently unaffected by block differences. It is only the chosen three-factor interaction in each replication, for example, *VRN* in Replication I, which is affected by block differences. It is general experience that an interaction of a high order (involving three or more factors) seldom turns out to be significant. Moreover, it becomes increasingly difficult to attach any physical interpretation to the higher

order interactions. For both these reasons it would not very much matter to the conclusions to be drawn from an experiment if an interaction of a high order were to be confounded with block differences, if by so doing the estimates of other effects such as the main effects and first order interactions can be secured with a sufficiently high degree of precision. It is these latter effects which need to be estimated with the maximum precision and should, therefore, be kept free from block differences. It is easy to achieve this in the present instance, since each replication is divided into only two blocks and a single contrast confounded with blocks. All that has to be done is to choose for confounding a high order interaction, such as one of the three-factor interactions or the four-factor interaction *VRNP*. After the contrast to be confounded is selected, the symbolic expression for it may be written out in full and all treatment combinations having the positive sign in the expanded difference may be assigned to one block and those having the negative sign to the other. The choice of confounding and the manner of allotment of treatments to plots are more intricate when a replication is to be divided into more than two blocks and in the case of factors at more than two levels. Once the treatment effects to be confounded are decided upon and all the treatment combinations are divided into the appropriate groups, the groups are assigned to the blocks within a replicate at random, as are also the treatment combinations belonging to each group to the plots in the assigned block. While a fuller treatment of the subject is beyond the scope of the book, confounding in experiments with three factors, each at three levels, is discussed in the later portion of the present chapter.

10*d*. 1 Analysis of a Confounded Design of 2^n Type

Having examined the layout of the 2^4 experiment which we have been considering, we now come to the statistical analysis of the results of the experiment. The total sum of squares for 63 degrees of freedom between the 64 plot yields and the block sum of squares for 7 degrees of freedom between the 8 blocks may be calculated in the usual manner. The computation of the treatment sum of squares, however, needs consideration. Between the 16 treatments there are 15 degrees of freedom which in fact are divisible into 4 main effects (*V*, *R*, *N* and *P*), 6 first

order interactions (*VR*, etc.), 4 second order interactions (*VRN*, etc.) and one third order interaction (*VRNP*). Each effect carries a single degree of freedom and is seen to be proportional to a linear function of plot yields of the type $(c_1y_1 + c_2y_2 + \ldots)$ where the sum of all the coefficients $(c_1 + c_2 + \ldots)$ is zero. The main effect *V*, for instance, is given by the summation over the 4 replications of the following linear function: $(v - 1)(r + 1)(n + 1)(p + 1)$, that is,

$$\begin{aligned}\{&v_2r_2n_1p_1 + v_2r_2n_1p_0 + v_2r_2n_0p_1 + v_2r_2n_0p_0 + v_2r_1n_1p_1 + v_2r_1n_1p_0\\ &+ v_2r_1n_0p_1 + v_2r_1n_0p_0\\ &- v_1r_2n_1p_1 - v_1r_2n_1p_0 - v_1r_2n_0p_1 - v_1r_2n_0p_0 - v_1r_1n_1p_1\\ &- v_1r_1n_1p_0 - v_1r_1n_0p_1 - v_1r_1n_0p_0\}\end{aligned}$$

where the symbols for treatment combinations stand for the plot totals under them taken over all replications. The sum of squares for such a linear function is given by dividing the square of the function by the sum of squares of the coefficients *c* of the individual terms in the function, that is, by

$$\frac{(c_1y_1 + c_2y_2 + \ldots)^2}{(c_1^2 + c_2^2 + \ldots)}$$

where $y_1, y_2, \ldots$ represent the individual plot yields. Since in the linear function representing *V* each plot has a coefficient + 1 or − 1, the value of c^2 is 1 for every plot yield and consequently the divisor for the sum of squares reduces to merely the total number of plots in all the replications from which the effect has been computed, namely, $16 \times 4 = 64$. Symbolically therefore we may express the sum of squares for *V* for the single degree of freedom to be

$$\frac{\{(v - 1)(r + 1)(n + 1)(p + 1)\}^2}{64}$$

in which the expression within the curled brackets in the numerator is to be expanded in full and the algebraic sum of the totals over all replications of the treatment combinations occurring in the expansion is to be taken and squared. The sum of squares for all the other main effects, all the two-factor interactions as well as the four-factor interaction are computed in exactly the same

manner from all the four replications and using the same divisor, 64. Calculations for the three-factor interactions, however, have to be modified since they are confounded. The interaction *VRN*, for instance, has been confounded in the first replication; its estimate from this replication cannot be obtained since it is affected by the natural soil differences between the two blocks. We confine ourselves, therefore, to the remaining three replications in which the effect is not confounded for estimating it. The estimate of *VRN* is also a linear function of plot yields but involves only 48 plots. The sum of squares is therefore given by the ratio

$$\frac{\{(v-1)(r-1)(n-1)(p+1)\}^2}{48}$$

where, in the numerator plot yields in the three replications II, III and IV only are taken. Since each of the other three three-factor interactions is also confounded only once, their sums of squares are also calculated from similar expressions. In this manner we obtain the sum of squares for the treatment effects. Lastly, the error sum of squares is computed by subtracting the block sum of squares and treatment sum of squares from the total sum of squares as usual. It will be noted that the error sum of squares carries (63 — 7 — 15 =) 41 degrees of freedom. In a randomized block design in four replications, the error degrees of freedom would have been 45. The 4 degrees of freedom from error have been removed in the present design as arising out of the difference between the two members of each of the 4 pairs of blocks into which the replications are divided. The final analysis of variance is then as in Table 10.2.

It will be seen that only the varietal differences, the response to nitrogen and the interaction between varieties and seed-rates are significant (at 1 per cent. level).

10*d*. 2 Efficiency

The object of confounding is to ensure that the estimates of unconfounded effects are obtained with increased precision by removal of the effect of possible heterogeneity within a large replicate by its subdivision into smaller blocks. In order to see

TABLE 10.2

Analysis of variance of 2^4 experiment in four replications with partial confounding of second order interactions

(Yield in lb. per plot)

Source of variation	D.F.*	S.S.	M.S.	F
Blocks ..	7	230·64	32·95	1·31
V ..	1	1321·32	1321·32	47·72**
R ..	1	60·84	60·84	2·20
N ..	1	290·70	290·70	10·50**
P ..	1	5·06	5·06	..
VR ..	1	495·06	495·06	17·88**
VN ..	1	1·10	1·10	..
VP ..	1	11·90	11·90	..
RN ..	1	6·25	6·25	..
RP ..	1	0·00	0·00	..
NP ..	1	26·52	26·52	..
VRN ..	1′	0·46	0·46	..
VRP ..	1′	12·92	12·92	..
VNP ..	1′	2·43	2·43	..
RNP ..	1′	2·61	2·61	..
VRNP ..	1	4·20	4·20	..
Error ..	41	1135·24	27·69	..
Total ..	63	3607·25	..	..

how far we have gained in this aim by adopting the more complicated layout, we may work out a rough estimate of the error variance per plot which would have been obtained, had a simple randomized block design been adopted and compare it with the error mean square per plot of the confounded experiment. The former estimate may be obtained by segregating the sum of squares between the four replications for 3 degrees of freedom from the block sum of squares for 7 degrees of freedom and adding the remaining sum of squares for 4 degrees of freedom to the error sum of squares of the confounded experiment. The pooled error sum of squares so obtained, divided by 41 + 4 = 45 degrees of freedom gives an approximate estimate of the error mean square for the corresponding randomized block design, on the assumption that the contributions of the confounded interactions to the

* Partially confounded degrees of freedom are indicated by dashes.

CHAPTER XI

SPLIT-PLOT AND STRIP-PLOT DESIGNS

11*a*.1 The Split-Plot Design

It will be recalled that, in the varietal-*cum*-manurial trial on wheat discussed in Chapter IX, the two main effects, namely, the differences between varieties and the responses to different levels of nitrogen as well as the interaction between the two factors were estimated with the same degree of precision. Further the differences in yield between different levels were of the same order of magnitude for both factors. It is however conceivable that some factors may produce much larger differences than others. Factors like sowing dates and irrigations show usually larger differences than others like spacings. When two such dissimilar factors are tested in an experiment it is possible that, with a simple factorial design the accuracy attained would be sufficient to bring out the significance of the difference between the levels of the factor capable of showing larger differences but the experimental error might be too high to permit detection of the smaller difference between the levels of the other factor. In planning a trial with such factors it is therefore desirable to modify the layout so as to increase the precision of the latter comparisons. In Chapter VII we have seen that one way of improving the precision of comparisons is the adoption of local control, that is, restriction of randomisation in such a way that the treatments to be compared are allotted to contiguous plots in a compact block. The same device can be extended a stage further. We may restrict the randomization within blocks in such a manner that the several levels of the second factor which we wish to compare with greater precision are assigned to contiguous plots with a common level of the first factor instead of scattering them over the entire replicate. We may do so by first dividing each block into *main plots* to be assigned to the levels of the first factor at random and next subdivide each main plot into *sub-plots* to be allotted randomly to the levels of the second factor. Such a layout in which one set of treatments is assigned to large plots,

called main plots, into which a replicate is divided and another set of treatments to subdivisions of the main plots called sub-plots, is termed a *split-plot* design.

Example 11.1

In Table 11.1 are given the plan and yields of a cultural trial on sugarcane in which a split-plot layout was adopted. The treatments consisted of four dates of planting, namely, planting in October (s_1), November (s_2), February (s_3) and March (s_4), and three methods of planting, namely, planting in trenches and turning into ridges (m_1), planting on flat and turning into ridges (m_2)

TABLE 11.1

Plan and yields of a split-plot experiment on sugarcane

(Yield in md.* per sub-plot; sub-plot size 122'×8')

Replication I			Replication II		
	m_3	7·7		m_2	8·3
s_2	m_1	5·2	s_3	m_1	5·4
	m_2	2·6		m_3	8·3
	m_1	0·9		m_3	7·3
s_4	m_2	2·1	s_1	m_1	7·8
	m_3	0·4		m_2	6·4
	m_2	4·2		m_1	1·3
s_3	m_1	1·3	s_4	m_2	1·0
	m_3	1·4		m_3	1·3
	m_1	6·8		m_1	5·5
s_1	m_2	6·9	s_2	m_3	2·6
	m_3	5·3		m_2	2·1

*1 maund equals 82·3 lb.

Main plot treatments:
Dates of planting
s_1 : October
s_2 : November
s_3 : February
s_4 : March

Sub-plot treatments:
Methods of planting
m_1: Planting in trenches and turning into ridges
m_2: Planting on flat and turning into ridges
m_3: Planting on flat and leaving as such

TABLE 11.1—*Continued*

	Replication III			Replication V	
	m_3	4·4		m_1	22·1
s_2	m_2	4·0	s_1	m_2	19·0
	m_1	3·7		m_3	14·9
	m_2	4·4		m_1	6·3
s_2	m_1	3·5	s_4	m_3	3·9
	m_3	1·2		m_2	1·6
	m_2	3·1		m_3	5·3
s_1	m_3	2·9	s_3	m_1	6·3
	m_1	4·8		m_2	4·3
	m_3	0·3		m_1	4·9
s_4	m_1	0·9	s_2	m_2	4·0
	m_2	1·4		m_3	5·1
	Replication IV			Replication VI	
	m_3	4·2		m_3	1·5
s_2	m_1	7·3	s_4	m_1	2·0
	m_2	7·5		m_2	2·1
	m_1	16·9		m_2	4·9
s_1	m_3	10·0	s_1	m_3	2·2
	m_2	9·5		m_1	3·3
	m_2	2·6		m_2	2·5
s_4	m_1	4·5	s_3	m_1	2·0
	m_3	4·0		m_3	2·9
	m_1	8·8		m_3	1·3
s_3	m_2	6·1	s_2	m_2	0·9
	m_3	5·9		m_1	2·3

and planting on flat and leaving as such (m_3). The 4 levels of the first factor were allotted to 4 main plots of each block at random and the 3 levels of the second factor were randomly assigned to the 3 sub-plots into which each main plot was divided.

11*a*.2 Advantages of the Design

It is instructive to compare the split-plot layout with an arrangement in randomized blocks and to observe how, through a more effective local control, a greater similarity of conditions for the comparisons between methods of planting is secured in the split-plot design than is possible in a randomized block design. It will also be noticed that the split-plot layout makes for the greater convenience of agricultural operations, when treatments like irrigation and other cultural treatments like sowing dates, spacing, ploughing, etc., are involved, as these can be assigned to the larger or main plots. Further, the design also results in the saving of experimental area and resources devoted to the provision of border rows; for in many cases it would suffice to have wide borders between main plots only. On the other hand, it is likely that the comparisons between main plot treatments, dates of planting in the present case, would be made with much less precision than in randomized blocks as they would be obtained exclusively from differences between main plots, which are further apart from one another on the average than would have been the case in the randomized block layout.

11*a*.3 Intraclass Correlation between Sub-Plots

The difference in precision of the two sets of comparisons, those between main plots and between sub-plots, may be regarded from another point of view, namely, as the consequence of the intraclass correlation between contiguous plots. We have seen in Chapter VIII on randomized blocks how we may visualize the reduction in experimental error brought about by grouping plots in blocks as being due to the effective exploitation of the intraclass correlation between contiguous plots within blocks. In the same way, the greater correlation between neighbouring sub-plots within main plots may be said to render the comparisons of sub-plot treatments more accurate than in the randomized block design. It can be seen from formula (6.35) of Chapter VI that the standard error for the comparisons between sub-plots will be

$$\sqrt{\sigma^2(1-\rho')} \qquad (11.1)$$

per plot instead of σ where σ^2 is the experimental error variance per plot in the corresponding randomized block design and ρ'

is the intraclass correlation between the sub-plots belonging to the same main plot apart from any correlation existing between plots in the block as a whole which is already reflected in σ^2. This correlation is usually positive and hence gives greater precision to comparisons between sub-plots. The comparison between main plot treatments, on the other hand, will be much less accurate than in the randomized block design, the standard error per plot on a sub-plot basis for such comparisons being

$$\sqrt{\sigma^2 [1 + (k - 1)\rho']} \qquad (11.2)$$

if each main plot is composed of k sub-plots. The interactions of the sub-plot treatments with the main plot treatments are estimated from differences within main plots; they are therefore estimated as accurately as the effect of the sub-plot treatments. The analysis of the split-plot experiment thus virtually amounts to the estimation of σ^2 and ρ' from which the standard errors for the different comparisons may be calculated. We can proceed to do this systematically by first carrying out the analysis of variance.

11*a*.4 Analysis of Variance

The sum of squares and the mean square for the various treatment effects as well as replications are calculated in the usual way. It should be noted however that the entire analysis is carried out on the basis of the sub-plot, the divisor for each sum of squares depending on the number of sub-plots of which each of the items which are to be squared is composed; for example, for dates of planting the divisor is 18 and not 6. We have to obtain in addition separate estimates of error variance for main plots as well as sub-plots. We have thus one mean square for the main plot error and another for the sub-plot error. The sum of squares for main plot error is obtained from the analysis of the main plot yields by subtracting from the total sum of squares for the two-way table of blocks and planting dates, the sum of squares for blocks and the sum of squares for planting dates. The calculations are shown on the next page.

TABLE 11.2

Two-way table of blocks and dates of planting in split-plot design

(Each figure is a total of 3 sub-plots)

Blocks	Dates of planting				*Total*
	s_1	s_2	s_3	s_4	
I ..	19·0	15·5	6·9	3·4	44·8
II ..	21·5	10·2	22·0	3·6	57·3
III ..	10·8	9·1	12·1	2·6	34·6
IV ..	36·4	19·0	20·8	11·1	87·3
V ..	56·0	14·0	15·9	11·8	97·7
VI ..	10·4	4·5	7·4	5·6	27·9
Total ..	154·1	72·3	85·1	38·1	349·6

$$C.F. = \frac{(349 \cdot 6)^2}{72} = 1697 \cdot 50$$

$$\text{Total } S.S. \text{ (23 d.f.)} = \frac{(19 \cdot 0)^2 + \ldots . + (5 \cdot 6)^2}{3} - 1697 \cdot 50 = 1054 \cdot 73$$

$$S.S. \text{ for Blocks (5 d.f.)} = \frac{(44 \cdot 8)^2 + \ldots . + (27 \cdot 9)^2}{12} - 1697 \cdot 50 = 338 \cdot 54$$

$$S.S. \text{ for Planting dates (3 d.f.)} = \frac{(154 \cdot 1)^2 + \ldots . + (38 \cdot 1)^2}{18} - 1697 \cdot 50 = 395 \cdot 15$$

S.S. for Main Plot Error,

[Error (*a*)] (23 − 5 − 3 = 15 d.f.) = 321·04

= (1054·73) − (338·54) − (395·15).

TABLE 11.3

Analysis of variance of main plot yields

(Unit : md. per sub-plot)

Source of variation	*D.F.*	*S.S.*	*M.S.*	*F*
Blocks ..	5	338·54	67·71	3·16
Sowing dates ..	3	395·15	131·72	6·16**
Error (a) ..	15	321·04	21·40	..
Total ..	23	1054·73	..	..

Next we turn to sub-plots. From the two-way table of planting dates and methods of planting (Table 11.4) the sum of squares for the main effect M and the interaction SM are calculated in the usual manner.

TABLE 11.4

Two-way table of treatment totals for split-plot experiment

(Totals of 6 plots each in md.)

Methods of planting	Dates of planting				
	s_1	s_2	s_3	s_4	*Total*
m_1	61·7	28·7	27·5	15·9	133·8
m_2	49·8	21·5	29·4	10·8	111·5
m_3	42·6	22·1	28·2	11·4	104·3
Total ..	154·1	72·3	85·1	38·1	349·6

We have

$$S.S. \text{ for methods of planting (2 d.f.)} = 19{\cdot}72$$

and

$$S.S. \text{ for interaction } SM \text{ (6 d.f.)} = 19{\cdot}51$$

Also from the entire table of all the 72 plot yields,

$$\text{Total } S.S. \text{ (71 d.f.)} = 1191{\cdot}58$$

The sum of squares for sub-plot error is obtained by subtracting from the total sum of squares the sum of squares for blocks and all the treatment effects as also the sum of squares for main plot error.

Sum of squares for sub-plot error, or Error (b) = Total sum of squares − Block sum of squares − Treatment sum of squares − Main plot error $= 1191{\cdot}58 - 338{\cdot}54 - 395{\cdot}15 - 19{\cdot}72 - 19{\cdot}51 - 321{\cdot}04 = 97{\cdot}62$.

We have then the following table of complete analysis of variance:

TABLE 11.5

Complete analysis of variance

(Unit : md. per sub-plot)

Source of variation	D.F.	S.S.	M.S.	F
Blocks ..	5	338·54	67·71	3·16
Planting dates (*S*) ..	3	395·15	131·72	6·16**
Error (*a*) ..	15	321·04	21·40	..
Methods of planting (*M*)	2	19·72	9·86	4·04*
Interaction Planting dates × methods (*SM*)	6	19·51	3·25	1·33
Error (*b*) ..	40	97·62	2·44	..
Total ..	71	1191·58	..	..

The mean square between planting dates is compared against mean square for the main plot error [Error (*a*)] for significance and those for the main response to methods of planting and the interaction, planting dates × methods of planting, being effects belonging to sub-plots, are tested against the mean square for the sub-plot error [Error (*b*)]. It is seen that both the main effects are significant, the differences between planting dates being significant at 1 per cent. level. We notice too from Table 11.4 that the differences between methods of planting are of a much smaller order than those between dates of planting.

11*a*.5 EFFICIENCY

The estimates of the two error variances as seen from Table 11.5 are 21·40 and 2·44 respectively, thus confirming our expectation that the main plot error is likely to be much larger than the sub-plot error. For comparing the efficiency of the split-plot design with the randomized block design, we may equate the two error means squares to their expected values. We have

$$\sigma^2 (1 + 2\rho') = 21 \cdot 40$$

and

$$\sigma^2 (1 - \rho') = 2 \cdot 44$$

from which we get the estimate of the correlation between contiguous sub-plots to be + 0·72 and that of the hypothetical

variance per plot of the corresponding randomized block design, to be 8·71. The efficiency of the sub-plot comparisons may therefore be measured by the ratio $8 \cdot 71/2 \cdot 44 = 3 \cdot 57$ which is equivalent to a gain of 257 per cent. The gain in the present example is more than is usually expected; but it nevertheless illustrates the increased efficiency of the sub-plot comparisons in a split-plot experiment. The efficiency of the split-plot design with regard to the main-plot comparisons, on the other hand, turns out to be $8 \cdot 71/21 \cdot 40$, or only 41 per cent. as compared with the randomized block design. It is also instructive to note that the test for the significance of the main plot treatment differences is not likely to be as sensitive as that for the differences among sub-plot treatments for the additional reason that the degrees of freedom for the main plot error are not likely to be at all as numerous as for the sub-plot error and may not usually be as ample as in the present instance. For both these reasons, the treatments assigned to the main plots are not likely to be compared with much precision. This limitation has to be borne in mind when employing a split-plot design. When, as in the present instance, one of the treatment factors is expected to show large differences not requiring a very sensitive test for detection, or is desired to be studied not so much for its own sake as for its possible interactions with the other factors, it would be most appropriate to adopt a split-plot design and place that treatment in the main plots. It will be noted from the preceding analysis of variance that the interaction of the main plot and sub-plot treatments is tested against the same error mean square as the sub-plot treatment responses themselves.

11*a*.6 Standard Errors of Comparisons

It remains now to estimate the standard errors of various treatment comparisons. In the first place, the standard error of the average difference between two levels of the main plot treatment, namely, dates of planting (averaged over all the levels of the sub-plot treatment) is estimated as

$$\sqrt{2s_a^2/rk} \qquad (11.3)$$

where s_a^2 is the mean square for main plot eror [Error (a)]$=21 \cdot 40$ in our case, r is the number of replications, 6 in the present example, and k is the number of sub-plots per main plot, namely, 3. The

value of the standard error for the difference of two main plot treatments in the present example is therefore 1·54 md./plot. Similarly, the standard error of the difference between any two levels of sub-plot treatment depends solely on the sub-plot error mean square s_b^2 [Error (b)] = 2·44 in the present case, and is given by

$$\sqrt{2s_b^2/rm} \qquad (11.4)$$

when averaged over all the levels of the main plot treatment, there being m such levels in all. This works out to 0·45 md./plot in the illustrative example. On the other hand, if two sub-plot treatments are compared at a fixed level of the main plot treatment, the standard error is given by the same formula but with $m = 1$. It will be noticed that the denominator whether we want to compare levels of a main plot or a sub-plot treatment is equal to the number of sub-plots on which is based each of the two mean values which are being compared.

For comparisons arising as a combination of comparisons between sub-plots within main plots, the standard error would involve only the mean square for sub-plot error. As an example, suppose we wish to compare planting in trenches with planting on flat (irrespective of turning into ridges or leaving as such) for October and November plantings. We have to test for this purpose the difference between m_1 and the average of m_2 and m_3, both quantities being themselves averaged over s_1 and s_2. Symbolically the comparison is expressible as

$$\tfrac{1}{2}(s_1 + s_2)\left(m_1 - \frac{m_2 + m_3}{2}\right) \qquad (11.5)$$

Now

$$\left(m_1 - \frac{m_2 + m_3}{2}\right)$$

can be put in the form of a combination of comparisons within main plots as

$$\tfrac{1}{2}[(m_1 - m_2) + (m_1 - m_3)]$$

The variance of the contrast represented by (11.5) would therefore involve s_b^2 only and may be calculated by an application of the formula already given, namely, that the variance of a linear

function of independent plot yields of the type $c_1y_1 + c_2y_2 + \ldots$ is given by

$$V(y) \times (c_1^2 + c_2^2 + \ldots) \tag{11.6}$$

Expanding the brackets, the contrast presented in (11.5) can be written out as

$$\begin{aligned}(\tfrac{1}{2})\, s_1m_1 + (-\tfrac{1}{4})\, s_1m_2 + (-\tfrac{1}{4})\, s_1m_3 + (\tfrac{1}{2})\, s_2m_1 \\ + (-\tfrac{1}{4})\, s_2m_2 + (-\tfrac{1}{4})\, s_2m_3\end{aligned} \tag{11.7}$$

where s_1m_1 stands for the mean yield over r replications of the treatment combination s_1m_1, etc. We notice that each such mean of r plot yields in the above expansion is entirely independent of others in the sense that it has no plot yields in common with any other mean. We can accordingly apply the formula for variance, summing up the squares of the coefficients and taking s_b^2/r as the estimate of the variance of a treatment mean. The variance is therefore equal to

$$\frac{s_b^2}{r}\left[\frac{1}{4} + \frac{1}{16} + \frac{1}{16} + \frac{1}{4} + \frac{1}{16} + \frac{1}{16}\right] = \frac{3s_b^2}{4r}$$

Whenever the contrast is expressed symbolically as a product of bracketed expressions as above, it is not even necessary to expand the brackets, to compute the estimate of variance. The detailed procedure given above reduces to the following simple rule: Square the coefficients of the symbols for levels in each bracket, add them up for each bracket and take the product. The resulting product multiplied by s_b^2/r gives the required variance. Applying this method, we find the variance of the contrast

$$\left(\frac{s_1 + s_2}{2}\right)\left(m_1 - \frac{m_2 + m_3}{2}\right)$$

to be

$$\frac{s_b^2}{r} \times [\{(\tfrac{1}{2})^2 + (\tfrac{1}{2})^2\}\ \{(1)^2 + (-\tfrac{1}{2})^2 + (-\tfrac{1}{2})^2\}] = \frac{3s_b^2}{4r}$$

as before. The standard error is consequently

$$\sqrt{\frac{3\,(2{\cdot}44)}{4 \times 6}} \text{ or } 0{\cdot}55 \text{ md. per sub-plot}$$

The standard error of a difference between the levels of main plot treatment, on the other hand, except when it is averaged over all the sub-plot levels, involves both the error mean squares s_a^2 and s_b^2. In a split-plot experiment in r replications with m levels of main plot treatment and k of sub-plot treatment, the standard error of a difference between a pair of levels of main plot treatment, averaged over k' of the sub-plot levels is given by the square-root of

$$\frac{2}{k.k'.r}\{k'.s_a^2 + (k - k')s_b^2\} \qquad (11.8)$$

Thus if we wish to compare October planting with November planting for the method of planting in trenches only, we have $k = 3$, $k' = 1$, $r = 6$, $s_a^2 = 21{\cdot}40$, $s_b^2 = 2{\cdot}44$, whence the standard error of the difference to be tested is the square-root of

$$\frac{2}{(3)(1)(6)}\{(1)(21{\cdot}40) + (3 - 1)(2{\cdot}44)\} \text{ or } 1{\cdot}71 \text{ md. per sub-plot.}$$

Since the standard error involves estimates of more than one single variance, the t test is not strictly applicable for testing the significance of the comparison. Nevertheless the test may serve as a good approximation to the criterion for significance.

11*a*.7 Extension of the Split-Plot Design

If necessary, the split-plot design could be extended to a further stage by splitting the sub-plots into what may be called *second order sub-plots* to be assigned at random to a further set of treatments. The analysis would follow the same lines as before with the additional estimations of the error mean square for the second order sub-plots. The main plot treatment differences would be compared against the main plot error, the main effect of the sub-plot treatments and their interactions with main plot treatments would be tested against the sub-plot error and the main effect of ultimate plot treatments and their interaction with all other factors would be tested against the ultimate plot error. These last mentioned effects would be estimated with the greatest precision as a result of most effective local control, since the second order sub-plots would be compared under more

homogeneous conditions than the first order sub-plots and the main plots.

11*b*.1 The Strip-Plot Design

We shall now consider a design, analogous to the split-plot arrangement, in which two different sets of treatments can be tried in large plots with one set of plots superposed over the other set at right angles. Such an arrangement may sometimes be convenient in a cultural experiment involving, for instance, factors like spacing and ploughing, where the use of small plots by splitting larger plots is not feasible. A block may be divided into strips in one direction to be allotted to one set of treatments, say different spacings, and into another set of strips, in a direction at right angles to the first, to be allotted to the second set of treatments, say ploughing. The arrangement is called a *strip-plot design*. The plots formed by the intersection of the strips may be further split or the entire primary strips belonging to one set may themselves be divided into further narrower strips for accommodating a yet further set of treatments. The allotment of the treatments to the strips or plots at each stage has of course to be made at random.

As a simple illustration of a strip-plot experiment we may consider the following example of an experiment carried out at a paddy research station in West Bengal.

Example 11.2

There were three dates of planting and three green manuring treatments, namely

Treatments	
Dates of transplanting	Green manuring
d_1: July 16th	g_1: Manuring with Sunn hemp
d_2: August 16th	g_2: Manuring with Dhaincha
d_3: September 16th	g_3: No manuring (control)

Obviously it made for convenience to have strips for both the factors. The plan and yields are given in Table 11.6.

TABLE 11.6

Plan and yield of strip-plot experiment

(Yield in oz. per plot ; plot of size 20' × 20')

	Replication I				Replication II		
	d_1	d_3	d_2		d_2	d_1	d_3
g_2	376	455	480	g_1	549	396	492
g_1	386	476	496	g_3	533	388	482
g_3	355	433	446	g_2	540	406	512

	Replication III				Replication IV		
	d_2	d_1	d_3		d_2	d_3	d_1
g_2	500	347	468	g_3	413	334	201
g_3	482	337	435	g_1	469	436	298
g_1	513	387	476	g_2	436	398	280

	Replication V				Replication VI		
	d_3	d_1	d_2		d_2	d_3	d_1
g_1	458	366	474	g_3	490	447	348
g_3	413	333	425	g_2	509	473	356
g_2	434	356	465	g_1	520	487	397

The statistical analysis is explained in the following section.

11*b*.2 STATISTICAL ANALYSIS

We have already seen in the case of the split-plot experiment that corresponding to each plot-size we have to estimate a separate error variance. This principle is applicable to the present design as well. It will be seen that in the present example, three different plot sizes are involved: the dates of transplanting treatments have been allotted to plots of one size, namely, the column strips; the manuring treatments have been assigned to plots of a second size, namely, the row strips and lastly the comparisons of the different combinations of the two treatments or the interaction comparisons have to be made from plots of a yet third size formed by the intersections of the two sets of strips. It

ws then that three different error mean squares have to be puted. The method of calculation is however the usual one, ely, that the error mean square corresponding to each effect ılculated as the interaction of that effect with replication.

First we prepare the three two-way tables for blocks × dates, ks × manuring and dates × manuring as shown in Table 11.7.

TABLE 11.7

Two-way tables of plot yields for strip-plot experiment

a) *Blocks* × *Dates of transplanting* (each figure is a total of 3 plots)

Block	d_1	d_2	d_3	*Total*
I	1117	1422	1364	3903
II	1190	1622	1486	4298
III	1071	1495	1379	3945
IV	779	1318	1168	3265
V	1055	1364	1305	3724
VI	1101	1519	1407	4027
Total ..	6313	8740	8109	23162

(*b*) *Blocks* × *Manuring* (each figure is a total of 3 plots)

Block	g_1	g_2	g_3	*Total*
I	1358	1311	1234	3903
II	1437	1458	1403	4298
III	1376	1315	1254	3945
IV	1203	1114	948	3265
V	1298	1255	1171	3724
VI	1404	1338	1285	4027
Total ..	8076	7791	7295	23162

(*c*) *Dates of transplanting* × *Manuring* (each figure is a total of 6 plots)

Manuring	d_1	d_2	d_3	*Total*
g_1	2230	3021	2825	8076
g_2	2121	2930	2740	7791
g_3	1962	2789	2544	7295
Total ..	6313	8740	8109	23162

From Table 11.7 (*a*), we obtain

$C.F. = 9934782 \cdot 30$

Corrected Total *S.S.* (17 d.f.) $= 10184367 \cdot 33 - C.F.$

$= 249585 \cdot 03$ (1)

S.S. between blocks (5 d.f.) $= 10001596 \cdot 44 - C.F.$

$= 66814 \cdot 14$ (2)

S.S. between dates (2 d.f.) $= 10110969 \cdot 44 - C.F.$

$= 176187 \cdot 14$ (3)

Whence *S.S.* for error (*a*) for date-strips (10 d.f.)

$= (1) - (2) - (3)$

$= 6583 \cdot 75$ (4)

From Table 11.7 (*b*), we obtain

Corrected Total *S.S.* (17 d.f.) $= 10023521 \cdot 33 - C.F.$

$= 88739 \cdot 03$ (5)

S.S. between blocks (5 d.f.) $= 10001596 \cdot 44 - C.F.$

$= 66814 \cdot 14$

S.S. between manuring (2 d.f.)

$= 9952137 \cdot 89 - C.F.$

$= 17355 \cdot 59$ (6)

Whence *S.S.* for error (*b*) for manuring-strips (10 d.f.)

$= (5) - (2) - (6)$

$= 4569 \cdot 30$ (7)

From Table 11.7 (*c*) we have

Total *S.S.* (8 d.f.) $= 10128501 \cdot 33 - C.F.$

$= 193719 \cdot 03$ (8)

S.S. for dates (2 d.f.) $= 10110969 \cdot 44 - C.F.$

$= 176187 \cdot 14$

S.S. for manuring (2 d.f.) $= 9952137 \cdot 89 - C.F.$

$= 17355 \cdot 59$

Whence *S.S.* for interaction dates × manuring (4 d.f.)

$$= (8)-(3)-(6)$$

$$= 176{\cdot}30 \qquad (9)$$

Finally, from the table of individual plot yields

Total *S.S.* (53 d.f.) $= 10208522{\cdot}00-C.F.$

$$= 273739{\cdot}70 \qquad (10)$$

$\therefore$ *S.S.* for error (*c*) for plots corresponding to the interaction dates × manuring (20 d.f.)

$$= (10)-(2)-(3)-(4)-(6)-(7)-(9)$$

$$= 2053{\cdot}48 \qquad (11)$$

Hence we have the following analysis of variance table:

TABLE 11.8

Analysis of variance of strip-plot experiment

(Unit: oz. per plot)

Source of variation	*D.F.*	*S.S.*	*M.S.*	*F*
Blocks	5	66814·14	13362·83	..
Dates of transplanting ..	2	176187·14	88093·57	133·80**
Error (*a*)	10	6583·75	658·38	..
Manuring	2	17355·59	8677·80	18·99**
Error (*b*)	10	4569·30	456·93	..
Dates × manuring	4	176·30	44·08	..
Error (*c*)	20	2053·48	102·67	..
Total ..	53	273739·70	..	..

The mean square for dates of transplanting is tested against error (*a*) for significance; that for manuring is tested against error (*b*) and that for interaction against error (*c*).

It will be noticed that the three errors are of different magnitudes, that for the smaller plot size being lower than the other two. This is usually to be expected. The test of significance for interaction is more sensitive also in virtue of the large number of degrees of freedom available for the corresponding error.

11*b*.3 Standard Errors of Comparisons

The standard errors of the various comparisons are computed in much the same manner as in the case of the split-plot design. Thus for any comparison between dates of transplanting averaged over all the levels of manuring the error (*a*) alone will be used; for example, the standard error for the difference of the average yields under two dates of transplanting

$$= \sqrt{\frac{2s_a^2}{6\times 3}}$$

$$= 8\cdot 6 \text{ oz. per plot}$$

since each average involves $6\times 3 = 18$ plot yields.

Similarly the standard error for any comparisons between different kinds of manuring averaged over all dates of transplanting involves only error (*b*). There are yet other comparisons belonging to the interaction between the factors in the two strips which involve the error mean square s_c^2 alone. The difference between the responses to green manuring with Dhaincha at the first two dates of transplanting for instance, which is symbolically expressible as $(d_1 - d_2)\ (g_2 - g_3)$ is composed of differences within date strips as well as differences within manuring strips. Consequently the variance of such a comparison will involve error (*c*) alone. The actual expression for the variance of the contrast is obtained by applying the same rule as in the split-plot experiment (*cf.* Section 11*a*.6) and is in the present case given by

$$\frac{s_e^2}{r} \times \left[\{(+1)^2 + (-1)^2\}\{(+1)^2 + (-1)^2\}\right] = \frac{4s_e^2}{r} = 68\cdot 45$$

The standard error of the comparison is therefore 8·3 oz. per plot.

CHAPTER XII

ANALYSIS OF COVARIANCE AND MISSING PLOT TECHNIQUE

12*a*.1 Use of Ancillary Information

In the last few chapters a number of designs for field experiments has been considered. The object of these designs is to permit treatment comparisons to be made with the greatest possible precision. This has been seen to be achieved in the various designs through the application in various degrees of the principle of local control. Correspondingly in the statistical analysis of the results, the variation under control, such as between blocks in a randomized block design, is segregated from the residual error variation. Reduction of experimental error through local control is thus duly reflected in the statistical analysis. There is, however, another method by which the error affecting the treatment comparisons may be lowered. Some of the variation from extraneous sources which is not controlled and contributes to the experimental error may yet be measurable. Measurements can be made of such characteristics in each plot in addition to its yield and these ancillary data utilised for making an allowance for the effect of these factors on the estimates of treatment differences on the one hand and for the contribution made by them to the estimated experimental error on the other. The precision of the estimates of the treatment averages and their comparisons is thereby increased. This form of control over the experimental error is spoken of as the *statistical control of error.*

One of the most useful forms of ancillary observations for improving precision through statistical control is provided by the data of a uniformity trial conducted over the experimental plots previously to the experimental trial proper. For each plot we then have two observations: (i) y the experimental yield, which is the resultant of the effect of both the experimental treatment given to the plot and the various extraneous factors giving rise to the experimental error, and (ii) x the uniformity trial yield, which,

in the absence of any treatment differences, can be taken as an index of variation from plot to plot due to factors such as soil heterogeneity which affect alike the uniformity trial and experimental sets of yields. Whatever the degree of the local control of error that may be achieved through the adoption of a suitable layout, the residual error is bound to contain a contribution from the plot to plot variation within blocks into which the layout is divided. The ancillary observations x can be utilised in estimating this contribution to the extent to which common factors affect the two sets of yield and in making appropriate adjustments to reduce the error affecting the treatment comparisons. Uniformity trial data for making such adjustments can be conveniently secured in horticultural and plantation crops. We may consider another situation in which the statistical control of error is more commonly possible and should be made. In plant breeding trials, for example, although the same number of seeds are sown per plot the number of plants in each plot will not be the same at the time of harvest owing to the failure of some of the seeds to germinate or subsequent deaths of plants. These losses vary from plot to plot even within the same progeny or variety. Such variation in plant numbers will necessarily make its contribution to the experimental error affecting the estimates of yield of different progenies or varieties. This effect can be allowed for statistically by recording the plant number or stand in addition to the total yield for each plot in the manner illustrated in Example 12.1.

Example 12.1

A varietal trial on cotton was conducted at the Institute of Plant Industry at Indore in 1944. The trial was laid out in 4 blocks. There were 16 varieties. On each plot were recorded the stand and the yield. Table 12.1 gives the data.

TABLE 12.1

Plot yields and plant numbers in cotton varietal trial

Randomized Blocks, Plot size 1/130 *acre*
(Upper figures denote yield of seed cotton per plot in oz.
Lower figures denote final stand per plot)

Variety	Block				*Total*
	I	II	III	IV	
(1)	14·0 214	7·0 238	17·0 240	4·0 137	42·0 829
(2)	17·0 396	11·5 205	8·0 349	11·0 165	47·5 1115
(3)	15·0 439	8·0 232	8·0 285	1·5 107	32·5 1063
(4)	16·0 282	7·0 213	17·0 357	8·0 194	48·0 1046
(5)	12·0 226	9·0 142	12·0 248	8·0 147	41·0 763
(6)	15·0 353	5·0 125	17·0 312	9·0 220	46·0 1010
(7)	15·0 375	3·0 97	16·0 279	8·0 189	42·0 940
(8)	14·0 279	12·0 353	16·0 329	7·0 248	49·0 1209
(9)	14·0 329	7·0 201	13·0 391	10·0 244	44·0 1165
(10)	15·0 259	4·0 141	9·0 281	3·0 146	31·0 827
(11)	17·0 387	6·0 186	15·0 280	8·0 187	46·0 1040
(12)	18·5 296	10·0 178	10·5 452	11·0 124	50·0 1050
(13)	17·0 279	4·0 88	9·0 162	8·0 171	38·0 700
(14)	14·0 373	3·0 210	10·0 223	6·0 243	33·0 1049
(15)	12·5 259	6·5 136	18·0 382	4·0 134	41·0 911
(16)	16·0 430	9·0 279	9·0 391	4·5 243	38·5 1343
Total Yield ..	242·0	112·0	204·5	111·0	669·5
Total Stand ..	5176	3024	4961	2899	16060

We shall first carry out the analysis of variance of the yield data (y) in the usual manner to compare the yields of the varieties without any adjustment for variation in plant numbers. The results of this analysis are given in Table 12.2.

TABLE 12.2

Analysis of variance of plot yields in cotton varietal trial (oz./plot)

Source of variation	*D.F.*	*S.S.*	*M.S.*	*F*
Blocks ..	3	824·5	274·83	..
Varieties ..	15	134·3	8·95	1·12
Error	45	359·3	7·98	..
Total ..	63	1318·1	..	..

It is seen from Table 12.2 that the varietal differences in yield are not significant. It will be noticed from Table 12.1, however, that the plant numbers in different plots have differed rather widely and this variation might have masked the significance of varietal differences. It is this influence of the variate x called the *concomitant* or *ancillary* variate (plant number) on the main variate y (yield) in contributing partly to the experimental error which we wish to assess and allow for through statistical analysis. One way of doing this which suggests itself immediately is to carry out the statistical analysis not on the plot yields but on the values of average yield per plant obtained for each plot by dividing the plot yield by the plant number. The yield is, however, not likely to be proportional to the number of plants per plot and such an adjustment might exaggerate the yield rate for plots with fewer plants. The more logical procedure is to correct the yields on the basis of the relation between the yield and plant number indicated by the data themselves, as revealed by the regression of y on x. The systematic procedure for adjustment of the data is known as the *analysis of covariance* which we shall proceed to illustrate with the help of the data given above.

12*a*.2. Computations for the Analysis of Covariance

The first step in the analysis consists in building up a table of sums of squares of deviations from the means of the variates x and y, namely, $S.S.\,(x^2)$ and $S.S.\,(y^2)$, as well as the sum of products of deviations from the means of x and y, $S.P.\,(xy)$, in which the total sum of squares or sum of products is partitioned into the components corresponding to blocks, varieties and error respectively. For the sum of squares of x and y this is done in the usual manner of an analysis of variance table. The method of obtaining the total sum of products (xy) and its components is the same as in the case of the sum of squares, except that wherever the single variate is squared, say (x^2), the product of the corresponding x and y observations is to be taken. Thus,

$$\text{Correction Factor} = \frac{\Sigma(x) \times \Sigma(y)}{rt}$$

where rt = total number of plots,

$$= \frac{(16060) \times (669 \cdot 5)}{64}$$

$$= 168002 \cdot 6$$

$$\text{Total } S.P.\,(xy) \text{ (for 63 } d.f.) = (214)(14 \cdot 0) + \ldots\ldots + (243)(4 \cdot 5) - C.F.$$

$$= 186579 \cdot 0 - 168002 \cdot 6$$

$$= 18576 \cdot 4$$

$S.P.\,(xy)$ between blocks (for 3 $d.f.$)

$$= \frac{(5176)(242 \cdot 0) + \ldots\ldots + (2899)(111 \cdot 0)}{16} - C.F.$$

$$= 182974 \cdot 6 - 168002 \cdot 6$$

$$= 14972 \cdot 0$$

S.P. (xy) between varieties (for 15 *d.f.*)

$$= \frac{(829)(42\cdot0) + \ldots + (1343)(38\cdot5)}{4} - C.F.$$

$$= 169127\cdot6 - 168002\cdot6$$

$$= 1125\cdot0$$

S.P. (xy) for error (for 45 *d.f.*) $= 18576\cdot4 - 14972\cdot0 - 1125\cdot0$

$$= 2479\cdot4$$

Accordingly we obtain the following table:

TABLE 12.3

Sums of squares and products

Source of variation	*D.F.*	*S.S.* (x^2)	*S.P.* (xy)	*S.S.* (y^2)
Blocks	3	279398·4	14972·0	824·5
Varieties	15	107775·3	1125·0	134·3
Error	45	156982·1	2479·4	359·3
Total ..	63	544155·8	18576·4	1318·1

12*a*.3 ADJUSTMENT FOR REGRESSION AND TEST OF SIGNIFICANCE

We now proceed to adjust the plot yields (y) for the influence of the plant number (x) in so far as its random variation from plot to plot is concerned. It is therefore the relation as reflected in the covariance between y and x in the 'Error' line which must be employed for the adjustment. The relation, in so far as it is linear, is expressed by the linear regression coefficient b of y on x. This has a value equal to 2479·4/156982·1 or 0·015794, and carries a single degree of freedom with a corresponding sum of squares given by

$$\frac{[S.P.\ (xy)]^2}{S.S.\ (x^2)} \qquad (12.1)$$

or 39·2. By removing this quantity from the error sum of squares for y we obtain the sum of squares for residual random variation in y called the *adjusted error sum of squares*. The corresponding degrees of freedom for error are also reduced by one. In the present instance; the adjusted error sum of squares has a value 359·3 − 39·2 = 320·1 with 44 degrees of freedom.

Next, the estimate of the variation between varieties has to be adjusted for the influence of the variation in stand. The derivation of the appropriate adjustment is beyond the scope of the book, but the arithmetical procedure to be followed is simple and may be explained with the help of the example under consideration. The first step in the procedure is to add up separately the sum of squares for x, the sum of products for x and y and the sum of squares for y for the 'Varieties' line and the 'Error' line in Table 12.3 to obtain the corresponding quantities for 'Varieties + Error' as in the following table:

Source of variation	*D.F.*	*S.S.* (x^2)	*S.P.* (xy)	*S.S.* (y^2)
Varieties	15	107775·3	1125·0	134·3
Error (unadjusted) ..	45	156982·1	2479·4	359·3
Varieties + Error ..	60	264757·4	3604·4	493·6

The 'Varieties + Error' sum of squares for y is then adjusted in the same manner as the 'Error' sum of squares for y by subtracting from it the quantity $[S.P.\ (xy)]^2/S.S.\ (x^2)$ from the same line, that is, $(3604{\cdot}4)^2/264757{\cdot}4 = 49{\cdot}1$ so that the adjusted sum of squares for y for 'Varieties + Error' = 493·6 − 49·1 = 444·5. The treatment sum of squares for y corrected for the regression of y on x is obtained simply as the difference between the corrected values of the sum of squares for y for the 'Varieties + Error' and 'Error' lines. It has the value 444·5 − 320·1 = 124·4 in the present case. It continues to carry 15 degrees of freedom since the degrees of freedom for the corrected sum of squares for y for 'Varieties + Error' are also reduced by 1 to allow for the adjustment on the basis of linear regression calculated from the same line. The analysis of the residual variation is carried out in the usual way and the familiar F test employed to

compare the adjusted mean square for varieties against the adjusted mean square for error in order to test the significance of varietal differences as a whole. The procedure is given in Table 12.4.

TABLE 12.4

Adjusted analysis of variance

Source of variation	*D.F.*	*S.S.*	*M.S.*	*F*
Varieties ..	15	124·4	8·30	1·14
Error	44	320·1	7·28	..
Varieties + Error ..	59	444·5	..	..

It should be noted that in applying the technique of analysis of covariance to factorial experiments the procedure of adjustment of treatment sum of squares has to be carried out separately with each component treatment effect such as a main effect or an interaction.

We notice from Table 12.4 that there has been no appreciable reduction in the error variance and the varietal yield differences remain non-significant even after allowing for the variation in plant number from plot to plot.

12*a*.4 INDIVIDUAL COMPARISONS

If the differences between yields had turned out significant, one would have proceeded to test the individual differences. Although such individual comparisons are clearly not warranted in the present example, we shall give the procedure for the sake of illustration. First of all, the varietal mean yields have to be adjusted for the regression of yield on stand. This is done by correcting all the mean yields to a standard value of the plant number, namely, the general mean $\bar{x}$ of plant number over all plots in the experiment. If $\bar{x}_i$ and $\bar{y}_i$ are the observed mean values of stand and plot yield respectively of the i-th variety, b the regression coefficient of y on x as estimated from the error line of Table 12.3, the adjusted mean yield for the i-th variety is given by

$$\bar{Y}_i = \bar{y}_i - b(\bar{x}_i - \bar{x}) \qquad (12.2)$$

and reduces in the present case to

$$\bar{Y}_i = \bar{y}_i - 0{\cdot}01579\,(\bar{x}_i - 250{\cdot}94)$$

The values of the average stand as well as the unadjusted and adjusted average yields for the 16 varieties are given in Table 12.5.

TABLE 12.5

Adjusted and unadjusted average yields in oz./plot and mean plant number/plot

Variety number	Average plant number	Average yield	
		Unadjusted	Adjusted
1	207·25	10·50	11·19
2	278·75	11·87	11·43
3	265·75	8·12	7·89
4	261·50	12·00	11·83
5	190·75	10·25	11·20
6	252·50	11·50	11·48
7	235·00	10·50	10·75
8	302·25	12·25	11·44
9	291·25	11·00	10·36
10	206·75	7·75	8·45
11	260·00	11·50	11·36
12	262·50	12·50	12·32
13	175·00	9·50	10·70
14	262·25	8·25	8·07
15	227·75	10·25	10·61
16	335·75	9·62	8·28

We have next to find the standard error of the difference of a pair of adjusted varietal means. It is to be expected that since the different varietal mean yields are adjusted to different extents depending on the varietal stand, the standard error will vary with the pair of varieties compared and will involve the

average values of the stand of the pair of varieties concerned. The expression for the variance of the difference of a pair of varieties i and j, with mean plant numbers $\bar{x}_i$ and $\bar{x}_j$ is given by

$$V(\bar{Y}_i - \bar{Y}_j) = s'^2 \left\{ \frac{2}{r} + \frac{(\bar{x}_i - \bar{x}_j)^2}{A} \right\} \qquad (12.3)$$

where s'^2 is the residual error mean square (7·28), r is the number of replications (4), and A the sum of squares for x for the 'Error' line (156982·1). The variance is thus seen to depend on the square of the difference between the mean plant numbers of the two varieties being compared and consequently, will vary from one pair of varieties to another. It is evident that the range of variation in the squares of the differences of mean plant numbers for different pairs of varieties is not large in comparison to the sum of squares for x for the error, and it would suffice for most purposes to take the average of the variance values as the basis for varietal comparisons. This does not imply that all the individual variance values have to be calculated first in order to arrive at the average value. The mean variance of the difference of a pair of varieties can be shown to be given directly by the expression

$$\overline{V} = \frac{2s'^2}{r} \left\{ 1 + \frac{1}{p-1} \cdot \frac{B}{A} \right\} \qquad (12.4)$$

where p is the number of varieties being compared (16), B is the sum of squares for x for the 'Varieties' line (107775·3), and A, as before, is the sum of squares for x for the 'Error' line (156982·1). The average variance works out to be 3·807 in the present case. If $(p-1)A$ is large compared with B, it will be sufficiently accurate to take $2s'^2/r$ as the average variance $\overline{V}$ of the difference of a pair of varieties. The critical difference (*C.D.*) can now be calculated in the usual manner as the product of the standard error $\sqrt{\overline{V}}$ with the value of $t_{5\%}$ for 44 degrees of freedom on which the estimate of residual variance (s'^2) is based. This critical difference turns out to be $1{\cdot}951 \times 2{\cdot}021 = 3{\cdot}94$. Since, however, the varieties taken together did not indicate significance of differences among themselves, it would be hazardous to utilise the critical difference in the present case to investigate the significance of individual differences.

It will be noticed that in the present discussion the variance of the difference between two adjusted means was calculated without obtaining first the standard error of a single mean. The latter is easily derived. The expression for the adjusted mean it will be recalled is given by

$$\bar{Y}_i = \bar{y}_i - b(\bar{x}_i - \bar{x})$$

The variance of this estimate is therefore equal to

$$V(\bar{y}_i) + (\bar{x}_i - \bar{x})^2 V_b$$

where $V(\bar{y}_i)$ and V_b denote the residual variance of the varietal mean $\bar{y}_i$ and the variance of the regression coefficient b respectively. Now

$$V(\bar{y}_i) = \frac{s'^2}{r}$$

and

$$V_b = \frac{s'^2}{A}$$

Hence the variance of the varietal mean is

$$s'^2 \left\{ \frac{1}{r} + \frac{(\bar{x}_i - \bar{x})^2}{A} \right\} \qquad (12.5)$$

It should be noted that the variance of the difference of two adjusted means cannot be expressed directly as the sum of the variances of the respective means for the obvious reason that the two means are no longer independent and the variance of the difference of two means has to be worked out directly in the manner given above. Similar is the situation with other complex cases to be considered later. We have for this reason confined ourselves to giving in such cases only the variances of the differences of treatment means.

12*a*.5 Efficiency of Covariance Procedure

We can obtain an idea of the gain in precision made in adopting the covariance analysis by comparing the variance of the difference of a pair of varietal means as obtained from the simple analysis of variance of yields without adjustment for the concomitant variate with the corresponding average variance

after adjustment. The former it can be seen has a value $2 \times 7 \cdot 98/4 = 3 \cdot 99$ in the example under discussion while the latter has been shown to be 3·81. The gain in precision is therefore seen to be extremely small. Usually however the technique of reduction of experimental error through the analysis of covariance whenever the concomitant variate is suitably chosen is very effective and in fact is capable of wide application. The area of saline patches in experimental plots in fields affected by salinity provides a good example of a concomitant variate and will be discussed in Chapter XV. In choosing a concomitant variate it should be noted that the variate need not be necessarily measurable. Even if it lends itself to classification such as good, bad, indifferent, etc., and can be suitably scored the gain in efficiency from covariance technique is often considerable. Weed infestation provides an example, the degree of weed infestation being suitably scored to serve as a concomitant variate. The technique of covariance provides a powerful method of controlling the experimental error and should be availed of by the experimenter wherever feasible.

12*a*.6 Extension of the Technique of Analysis of Covariance

The analysis of covariance as a means of the statistical control of error can be extended to more than one concomitant variate. The procedure is essentially the same except that the sums of squares for y for the 'Error' line as well as the 'Varieties + Error' line would be corrected by subtracting from each the corresponding sums b_1 *S.P.* $(x_1y) + b_2$ *S.P.* $(x_2y) + b_3$ *S.P.* $(x_3y) + \ldots$ taken over all the concomitant variates $x_1, x_2, x_3, \ldots$ where the b's are the partial regression coefficients of y on $x_1, x_2, x_3, \ldots$ calculated from the sum of squares and sum of products in the same line. The adjusted sum of squares for y for varieties would be obtained by taking the difference of the two corrected items. The number of degrees of freedom for the adjusted 'Error' sum of squares would be reduced from that for the unadjusted sum of squares by as many units as the number of concomitant variates taken. Thus with two concomitant variates x_1 and x_2, the quantities required for calculating the partial regression coefficients b_1 and b_2 from the error line would be

S.S. (x_1^2), *S.S.* (x_2^2), *S.P.* (x_1y), *S.P.* (x_2y) and *S.P.* (x_1x_2). Similar quantities would have to be computed from the 'Varieties + Error' line in order to obtain the adjusted sum of squares for y. The adjusted sum of squares for error will have 2 degrees of freedom less. For further details the reader may refer to the treatment of the subject by Snedecor (1946).

12*b*.1 Analysis of Incomplete Observations

It will be observed that in all the designs we have been dealing with every variety or treatment has been repeated the same number of times (usually once) in each replication. Such a distribution by keeping each set of effects independent from any other, makes for the simplicity of the statistical analysis of the resulting data and renders the calculation of the various components of the total variation straight-forward. Nevertheless, it does at times happen that, howsoever well the experiment may be designed and howsoever carefully it may be conducted, the observations for particular plots might be lost or, even if available, are so affected by some accidental extraneous cause that it would not be proper to regard them as normal experimental observations. Such a contingency may arise, when for instance, one or more plots located on the edge of the layout become subject to depredation by cattle or birds, or manure is dumped on one or two spots in the layout sometime before spreading. In the latter case the plots occupying these spots might show a vigorous growth of the crop. The yields of such plots cannot be regarded as being subject to the usual random variations alone, and should be rejected from the experimental data. The plots so affected from extraneous causes and the data for which have to be omitted from analysis are called *missing plots*. A word of caution is, however, necessary in treating the values as missing. Rejection of values should never be done merely because the figures appear too high or too low but only when external evidence shows that the values proposed to be rejected and no others are affected by some accidental factor.

The statistical analysis of incomplete observations when one or more plots are missing is necessarily somewhat complicated owing to the disturbance in the initially symmetrical distribution

of plots among different treatments on the one hand, and among different blocks on the other. In cases where the disturbance affects only one plot, the procedure of analysing the data reduces to (*a*) calculating with the aid of appropriate standard formulæ involving the available data an estimate of the missing value; (*b*) working out the estimates of the treatment averages and the error mean square in the usual manner after inserting the estimate for the missing plot; and (*c*) making a certain additional adjustment to the sum of squares for treatments. Before giving the arithmetical procedure and the necessary formulæ appropriate for different designs it is instructive to develop with the help of an example the method of analysis for the case of a single missing plot in a randomized block design through an ingenious application of the analysis of covariance technique due to Bartlett (1937).

12*b*.2 Bartlett's Technique for Missing Plots

Example 12.2

Data pertaining to a sugarcane varietal trial at Jullundur are given in Table 12.6. The statistical analysis is described in this and subsequent sections.

TABLE 12.6

Sugarcane varietal trial

(Yield of cane in seers* per plot of 1/40th acre)

Variety	Block				
	I	II	III	IV	V
1	270	307	271	238	230
2	228	260	231	194	240
3	225	236	260	187	142
4	105	131	..	117	53
5	20	86	85	23	15

* A seer is roughly equivalent to 2 lb.

The yield of the plot under the fourth variety in block III was reported missing. The first step in the analysis consists in

assuming an imaginary or pseudo-variate x to be associated with y, the plot yield. The value of x is taken to be zero for every plot excepting the missing plot. For the missing plot itself, x is given the value -1 and y the value zero. Table 12.7 gives these values of x and y together with the varietal and block totals for each variate.

TABLE 12.7

Values of x and y

Variety	Block					*Total*
	I	II	III	IV	V	
1	270 (0)	307 (0)	271 (0)	238 (0)	230 (0)	1316 (0)
2	228 (0)	260 (0)	231 (0)	194 (0)	240 (0)	1153 (0)
3	225 (0)	236 (0)	260 (0)	187 (0)	142 (0)	1050 (0)
4	105 (0)	131 (0)	0 (−1)	117 (0)	53 (0)	406 (−1)
5	20 (0)	86 (0)	85 (0)	23 (0)	15 (0)	229 (0)
Total ..	848 (0)	1020 (0)	847 (−1)	759 (0)	680 (0)	4154 (−1)

The next step is to work out formally the analysis of covariance of y on x, the yield being regarded as the dependent variate. The computations are simple and are given in sequence below:

TABLE 12.8

Table of sums of squares and sum of products of x and y

Source of variation	*D.F.*	*S.S.* (x^2)	*S.P.* (xy)	*S.S.* (y^2)
Blocks ..	4	0·16	−3·24	12850·16
Varieties ..	4	0·16	84·96	185979·76
Error ..	16	0·64	84·44	18769·44
Total ..	24	0·96	166·16	217599·36

TABLE 12.9

Adjustment for error sum of squares

Source of variation	*D.F.*	*S.S.*
Error (unadjusted) ..	16	18769·44
Regression of y on x..	1	11140·80
Error (adjusted)..	15	7628·64

TABLE 12.10

Adjustment for varietal sum of squares and test of significance

Source of variation	*D.F.*	*S.S.*	*M.S.*	*F*
Varieties + Error (unadjusted) ..	20	204749·20	..	..
Adjustment	1	35870·45	..	..
Varieties + *Error* (adjusted) ..	19	168878·75	..	..
Error (adjusted)	15	7628·64	..	..
Varieties (adjusted)	4	161250·11	40312·53	79·27**
Error (adjusted)	15	7628·64	508·58	..

The entire procedure of covariance of y on the pseudo-variate x has been applied systematically in order to obtain the appropriate sums of squares for the differences between varieties and for the error required for the test of significance of varietal differences. The degrees of freedom for error have been reduced by 1, since one plot yield is missing. The F test of significance is carried out as usual, and shows in the above instance that the varietal differences are highly significant.

12*b*.3 ADJUSTMENT OF AVERAGES

Next we turn to the adjustment of varietal means. This is done by applying the usual linear regression formula

$$\bar{Y}_i = \bar{y}_i - b(\bar{x}_i - \bar{x})$$

The coefficient of regression in this case can be verified as

$$b = \frac{(rB + pV - G)}{(p-1)(r-1)}$$

where B is the total of the available yields in the block containing the missing plot, V is the total of the available yields of the variety which includes the missing plot, G is the total of all available yields, and p and r are respectively the numbers of varieties and blocks in the experiment. The value x to which the yield is adjusted, is however not the general mean of x's, but the value zero, in order that the means of varieties without any missing plot may not be affected by the adjustment, as would appear to be but proper. The adjusted value of the mean for the variety with the missing plot is

$$\frac{1}{r}\left[V + \frac{(rB + pV - G)}{(p-1)(r-1)}\right]$$

or 107·59 in the example chosen. The expression is precisely the same as the ordinary mean yield of the variety, had the yield of the missing plot been

$$\frac{(rB + pV - G)}{(p-1)(r-1)} \qquad (12.6)$$

The latter expression, in fact, provides the easier method of estimating the value of the missing plot in the randomized block design. It should be noted, however, that this estimation is not done with a view, as it were, to supply the missing value as is often loosely stated, but in order to correct the unadjusted mean $V/(r-1)$ of the variety under consideration for the effect of block differences arising out of the disturbance in the uniform distribution of plots among blocks and varieties. If, for instance, the missing plot belonged to a particularly fertile block, the mean of the variety with the missing plot will be under-estimated in comparison with the other varietal means. It is such disturbances both in the comparison of individual means and the varietal differences as a whole, as also in the estimate of the plot to plot variation which is similarly affected by the missing plot, that the adjustments seek to rectify.

12*b*.4 Standard Errors of Comparisons

Lastly, the standard errors of the comparisons have to be adjusted wherever necessary. The standard error of the difference between two varieties neither of which contains the missing plot remains unaffected and is given by the square-root of $(2s'^2/r)$, s'^2 being the adjusted error mean square and has the value 14·26 here. The standard error of the comparisons of the mean for the variety with the missing plot with any other needs adjustment in the light of the correction of the former mean and is given by the square-root of

$$\frac{s'^2}{r}\left\{2+\frac{p}{(p-1)(r-1)}\right\} \qquad (12.7)$$

and has a value of 15·34 in the present case. The values of varietal means with the critical differences are given in Table 12.11.

TABLE 12.11

Average yields (in seers/plot)

Variety number	1	2	3	4	5
Average yield	263·2	230·6	210·0	107·6	45·8

C.D. between variety No. 4 and any other	..	32·7
C.D. between any two other varieties	..	30·5

12*c*.1 Method of Substitution for the Missing Value

We have developed and illustrated above the exact method of dealing with incomplete data with one missing observation. This method leads to the substitution for the missing plot the value

$$\frac{(rB+pV-G)}{(p-1)(r-1)}$$

for adjusting the varietal mean. The subsequent analysis is done as though the data were complete, the only deviation being the reduction of a single degree of freedom from the error degrees of freedom. While the procedure provides a proper estimate of the error variance per plot, the varietal sum of squares thus obtained is inflated and consequently the significance of the

varietal differences is exaggerated. Applying the method to the illustrative example (12.2) we have the following table of analysis of variance:

TABLE 12.12

Analysis of variance of 'completed data'

(Unit : Seers/plot)

Source of variation	*D.F.*	*S.S.*	*M.S.*	*F*
Blocks ..	4	16490·32	4122·58	..
Varieties ..	4	166346·14	41586·54	81·77
Error	15	7628·64	508·58	..
Total ..	23	190465·10	..	..

It will be seen by comparison with Table 12.10 that while the error sum of squares tallies, the varietal sum of squares is inflated. The latter may however be corrected by subtracting from it the quantity

$$\frac{(B + pV - G)^2}{p(p-1)(r-1)^2}$$

which in the present case amounts to 5096·0281. Thus with a single missing plot the correct analysis of a randomized block experiment may be carried out economically by (i) evaluating the adjustment for the varietal or treatment mean based on what is called the *missing plot* value from the formula

$$\frac{(rB + pV - G)}{(p-1)(r-1)}$$

(ii) inserting this value in the place of the missing plot and carrying out the ordinary analysis of variance of the completed data, (iii) deducting the quantity

$$\frac{(B + pV - G)^2}{p(p-1)(r-1)^2}$$

from the varietal sum of squares obtained in (ii) in order to obtain the correct varietal sum of squares, and (iv) testing for significance the correct varietal sum of squares against the error

in the analysis of variance table in the usual way. The procedure of adjusting the varietal means and evaluating the critical differences is already given in sections 12*b*.3 and 12*b*.4.

12*c*.2 Extension of the Covariance Technique to More than One Missing Value

The method of application of the analysis of covariance technique employing pseudo-variates can be easily extended to more than a single missing plot to obtain the exact estimates of treatment and error sums of squares. In the case of two missing plots, for instance, it is necessary to assume two pseudo-variates, x_1 and x_2, say, the first taking the value zero for all plots save the first missing plot for which it takes the value -1 and similarly the second taking the value zero for all plots except the second missing plot for which it takes the value -1. The yield variate y is assigned the value zero for all the missing plots and the covariance of y on x_1 and x_2 is worked out to obtain the adjusted sum of squares for treatment and error by following the partial regression technique appropriate for two independent variates. It will be seen, however, that the method becomes exceedingly complicated as the number of missing plots increases. The alternative method described in section 12*c*.1 is simpler but there the unadjusted estimate of treatment sum of squares is somewhat inflated. Ordinarily this need not vitiate conclusions except where the results are just significant. In most cases therefore it is sufficient to adopt the simpler method and carry out the ordinary analysis of variance. If the results are significant it is advisable to make a test of significance again after adjusting the treatment sum of squares.

The first step in this method consists in obtaining the missing values to be inserted. In order to illustrate the procedure, let us assume that in the illustrative example of section 12*b*.2 (Table 12.6) a second plot yield, namely, that of the plot in block I under variety 4 is also missing. Let a and b denote the two missing values to be evaluated. We write two equations by equating to each missing value in turn the expression

$$\frac{(rB + pV - G)}{(p-1)(r-1)}$$

where the values of V, B and G are written with the yield of the other missing plot represented by the appropriate algebraic symbol a or b. Thus we have

$$V_a = 301 + b,\ B_a = 847,\ V_b = 301 + a,\ B_b = 743,$$

$$G_a = 4049 + b,\ G_b = 4049 + a$$

Then

$$a = \frac{[5 \times 847] + [5\,(301 + b)] - [4049 + b]}{4 \times 4} \text{ or } 16a - 4b = 1691$$

and

$$b = \frac{[5 \times 743] + [5\,(301 + a)] - [4049 + a]}{4 \times 4} \text{ or } 16b - 4a = 1171$$

Solving these two simultaneous equations we get

$$a = 132{\cdot}25 \text{ and } b = 106{\cdot}25$$

12c.3 The Iterative Method

While the covariance method is exact and can be employed for dealing with more than two missing plots a rapid method for practical use is the iterative process. In this method the general mean of available yields is substituted as a first step for all missing values except one which is then estimated by the standard formula,

$$\frac{(rB + pV - G)}{(p - 1)(r - 1)}$$

for a single missing value. The process in turn is carried out for all the missing values giving the first approximations. At each subsequent stage, the approximate values estimated earlier are substituted in place of the general mean. The process is continued until the successive approximations agree closely.

Thus in the example above we substitute for variety 4 in block I the general mean 4049/23 and obtain for the first approximation for variety 4 in block III the value

$$a_1 = 141{\cdot}8$$

Then we substitute the general mean 4049/23 in place of variety 4 in block III and obtain for variety 4 in block I the first approximation given by

$$b_1 = 109{\cdot}3$$

The second approximations can be seen to be given by

$$a_2 = 133{\cdot}0$$

$$b_2 = 108{\cdot}6$$

while the third by

$$a_3 = 132{\cdot}8$$

$$b_3 = 106{\cdot}4$$

It will be noticed that the second and the third approximations are close to each other, making it unnecessary to proceed further. Usually two approximations are sufficient.

With the missing values thus estimated, the average yields of the treatments are calculated and the usual analysis of variance worked out. The exact expressions for the standard errors of the various differences between the average yields become very complex in the case of several missing values. It often suffices, however, to employ an approximate method in estimating the standard errors in such cases. This consists in assigning to each treatment what may be called an *effective number of replications* and then applying the familiar formula for the standard error of the difference of two means based on r_1 and r_2 replications respectively, namely

$$\sqrt{\frac{s^2}{r_1} + \frac{s^2}{r_2}}$$

where s^2 is the error mean square from the analysis of variance table, the degrees of freedom for which are reduced from the normal number by as many units as the total number of missing values. The effective numbers of replications depend on the two treatments being compared. If these be v_1 and v_2 say, the effective number of replications r_1, for v_1 is computed as follows:

Any observation on v_1 is given the score 1, if the corresponding replication contains both v_1 and v_2.

It is given the score $\frac{1}{2}$ if the replicate contains v_1 but v_2 plot yield is missing.

It is given the score 0, if in the replicate, v_1 itself is missing.

The total of such scores, taken over all the replications, gives the effective number r_1. Similarly r_2 is computed for v_2, taking account of v_1. As an example, consider the difference between v_1 and v_4 given in Table 12.6 in which the yield under v_4 in block I is also taken to be missing in addition to that under v_4 in block III. Taking block by block,

$$r_1 = \tfrac{1}{2} + 1 + \tfrac{1}{2} + 1 + 1 = 4$$

$$r_4 = 0 + 1 + 0 + 1 + 1 = 3$$

12*d*.1 Missing Plot in the Latin Square

We have confined ourselves to the randomized block design in the discussion so far. Other designs such as the latin square may be dealt with similarly by the application of the analysis of covariance technique or by the approximate method of completed data. For a single missing value in a latin square, the so-called missing value is given by the formula

$$\frac{[p(R + C + V) - 2G]}{(p-1)(p-2)} \tag{12.8}$$

where p is the number of varieties or treatments, R, C and V the totals of available yields in the row, column and treatment respectively containing the missing plot, and G is the total of all available yields. By inserting this value and working out the ordinary analysis the error sum of squares is correctly obtained, but the treatment sum of squares must be corrected by subtracting the quantity

$$\frac{[(p-1)V + R + C - G]^2}{(p-1)^2(p-2)^2}$$

from the value obtained. The standard error of the difference between the treatment containing the missing value and another is given by

$$\sqrt{s'^2\left\{\frac{2}{p} + \frac{1}{(p-1)(p-2)}\right\}} \tag{12.9}$$

and that between any other two treatments being given by $\sqrt{2s'^2/p}$ as usual.

In the case of several missing values, the standard error of the differences between average yields is estimated approximately by the same method of effective number of replications as in the case of the randomized block design. The rule for finding the effective number is as follows:

(i) Any observation on v_1 is assigned a score 1, if both the row and column containing the observation also contain an observation on the other treatment v_2 being compared with v_1;

(ii) It is given a score $\frac{2}{3}$, if either the row or the column alone includes a missing plot for v_2;

(iii) It is given a score $\frac{1}{3}$, if both the row and the column include a missing plot each for v_2;

(iv) Finally, it is assigned a score 0, if v_1 itself is missing.

The sum of such scores gives the effective number of replications.

More difficult cases like the missing of an entire row in a latin square involve complications which would take us beyond the scope of the present book.

CHAPTER XIII

DESIGNS FOR PLANT BREEDING TRIALS

13*a*.1 Introduction

As in the comparison of agronomic treatments or crop varieties plant breeding work also involves the use of experimental designs based on the principles explained in the previous chapters. The techniques appropriate to plant breeding trials are however of comparatively recent development, chiefly on account of the fact that for a long time plant breeding was considered an art depending for the most part on the judgment and skill of the plant-breeder. The necessity of objective testing was realised only gradually. The small and variable quantities of seed produced by plants presented a basic difficulty and it did not appear feasible to compare the genetic value of different progenies, that is, groups of plants derived from different parent plants, by the application of the principles of randomization and replication. Again in early stages of breeding the material is often very heterozygous and this presented one further difficulty in conducting replicated tests as the genetic variance would also contribute to experimental error. These difficulties appear to have dissuaded the plant-breeder from using modern experimental designs in his work. However, as a result of researches carried out at the Institute of Plant Industry at Indore, objective testing of plant breeding material right from the stage of single plant selections has been found to be practicable and effective and the technique of plant selection has now been placed on a sound footing. This technique enables the breeder not only to sift more material at each stage of selection and testing, but also to assess the scope for further improvement. The purpose of this chapter is to deal chiefly with the designs available for the field trials of plant breeding material.

13*a*.2 Replicated Progeny Rows

An important phase in the development of the plant breeding technique was the switchover from mass selection to progeny row

breeding. Mass selection, which was the method in use in old days, consists in choosing from the material under selection a number of plants that appear to have a superior value in respect of the character or characters under selection, bulking the seed from these, raising from the seed the next generation and continuing selection in this generation as was done in the previous generation. The method is inefficient since in each generation the selection is subject to the large amount of environmental or nongenetic variability present in the field. It was discovered by breeders that if, instead of bulking, the seed obtained from different selected plants was sown in separate progeny rows and selections made on the basis of the progeny means, the selections would be more reliable as representing superior genotypes. This is due to the fact that the progeny means are susceptible to only a fraction of the environmental variation to which individual plants are subject, the deviations of individual plant values in opposite directions from this cause mostly cancelling out in taking their mean. Adoption of this method led to the realisation that the comparison of the means of different progenies should be unbiassed and as accurate as possible. This object could be achieved only by carrying out randomized replicated trials of progenies instead of basing the comparison of the progeny means on single unreplicated rows. The difficulties in the way of such replicated tests referred to above have been overcome successfully. For unbiassed comparison of progenies it is necessary that the various progenies are randomized in each replication and for increased accuracy a larger number of replications rather than a larger plot size is desirable. The number of replications can be increased to the desired extent by reducing the plot size to a few plants per plot. The suitability of this type of layout has been established for quite a wide range of crops. Examples are cotton, sorghum, wheat, groundnut, linseed. It is found that even with the small plot size necessary in these trials the means based on 8 to 10 replications which are usually feasible can be estimated with standard errors of 10 per cent. or less for a highly variable character like yield, while for more stable characters like staple length or ginning percentage in cotton or oil content in oilseeds standard errors of only one or two per cent. can be attained. An example of

a progeny row trial on cotton laid out at Indore will help to illustrate the technique.

Example 13.1

A replicated progeny row trial with progenies of Malvi 1 strain of cotton was carried out at the Institute of Plant Industry, Indore. Progenies from the seed of 10 plants taken from the parental generation were grown for comparison so that there were 10 plots in each block. Each plot consisted of a single row of 5 plants (barring variation due to failure of germination or deaths due to wilt). The spacing between rows of plants was 2 feet, while that between plants was 1 foot. Two rows of bulk seed of the same strain were sown all round each block as border rows to avoid border effect on the experimental material. There were 10 replications. The field plan is reproduced in Fig. 13.1.

The analysis of variance of the data is done in the same way as that for a simple randomized block experiment, the different progenies being the different treatments to be compared. The results of the analysis of variance for four characters including yield are given in Table 13.1 (*a*) and (*b*).

TABLE 13.1 (*a*)

Summary of analysis of variance of original yield and yield adjusted for stand (in gm./plot)

Source of variation	Degrees of freedom	Mean square		*F*	*F*
		Unadjusted yield	Adjusted yield	$P = \cdot 05$	$P = \cdot 01$
Blocks	9	1188·3	..	2·02	2·31
Progenies	9	446·7	395·2	2·02	2·31
Error	81*	744·7	417·7	..	..
Standard error per cent. (mean of 10 plots) ..		14·9	11·2	..	..

* Number of degrees of freedom to be reduced by one for adjustment on stand.

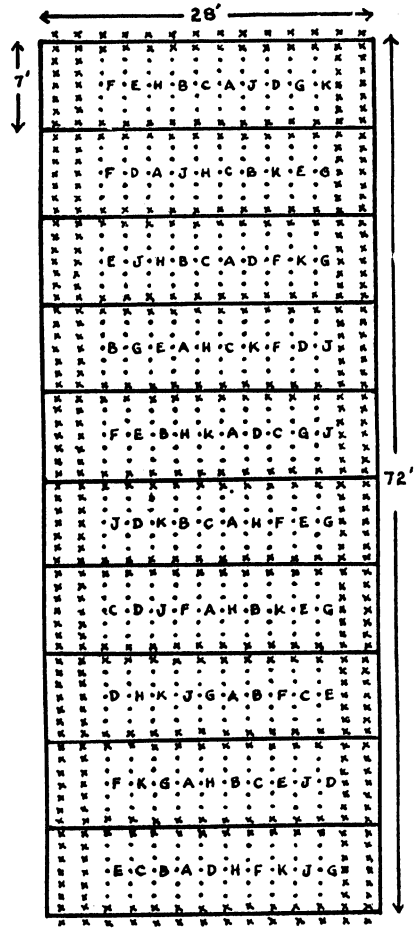


FIG. 13.1. Yield plan of replicated progeny row trial (Malvi 1, 1934).

..... Experimental plots

× × × × × × Guard rows

A–J Ten progenies

TABLE 13.1 (*b*)

Summary of analyses of variance for other characters

Source of variation	Degrees of freedom	Mean squares		
		Staple length in mm.	Ginning per cent.	Stand
Blocks	9	0·78	32·65	3·19
Progenies	9	0·74	24·38	2·11
Error	81	0·51	2·66	0·87
Standard error per cent. (mean of 10 plots)		1·0	1·8	7·2

The differences between progenies are found to be significant for the characters stand and ginning percentage. The strain Malvi 1 was the result of selection continued over 8 seasons. That the differences between progenies of a strain inbred for 8 generations could be detected by this trial shows the sensitiveness of the method. The plot yields partly depend upon the stand and so an allowance has to be made for variation in stand while analysing the results for yield. This is done by the application of the analysis of covariance between yield and plant number per plot. The analysis of yield thus adjusted is given in the table. It will be seen that the adjustment reduced the standard error for progeny mean from 14·9 to 11·2 per cent. Considering that yield is highly susceptible to environmental variations this is regarded as a fairly reasonable value.

13*a*.3 COMPACT FAMILY BLOCKS

In Example 13.1 all progenies belonged to a single strain. When, however, progenies tried in an experiment belong to a number of strains, families, crosses or field selections from different localities, two types of comparisons are involved: (1) the comparison between different families, crosses, etc. and (2) the comparisons between progenies within them. If there are t progenies per family, u families and r replications, all the tu plots being randomized together in each of the r blocks, the standard error of the mean of a single progeny will be $s/\sqrt{r}$ since the progeny mean is based on r plots; whereas the standard

error of a family mean will be $1/\sqrt{t}$ times this value, that is, $s/\sqrt{tr}$, the family mean being based on tr plots. The latter error should be used for family comparisons. For comparison of progenies within families it is desirable to analyse the various families separately and to test the progeny differences within each family against the error obtained from the data of that family. The error variance might differ significantly from family to family and consequently in taking a pooled estimate of error, the error might be over-estimated in respect of some families and under-estimated in others. Even if the error variances do not differ significantly between the families it would not be appropriate to carry out an overall test of significance of differences between progenies within families, as the progeny differences might be significant in some and non-significant in other families and the latter might mask the former.

When, as above, it is intended to study the progeny differences within families, it is advisable to use another design, one which utilises the principle of local control more fully than a randomized block design with progenies randomized all over the block, in order to reduce the error variance of progeny comparisons within families. In this design the progenies belonging to the same family are sown side by side in a family main plot. The family main plots are first randomized in each block and progeny plots, which constitute the sub-plots, are randomized within the main plots. This design, known as the *compact family block design*, is thus quite analogous to the split-plot design. Example 13.2 describes a compact family block trial with 30 progenies belonging to 6 families of Malvi 9, carried out at the Institute of Plant Industry, Indore.

Example 13.2

In this trial there were 5 progenies in each of 6 families and there were 10 replications. The progeny plots consisted of 5 plants spaced 1 foot apart and the distance between rows was 2 feet. There was a border row of bulk seed surrounding each block. The field plan of the trial is given in Fig. 13.2.

The families 17–34, 17–38, 19–45, 19–46, 20–51 and 20–58 were randomized within each block and progenies 1 to 5 of each

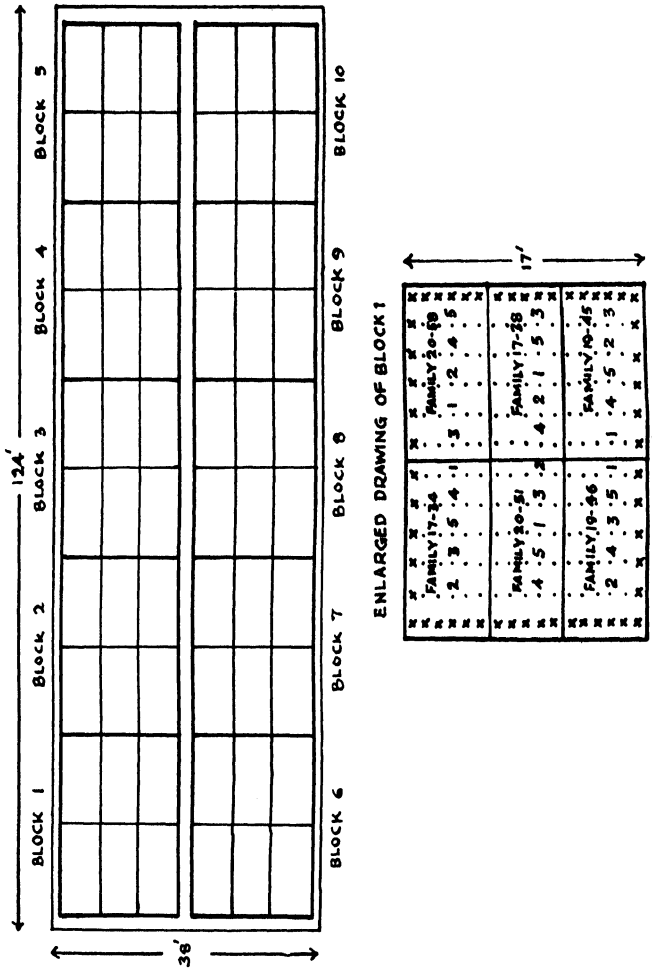


FIG. 13.2. Yield plan of compact family block trial (Malvi 9, 1935).
..... Experimental plots
× × × × × × Guard rows

family randomized within the family main plots. An enlarged drawing of replication I is given in Fig. 13.2 and would make the layout clear.

The analysis of the trial is carried out in two stages: First from the main plot values the variance between families and the corresponding error are calculated treating the experiment as one in simple randomized blocks; then the sub-plot values for progenies are analysed separately for each family to give the variance between progenies for each family and the corresponding error. In this respect the analysis differs from the analysis of a split-plot trial. The analyses of variance of yield for families and for progenies in each family are given in Table 13.2.

TABLE 13.2

Analyses of variance of yield in gm. per plot

Between families			Between progenies within families						
					Mean squares within families				
Due to	*D.F.*	*M.S.*	Due to	*D.F.*	17–34	17–38	19–45	19–46	20–51	20–58
Replications ..	9	30823	Blocks ..	9	1275	2150	2778	2908	3168	6413
Families ..	5	44508	Progenies	4	499	4350	5092	1415	782	921
Error ..	45	12527	*Error* ..	36	1011	1039	1541	1016	1501	985

It can be seen from the results that the differences between families are significant as also that differences between progenies are significant in families 17–38 and 19–45.

Owing to the random arrangement of family plots within blocks and also of progenies within family plots environmental contributions to the error variances within different families tend to be equalised. Hence a comparison of these error variances provides a rough idea of the relative magnitudes of genetic variability within progenies in the various families. However, it is necessary to carry out a test of homogeneity of error variances before making the comparison, since differences in the genetic variabilities could be assumed to exist only if significant heterogeneity between error variances is revealed by the test. Bartlett's

test of homogeneity of variances is useful for this purpose. The test is carried out as follows:

Let there be n mean squares, s_1^2, s_2^2, s_3^2, ..., s_n^2 based respectively on k_1, k_2, ..., k_n degrees of freedom. From these values a pooled estimate of variance $\bar{s}^2$ is calculated by the formula

$$\bar{s}^2 = \frac{1}{\sum_1^n k_r} \left(\sum_1^n k_r s_r^2 \right) \tag{13.1}$$

Next the quantity

$$\chi'^2 = \left\{ \left(\sum_1^n k_r \right) \log_e \bar{s}^2 - \sum_1^n k_r \log_e s_r^2 \right\} \tag{13.2}$$

is calculated. For convenience, the logarithms may be taken to the base 10 and the result multiplied by $\log_e 10$, that is, 2·3026, to get the quantity χ'^2. The quantity χ'^2 is distributed approximately as χ^2 with $(n - 1)$ degrees of freedom but is slightly biased upwards. It can be effectively corrected by dividing by the correction factor

$$C = 1 + \frac{1}{3(n-1)} \left\{ \sum_1^n \frac{1}{k_r} - \frac{1}{\sum_1^n k_r} \right\} \tag{13.3}$$

The quantity χ'^2/C is referred to the table of χ^2 with $n - 1$ degrees of freedom; if the value is significant, the variances are significantly heterogeneous. If k_r has the same value for each variance the formulæ are simplified. In the present case we have

$$\bar{s}^2 = 1182, \quad \chi'^2 = 4{\cdot}131, \quad C = 1{\cdot}011$$

and

$$\chi^2 (5\ d.f.) = 4{\cdot}052$$

On reference to the χ^2 tables we find this to be a non-significant value.

An important advantage of this method of analysis is that considerable amount of labour involved in computation can be saved by conducting a partial analysis. If from the results of family main plots some families are found to be inferior in some essential characters like germination, stand or vigour, the analysis

of progeny plots within these families can be dropped, while the results for the whole families can be included in the main plot analysis. The labour saved in this way can be considerable when a large number of families is tried. In fact we can go a step further and decide not to harvest separately the progenies of the inferior families should a spot inspection in the field before harvest indicate so. In that case, if the experiment is arranged in compact family blocks, all the progenies of the inferior families can be harvested in bulk and this would result in a considerable economy of labour in the field. The data for family bulks can be included in the main family analysis for completeness while saving the labour of separately harvesting and weighing of progeny plots. This device is particularly valuable where experimental facilities are poor as when experiments are conducted at out-stations away from the main research station.

Yet another advantage of the replicated designs for plant breeding lies in the way they help the experimenter in the eye judgment of observational characters like vigour, lodging or earliness. When these observations are based on plot-to-plot examination in randomized progeny row trials they can be interpreted with more confidence and even subjected to statistical analysis by recording the observations on a suitable quantitative scale. If a particular progeny fails as judged by any one or more of these characters in all or several replications, the fact that the progenies are randomized in each block ensures that this is not due to an accidental effect but to the inferior quality of the progeny itself.

13*a*.4 Efficient Utilisation of Seed

In progeny row breeding work the amount of seed available often varies widely among the plants selected as parents. Where the minimum seed available is sufficient to sow 10 or more plots of the size decided upon this variation is of no consequence. In other cases, the reduction of the number of replications to that for which the lowest yielding plant provides sufficient seed, will result in loss of efficiency. Where most of the selected parent plants have given about the same amount of seed, and two or three considerably less, a plot size and number of replications

may be employed which will use up nearly all the seed of the higher yielding group. The seed of progenies of the lower yielding group is used to sow as many plots as the amount available permits, and the remaining plots allotted to those progenies are sown with bulk seed, either surplus from the higher yielding plants or from related strains. In interpreting the results, however, the values for such dummy plots sown with bulk are excluded and the appropriate analysis of the remaining data carried out by the technique of incomplete observations described earlier.

The following example of a linseed progeny row trial laid out at Indore provides a good example.

Example 13.3

The trial included 23 progenies and was laid out in 7 randomized blocks. Of these five contained all 23 progenies. The sixth contained 17 progenies, and the seventh, 15. The layout is given in Fig. 13.3.

The 5 complete blocks were analysed separately for comparison. The summaries of analyses for the 5 complete blocks and for all the 7 blocks are given in Table 13.3.

TABLE 13.3

Analyses of variance of yields of linseed progenies

Due to	Five complete blocks		7 Blocks (Two incomplete)	
	D.F.	*M.S.*	*D.F.*	*M.S.*
Blocks	4	118·65	6	106·70
Progenies	22	10·69	22	14·10
Error	88	14·30	118 (132−14)	13·83
Standard error per cent.				
Mean of 5 plots ..	..	10·31	..	10·48
Mean of 7 plots ..	..	..	..	8·87

The number of degrees of freedom available for the estimation of error was 118 in the calculation over all the 7 blocks, that is, 14 less than the number that would have been available if there were no missing plots. The results show that the standard error

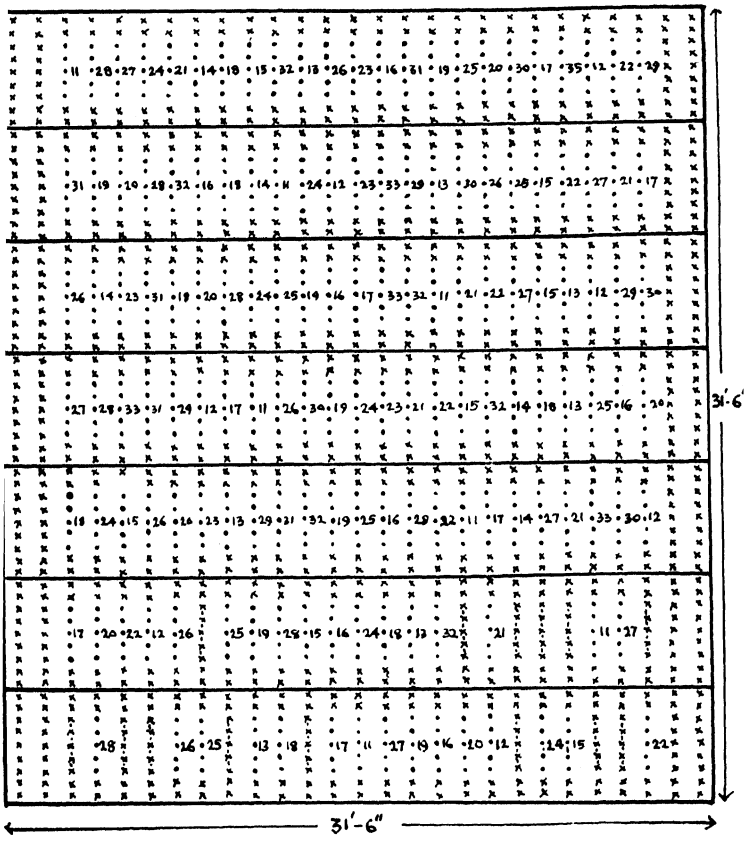


FIG. 13.3. Yield plan of replicated progeny row trial (linseed, 1934).

..... Experimental plots

×××××× Guard rows

×-×-×-×-× Plots sown with bulk

per cent. for those progenies which appeared in all blocks was reduced from 10·31 to 8·87, an appreciable gain. For those progenies which appeared in 5 blocks only, it remained practically unchanged. The example shows how variable quantities of seed can be utilised efficiently. If the amount of seed from selected parents varies more widely, another device can be used. In such

cases, the number of blocks might be made equal to the number of plots that can be sown with the seed available from the lowest yielding parent. Then parents yielding twice as much seed can be included in 2 plots per block, those yielding three times as much seed, in 3 plots per block, and so on. The analysis of variance of such data presents no essentially new feature, except that in calculating the sums of squares between progenies and also between families, if these are involved, appropriate divisors have to be employed depending on the total number of plots of each progeny or of each family. This precaution is also necessary when calculating the standard errors for comparison between individual progenies and families. If the seed of some parent is slightly less than that required to sow the requisite number of plots in all blocks one plot in one or two blocks can be sown with bulk and the data treated by the missing plot technique appropriate to the design. For more details about such devices as well as plant breeding technique in general reference should be made to Hutchinson and Panse (1937) and Panse (1940).

13*b*.1 Incomplete Block Designs

Plant breeding trials frequently involve a large number of progenies, and at a later stage in the breeding programme a large number of strains. If the number of varieties to be compared (using the term varieties in a very general sense) is moderate, say not more than 15 or 20, the trial may be laid out in simple randomized blocks. If a larger number of varieties is to be tested, or even with a moderate number if the soil heterogeneity is pronounced, the block accommodating a complete replicate might not be sufficiently homogeneous and the precision of the varietal comparisons would suffer. The situation is similar to that arising in factorial experiments when the number of treatment combinations is large. The same solution, namely, to spread a replication over more than one block, is applied here also. Thus the principle of reduction in error variation through local control is maintained by grouping the plots into compact homogeneous blocks of a moderate size, by accommodating in each block only a portion of the total number of varieties to be compared. Designs of this type are known as *incomplete block designs*.

In the case of confounded factorial designs the aim was to preserve the accuracy of the estimates of main effects and first order interactions at the cost of information on higher order interactions. In a varietal trial, however, we are equally interested in all the varietal comparisons and cannot afford to neglect any of these in planning the experiment. The device of partial confounding is therefore adopted, that is to say, that varietal comparisons which are confounded in one set of replications are kept free from any confounding in another set of replications and *vice versa.* Thus comparisons are available between all pairs of varieties in the experiment and owing to the reduced size of the block, these are generally more precise than if the same experiment had been laid out in simple randomized blocks. We shall describe in the remaining part of this chapter some simpler designs of this type.

13*b*.2 THE SIMPLE LATTICE DESIGN

The simplest incomplete block design is available when the number of varieties, v, is a perfect square, say $q^2 = 16, 25, 36, 49$, etc., and is known as the *simple lattice* or the *double lattice.* It involves a minimum of 2 replications in blocks of q plots each. A larger number of replications, which is a multiple of 2, is also possible, these being laid out by repeating the arrangement in the initial 2 replicates. The various steps in the construction of the double lattice design may be illustrated with 64 varieties, so that $q = 8$.

1. The varieties are numbered at random from 1 to 64, and set down in the form of a square array of side $q = 8$ as follows:

1	2	3	4	5	6	7	8
9	10	11	12	13	14	15	16
17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32
33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48
49	50	51	52	53	54	55	56
57	58	59	60	61	62	63	64

2. One replication, which may be called the X group, is formed by taking the row sections as blocks, that is, by placing

together in one block the 8 varieties occurring in the same row. Thus varieties 1 to 8 would form one set to be allotted to plots of one block. In this way the X group of 8 sets forming one replication is obtained.

3. A second replication, designated as the Y group, is obtained by taking column sections of the array. Varieties occurring in the same column are grouped into one set to be allotted to plots of the same block. For example, varieties 1, 9, 17, 25, 33, 41, 49 and 57 form a column set.

The two groups, X and Y, together constitute the double lattice design. It may be noted that of $(q^2 - 1)$ degrees of freedom for the varietal comparisons, a set of $(q - 1)$ degrees of freedom is confounded with the q blocks of the X group and another set of $(q - 1)$ degrees of freedom is confounded with the q blocks of the Y group. These two sets of degrees of freedom are different. The remaining $(q - 1)^2$ degrees of freedom are free from confounding in either group.

4. The 8 sets of varieties of the same group (X or Y) forming a replication should be allotted at random to 8 contiguous blocks of 8 plots each, to form a compact replicate. The 8 varieties allotted to a block are then assigned at random to the 8 plots in the block. If the design consists of a multiple of 2 replications the varieties are assigned to plots in the blocks of the additional replications with fresh randomization.

A typical plan for a simple lattice for 64 varieties in 4 replications is shown in Table 13.4.

13*b*.3 The Statistical Analysis of Incomplete Block Designs

The analysis of the incomplete block designs is complicated by the fact that varietal differences are partly confounded with block differences and have therefore to be adjusted for the block effects. Further, the block differences themselves being identical with certain varietal comparisons provide some information on the latter. It is desirable to recover this *interblock information* in order to make the most efficient use of the data collected. It should be pointed out, however, that provided the incomplete block design consists of contiguous blocks forming complete

TABLE 13.4

Plan and yields of experiment in 8×8 simple lattice design with 4 replications

(Serial number of variety in brackets, yield of grain in oz. per plot)

				REPLICATION I				
Block 1.	(38)	(34)	(35)	(39)	(40)	(33)	(37)	(36)
	9·0	6·5	11·8	5·0	9·5	9·5	7·3	7·3
Block 2.	(21)	(22)	(24)	(20)	(18)	(19)	(23)	(17)
	10·8	10·3	10·0	11·0	11·3	7·8	7·5	7·5
Block 3.	(14)	(10)	(12)	(13)	(11)	(9)	(16)	(15)
	7·8	12·0	9·0	6·3	9·3	8·3	8·8	9·5
Block 4.	(5)	(2)	(7)	(4)	(3)	(6)	(8)	(1)
	11·0	8·8	8·0	9·5	9·5	9·0	9·0	6·5
Block 5.	(29)	(31)	(26)	(23)	(30)	(32)	(25)	(27)
	11·8	10·5	12·5	8·0	9·5	10·0	6·3	6·3
Block 6.	(45)	(47)	(41)	(44)	(43)	(42)	(48)	(46)
	9·3	12·3	11·0	12·8	8·5	8·3	7·3	8·8
Block 7.	(64)	(57)	(59)	(61)	(63)	(58)	(60)	(62)
	12·3	12·3	10·0	10·0	10·0	10·8	9·3	10·5
Block 8.	(55)	(49)	(56)	(51)	(54)	(52)	(53)	(50)
	9·0	5·5	8·5	7·5	7·5	6·8	7·3	11·0
				REPLICATION II				
Block 9.	(43)	(35)	(51)	(27)	(59)	(11)	(19)	(3)
	8·5	15·0	7·0	7·0	10·0	6·3	8·5	8·0
Block 10.	(34)	(50)	(42)	(26)	(2)	(18)	(10)	(58)
	9·3	12·0	8·3	13·5	13·8	14·0	13·3	10·8
Block 11.	(37)	(29)	(53)	(45)	(21)	(61)	(13)	(5)
	11·3	7·8	10·0	9·3	10·3	10·0	3·8	14·0
Block 12.	(48)	(8)	(32)	(24)	(56)	(40)	(16)	(64)
	7·3	14·8	10·3	12·0	5·8	7·3	6·5	12·3
Block 13.	(23)	(7)	(55)	(31)	(63)	(15)	(47)	(39)
	7·5	8·0	9·0	10·5	10·0	9·5	8·0	5·0
Block 14.	(60)	(4)	(20)	(52)	(44)	(36)	(28)	(12)
	8·8	9·5	11·0	6·8	11·3	7·3	8·0	9·0
Block 15.	(22)	(46)	(14)	(62)	(38)	(6)	(30)	(54)
	12·3	8·8	9·5	10·5	14·8	9·5	11·3	9·5
Block 16.	(49)	(17)	(41)	(1)	(57)	(9)	(25)	(33)
	8·8	11·3	11·0	8·8	12·3	10·3	9·5	9·3
				REPLICATION III				
Block 17.	(11)	(12)	(15)	(13)	(9)	(16)	(10)	(14)
	5·0	11·3	12·3	8·3	11·8	7·8	10·0	7·0
Block 18.	(50)	(56)	(55)	(49)	(51)	(52)	(54)	(53)
	17·3	5·5	9·5	6·8	5·0	7·5	8·3	10·8
Block 19.	(26)	(31)	(27)	(29)	(30)	(32)	(25)	(28)
	10·8	12·0	6·8	7·0	6·8	9·5	10·3	11·5
Block 20.	(20)	(22)	(23)	(19)	(18)	(21)	(24)	(17)
	10·3	10·3	7·0	7·8	9·3	8·8	7·8	7·3
Block 21.	(47)	(41)	(42)	(45)	(48)	(46)	(44)	(43)
	10·3	12·0	6·0	10·5	1·5	7·8	6·8	6·8
Block 22.	(40)	(35)	(39)	(37)	(34)	(33)	(38)	(36)
	6·5	6·3	4·0	11·0	5·5	8·0	7·3	5·8
Block 23.	(62)	(57)	(60)	(64)	(61)	(63)	(58)	(59)
	11·0	10·0	9·0	16·8	10·8	9·9	11·8	8·0
Block 24.	(6)	(1)	(4)	(3)	(7)	(5)	(2)	(8)
	7·0	6·5	10·5	7·0	8·8	14·0	10·5	8·3

TABLE 13.4—(*Continued*)

REPLICATION IV

Block 25.	(64)	(56)	(48)	(8)	(24)	(32)	(40)	(16)
	8·3	6·8	1·5	8·3	7·8	9·0	6·5	9·8
Block 26.	(62)	(46)	(38)	(6)	(30)	(22)	(54)	(14)
	11·0	19·3	9·5	9·5	6·8	11·0	8·3	7·0
Block 27.	(53)	(45)	(37)	(61)	(13)	(5)	(29)	(21)
	11·3	10·5	11·0	7·0	7·5	14·0	6·3	8·8
Block 28.	(3)	(59)	(51)	(43)	(11)	(27)	(19)	(35)
	9·3	8·0	5·0	11·5	6·0	6·8	12·3	14·3
Block 29.	(31)	(7)	(63)	(23)	(15)	(55)	(39)	(47)
	12·0	13·3	9·9	7·3	12·3	9·5	7·0	12·5
Block 30.	(44)	(20)	(4)	(12)	(28)	(52)	(60)	(36)
	6·8	10·3	10·5	11·5	7·5	8·3	9·0	5·8
Block 31.	(1)	(25)	(17)	(9)	(57)	(41)	(33)	(49)
	9·8	10·3	9·8	11·8	10·0	7·0	9·0	6·8
Block 32.	(26)	(34)	(2)	(58)	(42)	(18)	(50)	(10)
	10·0	5·5	10·5	9·5	6·0	9·3	11·3	10·3

replications, such as the lattice design in Table 13.4, it is perfectly legitimate to carry out the straightforward analysis appropriate to a simple randomized block design, ignoring the incomplete blocks within replications. The conclusions drawn from the corresponding tests of significance are valid. The varietal comparisons based on such analysis would of course not be generally as accurate as those based on the exact analysis and may be resorted to only in an emergency arising from the need of saving labour or from the occurrence of missing plots.

The correct analysis of the lattice design is explained below:

13*b*.4 ANALYSIS OF A SIMPLE LATTICE DESIGN

Example 13.4

The data for illustrating the procedure of analysis of a simple or double lattice design pertain to an experiment carried out at Coimbatore to compare the yield of 64 strains of paddy. The entire design, consisting of X and Y groups of sets, was repeated twice so that there were in all 4 replications. The plot-wise yields of paddy are shown in Table 13.4.

It will be seen from the table that blocks in replications I and III are formed from row sections of the square array of the 64 varieties, while replications II and IV consist of blocks formed from column sections. We shall designate the block in replication I corresponding to row section 1 (varieties 1 to 8) as X_1',

that corresponding to row section 2 (varieties 9 to 16) as X_2' and so on; similarly the corresponding blocks in replication III will be named as X_1'', X_2'', ..., X_8''. The blocks corresponding to column sections 1, 2, ... 8 in replication II will be designated Y_1', Y_2', ..., Y_8' and similar blocks in replication IV, Y_1'', Y_2'', ..., Y_8''.

The statistical analysis of the plot yields consists of the following steps:

1. Form a table of varietal totals, arranging the totals for convenience in a square array as in Table 13.5. The figures recorded in the margins of this table are derived as explained later.

TABLE 13.5

Varietal totals for 4 replications (unadjusted) in 8 × 8 simple lattice

									Total	μC
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)		
	31·6	43·6	33·8	40·0	53·0	35·0	38·1	40·4	315·5	+1·50
	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)		
	42·2	45·6	25·8	40·8	25·9	31·3	43·6	32·9	288·1	−0·05
	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)		
	35·9	43·9	36·4	42·6	38·7	43·9	29·3	37·6	308·3	+1·01
	(25)	(26)	(27)	(28)	(29)	(30)	(31)	(32)		
	36·4	46·8	26·9	35·0	32·9	34·4	45·0	38·8	296·2	−0·16
	(33)	(34)	(35)	(36)	(37)	(38)	(39)	(40)		
	35·8	26·8	47·4	26·2	40·6	40·6	21·0	29·8	268·2	+1·49
	(41)	(42)	(43)	(44)	(45)	(46)	(47)	(48)		
	41·0	28·6	35·3	37·7	39·6	44·7	43·1	17·6	287·6	+0·41
	(49)	(50)	(51)	(52)	(53)	(54)	(55)	(56)		
	27·9	51·6	24·5	29·4	39·4	33·6	37·0	26·6	270·0	+0·13
	(57)	(58)	(59)	(60)	(61)	(62)	(63)	(64)		
	44·6	42·9	36·0	36·1	37·8	43·0	39·8	49·7	329·9	−0·82
Total	295·4	329·8	266·1	287·8	307·9	306·5	296·9	273.4	2363·8	
$\mu C'$	−0·87	−0·27	−1·04	+0·27	+0·11	−1·66	−0·31	+0·26		

2. Obtain the block totals and arrange separately the X block totals and the Y block totals in two 8×2 tables (in general $q \times 2$ tables) so that each row contains the totals of the 2 blocks having the same varieties, as indicated below:

	X Block totals			*Y* Block totals		
	Replication			Replication		
Block	I	III	Sum	II	IV	Sum
1	X_1'	X_1''	X_1	Y_1'	Y_1''	Y_1
2	.	.	.	.	.	.
3	.	.	.	.	.	.
4	.	.	.	.	.	.
5	.	.	.	.	.	.
6	.	.	.	.	.	.
7	.	.	.	.	.	.
8	X_8'	X_8''	X_8	Y_8'	Y_8''	Y_8
Grand Total ..	X'	X''	X	Y'	Y''	Y

The values obtained from the present data are shown in columns 1, 2, 3 and 6, 7, 8 of Table 13.6.

TABLE 13.6

Block totals and block corrections, 8×8 *simple lattice*

X Group of blocks

	Replication I X_i'	Replication III X_i''	Sum X_i	C_i	Block correction μC_i
	1	2	3	4	5
	71·3	72·6	143·9	+27·7	+1·50
	71·0	73·5	144·5	− 0·9	−0·05
	76·2	68·6	144·8	+18·7	+1·01
	74·9	74·7	149·6	− 3·0	−0·16
	65·9	54·4	120·3	+27·6	+1·49
	78·3	61·7	140·0	+ 7·6	+0·41
	63·1	70·7	133·8	+ 2·4	+0·13
	85·2	87·3	172·5	−15·1	−0·82
Total ..	585·9	563·5	1149·4	+65·0 R_c	+3·51

TABLE 13.6—(*Continued*)

Y Group of blocks

	Replication II Y_i'	Replication IV Y_i''	Sum Y_i	C_i'	Block correction $\mu C_i'$
	6	7	8	9	10
	81·3	74·5	155·8	−16·2	−0·87
	95·0	72·4	167·4	− 5·0	−0·27
	70·5	72·2	142·7	−19·3	−1·04
	71·7	69·7	141·4	+ 5·0	+0·27
	76·5	76·4	152·9	+ 2·1	+0·11
	86·2	82·4	168·6	−30·7	−1·66
	67·5	83·8	151·3	− 5·7	−0·31
	76·3	58·0	134·3	+ 4·8	+0·26
Total ..	625·0	589·4	1214·4	−65·0 R_c'	−3·51

3. From each of the tables for X and Y block totals, obtain the sum of squares corresponding to rows and columns interaction and add. This is equivalent to the application of the following formula, when the simple lattice for q^2 varieties is repeated p times ($p = 2$ in the present case):

$$\frac{(X_1'^2 + X_1''^2 + \ldots . X_8'^2 + X_8''^2) + (Y_1'^2 + Y_1''^2 + \ldots . + Y_8'^2 + Y_8''^2)}{q = 8}$$

$$- \frac{(X_1^2 + X_2^2 + \ldots . + X_8^2) + (Y_1^2 + Y_2^2 + \ldots . Y_8^2)}{pq = 2 \times 8 = 16}$$

$$- \frac{(X'^2 + X''^2) + (Y'^2 + Y''^2)}{q^2 = 64} + \frac{X^2 + Y^2}{pq^2 = 2 \times 64 = 128} \qquad (13.4)$$

for d.f. $2(p - 1)(q - 1) = 2(1)(7) = 14$

On substituting in this formula the X and Y values from Table 13.6 we obtain the sum of squares as

$$\frac{(83600\cdot58 + 93282\cdot56)}{8} - \frac{(166665\cdot44 + 185386\cdot20)}{16}$$

$$- \frac{(660811\cdot06 + 738017\cdot36)}{64} + \frac{(1321120\cdot36 + 1474767\cdot36)}{128}$$

$= 93\cdot34$ for 14 d.f.

A little reflection will show that the sum of squares so obtained is representative of block differences from which both the replication differences and varietal differences have been eliminated. We designate this sum of squares as component (*a*) of the block sum of squares adjusted for varietal differences. Note that this component exists only when the design is repeated.

4. Altogether, however, there are $4\times(8-1) = 28$ degrees of freedom [in general, $2p(q-1)$ degrees of freedom] for blocks within replications. The remaining component, designated component (*b*) of block sum of squares has to be calculated from the differences among X totals themselves for 7 degrees of freedom and Y totals themselves for another 7 degrees of freedom. Before calculating the sum of squares, however, allowance has to be made for the fact that the different sets of blocks contain different varieties. The adjustment involves the calculation of quantities C and C', shown in columns 4 and 9 of Table 13.6, by means of the following formulæ:

C_i = [Total over all replications of varieties occurring in the ith block of X group]—$2X_i$

and C_i' = [Total over all replications of varieties occurring in the ith block of Y group] $- 2Y_i$ (13.5)

For example,

$$C_1 = (v_1 + v_2 + \ldots . v_8) - 2X_1$$
$$= 315\cdot5 - 2(143\cdot9) = +27\cdot7$$

$$C_1' = (v_1 + v_9 + v_{17} + v_{25} + v_{33} + v_{41} + v_{49} + v_{57}) - 2Y_1$$
$$= 295\cdot4 - 2(155\cdot8) = -16\cdot2$$

Let $(C_1 + \ldots + C_8)$ be denoted by R_C, $(C_1' + \ldots + C_8')$ be denoted by R_C'. Then check $R_C + R_C' = 0$.

5. Calculate the sum of squares of deviations of each set of C's among themselves, add and divide by $2qp$, or more directly, apply the following formula:

Component (b) of block sum of squares (adjusted)

$$= \frac{\Sigma C_i^2}{2qp} - \frac{\Sigma Rc^2}{2q^2p} \qquad (13.6)$$

· for $2(q-1)$ d.f.

the summation being taken over both the groups. In the present case the formula reduces to

$$\frac{(C_1^2 + \ldots . C_8^2) + (C_1'^2 + C_2'^2 + \ldots . C_8'^2)}{2(8)(2)} - \frac{Rc^2 + Rc'^2}{2(64)(2)}$$

$$= \frac{3867 \cdot 44}{32} - \frac{8450 \cdot 00}{256}$$

$$= 87 \cdot 85 \text{ for } 2 \times 7 = 14 \text{ d.f.}$$

6. Add the two components (a) and (b) to obtain the sum of squares for blocks (adjusted) for $2(p-1)(q-1) + 2(q-1) = 2p(q-1)$ degrees of freedom, in the present case $4(8-1) = 28$ degrees of freedom, which is precisely the number of degrees of freedom for blocks within replications. The value of this sum of squares in the present example is $93 \cdot 34 + 87 \cdot 85 = 181 \cdot 19$.

7. Calculate the total sum of squares, sum of squares between replications and sum of squares between varieties (unadjusted) in the usual way and complete the analysis of variance table by inserting the sum of squares for blocks (adjusted) obtained above and obtaining by subtraction the sum of squares for intra-block error. The analysis of variance is given in Table 13.7.

8. The mean squares for blocks and error are designated respectively as E_b and E_e. If E_b happens to be less than E_e, pool the adjusted block sum of squares with error sum of squares and proceed with the analysis as if the experiment were laid out in simple randomized blocks with $2p$, (here 4), replications.

TABLE 13·7

Analysis of variance of simple lattice design in Example 13.4 *(oz./plot)*

Source of variation	D.F. General case (q^2 varieties p repetitions)	D.F. Present example	S.S.	M.S.	
Replications	$(2p-1)$	3	30·33	..	
Varieties (unadjusted) ..	(q^2-1)	63	868·63	..	
Blocks (adjusted)	$2p\,(q-1)$	28	181·19	6·47	E_b
Error (intra block) ..	$(q-1)\,(2pq-q-1)$	161	485·20	3·01	E_e
Total ..	$2pq^2-1$	255	1565·35	..	

If, as will usually be the case, E_b is greater than E_e, proceed to calculate μ, 'the weighting factor' to be used for adjusting the varietal totals for the block effects, by means of the following formula:

$$\mu = \frac{p\,(E_b - E_e)}{q\,[pE_b + (p-1)\,E_e]} \qquad (13.7)$$

which in the present case gives a value

$$\mu = \frac{2\,(E_b - E_e)}{8\,(2E_b + E_e)} = 0{\cdot}054$$

9. Multiply each C and C' value by μ to obtain block corrections μC and $\mu C'$ as shown in columns 5 and 10 of Table 13.6. Check that they add to zero, when taken over both the groups, except for errors of approximation.

10. Adjust each varietal total by adding a correction term to the unadjusted total for every group of blocks in which the variety occurs; for example, variety 29 occurs in the groups of blocks corresponding to C_4 and C_5'; the unadjusted varietal total 32·9 is therefore corrected by adding μC_4 and $\mu C_5'$, that is, $-0{\cdot}16$ and $+0{\cdot}11$ to obtain 32·85 as the adjusted varietal total. It is convenient to effect the corrections by putting down the correction terms in the marginal column and row in the table

of the unadjusted varietal totals itself as shown in Table 13.5. The correction then consists of merely adding to each unadjusted varietal total the two correction terms occurring in the same row and column of the square array as the variety. The adjusted varietal totals are shown in Table 13.8. The adjusted varietal means are simply calculated from the adjusted totals by dividing by the number of replications, 4 in the present case.

TABLE 13.8

Adjusted varietal totals, 8×8 simple lattice

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
32·2	44·8	34·3	41·8	54·6	34·8	39·3	42·2
(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
41·3	45·3	24·7	41·0	26·0	29·6	43·2	33·1
(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
36·0	44·6	36·4	43·9	39·8	43·2	30·0	38·9
(25)	(26)	(27)	(28)	(29)	(30)	(31)	(32)
35·4	46·4	25·7	35·1	32·8	32·6	44·5	38·9
(33)	(34)	(35)	(36)	(37)	(38)	(39)	(40)
36·4	28·0	47·8	28·0	42·2	40·4	22·2	31·0
(41)	(42)	(43)	(44)	(45)	(46)	(47)	(48)
40·5	28·7	34·7	38·4	40·1	43·4	43·2	18·3
(49)	(50)	(51)	(52)	(53)	(54)	(55)	(56)
27·2	51·5	23·6	29·8	39·6	32·1	36·8	27·0
(57)	(58)	(59)	(60)	(61)	(62)	(63)	(64)
42·9	41·8	34·1	35·0	37·1	40·5	38·7	49·1

11. The analysis of variance set out in Table 13.7 does not supply directly a test of significance for the overall differences between adjusted varietal means. No exact test for these differences is in fact available owing to the complication that two variances, the intrablock error E_e between plot to plot within blocks and the interblock error E_b between blocks, are involved. An approximate F-test can, however, be carried out as follows:

Compute the unadjusted sum of squares for component (*b*) of blocks which is the total of row sum of squares for each group X and Y, and call it B_u. Thus

$$B_u = \frac{(X_1^2 + X_2^2 + \ldots . X_8^2) + (Y_1^2 + \ldots . Y_8^2)}{pq = 16} - \frac{(X^2 + Y^2)}{pq^2 = 128}$$

$$= 160{\cdot}36 \qquad (13.8)$$

Then calculate the quantity

$$q\mu \left[\frac{2}{(1 + q\mu)} B_u - B_b \right] \qquad (13.9)$$

where B_b is the adjusted sum of squares for component (*b*) of blocks, 87·85 in the present case as calculated earlier. Since $q = 8$, $\mu = 0{\cdot}054$, and $B_u = 160{\cdot}36$, this quantity has the value 58·80. This quantity is to be subtracted from the unadjusted sum of squares for varieties, namely, 868·63 shown in Table 13.7.

The sum of squares for varieties adjusted for the approximate F-test is therefore 868·63 − 58·80 = 809·83.

Calculate the varietal mean square and compare it with the intrablock error mean square to give the ratio F as indicated in Table 13.9.

TABLE 13.9

Test of significance for adjusted varietal differences, 8×8 simple lattice

Source of variation		*D.F.*	*S.S.*	*M.S.*	*F*
Varieties (adjusted)	..	63	809·83	12·85	4·27**
Error E_e (Intrablock)	..	161	485·20	3·01	..

We conclude that the varieties on the whole show highly significant differences.

12. We next turn to the comparison of individual differences between the adjusted varietal means. This is complicated by the fact that the precision of comparisons between varieties which occur together in a block, as for instance v_1 and v_9 which occur

together in blocks 16 and 31 (Table 13.4), is greater than that of comparisons between varieties which are never together in a block, as for instance v_1 and v_{10}. Two standard errors have accordingly to be calculated. The corresponding variances are given by the following expressions:

V_1 = Variance of difference between adjusted means of two varieties occurring together in a block, *e.g.*, v_1 and v_9

$$= \frac{2}{r} E_e [1 + \mu] \qquad (13.10)$$

where r is the number of replications

$$= \frac{2}{4} (3{\cdot}01) [1 + 0{\cdot}054]$$

$= 1{\cdot}59$ in the present example.

V_2 = Variance of difference between adjusted means of two varieties not occurring together in a block anywhere in the design, *e.g.*, v_1 and v_{10}

$$= \frac{2}{r} E_e [1 + 2\mu] \qquad (13.11)$$

$$= \frac{2}{4} (3{\cdot}01) [1 + (2)(0{\cdot}054)]$$

$= 1{\cdot}67$ in our example.

Except in the case where the number of varieties is small, it is usually sufficient to use an average value of the variance for varietal comparisons. The average variance can be directly worked out by means of the following formula:

$$V = \frac{2E_e}{r} \left[1 + \frac{2q\mu}{q+1}\right], \; r = 2p \qquad (13.12)$$

which reduces to the value $1{\cdot}65$ in our example. The average critical difference between any two adjusted varietal means is therefore $\sqrt{1{\cdot}65} \times 1{\cdot}96 = 2{\cdot}52$.

13. Lastly, we may wish to assess the gain in efficiency resulting from the adoption of the incomplete block design in preference to the randomized block design. For this purpose, we calculate a quantity designated the *effective error mean square* s^2 for the incomplete block design, such that $2s^2/r$ gives the average variance for the difference of two adjusted varietal means. s^2 is therefore given by the expression

$$E_e\left[1 + \frac{2q\mu}{q+1}\right] \qquad (13.13)$$

and is equal to 3·10 in the present case. Next, we estimate the error mean square s'^2 which we would have got, had we adopted a simple randomized block design. This is done simply by pooling the adjusted sum of squares for blocks within replications with the error sum of squares. In the present case,

$$s'^2 = \frac{181{\cdot}19 + 485{\cdot}20}{28 + 161} = 3{\cdot}53$$

The relative efficiency of the lattice design in comparison with the simple randomized block design is given by the ratio s'^2/s^2. In the present example,

$$\frac{s'^2}{s^2} = \frac{3{\cdot}53}{3{\cdot}10} = 1{\cdot}14 \text{ or } 114 \text{ per cent.}$$

which corresponds to a gain of 14 per cent. in efficiency for the simple lattice design.

13*b*.5 Concept of Balancing

As we have seen in the foregoing sections, the complications which arise in the analysis of the simple lattice design are mainly due to the fact that adjustments have to be made to allow for the block differences in estimating varietal differences and the part of the information concerning varietal comparisons which is contained in the block differences has to be extracted. A further complication results from the fact that the precision of the different varietal comparisons is not the same. Varieties which occur in the same block are naturally compared with greater precision than those which do not occur together in a block.

This necessitates the computation of separate standard errors for the different comparisons.

This last complication can be eliminated if the incomplete block design adopted is such that every pair of varieties occurs together in a block the same number of times. In that case the precision of all the varietal comparisons is the same. Such designs are known as *balanced incomplete block designs.* In addition to simplifying the statistical analysis, these designs possess the advantage of increased overall precision on account of symmetry as compared with unsymmetrical incomplete block designs like the simple lattice with the same number of replications. The disadvantage of the balanced design is that it generally requires many more replications than can be afforded by the experimenter. Besides, balanced designs are available only for those numbers of varieties which satisfy certain conditions. It is not possible to discuss in this book the various types of balanced designs which are available. Reference on this point should be made to sources like Cochran and Cox (1950). We shall confine ourselves to a simple extension of the double lattice design which provides a balanced design known as the *balanced lattice.*

13*b*.6 THE BALANCED LATTICE

A balanced lattice design is possible only when the square-root of the number of varieties is either a prime number, that is, a number which is not divisible by any other number, for example, 3, 5, 7, 11, 13, ..., or a power of a prime, such as $4 = 2^2$, $9 = 3^2$, $8 = 2^3$ and so on. Thus a balanced lattice design can be had for $v = 16, 25, 49, 64, 81, \ldots$, but not for 36 or 100 since $\sqrt{36} = 6$ or $\sqrt{100} = 10$ are not prime numbers.

The construction of a balanced lattice for q^2 varieties requires the aid of special sets of latin squares of side q, called mutually *orthogonal latin squares.* The property of orthogonality of two latin squares can be best illustrated by the following example of 3×3 latin squares:

Square 1			Square 2		
A	*B*	*C*	*A'*	*B'*	*C'*
B	*C*	*A*	*C'*	*A'*	*B'*
C	*A*	*B*	*B'*	*C'*	*A'*

On superimposing square 2 over square 1, we get the following arrangement:

AA'	BB'	CC'
BC'	CA'	AB'
CB'	AC'	BA'

It will be observed that the letter A' of the second latin square occurs with letters A, B, C, each exactly once. So also B' and C' occur with A, B and C each once only. When in a pair of latin squares each letter of one latin square occurs just once with each letter of the other latin square the two latin squares are said to be orthogonal.

If a set of latin squares is such that any two latin squares in the set are orthogonal to each other, they are said to constitute a mutually orthogonal set. The following latin squares of side 4 are mutually orthogonal:

Square 1				Square 2				Square 3			
A_1	B_1	C_1	D_1	A_2	B_2	C_2	D_2	A_3	B_3	C_3	D_3
B_1	A_1	D_1	C_1	C_2	D_2	A_2	B_2	D_3	C_3	B_3	A_3
C_1	D_1	A_1	B_1	D_2	C_2	B_2	A_2	B_3	A_3	D_3	C_3
D_1	C_1	B_1	A_1	B_2	A_2	D_2	C_2	C_3	D_3	A_3	B_3

The maximum number of mutually orthogonal latin squares of side q, available only if q is either a prime number or a power of a prime, is $(q - 1)$. The two latin squares of side 3 thus constitute the full set of mutually orthogonal latin squares of that size and the three latin squares of side 4 given above also constitute the complete set. The complete sets of orthogonal latin squares of sides 3, 4, 5, 7, 8 and 9 are given in Table XVI of Fisher and Yates' *Statistical Tables* (1948).

The use of the sets of orthogonal latin squares in obtaining groups of sets of varieties for a lattice design will be illustrated in the case of 25 varieties. For this purpose we require the following four orthogonal latin squares of side 5, reproduced from Fisher and Yates' *Statistical Tables*:

Square 1

1	2	3	4	5
2	3	4	5	1
3	4	5	1	2
4	5	1	2	3
5	1	2	3	4

Square 2

1	2	3	4	5
3	4	5	1	2
5	1	2	3	4
2	3	4	5	1
4	5	1	2	3

Square 3

1	2	3	4	5
4	5	1	2	3
2	3	4	5	1
5	1	2	3	4
3	4	5	1	2

Square 4

1	2	3	4	5
5	1	2	3	4
4	5	1	2	3
3	4	5	1	2
2	3	4	5	1

The first two groups or replications, *viz.*, *X* and *Y* groups of sets, are obtained as in the case of simple lattice by means of row sections and column sections of the square array in which the variety numbers are arranged.

To obtain the third group of sets of varieties, the square array of variety numbers is superimposed on the latin square 1 above, and the different sets are formed by putting together in a set, varieties which occur with the same letter or number of the latin square. The variety numbers superimposed on square 1 are shown below in brackets against the latin square numbers:

1 (1)	2 (2)	3 (3)	4 (4)	5 (5)
2 (6)	3 (7)	4 (8)	5 (9)	1 (10)
3 (11)	4 (12)	5 (13)	1 (14)	2 (15)
4 (16)	5 (17)	1 (18)	2 (19)	3 (20)
5 (21)	1 (22)	2 (23)	3 (24)	4 (25)

We thus derive the following sets belonging to the third replication or group:

Set 1: 1, 10, 14, 18, 22.

Set 2: 2, 6, 15, 19, 23.

Set 3: 3, 7, 11, 20, 24.

Set 4: 4, 8, 12, 16, 25.

Set 5: 5, 9, 13, 17, 21.

In the same manner we derive the other groups, each corresponding to one of the orthogonal latin squares. With 25 varieties, there will be in all $5 + 1 = 6$ groups, possible, 2 of them being derived from the row and column sections of the square array of the varieties and the remaining $4 = (5 - 1)$ from the four orthogonal latin squares. In general, if q is a prime number or a power of a prime, we can have with $v = q^2$ varieties $(q + 1)$ groups, 2 from row and column sections of the varietal array and $(q - 1)$ from the $(q - 1)$ mutually orthogonal latin squares. We can take any number of these groups to derive a corresponding lattice design such as double, triple, etc. If all the $(q + 1)$ groups are taken we shall obtain a balanced lattice. The 6 groups of sets for the balanced lattice design for 25 varieties are shown below for illustration:

Group I (Row Section)	Group II (Column Section)
Set (1) 1, 2, 3, 4, 5.	Set (1) 1, 6, 11, 16, 21.
(2) 6, 7, 8, 9, 10.	(2) 2, 7, 12, 17, 22.
(3) 11, 12, 13, 14, 15.	(3) 3, 8, 13, 18, 23.
(4) 16, 17, 18, 19, 20.	(4) 4, 9, 14, 19, 24.
(5) 21, 22, 23, 24, 25.	(5) 5, 10. 15, 20, 25.
Group III (Latin sq. 1)	**Group IV (Latin sq. 2)**
Set (1) 1, 10, 14, 18, 22.	Set (1) 1, 9, 12, 20, 23.
(2) 2, 6, 15, 19, 23.	(2) 2, 10, 13, 16, 24.
(3) 3, 7, 11, 20, 24.	(3) 3, 6, 14, 17, 25.
(4) 4, 8, 12, 16, 25.	(4) 4, 7, 15, 18, 21.
(5) 5, 9, 13, 17, 21.	(5) 5, 8, 11, 19, 22.
Group V (Latin sq. 3)	**Group VI (Latin sq. 4)**
Set (1) 1, 8, 15, 17, 24.	Set (1) 1, 7, 13, 19, 25.
(2) 2, 9, 11, 18, 25.	(2) 2, 8, 14, 20, 21.
(3) 3, 10, 12, 19, 21.	(3) 3, 9, 15, 16, 22.
(4) 4, 6, 13, 20, 22.	(4) 4, 10, 11, 17, 23.
(5) 5, 7, 14, 16, 23.	(5) 5, 6, 12, 18, 24.

It may be verified that every pair of varieties occurs together in the same block exactly once in the whole design. For example, reference to the first set in each group will show

that variety 1 occurs once only with each of the remaining 24 varieties in the 6 blocks in which it is included. This is what constitutes balance.

As mentioned before, the sets constituting each group should be allotted at random to contiguous compact blocks to form a compact replication and the varieties in each set assigned at random to the plots of each block.

13*b*.7 Analysis of a Balanced Lattice Design

Example 13.5

We next illustrate the analysis of a balanced lattice with the help of data of a paddy varietal trial with 16 varieties conducted at Nagina Research Station in Uttar Pradesh. The plan and yields are given in Table 13.10. The statistical

TABLE 13.10

Plan and yields of 4×4 *balanced lattice*

Yield in lb. per plot of size 1/40 *acre. Experiment with paddy varieties at Nagina (U.P.)*

Replication I.	Block 1		Block 2	
	v_{11}	33	v_{13}	22
	v_1	25	v_4	29
	v_{16}	36	v_7	24
	v_6	32	v_{10}	22
	Block 3		Block 4	
	v_2	36	v_9	41
	v_{15}	13	v_8	29
	v_{12}	31	v_{14}	36
	v_5	15	v_3	25
Replication II.	Block 5		Block 6	
	v_3	18	v_{10}	18
	v_{13}	41	v_8	21
	v_6	23	v_1	14
	v_{12}	39	v_{15}	44
	Block 7		Block 8	
	v_5	33	v_9	27
	v_4	40	v_7	21
	v_{14}	25	v_2	30
	v_{11}	26	v_{16}	31

TABLE 13.10—(*Continued*)

Replication III.	Block 9		Block 10	
	v_4	19	v_1	19
	v_9	33	v_{12}	36
	v_{15}	20	v_{14}	30
	v_6	21	v_7	17
	Block 11		Block 12	
	v_{16}	36	v_2	29
	v_5	19	v_{13}	28
	v_3	13	v_8	31
	v_{10}	19	v_{11}	40
Replication IV.	Block 13		Block 14	
	v_{15}	50	v_{12}	55
	v_7	32	v_4	22
	v_3	23	v_8	23
	v_{11}	38	v_{16}	39
	Block 15		Block 16	
	v_1	15	v_{10}	34
	v_9	27	v_6	20
	v_{13}	19	v_{14}	28
	v_5	39	v_2	24
Replication V.	Block 17		Block 18	
	v_{14}	33	v_3	35
	v	41	v_1	17
	v_{15}	31	v_2	21
	v_{16}	27	v_4	18
	Block 19		Block 20	
	v_9	14	v_7	32
	v_{12}	38	v_5	11
	v_{11}	9	v_8	37
	v_{10}	25	v_6	16

analysis, which will be seen to be much more simple than that of the double lattice is as follows:

1. Prepare a two-way table of blocks and varieties as in Table 13.11. Form the block totals and replication totals as in the last rows. Next obtain the varietal totals V_j and simultaneously obtain the quantities B_j by summing the totals of all blocks which contain v_j, for example,

$$B_1 = 126 + 97 + 102 + 100 + 91 = 516$$

Block / Variety	REPLICATE I				II				III				IV				V				Total	
	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_{10}	B_{11}	B_{12}	B_{13}	B_{14}	B_{15}	B_{16}	B_{17}	B_{18}	B_{19}	B_{20}	V_j	B_j
v_1	25					14				19					15			17			90	516
v_2			36					30				29				24		21			140	529
v_3				25	18						13		23					35			114	573
v_4		29					40		19					22				18			128	544
v_5			15				33				19				39					11	117	502
v_6	32				23				21							20				16	112	542
v_7		24						21		17			32							32	126	547
v_8				29		21						31		23						37	141	591
v_9				41				27	33						27				14		142	519
v_{10}		22				18					19					34			25		118	473
v_{11}	33						26					40	38						9		146	607
v_{12}			31		39					36				55					38		199	543
v_{13}		22			41							28			19		41				151	578
v_{14}				36			25			30						28	33				152	595
v_{15}			13			44			20				50				31				158	560
v_{16}	36							31			36			39			27				169	593
Total Replication	126	97	95	131	121	97	124	109	93	102	87	128	143	139	100	106	132	91	86	96	2203	8812

Check that the B_j values all add to 4 times (in general q times) the grand total G.

2. Next work out the quantities

$$W_j = qV_j - (q+1)B_j + G \qquad (13.14)$$

for all varieties. Check that $\Sigma W_j = 0$. In the present example

$$W_j = 4V_j - 5B_j + G$$

W_j's are given in column 4 of Table 13.12.

TABLE 13.12

Varietal totals and adjusted varietal means (in lb. per plot)
4×4 balanced lattice

Variety number	Unadjusted varietal total $= V_j$	Sum of totals of blocks containing v_j $= B_j$	$4V_j - 5B_j + G = W_j$	Adjusted varietal total $= V_j + \mu W_j$	Adjusted varietal mean
1	2	3	4	5	6
1	90	516	− 17	89·80	17·96
2	140	529	+118	141·37	28·27
3	114	573	−206	111·61	22·32
4	128	544	− 5	127·94	25·59
5	117	502	+161	118·87	23·77
6	112	542	− 59	111·32	22·26
7	126	547	− 28	125·68	25·14
8	141	591	−188	138·82	27·76
9	142	519	+176	144·04	28·81
10	118	473	+310	121·60	24·32
11	146	607	−248	143·12	28·62
12	199	543	+284	202·29	40·46
13	151	578	− 83	150·04	30·01
14	152	595	−164	150·10	30·02
15	158	560	− 35	158·41	31·68
16	169	593	− 86	168·00	33·60
Total ..	2203	8812	0	2203·01	..

3. Calculate the sum of squares for blocks within replications, adjusted for varietal effects, from the formula

$$\frac{\sum_{1}^{q^2} W_j^2}{q^3 (q+1)} \qquad (13.15)$$

In the present case we obtain

$$\frac{(-17)^2 + \ldots\ldots + (-86)^2}{320} = 1355{\cdot}77$$

4. Calculate the total sum of squares, the sum of squares for replications and sum of squares for varieties (unadjusted for block effects) and complete the table of analysis of variance as shown in Table 13.13, working out the intrablock error sum of squares by subtraction in the usual way.

TABLE 13.13

Analysis of variance of 4×4 *balanced lattice*

Source of variation	*D.F.*	*S.S.*	*M.S.*
Replications	$4 = (q)$	289·33	72·33
Varieties (unadjusted) ..	$15 = (q^2 - 1)$	2035·89	135·72
Blocks within replications (adjusted)	$15 = (q^2 - 1)$	1355·77	90·38 (E_b)
Error (Intrablock)	$45 = (q-1)(q^2-1)$	3312·90	73·62 (E_e)
Total ..	$79 = q^2(q+1)-1$	6993·89	..

5. From the table of the analysis of variance we obtain estimates of plot to plot variation within blocks, that is, intrablock error designated E_e and also of the error variation between blocks within replications designated E_b to which the interblock components of varietal comparisons are subject. We calculate from these, the ' adjustment factor '

$$\mu = \frac{E_b - E_e}{q^2 \cdot E_b} \qquad (13.16)$$

In the present example we have

$$\mu = \frac{90 \cdot 38 - 73 \cdot 62}{16 \times 90 \cdot 38} = 0 \cdot 0116$$

6. The varietal total for V_j adjusted for block differences and including the information supplied by interblock differences, is given by

$$(V_j + \mu \cdot W_j) \qquad (13.17)$$

These values are given in column 5 of Table 13.12. Dividing the adjusted varietal total by the number of replications r, here 5, we obtain the adjusted varietal mean as given in the last column of the same table.

It should be noted that if E_b turns out to be less than E_e, μ is taken as zero, no adjustment is made to the varietal totals and the test of significance is carried out as in simple randomized block design, pooling the sum of squares for blocks within replications with error sum of squares which was the procedure recommended in the analysis of the simple lattice also.

7. For testing the overall differences between varieties an approximate F-test can be made. We calculate the sum of squares of deviations of the adjusted varietal totals from their own mean and divide by $(q^2 - 1)(q + 1)$, here (15) (5) = 75, to obtain the mean square, 139·11.

We have next to calculate a quantity designated the effective error mean square s^2 such that $2s^2/r$, where r is the number of replications, is equal to the variance of the difference of two varietal means. This is obtained from the formula

$$s^2 = E_e (1 + q\mu) \qquad (13.18)$$

and comes to

$$73 \cdot 62 (1 + 0 \cdot 0464) = 77 \cdot 04$$

in our example.

The F-test is carried out by taking the ratio of the adjusted varietal mean square to the effective error mean square and referring to the value of F for $n_1 = (q^2 - 1)$ and $n_2 = (q - 1) \times (q^2 - 1)$ degrees of freedom. In the present case the ratio is

$139 \cdot 11/77 \cdot 04 = 1 \cdot 81$ for $n_1 = 15$ and $n_2 = 45$ degrees of freedom and is not significant.

8. If the varietal differences are significant as a whole, we turn to the comparison of individual varietal differences for which we require the standard error of the difference of two adjusted varietal means. The variance of the difference is given by

$$V = \frac{2E_e}{r}(1 + q\mu) \qquad (13.19)$$

In the present example we have

$$V = \frac{2}{5} \times 77 \cdot 04 = 30 \cdot 81$$

The square-root of the variance, $5 \cdot 55$ in the present case, gives the standard error required.

9. Lastly, we may wish to compare the efficiency of the balanced lattice design with that of the simple randomized block design. An estimate of the error variance s'^2 that would have been obtained had the experiment been arranged in ordinary randomized blocks is provided by pooling the intrablock error sum of squares with the sum of squares for blocks within replications and dividing by the pooled degrees of freedom. Thus in our example

$$s'^2 = \frac{1355 \cdot 77 + 3312 \cdot 90}{15 + 45} = \frac{4668 \cdot 67}{60}$$

$$= 77 \cdot 81$$

This has to be compared with s^2, the effective error mean square for the balanced lattice, here $77 \cdot 04$. The ratio s'^2/s^2 turns out in the present case to be $1 \cdot 01$ which means a gain in precision of just one per cent. by adopting the balanced lattice in place of the simple randomized block layout.

13*b*.8 The Use of Incomplete Block Designs

Incomplete block designs are recommended for trials involving a large number of varieties. The usefulness of this type of design for a given experiment depends, however, upon not only the

number of varieties included but also on the size of the plot and the degree of heterogeneity of the experimental field. It is only when the full replicate becomes unduly large in size and is consequently rendered heterogeneous that a distinct gain is to be expected by subdividing the replicate into relatively more homogeneous blocks. In progeny row trials, the size of the plot is small, being generally a short length of a single row, and inclusion of even a large number of progenies in the experiment might not make the block too large to be efficient. Results of the study of uniformity trials on cotton at the Institute of Plant Industry, Indore, have shown that for plot sizes consisting of a single row 3 to 18 feet long incomplete block designs were not more efficient than simple randomized blocks even with upto 200 progenies. With larger plot sizes there was a moderate gain in efficiency from the use of these designs, which may thus be recommended for trials with the first stage of progeny bulks or small bulks as they are called.

The choice of a particular incomplete block design should among other considerations depend upon the availability of land and other experimental facilities. The balanced lattice, although simpler to analyse and relatively the most efficient among these designs, requires a minimum of $q + 1$ replicates and corresponding areas of land. Besides, in the earlier stages of plant breeding trials the quantity of seed of different varieties available for the experiment is usually limited. For these reasons the balanced lattice would generally be suitable when the number of varieties is less than 50. For a larger number of varieties, the double lattice is to be preferred owing to the limited replication required, although its statistical analysis is more complicated.

CHAPTER XIV
GROUPS OF EXPERIMENTS

14*a*.1 The Problem

In large-scale experimental programmes it is necessary to repeat the trial of a set of treatments like varieties or manures at a number of places and in a number of seasons. The places where the trial is repeated are usually the experimental stations located in the tract. The aim of repetitions is to study the susceptibility of the treatment effects to place and climatic variations. More generally, the aim of repetitions is to find out treatments suitable for particular tracts in which case the trials are carried out simultaneously on a representative selection of sites. The various considerations to be taken into account in planning a programme of trials on a representative selection of cultivators' fields will be dealt with in Chapter XVI. In either case, whether the trials are repeated at research farms or on a representative selection of cultivators' fields, it becomes necessary to make a joint statistical analysis of the data by combining the results of individual trials in order to obtain general conclusions in regard to the suitability of treatments and the extent of interaction of the different treatments with seasonal and soil conditions represented by the trials. It is only necessary to point out here that the problems presented by the analysis of the data and the techniques required for their solution are practically the same in both cases. It is proposed to deal with these common problems in the present chapter.

Outwardly the problems involved in the joint analysis of a set of trials are analogous to those presented by the analysis of a replicated trial. Thus if the replications in a trial can be divided into two or more groups of replications we have a miniature representation of a set of trials. The simple technique of analysis of variance may not be valid under these circumstances if the error variances in the different groups are not of the same order and interaction of the treatments with different groups is also significantly different. As stated earlier, the object of interest in analysing a set of trials is, in general, to estimate the average responses

to given treatments and to test the consistency of these at different places and in different seasons, that is to test the interaction of the responses with the latter factors. The utility and the significance of the estimates of average responses depend on whether the response is consistent from place to place or changes with it, in other words on the absence or the presence of the interaction. The results of a set of trials may, therefore, be considered as belonging to one of the following four types: (i) the experimental errors homogeneous and the interaction absent, (ii) the errors homogeneous and the interaction present, (iii) the errors heterogeneous and the interaction absent, and (iv) the errors heterogeneous and the interaction present. It follows that the first thing to determine in combining results of a set of experiments is to know whether the experimental errors are homogeneous and if the treatment responses are consistent. We shall deal with these points in the following sections.

14*b*.1. Test of Homogeneity of Experimental Errors and Treatment Comparisons

We shall first consider the type of data under heads (i) and (ii), namely where the experimental errors are homogeneous and illustrate the steps involved in the analysis and interpretation of such data with the help of an example.

Example 14.1

A set of six trials was carried out at six centres in Rajasthan to compare four improved wheat varieties against local. At each place a 5×6 randomized block trial was carried out in the same season. The analysis of each trial individually provided an error mean square based on 20 degrees of freedom. Table 14.1 gives the mean yield in lb. per acre for each variety at each place as well as the error mean square obtained in each trial.

We seek general conclusions regarding two points: (1) The overall differences in the performance of varieties in the tract represented by these places, and (2) the susceptibility of varietal differences to variation of place.

If the magnitude of the experimental errors could be assumed to be the same at each place a simple arithmetic mean of the

TABLE 14.1

Mean yield in lb. per acre of 5 varieties of wheat at six centres

Variety of wheat	Place					
	Bhilara	Bali	Pali	Khetri	Maknera	Taluji
Punjab 8 A ..	702	1498	583	2140	1454	2262
C. 518 ..	770	1266	614	2071	1194	2187
C. 591 ..	833	1429	723	2115	1276	2137
Cwn. 13 ..	527	1557	718	2056	1098	2133
Local ..	905	1482	626	2074	1454	2399
Error M.S. ..	5776	8028	4516	9526	7056	5535

treatment means (here varietal means, the varieties being the treatments) can be taken to provide an estimate of treatment effect over the tract represented by the places and a simple analysis of variance on the two-way table showing treatment and place means would give the required information. When, however, the experimental errors differ from place to place, the simple mean and analysis of variance cannot be used, equality of error variances being an assumption basic to such analysis, and a more complex analysis by methods given later would be necessary. In order, therefore, to decide whether a simple analysis of variance is adequate for our purpose it is necessary to examine first the homogeneity of error variances. This is done by Bartlett's test described in Section 13*a*.3.

With k's equal, equations (13.1), (13.2) and (13.3) reduce respectively to

$$\bar{s}^2 = \frac{1}{n} \sum_{1}^{n} s_r^2$$

$$\chi'^2 = k \left(n \log \bar{s}^2 - \sum_{1}^{n} \log s_r^2 \right)$$

and

$$C = 1 + \frac{n+1}{3nk}$$

We have

$$\bar{s}^2 = \tfrac{1}{6}(5776 + 8028 + \ldots\ldots + 5535) = 6740$$

$$n \log \bar{s}^2 = 6\,(8{\cdot}81582) = 52{\cdot}89492 \text{ and } \Sigma \log s^2 = 52{\cdot}70984$$

Hence

$$\chi'^2 = 20\,(52{\cdot}89492 - 52{\cdot}70984) = 3{\cdot}7016$$

and

$$C = 1 + \frac{6+1}{3\times 6\times 20} = 1{\cdot}0194$$

$$\chi^2\,(5\ D.F.) = \frac{3{\cdot}7016}{1{\cdot}0194} = 3{\cdot}6311$$

This is seen from the χ^2 table to be a non-significant value. Hence the experimental errors of individual trials may be regarded as being homogeneous. We can, therefore, carry out the analysis of variance on the data presented in Table 14.1, treating it as a two-way classification in an ordinary replicated trial. The partition of the degrees of freedom will be as follows:—

Due to	*D.F.*
Varieties	4
Places	5
Varieties × Places	20
Total ..	29

Further, we can pool the errors from the individual trials and obtain a joint estimate of the error variance, $\bar{s}^2$. The combined analysis of variance is given in Table 14.2.

TABLE 14.2

Combined analysis of variance of six trials in Table 14.1

Source	*D.F.*	*S.S.*	*M.S.*	
Varieties	4	88499	22124·8	
Places	5	10764355	2152871·0	
Varieties × Places ..	20	214209	10710·4	7306·8
Pooled *Error*	120	..	6739·5	

The next step in the analysis is to test for significance the variety and place interaction. It can be shown that the expectations of the mean squares for varieties, varieties × places and pooled error in the combined analysis are given by

Source of variation	Expected value of *M.S.*
Varieties	$\sigma_e^2 + r\sigma_m^2 + rp\sigma_v^2$
Varieties × Places ..	$\sigma_e^2 + r\sigma_m^2$
Pooled Error	σ_e^2

In the above expressions σ_e^2, σ_v^2 and σ_m^2 stand for the variance ascribable to experimental error, varietal effects and varieties × places interaction component respectively, r being the number of replications and p the number of places. It can be seen that the question whether the interaction is significant, that is whether the differences between varieties tend to vary from place to place can be settled by comparing the mean square for varieties × places with the estimate of error variance by the F-test. If this mean square is found to be non-significant it means that the interaction is absent or in other words $\sigma_m^2 = 0$. If this interaction is assumed to be non-existent the sum of squares for varieties × places and the error sum of squares can be pooled and a more precise estimate of the error, σ_e^2 can be obtained for testing the significance of the varietal and place differences. If, however, σ_m^2 cannot be assumed to be zero, or in other words if the interaction is present the appropriate mean square for testing the significance of varietal as well as place differences is the mean square due to varieties × places.

The F ratio for the interaction in the present example is

$$\frac{10710 \cdot 4}{6739 \cdot 5} = 1 \cdot 59$$

This is a non-significant value, the value required for significance at 5 per cent. level being 1·66. Assuming then the interaction to be non-existent we can pool the sum of squares for varieties × places and for error and obtain a more precise estimate of σ_e^2

based on 140 degrees of freedom given by

$$\frac{20\times 10710\cdot 4 + 120\times 6739\cdot 5}{140} = 7306\cdot 8$$

Testing the varietal mean square against the pooled estimate of experimental error thus obtained we get

$$F = \frac{22124\cdot 8}{7306\cdot 8} = 3\cdot 03$$

which shows the average varietal differences over the tract to be significant, the $F_{5\%}$ value for $n_1 = 4$ and $n_2 = 140$ being 2·43.

Although in the present example the interaction mean square is found to be non-significant, we notice that the observed F ratio is appreciably larger than 1 and in fact close to the $F_{5\%}$ value. In a situation like this it is safer to assume the existence of the interaction since all comparisons are unlikely to be consistent from place to place. The appropriate mean square for testing the significance of varietal mean square will then be the mean square for varieties × place interaction. The F ratio for varieties is therefore

$$F = \frac{22124\cdot 8}{10710\cdot 4} = 2\cdot 07$$

which is seen to be non-significant at $P = 0\cdot 05$ with $n_1 = 4$ and $n_2 = 20$, the value required for significance being 2·87. The test for places might be similarly made but is generally of no importance.

The two results do not contradict each other; but bear a somewhat different interpretation. In comparison to the experimental error average differences between varieties are large enough to be significant; but these differences are not sufficiently consistent from place to place to be shown significant in comparison to the interaction mean square. This means that while variety A, say, is on the average significantly superior to another variety B, it may happen that at certain places A is not superior to B or in an extreme case may even be significantly inferior to B. It would then be risky to recommend variety A over the entire tract represented by the places at which the trials were carried out. Similar

considerations regarding seasonal variation should receive due attention while making recommendations. The analysis with respect to seasonal variation will be similar to that for places, the varieties × seasons interaction being used to test the significance of varietal differences. If the test shows the varietal differences to be non-significant it would not be desirable to recommend a variety even though the varietal differences are large enough to be significant in comparison to the experimental error. This is of particular importance under Indian conditions where the cultivator, living on a small margin has no retaining capacity, and cannot therefore afford to adopt a crop variety or treatment that is liable to turn out markedly poorer than the locally established one in some seasons notwithstanding its superior performance on the average.

14*b*.2 Heterogeneity of Interactions

The analysis given above does not provide a detailed and complete picture of the situation. The test of significance of treatments and of interactions give the significance of average differences between treatments and of their average variation from place to place; but this carries the assumption that the differences or the interaction effects for all treatments are homogeneous. On the other hand, it may happen that the treatments tried can be divided into two or more groups each with certain common characters, differences between groups being significant. It is then possible that treatments within different groups may exhibit different degrees of interaction with places, or that differences between groups may exhibit significant interaction with places. A similar situation could arise with respect to places as happens to be the case in the example under consideration. The six places, it will be seen from Table 14.1, fall into three groups of two places each, Bhilara and Pali forming one group, Bali and Maknera another, and Khetri and Taluji the remaining group, the yield differences within groups being small. We might in such a case enquire whether between group differences are significant. Such a grouping of the places is, however, of little interest unless the places within a group are geographically contiguous or have some factor in common like soil, rainfall, etc. Referring to the varieties in our example, we might be

interested in the comparison local *versus* outside varieties, in which case the 4 degrees of freedom corresponding to varietal differences will be divided thus:

	D.F.
Local *vs.* rest	1
Within rest	3

There will be a corresponding partition of the interaction (varieties × places) degrees of freedom.

	D.F.
Interaction corresponding to Local *vs.* rest ..	5
Interaction corresponding to within rest ..	15

Calculating the sums of squares and mean squares appropriate for this partition the analysis of variance takes the following form:

	D.F.	*S.S.*	*M.S.*
Varieties			
Local *vs.* rest	1	48682	48682
Within rest	3	39817	13272
Interaction			
Local *vs.* rest	5	55195	11039
Within rest	15	159014	10601

This further subdivision of the degrees of freedom has made no appreciable difference to the interaction mean square, both the components being of the same order. It is seen, however, that the mean square for varieties has undergone considerable change, the comparison local *versus* rest removing more than its share of variation. The F ratio appropriate to the testing of this mean square is obtained by dividing this mean square by the interaction mean square for 5 degrees of freedom. The value obtained is 4·41. This value is non-significant for $n_1 = 1$ and $n_2 = 5$, the value required for significance being 6·61. In spite of an increase in the value of the F ratio, the ratio could not reach the level of significance, the value required for significance being also raised on account of reduction in the number of degrees of freedom n_1 and n_2. Nevertheless, this further analysis gives an indication that the local might on the whole be superior to extraneous varieties. The illustration demonstrates the procedure to be adopted

when heterogeneity of different comparisons is suspected. The importance of separating out such effects among the treatment comparisons should be borne in mind, particularly when heterogeneity among treatments is expected on *a priori* grounds. It should be noted that the test of a single degree of freedom corresponding to a difference between two groups of treatments or between two individual treatments can be more easily carried out by tabulating the means for the two groups or the two treatments at each place, taking the differences directly and calculating from these differences the t value as in Example 4.3 of Chapter IV. The t and F tests in this case are, of course, equivalent.

14*c*.1 Test of Interaction and Method of Combining Results

We now come to the analysis of a set of experiments in which the errors vary significantly from place to place. The questions of interest are the same as before: namely the average comparisons (responses) between treatments over the tract and the variation of treatment differences or responses from place to place. As already indicated a straightforward combined analysis of variance is not appropriate to answer these questions when the error variances are not homogeneous for all trials. The appropriate method is best illustrated with the help of an example.

Example 14.2

Consider the following set of 4 trials, carried out at 4 places in which 4 varieties of wheat including local were tried. There were 3 replications at each centre, the partition of the degrees of freedom at each centre being as follows:

Source of Variation	*D.F.*
Blocks	2
Varieties	3
Error	6
Total ..	11

The yield per acre at each place for each variety as well as the error variance based on 6 degrees of freedom from 3 replications at each place is given in Table 14.3.

TABLE 14.3

Yield in lb. per acre in 4 trials

Variety of wheat	Place				
	Basi	Renawal	Ajmer	Pawata	*Total*
Punjab 9 D ..	1680	1000	2080	456	5216
C. 518 ..	1630	872	2216	336	5054
C. 591 ..	1848	936	1976	192	4952
Local ..	2264	976	1800	408	5448
Total ..	7422	3784	8072	1392	20670
Error M.S. ..	221952	55488	15552	10800	..

Applying Bartlett's test to this set of variances we find that the corrected χ^2 for 3 degrees of freedom, has a value 15·59. This is a highly significant value and demonstrates the heterogeneity of errors.

With the heterogeneous error variances, the procedure to be followed for the test of treatment differences depends on the presence or absence of the treatment × place interaction. The next step to be followed therefore consists in making a test of significance for the interaction. This is made by the method of the *weighted analysis of variance* and is explained later on. If the interaction is present, we set down the means of each treatment at each place in a two-way table and carry out a simple analysis of variance and compare the treatment mean square with the interaction mean square for the significance of treatment differences. This procedure is known as the *unweighted analysis of variance*. The simple arithmetic mean of treatment means here provides the estimate of the treatment effect over the tract. The method is approximate, because we neglect the differences in precision with which the treatment means are estimated at different places, but provides a satisfactory solution of the problem under the circumstances. When, however, the interaction is absent, the method of weighted analysis of variance itself is available for testing the treatment differences. In this method, the treatment means at each place are weighted with weights inversely proportional to the error variances of these means to obtain the estimate

of the treatment effect for the tract represented by the places. This procedure results in giving greater weight to the more precise values, that is values with smaller errors. It is essential for the success of this procedure that the error mean squares used for the calculation of the weights should be good estimates of the corresponding error variances. When they are based on 15 degrees of freedom or more they may be taken to be satisfactory from this point of view. In the above example the number of degrees of freedom for the error mean squares is too small, being only 6. Nevertheless the data will be used to illustrate the procedure of the weighted analysis of variance.

For the purpose of the analysis we first calculate the weight

$$W_i = \frac{r}{s_i^2} \qquad (14.1)$$

for each experiment, r being the number of replications and s_i^2 the corresponding error mean square on per plot basis. Using these weights we next calculate for each place the quantities $W_i P_i$ where P_i's are the place totals and for each variety the quantities $\Sigma W_i t_i$ where t_i's are means for each variety at each place. Table 14.4 gives these along with the mean yields from the previous table.

TABLE 14.4

Calculations for weighted analysis of variance

Variety of wheat	Place				
	Basi	Renawal	Ajmer	Pawata	$\Sigma W_i t_i \times 10^4$
Punjab 9 D ..	1680	1000	2080	456	6047
C. 518 ..	1630	872	2216	336	5900
C. 591 ..	1848	936	1976	192	5101
Local ..	2264	976	1800	408	5439
$W_i \times 10^4$ ..	0·135	0·541	1·929	2·778	..
P_i ..	7422	3784	8072	1392	..
$W_i P_i \times 10^4$ ..	1002	2047	15571	3867	$10^4 \times G = 22487$
S_i ..	14020100	3589056	16381632	524160	..

G in the corner of the table stands for the sum of $\Sigma W_i t_i$ over all the varieties and can be obtained both by summing the row $W_i P_i$, or the column $\Sigma W_i t_i$. The procedure serves as a check. The last row of the table (S_i) gives the columnwise crude sum of squares obtained in each column. The various items in the analysis of variance are calculated as follows:—

$$\text{Total } S.S. = \Sigma W_i S_i - C \tag{14.2}$$

where

$$C = \frac{G^2}{t \Sigma W_i}$$

t being the number of treatments (4 in our case).

$$S.S. \text{ for places} = \frac{1}{t} \Sigma (W_i P_i^2) - C \tag{14.3}$$

$$S.S. \text{ for treatments} = \frac{\Sigma (\Sigma W_i t_i)^2}{\Sigma W_i} - C \tag{14.4}$$

The sum of squares for interaction I is calculated by subtraction. Thus we have in the present case,

$$C = \frac{(22487)^2}{4 (5 \cdot 383) \times 10^4} = 2348 \cdot 4$$

$$\text{Total } S.S. = \frac{1}{10^4} \{(0 \cdot 135) (14020100) + (0 \cdot 541) (3589056) + \ldots .\} - C$$
$$= 1340 \cdot 7$$

$$S.S. \text{ for places} = \frac{1}{4 \times 10^4} \{(7422) (1002) + (3784) (2047) + \ldots . \} - C$$
$$= 1308 \cdot 0$$

$$S.S. \text{ for varieties} = \frac{1}{5 \cdot 383 \times 10^4} \{(6047)^2 + (5900)^2 + \ldots . \} - C$$
$$= 10 \cdot 5$$

$S.S.$ for interaction. $I = 22 \cdot 2$ by subtraction.

For testing the significance of interaction we have to transform the sum of squares for interaction (I) into χ^2 using the formula

$$\chi^2 = \frac{(n-4)(n-2)}{n(n+t-3)} (I) \tag{14.5}$$

where n is the number of degrees of freedom on which the error mean square is based in each experiment, $n = 6$ in the present case. If n is not the same for all experiments we take simple arithmetic average $\bar{n}$ of all values. The χ^2 so obtained is tested with

$$\frac{(p-1)(t-1)(n-4)}{(n+t-3)} \qquad (14.6)$$

degrees of freedom, p being the number of trials. The formula (14.6) may give fractional degrees of freedom ascribed to the χ^2 and the number in this case has not the usual meaning. The χ^2 value has to be referred to the table with the nearest integral number of degrees of freedom. In the present case, we have

$$\chi^2 = \frac{2 \times 4}{6 \times 7}(22{\cdot}2) = 4{\cdot}23$$

$$\text{with } \frac{3 \times 3 \times 2}{7} = 2{\cdot}57 \text{ } D.F.$$

The χ^2 values for $P = {\cdot}05$ and 2 and 3 degrees of freedom are 5·99 and 7·82 respectively. The observed value of χ^2 and hence the interaction is therefore non-significant.

The procedure just described provides an overall test of significance of interaction and is consequently open to the same criticisms as a similar test in connection with the combined analysis of variance of experiments with homogeneous errors, namely that the interaction corresponding to different sets of treatments may not be homogeneous.

When heterogeneity of interaction is suspected, the significance of any component of the interaction which is of interest, can be tested by the present procedure by considering only that part of the table which is concerned with the particular component. If such interaction comes out to be significant, the significance of the relevant treatment differences can be tested by comparing the treatment and interaction mean squares obtained from an unweighted analysis. No general test of overall treatment differences appears to be available in the absence of interaction; but we can test individual degrees of freedom by the procedure explained in the next section.

14c.2 Test of Single Degree of Freedom in the Absence of Interaction

From each trial set down the responses corresponding to the degree of freedom in which we are interested for each trial, calculate from them a weighted mean

$$\bar{x}_w = \frac{\Sigma Wx}{\Sigma W} \tag{14.7}$$

and thence calculate the quantity

$$Q = \Sigma Wx^2 - \bar{x}_w \Sigma Wx \tag{14.8}$$

and a χ^2 given by

$$\chi^2 = (p-1) + \sqrt{\frac{n-4}{n-1}} \left\{ \frac{n-2}{n} Q - \overline{p-1} \right\} \tag{14.9}$$

p and n having the same meaning as before.

The weights W are the reciprocals of the estimates of error variances of the responses. The quantity χ^2 is approximately distributed as χ^2 with $p-1$ degrees of freedom and provides a test of the interaction of response with the places. If χ^2 is found to be non-significant we proceed to calculate t given by

$$t = \frac{\text{Response}}{S.E.\ \text{response}}$$

$$= \frac{\bar{x}_w}{\sqrt{1/\Sigma W}} \tag{14.10}$$

and refer the observed value to the t table with $p(n-1)$ degrees of freedom.

We shall illustrate the procedure for testing the difference between C. 591 and the local variety in Example 14.2. In Table 14.5 are given the differences x, weights W and products Wx from each trial.

TABLE 14.5

Calculation of Q

Place	Response x	Weight W	Wx
Basi	416	$0{\cdot}0676 \times 10^{-4}$	$28{\cdot}12 \times 10^{-4}$
Renawal	40	$0{\cdot}2703 \times 10^{-4}$	$10{\cdot}81 \times 10^{-4}$
Ajmer	−176	$0{\cdot}9645 \times 10^{-4}$	$-169{\cdot}75 \times 10^{-4}$
Pawata	216	$1{\cdot}3890 \times 10^{-4}$	$300{\cdot}02 \times 10^{-4}$

$$\Sigma W = 2{\cdot}6914 \times 10^{-4}, \quad \Sigma Wx = 169{\cdot}20 \times 10^{-4}$$

$$\bar{x}_{w} = \frac{169{\cdot}20 \times 10^{-4}}{2{\cdot}6914 \times 10^{-4}} = 62{\cdot}87$$

$$Q = 10{\cdot}6811 - 62{\cdot}87\,(0{\cdot}01692) = 9{\cdot}617$$

$$\chi^2\ (3\ D.F.) = 3 + \sqrt{\frac{2}{5}}\left\{\frac{4}{6}(9{\cdot}617) - 3\right\}$$

$$= 3 + 3{\cdot}41\,(0{\cdot}6325)$$

$$= 5{\cdot}157$$

It can be seen from the table of χ^2 that the value is non-significant.

When the interaction is non-significant, we next calculate t for testing the significance of the response under consideration. We have on substituting from Table 14.5 in formula (14.10)

$$t = \frac{62{\cdot}87}{60{\cdot}95} = 1{\cdot}03$$

which is an obviously non-significant value for 24 degrees of freedom.

When interaction cannot be assumed to be absent .a simple t test of the unweighted mean difference is alone possible for establishing the significance of difference.

The above tests for a single degree of freedom are very useful and can be generally applied to a set of experiments with varying errors as well as numbers of replications. It is necessary for the

purpose to set down correctly (taking into account the number of replications) the error variances of the response x at each place. We shall give one more example to illustrate the procedure by testing the significance of the linear response in a set of trials, in which 5 levels of nitrogen were tried in randomised blocks at a number of places. When graded doses of a manure are applied to a crop, it is possible to examine the regression relationship between quantity of manure and the corresponding mean response and test the significance of this relationship by separating from the treatment degrees of freedom and sum of squares, the values for the linear, quadratic and if necessary, higher degree components of the regression curve. With 5 levels of nitrogen, the treatment sum of squares will have 4 degrees of freedom, out of which the different regression components may be separated. Here we are concerned with only the linear component with one degree of freedom.

Example 14.3

Table 14.6 gives the mean yield of seed cotton (in lb. per acre) for each level of nitrogen at each of five places as well as the error mean squares based on 41 degrees of freedom in all trials except the one at Khandwa where the number of degrees of freedom was 22.

TABLE 14.6

Yield of seed cotton in lb. per acre of Nitrogen levels

Nitrogen level	Places				
	Jalgaon	Akola	Khandwa	Indore	Surat
0 N ..	503·4	201·6	143·8	125·4	403·8
20 N ..	576·6	197·4	278·1	161·0	465·9
40 N ..	669·5	217·7	325·0	236·7	564·2
60 N ..	797·7	233·9	390·6	277·4	594·4
80 N ..	824·5	240·1	396·9	264·2	644·4
Error M.S. ..	1255·7	109·6	411·0	357·9	409·8

Bartlett's test of homogeneity of a set of variances gives $\chi^2 = 56{\cdot}870$ with 4 degrees of freedom which clearly shows the heterogeneity of errors.

The linear component of the response to nitrogen at each place can be calculated from the formula

$$b = \frac{\Sigma y (x - \bar{x})}{\Sigma (x - \bar{x})^2}$$

and its variance from the formula

$$V(b) = \frac{\text{Residual } M.S.}{\Sigma (x - \bar{x})^2}$$

Thus for Jalgaon

$$b = \frac{40 (824 \cdot 5) + \ldots . + (- 40) (503 \cdot 4)}{(40)^2 + \ldots . + (- 40)^2}$$

$$= 4 \cdot 32$$

and

$$V(b) = \frac{1255 \cdot 7}{4000} = 0 \cdot 3140$$

The linear components of response, the variances of the response as represented by the linear regression coefficient, weights W and the products bW are given in Table 14.7 for each of the five places.

TABLE 14.7

Calculations for testing homogeneity of linear components of response to nitrogen

Place	Linear component (b)	Variance of b	Weight (W) of b	$W \times b$
Jalgaon	4·32	0·3140	3·18	13·74
Akola	0·57	0·0274	36·50	20·80
Khandwa ..	3·09	0·1028	9·73	30·07
Indore	1·97	0·0895	11·17	22·00
Surat	3·05	0·1025	9·76	29·77
Total ..	13·00	..	70·34	116·38

Hence

$$\bar{b}_w = \frac{116 \cdot 38}{70 \cdot 34} = 1 \cdot 654$$

and

$$Q = 298{\cdot}27 - 1{\cdot}654\,(116{\cdot}38) = 105{\cdot}78$$

For this value of Q, we get

$$\chi^2(4\ D.F.) = 4 + \left\{\frac{35{\cdot}2}{37{\cdot}2}\,(105{\cdot}78) - 4\right\}\sqrt{\frac{33{\cdot}2}{36{\cdot}2}}$$

$$= 96{\cdot}05$$

using $\bar{n} = 37{\cdot}2$ the average value of the number of degrees of freedom. This shows the existence of the interaction of response with places, the χ^2 value being highly significant. Therefore we shall infer the significance of the mean response by taking the unweighted mean of the five values, which is 2·60, and comparing it with its standard error obtained from the deviations of the individual responses from their mean. The estimate of error variance is 0·3959 for 4 degrees of freedom.

$$\therefore \quad S.E.\ (\text{mean response}) = \sqrt{0{\cdot}3959} = 0{\cdot}629$$

The ratio

$$\frac{\text{Response}}{S.E.\ \text{of response}} = 4{\cdot}13\ (\text{with } 4\ D.F.)\ \text{is significant at 5 per cent.}$$

These methods will, it is hoped, meet the requirements of the research worker for analysing the simpler types of data from groups of experiments with which he is commonly confronted. The adoption of a common design for experiments at several places or in different seasons is essential in order to maintain the simplicity of procedure of consolidation of results and should be aimed at wherever possible. Nevertheless, it should be noted that if a set of similar experiments has been carried out with differing designs or with different numbers of replications the results can still be consolidated. For help in more complex cases reference may be made to other sources, like Cochran and Cox (1950). It is necessary to emphasise that no analytic methods should be employed mechanically without first setting down the object of the analysis and without a proper appreciation of the statistical situation involved.

CHAPTER XV

PRACTICAL CONSIDERATIONS IN FIELD EXPERIMENTATION

15*a*.1 Introduction

It has been the aim of this book to equip the research worker with statistical techniques commonly required in his work without burdening him with excessive theoretical or mathematical detail. Statistical method is a tool to be used in the elucidation of biological principles and in the solution of agricultural problems and for employing it efficiently a close understanding of the subject to which it is applied is essential. Practical considerations are as important as theoretical requirements in determining the appropriate statistical approach to a problem. These practical aspects may be considered in relation to two more or less distinct phases of agricultural experimentation. In the initial phase, attention is chiefly devoted to the testing of treatments at the experimental station. In the second phase, the treatments are tried under cultivators' conditions to verify their suitability for adoption in farming practice. In this Chapter we shall set out the more important practical considerations to be borne in mind in conducting experiments at research stations.

15*a*.2 Objectives and Scope of Experiment

Perhaps the first point that should receive attention in planing an experiment is to be clear about the specific objectives of the experiment. For example, the problem may be one of assessing the manurial value of a compost prepared by a new method. In laying out an experiment in which the compost is tried against other competing manures and a no-manure control, it is necessary to be explicit regarding the conditions under which it is proposed to determine the manurial response. Is it proposed to study the reaction of the commonly grown variety of the crop or an improved variety which may have been recently recommended to growers? Similarly, is the manure to be tested under irrigated or rainfed conditions? These and other relevant issues have to

be settled before the experiment is planned to avoid the possibility that the experiment is later found to be inadequate in scope. From this point of view it is a useful practice for the experimenter to write down the objectives of his experiment, making mention of all relevant details. Thus, for the simplest type of experiment in the above example the specification might be as follows: "To study the response of wheat (local soft variety) to the new compost, as compared to farm-yard manure and ammonium sulphate each at 40 lb. nitrogen per acre, under irrigated condition." This procedure will help the experimenter to ensure in his experiment the presence of various factors which are necessary for answering satisfactorily the questions posed.

The proper appreciation of the scope of an experiment presumes a knowledge of the agricultural background of the problem, and it is desirable that agricultural aspects of the experimental design should receive full consideration along with the statistical aspects before it is finally settled. The point may be illustrated by reference to trials involving ammonium sulphate. This fertiliser needs adequate supplies of lime in the soil for its utilisation by the crop. Trial of the fertiliser on soils deficient in lime has given rise to some prejudice against the fertiliser. In trying ammonium sulphate on such soils it is obviously necessary to study the application of lime simultaneously in order to obtain a realistic picture of the fertiliser value of ammonium sulphate.

In the choice of experimental treatments care is necessary to ensure that the treatments are really capable of achieving the declared aim of the experiment. Thus trials to study optimum doses of manures and fertilisers were often laid out in the past with levels of nitrogen upto only 30 lb. or 40 lb. nitrogen per acre and it was observed that the response to nitrogen generally continued to increase more or less uniformly with the level of manuring. From such results it is not possible to find the optimum level of manuring. The restrictions imposed on the levels of manure or fertiliser tried were sought to be justified on the grounds of economic considerations. It has to be pointed out that such considerations should not weigh in experiments laid out to determine economically optimum quantities of manures for application. Studies on this subject (Crowther and Yates, 1941; Sukhatme,

1941; Panse, 1945) show that the consideration of the economics of manuring rests on two factors—the response curve, that is, the regression curve expressing the relation between level of manuring and response of the crop, and the prevailing prices of manure and crop produce. The latter are subject to continuous fluctuation and conclusions based on any particular set of prices can have only a temporary application in practice. The response curve, on the other hand, might be presumed to be a relatively stable character, at least as long as the agricultural conditions under which the crop is grown remain unchanged. The primary objective of such experiments should therefore be the determination of the response curve. It is necessary for this purpose to try a sufficiently wide range of doses going well beyond the optimum. This might appear to require a trial with a large number of treatments; but this need not be the case. Four or five well-spaced doses will suffice to determine the shape of the response curve, over the region of interest. If, for instance, the optimum dose of nitrogen is expected to be in the neighbourhood of 40 or 50 lb. nitrogen suitable doses for trial might be 0, 30, 60 and 90 lb. or 0, 25, 50, 75 and 100 lb. nitrogen.

Another mistake sometimes committed in the choice of treatments in manurial experiments is the trial of arbitrary quantities and combination of manures instead of comparable quantities in terms of elements N, P, K, etc. Thus 10 tons of farm-yard manure per acre might be tried against 250 lb. of ammonium sulphate per acre. Taking the nitrogen content of farm-yard manure roughly as 0·5 per cent. and of ammonium sulphate as 20 per cent., the nitrogen supply from the two manures would be 112 lb. and 50 lb. per acre respectively. There are, besides, the other important elements P and K in farm-yard manure which ammonium sulphate does not contain. Comparison of the effects of the two manures has thus hardly any logical basis. The trial of manures like niciphos containing both nitrogen and phosphorus against a purely nitrogenous manure like ammonium sulphate presents a similar difficulty. If the two treatments supply equal quantities of nitrogen it would be possible to judge the effect of supplementing nitrogen with phosphate; but when even the quantities of nitrogen are not equalised, the experiment can supply little useful information.

15*a*.3 Observation Plots

Before experimenting with a number of treatments in an elaborate trial, it is frequently advisable to try the treatments on a set of observation plots, as for example when the treatments are crop varieties imported from outside. Such preliminary trials often reveal the gross unsuitability of some varieties under the conditions in the tract. It may be that they are unable to stand the new weather conditions, or are specially susceptible to pests and diseases and hence unsuitable for general adoption. Laying out of observation plots may also be useful in bringing out and solving if possible difficulties that might be involved in the application of a treatment. Thus application of a manure like molasses involves many practical difficulties and it would be desirable to settle in preliminary trials the technique of applying molasses on land, before undertaking an elaborate replicated experiment. On the other hand, when the general suitability of the treatments can be presumed, it is not only unnecessary but wasteful to lay out observation plots. For example when strains selected from local crop varieties are to be compared for their field performance it is clearly unnecessary to lay out observation plots with these strains and thereby lose one season before carrying out replicated yield comparisons.

15*a*.4 Choice of Experimental Design

The design of an experiment depends largely on the number and nature of treatments proposed to be included in the experiment. Thus, if a large number of treatments is to be tested in a factorial scheme, it would generally be more efficient to adopt a confounded design than a simple randomised block layout. If a large number of varieties is to be tried it would be desirable to use an incomplete block design. Again in experiments involving a combination of treatments like irrigation and crop varieties, it might be advisable to adopt a split plot design with major plots given to irrigation. The reader is already familiar with the statistical considerations involved in the choice of various alternative designs available for carrying out a given experiment. There is another consideration also and that is the question of available resources. Complex designs can be successfully employed in agricultural experiments when field assistance with the requisite

skill and training is available. In the absence of such help mistakes might be made in the conduct of the experiment rendering the proper analysis of the results difficult and laborious or even impossible. The simple randomised block design possesses a remarkable simplicity and flexibility both in the lay-out and statistical analysis and might be preferred under circumstances where the successful employment of a more complex design is problematic owing to lack of necessary resources. Uniformity trial data indicate that with the plot size usually used namely 1/100 to 1/50th acre, the randomised block design may be adopted even with treatments numbering upto about 20 without appreciable loss of efficiency.

A point to which the research worker should pay particular attention at the stage of planning an experiment is to assure himself that the design adopted is sound. A standard design is always sound. This is, of course, true when the design chosen is appropriate for the experiment. Sometimes a standard design is arbitrarily modified to suit the special requirements of an experiment and mistakes are liable to be made in this process, resulting in a faulty design. Faults of the design are discovered when it is too late, either after the experiment is laid out or even when the data are to be analysed. A practice which serves as a safeguard against such a situation is that of setting out a skeleton analysis of variance for the proposed experimental design. This will help to reveal and correct defects in the design, if any, and thus avoid any unforeseen difficulty in the analysis of experimental results after the experiment has been completed and the data assembled for examination. Accidents resulting in missing plots can, of course, be dealt with by methods of Chapter XII.

15*a*.5 Uniformity of Site

After the choice of the design, the next question is the selection of a proper site for the experiment. Sometimes the experimenter is required to carry out the experiment on whatever land is available. Frequently, however, he is in a position to choose his own site. When this is the case the site selected should be as uniform as possible. By uniformity is meant uniform fertility of land and not simply topographical evenness of the surface of the field although the latter does indicate to some extent uniform fertility.

A good idea of the variation in fertility of a field can be obtained by observing a standing crop growing on the field under a uniform treatment, when patches of differential fertility are easily marked. The importance of uniformity of the land as a factor contributing to the efficiency of the experiment is liable to be overlooked owing to the fact that the various experimental designs are intended to overcome the difficulty caused by this lack of uniformity. It is true that the experimental design seeks to eliminate the grosser effects of such variation on the treatment comparisons; but it cannot eliminate such effects entirely. The more uniform the land the smaller is the experimental error and thus greater the precision of the treatment comparisons.

A uniformity trial on the site of the experiment in the preceding season can give a reasonably accurate idea of the uniformity of the land and further the data can be used to improve the accuracy of experimental comparisons by the analysis of covariance. It is not, however, always advisable to carry out uniformity trials for this purpose. Such a procedure involves almost doubling the expenditure and a year's delay in obtaining the results. With annual crops the reduction in error brought about by such adjustment is not, in general, correspondingly high and the resources would be better utilised in increasing the size of the experiment. With perennial crops such as tea and orchard trees, however, a preliminary uniformity trial, which merely consists in recording yields of the trees before the experimental treatments are applied is found to lead to an appreciable reduction in error. So also in the case of land newly acquired for experimental investigations it would be worthwhile to carry out a uniformity trial.

Sometimes the field in which the experiment is to be laid out is surrounded by trees and they might cast their shade over a portion of the experimental site. Such shade affects the growth of plants under it. Portions of border plots are more likely to receive the shade and for a longer time. Sufficient margin should therefore be left between the trees and the experimental site so that no experimental plots are affected by their shade.

15*a*.6. Size, Shape and Arrangement of Plots

The desirability of having compact blocks and long and narrow plots has been emphasised already. The size of plot

requires further consideration. While the coefficient of variation of plots decreases steadily as plot size increases (Table 7.2) this decrease is not proportionate and consequently an increase in the number of replications, with a reduced plot size if necessary, leads to a more precise comparison of treatments. In other words increasing the replication rather than plot size is more effective in improving the precision of treatment comparisons. The reduction in plot size cannot, however, be carried too far. The plot size has to be large enough in order that the normal cultivation operations by farm implements may be carried out conveniently. It might be possible to manage experiments with a small plot size by replacing the normal operations with mechanical or bullock power by hand operations, but this modification cannot be recommended where the object of an experiment is to try a set of treatments under field conditions and it is consequently necessary that the treatments should be subjected to these very conditions in the conduct of the experiment. If the conditions are changed radically the conclusions might cease to be applicable to normal farming practice.

Another consideration which limits the reduction of plot size is that of border effects and the consequent necessity of leaving non-experimental margins. The border effect, owing to which the yields or other characters of plants nearer the borders of plots differ from those of the central portions, arises in a number of ways. In varietal trials the border plants of more vigorous varieties gain by competition with plants of neighbouring plots—an advantage not available to the inner portions of the plot or under normal field conditions—and thus the difference between the performance of the varieties in the field is liable to be overestimated. Similarly, in manurial trials the manure from manured plots might seep upto some distance beyond the border of these plots and the resulting benefit to border plants in the unmanured plots might decrease the apparent response to the manure. Apart from the possible introduction of bias in the comparison of treatments, these border effects would lead to an inflation of error variance by increasing the heterogeneity among plots. Provision of non-experimental margins round the experimental plot thus becomes necessary (Hutchinson and Panse, 1935). This is done by leaving out one or more rows or a narrow strip of the crop,

depending on the magnitude of border effect, on either side of the experimental plot and a few feet at each end of the rows, when taking observations or when harvesting the plot for recording its yield. It can be seen easily that when the gross plot is smaller in size a greater proportion of the plot area would be discarded as non-experimental margins. It is important to note that the non-experimental margins should receive the same treatment as the experimental plot within, throughout the conduct of the experiment. In fact, the distinction between the two should be made only in marking out the experimental plot for observation or harvest.

The measurement of dimensions of a plot also requires attention, for although it is a simple matter, mistakes due to adoption of faulty procedures are not uncommon. While measuring the breadth of the plot it is necessary to include in it a distance equal to half the row-spacing beyond each of the extreme rows. For example, if there are 12 rows in a plot and the row-spacing is 2 feet, the breadth of the plot will be $12 \times 2 = 24$ feet, of this 22 feet will be the distance between the first and the last rows and 1 foot would be the half row-spacing beyond either of them. This can be followed more easily with the help of the diagram in Fig. 15.1.

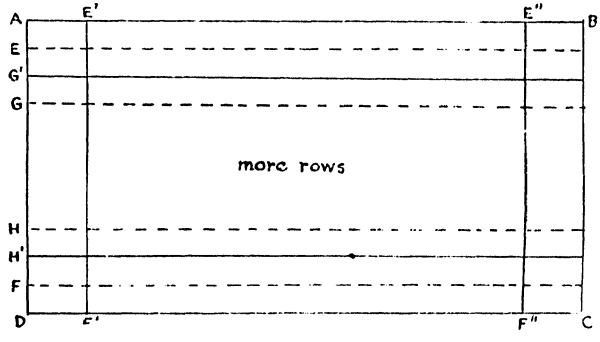


FIG. 15.1. Crop-plot with border rows.

ABCD represents the gross plot. E and F are the guard-rows. The distances AE and FD are each half the row-spacing. The portions AE'F'D and BE''F''C are the end margins to be excluded

from the experimental plots. The experimental plot has thus a length E′E″ and breadth G′H′, G′ and H′ being the mid-points of EG and FH respectively. The correctness of the procedure becomes apparent by considering that half of inter-row distance on each side of a row belongs to that particular row and the area covered by a row is thus the product of the length of the row and inter-row-spacing. The gross plot size is given by AB $\times$ AD and the experimental plot size by E′E″ $\times$ G′H′. If the rows consist of evenly spaced plants, the lines E′F′ and E″F″ representing the end boundaries of the experimental plot should pass through the midpoints between two sets of adjacent plants. While reporting the results of experiments, details regarding gross and net plot size should be given as was done in the examples in Chapter VIII.

15*a*.7. Arrangement of Blocks

In arranging blocks in the field and plots within blocks two points need attention. One is the necessity in certain types of trials such as those on cultivation, sowing dates, etc., of leaving room for turning agricultural implements at the end of a plot. The other point is that the arrangement of blocks should be such as to maximize differences between blocks whereas plots should be so arranged within blocks as to minimize the differences among them. The latter point has been dealt with already in Chapter VIII. If the fertility of a field is known to change steadily from point A to point B, as would be the case, for instance, if the field is sloping from A to B, it would be advantageous to place the blocks one after another along the gradient AB, leaving room between blocks for turning implements, as shown in Fig. 15.2.

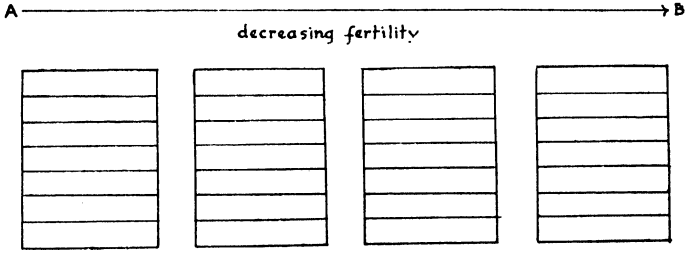


Fig. 15.2. Arrangement of blocks and plots in a field with a fertility gradient.

It would also be advantageous to lay out the plots in each block in a series at right angles to this direction as shown in the figure. If the number of plots is large and there is no marked fertility gradient in the field they may be arranged in two rows as shown in Fig. 15.3 to make the block compact.

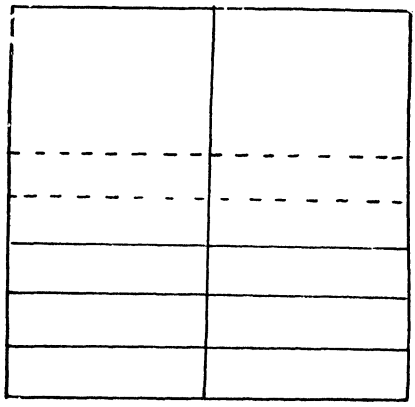


FIG. 15.3. A block with two rows of plots.

If there is a gradient, however, it would be desirable to have all the plots in a single series across the block as in Fig. 15.2. This would be the case when planning experiments on terraced fields on hill slopes. In this situation the compactness of blocks may also have to be sacrificed in order to locate the entire blocks at one level across the slope. Similarly, if there is some topographical feature in the field such as a ridge or a bund demarcating it into different portions, it should be made a dividing line between blocks so as to remove the differences between portions of the field separated by this feature as block differences. For the same reason whenever inclement weather or other reasons compel interruption in field operations like sowing or cultivation, care should be taken to interrupt such operations only on the completion of a whole block. If this precaution is not observed the differences consequent upon different parts of a block receiving the operations at somewhat different times would contribute to error and inflate it.

15*a*.8. Modifications of the Usual Layouts

The nature of treatments to be compared might necessitate a radical departure from the usual experimental layout. A good example is provided by the *haveli* system of cultivation. In this system the field is bunded at the advent of the monsoon and the rain-water is allowed to accumulate in it. Towards the end of the monsoon, the water is let out and the area sown with winter crops like gram and wheat. To compare this method of cultivation against, say, the ordinary method, the plots have to be large enough to accommodate bunds of the size commonly employed. Besides, the effect of bunding on the conservation of moisture would extend to some distance beyond the bunded area which must be excluded from unbunded plots. Such a trial might, therefore, be carried out by selecting a number of fields as replications, dividing each into two portions, leaving sufficient margin between the two and allotting the alternative systems randomly to the two portions. Similarly in experiments involving *contour-bunding*, in which large sections of land having a common drainage system, known as catchments, are contour-bunded for retaining more moisture in the soil, each such catchment would be subdivided into two or more portions for different treatments. Again in experiments on biological control of insect pests through the introduction of suitable predators the untreated plot would have to be well beyond the flying range of the predators. The usual method of laying out an experiment by dividing a field into a number of blocks and subdividing these into several small plots is inapplicable in such cases and might produce absurd results.

Experiments on spacing between crop plants necessitate an interesting modification of the usual experimental lay-out. Suppose we wish to compare 12, 18 and 24 inches row-spacings for some crop. It can be seen that to compare these spacings the experimental plot must be 72 inches (72 being the lowest common multiple of the numbers 12, 18 and 24) or a multiple of 72 inches in breadth as otherwise it cannot be fully occupied by each of the spacings to be tried. Similarly to compare the three spacings 18, 24 and 30 inches the minimum breadth of the experimental plot will have to be 360 inches or 30 feet. These breadths are for the net experimental plots which are thus of the same

size and shape. Further, in order to provide non-experimental margin with the same row-spacing as that tried in the plot the gross plots have to be of unequal size. Hence the blocks in such trials may have to be divided into plots of unequal sizes appropriate for each spacing, after determining the order of treatments in each block by randomisation. Where however the different spacings to be tried do not give plots of identical size but of slightly varying breadth the inequality in plot area should be adjusted by reducing the observations per plot to a standard area for analysis.

15*a*.9 CONDUCTING THE EXPERIMENT

A number of precautions is necessary during the progress of the experiment. The principal thing to aim at is to provide as uniform conditions as possible to all plots. Just as lack of uniformity in the fertility of the experimental plots contributes to error, want of uniformity of other conditions also has the same effect. However, uniformity of operations must be understood in the proper perspective. Considerations of uniformity should not lead to a change of operations so radically as to make them unrepresentative of field conditions. Thus replacing drill-sowing by hand-sowing to increase uniformity would reduce the practical utility of results obtained from the experiment. Without resorting to such a drastic change, uniformity of sowing can be increased reasonably by dividing the seed to be sown into equal portions for sowing in each drill row. Similarly, plots should be harvested as and when they mature. If all the plots are harvested simultaneously for the sake of uniformity, the harvesting being postponed until all plots have matured, it would put early maturing plots, representing early maturing varieties or treatments, at a disadvantage as these plots might shed some of their grain or their produce might be damaged in some other way.

The experiment needs to be protected at every stage from damage by cattle or other animals, birds, etc. Otherwise, damaged plots will have to be rejected and the resulting data will be incomplete. Such data can be completed by using the missing plot technique. This possibility should not, however, encourage the experimenter to regard loss of part of data with complacence. It should be clearly understood that methods of dealing with incomplete data are meant only to make the best of the remaining

data and that loss of data invariably leads to corresponding loss of information from the experiment. Each missing value makes the analysis of the data more and more laborious. The number of degrees of freedom for error is reduced owing to missing plots. The standard error of comparisons including missing plots is inflated and the test of treatment differences is rendered less sensitive. These considerations should make the experimenter vigilant in ensuring that the experiment does not suffer from disturbance at any stage.

15*a*.10 Ancillary Observations and Sampling

In field trials the principal character under study is the yield of crop. A variety of other observations such as on rate of growth, period of flowering, extent of damage by pests and diseases are also made. These are the ancillary observations. In recording them it is generally necessary to resort to sampling since it is not practicable to take observations on every plant in the experimental plots. The main problem in sampling for ancillary observations is one of deciding upon the unit of sampling. The choice of sampling unit depends to some extent on considerations of convenience and the nature of the crop. If plant growth is such that individual plants are distinct as for example in cotton or tobacco, a plant may serve as the sampling unit. With crops like wheat and sugarcane where the crop is grown in continuous rows and individual plants cannot be easily distinguished a yard or foot length of the row may be taken as the sampling unit. These units are located randomly in the experimental plots. An equal number of units is selected from each plot for observation and the data obtained from all units in a plot are pooled for analysis. Alternatively, the observations from individual sampling units are analysed to provide separate estimates of variation between plots and between sample units within plots. This information will enable the experimenter to examine whether the amount of sampling is adequate and to modify the sampling procedure if necessary in future trials. More sampling would be indicated where the mean square within plots approaches in magnitude that between plots.

Sometimes the experimenter wishes to take observations on characters that are difficult to measure quantitatively. A common

instance of this kind is the extent of damage by a plant pest or disease. The experimenter wishes to try some curative or preventive treatments and ascertain and compare the extent of damage in plots under different treatments. For this purpose the experimenter might mark suitable sample units in the plots for observation and classify the plants in these units into various grades according to the extent of damage, *e.g.*, slight, medium, severe, very severe and so on. This would give him for each plot the number of plants in each category. To obtain a single measure of damage in the plot the experimenter may assign scores to the different grades, as for example 1 for slight damage, 2 for medium and so on. These scores could be analysed in the usual manner for comparing the different treatments. The procedure provides a workable solution of the problem. Another approach available in such a case would be to invite a number of observers to arrange the plots under various treatments in the order of the extent of damage in each replication. The ranks could then be converted into scores for statistical analysis with the help of Table XX in the *Statistical Tables* by Fisher and Yates.

15*a*.11 Analysis and Interpretation of Results

The statistical analysis of experimental data and interpretation of results often present problems which require the exercise of much ingenuity and commonsense. Considerable economy in computation can often be effected by not falling into the habit of carrying out the analysis mechanically. If the purpose at hand is served by a simpler analysis more elaborate analysis may be dispensed with. Thus a varietal trial in incomplete blocks might as an approximation be analysed as in simple randomised blocks and progeny-wise analysis in a compact family block trial might be dropped in families which do not appear promising. Carrying out numerical calculations to appropriate number of decimal places (Appendix VII) would help in economising labour and at the same time guard against the possibility of discarding too many figures at the earlier stages in the calculations. Similar considerations would apply to the choice of suitable ancillary characters for adjustment of the results by the analysis of covariance. By experience the research worker can judge which characters would give appreciable reduction of error and restrict the analysis to

such characters when observations on a number of characters are available. An interesting example of how precision of an experiment can be improved appreciably by the use of suitable ancillary observations is provided by the results of a trial carried out on land affected by saline patches.

Example 15.1

Six strains of gram A, B, C, D, E and F were tried in six randomised blocks on alkaline soil. The crop was partially damaged by alkali patches. The yields obtained from various plots are given in Table 15.1 according to variety and block.

TABLE 15.1

Yield of gram strains in oz./plot

Variety \ Block	I	II	III	IV	V	VI
A ..	242	163	156	163	232	222
B ..	161	128	274	127	224	285
C ..	86	186	206	333	166	293
D ..	172	118	232	231	189	274
E ..	122	150	89	237	251	288
F ..	196	181	132	187	243	164

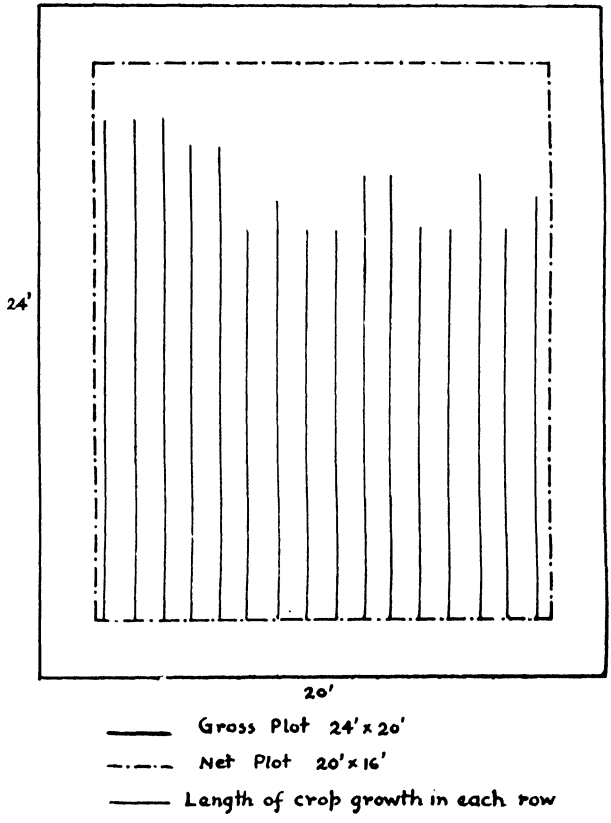
An inspection of Table 15.1 would show that the yields are highly variable. Thus variety C has given yields ranging from 86 to 333 oz. per plot. Similarly variation within certain blocks is also considerable. The analysis of the yield data gave the following results:—

TABLE 15.2

Analysis of variance of gram yields (oz./plot)

Source of variation	*D.F.*	*S.S.*	*M.S.*
Blocks ..	5	43009·2	..
Varieties ..	5	2908·5	581·7
Error ..	25	83824·0	3353·0
Total ..	35	129741·7	..

The varietal mean square is even less than the error mean square and so the varietal differences must be regarded as non-significant. The coefficient of variation is however large being 29·3. This large variation has presumably been due to alkali patches. In order to allow for the effects of damage due to these patches, the undamaged length of the crop was measured in each plot and the plot yields were adjusted by covariance with the undamaged area per plot. Figure 15.4 shows diagrammatically the crop-growth in one plot.



Distance between two adjoining rows is one foot.

FIG. 15.4. A plot of gram damaged by alkali patches.

The undamaged crop lengths for various plots are shown in Table 15.3.

TABLE 15.3

Undamaged length of crop rows in feet per plot

Variety \ Block	I	II	III	IV	V	VI
A ..	249	213	297	256	278	240
B ..	286	287	314	210	275	318
C ..	159	146	243	320	183	320
D ..	263	287	267	312	216	320
E ..	165	146	160	294	211	297
F ..	274	216	231	254	320	217

On adjustment of the analysis of yield data by covariance the following results were obtained for the varieties and error:

Source of variation	*D.F.*	*S.S.*	*M.S.*	*F* (observed)
Varieties ..	5	11850·3	2370·6	1·26
Error ..	24	45233·7	1884·7	..

It will be seen that there is a considerable reduction in the error mean square and although the variance ratio is non-significant, it is now greater than 1. The original and adjusted mean yields of different strains are shown in Table 15.4.

TABLE 15.4

Adjusted and unadjusted means of gram strains (lb./acre)

Variety ..	A	B	C	D	E	F	*S.E.*
Adjusted mean ..	182·3	166·1	217·1	171·9	208·4	173·2	16·8
Unadjusted ..	185·6	188·9	200·1	191·6	179·1	173·7	22·3

In the above example it was relatively easy to eliminate the effects of the principal cause of heterogeneity in the data. However this may not be always possible. It is sometimes found that one or two varieties or treatments yield much more or much less than the remaining and also differ from the latter in variability. Application of the simple analysis of variance to such data involving the pooling together of variances to which different comparisons are subject would not be justified as the common estimate of variance would under-estimate the variance for certain comparisons and over-estimate it for others. In this situation the varieties or treatments introducing disturbance might be omitted from the main analysis which would then provide more reasonable tests of significance for the remaining comparisons. It should be noted, however, that such omission of data from analysis can be justified only to take into account the deviation of the actual material from the hypothetical model for which the usual procedure of analysis of variance is valid and is not to be resorted to in order to obtain results which might appear satisfactory to the experimenter.

15*a*.12 TRANSFORMATIONS

A situation is encountered sometimes in which a straightforward analysis of variance of the data is not valid. This happens when the observations refer to small whole numbers or to percentages calculated from small whole numbers. Consider a treatment such as a germicidal spray intended to protect the crop against some disease. To test the efficacy of the treatment it would be natural to take counts of surviving plants in the treated and untreated plots. If the proportion of surviving plants in the untreated population is say half and each plot contains n plants initially, the number of surviving plants in the different untreated plots would be distributed binomially, with sampling variation equal to $n/4$. The proportion of surviving plants in the population for treated plots might be nearer one, say 5/6, and the sampling variance for the number of surviving plants per plot would be $(1/6)\ (5/6)\ n$. The inequality of the variances makes the direct analysis of variance inappropriate. The same considerations apply to percentages calculated from such numbers. The difficulty can be overcome largely by transforming the data on a suitable

scale, and analysing the transformed values. Thus when the observations are expected to follow the binomial distribution it is useful to carry out the inverse sine transformation. This consists in calculating the angle θ corresponding to the observed value of proportion P, such that $\sin \theta = \sqrt{P}$. This can be done with the help of tables of natural sines. To a first approximation, the treatment variances would be equalised on this scale, and the transformed values can then be analysed by the usual procedure of analysis of variance. Similarly when the mean and variance of different treatments appear to be associated, that is when treatments with greater mean show greater variance and *vice versa* as also when the data exhibit a high variability not ascribable to any particular set of treatments or other specific factors, it is useful to transform the data to logarithms and analyse the logarithms of individual observations. Tables suitable for use in making these transformations are included among *Statistical Tables* by Fisher and Yates (Tables XXVI and XXXI).

CHAPTER XVI

EXPERIMENTS IN CULTIVATORS' FIELDS

16*a*.1 The Place of Experiments in Cultivators' Fields in Research

The experimental designs described in the previous Chapters are mainly intended for use in agronomic and plant breeding research. Their planning and conduct presuppose the availability of certain minimum facilities and technical skill, such as are available at experimental stations. On the other hand, any conclusions based on the results of a group of experiments at research stations cannot be immediately recommended for general adoption under actual farming conditions in the country. Firstly, the number of experimental stations is small, and secondly, the fertility of the soil and the level of management at experimental farms are superior to those in cultivators' fields. The reader will recall that in interpreting the results of a group of experiments in Chapter XIV, it was pointed out that the conclusions can be generalized for application only to the type of fields represented by the experimental stations and not over the tract as a whole in which the experimental stations are located. The former is an indefinable and at the best a limited population and consequently restricts the utility in practice of the results obtained at the research stations. A satisfactory method of bridging this gulf between the results of research at experimental stations and their adoption by cultivators is to conduct experiments in fields representative of the entire tract.

16*a*.2 Practical Difficulties

The statistical principles of selecting fields representative of a tract or of any unambiguously specified type within it, form the subject-matter of the sampling theory of surveys (Sukhatme, 1953) and have been briefly explained in Chapter III. Simple as these appear, the idea of locating a group of experiments in a representative sample of cultivators' fields does not seem to have been put into practice on an extensive scale anywhere. On account of the apparent difficulties and the heavy cost involved in organizing an

experimental program of this type, the approach has been generally regarded as impracticable. These difficulties, in general, are the limited experimental facilities in the countryside and the apathetic, if not antagonistic, attitude towards experimentation on the part of the cultivator, apart of course, from the relative inaccessibility of many of the fields selected for the experimental program. An average cultivator is a poor man working on a small, usually unfenced, area of land and is preoccupied with his daily routine. He can therefore hardly be expected to divert any of his limited resources to experimental work which might disturb the normal operations on his field or in which there is a risk of incurring losses. A correct psychological approach to win his confidence and gain his co-operation thus becomes the first step before initiating a successful experimental program of this type. For this purpose it is necessary to ensure that the design of the experiment is simple enough to be conducted within the limited resources available on a cultivator's field and such that it can easily be fitted into the normal routine of his work. Secondly, utmost care is called for in choosing the treatments for experimentation so that he may not incur loss through the granting of facilities for trying these treatments on his field. As will be shown in the succeeding sections, it is possible to design very simple experiments under the cultivator's conditions which can fit into his time-table of operations on the field and yet collectively supply the information required. Similarly, with well-chosen agronomic treatments such as manures and fertilizers under conditions of an assured moisture supply, or with promising varieties of crops, the probability of loss can be reduced to the minimum. The manures, fertilizers or seed required for the experimental treatments would be supplied free. Payment would be made to the cultivator for any labour assistance which he may provide for purposes of the experiment. It will thus be seen that the practical difficulties referred to above are in no way insuperable and can and must be overcome in the interests of the need for obtaining experimental results under a truly representative set of farming conditions. Experience has, in fact, shown that where experimental programs on these lines have been taken up, the cultivators have soon come forward to pay the cost of experimentation.

16*a*.3 Need for using the Principle of Random Sampling in Selecting Sites

It has been mentioned in Chapter III that the method of random sampling can give a representative sample of fields and that it is only from experiments on such fields that results of general applicability and of measurable precision can be derived. Notwithstanding these advantages, non-random methods of selection, based upon the personal judgment of experimenters, are not infrequently used for selecting representative sites. It should however be emphasized that the advantages of deliberate selection are more apparent than real. Experience has shown that the use of non-random methods, even in the hands of experts, cannot be relied upon to give a sample representative of the population and consequently estimates of response obtained from experiments in such fields are liable to be seriously biased. Even if quotas are set up to represent the different categories like soil types, the ultimate selection of actual fields within each category is influenced by the personal judgment of the experimenter and the result is therefore likely to be biased. It is of course true that such methods are convenient to use in practice. Their cost is also low relative to that of the method of random sampling. However, unlike random sampling, these methods lack the means for judging the precision of the response obtained. Since an important primary object of experimentation on cultivators' fields is to estimate the average response to a given agricultural improvement measure over the tract and to test the consistency of this response in different parts of the tract the method of random sampling should be used in locating fields for experiments.

It is contended that the selection of experimental sites should not be made from all the fields in the tract but rather from, say, a given soil type or a climatic or agriculturally homogeneous zone. If a sufficiently detailed soil map is available, and if the experimental treatments are such as warrant their trial only on a specified soil type, there is nothing to prevent using the method of random sampling in selecting fields out of the given soil type. Where, however, detailed sampling frames like a soil map are not available, it is convenient to select from the totality of fields

in the tract and then assign the selected fields for experimentation by the desired types or zones or regroup the results in any desired manner. The principle of random sampling is just as valid whether we select sites from a population of fields within a given geographic region or from any given soil, climatic or agricultural zone.

Another possible objection to the use of the method of random sampling for the selection of experimental sites is that a particular experiment may be located on a site where a manurial treatment may be ineffective because of the operation of some limiting factor such as salinity. It must be remembered, however, that such factors form part of the conditions which affect the average yield of the tract and it is therefore necessary that the experimental treatments should be tested under all conditions obtaining in the tract. There is however nothing to prevent the experimenter from defining the population of fields in advance in a way considered most suitable for trying out the given experimental treatments. Thus in experimenting with manurial treatments one might confine the selection to fields receiving irrigation only. Or again, in experimenting with a new promising variety of a crop one might well have to experiment only in those areas where the growing season is long, if the variety to be tried is a late manuring one. Whatever be the population, whether it is the totality of fields from the tract or fields belonging to any given type within it, the experimental sites to represent the population should be selected using the principle of random sampling.

One objection of any substance to the use of the principle of random sampling arises from the limitations of communication. Thus, fields may be inaccessible during the rainy season, making transport of manure, fertilizer, seed, etc., difficult. Deviations from the principle of random sampling under such conditions may in extreme cases be inevitable, but even here the principle of random sampling can be approximated by sub-sampling randomly a small predetermined number of fields out of the initially selected fields in the sample whose omission appears unavoidable and making a determined effort to experiment thereon. Provided the omissions in the sub-sampling are few they will not seriously affect the validity of the results.

16*a*.4 The Design of Experiment

Apart from the choice of treatments which we have seen must be few in number and promising in their results, the design of an experiment on cultivators' fields must be extremely simple and possess demonstration value if we are to win the co-operation of the cultivator in any experimental program of this type. The simplest of experimental designs is the randomized block design. But even a randomized block design with its replication involving numerous small plots lying side by side in the field cannot fulfil the requirement mentioned in Section 16*a*.2, namely, that of enabling the cultivator to carry out his normal field operations undisturbed. A design which might appeal to the cultivator would be to divide his field into as many portions as there are treatments, apply the treatment over the whole of each of these portions and harvest plots of given dimensions at harvest time in the presence of the experimenter. Thus, with an experiment with five treatments, a field would be divided into five approximately equal portions; in one portion the crop would be grown according to the cultivator's normal practice and this would be the control treatment for purposes of experiment. In the other four portions of the field the experimental dressings would be superimposed on the cultivator's normal practice, namely, the control. If the field is too large a suitable section thereof, restricted to an upper limit of, say, one acre, could be selected for experimentation on these lines. In brief, the idea underlying this design would be that the whole field would be cultivated, seeded, etc., by the cultivator in his usual way, but four suitable portions of a given area would have experimental treatments superimposed on the normal. At harvest time plots of given dimensions would be marked in random positions within the different portions and the produce from these plots would be weighed and recorded.

The procedure may present a practical difficulty. This arises from the fact that the field staff are required to measure the areas of the different portions in order to determine the precise quantities of the treatments to be applied to each. It should not however be difficult to train the field staff in the measurement of the areas of the portions into which a field may be divided for purposes of experimentation. The alternative solution

is to standardize the size of the portions, say 1/10 acre each and arrange them in a compact block in the midst of the field of the cultivator. This arrangement could have the advantage of being economical in that a smaller amount of manure, fertilizer, seed, etc., would be required for use as experimental treatments and at the same time would not present any serious difficulty in letting the cultivator carry out his normal field operations undisturbed. Where the cultivator is co-operative and prepared to put up with some inconvenience to his normal operations entailed by this arrangement, this design has been tried with success.

In either case the procedure is open to obvious objections on statistical grounds. In the first place there is no replication. This objection can however be met by repeating the experiment on another field. In other words, fields rather than compact blocks within fields would constitute the replication for the experiment. The second objection is that the procedure does not allow an effective use to be made of the principle of local control in eliminating fertility variation from treatment comparisons. In an experimental program involving a number of experiments spread over a large tract, this is an unimportant factor, for the main object of the experimental program is to estimate the average response of the various treatments for the tract as a whole and not for any specified field. The accuracy of a single experiment thus plays only a secondary role in the whole scheme. That the experimental plots for harvest are not contiguous or of a shape considered advantageous in a field experiment at a research station is thus altogether unimportant owing to the small contribution of these factors to the response compared to the contribution of variation from other sources, namely, the variation between experiments. Even then, the fact that all treatments are grouped together in the same field provides a degree of local control. In regard to randomization it is important that the treatments should be allotted at random to the different portions or plots within a field. Further, the randomization should be carried out independently on different fields. It is sometimes considered that one random arrangement of treatments is satisfactory for all fields. This is a mistaken notion. Looked at from this angle it can be seen that the procedure of spreading the experiment over several fields, dividing each into as many portions as there are treatments to be tried subject to

certain upper limits for the field and for portions thereof, allocating the treatments to these portions in a random order and harvesting and weighing the produce from plots of the requisite dimensions marked in a random position in each portion at harvest time, not only would enable the cultivator to proceed with his normal operations undisturbed but at the same time would satisfy the basic principles of randomization, replication and local control of an experimental design. The design described above presupposes that the fields are large and divisible into portions of say 1/10 acre each. For certain crops and in certain regions however fields are likely to be small. Thus, in terraced regions paddy fields are long, narrow and small. Under these conditions a field itself might be the unit of treatment, 4 or 5 adjacent fields constituting the different portions of the experiment. A cluster of adjacent fields forms in this case the block. The arrangement has one distinct advantage over the arrangement of dividing a single field into different portions, in that it is no longer necessary to put up bunds between different portions of a field. On the other hand, the principle of local control is now less effective.

16*a*.5 The Number of Experimental Fields and its Distribution Between and Within Places

Fields for experiments will ordinarily be selected in two stages of sampling—places (usually villages) in the first stage and fields within the selected places in the second stage. The cost of repeating an experiment in one more field in the same place will obviously be smaller than that of locating it in a field in another randomly selected place. Likewise, the variation of the treatment response within a place will ordinarily be smaller than that between places. On the first of these two considerations the total cost of experimentation in a given number of fields would decrease as the number of fields per place is increased. The second consideration would pull in the opposite direction, the variance of the treatment response increasing with the number of fields in a place at the cost of the number of places. The aim of an experimenter should clearly be to so determine the number of experimental fields and its distribution between and within places that the treatment response over the tract is estimated with the minimum variance for a given budget, or alternatively, with the desired

precision at a minimum cost. It is therefore necessary as a first step, to know the relationships between (1) the total cost of experimentation and the number of places and fields per place, and (2) between the variance of the estimated response and the number of places and fields per place.

We have seen in Chapter XIV that the analysis of variance of plot yields of a group of experiments at n places with m fields (replications) per place takes the form shown in Table 16.1.

TABLE 16.1

Analysis of variance of plot yields from group of experiments

Source	*D.F.*	Mean Square	Estimate of
Treatments	$t-1$	..	..
Places	$n-1$	..	..
Fields within Places, *i.e.*, blocks ..	$n(m-1)$	..	..
Places × Treatments ..	$(n-1)(t-1)$	s_b^2	$\sigma_e^2 + m\sigma_m^2$
Blocks × Treatments ..	$n(m-1)(t-1)$	s_w^2	σ_e^2

We further saw that the error variance per plot of a treatment response will consist of two parts: (1) the error variance per plot at a place σ_e^2 and (2) the variance due to interaction of response with places σ_m^2, the variance of the average treatment response estimated from nm experiments spread over n places with m replications each being given by

$$V = 2\frac{(m\sigma_m^2 + \sigma_e^2)}{nm} \qquad (16.1)$$

This expression determines the relationship between the precision of the estimated response and the number of places and replications per place. Likewise, the total cost of experimentation can also be considered as made up of two components: (1) the cost of setting up an experimental place or centre, and (2) the cost of operations in the conduct of experiments. The cost of setting up an experimental place includes the cost of salary of the experimenter for the days of his visit to the place and the cost of his journey including transport of equipment. The total cost of experimentation can therefore be expressed as

$$C = c_1 n + c_2 nm \qquad (16.2)$$

c_1 representing the cost of setting up a place and c_2 the cost of conducting the experiment in a single field.

The problem of determining the number of experimental places n and the number of replications per place m is thus the problem of minimising the total cost for a prescribed value of the variance, V_0, with which the response to a given treatment is sought to be estimated. It can be shown that these values, known as the optimum values of n and m, are given by

$$\hat{m} = \sqrt{\frac{c_1}{c_2}} \cdot \frac{\sigma_e}{\sigma_m} \tag{16.3}$$

and hence

$$\hat{n} = \frac{2\left(\sigma_m^2 + \frac{\sigma_e^2}{\hat{m}}\right)}{V_0} \tag{16.4}$$

Experimental surveys on a pilot scale should provide the data to evaluate n and m. The value of the ratio σ_e/σ_m will depend on the relative magnitude of the true variance of response to treatment between places and between fields within places. As far as the other component $\sqrt{c_1/c_2}$ of formula (16.3) is concerned, c_1 will, of course, be larger than c_2, but the exact value of the ratio would depend upon the local conditions, the rates of pay, the cost of labour and the number of treatment plots. To illustrate the application of the formulæ where district agricultural staff is conducting the experiments, each within his jurisdiction and is paid on a monthly basis, his salary and travelling allowance, we might take $c_1/c_2 = 4$. Sufficient data for estimating the ratio of true variances between places to that within places is yet lacking; but the limited experience that is available indicates that although variable it might be assumed tentatively as 1 to 2. The optimum value of the number of replications (fields) per place can thus be taken to be 3. Substituting $\hat{m} = 3$ in formula (16.4) we should arrive at the number of places to be selected for experimentation for estimating response to treatment with a given precision. On the other hand, should treatment effects vary more widely from place to place, σ_m would be comparable or even larger than σ_e and the experimental plan might take the form of a larger number of places with fewer replications per place, a field always accommodating one replication and no more unless conditions

warrant otherwise. The minimum replication in that case could be one per place but this would deprive the experimenter of the opportunity to assess the variance within places for use in planning further investigations. Two replications per place would, therefore, seem advisable as the minimum.

In the analysis presented here it has been assumed that σ_e is constant and independent of the number of treatments. Actually the value of σ_e will vary with the number of treatments. In extending the results to planning of experiments with different numbers of treatments, therefore, care would be necessary to use the appropriate values for σ_e and σ_m in the formulæ for the number and distribution of replications between and within places. Any difference resulting from a change in the value of σ_e would however be ordinarily small and the result obtained on one set of treatments can be taken to serve as a rough guide to the number of replications and its distribution in order to plan a similar experimental program with the maximum statistical efficiency.

All through the discussion in this and the previous sections we have assumed that the tract would first be divided into homogeneous agricultural zones. Even within the agricultural zones the experiments may have to be confined to given soil and climatic types so that the cropping system is more or less uniform over the zone. Again, it is not sufficient to study the interaction of response with regions. It is equally important to study the interaction of response with seasons. In fact, no results can be recommended for adoption in practice unless the experimental program in cultivators' fields is continued in the same region for two or three years. This of course does not imply that the same fields and plots should continue to be experimented with year after year. On the contrary, fields for experiments should be chosen afresh each year. If it is desired to estimate residual effects this could be done by taking observations on a pre-selected fraction of the total number of fields of the previous year.

16*a*.6 Choice of Treatments

The choice of treatments is governed by three considerations: (1) they should be promising in results, as judged by past research at experimental stations, (2) they should be small in number so

as to avoid the need for laying out more than 5 or 6 plots per field and (3) they should form a self-contained set in the sense that easily intelligible comparisons of direct practical value can be made from the set of treatments adopted. A very common objective in manurial trials is the comparison of levels of nitrogen and phosphate, singly and in combination. Now if we take three levels of each of these components, no application of manure being one of the three, it would require 9 plots to lay out a factorial experiment with these two factors. Such a trial cannot possibly be carried out in cultivators' fields on any extensive scale. The treatments to be tried may therefore be divided into sets as follows:—

A. (i) o, n_1, n_2

(ii) o, p_1, p_1n_1, p_1n_2

(iii) o, p_2, p_2n_1, p_2n_2

B. (i) o, p_1, p_2

(ii) o, n_1, n_1p_1, n_1p_2

(iii) o, n_2, n_2p_1, n_2p_2

It will be seen that each group enables the experimenter to get information on the interaction NP and response to n or p as compared to the cultivator's normal practice, response to p being confounded in group (A) and response to n, in group (B). Depending on whether the experimenter is primarily interested in comparisons of n, or p the sets of treatments given in the first or the second group may be adopted, one-third the total number of experiments being devoted to each set of treatments in either case. If, however, information is required on the response of both n and p both groups of treatments may be adopted each set being allotted to one-sixth the total number of experiments. An obvious advantage of the suggested arrangement is that it does not call for trial of more than four treatments at a time while permitting all comparisons of interest to be made.

Keeping down the number of treatments to be tried in an experiment to 3 or 4 is of particular importance when the field-work is to be managed by relatively unskilled staff and when the co-operation of the cultivator has to be enlisted for the first time

in organising an experimental program in his fields. When however, trained and skilled staff is available for this type of work and there is adequate provision of supervision and the cultivator is appreciative of the value of such experiments more ambitious experimental programs can be undertaken. Even then it is usually necessary to confine the number of treatments to be tried to 5 or 6, and it calls for the exercise of considerable skill in the choice of treatments so as to provide comparisons of interest with this number of treatments. As an example we shall give illustrative sets of treatments from a program of fertilizer research in progress, sponsored jointly by the Government of India and the U.S. Technical Co-operation Mission. The objective is to compare different types and levels of nitrogenous and phosphatic fertilizers. Three nitrogenous and four phosphatic fertilizers, namely, n (Ammonium Sulphate), n' (Ammonium Nitrate), n'' (Urea), p (Super-phosphate), p' (Nitrophos), p'' (Ammonium Phosphate), p''' (Bone Meal), are included for trial. The numerical suffixes attached to the symbols indicate the level of nitrogen or phosphate at which the fertilizer is to be applied. The suggested sets are as follows:—

I. Comparison of levels and types of nitrogen.

(*a*) On soils not expected to respond to phosphatic manuring, the following sets of five treatments are to be tried, each in one-third of the total number of experiments in this category.

(i) $o,\ n_1,\ n_2,\ n_1',\ n_2'$

(ii) $o,\ n_1,\ n_2,\ n_1'',\ n_2''$

(iii) $o,\ n_1',\ n_2',\ n_1'',\ n_2''$

Nitrogen at 20 and 40 lb. per acre.

(*b*) On soils expected to respond to phosphate, the following sets of six treatments are to be tried each in one-third of the number of experiments of this type.

(i) $o,\ p,\ n_1p,\ n_2p,\ n_1'p,\ n_2'p$

(ii) $o,\ p,\ n_1p,\ n_2p,\ n_1''p,\ n_2''p$

(iii) $o,\ p,\ n_1'p,\ n_2'p,\ n_1''p,\ n_2''p$

P_2O_5 at 20 lb. per acre.

II. Comparison of type of nitrogen and effect of phosphate.

o, n, np, $n'p$, $n''p$

Nitrogen and P_2O_5 each at 20 lb. per acre.

III. Comparison of types of phosphate and effects of nitrogen (for unirrigated areas).

o, n, np, np', np'', np'''

On non-acid soils, the np''' treatment would be omitted.

On all plots except the untreated plot, nitrogen is to be made up to a total of 20 lb. per acre by the addition of sulphate of ammonia. P_2O_5 at 20 lb. per acre.

IV. Comparison of types and levels of phosphates and effect of nitrogen (irrigated areas).

The following sets of six treatments, each on one-sixth of the number of experiments of this type is to be tried.

(i) o, n, np_1, np_2, np_1', np_2'

(ii) o, n, np_1, np_2, np_1'', np_2''

(iii) o, n, np_1', np_2', np_1'', np_2''

(iv) o, n, np_1, np_2, np_1''', np_2'''

(v) o, n, np_1', np_2', np_1''', np_2'''

(vi) o, n, np_1'', np_2'', np_1''', np_2'''

P_2O_5 at 20 and 40 lb. per acre.

On all plots except the untreated plot, nitrogen should be made up to a total of 40 lb. by the addition of sulphate of ammonia. Sets (iv), (v) and (vi) are to be tried only on acid soils.

The grouping illustrated by these sets has been made possible by the judicious use of the information derived from past research, namely, that crops respond to nitrogen in some easily available form everywhere in the country whereas response to phosphate is limited to certain areas, while response to potash is generally absent. The main objective of these experiments is consequently the estimation of response to different forms and levels of nitrogenous fertilizers. Further, in phosphate-responsive areas

the experiments provide for information on the response to nitrogen when supplemented with phosphate. Different forms and levels of phosphate are therefore proposed for trial in conjunction with nitrogen. Treatments with forms and levels of phosphate alone are not included as the economic interest centres on the basis of past research on the effect of phosphate in the presence of nitrogen rather than supplied by itself. Finally, it should be noted that the sets of treatments given above are not the only sets possible for realising the objectives and alternative sets can be thought of. The general principle to bear in mind in planning an experimental program of this type is that the objectives should be defined clearly so that these can be realised by the experimental treatments proposed within the practical limitations imposed by the experimental design.

16*a*.7 Auxiliary Observations

An experimental program on the above lines would provide information on the mean responses for the tract and their interactions with the different subdivisions of the tract. If interaction is absent, the results would usually give adequate information for making recommendations applicable to the entire tract. If interaction is discovered, it would be necessary to make specific recommendations for different parts of the tract. Such recommendations would be strengthened by investigating causes responsible for the variations of response. This can be done by collecting suitable ancillary data on the soil and other environmental factors of the experimental fields for correlation with the experimental results. Thus, observations on the topography and soil characters with particular reference to depth of surface soil, depth to which roots are observed to penetrate, colour, texture, permeability, drainage of substratum and subsoil water-table would be of value. A laboratory analysis of soil samples taken from the experimental field would provide useful data. Details about the conduct of the experiment such as the date and method of sowing, seed rate, variety, whether the crop was irrigated and if so the amount of irrigation given, etc., should of course be recorded. Data on rainfall are important particularly for experiments on rainfed crops. Observations on the growth of crops and damage by diseases and pests should also be taken.

16*a*.8 The Analysis of Data

The methods of analysing the data of a group of experiments have been already dealt with in Chapter XIV. No new principles are involved in the analysis and interpretation of data from experiments on cultivators' fields except that we should add that even if the design is not uniform and treatments differ in some of the experiments, results for any particular comparison of interest can be consolidated from all experiments in which this comparison is possible by the method explained in the chapter. In this section we shall consider an example to illustrate the analysis of data obtained from a set of experiments in cultivators' fields.

Example 16.1

The example refers to the data obtained from experiments carried out as part of a wider program, initiated in Tanjore District of Madras State, for testing under cultivators' conditions treatments found promising at experimental stations. The data considered here relate to the value to the paddy crop of ammonium sulphate alone and in combination with superphosphate. The district was divided into four agriculturally homogeneous zones and in each zone a certain number of villages were selected randomly. In each village one field was selected randomly and this field was divided into three approximately equal parts. To one part was allotted ammonium sulphate (25 lb. nitrogen per acre), to the second ammonium sulphate (25 lb. nitrogen per acre) in combination with superphosphate (40 lb. P_2O_5 per acre) while the third was left untreated to serve as control representing the cultivators' own practice. For recording the yield of each treatment a plot measuring 7·26 cents was located in a random position and its produce harvested. In all there were 38 villages in which the experiment was carried out. Table 16.2 gives the yield of paddy for each of the three treatments in each village and for each zone.

TABLE 16.2

Experiments on cultivator's fields in Tanjore District of Madras State

Yield of dry paddy in lb. from plots of 7·26 cents area

	No. of village	Control	Ammonium Sulphate	Ammonium Sulphate + Super-phosphate	Zonal mean
Zone 1.	1	90	116	171	
	2	146	171	157	
	3	189	196	186	
	4	222	287	222	
	5	168	166	185	
	6	131	134	142	
	7	185	201	263	
	8	131	118	160	
	9	58	89	98	
	Mean	146·7	164·2	176·0	162·3
Zone 2.	10	170	227	183	
	11	218	244	196	
	12	114	154	186	
	13	138	114	126	
	14	104	150	126	
	15	180	215	225	
	16	162	197	214	
	17	152	193	213	
	18	136	153	159	
	19	191	203	210	
	Mean	156·5	185·1	183·8	175·1
Zone 3.	20	137	155	170	
	21	288	293	296	
	22	167	171	183	
	23	122	129	138	
	24	101	136	144	
	25	218	240	273	
	26	204	208	205	
	27	144	167	104	
	28	187	217	258	
	29	91	128	136	
	30	142	169	189	
	Mean	163·7	183·0	190·5	179·1
Zone 4.	31	181	198	210	
	32	202	204	206	
	33	249	263	277	
	34	216	230	268	
	35	186	154	171	
	36	203	216	230	
	37	188	176	200	
	38	194	203	196	
	Mean	202·4	205·5	219·8	209·2

Comparing the zonal mean values, interesting differences are observed. In the first place the general yield level is distinctly higher in the fourth zone than in the other three. Secondly, the first three zones have reacted differently to the treatments than the fourth, ammonium sulphate producing a distinctly larger response in the first three compared to the fourth, and phosphate applied in the presence of nitrogen producing a response only in the fourth zone as compared to the first three. This is brought out clearly in the analyses of variance for the individual zones given in Table 16.3. It will be seen that there are no replications within villages but such replication is not essential since for our present purpose the various villages within each zone constitute the replications and the groups of replications in the four zones can be regarded as four replicated trials. Splitting the two degrees of freedom for treatments into one for control *versus* fertilizer and the other for ammonium sulphate alone *versus* ammonium sulphate *plus* superphosphate, we see that the first has a significant mean square in the first three zones and not in the fourth, while the second has a significant mean square only in the fourth.

In considering the possibility of carrying out a composite or pooled analysis of variance, we notice that the error mean squares in individual zones (Table 16.3) vary appreciably; but applying Bartlett's test we get

$$\chi^2 = 4{\cdot}567 \text{ with } 3 \textit{ d.f.}$$

TABLE 16.3

Analysis of variance for individual zones (lb./plot)

Source	Zone 1 D.F.	Zone 1 M.S.	Zone 2 D.F.	Zone 2 M.S.	Zone 3 D.F.	Zone 3 M.S.	Zone 4 D.F.	Zone 4 M.S.
Between villages ..	8	7382**	9	3449**	10	9225**	7	2597**
Control *vs.* fertilizer (T_1)	1	3298*	1	5207**	1	3895**	1	560
Ammonium Sulphate *vs.* Ammonium Sulphate *plus* Super (T_2)	1	624	1	8	1	313	1	812*
Error	16	510	18	367	20	303	14	151

This is a non-significant value and there is no justification for assuming the errors to be heterogeneous. We can therefore carry out a simple analysis of variance of the entire data given in Table 16.2. This analysis is shown in Table 16.4. In view of the indication that the zones differ in their average yield and in the relative response to nitrogen and to a supplement of phosphate, the zones are divided into two groups, the first three together and the last, Z_1, indicating the differences within the first set and Z_2 the difference between this group and the last zone in the analysis of variance. The interaction of these groups with the two degrees of freedom for treatments, T_1 and T_2 is also shown separately.

TABLE 16.4

Analysis of variance pooled over all zones

Source	D.F.	M.S.
Between first three zones (Z_1)	2	2224**
Average of first three zones *vs.* fourth zone (Z_2) ..	1	25208**
Between villages within zones	34	5898**
Control *vs.* fertilizer (T_1)	1	11970**
Ammonium Sulphate *vs.* Ammonium Sulphate *plus* Super (T_2)	1	1107
Interaction (Z_1) × (T_1)	2	50
Interaction (Z_1) × (T_2)	2	214
Interaction (Z_2) × (T_1)	1	891
Interaction (Z_2) × (T_2)	1	223
Error	68	337

The results are somewhat unexpected in that none of the interaction components is significant, although the mean square for the interaction Z_2T_1 is appreciably higher than the pooled error mean square. From the analysis of individual zones, we should have expected this mean square to be significant. The interaction Z_2T_2 should also have shown significance. Apparently we are here dealing with a border line case in which the responses themselves are moderate and their interactions with the subdivisions of the tract are consequently not sufficiently marked to be demonstrated consistently. Nevertheless we must not overlook the inferences derived from the analysis of results

for the individual zones in planning future experimental work and in drawing conclusions for practical recommendations to be made to cultivators. This inference is that unlike in the first three zones a supplement of phosphate is likely to be necessary to the nitrogenous fertilizer in the fourth zone. This indication can be verified by further experimentation in the zone by adopting sets of treatments similar to those in group (A) mentioned in Section 16*a*.6.

According to the present combined analysis of variance, the standard errors of the treatment responses can be based on the mean square pooled from the 68 degrees of freedom for error and the 6 degrees of freedom for interaction between zones and treatments from Table 16.4, in view of the lack of significance of any of the interaction components.

REFERENCES

1. Bartlett, M. S. (1937) .. "Some examples of statistical methods of research in Agriculture and Applied Biology," *Supplement to the Journal of Royal Statistical Society*, **4**, 137–70.

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

3. Crowther, E. M. and Yates, F. (1941) "Fertiliser policy in war time—The fertiliser requirements of arable crops," *Empire Journal of Experimental Agriculture*, **9**, 77–97.

4. Deshpande, R. B. (1933) "Studies in Indian chillies," *Indian Journal of Agricultural Science*, **3**, 219–300.

5. Fisher, R. A. (1921) .. "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society of London*, A, **222**, 309–68.

6. ———— (1950) .. *Statistical Methods for Research Workers*, 11th Ed., Oliver and Boyd, London.

7. ———— and Yates, F. (1948) *Statistical Tables for Biological, Agricultural and Medical Research*, Oliver and Boyd, London.

8. Govande, G. K. (1947) "Linkage relations of lintlessness genes in Asiatic Cottons," *Indian Journal of Genetics and Plant Breeding*, **7**, 80–1 2.

9. Harris, J. A. (1913) .. "On the calculation of intraclass and interclass coefficients of correlation from class moments when the number of possible combinations is large," *Biometrika*, **9**, 446–72.

10. Hutchinson, J. B. (1934) "The genetics of cotton. Part X. The inheritance of leaf shape in Asiatic Gossypiums," *Journal of Genetics*, **28**, 437–513.

11. ———— and Panse, V. G. (1935) "Studies in the technique of field experiments—IV. A study of margin effect in variety trials with cotton and wheat," *Indian Journal of Agricultural Science*, **5**, 671–92.

12. ———— (1937) .. "Studies in plant breeding technique. II. The design of field tests of plant breeding materials," *Indian Journal of Agricultural Science*, **7**, 531–65.

13. Mather, K. (1938) .. *The Measurement of Linkage in Heredity*, Methuen and Co., Ltd., London.

14. Panse, V. G. (1940) .. "Inheritance of quantitative characters and plant breeding," *Journal of Genetics*, **40**, 283.

15. ———— (1945) .. *Manuring of Cotton in India*, Indian Central Cotton Committee, Bombay.

16. Pearson, K. (1930) .. *Tables for Statisticians and Biometricians, Part I,* 3rd Edition, Biometric Laboratory, University College, London.

17. Rhind, D. and Tin, U. (1933) "The effect of temperature on the breakage of rice during milling," *Indian Journal of Agricultural Science*, **3**, 658–62.

18. Shaw, F. J. F., Khan, A. R. and Alam, M. (1931) "Studies in Indian Oilseeds (5). The inheritance of characters in linseed," *Indian Journal of Agricultural Science*, **1**, 1–57.

19. Saxena, P. N. (1951) .. "A simplified method of enumerating latin squares by MacMahon's differential operators. Part II. The 7×7 latin squares," *Journal of the Indian Society of Agricultural Statistics*, **3**, 24–79.

20. Snedecor, G. W. (1946) *Statistical Methods*, 4th Ed. The Iowa State College Press, Ames, Iowa.

21. Sukhatme, P. V. (1941) "Economics of manuring," *Indian Journal of Agricultural Science*, **9**, 325–37.

22. ———— (1953) .. *Sampling Theory of Surveys with Applications*, Indian Society of Agricultural Statistics, New Delhi, and Iowa State College Press, Ames, Iowa, U.S.A.

23. Yates, F. (1937) .. *The Design and Analysis of Factorial Experiments, Technical Communication No.* 35. Imperial Bureau of Soil Science, Harpenden.

APPENDIX I

INTERPOLATION

A dependent variable y may be connected with an independent variable x by a simple algebraic relation such as

$$y = bx + c$$

or

$$y = ax^2, \text{ etc.}$$

The value of y corresponding to any particular value of x can therefore be found without difficulty. However, statistical analysis is concerned with many variables which are not related with one another in this simple manner, *e.g.*, a number and its logarithm. For facilitating practical work with such variables, tables are constructed giving values of the dependent variate (y) corresponding to certain selected values of the independent variate (x). Thus we have Table 2.1 giving the values of the probability corresponding to values of the normal deviate proceeding by steps of 0·1. In such tables, it is possible to find values of y corresponding to values of x which are not included in the table, but which are within its range. The process of finding values of y corresponding to values for x lying intermediate between the tabulated values is known as *interpolation* and so interpolation is sometimes described as the technique of reading between the lines of a numerical table.

Let x_1 and x_2 be any two consecutive values of x given in the table and y_1 and y_2 the corresponding values of y. Then it is clear that an increase in x by $(x_2 - x_1)$ corresponds to an increase in y by $(y_2 - y_1)$, the appropriate signs being taken into account. The value of y corresponding to a value of x lying between x_1 and x_2 may then be obtained by assuming to a first approximation that a change in y is proportional to the change in x throughout the interval. Suppose we want the value of y corresponding to $x_1 + \theta(x_2 - x_1)$, θ being less than 1. With the assumption of proportionality the change in y corresponding to a change of $\theta(x_2 - x_1)$ can be found by the simple rule of three and is given by

$$\theta(x_2 - x_1)\frac{(y_2 - y_1)}{(x_2 - x_1)} = \theta(y_2 - y_1)$$

The quantity $(y_2 - y_1)$ which is known as the *first difference* is denoted by Δy_1 and so the required value of y may be expressed as $y_1 + \theta \Delta y_1$.

Thus knowing that log 65 = 1·81291 and log 66 = 1·81954 we may get log 65·3 by interpolation. Here

$$\Delta y_1 = 1{\cdot}81954 - 1{\cdot}81291 = 0{\cdot}00663$$

and

$$\theta = 0{\cdot}3$$

Hence

$$\begin{aligned}\log(65{\cdot}3) &= 1{\cdot}81291 + (0{\cdot}3)(0{\cdot}00663)\\ &= 1{\cdot}81291 + 0{\cdot}00199\\ &= 1{\cdot}81490\end{aligned}$$

It may be seen easily that the operation can often be done mentally and the process is facilitated by the inclusion of differences Δy in the tables. The above process based on the assumption of a linear relation between x and y within the interval concerned is known as *linear interpolation* or *method of proportional parts.*

It may be pointed out that the value of log 65·3 correct to five decimal places is 1·81491 from which the calculated value differs slightly. This is due to the fact that the assumption of a linear relation between x and y is valid only within a very narrow range of x. The inaccuracies due to this cause are, however, usually negligible and linear interpolation is adequate for most purposes. If it is desired to obtain a closer approximation recourse can be had to more complex formulæ based on more than two tabular values such as for example, the four point interpolation formula explained by Fisher and Yates in the introduction to their *Statistical Tables* (1948).

Sometimes special methods may be employed in interpolation. We shall briefly indicate the methods suitable for interpolation in a number of statistical tables commonly used.

(1) Tables of Normal probability integral and normal deviate (Table 2.1 and Appendix III)—linear interpolation.

(2) Table of values of t (Appendix IV)—linear interpolation.

(3) Table of values of d for Fisher and Behrens' test (Table V_1—*Statistical Tables* by Fisher and Yates). It is found that the relation between the values of d and the reciprocals of the degrees of freedom n_1 and n_2 is approximately linear. For interpolation for intermediate values of n_1 and n_2 the reciprocals of the number of degrees of freedom concerned are therefore used. Thus to interpolate for $n_2 = 19$ ($n_1 = 8$,

$\theta = 15°$) we proceed thus. The reciprocal of 19, *i.e.*, 1/19 lies between 1/12 and 1/24, the reciprocals corresponding to $n_2 = 12$ and $n_2 = 24$. Hence

$$\theta = \frac{\left(\frac{1}{19} - \frac{1}{24}\right)}{\left(\frac{1}{12} - \frac{1}{24}\right)} = \frac{5}{19} = 0{\cdot}263$$

and

$$\Delta y = 2{\cdot}183 - 2{\cdot}077 = 0{\cdot}106$$

Therefore

$$d\,(n_2 = 19) = 2{\cdot}077 + (0{\cdot}263)\,(0{\cdot}106)$$
$$= 2{\cdot}077 + 0{\cdot}028 = 2{\cdot}105$$

(4) Table of values of χ^2-Transformation of probability to logarithms should be used. Thus to obtain the value of χ^2, for $P = 0{\cdot}025$ and $n = 8$, say, we find from the χ^2 table that for $P = 0{\cdot}05$, $\chi^2 = 15{\cdot}507$ and for $P = 0{\cdot}02$, $\chi^2 = 18{\cdot}168$. We take $\log(0{\cdot}05) = -1{\cdot}30103$, $\log(0{\cdot}025) = -1{\cdot}60206$ and $\log(0{\cdot}02) = -1{\cdot}69897$, and find

$$\theta = \frac{\log(0{\cdot}025) - \log(0{\cdot}02)}{\log(0{\cdot}05) - \log(0{\cdot}02)} = \frac{0{\cdot}09691}{0{\cdot}39794}$$
$$= 0{\cdot}2435$$

and

$$\Delta y = 15{\cdot}507 - 18{\cdot}168 = -2{\cdot}661$$

Hence we have, for $P = 0{\cdot}025$,

$$\chi^2 = 18{\cdot}168 + (0{\cdot}2435)\,(-2{\cdot}661)$$
$$= 18{\cdot}168 - 0{\cdot}648$$
$$= 17{\cdot}520$$

To calculate the probability corresponding to a χ^2 value we follow the converse process. Thus to calculate the probability corresponding to $\chi^2 = 17{\cdot}520$ with 8 degrees of freedom we find that it lies between $P = 0{\cdot}05$ and $P = 0{\cdot}02$ corresponding to χ^2 values 15·507 and 18·168. Hence we take

$$\theta = \frac{17{\cdot}520 - 18{\cdot}168}{15{\cdot}507 - 18{\cdot}168} = \frac{0{\cdot}648}{2{\cdot}661} = 0{\cdot}2435$$

and

$$\Delta y = \{\log (0{\cdot}05) - \log (0{\cdot}02)\} = 0{\cdot}39794$$

Hence

$$\begin{aligned} \log P &= \log (0{\cdot}02) + (0{\cdot}2435)(0{\cdot}39794) \\ &= -1{\cdot}69897 + 0{\cdot}09690 \\ &= -\bar{1}{\cdot}60207 \end{aligned}$$

or

$$P = 0{\cdot}025$$

(5) Tables of F (variance ratio)—linear interpolation with transformation of n_1, n_2 to reciprocals.

APPENDIX II

RANDOM NUMBERS—I

One-Digit Random Numbers

Columns:
(1) (2) (3) (4) (5) (6) (7) (8) (9) (10)(11)(12)(13)(14)(15)(16)(17)(18)(19)(20)(21)(22)

3	3	2	6	1	6	8	0	4	5	6	0	8	0	0	7	3	5	4	5	5	4
2	7	0	7	3	6	0	7	5	1	2	4	6	1	9	9	9	7	3	8	1	8
1	3	5	5	3	8	5	8	5	9	8	8	6	7	0	0	3	7	2	6	7	1
5	7	1	2	1	0	1	4	2	1	8	8	6	5	2	2	4	2	8	6	7	3
0	6	1	8	4	4	3	2	5	3	2	3	0	9	2	7	8	5	0	0	9	4
8	7	3	5	2	0	9	6	4	3	8	4	6	7	4	2	9	6	3	7	2	7
2	1	7	6	3	3	5	0	2	5	8	3	3	0	8	7	8	6	1	0	4	8
1	2	8	6	7	3	5	8	0	7	4	4	1	7	5	5	7	3	9	5	2	5
1	5	5	1	0	0	1	3	4	2	9	9	3	2	0	0	8	2	8	0	0	4
9	0	5	2	8	4	7	7	2	7	0	8	1	3	1	3	7	5	6	9	2	9
0	6	7	6	5	0	0	3	1	0	5	5	2	6	7	2	5	1	9	2	7	5
2	0	1	4	8	5	8	8	4	5	1	0	2	3	8	1	7	0	5	0	0	3
3	2	9	8	9	4	0	7	7	2	9	3	2	8	3	4	5	0	3	4	1	4
8	0	2	2	0	2	5	3	5	3	8	6	4	0	6	6	8	9	2	0	8	7
5	4	4	2	0	6	8	7	9	8	3	5	1	7	3	1	8	8	9	2	2	9
1	7	7	6	3	7	1	3	0	4	0	7	0	0	8	8	3	8	7	9	3	1
7	0	3	3	2	4	0	3	5	4	9	7	2	4	3	3	9	3	4	8	6	6
0	4	4	3	1	8	6	6	7	9	9	4	9	7	7	2	7	8	0	5	7	3
1	2	7	2	0	7	3	4	4	5	9	9	6	1	9	9	1	3	9	5	0	9
5	2	8	5	6	6	6	0	4	4	3	8	5	7	0	0	4	5	0	5	0	9
0	4	3	3	4	6	0	9	5	2	6	8	5	0	4	2	4	2	4	6	7	7
1	3	5	8	1	8	2	4	7	6	1	5	5	8	2	7	7	4	4	7	6	2
9	6	4	6	9	2	4	2	4	5	9	7	3	2	8	5	3	8	2	7	5	7
1	0	4	5	6	5	0	4	2	6	1	1	1	2	1	8	7	7	3	4	9	6
3	4	2	5	2	0	5	7	2	7	4	0	5	1	7	2	7	5	9	5	4	1
6	0	4	7	2	1	2	9	6	8	0	2	5	4	6	7	0	2	6	5	5	4
7	6	7	0	9	0	3	0	8	6	3	8	3	9	4	7	4	8	2	5	8	7
1	6	9	2	5	3	5	6	1	6	0	2	7	6	3	9	9	2	2	5	0	4
4	0	0	1	7	4	9	1	6	2	4	8	8	6	7	4	8	3	3	1	2	7
0	0	5	2	4	3	4	8	8	5	2	7	8	4	6	4	5	9	7	9	4	9

RANDOM NUMBERS—II

Two-Digit Random Numbers

Columns: (1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
51	51	00	83	63	22	55	39	65	36	63	70	77	45	85	50
68	97	87	64	81	07	83	73	71	98	16	04	29	18	94	51
30	79	20	69	22	40	98	72	20	56	20	11	72	65	71	08
81	69	40	23	72	51	39	75	17	26	99	76	89	37	20	70
90	60	73	96	53	97	86	37	48	60	82	29	81	30	15	39
46	15	38	26	61	70	04	68	08	02	80	72	83	75	46	30
99	05	48	67	26	43	18	14	23	98	61	67	70	52	85	01
98	35	55	03	36	67	68	49	08	96	21	44	25	27	99	41
11	53	44	10	13	85	57	78	37	06	08	43	63	61	62	42
06	71	95	06	79	88	54	37	21	34	17	68	86	96	83	23
83	45	19	90	70	99	00	14	29	09	34	04	87	83	07	55
49	90	65	97	38	20	46	68	43	28	06	36	49	52	83	51
39	84	51	67	11	52	49	10	43	67	29	70	80	62	80	03
16	17	17	95	70	45	80	44	38	88	39	54	86	97	37	44
13	74	63	52	52	01	41	90	59	59	19	51	85	39	52	85
68	93	60	61	97	22	61	41	47	10	25	62	97	05	31	03
01	07	98	99	46	50	47	91	94	14	63	19	75	89	11	47
74	97	76	38	03	29	63	80	06	54	18	66	09	18	94	06
19	33	53	05	70	53	30	67	72	77	63	48	84	08	31	55
43	70	02	87	40	41	45	59	40	24	13	27	79	26	88	86
95	80	35	14	97	35	33	05	90	35	89	95	01	61	16	96
82	15	94	51	33	41	67	44	43	80	69	98	46	68	05	14
65	31	91	51	80	32	44	61	81	31	96	82	00	57	25	60
85	23	65	09	29	75	63	42	88	07	10	05	24	98	65	63
65	79	20	71	53	20	25	77	94	30	05	39	28	10	99	00
81	06	01	82	77	45	12	78	83	19	76	16	94	11	68	84
00	52	53	43	37	15	26	87	76	59	61	81	43	63	64	61
50	28	11	39	03	34	25	91	43	05	96	47	55	78	99	95
53	32	40	36	40	96	76	84	97	77	72	73	09	62	06	65
69	84	99	63	22	32	98	87	41	60	76	83	44	88	96	07

RANDOM NUMBERS—II *Contd.*

Two-Digit Random Numbers

Columns: (17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)	(31)	(32)
65	87	08	13	50	63	04	23	25	47	51	91	13	52	62	24
43	65	42	78	66	28	55	80	47	46	41	90	08	55	98	78
62	60	53	51	57	32	22	27	12	72	72	27	77	44	67	32
54	96	72	66	86	65	64	60	56	59	75	36	75	46	44	33
46	95	86	19	83	52	47	53	65	00	51	93	51	30	80	05
18	08	51	51	78	57	26	17	34	87	96	23	95	89	99	93
13	37	00	79	68	96	26	60	70	39	83	66	56	62	03	55
40	25	24	73	52	93	70	50	48	21	47	74	63	17	27	27
90	65	77	63	99	25	69	02	09	04	03	35	78	19	79	95
53	09	48	86	28	30	02	35	71	30	32	06	47	93	74	21
40	69	80	97	96	47	59	97	56	33	24	87	36	17	18	16
96	06	68	93	41	69	96	07	97	50	81	79	59	42	37	13
07	76	21	40	24	74	36	42	40	33	04	46	24	35	63	02
27	78	37	06	06	16	25	98	17	78	80	36	85	26	41	77
44	66	88	97	81	26	03	89	39	46	67	21	17	98	10	39
58	91	63	65	99	59	97	84	90	14	79	61	55	56	16	88
00	97	26	16	91	21	32	41	60	22	66	72	17	31	85	33
00	51	72	62	03	89	26	32	35	27	99	18	15	78	12	03
15	00	41	92	25	73	40	38	37	11	05	75	16	98	81	99
30	92	30	45	51	94	69	04	00	84	14	36	37	95	66	39
19	94	91	67	48	57	10	25	19	64	82	84	62	74	29	92
10	70	49	92	05	12	07	23	02	41	46	04	44	31	52	43
23	13	67	95	07	76	30	55	85	66	96	28	28	30	62	58
63	71	54	50	06	44	76	68	45	19	69	59	35	14	82	56
19	29	56	23	27	19	03	69	31	46	29	85	18	88	26	95
39	79	11	28	94	15	52	37	31	61	28	98	94	61	47	03
86	57	77	55	33	62	02	66	42	19	24	94	13	13	38	69
51	26	35	96	29	00	45	33	65	78	12	35	91	59	11	38
07	21	02	84	48	51	97	76	32	06	19	35	22	95	30	19
86	33	49	90	21	60	74	43	33	42	02	59	20	39	84	95

RANDOM NUMBERS—III

Three-Digit Random Numbers

Columns: (1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
642	807	270	546	029	835	828	386	010	216	322	045
790	186	608	897	265	257	276	134	111	614	930	921
435	410	099	205	689	786	313	094	883	382	695	654
218	345	226	433	905	298	385	904	803	854	968	739
263	626	225	267	531	617	134	416	101	081	503	908
296	340	928	403	526	048	138	609	602	807	331	986
835	883	273	307	700	226	101	762	243	049	471	774
058	569	858	422	469	850	647	050	958	217	564	686
452	341	221	191	226	645	614	734	201	633	887	868
757	094	479	348	407	575	377	095	239	675	527	886
149	322	243	302	047	427	832	247	827	331	045	500
639	252	212	801	325	032	719	795	702	411	141	913
648	047	384	924	748	096	704	732	188	117	519	249
573	469	233	958	782	058	134	047	833	897	686	154
879	632	569	615	352	706	787	428	114	305	629	806
676	183	092	227	221	143	760	061	915	362	366	778
235	417	572	035	884	979	255	034	163	387	717	660
749	782	410	000	437	057	074	404	742	573	618	017
364	969	700	077	762	551	646	702	616	517	361	377
406	697	651	823	196	747	742	202	473	049	634	182
749	604	596	495	370	532	952	843	214	125	162	641
355	217	237	436	308	679	812	164	651	367	825	191
392	184	954	851	986	202	732	640	447	515	829	158
627	816	252	418	490	869	332	852	772	438	864	281
709	349	671	505	855	905	549	550	489	101	527	041
876	219	495	418	943	864	864	424	200	164	054	452
687	529	928	822	641	033	948	299	058	732	974	113
836	884	465	379	779	348	217	195	359	232	948	907
262	484	430	807	965	329	181	438	896	614	551	306
406	292	730	137	235	154	714	114	506	375	139	077

RANDOM NUMBERS –III *Contd.*

Three-Digit Random Numbers

Columns: (13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
288	302	957	018	109	053	044	058	849	285	898	732
965	943	462	554	146	318	313	540	090	553	340	096
870	654	605	967	968	085	370	252	657	094	698	056
813	728	351	266	619	151	079	473	763	886	097	893
506	662	573	866	835	785	689	529	992	283	964	416
304	855	222	564	247	726	626	370	569	002	759	996
232	804	271	605	536	173	607	504	020	357	975	079
547	746	272	659	500	487	039	821	904	130	633	750
579	419	722	753	519	962	836	477	033	320	248	817
113	008	777	675	351	395	656	463	578	647	736	959
526	559	446	464	308	899	620	172	197	937	171	423
224	878	732	433	005	993	355	727	995	421	816	713
199	107	231	637	192	397	865	512	072	863	904	818
491	049	367	154	956	911	777	635	102	349	675	392
674	920	950	500	232	289	553	962	844	902	272	428
857	512	776	644	719	415	362	900	851	169	852	504
102	072	305	756	036	523	026	453	977	744	132	319
519	302	585	845	931	731	642	365	632	333	831	719
648	414	669	196	462	612	192	781	061	420	943	216
284	604	628	959	985	898	494	235	935	259	394	334
627	443	283	351	188	946	131	915	229	203	877	693
429	152	062	486	826	147	338	911	530	984	319	317
922	430	588	568	966	031	699	384	192	956	384	030
461	744	365	022	401	067	667	423	957	158	754	211
039	060	686	065	021	808	697	314	744	220	369	155
122	309	242	226	403	441	624	875	320	402	098	046
486	341	451	395	054	268	134	740	902	999	108	084
494	983	308	978	800	884	383	530	025	978	343	269
631	946	604	987	264	596	101	084	367	788	322	601
507	294	628	614	385	914	324	632	069	382	626	724

APPENDIX III

VALUES OF NORMAL DEVIATE (x) FOR CHOSEN LEVELS OF *P*

	·01	·02	·03	·04	·05	·06	·07	·08	·09	·10
·00	2·5758	2·3263	2·1701	2·0537	1·9600	1·8808	1·8119	1·7507	1·6954	1·6449
·10	1·5982	1·5548	1·5141	1·4758	1·4395	1·4051	1·3722	1·3408	1·3106	1·2816
·20	1·2536	1·2265	1·2004	1·1750	1·1503	1·1264	1·1031	1·0803	1·0581	1·0364
·30	1·0152	·9945	·9741	·9542	·9346	·9154	·8965	·8779	·8596	·8416
·40	·8239	·8064	·7892	·7722	·7554	·7388	·7225	·7063	·6903	·6745
·50	·6588	·6433	·6230	·6128	·5978	·5828	·5681	·5534	·5388	·5244
·60	·5101	·4958	·4817	·4677	·4538	·4399	·4261	·4125	·3989	·3853
·70	·3719	·3585	·3451	·3319	·3186	·3055	·2924	·2793	·2663	·2533
·80	·2404	·2275	·2147	·2019	·1891	·1764	·1637	·1510	·1383	·1257
·90	·1130	·1004	·0878	·0753	·0627	·0502	·0376	·0251	·0125	0

The value of *P* for each entry is found by adding the column heading to the value in the left hand margin. The corresponding value of x is the deviation such that the probability of an observation falling outside the range from $-x$ to $+x$ is *P*. For example $P = \cdot 03$ for $x = 2 \cdot 1701$; so that 3 per cent. of normally distributed values will have positive or negative deviations exceeding the standard deviation in the ratio 2·1701 at least.

Note.—The above table is abridged from Table I of Fisher's *Statistical Methods for Research Workers* published by Oliver and Boyd Ltd., Edinburgh, by permission of the author and publishers.

APPENDIX IV

VALUES OF t FOR $P = \cdot 05$ AND $P = \cdot 01$

n'	$P = \cdot 05$	$P = \cdot 01$	n'	$P = \cdot 05$	$P = \cdot 01$
1	12·706	63·657	17	2·110	2·898
2	4·303	9·925	18	2·101	2·878
3	3·182	5·841	19	2·093	2·861
4	2·776	4·604	20	2·086	2·845
5	2·571	4·032	21	2·080	2·831
6	2·447	3·707	22	2·074	2·819
7	2·365	3·499	23	2·069	2·807
8	2·306	3·355	24	2·064	2·797
9	2·262	3·250	25	2·060	2·787
10	2·228	3·169	26	2·056	2·779
11	2·201	3·106	27	2·052	2·771
12	2·179	3·055	28	2·048	2·763
13	2·160	3·012	29	2·045	2·756
14	2·145	2·977	30	2·042	2·750
15	2·131	2·947			
16	2·120	2·921	∞	1·960	2·576

APPENDIX V

VALUES OF F (VARIANCE RATIO)

5 PER CENT. POINTS OF e^{2s}

n_2 \ n_1	1	2	3	4	5	6	8	12	24	∞
1	161·40	199·50	215·70	224·60	230·20	234·00	238·90	243·90	249·00	254·30
2	18·51	19·00	19·16	19·25	19·30	19·33	19·37	19·41	19·45	19·50
3	10·13	9·55	9·28	9·12	9·01	8·94	8·84	8·74	8·64	8·53
4	7·71	6·94	6·59	6·39	6·26	6·16	6·04	5·91	5·77	5·63
5	6·61	5·79	5·41	5·19	5·05	4·95	4·82	4·68	4·53	4·36
6	5·99	5·14	4·76	4·53	4·39	4·28	4·15	4·00	3·84	3·67
7	5·59	4·74	4·35	4·12	3·97	3·87	3·73	3·57	3·41	3·23
8	5·32	4·46	4·07	3·84	3·69	3·58	3·44	3·28	3·12	2·93
9	5·12	4·26	3·86	3·63	3·48	3·37	3·23	3·07	2·90	2·71
10	4·96	4·10	3·71	3·48	3·33	3·22	3·07	2·91	2·74	2·54
11	4·84	3·98	3·59	3·36	3·20	3·09	2·95	2·79	2·61	2·40
12	4·75	3·88	3·49	3·26	3·11	3·00	2·85	2·69	2·50	2·30
13	4·67	3·80	3·41	3·18	3·02	2·92	2·77	2·60	2·42	2·21
14	4·60	3·74	3·34	3·11	2·96	2·85	2·70	2·53	2·35	2·13
15	4·54	3·68	3·29	3·06	2·90	2·79	2·64	2·48	2·29	2·07
16	4·49	3·63	3·24	3·01	2·85	2·74	2·59	2·42	2·24	2·01
17	4·45	3·59	3·20	2·96	2·81	2·70	2·55	2·38	2·19	1·96
18	4·41	3·55	3·16	2·93	2·77	2·66	2·51	2·34	2·15	1·92
19	4·38	3·52	3·13	2·90	2·74	2·63	2·48	2·31	2·11	1·88
20	4·35	3·49	3·10	2·87	2·71	2·60	2·45	2·28	2·08	1·84
21	4·32	3·47	3·07	2·84	2·68	2·57	2·42	2·25	2·05	1·81
22	4·30	3·44	3·05	2·82	2·66	2·55	2·40	2·23	2·03	1·78
23	4·28	3·42	3·03	2·80	2·64	2·53	2·38	2·20	2·00	1·76
24	4·26	3·40	3·01	2·78	2·62	2·51	2·36	2·18	1·98	1·73
25	4·24	3·38	2·99	2·76	2·60	2·49	2·34	2·16	1·96	1·71
26	4·22	3·37	2·98	2·74	2·59	2·47	2·32	2·15	1·95	1·69
27	4·21	3·35	2·96	2·73	2·57	2·46	2·30	2·13	1·93	1·67
28	4·20	3·34	2·95	2·71	2·56	2·44	2·29	2·12	1·91	1·65
29	4·18	3·33	2·93	2·70	2·54	2·43	2·28	2·10	1·90	1·64
30	4·17	3·32	2·92	2·69	2·53	2·42	2·27	2·09	1·89	1·62
40	4·03	3·23	2·84	2·61	2·45	2·34	2·18	2·00	1·79	1·51
60	4·00	3·16	2·76	2·52	2·37	2·25	2·10	1·92	1·70	1·39
120	3·92	3·07	2·68	2·45	2·29	2·17	2·02	1·83	1·61	1·25
∞	3·84	2·99	2·60	2·37	2·21	2·09	1·94	1·75	1·52	1·00

Lower 5 per cent. points are found by interchange of n_1 and n_2, *i.e.*, n_1 must always correspond with the greater mean square.

VALUES OF F (VARIANCE RATIO)

1 PER CENT. POINTS OF e^{2z}

n_2 \ n_1	1	2	3	4	5	6	8	12	24	∞
1	4052	4999	5403	5625	5764	5859	5981	6106	6234	6366
2	98·49	99·01	99·17	99·25	99·30	99·33	99·36	99·42	99·46	99·50
3	34·12	30·81	29·46	28·71	28·24	27·91	27·49	27·05	26·60	26·12
4	21·20	18·00	16·69	15·98	15·52	15·21	14·80	14·37	13·93	13·46
5	16·26	13·27	12·06	11·39	10·97	10·67	10·27	9·89	9·47	9·02
6	13·74	10·92	9·78	9·15	8·75	8·47	8·10	7·72	7·31	6·88
7	12·25	9·55	8·45	7·85	7·46	7·19	6·84	6·47	6·07	5·65
8	11·26	8·65	7·59	7·01	6·63	6·37	6·03	5·67	5·28	4·86
9	10·56	8·02	6·99	6·42	6·06	5·80	5·47	5·11	4·73	4·31
10	10·04	7·56	6·55	5·99	5·64	5·39	5·06	4·71	4·33	3·91
11	9·65	7·20	6·22	5·67	5·32	5·07	4·74	4·40	4·02	3·60
12	9·33	6·93	5·95	5·41	5·06	4·82	4·50	4·16	3·78	3·36
13	9·07	6·70	5·74	5·20	4·86	4·62	4·30	3·96	3·59	3·16
14	8·86	6·51	5·56	5·03	4·69	4·46	4·14	3·80	3·43	3·00
15	8·68	6·36	5·42	4·89	4·56	4·32	4·00	3·67	3·29	2·87
16	8·53	6·23	5·29	4·77	4·44	4·20	3·89	3·55	3·18	2·75
17	8·40	6·11	5·18	4·67	4·34	4·10	3·79	3·45	3·08	2·65
18	8·28	6·01	5·09	4·58	4·25	4·01	3·71	3·37	3·00	2·57
19	8·18	5·93	5·01	4·50	4·17	3·94	3·63	3·30	2·92	2·49
20	8·10	5·85	4·94	4·43	4·10	3·87	3·56	3·23	2·86	2·42
21	8·02	5·78	4·87	4·37	4·04	3·81	3·51	3·17	2·80	2·36
22	7·94	5·72	4·82	4·31	3·99	3·76	3·45	3·12	2·75	2·31
23	7·88	5·66	4·76	4·26	3·94	3·71	3·41	3·07	2·70	2·26
24	7·82	5·61	4·72	4·22	3·90	3·67	3·36	3·03	2·66	2·21
25	7·77	5·57	4·68	4·18	3·86	3·63	3·32	2·99	2·62	2·17
26	7·72	5·53	4·64	4·14	3·82	3·59	3·29	2·96	2·58	2·13
27	7·68	5·49	4·60	4·11	3·78	3·56	3·26	2·93	2·55	2·10
28	7·64	5·45	4·57	4·07	3·75	3·53	3·23	2·90	2·52	2·06
29	7·60	5·42	4·54	4·04	3·73	3·50	3·20	2·87	2·49	2·03
30	7·56	5·39	4·51	4·02	3·70	3·47	3·17	2·84	2·47	2·01
40	7·31	5·18	4·31	3·83	3·51	3·29	2·99	2·66	2·29	1·80
60	7·08	4·98	4·13	3·65	3·34	3·12	2·82	2·50	2·12	1·60
120	6·85	4·79	3·95	3·48	3·17	2·96	2·66	2·34	1·95	1·38
∞	6·64	4·60	3·78	3·32	3·02	2·80	2·51	2·18	1·79	1·00

Lower 1 per cent. points are found by interchange of n_1, n_2, *i.e.*, n_1 must always correspond with the greater mean square.

APPENDIX VI

VALUES OF X^2

PROBABILITY

n	·99	·98	·95	·90	·80	·70	·50	·30	·20	·10	·05	·02	·01	·001
1	·000157	·000628	·00393	·0158	·0642	·148	·455	1·074	1·642	2·706	3·841	5·412	6·635	10·827
2	·0201	·0404	·103	·211	·446	·713	1·386	2·408	3·219	4·605	5·991	7·824	9·210	13·815
3	·115	·185	·352	·584	1·005	1·436	2·366	3·665	4·642	6·855	7·815	9·837	11·341	16·268
4	·297	·429	·711	1·064	1·649	2·195	3·357	4·878	5·989	7·779	9·488	11·668	13·277	18·465
5	·554	·752	1·145	1·610	2·343	3·000	4·351	6·064	7·289	9·236	11·070	13·388	15·086	20·517
6	·872	1·134	1·635	2·204	3·070	3·828	5·348	7·231	8·558	10·645	12·592	15·033	16·812	22·457
7	1·239	1·564	2·167	2·833	3·822	4·671	6·346	8·383	9·803	12·017	14·067	16·622	18·475	24·322
8	1·646	2·032	2·733	3·490	4·594	5·527	7·344	9·524	11·030	13·362	15·507	18·168	20·090	26·125
9	2·088	2·532	3·325	4·168	5·380	6·393	8·343	10·656	12·242	14·684	16·919	19·679	21·666	27·877
10	2·558	3·059	3·940	4·865	6·179	7·267	9·342	11·781	13·442	15·987	18·307	21·161	23·209	29·588
11	3·053	3·609	4·575	5·578	6·989	8·348	10·341	12·899	14·631	17·275	19·675	22·618	24·725	31·264
12	3·571	4·178	5·226	6·304	7·806	9·034	11·340	14·011	15·812	18·549	21·026	24·054	26·217	32·909
13	4·107	4·765	5·892	7·042	8·634	9·926	12·340	15·119	16·985	19·812	22·362	25·472	27·688	34·528
14	4·660	5·368	6·571	7·790	9·467	10·821	13·339	16·222	18·151	21·064	23·685	26·873	29·141	36·123
15	5·229	5·985	7·261	8·547	10·307	11·721	14·339	17·322	19·311	22·307	24·996	28·259	30·578	37·697

16	5·812	6·614	7·962	9·312	11·152	12·624	15·338	18·418	20·465	23·542	26·298	29·633	32·000	39·252
17	6·408	7·255	8·672	10·085	12·002	13·531	16·338	19·511	21·615	24·769	27·587	30·995	33·409	40·790
18	7·015	7·906	9·390	10·865	12·857	14·440	17·338	20·601	22·760	25·989	28·869	32·346	34·805	42·312
19	7·633	8·567	10·117	11·651	13·716	15·352	18·338	21·689	23·900	27·204	30·144	33·687	36·191	43·820
20	8·260	9·237	10·851	12·443	14·578	16·266	19·337	22·775	25·038	28·412	31·410	35·020	37·566	45·315
21	8·897	9·915	11·591	13·240	15·445	17·182	20·337	23·858	26·117	29·615	32·671	36·343	38·932	46·797
22	9·542	10·600	12·338	14·041	16·314	18·101	21·337	24·939	27·301	30·813	33·924	37·659	40·289	48·268
23	10·196	11·293	13·091	14·848	17·187	19·021	22·337	26·018	28·429	32·007	35·172	38·968	41·638	49·728
24	10·856	11·992	13·848	15·659	18·062	19·943	23·337	27·096	29·553	33·196	36·415	40·270	42·980	51·179
25	11·524	12·697	14·611	16·473	18·940	20·867	24·337	28·172	30·675	34·382	37·632	41·566	44·314	52·620
26	12·198	13·409	15·379	17·292	19·820	21·792	25·336	29·246	31·795	35·563	38·885	42·856	45·642	54·052
27	12·879	14·125	16·151	18·114	20·703	22·719	26·336	30·319	32·912	36·741	40·113	44·140	46·963	55·476
28	13·565	14·847	16·928	18·939	21·588	23·647	27·336	31·391	34·027	37·916	41·337	45·419	48·278	56·893
29	14·256	15·574	17·708	19·768	22·475	24·577	28·336	32·461	35·139	39·087	42·557	46·693	49·588	58·302
30	14·953	15·306	18·493	20·599	23·364	25·508	29·336	33·530	36·250	40·256	43·773	47·962	50·892	59·703

For larger values of n, the expression $\sqrt{2\chi^2} - \sqrt{2n-1}$ may be used as a normal deviate with unit variance.

Note.—The above table is reprinted from Table III of Fisher's *Statistical Methods for Research Workers* published by Oliver and Boyd Ltd., Edinburgh, by permission of the author and publishers.

APPENDIX VII

Notes on Computation

The aim of the computer should be to obtain results consistent with the data with a minimum amount of labour. While some carry the calculations to as many figures as possible, thus wasting labour, others make approximations in intermediate stages and give the final results to a considerable number of figures, giving the results a false appearance of accuracy. The computer should know exactly how many figures he should carry at every stage and be able to assess the accuracy of the result. The following notes would serve as a guide.

The Fundamental Operations

(1) *Addition.*—When some of the numbers to be added are correct to fewer decimal figures than others the result of addition would not be correct to more figures than in the least accurate of them. Hence the following rule should be followed. Round off the more accurate numbers to contain one more decimal figure than what are contained in the least accurate number, add, and round off the result to as many figures as the least accurate number.

Example :

To add 14·381, 6·237, 3·5026 and 8·17 given that each number is correct to its last figure. We add:

$$\begin{array}{r} 14\cdot381 \\ 6\cdot237 \\ 3\cdot503 \\ 8\cdot170 \\ \hline 32\cdot291 \\ \hline \end{array}$$

The result, rounded off to two decimal figures, is 32·29.

(2) *Subtraction.*—The numbers should be rounded off to as many decimal places as the number which has the least number of decimal places correct. The result will be correct to as many decimal places.

Example :

To subtract 12·2431 from 128·37 we have

$$128\cdot37 - 12\cdot24 = 116\cdot13$$

Significant figures.—The question of accuracy in case of products and quotients involves the idea of the number of significant figures.

The number of significant figures is the number of figures counting from the first figure not zero and excluding terminal zeroes when they are used to fill the place of unknown or discarded digits. Thus 237, 2·37 and 0·00237 all contain three significant figures. A number such as 23700 will also be said to contain only three significant figures if it is obtained by rounding off some other number such as 23721. However if it is known to be correct to the last figure it will be said to contain five significant figures. The correct way to represent the former number is to write it as $2{\cdot}37\times10^4$ and the latter as $2{\cdot}3700\times10^4$.

It should be noted that, if as a result of a subtraction a number of figures from the left disappear the result has a smaller number of significant figures. Thus the result of the subtraction (23491·3 − 23389·4), *viz.*, 101·9 contains only four significant figures though the subtraction involved numbers containing six significant figures.

(3) *Multiplication.*—When two or more numbers containing differing number of significant figures are to be multiplied, numbers with the larger number of significant figures should be rounded off to contain one more figure than the least accurate number and the product found. The product should then be rounded off to contain as many figures as the least accurate number.

Example :

To multiply 3429, 2·42 and 1·4 each correct to the last place, we round off the first number to three figures and multiply. We get

$$3430\times2{\cdot}42\times1{\cdot}4 = 11620{\cdot}84$$

Rounding it off to two figures, it will be written as $1{\cdot}2\times10^4$.

(4) *Division.*—Round off the more accurate number to contain one more figure than the less accurate. Result will be correct to as many figures as the less accurate number.

Example :

To divide 25·4372 by 12·1 we shall divide 25·44 by 12·1 and give the result to three significant figures, *i.e.*, quotient = 2·10

Powers and Roots

(5) *Powers.*—No more significant figures should be retained in the power than are contained in the number.

(6) *Roots.*—As many figures should be retained in the root as there are in the number. For this the root should be initially calculated to one more figure and then the result rounded off to contain one figure less.

STATISTICAL OPERATIONS

(7) *Mean.*—The mean is generally more correct than the values averaged, as the inaccuracies in the individual values tend to be of opposite signs. The mean should therefore be calculated to one more decimal place than what are given in the individual values.

(8) *Standard deviation.*—No more decimal places may be retained in the crude sum of squares than are contained in the original numbers, and the correction factor calculated to one more decimal place. While calculating the mean square by dividing by the number of degrees of freedom one more decimal place may be retained. The standard deviation might be given to as many significant figures as the mean square. However, it is not very useful to retain more than three figures as that is sufficient for making probability statements or specifying fiducial limits.

(9) *Regressions.*—While solving sets of equations and calculating the covariance matrices, the data should be treated as exact numbers; as many figures as could possibly be carried should be carried at each stage of the calculation and no rounding off or dropping of digits resorted to, subject to the limitations of computational devices (machines, etc.). The final result should be rounded off to contain as many figures as the original data.

No general rules can be laid down for securing the necessary accuracy in more complex operations. The accuracy that can be attained will have to be determined in the particular cases as required. It should be remembered, however, that the accuracy of the primary data sets a limit to the accuracy of the results of computations.

SUBJECT INDEX

A

Adjusted error sum of squares, 223.
Adjusted mean yield, 224.
Adjustment factor, 276.
Analysis of covariance, 220.
  extension of, 228.
Analysis of incomplete observations, 229.
Analysis of variance, 63–66.
  adjusted, 224.
  unweighted, 289.
  weighted, 289–292.
Ancillary information, use of, 217.
Ancillary observations, 310, 330.
Ancillary variate, 220.

B

Balanced lattice, 268–278.
  approximate F test in, 277.
  efficiency of, 278.
Balancing, concept of, 267.
Bartlett's technique for missing plots, 230.
Bartlett's test, for homogeneity of variances, 248–249, 282.
Bias, 41.
Biased estimate, 40.
Binomial distribution, 32–34.
Biometry, 1.
Blocks, arrangement of, 305–306.
  size and shape of, 159–160.
Border effect, 304–305.

C

Central value, 10.
Characters, qualitative, 1.
  quantitative, 1.
Chess-board arrangement of plots, 144.
Chi-square (χ^2), 72.
  partition of, 86.
  table of, 352–353.
  test of goodness of fit, 72.

C—(*contd.*)

  test of heterogeneity, 81.
  test of homogeneity of variances, 248–249, 282.
  test of independence, 75, 79, 83.
  use of, to determine size of experiments, 85.
Class interval, 3.
  choice of, 8.
Class value, 7.
Coding, 18.
Coefficient of variation, 14, 141.
Compact family block design, 245–250.
Computation,
  methods of, 15–21.
  notes on, 354–356.
Concomitant variate, 220.
Confidence coefficient, 45.
  interval, 45.
  limits, 44–45.
Confounding, 176–198.
  balanced, 191.
  efficiency of, 187–189, 197–198.
  partial, 184.
Consistency, criterion of, 94.
Consistency of response, 281.
Contingency tables, 75, 101.
Contour bunding, 308.
Correction factor, 65, 221, 249.
Correction, Sheppard's, 16, 105.
Correlation, 100.
  interclass, 131.
  intraclass, 131.
  multiple, 127.
  negative, 100.
  partial, 119–121.
  partial, significance of, 121.
  positive, 100
  test of significance of, 109.
  transformation to z of, 109–115.
Covariance, 103.
  analysis of, 220.
  matrix, 123.
Covariance procedure, efficiency of, 227.

C—(*contd.*)

Critical difference, 155–157.
Cultivators' fields, experiments on, 317.
  cost of, 324.
  choice of treatments in, 326.
  design for, 321–323.
  optimum number of places and fields for, 323.
Cumulative frequency, 24.
Curvilinear regression, 128.

D

d test (Fisher and Behren's), 67–69.
Decoding, 18.
Degrees of freedom, 48.
Design of experiments, 1.
  on cultivators' fields, 321–323.
Dichotomous classification, 35.
Dispersion, 10.
  measures of, 12–14.
Distribution,
  binomial, 32–34.
  multinomial, 35.
  normal, 25–26.
  of sample means, 42–44.
Double lattice design, 254–267.

E

Economics of manuring, 300.
Effective error mean square, 267–277.
Effective number of replications, 238.
Efficiency, criterion of, 94.
Efficiency of designs, 150–151.
  of balanced lattice, 278.
  of confounding, 187–189, 197–198.
  of covariance procedure, 227.
  of randomized blocks, 159–160.
  of simple lattice, 257.
  of split plots, 206.
Efficient utilisation of seed, 250–253.
Eliminated variate, 121.
Emerson's method of linkage estimation, 95–96.
Error mean square, 63.
Estimation, 94.
Expected frequencies, 72, 76, 83.
  for double backcross, 94.
  for selfed heterozygote, 95.

E—(*contd.*)

Expected value, 40.
Experimental designs, 150.
  choice of, 301–302.
Experimental error, 141.

F

F, tables of, 350–351.
F test, 66.
Factorial concept, 166–167.
Factorial experiments, 166–175.
  advantages of, 174.
Fertility,
  contour map, 140.
  gradient, 306–307.
First difference, 339.
Frequency, 2.
  curve, 9, 22.
  cumulative, 24.
  distribution, 2.
  polygon, 9.

G

Genetic experiments, size of, 85.
Gross plot size, 306.
Groups of experiments, 280–297, 324.

H

Haveli system of cultivation, 308.
Heterogeneity,
  of experimental errors, 288–297.
  of interactions, 286–288.
  soil, control of, 176–177.
  test of, between segregating families, 81–83.
  test of, in linkage data, 93.
Histogram, 9.
Homogeneity of variances, test of, 248–249, 282.

I

I-table, 195.
Incomplete block designs, 253–279.
  balanced, 267–278.
  use of, 278–279.

I—(*contd.*)

Incomplete observations, analysis of, 229.
Independent samples, 56–57.
Information, amount of, 151, 191.
  interblock, 255.
Interaction, 64, 171–173.
  of first, second or higher orders, 183.
  heterogeneity of, 286–288.
  test of, in groups of experiments, 288.
Interclass correlation, 131.
Interpolation, 339–342.
Intrablock error, 262.
Intraclass correlation, 131, 150, 160, 202.

J

J-table, 195.

K

Kurtosis, 10.

L

Latin square, 161.
  comparison with randomized blocks, 165.
  conjugate of, 162.
Lattice designs,
  double, 254–267.
  balanced, 268–278.
Least squares, principle of, 114.
Linear component of response, 296.
Linear interpolation, 29, 340.
Linear regression, 114.
Linkage, 89.
  Chi-square due to, 91.
  detection of, 88.
  estimation of, 93–98.
  heterogeneity of, 93.
Local control, 148, 150, 176, 199, 264, 253, 322.

M

Main effect, 171.
Main plot, 199.
Maximum likelihood, method of, 96–98.
Mean, 10.
Mean deviation, 13.

M—(*contd.*)

Mean product moment, 103.
Mean square, 63.
Measurement of plot dimensions, 305.
Median, 11.
Missing plots, 229.
  adjustment of averages with, 232.
  Bartlett's technique for, 230.
  bias in treatment sum of squares with, 235.
  extension of method for, to more than one missing value, 236.
  in latin square, 239–240.
  iterative method for, 237.
  method of substitution for, 234.
  standard errors of comparisons with, 234.
Mode, 11.
Multinomial distribution, 35.
Multiple correlation, 127.

N

Net plot size, 306.
Non-experimental margins, 304.
Normal curve, 10, 25–26.
Normal deviate, 26, 52, 55, 70, 73, 86.
  table of, 348.
Normal probability integral table, 27.
Null hypothesis, 53, 145.

O

Observation plots, 301.
Orthogonal functions of frequencies, 89.
Orthogonal latin squares, 268.

P

Plots, size, shape and arrangement of, 159–160, 303–304.
Polynomial, 128.
Population, 1.
Probability, 27.
Probability integral, 22–24.
Product ratio method of estimation of linkage, 98.
Progeny row trials, replicated, 241.
Proportional parts, method of, 29, 340.
Pseudo-variate, 231.

Q

Qualitative characters, 1.
Quantitative characters, 1.

R

Random numbers,
  use of, 38–40.
  tables of, 343–347.
Random sampling, 36–37, 319.
Randomization, 143, 322.
Randomized blocks, 150, 152–161.
  efficiency of, 159–160.
  vs. latin square, 165.
Range, 13.
Ranks, 311.
Recombination fraction, 93.
Regression, 113.
  curvilinear, 127–130.
  curvilinear, significance of, 129.
  equation, 113.
  linear, 114.
  linear, significance of, 116–118.
  partial (multiple), 121–123.
  partial, significance of, 123.
  relation with correlation, 115–116.
Replicated progeny rows, 241–245.
Replication, 142–143, 322.
  effective number of, 238.
  number of, in randomized blocks, 157–159.
Residual variance, 117.
Response curve, 300.

S

Saline patches, 228, 312.
Sample, 36.
Sampling, 36, 310.
  biased, 40–41.
  for ancillary observations, 310.
  random, 36–37, 319.
  sub-, 42, 320.
  variance, 44.
Scores, 228, 311.
Sheppard's correction, 16, 105.
Significance, level of, 53.
  test of, 51, 53.
Significant figures, 354.

S—(*contd.*)

Simple lattice design, 254–267.
  analysis of, 257–267.
  approximate *F* test in, 265.
  efficiency of, 267.
  variances in, 266.
Skewness, 10.
Spacing, experiments on, 308.
Split-plot design, 199–211.
  advantages of, 202.
  efficiency of, 206–207.
  extension of, 210–211.
  standard errors in, 207–210.
Standard deviation, 13.
Standard error, 44.
  of proportion, 48.
  of linear function of independent variates, 47.
  of sum and difference of means, 45–47.
Statistical control of error, 217.
Statistical methods, 1.
Statistics, 1.
Strip-plot design, 211.
  standard errors in, 216.
Subplots, 119.
Sum of products of deviations, 106, 221.
Sum of squares of deviations, 63.
  for a linear function, 186.

T

t, table of, 349.
t test, 57–63.
  in paired samples, 62–63.
Total correlation coefficient, 120.
Transformations, 315–316.
Treatments, 137.

U

Uniformity, of site, 302.
  of operations, 309.
Uniformity trials, 137, 217, 302, 303.
Unit of sampling, 310.
Unweighted analysis of variance, 289.

V

Variance, 14.
  analysis of, 63–66.
  residual, 117.
  sample estimate of, 48.

V—(*contd.*)

Variance ratio, 66.
  tables of, 350–351.
Variate, 1.
  dependent, 114, 339.
  independent, 114, 339.
Variation,
  coefficient of, 14, 141.
  continuous, 1.
  discrete, 1.

W

Weighted analysis of variance, 289.
Weighted mean, 112.
Weighting factor, 263.

Y

Yates' correction, 78.

Z

z transformation, 109, 121.

1343-56 Printed at The Bangalore Press, Bangalore City,
by C. Vasudeva Rao, Superintendent.