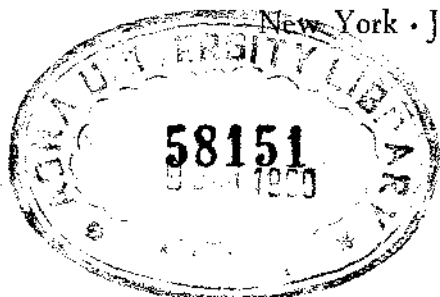# Technological
Applications
of Statistics

L. H. C. TIPPETT
Head of the Mechanical Processing Division
British Cotton Industry Research Association

Foreword

LIKE OTHER UNIVERSITIES, THE MASSACHUSETTS INSTITUTE OF
Technology finds it desirable from time to time to invite outsiders to
give lectures to its students. It is now more than ten years since one
such series of lectures, on experimental statistics, was given at the
Institute by Mr. L. H. C. Tippett, Statistician to the British Cotton
Industry Research Association. This first set of lectures was successful,
and we were happy to invite Mr. Tippett to give a second group on the
tenth anniversary of the first, that is, in the spring of 1948.

Another thing which we find desirable from time to time is to share
the benefits of such lectures with a wider audience. We are therefore
happy that the contribution of Mr. Tippett to this field now finds a
wider audience through the current publication by Williams and Norgate
Ltd., and John Wiley & Sons. We are pleased to have been associated
with Mr. Tippett in the original venture and glad that it has had this
outcome.

<div style="text-align:right">

JOHN E. BURCHARD
Dean of Humanities
Massachusetts Institute of Technology

</div>

Cambridge, Massachusetts

v

Preface

THIS BOOK IS A "WRITE-UP" OF A COURSE OF LECTURES GIVEN AT the Massachusetts Institute of Technology to a mixed audience consisting of industrialists, some of whom had little more than a general appreciation of statistics, students, and practised statisticians working largely in industry. I could not expect to satisfy every member of this audience completely but hoped that almost everyone got enough from the lectures to make him feel that he had not wasted his time. I offer this book to a wider audience in the same spirit.

The book is perhaps best regarded as an introduction and companion to a systematic text-book or course of study in applied statistics. For the beginner I have tried to present the logic of the statistical methods employed and of their application to technological problems. Parts of the book may be found difficult by the beginner, but he should not linger over them at a first reading; he should read on and be content to achieve some understanding of the general ideas. For the student in full course particular attention is paid to the basic assumptions and their implications for the technologist, and to points that arise in the practical application of the methods. The exposition is developed from particular examples, and the development is more through arithmetical procedures than through mathematical proofs. I believe that it is easier for the technician to attain a "feel" for the subject by the arithmetical method—the mathematics can follow later. But the method can only be really successful if the student works through the examples. They are treated as problems in technology as well as in statistics.

Some of the data for the examples have been taken from other publications, which are fully specified; special mention should be made of *Statistical Methods in Industry* (published by the British Iron and Steel Federation), which has been much drawn upon for data. I cordially thank the various authors and publishers for permission to use them. Other data have arisen in the course of my work at the Shirley Institute, Manchester, England; I thank those colleagues who have co-operated in producing these data and the Director of the British

vii

Cotton Industry Research Association for allowing me to use them.

I am grateful to Professor Harold A. Freeman for seeing this book through the press and to Mr. Louis C. Young for reading the proofs.

L. H. C. T.

September, 1950

Contents

ix

PART I. THE ROUTINE CONTROL OF QUALITY

Chapter 1. THE MEASUREMENT OF QUALITY

The control of quality presupposes its measurement. This has two aspects, technical and statistical.

Technical Measures of Quality

We understand by quality any characteristic of the products of a factory, of intermediate products, or of the raw material that is of interest. It may be a quality in the popular sense of the term—something of which there can not be too much; it may be something like a size which is required to be neither more than nor less than a certain value, or it may be something that is undesirable such as the appearance of a flaw; it may be described qualitatively or in terms of a numerical measure.

The specification and measurement of the quality of individual things is entirely a technical problem. Some characteristics, like the dimensions of machined parts, can be directly measured and their measures directly interpreted. The designer knows and can say how big the parts must be, and they can be readily measured by gauges. Other qualities are harder to define or can not be measured quickly or cheaply enough for purposes of production. For example, durability in use is usually a complex quality depending on the type of experience the article has to undergo. The experience may be made up of shocks, abrasion, vibration, corrosion, and so on, each element being describable by a range of forces, frequencies, and so on, and the elements being combined in certain proportions. Even if in use all the articles of a given kind undergo the same experience it is very hard to devise a laboratory test to measure the corresponding durability precisely, and if the experience in use varies from one article to another the case is almost hopeless. In circumstances like these laboratory tests are devised, the results of which are presumed to be approximately related to the desired quality. However, all too often such tests are chosen without adequate investigation, the presumption is made all too easily, and the technician is unable to state what values of test results correspond to satisfactory performance. The technical problem of devising good

1

measures is often exceedingly difficult, and the technician has to do the best he can in the circumstances; but the methods of control described herein are effective only if the measure of quality is technically suitable. The technician must be able to state precisely what he requires before the statistician can say precisely what to do. We shall assume that the technical part of the problem is solved.

Frequency Distribution

When articles are mass-produced, it is not practicable to deal with individuals. From the individual measurements we need to evolve measures of the quality of the bulk; and this is where the statistician enters the field.

Consider the data in Table I. They are the results of tests of count (a measure of fineness, being the number of hanks of 840 yards per pound) made on 200 test specimens ("skeins" or "leas") taken from a batch of a certain cotton yarn L. From the statistical viewpoint they

TABLE I

COUNT MEASURED ON SINGLE LEAS OF COTTON YARN L

36.6	38.1	35.0	37.3	36.1	38.7	37.9	37.8	38.2	36.7
38.5	37.6	37.8	36.3	36.6	36.2	37.8	37.3	37.4	35.4
35.1	37.9	36.0	38.2	38.2	38.4	35.1	36.2	36.4	36.9
37.3	37.9	36.5	36.1	38.3	38.6	38.4	37.3	37.7	37.3
36.4	36.6	37.7	37.2	38.8	38.4	35.8	38.9	37.2	37.9
38.3	37.4	38.3	38.4	37.2	36.9	36.5	39.0	36.5	36.9
37.2	35.4	39.6	39.6	37.9	36.2	37.4	37.2	36.6	37.4
36.6	38.5	38.1	37.5	36.6	37.5	36.2	38.0	36.1	37.0
38.0	37.3	36.9	36.0	38.1	36.4	34.9	37.0	36.4	37.1
38.7	36.3	37.3	37.5	37.1	35.8	37.0	37.0	37.7	36.6
37.4	38.1	36.4	38.3	37.3	37.7	37.3	36.0	35.2	38.4
36.7	35.0	37.9	37.2	37.6	38.2	36.1	37.7	36.3	36.1
36.1	37.8	37.2	38.2	39.6	37.3	37.1	37.8	38.0	36.3
37.1	38.3	37.3	37.3	37.5	36.6	36.8	37.2	36.7	37.8
36.5	37.0	36.6	38.2	36.2	37.6	36.2	35.8	36.2	38.1
35.4	38.2	37.3	37.2	37.5	37.8	36.5	37.9	37.4	36.5
37.6	37.0	37.0	37.8	38.1	35.6	37.5	38.2	36.6	37.8
39.5	35.6	36.7	38.0	37.2	37.5	37.3	37.9	37.2	37.7
35.1	37.5	37.6	38.1	37.1	36.6	37.4	38.1	37.1	37.8
37.0	37.2	37.0	37.5	37.3	37.1	37.3	35.8	37.3	36.9

might just as well be a dimension of 200 mass-produced parts, the life or efficiency of 200 electric lamps, or the strengths of 200 test specimens of steel. In statistical language the leas of Table I are termed *individuals*. The count is the quality of the individual lea; the problem is how to specify the quality of the batch as a whole.

You will notice that count varies widely—from 34.9 to 39.5—and that the order of the results has no apparent significance; large values are indiscriminately followed by small, medium, or large values. In order to see the data as a whole we carry out two processes. We put the results into an order and summarise them. In order of magnitude the values are

34.9	35.0	35.0	35.1	35.1	35.1	35.2	35.4	35.4	35.4	. . . .
37.2	37.2	37.2	37.3	37.3	37.3	37.3	37.3	37.3	37.3	. . . .
38.6	38.7	38.7	38.8	38.9	39.0	39.5	39.6	39.6	39.6	

the first row giving the first ten, the second row ten from near the centre of the range, and the third row the last ten. We notice that the values tend to be spread out at the extremes of the series, quite large gaps occurring sometimes between one value and the next (e.g., 39.0 is followed by 39.5), whereas towards the centre the values tend to be more bunched, the same value occurring many times and no gap between consecutive values exceeding 0.1 (e.g., the last 37.2 is followed by 37.3). We see this more clearly if we summarise by ignoring small changes in the value; after all, compared with a variation extending from 34.9 to 39.6, it does not much matter whether any given value is, say, 35.4 or 35.5 or 35.3. Accordingly we may divide the whole range of variation into sub-ranges of 0.5 count, letting the first boundary separate 34.9 from 35.0 by putting it at 34.95, putting the second at 35.45, and so on; some of these boundaries are marked in the above series. If this is done for the whole series and the numbers of values between the boundaries are counted and recorded, a *frequency table* as shown in Table II results. This is a tabular representation of a *frequency distribution*. As an alternative to ordering the values, a blank form like Table II, without the frequencies, can be made and a dot or stroke be put opposite the appropriate sub-range for each value in Table I. The dots or strokes can then be counted to give the frequencies. If you are not already familiar with frequency distributions you should do this, because you will gradually get the "feeling" of a distribution as the dots pile up.

The distribution may be represented graphically in several ways, one of the best of which is shown in Fig. 1 for the distribution of Table

TABLE II

FREQUENCY TABLE OF COUNT, YARN L

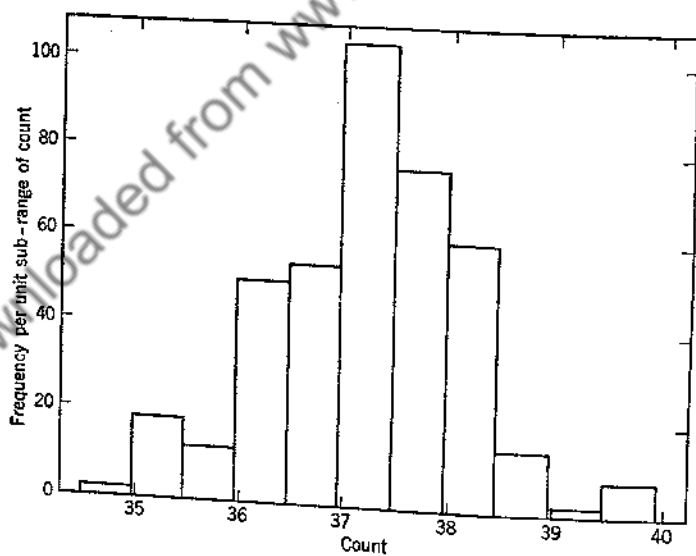| Boundaries of Sub-ranges (Count)	Frequency of Values
34.45–34.95	1
34.95–35.45	9
35.45–35.95	6
35.95–36.45	25
36.45–36.95	27
36.95–37.45	52
37.45–37.95	38
37.95–38.45	30
38.45–38.95	7
38.95–39.45	1
39.45–39.95	4
Total	200

Fig. 1.

II. There each column is proportional in area to the frequency in the sub-range represented by the values of count marked by its boundaries. This is a *histogram*, which is a particular form of a frequency diagram, and from it we see clearly how the values of yarn count are distributed, few being at the extremes, and most being towards the centre of the
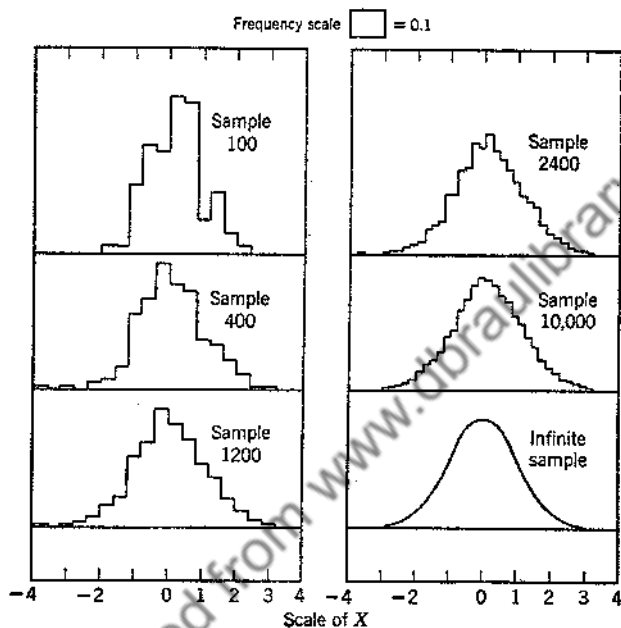

Fig. 2.

total range. This distribution represents fully the results of Table I, and no waving of a statistical wand can tell us anything more about them (provided their order is random in Table I).

You will notice that the outline of the diagram is somewhat irregular as well as step-like, and this, experience suggests, is because there are only 200 results instead of several thousands or millions. Statistical methods are based on the concept of a distribution of an exceedingly large number of observations, which the statistician terms an *infinite population*. Figure 2 shows how that concept arises. The data were artificially constructed to measured values of an unspecified variable, X. Samples of 100, 400, 1200, 2400, and 10,000 readings respectively were taken, and the corresponding histograms are shown in Fig. 2. For

the sample of 100 the steps are wide and irregular; for that of 400 the outline is more regular; and, as the size of the sample increases, it is appropriate to use narrower sub-ranges, and the outline becomes smoother and more regular. We can imagine that, if the sample size could be increased indefinitely, the outline of the histogram would merge into the smooth curve shown in Fig. 2 for an "infinite sample." This is the *frequency curve* of the infinite population and is assumed to be substantially the result that would be obtained if all the articles in a mass-produced bulk or lot were tested, the very large number in a bulk being to all practical intents and purposes equivalent to infinity. It is the population or lot we are interested in describing and studying, and we regard any finite number of individuals, as used in Table I and Fig. 1, as representing it approximately; the representative finite number is termed the *sample*. Sometimes we may use a sample with so few individuals that the idea of representing them by a frequency distribution seems to be ludicrous, or we may represent the data by an average of one sort or another; but even in these circumstances we must always think of the full frequency distribution of the infinite population as lying behind the representation, however inadequate that representation may be.

The form of distribution represented roughly in Fig. 1 and ideally by the smooth curve in Fig. 2 is very commonly encountered. The rise to the hump at the centre of the total range shows whereabouts the values tend to be concentrated; the tailing off towards the extremes shows the extent of the variation and that the farther the values deviate from the centre the fewer do they become; and the symmetry shows that the tendency to variation is the same above and below the centre. The smooth curve of Fig. 2 is termed the *Gaussian* or *Normal* distribution curve, and on it is based the statistical theory we shall use. When the distribution is of this form, there are only two characteristics to notice: the central value about which the results tend to be concentrated and the extent of the "spread" or variation.

If a distribution differs from the Normal form, some of the methods described herein do not apply and special methods have to be used, but this does not often happen.

Summary Statistical Measures

We always think of the quality of masses of things, or populations of individuals, in terms of frequency distributions; but we seldom use the full distribution in routine control. We use statistical measures that summarise the information given by a distribution. The most

commonly used measures describe (a) the central value around which the individual values are grouped and about which they are scattered, (b) the "spread" or degree of variation, and (c) the fraction of individuals lying between certain limits of the variable.

The Mean. The central value is almost always represented by the ordinary average or *arithmetic mean*, which for the bulk or population is denoted by the symbol $\overline{X}'$, the value estimated from a single sample being denoted by $\overline{X}$, or from a combination of several samples, by $\overline{\overline{X}}$.

The Standard Deviation. The degree of variation is described by a number of alternative measures of which the most fundamental from the standpoint of statistical theory is the standard deviation. It may be calculated by subtracting the arithmetic mean from each value in the sample in turn, squaring and adding the resulting differences, dividing the sum of the squared differences by the number of observations, and extracting the square root. It is represented by the Greek letter σ, and for a sample of size n observations of the variable X it may be defined algebraically by the equation

$$\sigma = \sqrt{\frac{\Sigma(X - \overline{X})^2}{n}} \tag{1}$$

where $\overline{X}$ is the arithmetic mean and Σ means "sum the quantity for the n values of X." It can be seen in a general way that for a large degree of variation the differences $(X - \overline{X})$ are large and hence so is σ. As n approaches the number in the bulk or population (i.e., infinity), σ approaches the standard deviation for the bulk or population, which is represented by the symbol σ'.

For the count of yarn L of Table I and Fig. 1, the mean $\overline{X}$ is 37.22 and the standard deviation σ is 0.93. The values of count that are one standard deviation above and below the mean are 38.15 and 36.29, and reference to Fig. 1 will show where these values come in the total spread; there is considerable variation outside these limits. Values at two standard deviations above and below the mean are 39.08 and 35.36, and Fig. 1 shows that between them is contained most of the variation. Values at three standard deviations above and below the mean, 40.01 and 34.43, contain all the variation. It is by this kind of argument that the significance of the standard deviation as a measure of variation comes to be appreciated.

This appreciation can be cast in a more general form if we assume a form for the population frequency distribution, and we shall assume the Normal form shown in Fig. 3. This is drawn on the same pattern

B

as Fig. 1, with the variable X represented along the abscissa and a frequency density, or frequency per unit range of X, represented along the ordinate, so that frequencies are represented by areas. It is convenient to represent proportionate frequencies in this way, so that the total area under the curve is unity and the fraction of the area contained between any two ordinates is the fraction of the total observations having values between those corresponding to the two ordinates. In
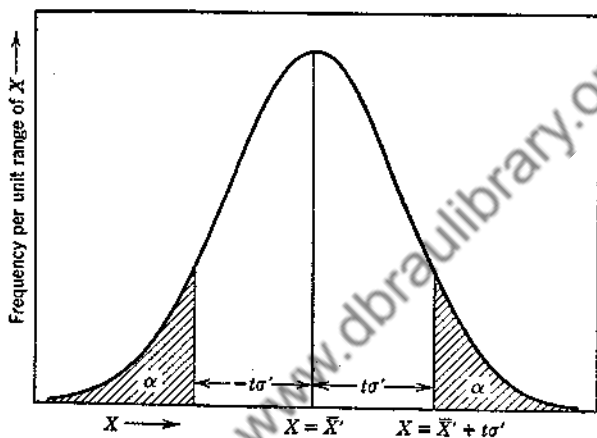
Fig. 3.

Fig. 3, the central ordinate is drawn at $\overline{X}'$, the mean value of X, and there is an ordinate drawn at $\overline{X}' + t\sigma'$ (i.e., at a value of X greater than the mean by t times σ'). The shaded area to the right of this second ordinate is denoted by α and is proportional to the frequency of observations having values greater than $\overline{X}' + t\sigma'$. Since the curve is symmetrical, the area to the left of a corresponding ordinate drawn at $\overline{X}' - t\sigma'$ is also α. There are very full tables showing the relationship between t and α for the Normal distribution, and a few important results are in Table III.

TABLE III

t	1		2		3
α	0.159		0.023		0.0013
α	0.05	0.025	0.01	0.005	0.001
t	1.645	1.96	2.33	2.58	3.09

Theoretically the Normal distribution extends to infinity in both directions of X, so that for no value of t is α absolutely zero. That would be an absurd result from a practical point of view were it not that, for large values of t, α becomes very small. For $t = 3$, $\alpha = 0.0013$, so that 0.0026, or 2.6 in a thousand of the values, are beyond limits set at three times the standard deviation above and below the mean, the remaining 997.4 in a thousand being contained within those limits. Thus for most practical purposes a range of six times the standard deviation may be regarded as just about containing all the values in a large bulk. The lower half of Table III is useful where α has some simple value and the corresponding value of t is required.

The Range. So far the standard deviation as calculated from a large number of observations has been described, but often in practice there is only a small number, or a collection of groups of small numbers. The standard deviation may then be calculated for each sample according to equation (1), and the values of σ averaged to give $\bar{\sigma}$, but this is related in a complicated way to the population value σ'. It is more convenient (and, if the number per sample is less than 20, advisable) to use as a measure of variation the *mean range*, denoted by the symbol $\bar{R}$. For example, the results of Table I are printed in groups of four and the ranges, being the differences between the highest and lowest values in the sub-ranges, are (reading down the columns) $3.85 - 3.51 = 0.34$, $3.83 - 3.64 = 0.19$, and so on; these are the 50 individual ranges R, and their average $\bar{R}$ is 1.89.

The mean range is a convenient measure, and it is much used, but it should be handled carefully. Other things being equal, it depends on the number of observations per sample, which must therefore be specified. If a population of observations is distributed Normally, the standard deviation calculated according to equation (1) (n very large) and the mean range in a very large number of samples *drawn at random* are related in the way shown in Table IV. For the count of yarn L,

<div align="center">TABLE IV</div>

Number in sample	2	3	4	5	10
Mean range ÷ standard deviation $(\bar{R}' \div \sigma')$	1.128	1.693	2.059	2.326	3.078

$\bar{R}/\sigma = 1.89/0.93 = 2.03$, which is reasonably close to the ratio of population values in Table IV for samples of 4. The assumption of Normality

is not very important practically, as the relationships are not very different from those given in Table IV even when the departures from Normality are quite large. But the proviso "drawn at random" is important.

The Fraction Defective. Often the technician is not interested in the whole frequency distribution of values or in any measure of central tendency or variation; he merely wants to control the proportion of values below a certain limit, above a certain limit, or between two limits. The manufacturer of washers made to specified tolerances of thickness is usually concerned only with the proportion having thicknesses outside the tolerance limits; and the "quality" of a batch of electrical insulators may be the proportion that break down at or below a certain specified voltage. This proportion is termed the *fraction defective*. It is represented by the symbol p' when determined for a population and by p for a sample.

The fraction defective according to any specified limit or limits can be deduced from a full frequency distribution of measured values, but in routine control it will usually be determined by some simplified procedure such as go, no-go gauges make possible. Statistically the fraction defective and the fraction not defective are equivalent, and indeed the same methods of expression and analysis apply to the fraction of individuals in a population having any characteristic, good (from the technical point of view) or bad. Some more general term for this fraction would be preferable, but the qualifying term "defective" is now well established in quality control, and we shall use it, but you must remember that statistically it may be interpreted in the widest possible way. The fraction is often expressed as a percentage, and you should remember that we are here concerned with fractions or percentages of independent individuals; quantities such as a percentage of ash in coal are measurable quantities denoted by X.

If the fraction defective is defined by limits of some measurable quantity distributed Normally, it is related to the mean $\overline{X}'$, the standard deviation, σ', and the limits $\overline{X}' \pm t\sigma'$ according to Fig. 3 and Table III, or to one of the fuller tables of which Table III is an extract. This result has practical use in setting design tolerances and investigating their effects, as will be illustrated in the following example based on a paper by Mr. Edmond E. Bates.*

A certain shaft had to be made to a certain diameter with tolerances of 0.000 and —0.001 inch, and 0.45 of those made were defective, being outside those limits. Was this distressingly high fraction defective due

* *The Iron Age*, July 3, 1947, p. 58.

to the general variability of the working of the lathe, or was it due to things occasionally going wrong that could be corrected by better supervision and control by the operator? The supervisor of the shop believed the second alternative, but it is a general experience that lathes do produce diameters with an uncontrollable variation that is characteristic of the machine in its particular state of repair and maintenance, and often this variation is distributed approximately Normally. What would be the fraction defective if this were true of our lathe?

Measurements on 12 samples of 5 shafts gave a mean range of 0.00137 inch, leading to an estimate for the standard deviation of $0.00137/2.326 = 0.00059$ inch (see Table IV). The most favourable machine setting would produce a mean diameter 0.0005 inch less than the specified diameter, so that the tolerance limits would be at 0.0005 inch above and below the mean and t (see Fig. 3) would be $0.0005/0.00059 = 0.85$. From Table III we see that for the Normal distribution the corresponding α is greater than 0.159, and the full tables show it to be very near 0.20. Thus, on the "Normal general variation" theory, the fraction defective would be 0.40, which is reasonably near the observed fraction of 0.45. Since the 5 shafts in each sample were made consecutively, their range measured a general variation that is scarcely controllable, and it was suggested that the machine should be repaired. This was (apparently reluctantly) agreed to by the supervisor, and the standard deviation was reduced to a value that is not stated but appears (from some graphs in the paper) to be about 0.00024 inch. This would give a value of t of $0.0005/0.00024 = 2$ (approx.) with about 5 per cent defectives (see Table III). This was an improvement, but was not good enough, since inevitable changes in the mean diameter due to tool wear add to the defectives. Attempts were made to reduce the variability further, but this could only be done by reducing the speed, which was unacceptable to the management.

Attention was then paid to the possibility of increasing the tolerance range. The shaft had to work in a ring, and the designer had specified the tolerances such that there would be absolutely no misfits. But, if the variation was Normal, the fraction of shafts and rings at the extremes of diameter was small, and the fraction of misfits in the assembly of shafts and rings taken at random, owing to large shafts being matched with small rings and vice-versa, was smaller. For example, if according to the strict tolerances and the variability of manufacture there were, say, 0.05 defective shafts and 0.01 defective rings, the fraction of defective assemblies was only $0.05 \times 0.01 =$

0.0005, and, if such a small fraction was acceptable, the tolerance limits for the shafts, say, could be relaxed so that most of the 0.05 previously rejected could be passed forward as non-defective. According to Mr. Bates, an acceptable scheme was prepared along these lines, with relaxed tolerances (and the consequent reduction in manufacturing costs), to give a calculated fraction of defective assemblies of only 4 in 1000.

For an investigation such as that just described we need to know the fraction defective for various values of the standard deviation and tolerance limits, and in order to determine this directly from the measurements several hundred shafts and rings would have to be measured for each frequency distribution. In fact, the results reported by Mr. Bates were obtained from measurements on only 210 parts altogether, and this was made possible only by estimating standard deviations and assuming the Normal distribution. The estimates must have been subject to substantial errors, since they were based on only 30 or 60 measurements per distribution, and the distributions were only approximately Normal, at best. However, great precision was not required, and it would not have mattered much if the final fraction of misfit assemblies turned out to be 2 or 3, or even 5 or 6, per thousand instead of the theoretically calculated 4.

It is clear that in circumstances like those just exemplified, the standard deviation of the articles produced by a machine over a short time is an important characteristic that indicates for what tolerance limits of manufacture the machine is suitable. It has been found good, in a large shop, to know the standard deviation for each machine and to allocate jobs with various tolerance limits accordingly.

Other Measures. There are other measures of frequency distributions that may be used in unusual circumstances (e.g., when the distribution is asymmetrical in shape). We need not consider them now.

Chapter 2. THEORY OF THE CONTROL CHART

The methods of quality control with which we shall deal have been developed primarily to meet the needs of the mass production of discrete articles, although we shall see that they can be adapted to meet other needs.

The ideal of mass production is to know and maintain at appropriate levels all the factors that determine quality so that all the articles produced have the required qualities. According to this ideal, for example, a spinner of cotton yarn would supply the correct staple and growth of raw cotton, would subject it all to the same processing with carefully defined drafts, roller settings, spindle speeds, and so on, and would expect every bobbin of yarn produced to have the required fineness, strength, cleanliness, and so on.

In practice, however, it is either impossible, or impracticable, or uneconomic to control all the conditions precisely. Raw materials, particularly if they are of biological origin, can not be made uniform. Processes have to be controlled by human operators, and it is impracticable to exercise supervision close enough to eliminate the effects of the human factor entirely. Machines and tools wear, and it is uneconomic to use them only for that brief period between the completion of running in and the appearance of the first signs of slight wear. All these sources of variation are natural to the process, and since we have to put up with them we may describe the resulting variations in quality as *allowable*. Sometimes super-imposed on these are variations that occur when things go wrong or when there is some change in conditions that can be identified and controlled. We may term these *preventable* variations. The aim of the statistical methods of quality control is to separate allowable from preventable variations so that we may know when causes of preventable variations, which Dr. Shewhart has termed *assignable causes*, are in operation. In this and the next two chapters attention will be confined to qualities that are measurable, not to those that are merely expressible as a fraction defective.

Statistical Control

First we must conceptually divide the production of a factory into sections in such a way that the system of causes of variation remains

13

constant for the production of the articles of any one section. Then only allowable variations occur within a section, and assignable causes, if they exist, produce variations only from one section to another. Sections so formed are termed *rational sub-groups*. The most common rational sub-group is the product of one machine or a homogeneous group of machines for a short time of one or two hours or a shift, de-

Fig. 4.
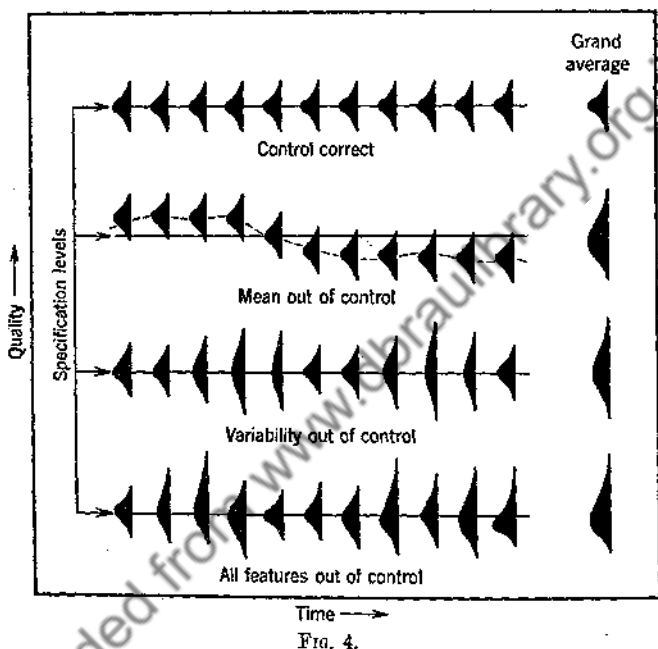
pending on the rate of production, the time being at least short enough to permit of no sensible change in the system of causes affecting quality. If, then, very large numbers of articles are tested or measured, the results may be as represented in Fig. 4.

For the first row in this diagram, a line is drawn to represent the specification level of some quality of the articles, and the allowable variations are represented by the small black frequency distributions shown sideways (as compared with the disposition in Figs. 1 to 3). At each time of testing, the distribution is exactly the same and is centred on the specification level, and a frequency distribution obtained by pooling results from many times of testing is also the same. The production so represented is said to be statistically *in control*.

For the second row in Fig. 4, the specification level is shown as before and the frequency distributions at different times are of the same shape and spread but are centred on different levels, following the dotted line. The frequency distribution of the pooled results from several times of testing will show more spread than the constituent distributions and may also be different in shape. That represents a simple type of lack of control, and the changes in level are the effects of the assignable causes.

The third row of Fig. 4 shows a second type of lack of control, where the distributions are all centred on the specification level and have the same shape but vary in spread, the composite distribution for pooled results having the same shape and centre but a spread intermediate between those of the constituent distributions.

In the fourth row of Fig. 4, everything varies from time to time—the level, the spread, and the shape of the distribution, and the composite distribution is related to its constituents in a way that can not be simply described; the specification level serves only to show the aim and represents no feature of the performance.

The Control Chart. Control Limits

In order to investigate the state of control for a process we use, instead of a chart of frequency distributions as shown in Fig. 4, a chart in which one or more of the statistical measures is plotted for the successive rational sub-groups. If the result for each sub-group is based on a very large number of tests and the process is in the state shown in the top row of Fig. 4 (i.e., in control), each measure is constant and the points fall substantially on a horizontal straight line. If the process is out of control, the points for either the mean or one of the measures of variability, or both, move up and down on the chart. In practice, however, the number of observations per sub-group is not large, since for reasons of economy we have to use relatively small samples. It is a matter of experience that the means, standard deviations, and ranges of successive samples from the same bulk vary, and so, even if a process is in control, the points on a chart will move up and down on account of sampling variations. We need some way of determining and representing the extent of sampling variations so that they can be distinguished from the results of variations in the process due to lack of control. Let us make a chart for a process that is in control and see what happens.

The results for Table I have been well mixed and are in random order, so that the successive groups of 4 may be regarded as small

samples from sub-groups of a process that is in control. There are 50 samples, and the individual results are plotted in the upper part of Fig. 5 as small dots, the four dots for each sample being distributed in a line along the abscissa. The 200 points are piled into the frequency distribution shown on its side at the end of the chart, and outlined in
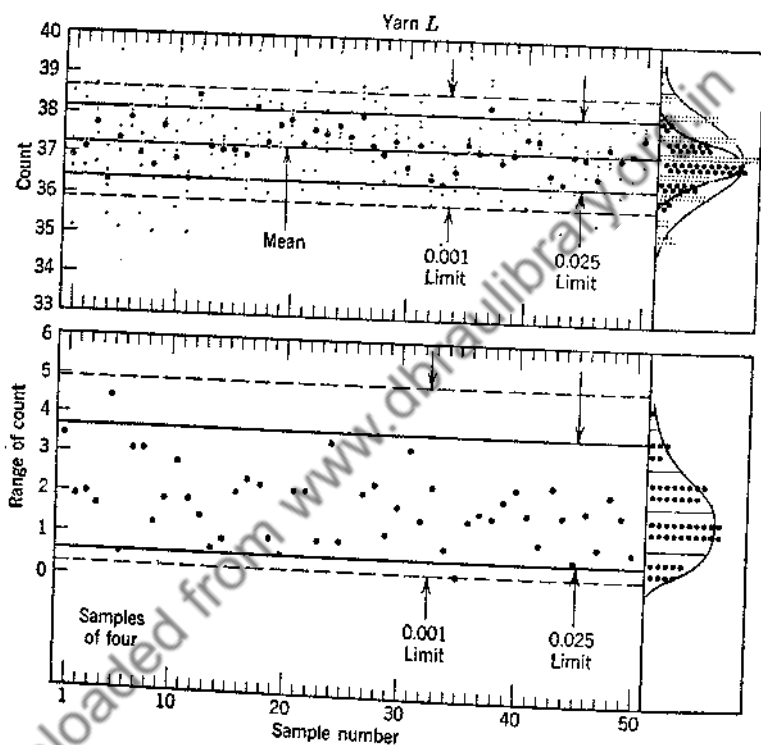
FIG. 5.

its imagined population form by the thin line. This distribution is that of Fig. 1 in a slightly different form, the smooth curve being the Normal distribution with the same mean and standard deviation as the actual distribution. The heavy dots represent the means of the samples of 4, and they vary, but not so much as do the individual values. They have, in Fig. 5, been formed into a frequency distribution of means, known as a *sampling distribution of the mean*, the thick, smooth curve outlining the theoretical sampling distribution for a population of means of 4. It is a Normal distribution centred on the same mean as

the individual values, with a standard deviation equal to that of the individual values divided by the square root of the number per sample. The standard deviation of a sampling distribution of the mean is termed the *standard error of the mean*, but, in spite of its special name, it may be interpreted in terms of proportionate frequencies like any other standard deviation with the aid of Table III. Generally the standard error of the mean of random samples of n observations is related to σ', the standard deviation of the population of individual values, by the formula

$$\text{Standard error of mean} = \frac{\sigma'}{\sqrt{n}} \qquad (2)$$

When we do not know the population value σ', we have to substitute as an approximation the best value available, which for the count of yarn L is $\sigma = 0.93$, giving as an estimate of the standard error in samples of 4 a yarn count of 0.465. The standard deviation of the 50 means should not be very different from 0.465 (you may care to perform the calculation). Likewise, as the centre of the sampling distribution we use the estimate $\overline{\overline{X}} = 37.22$, the double bar indicating that we have taken the mean of the 50 sample means.

It is not convenient in routine quality control, however, to calculate the standard deviation of a lot of means and see whether the result is the same as that given by equation (2). A better way is to choose limits of variation in the sample means that, under conditions of statistical control, should be exceeded by a certain small proportion of the samples, and test if the due proportion of actual points shows variation outside those limits. One set of limits commonly used is chosen so that under conditions of statistical control 0.025 of the points lie below the lower and 0.025 above the upper limit, and from Table III (p. 8) we see that these are at 1.96 times the standard error above and below the mean. For our data of Fig. 5, the grand mean is 37.22, the standard error is 0.465, and the limits are at $37.22 \pm 0.91 = 38.13$ and 36.31; horizontal lines are drawn in Fig. 5 to represent these values. Another pair of commonly used limits are the 0.001 limits at 3.09 times the standard error from the grand mean (see Table III); these for our example are at $37.22 \pm 1.44 = 38.66$ and 35.78 and are represented in Fig. 5 by dotted lines.

You will note that these two sets of limits are almost at two and three times the standard error from the mean, and some people prefer to use those simple multiples, sometimes calling them "two- or three-sigma" limits, in this context meaning "standard error." The cor-

responding proportions are 0.023 instead of 0.025 and 0.0013 instead of 0.001. Clearly this difference in practice is of no theoretical importance.

Now we would expect in the long run 0.05, or 2½ in 50 of the sample mean points in Fig. 5 to lie outside the inner limits; in fact 3 lie above the upper limit, 1 below the lower, and 2 almost exactly on the lower. The agreement is not bad, as these things go, the discrepancy being due to the fact that we have a relatively small number of samples. The theoretical proportions are expected to be satisfied closely only when the number of samples is large. No sample mean points should fall outside the outer limits, and none does.

A chart like that in the upper part of Fig. 5 is termed the *control chart* of the mean. It is a form of presentation of data first proposed by Dr. Shewhart, and although (perhaps because) it is simple it has proved a most valuable statistical tool. Essentially it is a graph with the value of a variable plotted along the ordinate and some designation of the rational sub-group along the abscissa. It may have a central line corresponding to the heavy line in Fig. 5, which may represent a grand mean or a specification level; we shall discuss this later. But a control chart must also have so-called *control limits* that under conditions of control contain a known proportion of the points and so provide a criterion of what variations in the sample means are allowable and compatible with control, and what variations indicate a lack of control. We shall discuss later the principles by which particular control limits are chosen.

A control chart constructed after the manner of Fig. 5 is useful for determining, after the event, whether a given set of data show lack of control. To be of use during production, the chart is extended along the y axis (which in such cases usually represents time in some form), and points are added one by one, samples being tested as production proceeds. Then each point that falls within appropriate control limits is taken as evidence of a continuance of control; if a point falls outside it is taken as an early indication of something going wrong, and investigatory or corrective action is taken. Even if control is maintained, some points should fall outside the limits, as we have seen in Fig. 5, and this procedure will sometimes lead to unnecessary action. Sometimes there may be a real change in the production, and the next point may remain within the limits, so that necessary action may be missed. The control procedure is thus not certain in its operation; but we shall discuss all this later.

When adding the points to a control chart one at a time as production proceeds, and taking or withholding action on the evidence of each

point, it is appropriate to think not of the proportion outside the control limits when the process is in control, but of the probability that an individual will fall outside. This is a change in language corresponding to a change in conception, but it involves no change in statistical procedure.

A companion to the control chart for the mean is that for the range, as shown in the lower part of Fig. 5. There the 50 individual ranges for the group of 4 in Table I have been plotted in an extended form against the group or sample number and consolidated into the form of a frequency distribution, the outline of which is represented by a smooth curve—the sampling distribution of the range. As this is not a Normal distribution, the mean range and standard deviation of ranges do not describe it sufficiently and the proportionate frequencies or limits corresponding to various probability levels cannot be determined from Table III. However, Dr. Dudding and Mr. Jennett's book, *Quality Control Charts* (B.S. 600R), gives, for various group or sample sizes, limits corresponding to probabilities of 0.025 and 0.001. For samples of 4 the limits corresponding to a probability of 0.025 are 0.29 and 1.93 times the population mean range, and, for the data of Table I which have a computed mean range of 1.89, the best estimate we have of these limits is 0.55 and 3.64; they are marked by the continuous lines on the range chart of Fig. 5. The corresponding limits for a probability level of 0.001 are 0.10 and 2.57 times the mean range, and for yarn *L* they are estimated to be at 0.19 and 4.85, as shown by the dotted lines on the range chart. Two points of the 50 lie outside the inner limits and 1 point on the lower 0.025 limit, and 1 point in 50 lies on the lower 0.001 limit; the agreement with theory for a range statistically in control is good.

Some writers use three-sigma limits for the range as for the mean, in spite of the fact that the sampling distribution of the range is not Normal. This procedure may be justified on the empirical ground that it works, or on the more theoretical ground that it gives a close enough approximation to the theoretical limits. It is necessary to know "sigma," the standard error of the range, and this has been tabulated for different sizes of sample. For convenient use in quality control, tables also give factors by which to multiply the mean range or the standard deviation in order to obtain the three-sigma control limits. For small samples the lower limit would be at a negative value of the range, which is absurd; in such instances the lower limit is conventionally put at zero. For samples of 4, the upper three-sigma limit is

2.28 times the mean range, which is between the 0.025 and 0.001 limits (see, for example, Professor Grant's *Statistical Quality Control*).

Control charts may be formed of the standard deviation, which is preferred to the range when the number of observations per sub-group is greater than 20 on the ground that the standard deviation is more likely to give early indication of any lack of control in the variability, and the superiority of the standard deviation in this respect increases as the number of observations increases. You will find the tables for making these charts in the text-books.

Control charts can be made for other statistical measures of populations, but the only one that is in common use is the control chart of the fraction defective; this will be discussed in Chapter 5.

Chapter 3. PRACTICAL APPLICATION OF THE CONTROL CHART PROCEDURE

The control chart is a statistical tool, but quality control is a part of technical management and, as I regard it, is a function related to routine production rather than to research and development. Statisticians like myself are apt to call the subject *statistical quality control*, hoping thereby to absolve themselves from the duty of considering management aspects. But, if the methods are to be useful, technical managers have got to believe that they can be of use, to be willing to use them, and to know how to use them.

In almost any factory the number of control charts that could be introduced is legion. Which of these are worth introducing? Under what circumstances are the various methods of quality control suitable? How are they to be used by management in the factory routine? What results come from their use? These kinds of questions have to be answered by someone.

The following discussion will give little specific guidance—it contains vague generalities rather than "brass tacks." A complete science of quality control has yet to be developed. As a consequence the technical manager must learn to use the methods by trial: he must be something of a pioneer. If he is to achieve results, he must experiment patiently and persistently and must be prepared for the risk of all experimental approaches—the risk of failure. Nevertheless the methods have had successful application in a wide variety of circumstances, and enterprise in this direction is likely to be well rewarded in one way or another.

Four types of use of the control chart procedure will now be discussed.

Tracing Causes of Variation

The first type of uses of the control chart is as an instrument of investigation when a process is out of control. If the process is new, all the conditions that affect quality may not be known or may not have been brought under control; or, if a process is established, things may go wrong. In either event it is helpful to know when and where preventable variations are occurring, so that assignable causes can be traced and eliminated.

21

The procedure is to collect data for a number of sub-groups, say 50 or more, and form them into a control chart. The chart is then continued with the control limits so calculated, points being added as production proceeds. If a point falls outside the chosen control limits, the assignable cause is sought and, if possible, eliminated. If during the investigation improvements are effected, a new chart with new limits is calculated from a later set of data, and the process is repeated until the chart shows that statistical control is established. This first use for the control chart then ceases; it is for the management to maintain the conditions necessary for control (machine settings, temperature, and so on). The chart shows when to look for assignable causes if the division into rational sub-groups is according to time, and where if it is according to machines, operatives, batches of raw material, or other manufacturing units. The mean chart is almost always used in these circumstances, but it is usually desirable to use also the range chart.

Sometimes the search, which tends to be of a "trouble-shooting" type, leads to the discovery of assignable causes that can be eliminated by the production people—causes that the production people often claim to have known about all the time! Such a result shows quality control at its most obviously successful. Sometimes control can not be established so easily, and the whole problem has to be referred to the research department or its equivalent; then the statistical methods dealt with in Part II come into their own. Sometimes the assignable causes can not be eliminated and have to be accepted as part of the process. In cotton fabric manufacture, for example, "piece" lengths from different looms and from different warps in the same loom are probably out of control; yet it is doubtful if all the causes are known, and it is certainly impracticable to control all of them. Even then, however, the knowledge given by control charts may be useful as suggesting lines of research and giving a basis for a different division into rational sub-groups which include these unpreventable variations, if control charts are required for one of the other uses mentioned below.

Where the assignable causes can be eliminated, not only is the process brought into control, but also the level of quality often improves and the variability within the rational sub-group is reduced. The improvement in quality is to be expected, since the control will always be operated to bring the level of quality nearer that desired. But any reduction in within-group variability must be indirect and can occur only when assignable causes of between-group variations happen also to be causes of within-group variations.

A control chart does not purport to tell what assignable causes are operating, much less does it tell the remedy; nevertheless it may often give useful indications. The pattern of points may give a clue to the cause. The simplest example is in machining metal parts to required dimensions, where a trend in the mean alone indicates a change in setting and a change in the variability sometimes indicates tool wear or some mechanical fault with the machine. I have heard of a girl, who is not an engineer, in charge of the control charts in a workshop; she has developed what seems to be an almost uncanny flair for telling the technical people what is wrong from an examination of the chart.

It is always a matter of technical and managerial judgment to decide how many control charts to keep—whether one for each machine or operator—and how finely to divide the sub-groups, but for investigational purposes a fairly high degree of detail is usually called for. The increase in expense that this entails is mitigated by the fact that the investigation is, or should be, of limited duration.

Although there have been very many successful applications of this first use of quality control, there have also been failures, and it would be useful to know, generally, under what conditions success is to be looked for. Obviously results will be expected only where lack of control is known to exist or is suspected. Success is most likely to be achieved where the process is fairly simple and assignable causes are not too hard to find or too remote in space or time to be controlled. A finished fabric, for example, is several months in process from the raw cotton and goes through a very complex series of processes; clearly there is not much hope of finding in the earlier processes assignable causes of variation that are detected only in the finished fabric. This is not to say, however, that control charts have no use in such circumstances.

This use of control charts is perhaps the most exciting of all the uses and is the one usually illustrated in the case histories given in the literature. However, it is not the most important use.

Routine Control

A second use of control charts is for the routine control of quality when the conditions can not be determined and controlled *a priori*, but the products have to be tested and the process adjusted accordingly. In cotton spinning, for example, all conditions may be kept uniform as far as practicable, and yet the yarn changes from time to time in count or fineness. Accordingly yarn is tested periodically. If the tested count is within certain limits, no adjustment is made to the spinning

C

machine; otherwise a pinion is changed to correct for the change in count. This procedure is traditional in the cotton textile industry and is very close to the control chart procedure. Machine tools are set in somewhat the same way, the final adjustments being made after parts have been measured. Sometimes the raw materials of a process may be tested, and the process may be adjusted for each batch to give the required quality in the product.

In established processes of these kinds methods of control must have been developed long before statistical quality control was thought of, but for efficient control it is necessary to have not only a test result but also a measure of its precision and a criterion for deciding what variations to ignore and for what variations to make adjustments. Such a criterion is provided by the control limits; but whether for routine purposes a chart is better than records kept in a book, or even than unrecorded test results, is a matter for special consideration by each management concerned. Action is stereotyped, and a simple testing and recording procedure should usually suffice. It will depend on special technical circumstances, too, whether there shall be separate control for each machine or group control, and whether the tests shall be made frequently or infrequently and at regular intervals or irregularly as certain conditions, such as the batch of raw materials, change.

A special case of the second use of control charts occurs when a process is non- or semi-automatic and the quality depends on the manipulation of an operator (e.g., on the speed or pressure with which a handle is operated). A control chart in an accessible position can help the operator to maintain a uniform performance.

Quality Assurance

The third type of use of control charts is found where perfect control of all the conditions for quality is attainable and has been attained and evidence is required of the fact. Maintenance of control requires continuous conscientious attention to machine settings, temperatures, cleanliness, and so on, by operatives, machine fixers, and supervisors; and if they perform their work nothing should go wrong. But they are human; mistakes can be made, and they can be discovered by control charts; the visible evidence of control given by a chart can stimulate interest in the matter and be a source of legitimate pride; and the knowledge that quality is being recorded can stimulate conscientiousness. Control charts may be useful, too, to the higher ranks of management as their only means of "keeping tabs" on the quality produced by the factory. Generally the amount of detail required will decrease as we go from the

factory floor to the president's office. In the factory it may be worth keeping a set of charts for the processes under each supervisor; the factory manager or superintendent may require a set for each department and product; and the president may be content with a composite chart summarising the quality of all products of the factory.

Another agent for whom control charts may be useful merely as evidence of control is the "consumer," who may be an ultimate consumer, another manufacturer who uses the product of one factory as the raw material of his own, or a department in the same factory "consuming" the products of an earlier department in the production line. Whenever material passes from one responsibility to another, or passes a stage beyond which faults can not be traced backwards, it is wise to have an assurance of the quality. Commonly this is achieved by inspecting each batch separately. This is often expensive if the control is to be really effective; otherwise it is uncertain in its effects, leading to the rejection of satisfactory batches and the acceptance of unsatisfactory ones. If, however, the process is in control and the characteristics of the product are known (usually the mean and standard deviation or mean range of the quality), the consumer has as much information as he can possible have; and control charts provide that information in the most economical way. There is a marked tendency for large consumers to accept the evidence of quality control given by control charts in preference to batch-by-batch inspection.

Usually it is best for the producer to keep these charts, for he is in the best position to choose the rational sub-groups. But, if this can not be done, the consumer may find it useful to keep a set of charts of things from each source of supply, using some batching units as sub-groups; he will thus have an assurance of control where it exists and a good basis for making complaints when they are necessary.

The assurance of control is required when a measure of variability is used to relate tolerance limits to the fraction defective, as described in Chapter 1 (p. 10), for the calculations are valid only if the mean and variability are both in control. We shall describe in Chapter 5 how to deal with the situation when the mean is not (and need not be) in control, but control of the variability is always essential.

We may extend the definition of "quality" to include indexes or measures of operating performance, such as fuel or power consumption, the proportion of time the machines are operating, and the numbers of machine breakdowns; or even statistics of absenteeism, sales, profits, and so on. It is always possible to plot such data on time charts and often to form rational sub-groups so that control limits can be fixed.

Such an extension of definition thus considerably widens the use of statistical "quality" control in management.

The same set of control charts may sometimes be put to the three types of use mentioned so far. Charts used for the routine control of a process (the second use) can also provide evidence of control (the third use), and charts used as evidence also show when or where assignable causes are operating (the first use), if they operate occasionally.

The control chart applies, *par excellence*, to continuous mass production. Does it apply where there are short runs? The first use can be made where the quantity of production gives as few as 20 sub-groups. The second and third uses can be profitable with such short runs if the quality of the products of successive runs can be reduced to a common measure. A machined dimension measured as a deviation from the specified dimension can often be used in this way, and the fraction defective is a common measure that is widely used. Control charts for this quality will be described in Chapter 5.

Reconciliation of Design and Manufacture

The fourth type of use does not really belong particularly to the control chart procedure, but, since it is usually regarded as an application of quality control, this is a convenient place to mention it. It is exemplified in Chapter 1, where a knowledge of the variability of articles turned by a lathe was used in specifying in the design tolerance limits that give an adequate performance and yet could be held in manufacture without producing an undue fraction of defective articles. Generally, when products vary statistically, it is as necessary, for satisfactory design, to know the statistical characteristics of the variation as to know, say, the strength of the materials to be used; and, conversely, just as the required strength of the product will somewhat determine the materials specified, so will the variability that is tolerable to the designer somewhat determine on which machine articles will be made or which particular process will be used, if there is a choice. Moreover the measures of variability and so on have no meaning and are not useful in this connection unless the qualities are stable—in control. The control chart not only establishes this but also provides the various measures required.

Administrative Details

There are a number of important administrative and related details that can only be mentioned here. They are: the size and qualifications

of the staff required for quality control; the precise definition of its duties and responsibilities; its relation to the management and operating staff; the decision regarding which department should be responsible for making tests; the placing of the charts; the "selling" of the idea of quality control; the ensuring of co-operative action.

Chapter 4. STATISTICAL AND TECHNICAL DETAILS IN APPLYING THE CONTROL CHART PROCEDURE

The details in the design of control charts must now be considered.

Rational Sub-groups and Statistical Individuals

According to our definition, the conceptual division of the products of a factory into rational sub-groups must be so made that within each sub-group only allowable variations occur, any preventable variation due to assignable causes being seen between sub-groups. The application of this principle to particular cases requires technical knowledge of likely kinds of causes, consideration of the purpose for which the control chart is being made, good judgment, and a clear head. It is helpful to have some preliminary idea of whether any assignable causes are likely to affect all parts of the process equally; whether they are likely to produce sudden changes or trends; and if the effects are sudden whether they occur frequently, or if they are trends whether they are slow.

Generally the larger the sub-group the more numerous are the classes of variation included within it, and the requirement for excluding preventable variations sets an upper limit to the size. The lower limit is set by considerations of economy. Each sub-group has to be sampled or tested separately, and a statistical measure or set of measures has to be computed to form one or more points on a chart; clearly the fewer the sub-groups the lower is the cost of testing.

All this advice is very general and somewhat vague, but the following particular example may be helpful.

The mule is a machine used for spinning cotton yarn very much more in Lancashire (England) than anywhere else in the world. It has about 1200 spindles all simultaneously spinning cops of yarn, each cop containing some 700 to 2500 yards according to its size and the fineness of the yarn. For testing purposes the yarn is divided into lengths of 120 yards, called leas, so that each cop contains from 6 to 20 leas. As a sample of the production, cops were taken from 5 spindles of mule 38, and 2 leas were taken from each cop, according to the scheme of Table V; this was repeated for mule 42, and for twenty-two others.

28

TABLE V

Weights of Leas of Mule-Spun Cotton Yarn

Mule No.	Spindle (Cop) No.	Lea No.	Weight	Range (Leas)	Mean Weight (Spindles)	Range (Spindle Means)	Mean Weight (Mules)
38	1	1	386	7	382.5	17.5	381.0
		2	379				
	2	1	369	1	369.5		
		2	370				
	3	1	392	12	386		
		2	380				
	4	1	389	18	380		
		2	371				
	5	1	387	0	387		
		2	387				
42	1	1	377	7	380.5	27	378.8
		2	384				
	2	1	382	15	389.5		
		2	397				
	3	1	370	15	362.5		
		2	355				
	4	1	384	3	385.5		
		2	387				
	5	1	378	4	376		
		2	374				

And so on for 24 mules

The quantity tested was the lea weight (the units are not specified because they do not matter here). There are possible variations (a) between leas from the same cop, (b) between cops from the same mule, and (c) between mules. The variations between mules are preventable in routine production, and in an investigation made at one time each mule could form a rational sub-group. The spindle variations can not

be controlled in routine production since there are far too many spindles to be treated individually, but we might study them as a research project, perhaps using a control chart for selecting a few spindles giving high, medium, and low values, discovering the causes of those variations, and then eliminating or reducing them by improved mule manufacture or improved control at the earlier process that makes the "roving" from which the cotton yarn is spun. Then spindle variations would be formally preventable, and the cops would form separate sub-groups. In such a research it would be well to use only one mule in order to avoid the confusion that would arise if mule variations were superimposed.

Sometimes care is needed in the sub-division of the sub-group into the ultimate statistical individuals or elements, whose variation is measured by the standard deviation or mean range and whose number n occurs in the denominator of equation (2). When the things under investigation are mass-produced articles, the articles seem obviously to be the statistical individuals, and often they are. Frequently, however, the production of each rational sub-group may be divided into strata or sub-sub-groups, which could, but for one reason or another do not, form rational sub-groups for quality control. In Table V, for example, if we are seeking to control mule variations, the spindles or cops form a basis for sub-sub-groups of leas. There are doubtless other multi-spindled machines that produce in batches of one article per spindle, and the batches form sub-sub-groups if the sub-groups are the articles made by the machine in successive intervals of time; or the same effect may be produced if a machine is frequently recharged with raw material and each charge produces a batch.

In such instances, the statistical individuals should be the largest natural sub-sub-groups that can be formed. In Table V, for example, for the purpose of controlling mule variations the individuals should be the spindle means shown in the sixth column; the mean range should be calculated from the ranges 17.5, 27, etc., in the seventh column, and the n for equation (2) should be 5. The control chart for the 24 mules is given in Fig. 6 with "correct" limits on the 0.001 probability levels shown in full lines; only the mean for mule 36 is out of control. Incorrect limits could be calculated from the individual lea results in two ways. The range for each mule could be obtained from the ten values in the fourth column of Table V, a mean range could be found and converted to the standard deviation, and this could be used in equation (2) with $n = 10$. This procedure would be very wrong, for the mean range

would measure a heterogeneous mixture of cop and lea variations. Another incorrect procedure would be to use the ranges of the fifth column of Table V to measure the within-cop variation, and put $n = 10$ in equation (2). This has been done to give the limits in dotted lines in Fig. 6; 6 of the 24 points are outside the incorrect 0.001 limits and 2 are on the lower one. This merely shows that the mule means vary more than by an amount that can be explained by the within-cop variation. This is not a very useful result to the technician who knows
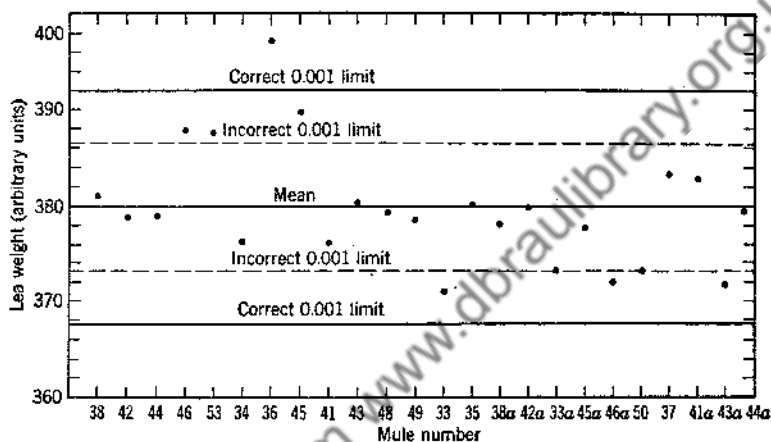

Fig. 6.

about spindle variations; these limits are perhaps inappropriate rather than incorrect.

If the production can be divided into sub-sub-groups which happen to show no real variation, the division is purely nominal and has no statistical meaning; then it is permissible to use the ultimate articles as individuals. But in general we do not know this to be the case, and since no harm is ever done by treating the sub-sub-groups as statistical individuals, the rule given here is a good one.

The statistical theory on which the control chart is based assumes that the statistical individuals in each sub-group are random in the sense that no significance can be attached to the order or to any groupings in which they occur. In the above example, the effect of the association of the results coming from the same spindle is destroyed by regarding the spindles as randomly distributed individuals. Commonly, however, the articles tested are a series in time, and then

seldom if ever has the order of occurrence absolutely no significance. Almost inevitably there are patterns of variation—trends or quasi-cyclical movements—on which are super-imposed random fluctuations, and the usual interpretation of the control chart is valid only if either the variability associated with the pattern is negligibly small compared with the random variation, or if the sequence can be broken into sub-sub-groups so as to turn the pattern into a form of variation that is indistinguishable from random. The limited possibilities of doing this are discussed in the next paragraph.

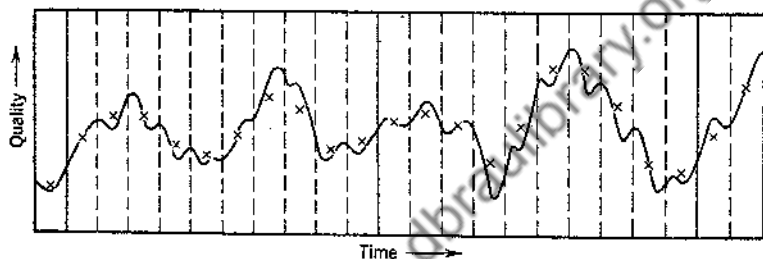Figure 7 is a plot of the quality of part of an imaginary sequence of

Fig. 7.

articles in one sub-group; the sequence is represented by a continuous curve which we may take as representing the underlying pattern of variation, the actual values for the individual articles being imagined as a swarm of points scattered about the curve. The problem is how to form sub-sub-groups which will constitute virtually random individuals. If the sequence is broken up according to the dotted vertical lines, the average quality, say, for each section constituting a statistical individual, the section averages, represented approximately in Fig. 7 by crosses, follow a recognisable pattern and so are not random. If, however, the division is in units of time that are long enough, say, according to the continuous vertical lines, the section averages (there are only two complete sections shown in Fig. 7) will not show a recognisable pattern. A satisfactory sub-division from the statistical viewpoint can be determined only after a full investigation, such as is seldom undertaken; but the engineer, knowing something of the physical causes underlying the process, can often say over what length of time the effects of single causes are likely to persist, and how extensive the sub-divisions must be in order to destroy the physical continuity. The same general considerations apply when sequences are distributed in space in one, two, or three dimensions.

In all this discussion we have ceased to regard the individual articles as statistical individuals; the articles have lost their statistical meaning. It follows, then, that the treatment discussed in the previous paragraphs can be applied to continuous products such as sand, chemicals, glass, and wire. If the continuous flow can be broken up into sub-sub-groups that exhibit only random fluctuations, these can be treated as statistical individuals.

For the simple application of the statistical theory, not only must the sequence of results within each sub-group follow no recognisable pattern, but also it must follow no form of variation that is repeated from one sub-group to another. For example, we may have a machine with eight heads or spindles numbered 1 to 8, each spindle producing at a slightly different level of quality, such that the eight levels in order of head number show no simple pattern of variation. Nevertheless, if the production of all heads of the machine at one time forms a rational sub-group, the head averages forming the individuals, the form of variation between heads will be repeated from one sub-group to another and so will not be random. We may class such a form also as a pattern—a complex one. An engineer would not often tolerate the continued existence of such head variations, but the example will help you to visualise the point.

What are the effects of patterns of variation within the sub-groups on the control chart? The answer to that question is doubtless very complicated; it will depend on the kind of pattern, and the effects have never been fully investigated.

Sometimes the pattern can obscure a real change in level. In order to illustrate this let us suppose an extreme case in which the eight-headed machine mentioned above produces articles with practically no random variation, and no variation between sub-groups, the only important variation being that between heads; and let there be one result per head in each sub-group. Then there will be a range which will be the same for each sub-group, and control limits for the mean chart will be set some distance on either side of the line for the grand mean; but all the sub-group means will lie on the mean line; they will not be scattered within the control limits. Indeed, the mean could shift somewhat and yet the points never go outside the limits. Thus in this instance the effect of treating the pattern of variation as though it were random is to over-estimate the effect of the allowable variation on the sub-group means, and this effect will be present in some degree if a random variation is super-imposed on the pattern. In actual cases the sub-group means can be too closely clustered about the grand mean

line, having regard to the position of the control limits, or there can be an actual variation between sub-group means just about enough to produce the appearance of statistical control. Similarly the effect of the pattern on the range chart is to cluster the points too closely around the line for the mean range and to obscure the effects of real changes in variability.

A pattern of variation within the sub-groups is not likely to produce the appearance of a variation between sub-groups that does not exist. The variation between sub-groups may sometimes be a continuation of the pattern within, each sub-group containing a portion of a larger pattern; but the difference between that and the ordinary interpretation of a control chart is likely to be theoretically interesting rather than practically important.

The discussion of this section has been long and rather complicated because the problem is complicated; but you should not make too much of the questions raised, particularly in the early stages of quality control practice. More often than not the random element in the variation predominates, and in any event the control chart is only an aid; you are not likely to go seriously astray if you let technical knowledge be the ultimate guide to action. Sometimes, however, you may obtain puzzling results, and a reference to this discussion may help you to unravel the puzzle and keep your faith in statistics. Or you may wish later to take the complications into account.

Sampling Method

In quality control, as in most other applied statistics, we work largely by samples, and this practice raises a number of problems.

Bias is one potential source of difficulty which does not arise for many industrial products. The variations in the dimensions of machined articles are usually too small to cause any appreciable tendency to select either the large or the small ones; no differences between electric lamps are likely to cause the sampler to prefer those with, say, long life; and so on for a wide variety of products.

However, bias in selection can exist. If textile fibres are selected individually, the longer ones have a greater chance of being included than the shorter ones unless the sampler consciously tries to correct for this bias, and then anything may happen. In a random selection of lumps of coal of various sizes, bias in size of lump may affect the apparent ash content. Particles of sand and other substances in bulk

tend to stratify according to size and weight, giving the risk of a bias due to position.

Bias can often be avoided, if necessary, by adopting some sampling method specially adapted to the technical and physical conditions. For example, in sampling cotton fibres, if they are taken in tufts of a few hundred, there is virtually no length bias. Then each tuft can be halved, one half selected by a toss of a coin can be discarded and the other further halved, and the process can be repeated until the residual tuft contains only a few fibres. Then, if several tufts are combined, there results a single representative sample, free from bias and convenient in size for testing. This illustrates the general principle of taking aggregates large enough to eliminate bias (suitable only where size does in fact eliminate bias) and fractionating them to produce a sample of a suitable reduced size.

Another general way of avoiding bias is to ensure that, when the subject varies in zones or strata, every zone is represented. In sampling particles in a stratified bulk, the sample may contain particles from all layers, and sampling tools are sometimes designed to secure this end.

Bias is not often important in quality control, even where it exists, for the technician is interested in comparisons between sub-groups.

Sometimes only a sample of sub-groups is included in the quality control scheme, as when a few articles are taken together every hour or every shift instead of more or less continuously. This, however, raises no questions, for the sub-groups actually tested are identified and dealt with, if out of control, they are not necessarily regarded as typical. There is only the remote danger that they may be taken periodically at intervals that coincide exactly with some periodicity in the quality; then the periodic variations would be missed.

Each sub-group is regarded as a random sample from an infinite population. When the total number of individuals in the rational sub-group is small (for example the number of values per spindle in Table V can not be more than about 20), it seems far-fetched to assume an infinite population. A similar situation arises when we throw a die, and we have no difficulty in regarding a limited number of throws as a random sample of the many throws that could be, but have not been, made—a sample of a hypothetical infinite population of throws. Likewise we can easily imagine an infinite population of within-group variations, the hypothetical result of the system of causes of allowable variations; the few actual values can then be regarded as a random

sample from this imaginary population. The infinite population becomes merely a statistical model of the complex of within-group causes.

If the sub-group is divided into sub-sub-groups in the way outlined in the previous section, to form statistical individuals that are independent in the sense that they conform to no regular or repeatable pattern in time or space, a sample of these sub-sub-groups (or *clusters* as they have been termed when sampling in other fields) taken in any way is a random sample for the purposes of control. For example, of the 1200 spindles on each mule (Table V) only 5 are tested. If the spindles vary at random, it does not matter whether the 5 are the first five, the middle five, five chosen at intervals of 300 spindles, or five chosen at random. If the spindles vary according to some pattern, the simple control chart procedure does not apply anyway.

The sub-sub-group is often represented by a sample, just as in Table V each spindle is represented by only 2 leas out of the 20 or so on the cop. Such a sample is not necessarily chosen at random, and it is better that it should not be so. If the variation within the sub-sub-group is entirely random, it does not matter how the sample is taken, and, if the variation follows a pattern, the use of a non-random sample improves the precision of control. Consider, for example, the cotton yarn of Table V and suppose that a cop from one spindle is a sub-sub-group forming a statistical individual. If there is, say, a trend in weight as successive leas are taken from a cop and leas are tested from specified places, say the first few leas or the first and then every fifth lea, the trend will not contribute to the apparent cop-to-cop variation, and this will be a good thing. The within-cop variation is not under investigation (if it were the cops would be the rational sub-groups), and to eliminate part of it (the systematic part) is to narrow the control limits for a given probability level and so to improve the precision of control. There are no general theoretical rules for deciding how to arrange the systematic sample, and practical consideration will usually determine the matter; for example, it is not possible to get at lea 5, say, until leas 1 to 4 have been wound off.

A random sample of leas from the cop would be perfectly sound, theoretically, since the within-cop systematic variation would thus be turned into a random variation and would contribute to the between-cop variation in the kind of way assumed by the statistical theory. The random sample would be sound but, for the purposes of control, inefficient.

Another kind of sample would be exemplified by taking, say, the first 2 leas from cop 1, leas 3 and 4 from cop 2, leas 5 and 6 from cop 3,

and so on, returning to leas 1 and 2 for cop 11; you can easily think up other patterned samples. This is theoretically unsound. The within-cop variation contributes to the variation between the cop means, but not randomly, and the results on the control chart can be affected by the within-cop pattern of variation.

Now, as a relief after this long and intricate discussion, let us consider a kind of situation that is fairly common and that is sometimes discussed. Suppose that articles are coming from the production line in a stream, that 5 articles are to be tested every hour, and that each set of 5 consecutive results is to represent a rational sub-group. Should the articles be taken: (a) regularly at intervals of 12 minutes, or (b) 5 at random from each hour's production mixed, say, in a bin, or (c) 5 produced closely together at the end of each hour's run? Method (a) is equivalent to regarding the whole of each hour's production as a sub-group; any pattern of variation that repeats from hour to hour will not contribute to the variation between the means of 5; any pattern of variation that changes from hour to hour together with random variations will make a contribution. Method (b) is also equivalent to regarding the hour's production as a sub-group, but all within-group variations will contribute to the variations between means of 5. If the repeatable pattern of variation is appreciable, method (b) will be less precise than method (a) for the purpose of controlling the hourly level of quality. Usually the repeatable pattern will not be very important, and the two methods will give substantially the same results. Method (c) is equivalent to dividing the production into very many very small rational sub-groups of 5 consecutive articles and testing only one sub-group every hour. Then the allowable variation is reduced to the very small variation between articles produced in a small space of time, and the control of average level of quality will be relatively precise. Where the only important sources of variation are a local random variation and preventable trends such as those due to tool wear, this will be the best method; where there are uncontrollable trends, this method will cause too much searching for assignable causes.

Sample Size and Control Limits

There are two approaches to the problem of deciding on the size of the sample per sub-group and setting control limits; we may term them the empirical and the statistical.

According to the empirical approach, a sample size is chosen on the ground that it seems reasonable to the practical man. It conforms to

past practice, or just about keeps a reasonable inspection or testing staff busy, or is of such a size that testing results are available soon enough for controlling the process. Where the articles are measured, the number per sample is usually somewhere in the neighborhood of 5 or 10.

The three-sigma limits are used on the control chart primarily on the ground that a wide range of experience has shown that a point outside such limits indicates some assignable cause of variation which it is economically profitable to investigate and eliminate. That these limits correspond to certain probability levels is, in the mind of the empiricist, interesting (possibly) but unimportant. Let us examine the statistical approach.

When we use the control chart as a basis for action in connection with assignable causes, we are liable to go wrong in one of two ways. First, a point may be outside whatever limits we adopt, although the population mean may be on the control level; the probability of this happening (i.e., of action being taken unnecessarily) is obtained from the probability level corresponding to the limits. For the 0.001 limits this probability is 0.002, if we take action whenever a point goes above the upper or below the lower limit; and, generally, the wider the spacing of the limits, the lower is this probability. The second type of error arises when the population value goes out of control but the point on the control chart remains within the limits and no action is taken (i.e., when necessary action is omitted). The probability of this error depends on, among other things, the extent of the shift in population value, and its value can not be stated generally. Clearly, however, the more widely the limits are spaced the greater is this probability. For a given probability of the first type of error (i.e., for control limits on a given probability level), the probability of the second type of error can be reduced by reducing the standard error of the quantity plotted, either by increasing n (the sample size) or by reducing σ' (the standard deviation of the individual values) or by doing both. This consideration provides a theoretical basis for deciding the sample size. We have discussed in the previous section some ways of reducing σ'.

If the technician can state the risks he is willing to run of taking unnecessary action, and of failing to take necessary action for a given shift in the control level, the statistician can easily calculate the requisite sample size. Alternatively the technician may be able to state how much it costs when he takes unnecessary action, and this multiplied by the probability of that type of error for a given sample size and control limits gives the "expectation" of cost of the first type of error.

Likewise the cost of failing to take action on a given shift in quality level multiplied by the corresponding probability of that type of error for the given sample and control limits gives the expectation of cost of the second type of error; it may be necessary to average this for all possible shifts in quality level, weighting the average according to the frequency of the various shifts. The sum of the two expectations gives the total cost due to the two types of error for the given sample size and control limits, and it is theoretically possible to find optimum limits for which the total cost is a minimum. The total cost is the sum of these minimum costs due to the two types of error and the sample and testing cost which increases with the sample size. Theoretically it is usually possible to find a sample size for which this total cost is a minimum; this is the optimum sample size.

If the development of the process in time is being followed and a shift in quality persists until corrected, the effect of the second type of error is merely to delay action; sooner or later the shift will be detected whatever the sample size and limits (within reason). This delay can be calculated and can be used in the above calculations instead of the probability of the second type of error.

Calculations of these kinds are relevant when control charts are being put to the first and second uses described in Chapter 3; on what rational basis one can determine for the third use when the evidence of control is strong enough is not clear. These calculations are complicated, and they involve a good deal more knowledge of costs than is generally available. I do not know of any instance where optimum control limits and sample sizes are determined on this basis, and even people who claim to set limits on consideration of probabilities use the probabilities only vaguely and choose them more or less arbitrarily. Thus, if a process is well "engineered," good control is to be expected, a little delay in tracing a small shift in level is not very important, and the limits are set at a low probability level so as to make unnecessary action rare. This is probably the kind of situation in which the three-sigma limits, corresponding to a probability of twice 0.0013, have been found satisfactory. In British publications inner limits at probability levels in the neighbourhood of 0.025 are sometimes proposed as "warning" limits; to take warning without taking drastic action is not very costly. Where a process is new and technical difficulties are being located and eliminated during the course of manufacture, it may be desirable to use such inner limits as a basis for action. Alternatively some help is given by looking out for trends and runs within the control limits.

D

The two approaches to the problems of setting control limits and sample sizes—the empirical and the statistical—appeal differently to different people. There has been a slight tendency for American writers to prefer the empirical basis and for English writers to prefer the other. My own view is that present-day practice is fundamentally empirical anyway, since to choose a probability level more or less arbitrarily on the basis of a vague experience is no less empirical than to choose three-sigma or any other limits on the same basis. The present justification for any control system is the pragmatic one that it works. The two approaches do not lead to seriously different practices; and I doubt that the measurement of the effects of control is at present precise enough to evaluate the effects of even quite substantial differences in control practice. But I have a temperamental objection to relying on empiricism for longer than necessary. We may at any time encounter new experiences for which the old empirical basis may be false. And if progress is to be made in knowledge of the economics of quality control and improved precision is to be attained, it can only be through investigation based on a statistical analysis of the quantities involved. For the sake of future progress, therefore, I lay emphasis on the statistical approach.

Whatever the basis of choice of the control limits, the methods of evaluation are strictly valid only when the population values $\bar{X}'$ and σ' of the mean and standard deviation are known. Sometimes $\bar{X}'$ is given by the specification to which the product is being made, and very occasionally σ' is known *a priori*. When available, such *a priori* knowledge should be used.

When the process is in control, there is no difficulty in estimating the unknown quantities from past test data; it is only necessary for the amount of data to be enough to give reliable estimates. For practical purposes we need not go deeply into this question; only very exceptionally should we be content with fewer than 20 sub-groups, and preferably there should be about 50. When there are 10 sub-groups or fewer, certain theoretical difficulties begin to have importance.

When a process is out of control, the procedure described presents logical difficulties. The grand mean $\bar{X}$ estimates a population mean that can be interpreted only artificially, and, if the variability changes from one sub-group to another, the population standard deviation σ' has no existence. It would be possible to think up peculiar effects of assignable causes that would give charts showing the appearance of control. Nevertheless this does not often occur in practice, and the

procedure usually discloses lack of control, if it exists, and so serves its purpose.

An Example

A paper by Mr. E. W. Harding * is specially interesting because it deals with an application to a subject that at first sight looks unpromising, it describes fully an experimental step-by-step approach, and it raises many of the problems that arise in quality control work. A discussion of this paper will illustrate as well as recapitulate this and the previous chapter.

Meehanite Metal, a high duty iron, is made at a number of foundries, and one objective is to produce at all of them iron having standard chemical and physical properties. The iron is melted in a cupola, which is charged continuously with iron and coke, a ladleful of molten metal being periodically taken away to pour into moulds for castings. In the product there are variations due to variations in the raw material, in the charge weighing and practice, in combustion conditions, and so on.

The past procedures left something to be desired because the operation of the technical controls was left to personal judgment, preventable variations could not be distinguished from those inherent in the process, there were no guides to tracing causes of variations and estimating their importance, and there was no quantitative measure of control by which the progress of a foundry could be followed or one foundry be compared with another.

The first step was to evaluate from past test data the mean and standard deviation of each property for each foundry. Differences in the standard deviation between foundries suggested that this was a good measure of the degree of control achieved at each foundry, and the lack of any relation between the standard deviation and mean suggested that a "standard" value of the variability of each property could be specified for all foundries, irrespective of the type of iron produced. "Standard" values were based on those for the best foundries. Consideration of the circumstances at the foundries suggested that the effectiveness of the control depended less on the control facilities available than on the conscientiousness and consistency of the efforts of the staff. The "standard" values of variability included

---

* "Statistical Control Applied to High Duty Iron Production," Supplement to *The Journal of the Royal Statistical Society*, Vol. 8, 1946, p. 233. A version of the paper, under a similar title, appears also in *The Foundry Trade Journal*, March 16, 23, and 30, 1944.

a large contribution due to testing errors, so that control would include control of the testing methods.

Control charts were then made. For each type of iron the rational sub-group was the quantity produced in the space of a few days, the variation within that time at a good foundry being acceptable as inherent in the process. For testing purposes a specimen of iron was taken from each of about one-seventh of the ladles, the proportion being designed to give 6 selected ladles per sub-group, which therefore contained about 42 ladles. The test result from each specimen formed an individual for the measurement of range. The sample size of 6 was chosen because it involved an amount of testing that was regarded as reasonable.

The selection of ladles was made according to a rota, the first ladle in the first group of 7 being chosen, the second in the second group, and so on, the seventh in the seventh group (belonging to the next rational sub-group), the first in the eighth group, and so on. This calls for comment. Had the variation among the 42 or so ladles within a sub-group been entirely random (i.e., free from trends or other pattern), the method of selection would not have mattered; the adoption of an elaborate method suggested that the technicians were not prepared to assume randomness. It might have been possible to assume that any trends would be well contained within 7 consecutive ladles, so that 7 ladles would form a cluster and the 6 sets could have been regarded as a random sample from an imaginary population of "might-have-been" clusters. Then, in order to have made strictly valid charts, it would have been better to have taken always the same number of ladle in each cluster, say ladle 1 or 4 (see p. 36). The method of selection adopted enhanced the apparent random variation by including any non-random element there might be within the cluster; and since this non-random element did not necessarily contribute to the variations between sub-groups, the control limits might consequently be set too widely to detect some variations between sub-groups. In this instance the method, although perhaps somewhat unorthodox, proved of practical use, as we shall see.

In order to make the system simple, only three properties were measured on each specimen—two chemical and one physical; and the mean and range for all three were plotted on the same chart. Moreover, each foundry was discouraged from making charts for more than three main types of metal. These were concessions to practicability.

It was found that individual foundries could be in control when the control limits were calculated from a grand mean $\bar{\bar{X}}$ and mean range $\bar{R}$

obtained from the records of that foundry, but out of control when the limits were based on the "standards" adopted for the mean and standard deviation of the various properties. The aim was, of course, to achieve control within "standard" limits. The effect of the charts was to stimulate a conscious effort to meet the limits and to help by giving early warning of impending changes, with the result that there was a better control of the average level and a marked reduction in variability.

As a further aid to control, attention was paid to interpreting the data on the charts. One example is given where the range was shown to be out of control, an occasional range being too high. The subgroups for which this obtained were those containing the last ladle in the "heat," and, when the data for the early and late ladles were treated separately, only the latter were found to be out of control. This effect could have been discovered only through the adoption of the rota method of sampling, and it was rightly attributed to a nonrandom within-group effect. Thus the unorthodox sampling scheme coupled with an unorthodox interpretation of results proved to be fruitful. The effect was associated with causes coming into operation at the end of the heat.

Other clues to assignable causes were given according as changes affected all qualities of iron at the foundry or only one, as they affected all properties or only one, as they affected the average or variability or both, as they were abrupt or gradual, and so on. The work was helped by recording on a chart parallel to the control chart all operating changes as they were made.

The whole picture given by the paper is of a scheme that was developed experimentally, not installed ready-made; perhaps the most significant feature is that there is a section headed "Future Developments."

Chapter 5. CONTROL OF THE FRACTION DEFECTIVE

Control Charts for the Fraction Defective

The three preceding chapters contain a discussion of quality control when the quality of the individuals is some measurable quantity and that of the mass is a mean or some measure of variability. Exactly parallel procedures are available when the quality of a mass is specified by the fraction of defective articles. The production may be divided into rational sub-groups, a sample may be taken from each sub-group, and the fraction or number of defective articles may be determined and plotted on a chart. If the number of articles in each sample is the same for all sub-groups, it is convenient to plot the number of defectives, and this procedure will be dealt with here. The methods have been developed for application where the number varies from one sample to another, but they are more troublesome, and this kind of situation should be avoided if possible.

When the quality is measurable, the control limits for the mean or range chart depend on σ', an empirically determined measure of the variation that happens to exist within sub-groups. For the fraction defective, the allowable variation is more fundamentally defined as the kind of variation that occurs in equivalent idealised games of chance. Let us suppose the fraction of defective articles in the population to be p' and the number per sample to be n. Then we may imagine a large churn containing millions of exactly similar tickets except that a fraction p' are marked in some way. If these tickets are thoroughly mixed, samples of n are drawn at random, and the numbers of marked tickets, pn, are plotted on a control chart (p varies from sample to sample); the scatter of the points shows the allowable variation in the number defective in samples of n from a process that is statistically in control. In order to define control limits, we need to know the probability distribution of the number pn.

This distribution has been deduced theoretically and is known as the *binomial distribution*. It depends only on p' and n, and the mean number of defectives per sample, $\bar{p}n$, approaches $p'n$ as the number of samples becomes very large. The distribution is, of course, not Normal, but its exact application presents two difficulties: the probabilities are

44

laborious to calculate, and only exceptionally can limits be found corresponding to any particular probabilities such as 0.025 or 0.001. The second difficulty arises because the number of defectives per sample is a discrete variable—it must be a whole number. Accordingly, approximate control chart methods have been developed.

The first, which is described in British Standards Institution publications on quality control, is based on the fact that, when p' is small (say less than 0.05), the binomial approximates another distribution, the *Poisson distribution*, which is defined not by p' and n separately, but by $p'n$, which is termed the population or expected number of defectives per sample; and tables and charts are available giving the probabilities for various values of $p'n$. The British Standard 1313:1947 gives for various values of the "average number of defectives expected in the sample" (which is np' in our notation), the upper control limit corresponding to a probability level of 0.005. If the production is in control, 1 point in 200 on the average will lie on or above this limit. The limit is given as a whole number of defectives per sample, and only for values of np' for which a whole number defines the 0.005 limit exactly. For other values of np', the next higher value in the table is used and the probability is somewhat less than 0.005.

The second approximate method is the one mostly described in American publications; it is by far the more convenient. The standard deviation (or standard error) of the number of defectives per sample for perfect control is

$$\text{Standard error of } np = \sqrt{np'(1 - p')} \qquad (3)$$

so that if p' and n are known, this can be evaluated, and three-sigma limits can be set at 3 times this standard error above and below the "expected" number $p'n$. This procedure may sometimes give a negative value for the lower limit, which is then conventionally set at zero. Limits will usually come at fractional numbers defective per sample, but, if they are drawn thus, the actual number for each sample will always be definitely inside or outside them. This procedure can be regarded as approximating the exact procedure, since the binomial distribution approaches the Normal form as n becomes very large, provided p' does not become very small. However, the preferred justification is usually that, as in other kinds of chart, three-sigma limits have been found to give good results.

A straightforward extension of either method serves for the formation of charts of the fraction or percentage defective in the sample.

One of these quantities must be used if the total number of articles in the sample varies.

The population fraction defective p' may be known or postulated, or it may be determined as a grand mean $\bar{p}$ obtained from a number of samples (usually 20 or more). If $\bar{p}$ is used as an estimate of p', it is preferable if at all possible for it to be calculated from a series taken when production is in control and p' is constant at the required level. Otherwise the limits will tend to conform somewhat to the actual variations, and the chart will detect changes in the level of control less sensitively.

The same general considerations arise in determining the best sample size for the fraction or number defective as for the mean and range, but no quantitative investigations appear to have been made, and advice on sample size is somewhat arbitrary and is based on vague general experience. A good working rule is to make the samples large enough to give between 1 and 3 defectives per sample on the average. Other problems of application are, in general, the same as for mean and range charts.

Compressed Limits

When the quality of the articles is some dimension, the fraction or number that are defective because they fall outside the limits of go, no-go gauges are alternatives to the mean and standard deviation as a statistical measure for control purposes. Statistically, however, it is a less efficient measure, because for a given sample size and probability level for the control limits there is a higher probability of missing a real change in the control level. This difference in efficiency can be countered by having a larger sample for the fraction defective than for the range and standard deviation in order to give the same precision of control. Then gauging is economical if it is intrinsically so much quicker and cheaper to perform than measuring that the larger sample takes less time and costs less. With the low fractions of defectives usually encountered in inspection, say up to 0.05, the gauged sample needs to be very much larger than the measured sample for equivalent precision, but gauging nevertheless remains very popular on account of its convenience and simplicity.

If the measured quality is distributed Normally, control by gauging is statistically equivalent to estimating the mean and standard deviation from the fractional frequencies α_1, α_2, and α_3 respectively below one limit of the variable, between that and a second limit, and above the second limit. This method of estimation is less efficient statistically

than the methods involving the calculation of the mean and standard deviation because it does not utilise all the information of the full frequency distribution; it merely utilises the three proportionate frequencies α_1, α_2, and α_3. The statistical efficiency of the gauging method has been worked out for large samples, and it is found that there are optimum values of the three proportionate frequencies. For giving warning of small changes in the level of quality, the gauging method is at its best when the gauges are so set that $\alpha_1 = \alpha_3 = 0.25$ (approx.) (i.e., so set that there are about 50 per cent defectives); then samples of 124 articles give as good control as samples of 100 measured and averaged. For smaller samples down to about 20, the numbers in the same ratio give equal control by the two methods. For indicating lack of control due to changes in variability the gauging method is at its best when $\alpha_1 = \alpha_3 = 0.05$ (approx.) (i.e., when there are about 10 per cent of defectives); then samples of 156 articles give as good control against this kind of change as 100 articles measured for the calculation of the standard deviation. If the production is liable to go out of control through a change in either average level or variability or both, a good compromise setting of the gauges would make $\alpha_1 = \alpha_3 = 0.1$, giving about 20 per cent defectives on the average; then about 160 articles gauged give as good control as 100 measured.

Thus gauges set for the purpose of control and not for the purposes of detecting defectives according to the designer's criteria can economically give good control. Such are termed *compressed limit gauges*. It is not necessary for α_1, α_2, and α_3 to have any exact values; it is merely that the method is at its best when the values are near those stated. Once the gauges are set, the results can be put on a control chart for the fraction "defective" in the ordinary way. Since the average fraction is higher than 0.1, the approximation of using the Poisson distribution to calculate the probability limits according to the methods in the British Standards Institution publications is not good, whereas the three-sigma limits are a better approximation to the corresponding Normal probability levels than when the fraction is low.

Where the variability is known to be in control and it is desired to control only the level, a single gauge may be made to the required mean dimension; then, if the distribution is symmetrical in form and the production is at the required level, 50 per cent of the articles will pass through the gauge. Any departure in the fraction "defective" from 0.5 will indicate a change in level. Control by this method is statistically equivalent to estimating the mean of a distribution from the

median, and under most conditions a sample of 160 articles gauged gives as good control as one of 100 measured and averaged.

Two-Way and Related Charts

One drawback of ordinary fraction defective charts is that they give no clue as to what is wrong when a process goes out of control, in the way that mean and range charts do. This can be largely eliminated by having separate charts for np_1, the number of articles below the
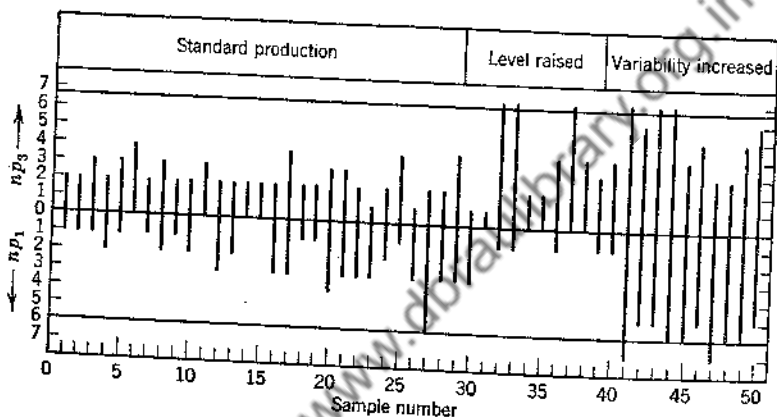
Fig. 8.

lower limit of size, and np_3, the number above the upper limit. The interpretation of such charts is made easier if they are plotted with a common base line for $np_1 = np_3 = 0$, and scales for np_3 and np_1, going respectively upwards and downwards, as shown in Fig. 8. Then, if the production goes out of control through an increase in the mean of the underlying measurable character with no change in the variability, np_3 increases and np_1 decreases, as shown for samples 31 to 40 in Fig. 8; a decrease in the mean produces a decrease in np_3 and an increase in np_1. If the variability of the underlying measurable character increases, with no change in the mean, then both np_3 and np_1, increase, as shown for samples 41 to 50 in Fig. 8.

It may be interesting to consider how Fig. 8 was arrived at. It presents the results of an artificial sampling experiment. For the "standard production" of samples 1 to 30, 30 sets of 20 two-digit numbers were taken from one of the existing tables of *random numbers* (i.e., numbers formed by combining the digits 0 to 9 in all sorts of

ways entirely at random). In each set of 20 the number of numbers between 90 and 99 (inclusive) was counted and entered as np_3; thus the population fraction p'_3 was 0.10. We may imagine np_3 to be the numbers of articles, in 30 samples of 20 taken from a bulk production in control, that are larger than a limit of size so set that in the bulk a fraction 0.10 are larger. Thus the values np_3 are the sample numbers of defectives when the population fraction is 0.10. In the same 30 sets of 20 numbers, those between 00 and 09 (inclusive) were entered as np_1; they correspond to $p'_1 = 0.10$. Then for samples 31 to 40 the same procedure was adopted, except that it was imagined that the mean dimension had increased so as to make $p'_3 = 0.25$ and $p'_1 = 0.03$. The fraction p'_3 was chosen arbitrarily to represent a substantial change, and p'_1 was calculated to correspond, a Normal distribution of the underlying quality being assumed. The change from $p'_1 = p'_3 = 0.10$ to $p'_1 = 0.03$ and $p'_3 = 0.25$ corresponds to an increase in mean of 0.61 times the standard deviation. For samples 41 to 50 the experiment was continued with $p'_1 = p'_3 = 0.25$, the change from the "standard production" corresponding to increasing the standard deviation to 1.9 times its original value.

Next comes the calculation of the control limits. In Fig. 8 are shown three-sigma limits. These could be calculated from the known p'_1 and $p'_3 = 0.10$ for standard production, but in order to simulate somewhat practical conditions $n\bar{p}_1$ and $n\bar{p}_3$ were calculated from the first 30 results. They are: $n\bar{p}_1 = 1.97$, whence $\bar{p}_1 = 1.97/20 = 0.0985$ and, according to equation (3), the standard error is 1.33, giving the upper three-sigma limit at 6.0; $n\bar{p}_3 = 2.33$, whence $\bar{p}_3 = 0.1165$ and the standard error is 1.44, giving the upper three-sigma limit at 6.6.

Now it is seen from Fig. 8 that for samples 1 to 30 all the fluctuations are within the limits except that for sample 27 np_1 approaches its limit; for samples 31 to 40 np_3 goes outside its limit three times, giving clear evidence of the change in mean; and for samples 41 to 50 both np_1 and np_3 approach or cross their limits several times. It will be noticed, however, that although the changes underlying the results of Fig. 8 are substantial, several results are within the three-sigma limits, especially those for samples 31 to 40. The samples of 20 are only just large enough to detect the changes if three-sigma limits are used and interpreted strictly.

The limits could have been calculated from $\bar{p}_1$ and $\bar{p}_3$ calculated from the 50 sets of results. Then the limits would have been even more widely spaced and the evidence of change even less clear.

An alternative method of expression of the results of a two-way chart is to plot

$$a = np_3 - np_1$$
$$v = np_3 + np_1 \tag{4}$$

If the mean of the underlying quality increases, so does a, whereas v changes very little; a chart of a thus corresponds to the mean chart (as a mnemonic you may remember that a is the initial letter of average). If the variability alone increases, a is almost unchanged and v increases. Values of a and v for the results of Fig. 8 are plotted in Fig. 9.
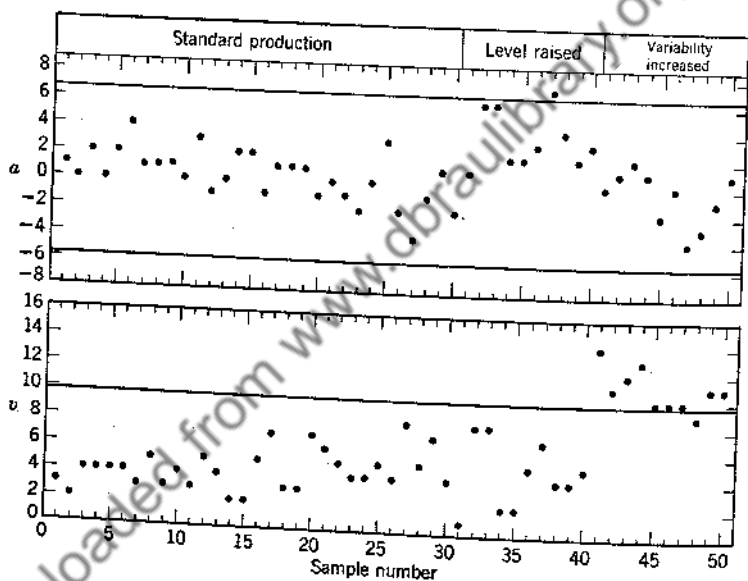

FIG. 9.

For the calculation of the control limits, use may be made of equations (5).

$$\text{Standard error of } a = \sqrt{\left(v' - \frac{\overline{a'^2}}{n}\right)}$$
$$\text{Standard error of } v = \sqrt{\frac{\overline{v'(n - v')}}{n}} \tag{5}$$

where a' and v' are population mean values of a and v; that is,

$$a' = np'_3 - np'_1$$

$$v' = np'_3 + np'_1$$

If a' and v' are not known, $\bar{a}$ and $\bar{v}$, averages estimated from a series of actual results, are used instead. Equations (5) apply only when n is "large" and do not give good approximations when n is less than about 20. Indeed, v is merely the number of "defective" articles in a sample and the second of equations (5) is exactly equivalent to equation (3).

For the data of Fig. 8 we know that, during the period of "standard production," $a' = 0$ and $v' = 4$, but the limits drawn in the figure have been calculated from $\bar{a} = 0.367$ and $\bar{v} = 4.300$, calculated from the data. The standard error of a is then 2.07 and the three-sigma are at $0.37 \pm 6.21 = 6.58$ and -5.84. The standard error of v is 1.84, and the three-sigma limits are at $4.30 \pm 5.52 = 9.82$ and -1.22. Because a negative value for v is impossible, we set the lower limit at $v = 0$. All the above calculations were done with a slide-rule, and the second decimal place may be two or three units in error.

The movements in a and v relative to the control limits on Fig. 9 lead to substantially the same conclusions as the two-way chart of Fig. 8. Any preference for one chart or the other is perhaps a matter of taste.

A full theoretical discussion of the methods of these last two sections is given in a paper by Mr. W. L. Stevens.* There, among other things, are given rather more refined methods for calculating control limits.

These methods of control by gauging are not widely used, although they have been much used by one or two people. One possible limitation mentioned by a contributor to the discussion of Stevens's paper is the effect of errors due to wear in gauges and the human variations in their operation. The application of the methods thus calls for more practical investigation, but they are promising enough to deserve more extensive trial.

* "Control by gauging," *Journal of the Royal Statistical Society,* Series B, Vol. 10, 1948, p. 54.

Chapter 6. SPECIAL APPLICATIONS AND ADAPTATIONS OF THE CONTROL CHART

Time Series

Sometimes data are presented in the form of a time series in which there is one result for each of a number of successive times, and one wants to know if there are any evidences of non-random variations attributable to assignable causes. Such data arise perhaps more often in efficiency and production control than in quality control.
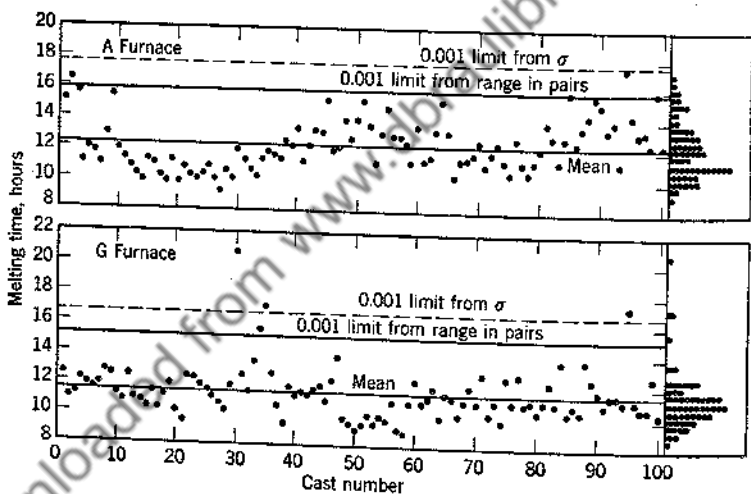
Fig. 10.

In Fig. 10 are plotted for two steel furnaces the times to melt 100 successive casts (the data are from *Statistical Methods in Industry*). Since there is only one value for each cast, there is a problem in deciding how to calculate control limits. The data are accumulated into the form of frequency distributions to the right of each chart; the distributions are not Normal, that for G Furnace being markedly asymmetrical—but that does not necessarily signify the presence of assignable causes. If the standard deviation is calculated from the root-mean-square deviation from the mean [equation (1), p. 7] and control

52

limits determined from this, the existence of an undue number of points outside the limits will merely signify the departure from Normality. For A Furnace the mean is 12.27 hours, the standard deviation so estimated is 1.718 hours, and the 0.001 upper limit is at $12.27 + (3.09 \times 1.718) = 17.58$ hours. This is drawn in a dotted line in Fig. 10, and only one value comes near it; such moderate departures from Normality as that of the distribution of A Furnace are not clearly demonstrated in this way. For G Furnace the mean is 11.50 hours, the standard deviation is 1.687 hours, and the 0.001 upper limit is 16.71 hours. This is drawn in a dotted line in Fig. 10, and there are three values well beyond it, owing to the extreme departure from Normality. The lower 0.001 limits have not been drawn, because there are no large deviations below the mean value.

The chart for A Furnace gives a strong suggestion of systematic changes, the melting time starting high for the first 2 casts, then remaining low for casts 10 to 30, increasing to about the mean level for casts 30 to 80 or so, and finally increasing still further. Superimposed on these slow changes are apparently random variations, which we may regard as the natural allowable variations of the process. How may we evaluate them? The difference between successive pairs of results (i.e., between the first and second, the second and third, and so on) is affected only by the random variation plus the variation due to the extent of the slow movement from one result to the next. This second contribution is negligibly small for A Furnace, so that for our purposes the mean difference between pairs may be regarded as a mean range in samples of 2, estimating the random variation. For A Furnace this mean range is 1.325 hours, and the estimated standard deviation (Table IV) is $1.325 \div 1.128 = 1.175$ hours. This is very much less than the value of 1.718 hours which measures the random variation plus that due to the slow movements. The 0.001 upper control limit for the random variation is $12.27 + (3.09 \times 1.175) = 15.90$ hours, and this is drawn in Fig. 10. Four points are outside this limit, demonstrating the reality of the slow movement. In this instance, the control chart with its limit perhaps tells us little more than an inspection of the points without the limits tells, but the conclusion is more clearly established, and with the limit the technician will know when to investigate or take action as the chart is continued and subsequent points are added.

The slow movements are not so apparent for G Furnace, but the mean range in pairs is 1.358 giving an estimated standard deviation of $1.358 \div 1.128 = 1.204$ hours. This also is less than the estimate of 1.687 hours from the full distribution, suggesting that high values are

not entirely "bolts from the blue"; the chart usually climbs to the high value and climbs down from it through the two or three values before and after. The control limit based on the lower standard deviation for G Furnace in Fig. 10 does not appreciably alter the conclusions reached with the aid of the other 0.001 limit.

This simple adaptation of the control chart procedure for dealing with time series does not always "work"; it fails entirely when high and low values tend to alternate. But it often proves a useful adjunct to the visual examination of the trends on a time chart. There are, of course, other and much more elaborate ways of dealing with time series.

The Group Control Chart

The group control chart was first described by Dr. E. H. Sealy.* It is designed to give control of a group of similar and independently adjustable machines or heads on a machine in a way that is more economical than making a separate control chart for each. At each testing time, two or more articles are tested from each machine, and the separate means and ranges are calculated just as though they were to be put on separate charts for the machines. For the group chart, the limits are calculated from the common mean and a common mean range for all machines. Then on each occasion the highest and lowest of the machine means are plotted on the mean chart, and the corresponding machine number is also recorded. On the range chart the highest of the machine ranges is plotted, and again the machine number is also recorded. The chart indicates when any machine in the group goes out of control and which one it is, and the record of the numbers shows whether any machines have been giving more trouble than the others.

The method does not appear to be very good where several machines are likely to go out of control at the same time.

Modified Control Limits

All the procedures so far described are based on the theory that perfect statistical control as defined in Chapter 2 is practicable and economical; that the rational sub-groups are arranged to contain all the allowable variation. In machining articles to specified dimensions with tolerances, however, tool wear and trends due to other causes are inevitable, and it is uneconomical to attempt by frequent re-settings and replacements to correct for the consequent departures from perfect

* *A First Guide to Quality Control for Engineers*, London, Ministry of Supply.

statistical control. It is necessary to maintain a control of the process that is sufficient but short of perfect control. This situation will be discussed in terms of machining, but it doubtless arises also in some other processes.

The situation is set out diagrammatically in Fig. 11. The dimension is represented along the y-axis and the drawing dimension and tolerance
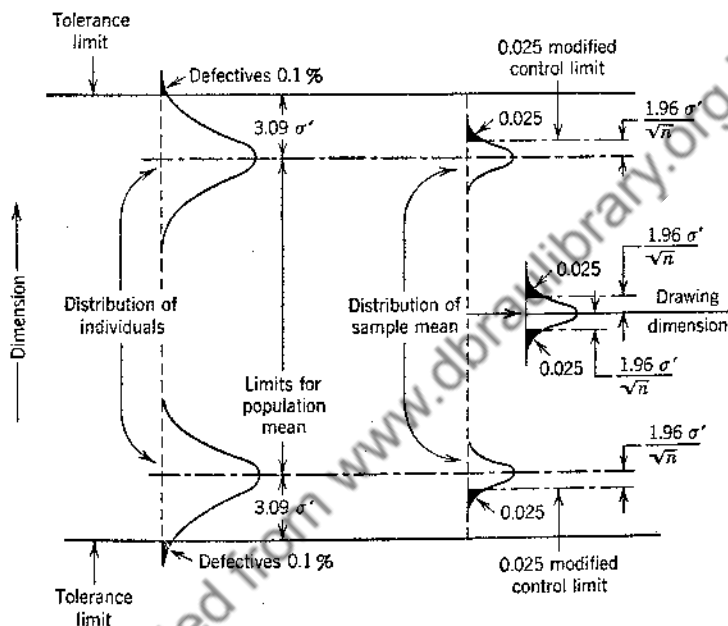
Fig. 11.

limits are represented by continuous straight lines. It is assumed that the production needs to be maintained so that at no time are more than 0.1 per cent of the articles over or under size, and this requires that at all times the "population" mean dimension shall be between limits within the tolerance limits. The imagined frequency distributions of the individual articles are drawn in Fig. 11 when the population mean is at the two levels that give 0.1 per cent defectives. If the distribution is Normal with a standard deviation of σ', these levels are $3.09\sigma'$ within the tolerance limits (Table III); they are marked in Fig. 11 as "limits for population mean." The production requirements are satisfied if the population or lot mean dimension of the articles produced at any one time is between these limits.

E

Now suppose that the number of articles per sample is n and that the 0.025 control limits are in use. For perfect statistical control the limits would be set at $1.96\sigma'/\sqrt{n}$ above and below the drawing dimension, as shown in Fig. 11. Since, however, the population mean may move between certain limits, the control limits may be drawn at $1.96\sigma'/\sqrt{n}$ above the upper limit for the population mean and at a similar distance below the lower population limit. These are called *modified control limits*, and they are drawn in Fig. 11. Corresponding limits may, of course, be calculated for other percentages of defectives and probability levels, and (more rarely) for other forms of frequency distribution.

The modified control limits are used just like any other control limits, the machine being re-set only when a point goes outside them, and such charts have apparently been satisfactory in practice. Clearly the variability must be strictly in control. Practical experience, however, is not always a good test of theoretical validity, and there are theoretical difficulties about modified control limits. It is logically fallacious to transfer all the arguments that apply when there is statistical control to the uncontrolled situation. Why are the modified control limits set outside the limits for the population mean rather than inside? The population mean is almost sure to move towards one or other population limit and may easily move beyond before a sample mean goes outside the modified control limits as shown. If there is statistical control, 0.05 of the sample points lie outside the conventional 0.025 control limits; the proportion of points lying outside the 0.025 modified limits will be 0.025 if the population mean remains exactly on one or other of the limits for the population mean, and less if it spends any appreciable time between those limits. It will only reach 0.05 if the population mean moves outside its limits. Indeed, in order to describe the results of any state of "controlled uncontrol" in terms of probabilities corresponding to those used in describing the state of control, it is necessary to make assumptions about the form of the allowed variation in the population mean.

One simple assumption is that the population mean changes linearly with time in the neighbourhood of the time when action is taken. For the sake of a concrete picture let us imagine a tool wearing so that the mean dimension of the articles produced increases linearly with time, and at a certain stage let the tool be discarded and the machine be fitted with a new tool and re-set. Then it is possible to work out a solution in terms of the following quantities:

The rate of wear of the tool measured by the rate of change of the population mean with time.

The intervals at which samples are tested.

The sampling distribution of the sample mean: especially its standard error.

The control limit of sample mean value at which the tool is discarded.

The probability distribution of the state of wear at which the tools are discarded, and in particular the level of population mean dimension that is exceeded with a given (low) probability.

If any four of these sets of quantities are known or can be assumed, the fifth can be deduced. The subject is dealt with and tables are given in a paper, "The Control of Industrial Processes Subject to Trends in Quality." † I have not heard of anyone who has made successful use of these results, but, if quality control is to develop soundly and to give greater precision, not only must practical application be successful but also the theoretical basis must be sound.

It is possible to use go, no-go gauges to give the kind of control achieved with the aid of modified limits of a measured dimension. If the frequency distribution of the dimension of individual articles is symmetrical (it need not be Normal) and limit gauges are set to the two dimensions designated in Fig. 11 as limits for population mean, the population mean will satisfy the required conditions if no more than 50 per cent of the articles lie above the limit set by the larger gauge or below that set by the smaller. This condition can be established by forming a two-way fraction defective chart for the two gauges, similar to that of Fig. 8, calculating the limits on the basis of $p'_1 = p'_3 = 0.5$.

† *Biometrika*, Vol. XXXIII, 1944, p. 163.

Chapter 7. ACCEPTANCE SAMPLING

In quality control as so far described, test data are used to trace backwards towards the manufacturing source assignable causes of variation. Another use is to decide what to do with the products in subsequent processes. If successive lots of any product are the same in quality, whether that is measured statistically or not, they will all be treated the same, and the only purpose of the test data may be to give assurance that the quality is the same (i.e., in control)—such assurance a control chart gives. The evidence of a control chart, where available, is much preferable to that supplied by the best acceptance sampling scheme. Sometimes, however, there is no evidence or expectation of control, as when lots of raw material are presented from unknown sources, and test results are then required as a guide to subsequent action. Such action may involve rejection if the lot is unsuitable, allocating the lot to a particular grade for subsequent use (e.g., a given lot of bolting silk will be classified and used according to the size of holes in the mesh), or adapting the later process to the quality of the material (as in some metallurgical and chemical manufacture where the quantities of the various constituents are decided according to the analytical characteristics of the lots supplied).

If all the material in the lot is tested, as when there is "100 per cent inspection," the appropriate action is purely a technical question. Statistics enters when the quality of the lot is appraised from a sample, as must occur when the test is destructive, and as often occurs for reasons of economy. The sample does not represent the lot exactly, and, if action is taken on the basis of the sample result, there is inevitably a risk that the action may be unsuitable; it is the business of statistics to calculate such risks.

There is an attitude that in some circumstances such risks can not be run and that then the lot must be tested *in toto*. The fact that serious accidents and even disasters occur shows that we can not eliminate all risks from life, but it is true that there are circumstances in which risks must be reduced to such a low level that it is impracticable to evaluate them by ordinary statistical calculations. None of us is prepared to travel by aeroplanes inspected on a system that consciously leads to any appreciable risk of the machine failing. But most factory inspection systems are subject to human fallibility, and this

58

sometimes introduces risks that are comparable in magnitude with statistical risks arising in sampling schemes. In more than one actual experience a lot of articles has been subjected to 100 per cent inspection for sorting into defectives and non-defectives, and subsequent re-inspection has disclosed many defective articles among the "non-defectives" of the first inspection and many non-defective articles among the "defectives." In such circumstances a reduction in the amount of inspection made possible by the adoption of a sample scheme, with the resulting improvement in the quality of the inspection, may reduce inspection errors by so much that the sample scheme gives improved control in spite of sampling errors. The non-statistical risks have not been studied systematically, and they must be left for special investigation. We must assume that the quality of inspection is perfect and deal only with the statistical risks associated with the use of samples.

Action as a result of inspection may be a simple choice between the alternatives of acceptance and rejection, the alternatives may also include taking a further sample or subjecting the lot to 100 per cent inspection for the replacement or rectification of defective articles, or some complicated quantitative adjustment may be made at some subsequent process. The complete subject covering all these possibilities would be vast and has not been fully developed. Even so, the part that has been developed is considerable, and here we deal with only a few simple situations for the purpose of introducing the main ideas and quantities used in sampling inspection.

Our discussion will be confined to terms of the sampling of articles or manufactured pieces, but the theory applies equally to statistical individuals or sub-sub-groups as described in the first section of Chapter 4.

The simplest situation is one in which a "producer" offers a lot which the "consumer" either accepts or rejects. You will have no difficulty in generalising these terms. The producer may be a vendor of raw materials or a manufacturer; the consumer may be a manufacturer who uses the products in his work or the ultimate consumer; and the two may belong to the same or separate concerns.

Inspection by Measurable Quantities

The exposition of this section is based on an example taken from a paper by Mr. W. T. Hale.* Mr. Hale's data are used to expound the

---

* "A statistical sampling plan for refractory products with special reference to silica bricks," *Transactions of the Ceramic Society,* Vol. 46, 1947, p. 147.

principles of sampling, not to expound his scheme; for a full discussion
of his problem you should refer to the original paper.

A property of the bricks that interests the consumer is the specific
gravity, which is required to be low. Mr. Hale suggests a sampling
inspection scheme of the simplest possible type (viz., that 4 bricks
should be tested from each lot—in this case a load; if the mean specific
gravity of these 4 is less than 2.365 the lot should be accepted, other-
wise it should be rejected). Let us consider the consequences of such
a scheme. In practice, one wishes to specify a scheme to satisfy cer-
tain conditions, but it is easier for our exposition to work in the op-
posite direction.

I have calculated from the data given in Mr. Hale's paper that, when
the effects of systematic variations between and within kilns are elimi-
nated, the standard deviation of specific gravity is 0.0132 unit, so that
the standard error of the mean of 4 bricks is, by equation (2), 0.0066
unit. This is the standard error with which a sample of 4 bricks esti-
mates a lot mean, provided that (1) the lot is comprised of all the
bricks made at one firing in one kiln, (2) the 4 bricks are always taken
from the same positions in the kiln, and (3) the systematic position
effect is known. This is referred to later, in Chapter 10 (p. 120).

Now suppose that many lots are presented, each with a "true" mean
specific gravity of 2.358 say, and that 1 sample of 4 is taken from each;
then the frequency distribution of the sample means from these lots
will be the distribution in the upper part of Fig. 12 centred on 2.358.
The proportion of loads that will be accepted is represented by the
area under the frequency curve to the left of the ordinate at 2.365; this
area is bounded by a thick line and may be calculated from an extended
version of Table III. The ratio $t = (2.365 - 2.358) \div 0.0066 = 1.06$
and α (from the extended tables) is 0.145, so that the proportion of
area to the left of the bounding ordinate is 0.855. This is plotted
against the true mean for these lots, the population mean, in the lower
part of Fig. 12, the point being marked by a cross. The fraction 0.855
is the probability that any load having a mean specific gravity of
2.358 will be accepted according to the specified sampling scheme.

The distribution of means for samples from lots having a true mean
of 2.374 is also represented in the upper part of Fig. 12, and the proba-
bility of acceptance of 0.087 is plotted against 2.374 in the lower part.
Corresponding probabilities of acceptance for other values of the lot
mean can be similarly deduced, and when plotted they fall on the curve
shown in the lower part of Fig. 12. This is termed the *operating char-
acteristic* (or OC) *curve* of the sampling scheme, and it gives a complete

statistical description of the consequences of the scheme. The probability of accepting a lot of bricks can be read directly from the diagram if the true mean is known, and the probability of rejection is one minus the probability of acceptance.
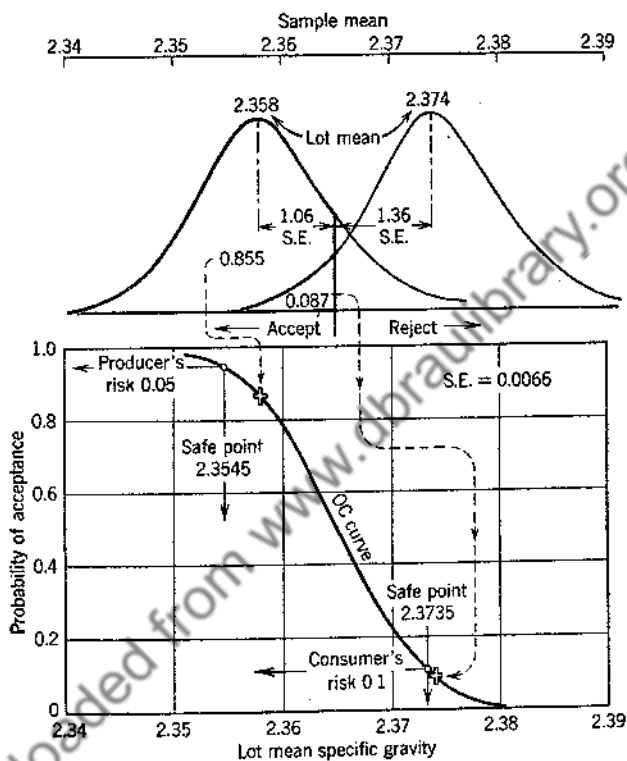
Fig. 12.

From Fig. 12 we see that any lot with a true mean specific gravity of 2.35 or less is almost sure to be accepted, and any with a true mean greater than 2.38 is almost sure to be rejected; lots between 2.35 and 2.36 have a good chance of being accepted, and those between 2.37 and 2.38 have a poor chance of being accepted (or a good chance of being rejected); between 2.36 and 2.37 is a region of great uncertainty, the lot having a moderate chance of acceptance. The effect of the errors of random sampling is to produce this region of uncertainty or poor discrimination between lots with low and high mean specific gravities;

the greater the errors the broader is the region of uncertainty. For perfect discrimination with no sampling errors, the operating characteristic curve would be like this ⌐⌐ , the probability of acceptance dropping suddenly from 1.0 to 0.0 at a value of the lot mean that distinguishes good from bad lots.

It is not feasible to make use of a whole characteristic curve, and so we confine attention to one or two points on it. Let us look at things first from the consumer's point of view. He would prefer not to accept lots with a high mean specific gravity but, knowing that certainty is impossible, is willing to run a small risk of accepting lots with a high mean; this is termed the *consumer's risk*, and the corresponding value of the lot mean is the *consumer's safe point*. For a consumer's risk of 0.1 the safe point is 2.3735, and this point is marked on the OC curve in Fig. 12. The probability of accepting a load with a true mean greater than 2.3735 is less than 0.1, and in this sense the consumer is safeguarded by the scheme against accepting lots as bad as or worse than this; he is fairly safe to assume that the mean specific gravity of bricks in accepted loads will be less than 2.3735.

Generally, if the sampling distribution is Normal and a low value of the variable is required, it can be shown that

$$\text{Consumer's safe point} = L + (t_c \times \text{S.E.}) \qquad (6)$$

where L is the limiting sample value on the borderline between rejection and acceptance

     t_c is the Normal deviation corresponding to the given consumer's risk, as given by Table III

   S.E. is the standard error of the mean.

Here, for a risk of 0.1, $t_c = 1.28$, S.E. $= 0.0066$, $L = 2.365$, and the consumer's safe point $= 2.365 + (1.28 \times 0.0066) = 2.3735$.

In practice, of course, the consumer's risk and safe point are chosen, *a priori*, and the limit L is deduced from equation (6) to give the required safeguard. The choice of the safe point is not difficult in principle, if we have sufficient technical information. For example, the level of specific gravity above which an individual brick would be unsatisfactory in service, and hence a defective, is known; and, from a study of the frequency distribution of specific gravities, the mean specific gravity corresponding to an acceptably low fraction of defectives can be stated; this would be the consumer's safe point. The choice of the corresponding risk is also a matter for the technician rather than for the statistician, but it is more difficult. If the conse-

quences of accepting unsatisfactory lots are serious, the risk will be low, but no attempts have been made to establish such risks objectively, and the choice is based primarily on subjective judgment. A risk of 0.1 is quite commonly acceptable. Until there is a firm basis for choosing the risk, the design of a sampling scheme can not be truly scientific.

In the foregoing, the standard error of the sample mean (and hence the sample size) is taken as given, and within very wide limits, whatever the standard error, an acceptance/rejection limit can be found to give the consumer any required safeguard. When considering the reasonableness and economy of the scheme, the sample size must be taken into account, and to do this we must now look at things from the producer's point of view.

The producer wants to avoid the rejection of loads of bricks but is willing to run a small risk of rejection, called the *producer's risk*, which for the sake of example we may put at 0.05. The corresponding probability of acceptance is 0.95, and from the OC curve (Fig. 12) we see that the corresponding lot mean specific gravity is 2.3545; this is the *producer's safe point*. It is a safe point in the sense that, if, perhaps with the aid of quality control charts, he always produces bricks at a mean level of specific gravity, he runs only the acceptable risk of having each load rejected; in our example, only 5 per cent of the lots will be rejected.

Generally, if the sampling distribution is Normal and a low value of the variable is required, we have as a complement to equation (6) the following:

$$\text{Producer's safe point} = L - (t_p \times \text{S.E.}) \qquad (7)$$

where t_p is the Normal deviation corresponding to the given producer's risk, as given by Table III. Usually L will be given by the consumer's requirements; the choice of the producer's risk is subject to the same general considerations as that of the consumer's risk.

By combining equations (6) and (7) we see that

Difference between consumer's and producer's safe points

$$= (t_c + t_p)\text{S.E.} \qquad (8)$$

Equation (8) shows that, as a result of sampling errors, the producer must make to a higher quality (a lower mean specific gravity) than the consumer feels safe to assume in using the bricks, and this difference will usually cause economic loss or add to the ultimate costs. For example, the reduction in specific gravity of the bricks requires more

prolonged firing. This loss can not be reduced by reducing t_c and t_p, for that will merely involve increasing the risks, and such would be an ostrich-with-its-head-in-the-sand policy. The only way is to reduce the standard error. The adoption of a good sampling technique can often go some way towards this by reducing the value of the standard deviation s' that goes into equation (2). But, for a given material sampled by a given technique, the standard error is determined by the number in the sample. The larger the sample size, the costlier is the inspection, but the smaller is the loss due to the difference between the consumer's and producer's safe points; it is easy to see that usually there will be a most economical size of sample. It is doubtful that costings investigations are often done to determine this size, but perhaps they will be in the future as the principles of sampling become more widely understood.

In the absence of knowledge for determining an optimum sampling scheme, it is necessary to choose the various basic quantities in some way, perhaps by estimation from incomplete data, or by inspired guessing, or even arbitrarily; then the calculation of the details of a scheme is easy. Thus, if the producer's and consumer's risks and safe points are known, equation (8) can be used to determine the standard error, and, if the standard deviation is known, the sample size can be calculated according to equation (2). Then either equation (6) or equation (7) can be used to calculate L, and the sample scheme is complete. For example, suppose that the consumer's risk is 0.05 ($t_c = 1.645$), the consumer's safe point is 2.37, the producer's risk is 0.01 ($t_p = 2.33$), and the producer's safe point is 2.36. Then from equation (8), $0.01 = 3.975 \times S.E.$, and S.E. $= 0.0025$. The standard deviation is 0.0132, and the necessary sample size is $n = 0.0132^2 \div 0.0025^2 = 5.28^2 = 28$ (approx.). Then, if we use a sample size of 28 exactly, the standard error is $0.0132/\sqrt{28} = 0.00250$, and, from equation (7), $L = 2.37 - (1.645 \times 0.00250) = 2.366$ units of specific gravity. This will not give exactly the specified producer's safe point because of the approximations used in the calculations. In practice there would be some "juggling" with the quantities to give a convenient sample size in the neighbourhood of 28, perhaps 30, and a convenient "round" figure for L.

You may wonder why it is necessary to go through this elaborate procedure when there is an element of guess-work or judgment in the choice of some of the basic quantities, and whether it would not be just as good in such circumstances to choose the final scheme directly by estimation. The kind of procedure outlined is best; it gives the

best scheme that can be devised in the light of such knowledge as exists at the time, and it focusses attention on the points on which more data are required.

The equations of this section are limited to the case where assurance is required that the lot mean value is below some limiting value; it is not difficult to modify them for the case where a high value of the mean is preferred. The situation is somewhat more complicated when the lot mean is required to be between two limits, but the same general principles and ideas apply.

When the quality of the lot is specified by one of the measures of variability, the same general considerations apply as for the mean. The sampling distribution of the measure enables the operating characteristic curve to be calculated, and this can be interpreted as any other operating characteristic curve. Indeed, the situation will usually be much the same as for the mean specific gravity of bricks in that the sampling scheme will be usually required to protect the consumer against the value being greater than some chosen limit; but equations (6) to (8) will not apply because of the difference in the sampling distribution. It will usually be more difficult to find technical grounds for choosing the various acceptable limits of variability.

The problem of dealing with the situation when quality is specified by more than one measure is too complicated to be dealt with here.

You will notice that the standard error depends on n, the number in the sample, and not on the proportion the sample is of the lot. The relation between the size of sample and that of the lot becomes important only when a large proportion of the lot is inspected.

Acceptance/Rejection Procedure for Fraction Defective

When the quality of a lot of articles is assessed by the fraction that are "defective" (in the generalised sense described on p. 10), sampling schemes analogous to those just described can be evolved. In the simplest form of scheme, a sample of definite size, say n, is taken and inspected, and the lot is accepted if there are up to, say, c defectives, and rejected if there are more than c, where c may be 0, or 1, or 2, or any number less than n. The total consequences of such a scheme are entirely described by the operating characteristic (OC) curve. The number c is termed the *acceptance number*.

If the number in the lot is large compared with that in the sample (say more than 4 or 5 times as large), the probability of acceptance may be computed from the theoretical sampling distribution known as the binomial distribution, which leads to the following formulae:

Probability of acceptance of lot =

$(1 - p')^n$ when $c = 0$, or

$(1 - p')^n + np'(1 - p')^{n-1}$ when $c = 1$, or   (9)

$(1 - p')^n + np'(1 - p')^{n-1} + \dfrac{n(n-1)}{2} p'^2(1 - p')^{n-2}$ when $c = 2$

and so on, where p' is the fraction defective in the lot. From equations (9) the OC curve can be calculated for any simple sampling scheme (i.e., for any given value of n and c). It will be noted that
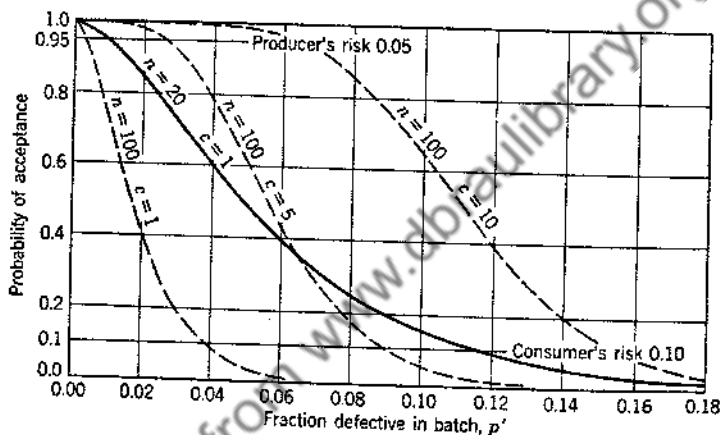
Fig. 13.

whereas we require three quantities to determine the OC curve when the quantity is measurable—the mean, the standard deviation of individuals, and the sample number—we need only two when the classification is for defectives—the fraction defective in the lot and the sample number. This is a result of statistical theory.

Some operating characteristic curves are given in Fig. 13, and, as before, we may choose consumer's and producer's risks and read off the corresponding safe points. The producer's safe point is termed the *acceptable quality level* (AQL) and the consumer's, the *lot tolerance per cent defective* (LTPD). In some writings these definitions are limited to values associated with specified risks.

Let us choose for illustration a producer's risk of 0.05 (probability of acceptance 0.95) and a consumer's risk of 0.1. When the sample size is 20 and each lot is accepted if the sample contains up to 1 (or 5

per cent) defective, the producer's safe point is at about $p' = 0.01$ and the consumer's is at about $p' = 0.115$ (bearing in mind the values of the chosen risks we may say that lots with fewer than 1 per cent defectives will probably be accepted, those with more than $11\frac{1}{2}$ per cent will probably be rejected, and those with values in between will have a very uncertain fate). Clearly this scheme is not very discriminative.

If n is increased to 100 and c is 20 (i.e., is retained at 5 per cent), the OC curve approaches somewhat more closely the ideal for certain discrimination, falling more steeply from a high to a low probability of acceptance, and the two safe points are closer together at $p' = 0.025$ ($2\frac{1}{2}$ per cent) and $p' = 0.09$ (9 per cent). We are now quite used to the idea that increased sample size makes possible improved discrimination between good and bad lots.

It is interesting to see what happens when c alone is changed, n being kept constant. Figure 13 gives OC curves for $n = 100$ and $c = 1, 5,$ and 10. The safe points are:

	Producer's 0.05 Safe Point		Consumer's 0.10 Safe Point	
$c = 1$	$p' = 0.002$	AQL = 0.2	$p' = 0.040$	LTPD = 4.0
$c = 5$	$p' = 0.025$	AQL = 2.5	$p' = 0.090$	LTPD = 9.0
$c = 10$	$p' = 0.065$	AQL = 6.5	$p' = 0.155$	LTPD = 15.5

The effect of increasing c is to raise both the consumer's and the producer's safe points for given risks, the former more than the latter for the small values of c that are commonly used, so that the range of values of p' in the region of uncertain fate is increased.

It is useful to be able to work out the consequences of a given sampling scheme defined by n and c, as we have done in Fig. 13, but it is more useful to state the requirements and determine the appropriate scheme. It is not difficult from published charts and tables of the binominal distribution to deduce a full range of values of n, c, and p' corresponding to given consumer's and producer's risks, so that, if the two values of p' are specified, n and c can be defined. There are no published charts or tables specifically arranged for the type of simple sampling described here.

Since n and c can vary only in units, it will not often be possible to find values corresponding exactly to chosen producer's and consumer's safe points and risks, but this presents no practical difficulties.

It will be noticed again that, provided the sample is small compared with the size of the lot, the number in the lot does not enter into the calculations; the sample size rather than its proportion to the number

in the lot determines the power of the scheme to discriminate between good and bad lots.

Acceptance/Rectification Procedure for Fraction Defective

All the practices and terminology of sampling owe much to workers associated with the Bell Telephone Laboratories, and notably to Mr. H. F. Dodge and Dr. H. G. Romig, but this section owes almost everything to them.

The schemes described in the previous section are sometimes termed *batch sentencing* schemes as opposed to *acceptance/rectification* or *screening* schemes. The latter are specially appropriate where lots are presented in a stream, as in mass production in a factory. If for each lot the sample has c defectives or fewer, the lot is accepted or passed forward without change; if the sample contains more than c defectives, the lot is subjected to 100 per cent inspection, and defective articles are either replaced or corrected (rectified) before it is passed forward. Operating characteristic curves can be calculated from equations (9), and hence the safe points corresponding to various risks, but when the lots are presented in a stream two new ideas and quantities arise.

The first is the average sample number (ASN). The probability of acceptance is the proportion of lots that are passed forward as a result of sample inspection only, for which the sample size is n; the remainder are fully inspected and have a "sample" size of N, say, where N is the number in the lot (this is the first time we have had to consider N). In the long run the average sample size (or number—the term average sample size does not lend itself happily to abbreviation) is the weighted mean of n and N, the weights being the two proportions. If p' is very low, few defectives appear in the samples, most lots are accepted without full inspection, and the ASN is little greater than n; as p' increases, the ASN increases, until at $p' = 1.0$ it becomes N (it becomes near to N long before that!).

There are a large number of sample schemes (combinations of n and c) that give approximately the same consumer's safe point (or LTPD) for a given risk. For example, it can easily be seen from Fig. 13 that at a risk of 0.10 a safe point of $p' = 0.12$ is given when $n = 20$ and $c = 1$, and when $n = 100$ and $c = 7$ (approx.). In acceptance/rejection sampling an additional criterion, for choosing one among the many schemes that satisfy the customer's requirements, is provided by the producer's safe point (or AQL). In acceptance/rectification inspection this has a less important meaning, since no lots are rejected. The

alternative additional criterion, used by Dodge and Romig in their *Sampling Inspection Tables*, is the minimum ASN. For each combination of n and c satisfying given consumer's requirements there is a different ASN-p' curve. In the Dodge-Romig schemes the value of p' in the general run of lots is presumed to be known within fairly broad limits (multiplied by 100 it is termed the *process average* per cent defective); for each range of p' one combination of n and c satisfying given consumer's requirements has a lower ASN than all the others, and that one is chosen for the scheme. It involves the least total amount of inspection and, assuming that it costs the same per article to inspect a sample as the lot, is the most economical scheme. With this criterion, therefore, it is only necessary to specify the consumer's requirements, which in the Dodge-Romig tables are described as the lot tolerance per cent defective corresponding to a risk of 0.10.

Another new quantity characteristic of acceptance/rectification schemes provides a criterion alternative to the consumer's safe point— the *average outgoing quality* (the AOQ). This is the average fraction (or percentage) of defectives in lots passed forward, for a given fraction p' in lots presented. Some of the lots will be passed forward without change as a result of the first sample inspection and will have a fraction p' defectives; other lots will have been subjected to full inspection and rectification and will have zero defectives; the AOQ will therefore be something less than p' (or, as a percentage, less than $100\ p'$).

As a simple example let us suppose that the sample size is n, that $c = 1$, that the number in the lot is large compared with n, and that all lots presented have a fraction of defectives equal to p'. Then according to equations (9) the proportion of lots passed forward without change is

$$(1 - p')^n + np'(1 - p')^{n-1}$$

and the average fraction defective in the lots passed forward is p' times this proportion plus zero times the proportion of lots rectified; that is,

$$\text{AOQ} = p'(1 - p')^n + np'^2(1 - p')^{n-1} \tag{10}$$

when $c = 1$. Corresponding equations may be deduced for other values of c. The curve of the AOQ plotted against p' is for a rectifying scheme what the OC curve is for an acceptance/rejection scheme. Such

curves for $n = 20$, $c = 1$, and for $n = 100$, $c = 1, 5$, and 10 are plotted in Fig. 14.

The line $AOQ = p'$ represents what would happen if there were no inspection; all the curves fall below it. For any given value of n and c, the AOQ curve follows the line at first, since very few lots are subjected to full inspection and rectification. As p' increases, the propor-
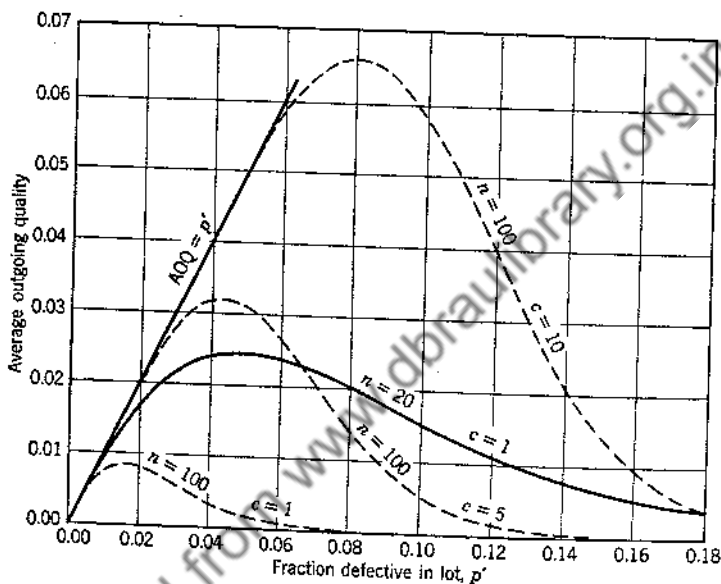
FIG. 14.

tion of rectified lots increases, and so the curve falls farther below the $AOQ = p'$ line. At first the effect of the reduction in defectives due to rectification is less than that of the increase in p', and the AOQ value rises; but after a certain value of p' the effect of rectification predominates, and the AOQ falls. For any given n and c, therefore, there is a maximum value of the AOQ beyond which the average fraction of defectives passed forward can not rise, whatever the fraction of defectives in the lots as presented. This is the *average outgoing quality limit*, or the AOQL, and is sometimes preferred to the lot tolerance as an assessment of the long-run quality of production.

Again, for approximately the same AOQL, there are many combinations of n and c (e.g., $n = 20$, $c = 1$ and $n = 100$, $c = 4$ have AOQL $= 0.025$ approx.), each having a different ASN-p' curve; and

for each range of p' that combination having the lowest ASN defines the chosen scheme.

Thus in order to use the Dodge-Romig single sampling inspection tables it is necessary to know: the process average per cent defectives and the number in the lot; it is necessary to choose (on technical grounds): the lot tolerance per cent defectives corresponding to a consumer's risk of 0.10, or the average outgoing quality limit. The tables then give the values of n and c that minimise the total amount of inspection. Moreover, if the lot tolerance is chosen, the corresponding AOQL is given, and vice-versa.

Engineers, especially Dodge and Romig, have, by their work, added a new view-point to statistical sampling. The classical view regards each sample separately as belonging to a separate lot or population; Dodge and Romig have taught us to think of a sampling procedure as a kind of screen through which factory production is continuously passed and which modifies the average quality.

Multiple and Sequential Sampling

The next stage in complication after single sampling is double sampling. According to this procedure, as a result of the inspection of a first sample the lot may be accepted or rejected without further test, or a second sample may be taken on the result of which the lot is accepted or rejected. If the scheme is an acceptance/rectification one, full inspection with rectification is substituted for rejection.

When the quality of the lot is specified by the mean of some measured quantity, such as the mean specific gravity of bricks, a low value of which is required, the lot is rejected if the sample mean is above a certain value L_1, say, or accepted if it is below another, lower, value L_2; or if the sample mean is between L_1 and L_2 a second sample is taken, and the lot is rejected if the combined mean for the two samples is above a new value L_3, say, or accepted if it is below L_3. For a given value of the lot mean the probability of ultimate acceptance by such a scheme can be calculated, and hence an OC curve be deduced, and in particular consumer's and producer's safe points. Curves of the ASN can also be deduced, since some lots will be appraised on the first sample of n_1, say, whereas others will require two samples of n_1 and n_2. The chief advantage of such over a single sampling scheme is a reduction in the average sample size for given consumer's and producer's safe points. This happens because a decision can be taken on one sample if the lot is either very good or very bad; the second sample is taken only from those lots that are moderately good or bad and

F

require more careful discrimination. The scheme to choose is that which gives the lowest ASN for the required protection. Since L_1, L_2, L_3, n_1, and n_2 can be adjusted (not independently, and only within limits), the determination of an optimum scheme is a fairly complicated business, and it has not been done extensively or systematically.

Double sampling schemes for the fraction defective have been more used, and the Dodge-Romig tables for acceptance/rectification inspection include such.

The Dodge-Romig double sampling procedure is set out schematically as follows: †

```
┌─────────────────────────────────┐
│ Inspect a first sample of n₁    │
└─────────────────────────────────┘
              ↓
If the number defective pieces in this sample
```

| does not exceed c_1 | exceeds c_1 but does not exceed c_2 | exceeds c_2 |

```
              ┌─────────────────────────────────┐
              │ Inspect a second sample of n₂   │
              │ pieces                          │
              └─────────────────────────────────┘
                            ↓
              If the number of defective pieces
                in the two samples combined
```

| does not exceed c_2 | exceeds c_2 |

```
┌─────────────────────┐      ┌──────────────────────────────────┐
│   Accept the lot    │      │ Inspect all pieces remaining in   │
│                     │      │ the lot and correct or replace all│
│                     │      │ defectives                        │
└─────────────────────┘      └──────────────────────────────────┘
```
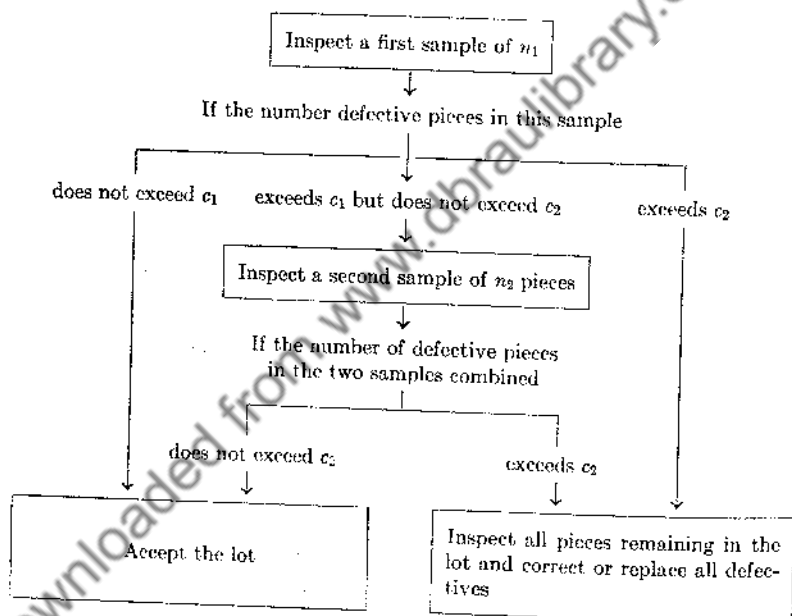
A double sampling scheme has, in addition to its OC curve and lot tolerance, its curve of AOQ, its AOQL, and its ASN curve (the ASN depending on the proportion of lots that are accepted after testing one sample of n_1, accepted after testing two samples of n_1 and n_2, and subjected to full inspection and rectification). The Dodge-Romig tables give for various lot sizes and ranges of process average values of n_1, n_2,

† Reproduced by permission of Bell Telephone Laboratories, Inc., from *Sampling Inspection Tables*, by H. F. Dodge and H. G. Romig, published by John Wiley & Sons, Inc., 1944.

c_1, and c_2 that minimise the amount of inspection and give either various chosen lot tolerances with a risk of 0.1 or various chosen values of

Inspect a sample
of n_1 articles

If the quality

exceeds
L_1

exceeds L'_1 but does
not exceed L_1

does not
exceed L'_1

Inspect a second
sample of n_2 articles

If the mean quality
of the two samples

exceeds
L_2

exceeds L'_2 but does
not exceed L_2

does not
exceed L'_2

Inspect a third
sample of n_3 articles

If the mean quality
of the three samples

exceeds
L_3

exceeds L'_3 but does
not exceed L_3

does not
exceed L'_3

Inspect a fourth
sample of n_4 articles

Etc.

Inspect a mth
sample of n_m articles

If the mean quality
of the m samples

exceeds
L_m

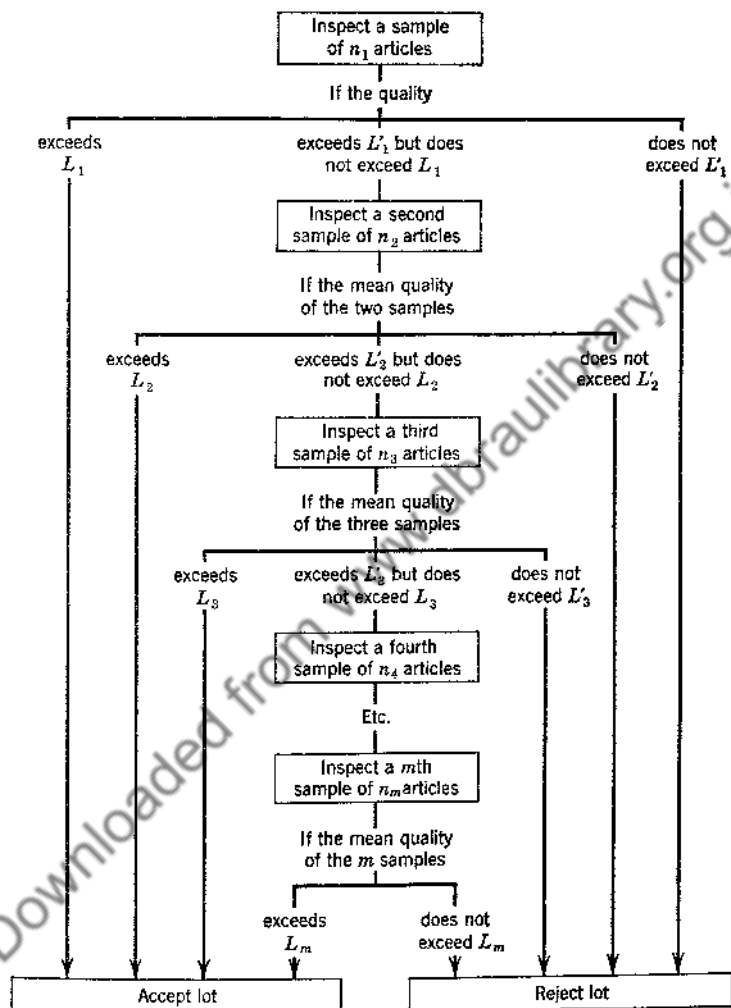does not
exceed L_m

Accept lot

Reject lot

Fig. 15.

the AOQL. Other tables for double sampling schemes are given in the book *Sampling Inspection* referred to in the bibliography. In this book the ASN is the average sample number for acceptance or rejection (not rectification), and account is taken of the reduction in sample size made

possible if the inspection is *curtailed* immediately the acceptance number of defectives is exceeded by unity.

It is easy to imagine an extension of the idea of deferring final judgment on a lot to the adoption of triple, quadruple, or higher multiple procedures, which may even extend to the limiting form in which an assessment is made after inspecting each article. These multiple procedures are termed *sequential sampling schemes*. A typical scheme is set out in Fig. 15; it is closed by forcing a decision to accept or reject after a certain number of samples.

The theory and practice of sequential sampling for acceptance/rejection was developed during World War II. Each scheme has its OC and ASN curve, and optimum schemes have been worked out, together with convenient graphical procedures. They carry practically to the limit the economy achievable in the number of articles inspected for a given degree of control by continuously examining results as they accumulate and coming to a decision at the earliest possible moment.

Some people prefer single sampling schemes on account of their administrative simplicity. Others prefer double sampling schemes partly because of their economy and partly because it is sometimes psychologically satisfying to the practical man to give the lot a second chance. Sequential schemes are considered to require most care and supervision in operation, but, where the inspection or testing costs per article are high, utmost economy in the number of articles inspected is important, and often outweighs administrative convenience.

You should re-read Chapter 4, particularly the section, "Sampling Method," and remember that all that has been said in this chapter about the sampling of articles and pieces applies to all "statistical individuals" as described in Chapter 4, and that all that has been said in Chapter 4 on sampling for quality control applies to sampling for routine inspection.

PART II. INVESTIGATION AND EXPERIMENTATION

Chapter 8. EXPERIMENTATION AND THE STATISTICAL THEORY OF ERRORS

Routine quality control in the factory requires the backing of technical research and investigation in order to discover the technical conditions necessary for the economical production of goods of the required quality, to determine what factors to control and how to control them, and to discover assignable causes of uncontrol. Investigations range from empirical experiments made on the factory floor to fundamental research at the university or in the research institute, but they are all based on the scientific method of varying a few factors in a controlled way and inferring from the results what are the causal effects of the factors.

Such inferences are often impeded by experimental errors. When working under the best conditions in a laboratory it is impracticable or impossible to make the experimental factors vary exactly as planned and keep the constant factors absolutely constant; the results are the effects of a more complicated system of causes than that planned: they include the effects of experimental errors. In the laboratory these errors can often be reduced to negligibly small proportions, when they cause little or no difficulty; it is an important part of the experimenter's art to bring about this state of affairs. But in technical investigation the effects of experimental errors are often comparable in magnitude to those under investigation, and they have to be taken into account systematically. The only known way of achieving this is to apply the statistical theory of errors, described in this chapter; a discussion of the assumptions involved in its practical application is left to the next chapter. You will frequently ask why various steps are taken in developing the theory, but for an answer you should consult a text-book on statistics.

The Principles of Significance Testing

Table VI gives the results of some experiments made by the third Lord Rayleigh to determine the density of nitrogen prepared in various ways; the figures in the body of the table give the weights, in unspeci-

TABLE VI

Weights of Nitrogen from Various Sources

Source	Nitric Oxide	Nitrous Oxide	Ammonium Nitrite	Air	Air
Reduced by:	Iron	Iron	Iron	Iron	Copper
Date	Nov.–Dec. 1893	Dec. 1893	Jan. 1894	Dec. 1893	Aug.–Sept. 1892
	2.29816 2.29890 2.30143 2.30182	2.29869 2.29940	2.29849 2.29889	2.30986 2.31001 2.31010 2.31017	2.31012 2.31024 2.31026 2.31027 2.31035
Mean	2.29947 (a)			2.31004 (b)	2.31025 (c)

fied units, of successive determinations made under standard conditions, of nitrogen in a standard bulb, so that variations in weight are due to variations in density plus the uncontrolled variations we term experimental error. The problem is to decide whether the source affects the weight, and hence the density, of the nitrogen. Rayleigh decided from these and similar results that it did: that "atmospheric" nitrogen had a higher density than "chemical" nitrogen; and thus he was led to the discovery of the rarer gases in the atmosphere. Let us examine the data somewhat as they might have appeared to Rayleigh before he reached this conclusion.

We may regard the first three columns of figures as one series for nitrogen reduced from chemical sources by iron, and those in the next two columns as two series for nitrogen reduced from the atmosphere by iron and copper respectively. We shall refer to these series as (a), (b), and (c). The data enable us to separate the effects of the source and of the reducing agent.

First, by comparing (a) with (b) we see the effect of source alone. The results for (a) range from 2.29816 to 2.30182, those for (b) range from 2.30986 to 2.31017; there is no overlap between the series, and the difference between the means $(2.31004 - 2.29947)$ is much larger than the spread for either series. Commonsense tells us that experimental

errors (which are measured by the spread within each series) could not account for the difference between the two series; that nitrogen from the atmosphere is really heavier than that from chemical sources.

When we compare (b) with (c) we see that the two series overlap (the ranges are 2.30986 — 2.31017 and 2.31012 — 2.31035), and that the difference between the means (2.31004 — 2.31025) is much the same as the spread within each series. A cautious commonsense would tell us that the results give no clear evidence that the reducing agent affects the density of the nitrogen reduced from the air, or a rash commonsense that the reducing agent has no effect.

How does the statistician handle this kind of situation? For Table VI it seems to be hardly necessary to consult the statistician, but the results of technical experiments are not always as clear-cut, and then we may require to use statistical arguments. It will help to make them clear if we apply them to Table VI.

The statistician argues: let us regard the variations within each series (due to experimental errors) as chance variations and test the *hypothesis* that the differences between the two series in each pair are due to these chance variations. This can be done in several ways.

First let us treat series (a) and (b) by writing down the 12 results in order of magnitude and substituting for each value the letter a or b according to the series it belongs to; thus we arrive at the arrangement

$$a \quad a \quad a \quad a \quad a \quad a \quad a \quad b \quad b \quad b \quad b$$

Of all possible arrangements of eight a's and four b's (there are 495 of them) this is a very peculiar one. Only one other arrangement has the same kind of peculiarity—that which gives first the four b's and then the eight a's. We express our recognition of this peculiarity by saying that the probability of chance giving such an arrangement is only $2/495 = 0.004$. This is very low; chance is unlikely to have given the arrangement; some factor other than chance is probably responsible for the distribution of the twelve values between the two series. That is the statistician's verdict, and that is as far as he goes. He does not say what the other factor is.

You may object that each one of the 495 arrangements is in one way or another unique, and that the same argument would establish it as due to something other than chance. Only if the arrangement follows a pattern corresponding to a factor of technical significance do we accept it as possibly due to something other than chance. The above arrangement of a's and b's could be due to a change in density from

series (a) to series (b). Other patterns might have some other technical significance; for example,

$$a \quad b \quad a \quad a \quad b \quad a \quad a \quad b \quad a \quad a \quad b \quad a$$

If an experiment gave this, we might be sceptical of it being due to chance, although it is hard to say what it would be due to.

When we compare series (b) and (c) in the same way we have

$$b \quad b \quad b \quad c \quad b \quad c \quad c \quad c \quad c$$

We need a single figure to characterise this kind of arrangement. For this Miss Swed and Dr. Eisenhart * have proposed the number of runs of the same letter. Thus, in the above there are 4 runs, the three b's counting as one, the one c as another, the one b as a third, and the four c's as a fourth. This criterion satisfies commonsense, since a low number of runs corresponds to a tendency for the series to separate as they would do if the experimentally imposed change in conditions affected the density of the nitrogen. The lowest possible number of runs is 2; Swed and Eisenhart have given probability tables for the various numbers of runs, and from these we find that the probability, for four b's and five c's, of chance giving 4 runs or fewer is 0.262. This is not low enough for us to dismiss chance as the possible factor. You will note that we say "*4 runs or fewer*," which is equivalent to saying "as great or a greater degree of differentiation."

When a statistician subjects his chance hypothesis to the kind of trial described, he is said to test the *statistical significance* of the results. If the verdict goes against chance, the difference between the two series is said to be *statistically significant;* otherwise it is statistically *not significant*. As a convenient shorthand the words significant and not significant are often used without the adverb, but the adverb should not be forgotten, for there is an important difference between statistical and technical significance, as we shall see.

The number of runs does not provide an entirely satisfactory criterion of the differentiation between the two series, (b) and (c). There are several arrangements with 4 runs, but they are all treated as of the same class, whereas some of them correspond to a greater differentiation than others. Compare, for example,

and

$$b \quad b \quad b \quad c \quad b \quad c \quad c \quad c \quad c$$

$$b \quad b \quad c \quad c \quad c \quad b \quad b \quad c \quad c$$

* *Annals of Mathematical Statistics*, Vol. XIV, 1943, p. 66.

The kind of effect we are investigating experimentally would be expected to affect most markedly the difference in means, which therefore seems to be a good criterion of differentiation. Of the two arrangements of eight a's and four b's that provide complete separation, namely,

$$a \quad a \quad a \quad a \quad a \quad a \quad a \quad a \quad b \quad b \quad b \quad b$$

$$b \quad b \quad b \quad b \quad a \quad a \quad a \quad a \quad a \quad a \quad a \quad a$$

the first when applied to the 12 results of Table VI arranged in order gives the larger difference between the means for series (a) and (b). This is the difference corresponding to the actual arrangement in Table VI, and the probability of chance giving as large a difference is therefore $1 \div 495 = 0.002$. There are 126 possible arrangements of the 9 values of series (b) and (c) into two series of 4 and 5 respectively, and the 3 leading to the greatest difference between means, together with the means, are

$$b \quad b \quad b \quad b \quad c \quad c \quad c \quad c \quad c \quad (2.31026\text{--}2.31002)$$

$$c \quad c \quad c \quad c \quad c \quad b \quad b \quad b \quad b \quad (2.31028\text{--}2.31005)$$

$$b \quad b \quad b \quad c \quad b \quad c \quad c \quad c \quad c \quad (2.31025\text{--}2.31007)$$

The last arrangement results from the actual experiment and is therefore special in that it is among the three chance arrangements producing the greatest differentiation. The probability according to this test of chance producing the degree of differentiation between (b) and (c) of Table VI, or a greater degree, is $3 \div 126 = 0.024$. This is low, and we now suspect that chance may not be a sufficient explanation of the difference between series (b) and (c).

This test based on means is more discriminatory than that based merely on runs because it narrows the class of arrangements that are regarded as having a possible technical significance through using more of the information-provided by the results; it uses the actual values. It leads us to suspect a difference between series (b) and (c) that the run-test dismissed as easily attributable to chance. Even so, it is not entirely satisfactory. For series (a) and (b) it can never give a probability lower than 0.002; any two sets of a given number of results that fail to overlap have the same degree of statistical significance on this test, however widely they may be separated. This does not satisfy commonsense. A more satisfactory test is provided by the so-called t test.

The *t* Test of Significance

According to this test the results for each series are regarded as a random sample of an infinite population of results that could have been obtained had the experiment been repeated indefinitely under the same controls that were applied from November, 1893, to January, 1894, to obtain the results of Table VI. This population is purely a concept; nothing even remotely approaching it can be realised physically owing to the limitations of human powers and physical materials; and a set of experimental results can only be likened to a statistical sample by a stretch of the imagination. This for some people has been a stumbling-block to the application of statistical theory to physical experiments. I find no such difficulty. The application has its difficulties, but they are not conceptual. Experience has led to the elaboration of the theory to cover the kind of situation that arises in the life of the experimenter and has shown that the guidance given by such elaborated theory is good. You should be able to regard the results of Table VI as a random statistical sample of experimental errors, plus "real" variations resulting from the experimentally imposed variations; and as this text develops you will agree as to the reasonableness of the methods and conclusions based on this assumption as a starting point.

A further assumption, of a mathematical kind, is that the frequency distribution of errors in the infinite population is Normal in form.

Then in comparing series (*a*) and (*b*), say, we tentatively adopt the hypothesis that they are random samples from the same population and calculate the probability that chance would give a difference in means as great as or greater than that observed. This test is similar to that used above, but it is based on a different model—on samples from an infinite population rather than on arrangements of a finite number of results.

Suppose that we take at random a pair of samples of n_1 and n_2 individuals from a population in which the "true" standard deviation is σ', and that the corresponding sample means are $\bar{X}_1$ and $\bar{X}_2$, giving a difference in means of $d = \bar{X}_1 - \bar{X}_2$. If we repeat such a sampling experiment many times we get many values of d which may be formed into a frequency distribution, the sampling distribution of the difference between two means. Theory states, and experience amply verifies, that under our assumptions this sampling distribution is Normal with a mean value equal to zero and a standard error given by the formula

$$\text{S.E.}'_d = \sigma' \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{11}$$

This is the equivalent of the formula for the standard error of a single mean [equation (2), p. 17]. Thus, if σ' is known, the probability distribution of d is the same as that of

$$t = \frac{d}{\text{S.E.}'_d} \tag{12}$$

tabulated briefly in Table III (p. 8). Usually there is no prior reason for supposing $\overline{X}_1$ to be greater or less than $\overline{X}_2$, and a large positive value of d (and hence of t) is equally significant with an equal large negative value. The probability of d or t exceeding any given value, positive or negative, is 2α, and this is used mostly in testing significances. For example, the probability of d exceeding twice its standard error (of t being greater than 2) is approximately 0.05.

Usually, however, we do not know σ', and so we use the best estimate obtainable from the sample, which is †

$$\sigma = \sqrt{\frac{\Sigma_1(X - \overline{X}_1)^2 + \Sigma_2(X - \overline{X}_2)^2}{n_1 + n_2 - 2}} \tag{13}$$

where X represents successively the individual values

$\overline{X}_1$ and $\overline{X}_2$ are respectively the means of the two series

Σ_1 means "sum over all the observations in series 1"

Σ_2 has a corresponding meaning for series 2

n_1 and n_2 are respectively the numbers of observations in the two series.

Equation (13) corresponds to equation (1) (p. 7) except that in (13) there are the sums of two sets of squares in the numerator, and in the denominator appears the number of observations reduced by 2. The latter quantity is termed the number of *degrees of freedom*, and its use instead of the number of observations results from a refinement of statistical theory that can not be dealt with here and that is important only when $n_1 + n_2$ is small.

The estimate of the standard error of the difference now becomes

$$\text{S.E.}_d = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{14}$$

† I am adhering to the symbolism of the ASTM standards, distinguishing between "true" or population values and sample estimates by adding a prime for the former. In most statistical literature the population value is represented by a Greek letter (e.g., σ) and the sample estimate by the corresponding Latin letter (s).

(note that n_1 and n_2 appear here, not the degrees of freedom), and becomes

$$t = \frac{d}{\text{S.E.}_d} \tag{15}$$

In applying the test, sample 1 is usually taken as that with the greater mean, so that t is usually positive.

There is a different probability table for t as defined by equation (15) according to the degrees of freedom on which the standard deviation is estimated, Table III being merely the special case for an infinite number of degrees (which would be necessary to make the sample estimate σ equal to the "true" value of σ'). Most tables give values of t corresponding to a few chosen probabilities after the manner of the lower half of Table III (except that $2\alpha = P$, say, is used). Generally any difference that chance would cause to be exceeded with a probability of P is said to lie on the P *level of significance;* and P may be quoted as a decimal fraction or a percentage. A large difference between two means relative to its standard error gives a low value of P, and a low value of P is said to correspond to a high level of significance.

The t test is satisfying partly because it corresponds closely to the commonsense procedure of assessing the difference between the means in relation to the variation within each series. Let us now apply it to Table VI. The sums of squares of deviations from the respective means $(\Sigma_1(X - \overline{X}_1)^2$, etc.)—ignore the decimal points in Table VI—are

Series (a)	133,152
Series (b)	538
Series (c)	275

For comparing series (a) and (b),

$$\sigma = \sqrt{\frac{133,690}{10}}$$

$$\text{S.E.}_d = \sqrt{\frac{133,690}{10}\left(\frac{1}{8} + \frac{1}{4}\right)} = 70.8$$

$$t = \frac{1057}{70.8} = 14.9$$

and there are 10 degrees of freedom. (Note that t is a dimensionless ratio, so that we need make no correction for the change in units resulting from ignoring the decimal points.) According to the tables, the

value of t that is exceeded by chance with a probability of 0.01 is 3.17; the above value is much greater than this, and so it is highly significant. It is difficult to calculate the probability accurately from the published tables, but it is certainly less than 0.0003, so that the difference between series (a) and (b) is much more highly significant on this test than it could possibly be on the other two.

For comparing series (b) and (c),

$$\sigma = \sqrt{\frac{813}{7}}$$

$$\text{S.E.}_d = \sqrt{\frac{813}{7}\left(\frac{1}{4} + \frac{1}{5}\right)} = 7.23$$

$$t = \frac{21}{7.23} = 2.91$$

and there are 7 degrees of freedom. From the tables it is calculated that the corresponding probability is approximately 0.022, which is near the value given by the previous test made by considering the differences in means given by various arrangements of the two series of results.

The probability calculated in applying the t test is usually the sum of the two "tails" giving the probabilities for positive and negative deviations because, as stated above, there is usually no *a priori* reason for supposing one mean to be greater than another. When there is such a reason—for example, it might be theoretically inconceivable that the true mean for series (b) in Table VI could be less than that for series (a)—the probability is that given by one tail.

In applying the t test, we use a pooled estimate of σ' obtained from the two series under test. It would be equally reasonable to regard all the results of Table VI as subject to the same experimental errors and to use as a pooled estimate of the standard deviation

$$\sigma = \sqrt{\frac{133,152 + 538 + 275}{8 + 4 + 5 - 3}} = \sqrt{\frac{133,965}{14}}$$

the degrees of freedom in the denominator being 3 less than the number of observations because the deviations are measured from 3 sample means (consult the text-books to discover why). Then, for comparing (b) and (c),

$$\text{S.E.}_d = \sqrt{\frac{133,965}{14}\left(\frac{1}{4} + \frac{1}{5}\right)} = 65.6$$

$$t = \frac{21}{65.6} = 0.32$$

and there are 14 degrees of freedom. From the tables we find that P is between 0.7 and 0.8, and we have lost in significance because the large contribution to the sum of squares of series (a) has increased S.E.$_d$. Even so, *provided that the pooling of all results to estimate σ is justified,* the last procedure gives the better test, since it uses all the available information, and it is more likely in the long run to show up a real effect, whatever it does in one particular case. But the *proviso* is important and is scarcely satisfied here. Mere inspection of Table VI shows that the results of series (a) are much more variable than those of series (b) and (c), and the errors of the experiments made with nitric oxide seem to have been specially large. It is not justifiable, on the face of it, to pool these with the errors of the other series. There are statistical tests for deciding whether pooling is justified, but they will not be discussed here. If the pooling of the errors of series (a) with those of (b) and (c) is unjustified, so is the pooling with those of series (b) for the ordinary t test of the difference between the two series (a) and (b). This will be referred to later.

Note that it would be quite wrong to test the significance of the difference between series (b) and (c) first with and then without the pooling of the errors of all three series, and then to choose the result which gives the lower probability. The decision to pool or not to pool must be taken without reference to its effect on the test of significance in the particular instance; it should preferably be taken before the t test is applied.

Tests of significance, of which the t test is typical, bulk unduly large in the theory of statistics as applied to experimentation. They merely inform the experimenter whether or not his results tell him anything about the subject under investigation, and this does not carry him far. In the early days of applied statistics, experimenters were likely to think that their results signified more than they did, and different experimenters obtained results that were discordant because of insufficiently appreciated errors. In those circumstances the statistician did a service in calling attention to the errors and providing tests of significance. The experimenter who is past this elementary stage in methodological evolution wants to know something more. He wants

to know *what* his results tell him.   In particular, if the difference between two means is statistically significant, he wants to know the *magnitude* of the difference and the *precision* with which that magnitude is estimated.   If it is not statistically significant he wants to know how large the difference could be and yet elude significance.   And in general he wants to know how to design an experiment so as to attain a given precision and make a given difference (chosen as the smallest that is technically important) statistically significant.   The discussion of these points is begun in the next section.

The Precision of Estimates—Confidence or Fiducial Limits

In this section a new illustration is considered: the measurement of the "standard fibre weight" of cotton.‡   The method of sampling and testing has been carefully standardised so that the errors of the determination have been stabilised, and it has been found from replicate tests made on a large number of cottons that the standard deviation of a single determination is 8.5 units; the experience on which this estimate is based is so large that we may regard this as the "true" value, σ' Now let us suppose that as a routine three determinations are made for each cotton, and that pairs of cottons are compared by the corresponding means.   Then from equation (11) the standard error of the difference d is $8.5 \sqrt{\frac{1}{3} + \frac{1}{3}} = 6.94$ units.

Now suppose that we have two particular cottons for which the true difference is d', and that we make many sets of three determinations from each, each pair of sets producing a value of the difference d between sample means.   Then the values of d will be distributed Normally about a grand mean value equal to d' with a standard deviation (or standard error) of 6.94 units, and we see from Table III that 95 per cent of the values of d will lie between $d = d' - 1.96 \times 6.94$ and $d = d' + 1.96 \times 6.94$ (i.e., within the limits $d = d' \pm 13.6$ units).

If, for example, d' for the two cottons is 14 units, the values of d will be spread along the dotted line drawn through $d' = 14$ in Fig. 16, the centre being at $d = 14$ and 95 per cent of the values being between $d = 0.4$ and $d = 27.6$ units.   Another pair of cottons might have a "true" difference of $d' = -10$, say, and the values of d would be spread along the corresponding dotted line in Fig. 16, with a centre or grand mean value at $d = -10$, and 95 per cent of the values between -10 ± 13.6.   Generally, for any pair of cottons (i.e., for any given

‡ The measure is described in the *Shirley Institute Memoirs*, Vol. **17**, 1939, p. 25, or *Journal of the Textile Institute*, Vol. 30, 1939, p. T173.

d'), the centre of the distribution will lie on the line $d = d'$, and 95 per cent of the values will lie between the limits at which the vertical line drawn through the value of d' cuts the two lines AB and CD in Fig. 16. It follows that if we test many pairs of cottons with different values of d', 95 per cent of the values of d will lie in the belt contained between the lines AB and CD. And this is true whatever the relative fre-
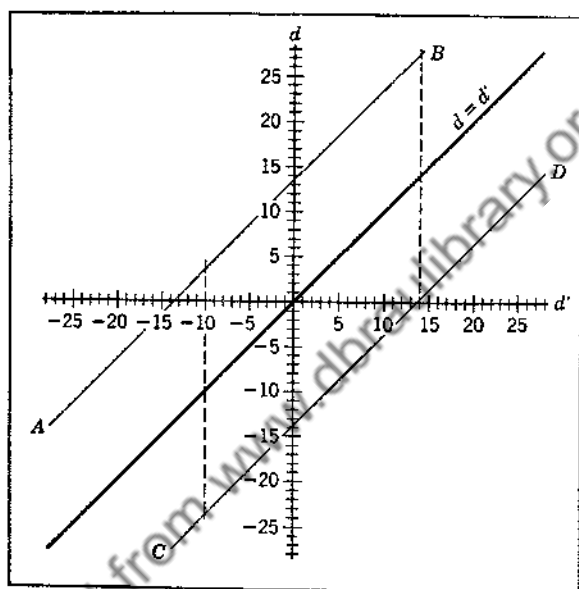
Fig. 16.

quencies of the values of d' and whatever the number of values of d for each pair of cottons.

The foregoing argument states something about d if the true difference d' is known. But in practice we do not know d'; we have one value of d and want to say something about d'. If 95 per cent of the values of (d, d') lie between the lines AB and CD, then for any given value of d 95 per cent of the points spread along the corresponding horizontal line in Fig. 16 will lie between the limits at which this line cuts AB and CD, and it is easy to see from the geometry of Fig. 16 that these limits are at $d \pm 13.6$. In other words, for a given value of d 95 per cent of the values of d' will lie within the limits $d \pm 13.6$; and, if for each determination of d for a pair of cottons we infer that the

true value d' lies between these limits, we shall be right in 95 per cent of the cases and wrong in 5 per cent.

The figure 95 per cent is a measure of our confidence that the true value d' lies within the stated limits as determined from the sample value of d. This has been termed by Dr. Neyman the *confidence co-efficient* and by Professor Fisher the *fiducial probability*. The corresponding limits are termed the 95 per cent (or 0.95) *confidence* or *fiducial limits*, and the belt in Fig. 16 between AB and CD is the 95 per cent *confidence belt*. Dr. Neyman's and Professor Fisher's theories have a different mathematical and logical basis and deal differently with more complex situations. It is only in simple situations such as that introduced here that they can be regarded as practically equivalent. The confidence coefficient and fiducial probability are superficially very like ordinary statistical probability, but they are not quite the same and all three quantities have to be handled differently in theory. However, any experimenter who is prepared to accept and think of them in the same way will not go very *far astray* in practical life.

Figure 16 was derived by combining experiences with many cottons, but, if for every experience in all fields we correctly calculate the 95 per cent confidence limits and infer that the true value lies within these limits, we shall tend in the long run to be right 95 and wrong 5 times in 100 such inferences. Limits corresponding to other confidence coefficients can be calculated (e.g., the 99 per cent confidence limits for a Normal sampling distribution are at $d \pm 2.58 \mathrm{S.E.'}$). Generally we arrive at the obvious conclusion that the more widely the limits are spaced the more confidence have we that the true value lies within them.

Now, for example, suppose that we have tested a pair of cottons and find that $d = 6$; then we have a 95 per cent confidence that d' lies somewhere between 6 ± 13.6 (i.e., between -7.6 and 19.6), but we are not prepared to say where. The value $d' = 0$ lies within the limits, and so we arrive in another way at the conclusion that such a difference is not statistically significant. On the other hand the difference might be as large as 19.6 units.

We can make the confidence belt narrower, and hence increase our powers of discriminating between cottons, by increasing the size of the sample. The size of sample necessary for any specified width of band can easily be calculated. We may be technically interested in differences between cottons in standard fibre weight of 8 units or greater,

G

and this may be the half-width of the 95 per cent confidence belt. Then, if $n_1 = n_2 = n$ is the number of tests per cotton,

$$1.96 \times S.E.' = 8$$

$$S.E.' = 8.5 \sqrt{\frac{2}{n}}$$

and

$$n = 8.66$$

There would be 9 tests per cotton.

In all the foregoing a knowledge of the true standard deviation σ' has been assumed; it is because of this that a Normal distribution for the means may be reasonably assumed, and only for that distribution do the factors 1.96 and 2.58 correspond to the 95 and 99 per cent confidence limits. Results based on these assumptions are quite useful in industrial practice, for quite often one has from long experience a good idea of the value of σ'.

Sometimes, however, such knowledge is lacking, and all that is available is an estimate σ based on a few degrees of freedom. Then the sampling distribution of $d/S.E.$ is not Normal; it is the distribution of the t used in the t test of significance, and the factor for the chosen confidence limits must be taken from the appropriate tables.

Thus for 7 degrees of freedom the factor for the 99 per cent confidence limits is 3.50 (not 2.58 as it would be if σ' and S.E.' were known). If we apply this to our test of the difference in means between series (b) and (c) in Table VI, we find that the confidence limits are at

$$0.00001(21 \pm 3.50 \times 7.23) = -0.00004 \text{ and } +0.00046$$

Thus, if we space limits to give a fairly high degree of confidence, we infer that the difference between series (b) and (c) could be zero, although the mean for (c) could exceed that for (b) by as much as 0.00046 of the units of weight. For 10 degrees of freedom the corresponding t factor is 3.17, and the same calculation applied to the comparison between (a) and (b) leads to the limits

$$0.00001(1057 \pm 3.17 \times 70.8) = 0.00833 \text{ and } 0.01281$$

These limits specify the precision with which the difference in mean weights is estimated, or they would were it not that the excessive variability of series (a) casts doubt on the procedure of making a pooled estimate of the standard deviation.

When σ' is not known and only an estimate σ is available, the decision as to the number of observations necessary for a given precision or width of confidence belt is not easy. However, it is not important to determine the number at all accurately. If σ is estimated from a limited number of degrees of freedom, an upper confidence limit for σ' can be determined (refer to the text-books for this), and, if this limit is used in the way illustrated here, the calculated number of observations will give a sufficient guide for practical purposes.

In Fig. 16 the confidence belt is symmetrically placed about the line $d' = d$. When errors can be positive or negative and the sampling distribution is symmetrical, no other disposition seems reasonable. When there is asymmetry in the basic conditions, the problem of placing the confidence belt is more difficult.

Errors of a Single Mean

So far the statistical theory of errors has been described with reference to the difference between the means of two series of results; but the same methods apply to the mean of a single series. The results in a series are regarded as a random sample from a hypothetical infinite population of results, the sample mean $\overline{X}$, being an estimate of the true or population mean, $\overline{X}'$. The sampling distribution of $\overline{X}$ is Normal with a grand mean at $\overline{X}'$ and a standard error as given by equation (2). We may test whether $\overline{X}$ differs significantly from some hypothetical true value, or calculate confidence limits for $\overline{X}'$, given $\overline{X}$. If the standard deviation σ is estimated from a limited number of degrees of freedom, $n - 1$ must be used in place of n in equation (1) and the distribution of t must be used instead of the Normal distribution.

Let us take for example the series (a) of Table VI. The mean weight is 2.29947 and its standard error

$$0.00001 \sqrt{\frac{133,152}{7 \times 8}} = 0.00049$$

based on 7 degrees of freedom. Suppose that, from the known molecular weight of nitrogen, the dimensions of the bulb, and the temperature and other conditions, it were possible to calculate the theoretical weight, say 2.30000 for the sake of argument. Then we can use the t test to determine whether the experimental result differs significantly from the theoretical. The difference between the two is 0.00053, t is $0.00053 \div 0.00049 = 1.08$; and for 7 degrees the probability of chance causing t to exceed this is between 0.3 and 0.4. This is fairly high, and

the difference could be due to the chance effect of experimental errors.

In fact we have no data on which to make a theoretical calculation; let us calculate within what limits the true value might lie with a 99 per cent confidence coefficient. These are above and below the experimental mean by 3.50 times the standard error based on 7 degrees of freedom, and are therefore at

$$2.29947 \pm 3.50 \times 0.00049 = 2.29775 \text{ and } 2.30119$$

The inference that the true weight probably lies between these limits depends on a number of assumptions that are discussed in the next chapter, and that are of dubious validity when applied to situations of this kind.

A special case occurs when there are two series of results, the members of which can be associated in pairs so that each pair provides an independent measure of difference. Table VII gives an example. The

TABLE VII [1]

PERCENTAGE OF FAT IN DIFFERENT SAMPLES OF MEAT ESTIMATED BY STANDARD AOAC AND MODIFIED BABCOCK METHODS

AOAC Method	Babcock Method	Difference	AOAC Method	Babcock Method	Difference
22.0	22.3	0.3	26.0	26.3	0.3
22.1	21.8	−0.3	26.2	24.9	−1.3
22.1	22.4	0.3	27.0	26.9	−0.1
22.2	22.5	0.3	27.3	28.4	1.1
24.6	24.9	0.3	27.7	27.1	−0.6
25.3	25.6	0.3	41.5	41.4	−0.1
25.3	25.8	0.5	41.6	41.4	−0.2
25.6	26.2	0.6	45.5	45.5	0.0
25.6	26.1	0.5	48.5	48.2	−0.3
25.9	26.7	0.8	49.1	47.5	−1.6

[1] Taken from Dr. W. J. Youden, *Analytical Chemistry*, Vol. XIX, 1947, p. 946.

object was to compare two analytical methods for determining the fat in meat, and 20 pairs of determinations were made on 20 specimens of meat. The investigation could have been made by choosing 20 specimens at random for the AOAC method, and 20 different specimens for the Babcock method, using the *t* test as described in previous sections

for the significance of the difference between the two means. Had this been done, the standard error would have been as estimated from the sets of values in columns 1, 2, 4, and 5 of Table VII, and we would have

Sum of squares of deviations from mean:

AOAC series	1644.61
Babcock series	1541.85
Total	3186.46
Standard deviation σ	9.16
Standard error of difference	$9.16\sqrt{\frac{1}{20} + \frac{1}{20}} = 2.89$ units

In this instance such a procedure would be most wasteful of effort. The meats vary in fat content very widely, and the standard error of the difference between the means would be partly due to the unequal representation of the different meats in the two samples produced by the random selection. This contribution to error can be eliminated by ensuring that the same meats are in the two samples and using the two methods of testing in parallel, as was in fact done for Table VII.

Each pair of results gives a difference, and the differences are very much less variable than the separate analytical results, since they are unaffected by the differences between the meats. These differences are due to: (1) errors in the individual results arising from experimental errors and the fact that the two tests can not be done on exactly the same morsel of meat; (2) a possible consistent difference due to the two methods; and (3) possible real differences between the results of the two tests that vary from meat to meat (e.g., the difference may tend to be large when the fat content is high and low when the fat content is low). For the time we shall ignore the possibility of (3) and test whether the difference (2) exists.

In order to do this we treat the differences as a single series and adopt the hypothesis that they are a random sample from a population whose true mean is zero. The mean difference is 0.045, the sum of squares of deviations from this mean is 8.0595, the estimate of standard deviation based on 19 degrees of freedom (one less than the number of differences because the deviations are measured from one mean difference) is $\sqrt{8.0595/19}$, the estimate of the standard error is $\sqrt{8.0595/(19 \times 20)} = 0.146$, and t is $0.045/0.146 = 0.31$. Such a low value of t is exceeded by chance with a fairly high probability, and if our assumptions are correct there is no reason from these data for supposing that the two methods give different results.

It will be noted that the result of comparing the two analytical methods on the same meats is to reduce the standard error of the mean difference from 2.89 to 0.146 without increasing the number of tests, and so to give a greatly improved power of discrimination. It is true that the degrees of freedom are reduced from 38 to 19, but this is a mere bagatelle compared with the reduction in standard error. This illustrates the gain in precision and economy that may result if an experiment can be arranged to give comparisons between similar individuals or specimens in pairs, so that differences between the pairs do not affect the comparisons. The gain depends on the extent of the eliminated variation between pairs. We shall discuss this further in Chapter 13.

Chapter 9. PRACTICAL APPLICATION OF THE STATISTICAL THEORY OF ERRORS

The preceding chapter describes statistical theory without much consideration of the practical background. Assumptions are stated and statistical significances are tested with little consideration of the practical implications of the assumptions and the technical interpretation of the results of the tests. This chapter is intended to repair the omission.

First let us note that the errors of which the t test takes account are technically very complicated. In data such as the nitrogen determinations of Table VI they are purely experimental errors as ordinarily understood (errors due to small departures from perfect temperature and pressure control or correction), errors (personal, temporal, and random) in reading instruments, and so on. To such may be added in some experiments real variations in the material under investigation, as in the example of fat analysis of Table VII. Here if the meats were chosen at random for the two methods of analysis and the ordinary two-mean t test for significance performed, the large real variation between meats would contribute to the errors. But experimental errors alone account for the variations in the differences of the third and sixth columns of Table VII. The statistical analysis makes no distinction for the types of error and variation; it treats them all alike. Any differentiation is a matter of technical interpretation.

Choice of Significance Level

In testing the significance of a difference we calculate a probability that the difference could be due to chance, but we can not act in a probable way; action must be definite. Accordingly, if the probability is below a certain level, we come to the verdict "statistically significant" and act as though the difference was real. In doing this we may err in one of two ways: (1) the verdict may be "significant" although the true difference is zero, or (2) the verdict may be "not significant" although there may be a real difference.

The risk of the first kind of error can be made as low as we please by choosing a sufficiently low probability P as the level of significance;

93

for then of the verdicts made when the true difference is zero a fraction $1 - P$ will be correct and only P will be wrong. A reduction in P is achieved by increasing the value of t necessary for significance.

Unfortunately, the lower the value of P, the greater is the risk of the second kind of error: the error of overlooking a real difference. The only way of reducing this risk for a given significance level is to reduce the standard error of the difference, by either reducing σ' or increasing n_1 and n_2; how to calculate the values of n_1 and n_2 necessary to show up given differences as significant has already been shown (p. 88).

For given samples, however, what should be the significance level? The short answer is: it all depends on the circumstances.

Often a good deal is known about the field of investigation; the hypothesis is well founded and would only be discarded on very strong evidence. Then a very low probability level is appropriate. Sometimes the experiment is made to investigate the existence of some new effect, as when Rayleigh measured the density of nitrogen from the air and from chemical sources. Then we usually adopt a "hard-boiled" attitude, acknowledging the existence of the new effect only when it is well established. (Occam's principle that entities should not be multiplied unduly is at the very roots of our scientific thinking.) Again we use a low probability level for significance. In an experiment like Rayleigh's, where much hangs on the correctness of the conclusions, we might adopt a level of $P = 0.001$. In many practical investigations we are willing to be somewhat more venturesome, and any result that gives a value of P below 0.01 is usually regarded as highly significant, and a result with $P = 0.05$ as just significant. There is a certain amount of arbitrariness in the choice of these levels, but they are also justified for general use by long experience.

In industrial life, however, we often meet circumstances in which less stringency is desirable. Mr. A. W. Swan * mentions the choice between two suppliers, A and B, of grinding wheels, on the basis of life and cost tests done on a few wheels and a t test of the significance of the difference in the mean of a "cost index" which incorporates life and cost. If there are no other reasons for preferring A to B, the one with the lower mean cost index on test would be chosen, however small the difference; the probability level for significance would be 0.5. On the other hand, A may be the present supplier, and a change to B would involve some trouble that would be justified only if there were a fairly strong expectation that B's wheels were preferable; the assurance would not need to be high, and a value of P as low as 0.2 or 0.1 might suffice

* *Journal of the Iron and Steel Institute,* Sept. 1948.

for significance. If the change from A to B involved some expensive change in equipment, assurance of superiority would need to be stronger and a probability level of 0.05 or lower might be required.

I was once confronted with some clinical data collected at a hospital with great trouble over many years, which seemed to show an interesting and important medical effect. The level of significance was only about 0.2, and the effect would usually be dismissed as not significant; my opinion was asked. I had been used to regard effects with a P greater than 0.05 as not significant, but I hesitated. It seemed wrong to ignore the evidence entirely and dismiss the discovery ruthlessly as a mare's nest. The effect was far from being established, but the indications were strong enough to be followed up by further investigations and observations.

The issues that arise in choosing levels of significance and control limits for control charts are similar; and it is important that we should not in all circumstances adhere rigidly to conventional values. Nevertheless it is inevitable that investigators will tend to adopt one or two probability levels, and in the absence of other considerations the conventional low levels of $P = 0.05$ and 0.01 will be found most suitable. An experiment is usually so constructed that, if the chance hypothesis survives, the results are in accordance with existing knowledge and ideas. The adoption of a low probability level puts a premium on the maintenance of the *status quo* in knowledge and retards the admission of new knowledge; but this is usually good. The hypothesis that survives a test through the verdict "not significant" survives only to be tested another day; hypotheses are not thus established irrevocably. On the other hand, when we discard a hypothesis we tend to do so finally; and this should be done only with great care. Progress towards new knowledge by experiments subject to error is like progress in a car towards a destination, through a fog. The faster the car goes, the more likely is it to take the wrong turning or to come to grief through bumping into an obstacle; the slower it goes, the safer is the journey but the longer it takes. A low probability level of significance corresponds to a slow speed of the car.

Choice of Hypothesis

In testing the significance of the difference in the series of determinations of the weights of nitrogen (Table VI) we postulated the hypothesis that the true difference is zero. This choice was made in accordance with the usual scientific method of regarding the data as telling nothing about the matter under investigation until the contrary

is proved—the attitude of the "devil's advocate." The choice is made on general scientific or technical, not statistical, grounds.

These considerations do not always lead us to postulate a zero difference. For example, from the (now) known composition of the atmosphere and the molecular weights of the constituents, we can calculate theoretically the density of "atmospheric" nitrogen compared with "chemical" nitrogen and so can calculate a theoretical value, d' say, for the difference in means between series (a) and (b) of Table VI. Then, in order to test the compatibility of Rayleigh's results with this theory, we would test whether the observed difference d is significantly different from d'.

Assumptions and Hypothesis

After the statistical test of the appropriate hypothesis has been made and the significance or non-significance established at the appropriate level comes the question of interpretation. In fact we establish a model consisting of assumptions plus hypothesis, and the test is made of the whole model. The assumptions may be wrong, or the hypothesis, or both; and this must be borne in mind when interpreting results.

The separation of the elements of the model into assumptions and hypothesis is arbitrary from a statistical point of view; it is a practical convenience. Broadly, the assumptions are the things we take for granted (including the accuracy of the arithmetic!); the hypothesis we are willing to hold to doubt is the thing under investigation. As far as possible, the tests are made insensitive to errors in the assumptions, but the possibility of such errors having important effects must be borne in mind.

A verdict of "statistically not significant" is not translated to mean that the effect under investigation does not exist. It means that the observed difference could be due to the chance effects of experimental errors; that the results give no information about the effect. It is doubtful if the experimentalist who was most sceptical of the applicability of statistical ideas to physical experiments would have the effrontery to claim any significance for his results in the face of such a verdict. The verdict casts doubt, and in the scientific world it requires less justification to cast doubt than to make a positive assertion. Nevertheless errors in the assumptions can render insignificant differences that would be significant were the errors removed. No permanent harm is done if this occurs, but there is a waste of effort, so it is worth while to consider the assumptions from this point of view.

When the verdict is "statistically significant," it is very important

to satisfy ourselves that it is not due to errors in the assumptions be-
fore embarking upon technical interpretations. The effects of these
errors will now be discussed.

The Assumption of Normality

In applying the t test we have assumed that the individual results
have a Normal distribution because that gives a Normal sampling
distribution for the mean. But, even if the distribution of the indi-
vidual results is not Normal, that of the means of samples of quite
moderate size (say of 4 or 5) is nearly Normal, and the t test then
gives results that are substantially correct. Only exceptionally are
departures from Normality in the distribution likely to lead to wrong
conclusions from the t test of significance.

The Assumption of Equal Variability

In applying the t test we obtain a pooled estimate of standard error
on the assumption that the two samples estimate the same true stand-
ard deviation, σ'. This is apparently not true of series (a) and (b) in
Table VI; there are statistical tests for deciding whether the difference
between two estimates of standard deviation is significant, for which
you should refer to the text-books. Until the data are further exam-
ined we must bear in mind the possibility that the significant difference
between series (a) and (b) is due to the difference in variability rather
than in means, although such examination would undoubtedly leave
the difference in means as significant.

Significant differences in standard deviation can depress or enhance
the apparent significance of a difference according as the larger or
smaller sample has the larger standard deviation. But the effect is
small if the samples are of equal size—a condition that can usually
be satisfied in planning the experiment.

There are ways of dealing with the data when the assumption of
equal variability is not justified, but they are not acceptable to all
statisticians, and it is beyond the scope of this book to describe them.
(The so-called Fisher-Behrens test is one of these methods.)

The assumption of equal variability does not, of course, arise when
there is only one series of values and one mean, or when the individuals
in one series can be paired to give a single series of differences, as in
Table VII. In Table VII the variability for methods A and B could be
quite different without invalidating the procedure described in the last
section of Chapter 8.

The Assumption of Randomness

There are no tests by which randomness can be established; we can only look for evidences of particular forms of non-randomness and, if they are absent, assume that the results within each series are substantially random.

One form of non-randomness is a trend or pattern within each series. The results in each series of Table VI are obviously given in order of magnitude and not of the time of the determination. But, had the order been also one of time, with the strong trend shown in Table VI, the application of the t test would have been quite unjustified. There are tests for the significance of such patterns, and, when they are discovered, an examination of the technical details of the experiment may disclose some hitherto unsuspected lack of control that can be removed.

A pattern of variation within a series may reduce the apparent significance of a difference between means if it is stable and common to both series; if it is not stable its effect may either enhance or reduce the apparent difference in means. You will be able to appreciate the effects of such patterns if you re-read the first section of Chapter 4, for a significant effect is, from the statistical view-point, closely parallel to lack of control in manufacturing.

Another form of non-randomness occurs when the results in each series tend to occur in groups, the variation within the groups being relatively slight, but each group being centred on a different level of the quality under investigation. This would occur if two mule-spun cotton yarns were being compared for almost any quality, there being several mule cops per series and several tests on each cop (see Table V, for example). The correct procedure in such instances is by combination to obtain one result for each independent group and to treat the group values as individuals, as has already been discussed in Chapter 4.

A third type of non-randomness is heterogeneity of the variability within each series, when, in the words of quality control, the variability is out of control. This would happen in the nitrogen experiment of Table VI, for example, if some determinations were made by Lord Rayleigh himself and others by a less skilled assistant. Indeed there are signs of such heterogeneity in Table VI. Within the eight determinations for "chemical" nitrogen [series (a)], the results for nitric oxide are apparently more variable than those for nitrous oxide and ammonium nitrite. Without close examination one can not be sure of this (there are four readings for one and only two each for the other two), but the figures are enough to illustrate the point. Just what is

the effect of such heterogeneity on the t test of significance I do not know.

The Assumption of the Inclusiveness of the Estimate of Error

This assumption is closely related to the previous one. When a difference is statistically significant, it is correct to infer that it can not be attributed to the chance effect of such errors as cause the variations within each series; provided that all other assumptions are satisfied closely enough, the difference might be due to the effect under investigation or to errors that do not add to the standard error. In order to eliminate the second possibility and arrive at useful (as well as valid) conclusions, all the errors that enter into the comparison between the two series should have full play in causing variations within each series.

This proviso seems obvious when stated, but it is often overlooked. Analysts make replicate determinations in order to arrive at some idea of the error, but they do not always replicate the whole procedure right from the start. Perhaps they will take two or three specimens, treat them together in the analytical process, and merely do the final weighings independently; the difference between these weighings will not be subject to all the sources of experimental error.

We have seen that the difference between series (b) and (c) in Table VI, although not highly significant, is too large to be easily dismissed as due to chance. We see from Table VI that the two series of tests were done at very different times: August–September, 1892, and December, 1893, and it is quite conceivable that some change in the uncontrolled experimental conditions had occurred, and that that change rather than the change in reducing agent was responsible for the difference in means. Had Lord Rayleigh been aware of the modern statistical theory of experimentation he might have arranged his experiment so that such a source of error would contribute to the standard error, and so would be taken into account.

When any factor other than that under experimental investigation varies between the two series in such a way that it does not contribute to the errors, it is said to be *confounded* with the experimental factor. Generally factors that vary together so that their effects can not be separated are confounded.

The Appropriateness of the Statistical Model

It may be helpful to consider the assumptions underlying the t test as summed up in a certain statistical model, and to consider other

models that do occur to which the methods are inapplicable. The model is the relatively simple one of two random samples from populations having the same standard deviation and differing (if they differ in any way) only in the mean. Accordingly any factor that affects the two series differently merely adds or subtracts a constant amount to the quality of each individual.

We may see what this means graphically by imagining that we can pair the individuals in the two series, as in Table VII. Then if the
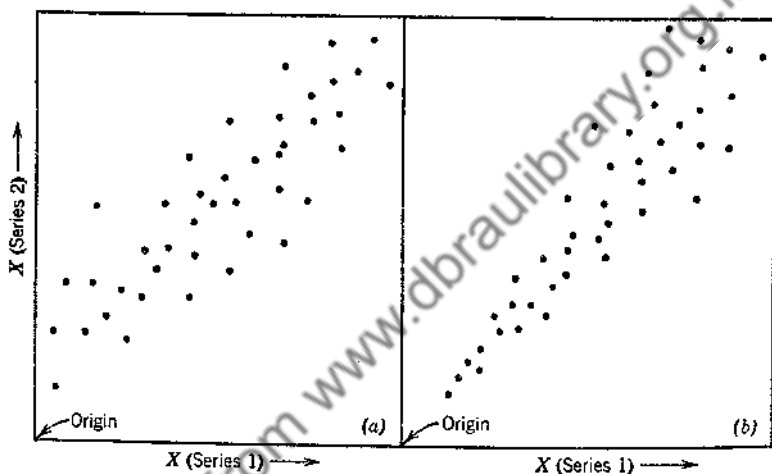

Fig. 17.

model is appropriate and we plot the results for, say, series 1 against those for series 2, using the same scale for both, we obtain a graph similar to that in Fig. 17(a). The points are all scattered about a line sloped at 45° to the axes, the degree of scatter being the same at all parts. If the line has an intercept on either axis there is a difference between the means for the two series, and if it goes through the origin the difference is zero; this is tested by the t test. If the scatter about the line is small, the variation that is common to the members of a pair is large compared with the variation in the differences between the members; if the scatter is so large that no trend can be discerned, there is no more statistical basis for the association in pairs than if the pairs were combined at random, and the t test for two independent series is appropriate.

Another simple model is that shown in Fig. 17(b). The points are scattered about a line that goes through the origin and may or may

not have a slope of 45°; and the degree of scatter increases as one proceeds away from the origin in such a way that the variation about the line measured on any linear scale is proportional to the distance from the origin. This model arises if each individual in series 1 tends to bear a constant ratio to the corresponding individual in series 2, the variations in that ratio being random and homogeneous over the whole range. If the two series are equal, the line has a slope of 45°; if not, the slope is other than 45°; but according to this model the line must go through the origin. Results conforming to this model may be treated by performing a t test on the ratios (or percentages) between the members of the pairs. Alternatively, instead of dealing with the original variable, X say, the ordinary test of significance may be performed on $Y = \log X$, since if X conforms to the model of Fig. 17(b) Y conforms to that of Fig. 17(a). The test on the ratios can only be performed if the individuals in the two series can be associated in pairs; that on the transformed variable Y can be performed whether or not this can be done.

Other, more complicated, models are possible. The points can be scattered about a line, sloped at some angle other than 45° and having an intercept on one axis; and the degree of scatter can increase with the distance from the origin, but not in proportion; or the points can be scattered about a curved line. If the appropriate model is known, the data can sometimes be transformed by some mathematical function other than simple logarithms, so that the transformed values conform to the model of Fig. 17(a); or some special treatment may be necessary. This leads us into realms of complication whither, fortunately, it is not often necessary to follow.

In this discussion it has been necessary to assume the possibility of associating the individuals in the two series in pairs. If this can be done it is a good plan to plot the results. The plot will show any gross departures from the assumed model or will give confidence that the t test can be safely applied. More often than not, the results are so few and the scatter is so great that the simple model of Fig. 17(a) seems to be as good as any other, and the t test may be used. In many of the remaining cases, the model of Fig. 17(b) is obviously more appropriate, and ratios, percentages, or logarithms may be used.

If the two series are independent, there is nothing to show whether the simple model applies. If from technical or other knowledge the model of Fig. 17(b) seems appropriate, the logarithmic transformation may be used. Otherwise there is nothing for it but to apply the ordinary t test and to realise that a significant value of t may be due,

not to a simple difference between the means of the two series, but to some more or less complicated departure from the assumed model—another way of saying what has already been said, that a significant value of t may indicate that the assumptions, or the hypothesis, or both are wrong.

The Precision of an Estimate

The foregoing considerations also arise in the use of confidence or fiducial limits for describing the precision of a mean or a mean difference, or a difference between two means.

The confidence coefficient or fiducial probability that determines the limits has to be chosen, and a kind of principle of indeterminacy arises. If the limits are widely spaced, we can have considerable confidence that the true value lies between them; if they are narrowly spaced, that confidence is low. The closer the approach to exactness in our knowledge the lower is our confidence; the higher the confidence, the less exact is the information. The levels of 0.95 and 0.99 have usually been found satisfactory in practice.

The assumptions of Normality, of equal variability, of randomness, and of the inclusiveness of the estimate of error have the same importance in this connection as in the testing of significances. The last assumption requires further discussion.

When most of the variations are not experimental error but are due to real variations that occur naturally in some material, it is usually not difficult to satisfy the assumption by the adoption of a suitable sampling procedure such as that discussed on pages 119 to 122. But, when the variations are largely experimental errors, it is difficult to be sure that everything has been taken into account.

Suppose, for example, that it is desired to estimate a physical constant. The result of a particular determination may be the algebraic sum of the following terms:

Observed value = (1) true value + (2) method error + (3) laboratory error + (4) observer error + (5) observer's time error + (6) random error

The true value is what we want to know, and, if we can not know it exactly, we want to know within what limits it lies. There are often different methods of determining the same quantity that, however well they are conducted, give slightly different results; the difference between the result by any one method and the true value is the method error. Furthermore, different laboratories using the same method often

arrive at different results, and so there is superimposed a laboratory error; and since each laboratory will employ different observers there may also be superimposed an observer error. Moreover a given observer will often show a secular change in his results, so that any one particular result is also affected by a time error; and finally we include all the residual errors in the random error.

The general idea is that the various errors are positive and negative and cancel out in the long run (this can be made a consequence of definition rather than an assumption), and that, if the determinations are replicated a number of times and the results averaged, the confidence limits show within what limits the true value probably lies. The statistical method does not pretend to do more than say within what limits a certain population mean lies, and the argument involves the identification of the population mean with the true value. Let us examine this argument, ignoring as a separate issue the difficulty of bringing secular errors into a random scheme [item (5)].

It is not difficult for an observer at one laboratory, by one method, to make several determinations at a time and so deduce within what limits a population mean comprised of the sum of items (1) to (5) lies. But this is not very useful information; it includes as an unknown the observer's error and his time error, and we certainly are not interested in those. By allowing the observer to repeat his determinations at different times and allowing several observers at one laboratory to do this, we bring items (4) to (6) into the estimate of error, and we can say within what limits the sum of items (1) to (3) lies. This may be of some limited interest for practical purposes, especially if the laboratory is an important one and often makes determinations of the kind in question. If successive lots of a material are sent to a laboratory for analysis, we would not be perturbed at a consistent laboratory error, provided that observer, time, and random errors could be taken into account. But we would be far from knowing the true value, or how near our estimate was to it. The result for the weight of nitrogen deduced on page 90 probably includes an unknown method and laboratory error. We could, of course, go further and have the determination made at several laboratories and so estimate (1) plus (2). But I do not see how to estimate the method error and so arrive at the true value; it is by no means certain that the errors inherent in all the known methods cancel out, but perhaps the analysis of such data into true value and method errors becomes an abstraction that is philosophically dubious.

H

The point of all this is that the application of statistical theory to purely experimental errors (as opposed to real or natural variations) is beset with doubts and difficulties, and it is far better to improve experimental technique so as to reduce such errors to unimportant limits. If this can not be done, it is better to use statistical theory cautiously than to do nothing.

Chapter 10. APPLICATIONS OF THE ANALYSIS OF VARIANCE: BASIC FORMS

In the previous two chapters most of the statistical principles involved in analysing the data of technical experiments and investigations have been introduced; the rest is mostly added complication to deal with the more complicated situations that so often arise in working life.

The more complicated (and powerful) methods are based on a method known as the *analysis of variance*. The variance is a measure of variability and is the square of the standard deviation. Its use instead of the standard deviation is entirely for arithmetical and algebraic convenience, but the convenience is very great. If two or more independent sources of variability are operating on a product, the variance of the combined effect is the sum of the variances due to the separate sources. The same is not true of the standard deviation.

The following sections will illustrate the methods of analysis and problems that arise by presenting the data of some actual experiments and investigations—data that appear in a number of standard forms. It is not our aim to deal with the subject exhaustively or in detail, but rather to display some of the most important tools in the statistician's chest and show what they can do.

Single-Factor Form

Table VIII gives the results of a weaving experiment that was conducted in a factory. There were 6 lots of warp yarn labelled respectively *AL*, *AM*, etc. They were spun from two growths of cotton, *A* and *B*, and each cotton was spun to three twists (i.e., the number of turns in the yarn per inch): low (*L*), medium (*M*), and high (*H*). The combinations of these factors give the 6 kinds of yarn, which are the experimental treatments. From each yarn were prepared 9 warps (a warp is a quantity of warp yarn that goes into one loom as a unit), and, as a loom came available in the course of events, a warp chosen at random from the 54 was assigned to it, until ultimately all 54 were disposed of. More than one warp was woven in some looms, but that did not upset the randomness of the distribution. The number of warp

TABLE VIII

WARP BREAKAGE RATES FOR INDIVIDUAL WARPS

	Yarn					
	AL	AM	AH	BL	BM	BH
	26	18	36	27	42	20
	30	21	21	14	26	21
	54	29	24	29	19	24
	25	17	18	19	16	17
	70	12	10	29	39	13
	52	18	43	31	28	15
	51	35	28	41	21	15
	26	30	15	20	39	16
	67	36	26	44	29	28
Totals	401	216	221	254	259	169
Means	44.56	24.00	24.56	28.22	28.78	18.78
Grand mean	28.15					

threads that broke during the weaving of each warp was counted and expressed as a rate of so many breaks per unit length of warp. These are the figures given in the body of Table VIII.

The weaving quality of each yarn is expressed by the corresponding mean at the foot of the table, but before considering the technical implications of the results (whether yarn *BL*, for example, which is more costly than *AL*, is worth the extra cost by reason of the lower mean breakage rate) we want to know how far errors can account for the observed differences. We might apply the *t* test to every pair of means, but this would be laborious (there are 15 pairs), and the conclusions would be doubtful. The probability calculated from the *t* test can only be interpreted in the way of Chapter 8 if there is only one difference under test, and to extend the argument to a set of 15 probabilities is difficult. Thus, even if all the variations in Table VIII were purely random, the simple *t* test would probably show the largest of the 15 differences to be "statistically significant."

One possibility would be to make a control chart of the treatment means, regarding their variation as significant if values fall outside,

say, the three-sigma limits. This is not a good procedure, as the control chart is not a very precise instrument when the observations are so few.

The best way is to conduct an analysis of variance as shown in Table IX. This enables us to test whether the differences between the 6 means, as a whole, are significant compared with the variations within each series.

TABLE IX

ANALYSIS OF VARIANCE OF WARP BREAKAGE RATES

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Treatments	3487.7	5	698
Error	5745.1	48	120
Total	9232.8	53	

The sum of squares in the row labelled "Total" is the sum of squares of the deviations of the individual readings from the grand mean; that is, $(26 - 28.15)^2 + (30 - 28.15)^2 + \cdots$, etc. The degrees of freedom, 53, is one less than the number of observations, and 9232.8/53 is the σ^2 measuring the total variability of the 54 observations. It is not entered in Table IX because it is of no interest. The sum of squares labelled "Error" is the sum of the squares of the deviations of the individual values from the treatment means; that is, $(26 - 44.56)^2 + (30 - 44.56)^2 + \cdots + (18 - 24.00)^2 + \cdots$, etc. The degrees of freedom is 6 less than the number of observations to allow for the fact that the deviations are measured from 6 means or, alternatively, each column contributes 8 degrees of freedom and $6 \times 8 = 48$. The mean square of $120 = 5745.1/48$ is a pooled estimate of the σ^2 measuring the within-treatment variability—the variability due to the experimental errors. The remaining sum of squares is 9 times the sum of squares of the deviations of the treatment means from the grand mean; that is, $3487.7 = 9 [(44.56 - 28.15)^2 + (24.00 - 28.15)^2 + \cdots]$. One reason for multiplying by 9 is that the two sums of squares add up to the total; the result is the same as if the appropriate treatment mean were substituted for each of the 54 readings, and the 54 values measured as deviations from the grand mean, squared, and added. The degrees of freedom are 1 less than the number of treatments; and the mean square

of 698 = 3497.7/5 is in some way a measure of the variability between treatment means. Let us consider this further.

Suppose that there was no treatment effect and that the apparent variation between the treatment means was due entirely to random variation resulting from the within-treatment variation and the fact that each mean is a mean of only 9 values. Then the standard deviation of the treatment means would be the standard error $\sigma'/\sqrt{9}$ (where σ' is the within-treatment standard deviation), the variance would be $\sigma'^2/9$, and 9 times the variance, which is the value entered as 698 in Table IX, would be σ'^2. In those circumstances the "Treatments" and "Error" mean squares would be estimates of the same variance σ'^2. But the direct estimate of σ'^2 is 120, and, since 698 is greater than this, there is an indication that the variation between treatment means is greater than that due to the effect of the within-treatment variations alone (i.e., that the treatments as a whole have an effect). Of course, since the two mean squares are only estimated on a few degrees of freedom, they would not be expected to be exactly equal even if there were no treatment effect. To establish a real treatment effect, the treatment mean square must not only be greater than the error, it must be significantly greater. The significance is tested with the aid of the ratio of the variances, termed F, as a criterion. If the two mean squares are estimates of the same variance, F has a certain sampling distribution depending on the two numbers of degrees of freedom and there are tables of values of F lying on various levels of significance. For 5 and 48 degrees, $F = 3.42$ lies on the 0.01 level. The value of F for Table IX is $698/120 = 5.8$, and it lies well above the 0.01 level; it is highly significant, and we infer that the treatment effects in Table VIII predominate well above the effects of error.

The F test is really an extension of the t test of significance, for F is the ratio of the variation between a number of means to the error variation and t is the ratio of the difference between two means to the error variation. If the F test were applied to testing the effect of two treatments, it would be found that $F = t^2$ exactly. So if you have mastered the concepts behind the t test, you may think of the F test in much the same way.

For a large number of treatments, the table of analysis of variance tells much the same tale as a control chart. Exact equality of the mean squares ($F = 1$) corresponds to 5 per cent of the points lying outside the two 0.025 control limits.

After the statistical significance of the difference between the yarns is established, the next step is the examination of the differences for

technical significance. Because they are significant as a whole, it does not follow that every difference is so; nor if some differences are small do they necessarily correspond to small real effects. Statistical theory provides a measure of error which gives some guidance, but it does not remove the uncertainty due to errors. The measure of error in this instance is the standard error of the difference between any pair of means, of which the pooled estimate, based on 48 degrees of freedom is $\sqrt{120(\frac{1}{9} + \frac{1}{9})} = 5.17$; and the 0.95 confidence belt has a width of $\pm 2.01 \times 5.17 = \pm 10.4$. Now let us examine the means of Table VIII.

Yarn AL is undoubtedly the worst (i.e., its mean breakage rate differs from that of the next highest by about 16). Whether yarn BH is significantly the best is not so indubitable; we may test it in the following way. The combined mean breakage rate for yarns AM, AH, BL, and BM is 26.39, which is 7.61 greater than that for yarn BH. The standard error of this difference is $\sqrt{120(\frac{1}{36} + \frac{1}{9})} = 4.08$, and t is $7.61/4.08 = 1.86$, which is between the 0.05 and 0.10 levels of significance for 48 degrees of freedom. Thus we can say with reasonable confidence that yarn AL is the worst; yarn BH is possibly the best, although a second experiment might not confirm this; and the differences between the other yarns are small compared with the errors.

There is another possible line of interpretation. We may be interested, not to pick out individual 'good and bad yarns, but to learn something of the effects of cotton growth and twist. It is conceivable that the results of Table VIII are explicable by a difference at all twists due to the cottons A and B, and for each cotton a decrease in breakage rate as the twist increases from L to M and thence to H; the mean for AM might easily have been about 28 and that for BM about 24, and then this possibility would have been more apparent. To test this possibility requires a more complicated statistical analysis than we have so far made, and we shall return to this in the next chapter (p. 135).

You will notice that we have made two kinds of approach to the results of Table VIII, and there may be others. The choice between them depends on non-statistical considerations.

Before we leave the discussion of Table VIII, some attention must be paid to the assumptions, which are of the same kind as those used in applying the t test. Strictly the assumption of homogeneity of errors is probably not justified. Thousands of warp breaks have been observed, and it has been found that, on the average, the error variance is proportional to the mean value. The results of Table VIII are in ac-

cordance with this, for, if we take the range as a rough measure of error variation, it is 45 for yarn AL, it averages 28 for yarns AM, AH, BL, and BM, and it is 15 for yarn BH. This does not invalidate the test of significance based on Table IX, because (a) the number of warps is the same for all yarns and (b) the assumption of equal error variances is satisfied if the hypothesis of equal means is true; the significance of the departure from randomness means a variation in both means and error variances (or standard deviations). But the pooled estimate of the error variance is not a good basis for testing the significance of the difference between yarn BH on the one hand and yarns AM, AH, BL, and BM on the other. A t test of that difference based only on the warps for the five yarns in question would probably be nearer the 0.05 level of significance (this has not been investigated). Alternatively, when the error variance is proportional to the mean, homogeneity is attained by analysing the square root of the observed values instead of the values themselves (see p. 180).

The random distribution of the warps among the looms ensures that errors due to differences between the looms and weavers affect the replicates for each yarn as much as the yarn differences. But there is one source of error not taken into account. The warps for each yarn belong to a different "set"; and the sets have gone separately through the preparatory processes, such as sizing, and on that account may differ by more than the warps of any one set. In this experiment set differences are confounded with yarn differences. Apart from any speculations one may be willing to make as a result of previous experience with weaving experiments, the only way of determining the set error is by replicating the sets, perhaps by repeating the experiment once or twice (a repetition is necessary, in any event, to reduce the effects of the within-set error). This idea prompts the reflection that it might be better to have fewer warps per set and more sets. Unfortunately it costs a good deal more to prepare 9 warps in 3 sets than in 1, and it upsets the factory routine, and, as the investigators were in this factory on sufferance, it would scarcely have been practicable to have had fewer warps per set. It might have been better in the first experiment, however, to have had fewer yarns, say two cottons each at two twists, and to have had two sets per yarn. That would have given information about the four treatments and about all the errors and would have cleared the way for a fuller investigation with other twist factors. But this, to some extent, is wisdom after the event.

Only some of the technical background and details of the weaving experiment of Table VIII are given here; to present all would be to

waste your time and, what is worse, add needlessly to your confusion. Enough has been given to illustrate the points under discussion.

The data of Table VIII may appropriately be said to be in the *single-factor form*, and the analysis of Table IX to constitute a *single-factor analysis* because there is one factor additional to the error.

Table VIII gives the results of an arranged experiment in which there is interest in the individual treatment means; Table X gives the result of a single-factor analysis made to investigate variations as they occur rather than as they were produced experimentally.

Mule cops of cotton yarn were collected in blocks of 20, each block being from a different mule (the technical terms are explained on p. 28). Two leas were weighed from each cop, giving for each block 39 "total" degrees of freedom, 19 "between cops" degrees, and 20 "within cops" degrees, together with corresponding sums of squares. There were 6 blocks, and the 6 sets of sums of squares and degrees of freedom were added to give those entered in Table X. The mean squares give

TABLE X

ANALYSIS OF VARIANCE OF WEIGHTS OF LEAS OF COTTON YARN

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Between cops (within blocks)	19,138.85	114	167.88
Within cops	5,681.00	120	47.34
Total	24,819.85	234	

an analysis of the variations within the blocks, and these are interesting because they are due to factors not easily controllable. The block-to-block variations are not studied because they can be easily eliminated by careful adjustment of the mules.

In Table X "between cops" corresponds to the factor and "within cops" to the error; but it is inappropriate to speak of error when the investigation is not a controlled, or partially controlled, experiment.

The "between cop" mean square is much greater than the other, and there can be no doubt of the "cop effect." But there is no point in examining the 120 individual cop means to pick out the high and low ones; a mule spins a thousand or more cops in a day, and it is im-

practicable to "keep tabs" on all those individually. What is needed is a statistical description of the cop variation, and this is provided by the *corrected cop variance.*

The mean square of 167.88 is n times ($n = 2$, the number of leas per cop) the apparent variance between cop means. But it is enhanced by the within-cop variance: it would equal 47.34 if there were no cop effect. The corrected cop variance, free from the effect of the within-cop variation, is $(167.88 - 47.34)/2 = 60.27$; this is an estimate of the variance between cop means if an infinity of leas could be tested per cop.

Perhaps the best way of expressing to the technical man the importance of the cop effect is to state that with it, if a large number of leas are taken at random, 1 per cop, the variance is $60.27 + 47.34 = 107.61$ and the corresponding standard deviation is 10.37. If the cop effect is eliminated, the corresponding variance is reduced to 47.34 and the standard deviation to 6.88. The statistical justification for these statements can be found in text-books. The appreciation of the mysteries of an analysis of variance can thus be reduced to the appreciation of two standard deviations; the technician, manager, or executive who has no "feel" for the standard deviation is in a hopeless case for appreciating the results of a large area of statistical investigation.

These results can be used in a slightly different way. When yarn is used as weft or filling, that from one cop forms adjacent threads in a width-ways strip of cloth a few inches long; and, in a curtain material that may be viewed by transmitted light, the strip containing yarn from a cop with a high mean weight per lea appears denser than that corresponding to a low mean lea weight. The variability between cops of mean lea weight when all the leas on a cop are weighed is thus related to the appearance of stripiness. There are 25 leas per cop, and the corresponding weight variance is thus $60.27 + (47.34/25) = 62.16$, giving a standard deviation of 7.88. Thus we have been able to deduce from tests on 2 leas per cop an index of the stripiness that shows when whole cops are woven. In an investigation, for example, it would be feasible to test 2 leas from each cop of a batch in order to obtain the standard deviation, and to weave the remainder in order to obtain the corresponding cloth. It has been assumed, quite reasonably, that there is no important pattern of variation within the cops.

All these variances and standard deviations are only estimates based on limited numbers of degrees of freedom, of population or true variances and standard deviations; and they are not very precise unless there are many degrees of freedom.

TABLE XI

YIELD POINT OF STEEL DISCS (TONS PER SQUARE INCH)

Cast No.	Ingot Means			Cast Means
Heat Treatment I				
1	20.1	. . .	. . .	20.10
2	22.6	22.4	. . .	22.50
3	24.6	22.55	. . .	23.57
5	20.8	21.0	20.75	20.85
7	22.4	. . .	. . .	22.40
8	21.9	. . .	. . .	21.90
Heat Treatment II				
4	21.25	21.9	. . .	21.57
6	21.15	. . .	. . .	21.15
Heat Treatment III				
5	19.45	18.4	. . .	18.92
6	21.9	. . .	. . .	21.90
Heat Treatment IV				
3	20.1	19.55	. . .	19.82
Heat Treatment V				
8	20.1	. . .	. . .	20.10
10	20.75	20.15	. . .	20.45

In the last two examples, the number of observations per yarn or cop is uniform; this arrangement facilitates the analysis and interpretation. Sometimes, however, especially when existing works records are used, the data can not be arranged in this convenient form. Table XI, which is taken from *Statistical Methods in Industry*, was obtained from works

records of routine tests of yield point made on specimens taken from steel discs—one test per disc. Several of the type of disc in question were made from an ingot, and, in order to eliminate the effect of any possible pattern of variation through the ingot, only those ingots were included for which the first four discs had been tested. The ingot means in Table XI are means for the first four discs per ingot. Furthermore, several ingots would be cast at the same tapping of the furnace, and sometimes more than one from the same cast would be used to make discs yielding results in Table XI; the ingots are therefore identified by the cast number. There were one ingot from cast 1, two from cast 2, and so on. The discs had also been subjected to different heat treatments, labelled I to V. Heat treatment is known to affect yield point, and so it is hardly worth while investigating that effect, but Table XI does give some information on whether there are any uncontrolled cast-to-cast variations that affect yield point. For any one heat treatment the cast means vary: are these variations greater than can be attributed to within-cast variations? This is the sort of question a control chart might answer were the data more extensive and systematic; as things are, an analysis of variance provides a good method.

Since we are not interested in treatment variations, we may regard each one as providing a separate block, measuring all the deviations from the treatment mean, and finally adding the sums of squares and degrees of freedom for the treatments. Casts numbered 1, 7, and 8 give no information on the within-cast variation, but they add to the information on the between-cast variation. Cast numbered 3 in treatment IV gives no information on the between-cast variation, since that is the only cast associated with that treatment, but it adds to the information on the within-cast variation. Thus all the data of Table XI add variously to our knowledge.

The results of the analysis are in Table XII. The calculation of the sum of squares for casts is somewhat tricky. The treatment means to be used are means of the individual ingot values [e.g., $(20.1 + 22.6 + \cdots)/10$ for treatment I], and the deviations of the cast means from the appropriate treatment means are squared, multiplied by the number of ingots per cast, and then summed. The ingot values are measured as deviations from the cast means, squared and summed in the ordinary way to obtain the result 3.25. The 8 degrees of freedom for the between-cast sum is made up of 5 from heat treatment I, 1 from II, 1 from III, 0 from IV, and 1 from V. The 8 within-cast degrees of freedom are made up of 1 each from casts numbered 2,

TABLE XII

ANALYSIS OF VARIANCE OF YIELD POINT OF STEEL DISCS (HEAT TREATMENT CONSTANT)

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Between casts	19.23	8	2.40
Between ingots within casts	3.25	8	0.41
Total	22.48	16	

3 (in heat treatments I and IV), 4, 5 (in heat treatment III), and 10, and 2 from cast numbered 5 (in heat treatment I). The mean squares in Table XII give a value of $F = 9.61/1.62 = 5.93$, which on 8 and 8 degrees of freedom lies near the 0.01 level of significance. The cast effect is well established, and the onus is now on the technician to discover the causes of the variation. He can make use of the knowledge that the standard error of the difference between two cast means belonging to the same heat treatment is $\sqrt{0.41\left[(1/n_1) + (1/n_2)\right]}$, where n_1 and n_2 are the numbers of ingots per cast, and that the 0.95 confidence limits are at 2.31 times this above and below the observed difference (2.31 being the value of t on the 0.05 level of significance for 8 degrees of freedom). With such varied numbers of ingots per cast it is very difficult to estimate the corrected cast variance.

These results depend on the usual assumptions applying, but when the data are so scanty the conclusions are not very precise, and only departures from the assumptions that are obvious are likely to be important. No such departures are apparent in Table XI.

In all the foregoing examples, the mean square associated with the main factor under investigation is greater than that associated with error or within factors. What do we infer when the latter is greater than the former? The short answer is that, if the difference is statistically significant, it indicates a non-random pattern of variation within the factor units.

Two-Factor Basic Form

An example of data in the two-factor basic form is in Table XIII, which is also taken from *Statistical Methods in Industry*, and which

gives the yield point of specimens of steel taken one each from a number of steel discs. In Table XIII values are given for each of the first

TABLE XIII

YIELD POINT OF SPECIMENS FROM STEEL DISCS (TONS PER SQUARE INCH)

Ingot No.	Order of Disc from Ingot				Mean
	1	2	3	4	
1F	20.4	20.4	19.2	20.4	20.1
2A	22.8	22.8	22.0	22.8	22.6
2B	21.2	21.6	22.8	24.0	22.4
3D	19.2	20.4	19.6	21.2	20.1
3H	20.8	20.4	18.4	20.8	20.1
3I	19.6	20.0	19.0	19.6	19.55
3I	21.4	22.2	26.0	28.8	24.6
3K	20.6	19.2	23.3	21.6	22.55
3L	18.8	19.0	17.6	20.4	18.95
3O	19.9	20.1	19.4	21.2	20.15
4B	19.8	20.4	22.0	22.8	21.25
4C	20.8	20.8	22.8	23.2	21.9
5A	18.0	19.2	20.2	20.4	19.45
5B	18.4	17.6	18.6	19.0	18.4
5E	21.4	20.4	20.6	20.8	20.8
5F	21.0	20.4	21.0	21.6	21.0
5I	21.4	19.2	20.8	21.6	20.75
6A	20.4	20.6	22.0	21.6	21.15
6B	22.8	21.4	21.8	21.6	21.9
7D	22.8	23.6	22.4	20.8	22.4
8B	21.2	22.0	21.4	23.0	21.9
8D	19.4	21.2	20.8	21.4	20.7
8E	18.8	20.0	20.8	21.2	20.2
8F	20.2	20.8	19.2	20.2	20.1
8G	20.0	21.0	21.5	20.6	20.8
9B	18.4	18.8	21.0	21.2	19.85
10B	21.2	21.0	20.0	20.8	20.75
10C	19.8	19.6	20.8	20.4	20.15
Mean	20.38	20.50	21.09	21.54	20.88

4 discs made from 28 ingots. The column of means shows an apparent variation from ingot to ingot, and the row of means suggests a steady increase in yield point from the first to the fourth order. We may

postulate that each yield point is made up of: the grand mean yield point + a deviation due to the ingot variation + a deviation due to the order variation + a random or residual deviation, regarding any one ingot deviation as being the same for all discs from that ingot, and any particular order deviation as the same for all ingots. Numerically this postulate is expressed for disc 1 from ingot $1F$ as follows (the terms being in the same order as before):

$$20.4 = 20.88 - 0.78 - 0.50 + 0.80$$

For disc 2 from ingot $1F$ the equation is

$$20.4 = 20.88 - 0.78 - 0.38 + 0.68$$

and for disc 1 from ingot $2A$ it is

$$22.8 = 20.88 + 1.72 - 0.50 + 0.70$$

According to this model the variation is due to the effect of two factors, ingots and orders, superimposed on the random variation; hence the term *two-factor* form. This form is also termed *basic* in order to distinguish from other forms, in which the effects of two factors may be superimposed, which will be illustrated in the next chapter. I also prefer to term the random effect *residual* because, as we shall see, it may be due to assignable causes and it is unsound at this stage to prejudice the issue; the variation is in fact the residual left over after the effects of the factors have been accounted for.

This analysis is purely conjectural, and we must subject it to tests in order to discover if the postulated effects are statistically significant. After all, if the figures in the body of Table XIII were distributed entirely at random, there would be certain differences between the ingot and order means, and we need to know whether the observed differences are greater. This we determine by the analysis of variance shown in Table XIV. The total sum of squares is that of the deviations of the individual results from the grand mean, and the degrees of freedom are one fewer than the total number of results. The ingots sum of squares is 4 times the sum $(20.1 - 20.88)^2 + (22.6 - 20.88)^2 + \cdots$, etc., and the degrees of freedom are one fewer than the number of ingots. The order sum of squares is 28 times the sum $(20.38 - 20.88)^2 + (20.50 - 20.88)^2 + \cdots$, etc., and the degrees of freedom are one fewer than the number of orders. The residual sum of squares is obtained from the residual deviations calculated above and is $0.80^2 + 0.68^2 + \cdots + 0.70^2 + \cdots$, etc.; alternatively it may be obtained by subtracting from the total the ingots and order sums of squares. The residual

TABLE XIV

Analysis of Variance of Yield Point of Steel Discs

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Ingots	185.3874	27	6.87
Order	24.3717	3	8.12
Residual	130.7858	81	1.62
Total	340.5449	111	

degrees of freedom are likewise obtained by subtraction. The beginner who carries through the calculation will arrive at a better appreciation of the basis of the analysis; for a full justification he must refer to the text-books.

Each ingot has one of each order, so that ingot differences are unaffected by the order effect, and the ingot mean square in Table XIV measures the ingot effect plus the residual variation; it is free from the order effect. Likewise the order mean square is free from the ingot effect. The virtue of the arrangement of the data is that it enables us to separate the effects of the two factors.

If there is no ingot effect, the ingot mean square will equal the residual, within the limits of sampling errors; the ratio F is $6.87/1.62 = 4.24$, and on the basis of 27 and 81 degrees of freedom it lies above the 0.001 level of significance. The F for order is 5.01 and for 3 and 81 degrees lies rather above the 0.01 level of significance. Both effects can therefore be accepted with some confidence.

The establishment of the statistical significance of the effects is, as usual, the beginning rather than the end of the full investigation. Individual ingots may be compared with a standard error of a difference in means equal to $\sqrt{1.62(\tfrac{1}{4} + \tfrac{1}{4})}$. The standard error of the difference between any two order means is $\sqrt{1.62(\tfrac{1}{28} + \tfrac{1}{28})}$; but the technician will probably be less concerned to compare pairs of orders than to note the steady trend in yield point down the ingot.

The ingots may reasonably be regarded as a random sample from a population, so that the corrected variance of the ingot effect has some meaning. By a simple extension of the argument applied to the single-factor form we have

$$\text{Corrected ingot variance} = \frac{6.87 - 1.62}{4} = 1.31$$

The importance of the ingot effect relative to the residual may be expressed by the following standard deviations:

Standard deviation (residual
variations alone) $= \sqrt{1.62} = 1.29$

Standard deviation (residual
plus ingot variations) $= \sqrt{1.62 + 1.31} = 1.71$

In view of the pronounced trend of yield point with order, and of the limited number of discs that can be forged from an ingot, it would be quite unreasonable to regard the four orders as a random sample from an infinite, or even a large population, and the corrected variance for orders has not the same meaning as that just given for ingots. However, it has a meaning which will be described in the next section.

Application to Sampling Inspection

I do not know whether steel discs are dealt with commercially in the following way, but many articles that show statistically the same kinds of variation are, so this discussion has some practical relevance, even if there is none to steel discs.

We may imagine a customer receiving deliveries of discs in large lots all mixed up, so that he has no means of identifying their ingot or order number. He is interested in the variability of the yield point values, and he knows that both ingot and order effects contribute to it. If all the discs are taken at random from the first four orders so that it is left to chance how many there are from each, the contribution of the order effect is the corrected order variance of

$$\frac{8.12 - 1.62}{28} = 0.23$$

a modest contribution. Then the customer would find the following standard deviation to compare with those given above:

Standard deviation (residual + ingot
+ order variations) $= \sqrt{1.62 + 1.31 + 0.23}$

$$= 1.78$$

If each order is equally represented in each lot, the corrected order variance for use in this connection is $\frac{1}{3}$ of 0.23.

I

These results may have another application. Each lot of discs may be accepted or rejected on the basis of the mean yield point determined on a sample. If n discs are tested at random from ingots and orders taken at random, the standard error of the mean is $1.78/\sqrt{n}$. If the n discs are all first-order discs, the order variation does not contribute to the sampling errors for the purposes of comparison and control, and the standard error is $1.71/\sqrt{n}$. The reduction is small because the order effect is small; but in some applications such an effect can be large, as the next example will illustrate.

The above arguments apply only to the first 4 discs from each ingot. Until the other orders of discs have been investigated we can say nothing about them.

The next example concerns the bricks used in Chapter 7 to illustrate sampling schemes. Table XV, which is adapted from one given by

TABLE XV

SPECIFIC GRAVITY OF SILICA BRICKS (DEVIATIONS FROM 2.30 MULTIPLIED BY 200)

Zone	Kiln Number									
	1	2	4	5	6	(7)	8	9	10	Total [1]
R1	4	2	5	2	8	(14)	13	4	4	42
L1	6	18	3	4	3	(11)	4	4	6	48
W1	8	14	10	12	6	(16)	10	7	7	74
C1	2	4	4	2	3	(5)	2	4	2	23
B1	4	10	3	4	4	(15)	6	10	6	47
R2	6	2	5	6	3	(12)	6	6	2	36
L2	4	4	4	4	4	(8)	2	6	4	32
W2	12	8	8	4	6	(10)	4	4	9	55
C2	4	3	4	4	3	(6)	2	3	3	26
B2	4	4	2	4	4	(10)	6	4	4	32
R3	4	5	4	7	7	(13)	7	4	6	44
L3	5	4	8	8	9	(10)	5	2	4	45
W3	16	4	10	8	8	(16)	6	8	5	65
C3	6	7	6	8	12	(13)	6	6	4	55
B3	0	4	4	6	6	(10)	5	4	5	34
Total	85	93	80	83	86	(169)	84	76	71	658

[1] Excluding kiln 7.

Mr. W. T. Hale in the paper already referred to,* gives the specific gravity of individual bricks in units chosen because they yield small numbers which are easy to handle arithmetically. The bricks were taken from 9 kilns, and in each kiln from 15 different sections, termed zones, into which the kiln is divided. The symbols for the zones have a significance which will be described in the next chapter. The zones are common to the kilns so that the data are in the two-factor basic form. Cursory examination shows that the results for kiln 7 are very different from those for the others, and so they are omitted from the analysis. The rejection of data merely because they are different is a dubious procedure, as it is difficult to know where to draw the line, and there is a danger of rejecting data that should be included. Here, however, the difference is so marked that it seems unreasonable to regard kiln 7 as belonging to the same population as the others, and there are plenty of degrees of freedom remaining.

TABLE XVI

ANALYSIS OF VARIANCE OF SPECIFIC GRAVITY OF BRICKS (UNITS × 200)

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Kilns	20.77	7	2.97
Zones	353.72	14	25.27
Residual	679.48	98	6.93
Total	1053.97	119	

The analysis of variance shows that there is no kiln effect (the kiln mean square is less than the residual, but not significantly so); the zone effect is statistically significant ($F = 25.27/6.93 = 3.65$ and is substantially greater than the value on the 0.01 level of significance for 14 and 98 degrees of freedom). These results suggest that it is possible to control the firing of each kiln so as to keep the mean specific gravity substantially the same from kiln to kiln. With such knowledge, with knowledge of the steps taken by the producer to maintain control, and with the evidence of a control chart, the customer might well have sufficient confidence to accept different lots of bricks without any acceptance/rejection scheme of inspection. Without this confidence he will probably use a sample scheme.

* *Transactions of the Ceramic Society,* Vol. 46, 1947, p. 147.

If the bricks are delivered in mixed lots so that the producer does not know from which zone any bricks he selects for test came, the zone and residual variation will contribute to the variance from which the standard error is calculated. The zone corrected variance is $(25.27 - 6.93)/8 = 2.29$, and the variance of bricks taken at random is $2.29 + 6.93 = 9.22$. The standard error of the mean of n bricks is then $\sqrt{9.22/n} = 3.04/\sqrt{n}$. If the sample of n bricks can always be taken from the same zones, the zone variation does not contribute to the random error of the mean, and the standard error is $\sqrt{6.93/n} = 2.63/\sqrt{n}$. This reduction in standard error shows the economy possible in this instance through introducing a degree of representativeness or *stratification* (as it is called, the zones being the *strata*) in the sampling. Often only the producer has the knowledge of the origin of the bricks necessary to make such a scheme possible. The improvement in precision depends on the magnitude of the variation between the strata, relative to the residual, and there is an art in using technical knowledge to arrange the strata so that as much of the variation as possible is between strata. When the field of sampling is more or less continuous, as is the internal volume of a brick kiln, so that the strata have to be defined artificially, it is a matter for investigation to discover, in each instance, the best number and arrangement of strata. In all such investigations and in the presentation of the final results, the analysis of variance is an invaluable tool.

The foregoing conclusions for the bricks are true only if each brick is taken at random from all those belonging to its zone, or if the systematic (as opposed to the random) variation within each zone is relatively small. Here, with as many as 15 zones and a not highly powerful zone effect, the second condition is likely to apply.

The standard error of the mean of 4 bricks, with the zone effect eliminated, is $2.64/2 = 1.32$ in the units of Table XV or, in the units of specific gravity, $1.32/200 = 0.0066$. This is the standard error used for the sampling scheme of Chapter 7 (p. 60).

We shall consider the zone variation from another point of view in the next chapter.

Three-Factor Basic Form

The three-factor basic form of data is also termed the *Latin square*, and a simple example is given in Table XVII, taken from a paper by Main and Tippett.[†] A weaving experiment was done in experimental

[†] *Shirley Institute Memoirs*, Vol. 18, 1941, p. 109, or *Journal of the Textile Institute*, Vol. 32, 1941, p. T209.

TABLE XVII

LATIN SQUARE ARRANGEMENT IN LOOMS (1), (3), (4), AND (7),
AND WARP BREAKAGE RATES

Warp No.	426	427	428	429
Period 1	5.52 (1)	2.87 (4)	9.76 (7)	6.69 (3)
Period 2	6.02 (4)	6.25 (7)	5.14 (3)	9.16 (1)
Period 3	8.90 (7)	2.91 (3)	5.77 (1)	6.53 (4)
Period 4	6.09 (3)	5.07 (1)	2.83 (4)	9.77 (7)

workrooms to determine the weaving quality of 4 warps, numbered 426
to 429, each of which had been treated differently. The warps were
woven simultaneously in 4 looms, and the total weaving time was
divided into four periods. At the end of each period the warps were
interchanged between looms according to the plan of Table XVII, so
that by the end of the experiment each warp had spent one period in
each loom. This experiment differs from that described in connection
with Table VIII, because here there is only one warp per treatment,
the unit of production is the weaving of one period, and warps are
interchanged between looms.

A characterisic and virtue of the arrangement is that, if warp,
period, and loom variations operate by adding or subtracting a con-
stant amount for each warp, period, and loom, the differences between
the warp means are unaffected by the other two factors as are those
between period means and loom means. The warp, period, and loom
effects are separated from each other.

A Latin square may have any number of rows or columns provided
that the number of rows equals the number of columns.

The analysis of variance is performed by a simple extension of that
for the two-factor analysis; the results for the warp breaks are in
Table XVIII. There are 16 results in Table XVII, and hence 15 de-
grees of freedom altogether. There are 4 each of warps, periods, and
looms, and hence 3 degrees for each, leaving 6 degrees for the residual.
The residual mean square estimates the errors for each of the factor
effects. The period mean square is less than the residual, but not sig-
nificantly so, and there is no appreciable period effect. The values of
F for the warp and loom mean squares are respectively 7.06 and 9.39,
and they lie between the 0.05 and 0.01 levels of significance for 3 and
6 degrees of freedom. The warp and loom effects are probably real,

TABLE XVIII

ANALYSIS OF VARIANCE OF WARP BREAKAGE RATES

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Warps	29.4984	3	9.83
Periods	1.1726	3	0.39
Looms	39.1986	3	13.07
Residual	8.3138	6	1.39
Total	78.1834	15	

and their mean values should be examined. It is to be noted that warp and treatment differences are confounded, and the data do not show whether the treatments are responsible for the differences between the warp means. The loom effect is not very interesting technically; it is a source of variation that has been eliminated from the warp comparisons.

Data of naturally occurring variations can sometimes be collected in the three-factor basic form, and corrected variances be estimated. But this is not often done, and the most useful application of the form is to experiments.

Multi-factor Basic Forms

Arrangements in which there are as many rows as columns can be extended to cover 4 or more factors, but for these higher forms there are restrictions as to the numbers of rows and columns. The four-factor form is known as the *Graeco-Latin square;* since these forms introduce complications without any new statistical principles, and since they are not widely used, they will not be dealt with here.

General Discussion

It should always be borne in mind that the interpretations given in this chapter of the results of the analyses of variance depend on the same general assumptions as those discussed in connection with the *t* test, and that the general statistical model is of a homogeneous random variation with the effects of the factors as simple arithmetical additions. The hypothesis tested by comparing the variances is that each factor effect is zero, but a significant excess of a factor variance

over the residual may indicate either that the hypothesis is untenable or that the model is inappropriate. All this has been said in connection with the t test, but it will bear reiteration; it should never be forgotten.

The most important assumption is probably the homogeneity of the residual variations. The examination of this is easy when the data are in the single-factor form, but it becomes progressively more difficult as the form becomes more complex; these forms are usually analysed without much consideration being given to the validity of the assumptions. This is a weakness in our practice, but, if the possibility of error is borne in mind and obvious gross departures from the assumptions are not ignored, we are not likely to go seriously wrong.

The kind of data for which the assumption is likely to be wrong arises where there is a natural floor or ceiling to the possible values and the variations extend away from very near that floor or ceiling. For example, the floor is zero if negative values of the variable are impossible, and the ceiling is often 1.0 when the variable is a ratio. In these circumstances, the residual variation about a factor mean that is near the floor or ceiling is likely to be less than that about a factor mean well removed from it. There are mathematical transformations of the variable that can often be used in these circumstances.

The factor effects that the analyses disclose may be the effects of experimentally imposed variations, as for the yarn effect in Table IX, in which case the individual means are examined after significance is established; or they may be naturally occurring variations, as for mule cop effect in Table X and the cast effect in Table XII, in which case either the individual means may be examined or an estimate of "corrected variance" be made. If "corrected variances" are estimated they may be used to indicate the importance of the factors or to estimate standard errors associated with various sampling schemes.

In the interpretation the residual variance has so far been regarded as due to variations that, in the language of quality control, may be left to chance—to errors and other random variations. For the more complex forms the possibility of another interpretation emerges. Consider the steel discs of Tables XIII and XIV. Since only one test of yield point was made for each disc, at least part of the residual variation is due to testing errors and to the difference between the single specimen of steel tested and the whole wheel—a conglomeration of effects we may call error. But in addition the order effect might be really different for the different ingots; we can imagine a very large number of tests done on each disc so that the error in the mean is negligibly small, and yet it is possible for an analysis of variance on

the data to give a residual variance. When this occurs, the average ingot effect measured by the variation in ingot means for all orders is termed a *main effect*, the main ingot effect; the average order effect for all ingots is the main order effect; and the residual effect, measured by the deviations of the individual true values from the combined ingot and order means, is the *interaction* between ingots and orders. In a two-factor basic analysis the interaction is confounded with error.

When there are more than two main factors, the system of interactions becomes much more complicated, but, since in the basic forms it cannot be analysed, this is not the best place to discuss these complications. It is sufficient to remember that in the analysis of the more complex basic forms the residual variance is due to error plus a complex of interactions.

A fairly common difficulty arises in connection with the analysis of Table XVI. The variance for kilns is not significantly different from the residual: may we add the two sums of squares (700.25), the two degrees of freedom (105), and obtain an improved estimate of the residual variance (6.68) based on more degrees of freedom? In this instance the temptation to do this is not very great since the increase in degrees of freedom is not great; but sometimes this pooling would give a material increase in degrees. Is pooling justified? The position is complicated and obscure. At first sight it would seem permissible to adopt some level of significance for the F test and to pool with the residual mean square values that do not differ from it according to this criterion. Then, whatever level of significance is chosen, some estimates will be wrongly pooled, others that should be pooled will not be, and except in the rare circumstances of the two effects balancing exactly, the final pooled estimates will be biassed. The uncertainty of the bias is an argument against pooling. Again: if the effect of pooling is to make little difference to the final estimate, why pool? And, if the effect is to make a considerable difference, it is questionable if the pooled estimate is an improved one. Nevertheless, when the data are in a complex form there are many sources each contributing a few degrees of freedom, and many corresponding to no true technical effect; and often there are only a few degrees for the residual. Then, pooling may make a considerable difference to the degrees available. Perhaps a good working rule is to decide on technical grounds, before examining the data, which sources are likely to contribute no more than random variation (sources that are generally known to be "in control" and "high order" interactions), and to pool the corresponding sums of squares, and so on, with the residual unless the mean square for any

one is very different from that for the residual—say at a level of significance beyond 0.01, leaving the sums of squares, and so on, for the other sources unpooled, whatever the values of their mean squares. In other words, the main basis for a decision on pooling should, according to this suggestion, be technical rather than statistical, the statistical results being used only to prevent the pooling of estimates that are very different.

In the analyses of the more complicated forms of data, there are several mean squares to be compared with the residual, and, in the very complex forms that will be illustrated in the next chapter, there may be many. The same difficulties arise in testing the significance of these as of the significances of differences between several pairs of means. The difficulty is tied to the general one of having a list of probabilities and deciding which correspond to significance and which do not; and the solution, which has yet to be obtained, depends somewhat on the number of probabilities in the list.

It will be noted that the basic forms are balanced in the sense that each value of each factor is combined once with each value of every other factor. This simplifies the arithmetic and algebra of the analysis enormously, an effect that has to be considered in designing the investigation (Chapter 13).

Chapter 11. APPLICATIONS OF THE ANALYSIS OF VARIANCE: COMPOSITE FORMS

There is scarcely any limit to the number of forms in which data can be arranged for the analysis of variance, nor to the complexity of the forms. Many of them, however, are combinations of the basic forms introduced in the last chapter, and these will now be illustrated.

Two-Factor Composite Form

Data illustrating the two-factor composite form are in Table XIX, and are taken from a paper by Dr. B. P. Dudding and Mr. W. J.

TABLE XIX

RESULTS OF TESTS ON ARTICLES MADE ON MULTI-HEADED MACHINE

Side	Left						Right					
Head No.	1	2	3	4	5	6	7	8	9	10	11	12
	31	28	25	30	23	31	33	34	35	27	37	31
	33	36	30	30	28	27	36	29	35	36	29	39
	29	35	28	29	31	28	23	31	29	30	32	45
	31	29	31	33	32	32	43	39	35	23	42	36
	36	31	31	33	27	28	41	35	32	37	41	36
	34	39	29	37	31	31	33	34	35	39	32	39
	31	35	31	36	35	37	23	31	31	53	33	36
	24	33	36	45	23	43	33	27	39	38	37	51
Totals	249	266	241	273	230	257	265	260	271	283	283	313

Jennett.* Articles were made on a machine with twelve heads and tested. In the original paper the number of articles per head varies, but Table XIX contains only the first 8 results belonging to each head. The units of the test results are unspecified.

* *Journal of the Institute of Electrical Engineers*, Vol. 87, No. 523, 1940.

First we may regard Table XIX as being in the single-factor basic form, and analyse the variance into two parts: between the 12 heads (11 degrees) and a residual within heads ($12 \times 7 = 84$ degrees); the results are to the left of the columns of Table XX. The between-head

TABLE XX

ANALYSIS OF VARIANCE OF TEST RESULTS OF TABLE XIX

Source of Variation		Sum of Squares		Degrees of Freedom		Mean Square	
Between heads	Between sides		263.3		1		263.3
	Between heads within sides	648.6	385.3	11	10	59.9	38.5
Within heads		2330.9		84		27.8	
Total		2979.5		95			

mean square is greater than the residual at a level of significance between the 0.05 and 0.01 levels.

The head means themselves may be arranged in the single-factor form, as shown in Table XXI, where the first mean of 31.125 is the first total of Table XIX, 249, divided by 8, and so on. This may be treated like any other table in the single-factor basic form, and the variance

TABLE XXI

MEANS OF TABLE XIX FOR THE VARIOUS HEADS

Side	Left	Right
	31.125	33.125
	33.250	32.500
	30.125	33.875
	34.125	35.375
	28.750	35.375
	32.125	39.125
Side means	31.583	34.896
Grand mean	33.240	

may be analysed into two parts: between sides (1 degree) and a residual within sides, but between heads ($2 \times 5 = 10$ degrees). For inclusion in Table XX it is convenient to modify the arithmetical procedure slightly. When performing the basic analyses, we summed the squares of the deviations from some grand mean of the various factor means and multiplied that sum by the number of observations per mean in order to make the mean square directly comparable with that for the residual. This practice is extended, and the various sums of squares in the analysis of Table XXI are multiplied by the number of original articles or values per mean. The results are in Table XX. The sums of squares are

$$263.3 = 48[(31.583 - 33.240)^2 + (34.896 - 33.240)^2]$$

$$385.3 = 8[(31.125 - 31.583)^2 + \cdots + (33.125 - 34.896)^2 + \cdots]$$

$$648.6 = 8[(31.125 - 33.240)^2 + \cdots + (33.125 - 33.240)^2 + \cdots]$$

The multipliers arise because there are 48 articles per side and 8 per head. Now let us consider the results of Table XX.

The distinction between the two sides of the machine has been made because it was known in advance that there is some technical factor common to all heads on a side but different for the two sides. Dr. Dudding and Mr. Jennett do not state what the factor is, but in order to give a concrete picture we may imagine the following situation, which may be revolting to the engineer but will illustrate the interpretation of the analysis. We may imagine that the heads of each side are driven from a common shaft which is driven through a separate pinion geared to the main drive and that, owing perhaps to wear, the two "side" pinions may not be quite the same and, further, that any differences between the pinions cause only a difference in average quality between the articles produced on the two sides; such a difference is the side effect. Likewise we may imagine that the drive to the heads from each "side" shaft is through 6 "head" pinions which may vary in such a way as to cause differences in average quality between the articles produced on the 6 heads on a side; these differences we shall call the head effect. Finally a complex of random causes produces the residual variations between the articles produced on any one head. The effect of the arithmetical procedure behind Table XX is such that, if there were no head effect, the mean square 38.5 would be *statistically equal to* (if that term may be used as short for "not statistically significantly different from") the mean square 27.8. In fact, the difference lies well below the 0.05 level of significance, and there is no evidence

of a head effect (although we can not of course say that the head effect is non-existent). If there were no side effect, the mean square 263.3 would be statistically equal to 38.5, irrespective of whether there was or was not a head effect; and it would be statistically equal to 27.8 if there was no head effect. The value $F = 263.3/38.5 = 6.8$ lies between the 0.05 and 0.01 levels of significance for 1 and 10 degrees of freedom; the side effect is probably real.

It may be asked, why not compare the mean square 263.3 with 27.8? It seems clear that here, as long as the possibility of a head effect is entertained, the mean square of 27.8 under-estimates the error with which the side effect is measured. Differences between the head pinions contribute to the apparent difference between sides, and their effect must be taken into account in estimating the error (i.e., the denominator for F must be 38.5). But, it may be urged, we have already shown the head effect to be statistically not significant; may we not make use of that conclusion and test the head effect by $F = 263.3/27.8 = 9.5$? There is a strong temptation to do this because this value of F lies well above the 0.01 level of significance for 1 and 84 degrees of freedom; and since the head effect is so far from being significant this seems to be a reasonable thing to do. However, if the force of these arguments is admitted, it would be logical to combine the estimates 38.5 and 27.8 and obtain a pooled estimate of the residual mean square based on 94 degrees of freedom; and when this is considered it will be seen that the whole issue is the same as arises in the issue of the pooling of estimates, discussed in the last chapter (p. 126).

Sometimes, in this kind of analysis, the mean square occupying the place of 263.3 in Table XX can be greater than that occupying the place of 38.5, but not significantly so; that corresponding to 38.5 can be greater than that corresponding to 27.8, but not significantly so; and yet that corresponding to 263.3 can be significantly greater than that corresponding to 27.8. This is a tantalising situation, and the only certain way out is to obtain more data; although technical knowledge may sometimes show a way out of the dilemma.

After the significance of any effect has been established there follows, as always, the step of technical interpretation and action. Here the mere knowledge that a side effect exists may suffice to call the engineer's attention to something that needs examination and correction; he may or may not be helped in this by the knowledge that the right side produces articles of the higher mean quality. It is interesting to record that, when the machine of Table XIX was examined, the cause of the difference between the two sides was discovered and eliminated.

If required, confidence limits to the difference between any two means can be set in terms of the appropriate error variance, in the way illustrated in the last chapter.

Alternatively we may imagine a large number of machines like that of Table XIX, with side and head pinions that vary at random. Then there are three contributions to the total variability, with variances due to the side effect (σ_s^2 say), the head effect (σ_h^2 say) and the random effect (σ_r^2 say). Then the mean squares of Table XX are estimates of the following variances:

$$263.3 \to 48\sigma_s^2 + 8\sigma_h^2 + \sigma_r^2$$
$$38.5 \to 8\sigma_h^2 + \sigma_r^2$$
$$27.8 \to \sigma_r^2$$

where the arrow sign $\to$ means "estimates" and the multipliers 48 and 8 are the numbers of articles per mean. This can be fairly easily shown by an extension of the arguments on which corrected variances were derived in the last chapter. It is easy to solve these equations and obtain the estimates of the corrected variances σ_s^2, σ_h^2, and σ_r^2, and these may be used in any of the ways illustrated in the last chapter. Here the estimate of σ_s^2 is almost valueless since there is only 1 degree of freedom for sides; but, if the results for several machines could be combined, a better estimate would result. In many investigations, of course, there will be more than two members of the second factor.

It is well to emphasise the assumptions that underlie this analysis; we have encountered their kind before. We are assuming that the difference between sides is, on the average, the same for all heads on each side, and that there is no other effect of sides. Thus differences in the state of wear of the side pinions could affect the variability of the articles produced on a side; we are assuming that this is not so and that the within-side variations are the same for both sides. Corresponding assumptions are made for the head effects.

Table XIX may be looked at in a different way; it may be regarded as giving 2 sets of 48 results to which the t test might be applied to measure the side effect. In applying this we would assume the variation within each side to be random; the analysis of Table XX tests this by investigating the significance of the head effect. Were this significance to be established, it would be necessary to regard the results for each head as forming a sub-sub-group or a cluster and to apply the t test to the 2 sets of 6 means of Table XXI. Incidentally, when this is done, t is found to be $\sqrt{263.3/38.5} = 2.62$, the degrees of freedom are

10, and the probability level of significance must be the same as that for the corresponding F.

It is worth while comparing Tables XI (p. 113) and XIX in order to see the difference between the two-factor basic and composite forms. The super-imposition of factors one on another can be extended indefinitely. For example, if data like those of Table XIX were available for several machines, and if there was nothing common to the left sides of the machines, and to the right sides (i.e., if the side pinions were combined at random), the form would be of three factors super-imposed on one another in single-factor basic forms. This may be set out diagrammatically as follows:

etc. for several machines
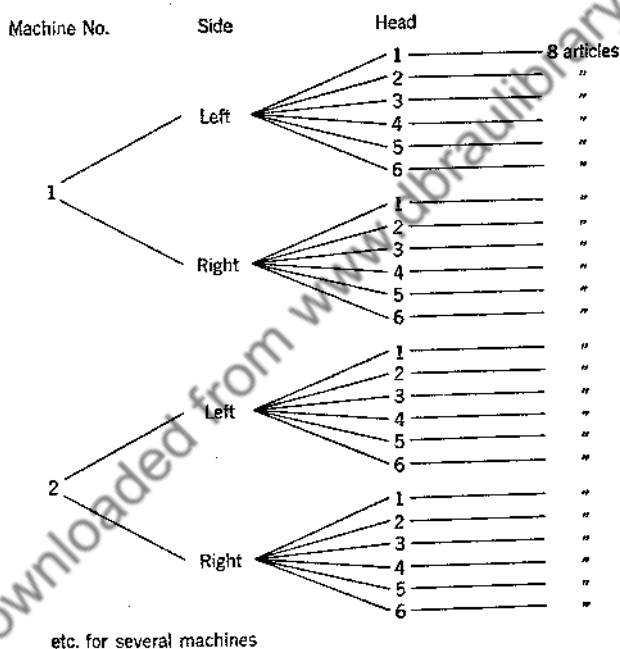
Three-Factor Composite Forms

One three-factor composite form is that illustrated at the end of the last section; you may call it the single-factor on single-factor on single-factor form if you wish.

In a systematic development of the subject the next three-factor composite form would probably be the single-factor on two-factor form, such as would result were there several blocks of results like

those of Tables XIII (p. 116) and XV (p. 120), with the block means themselves constituting a single-factor form and with nothing common between the factors for the blocks. In Table XIII, for example, this could occur for ingots, since each block would necessarily refer to different ingots, but it could not occur for orders, since each block would contain the same 4 orders. I cannot remember ever coming across the pure single-factor on two-factor composite form, and can give no example here.

The two-factor on single-factor form is fairly common; it is exemplified by Table VIII (p. 106). The yarn totals may be arranged in the way shown in Table XXII, which is of the two-factor form. Previously means rather than totals have been treated for ease of exposition; but the computation is easier if totals are used, and as the analysis becomes more complex this consideration becomes more important. It will be assumed that you understand the analysis of variance enough to permit the use of the totals in all subsequent analyses in this chapter. The computations will not be explained, but please notice that, where in treating means a sum of squares would be multiplied by a number (the number of observations per mean), in treating totals the sum of squares is divided by the same number.

TABLE XXII

TOTALS OF BREAKAGE RATES FOR YARNS OF TABLE VIII

| | Twist | | | Cotton Totals |
	Low	Medium	High	
Cotton A	401	216	221	838
Cotton B	254	259	169	682
Twist totals	655	475	390	1520

You may remember that the 6 yarns of Table VIII were spun from two cottons, A and B, each with three twists, L, M, and H; and that we considered as one possible technical explanation of the significant yarn differences that cotton A may on the average be worse than cotton B for all twists as shown by the two total breakage rates in the last column of Table XXII, and that the quality may deteriorate steadily

for both yarns as the twist is decreased as shown by the three twist totals of the last row of Table XXII. The further analysis of variance will enable us to examine this possibility. The full analysis is in Table

TABLE XXIII

ANALYSIS OF VARIANCE WARP BREAKAGE RATES

Source of Variation		Sum of Squares		Degrees of Freedom		Mean Square	
Between yarns	Cottons	3487.7	450.7	5	1	698	451
	Twists		2034.2		2		1017
	Residual		1002.8		2		501
Residual (within yarns)		5745.1		48		120	
Total		9232.8		53			

XXIII, where the results on the left of each column are transferred from Table IX. The sums of squares for cottons and twists are

$$450.7 = \frac{838^2 + 682^2}{27} - \frac{1520^2}{54}$$

and

$$2034.2 = \frac{655^2 + 475^2 + 390^2}{18} - \frac{1520^2}{54}$$

The residual mean square of 501 between yarns is due to error plus interaction between cottons and twists. It is greater than the residual within yarns, the value of $F = 4.2$ lying between the values on the 0.05 and 0.01 levels of significance for 2 and 48 degrees of freedom. Thus, either (1) there is an interaction, or (2) the error between yarns is greater than that measured by the variation between warps of the same yarn (the possibility of a set variation which is confounded with the yarn variation is mentioned on p. 110), or (3) both (1) and (2) are operating. Had there been complete replication, factor (2) would be inoperative, and the analysis would have allowed us to investigate the interaction.

With which residual should the cotton and twist mean squares be compared? There can be no question that the residual between yarns must be used, since it possibly measures a source of error not measured

by the residual within yarns. On this basis neither the cotton nor the twist effect is significant, and we are left with the result that the yarns differ, but we do not know with what factors to associate the differences. But suppose that the experiment had been so arranged that the residual within yarns correctly measured the error; what then? Then the excess over 120 of the mean square of 501 would be evidence of interaction and not error; and no ordinarily useful technical purpose would be served by the analysis in terms of the two main effects plus interaction. The usefulness of the analysis in these terms is that, if there proved to be no significant interaction, there would be a justified simplification in presenting the results as the two cotton means and the three twist means measuring the two main effects. Then the error mean square to be used in testing the significance of the main effects would be the residual within yarns (except for the considerations raised on p. 110).

Multi-factor Composite Forms

No useful purpose would be served by attempting to extend further the classification of the composite forms or to illustrate each one. As the number of factors increases, the number of types and the complication of the analysis increases enormously, and experience is necessary before one can move with any confidence in this field. Moreover, although the very complex forms are sometimes used and occasionally may be useful, they are not likely to have a very wide application in industry. The difficulty of checking the assumptions and of making the results of an analysis mean something tangible to the technical man usually makes it preferable to break down a complex field into relatively simple parts and to investigate the parts separately, rather than to combine everything into an omnibus analysis.

Nevertheless complex arrangements are sometimes desirable or inevitable, and the following two examples are of interest.

The first involves a further analysis of Table XV (p. 120) giving the specific gravities of bricks. For each kiln the zones are in three layers: 1, 2, and 3; and in each layer are 5 places: the right (R), left (L), near the wicket (W), the centre (C), and the back (B); so that the 15 zones are the combination of two main factors, layers and places, and the zone totals may be arranged in the two-factor form of Table XXIV. The results of the almost complete analysis are in Table XXV, where the results of Table XVI are given to the left of each column. The other sums of squares are

$$214.05 = \frac{122^2 + 125^2 + \cdots}{24} - \frac{658^2}{120}$$

$$56.12 = \frac{234^2 + 181^2 + 243^2}{40} - \frac{658^2}{120}$$

and

83.55 is obtained by subtraction

TABLE XXIV

TOTALS OF SPECIFIC GRAVITIES OF BRICKS FOR ZONES OF TABLE XV

	Place					Layer Totals
	R	L	W	C	B	
Layer 1	42	48	74	23	47	234
Layer 2	36	32	55	26	32	181
Layer 3	44	45	65	55	34	243
Place totals	122	125	194	104	113	658

TABLE XXV

ANALYSIS OF VARIANCE OF SPECIFIC GRAVITY OF BRICKS
(UNITS OF SPECIFIC GRAVITY MULTIPLIED BY 200)

Source of Variation	Sum of Squares		Degrees of Freedom		Mean Square	
Kilns	20.77		7		3.0	
Zones { Places	353.72	214.05	14 { 4		25.3 { 53.5	
Layers		56.12	{ 2		{ 28.1	
Residual I		83.55	{ 8		{ 10.4	
Residual II	679.48		98		6.9	
Total	1053.97		119			

The residual I contains residual II plus the effect of the interaction between places and layers, but this interaction is not statistically significant ($F = 10.4/6.9 = 1.5$ is well below the 0.05 level for 8 and 98 degrees). Here there is no error, and the place and layer mean squares must be tested against that for residual I. Only the place effect is statistically significant, but, of course, we can not say that there is no layer effect. If, for the sake of argument, we ignore questions of significance and treat all the effects as statistically significant, we may calculate the corrected variances as follows:

Kilns	Zero
Places (main effect)	$(53.5 - 10.4)/24 = 1.8$
Layers (main effect)	$(28.1 - 10.4)/40 = 0.4$
Place $\times$ layer interaction	$(10.4 - 6.9)/8 = 0.4$
Kiln $\times$ zone interaction + random variation within zones	6.9

These may be used in the ways illustrated in the previous chapter to measure the importance of the various effects as contributing to the variability of the bricks.

It was stated above that Table XXV presents an analysis that is almost complete; for completeness the residual II must be split up. For each kiln of Table XV we can find 5 place totals and the 40 totals can be put into the two-factor form and the sum of squares be analysed into parts associated with places (4 degrees of freedom), kilns (7 degrees), and places $\times$ kilns interaction ($39 - 4 - 7 = 28$ degrees). Similarly we can find for each kiln 3 layer totals and analyse the sum of squares into parts associated with layers (2 degrees), kilns (7 degrees), and layers $\times$ kilns interaction ($23 - 2 - 7 = 14$ degrees). The results for places, layers, and kilns are already entered in Table XXV; the two interactions form parts of residual II, the remaining part based on $98 - 28 - 14 = 56$ degrees of freedom being associated with the places $\times$ layers $\times$ kilns *second-order interaction*. This second-order interaction measures the variation in the places $\times$ layers interaction from kiln to kiln, or the variation in the places $\times$ kilns interaction from layer to layer, or the variation in the layers $\times$ kilns interaction from place to place. It is a complicated thing to think about theoretically, it can not easily be given any great technical significance, and its existence is almost unbelievable on technical grounds, when there is no main kiln effect (although such existence is theoretically possible). Therefore it will not be profitable here to pursue the actual analysis so far.

TABLE XXVI

WARP BREAKAGE RATES

Setting Type I	Gaiting	Setting. Types II and III									Totals
		IIa IIIa	IIa IIIb	IIa IIIc	IIb IIIa	IIb IIIb	IIb IIIc	IIc IIIa	IIc IIIb	IIc IIIc	
Ia	1	5.6	2.7	2.1	0.6	1.8	0.6	3.4	1.4	2.6	20.8
	2	3.4	2.5	2.2	0.6	2.5	0.3	0.8	1.4	1.7	15.4
	3	7.8	2.8	2.0	2.8	0.8	3.1	2.0	2.0	0.6	23.9
	Totals	16.8	8.0	6.3	4.0	5.1	4.0	6.2	4.8	4.9	60.1
Ib	4	2.2	2.2	2.4	4.5	4.8	4.2	2.3	1.1	0.8	24.4
	5	8.4	2.5	5.0	5.3	2.2	0.8	0.6	0.8	1.7	27.3
	6	3.1	1.7	2.5	2.0	1.7	1.1	1.4	1.7	0.6	15.8
	Totals	13.7	6.4	9.9	11.8	8.7	6.1	4.2	3.6	3.1	67.5
Ic	7	1.1	2.5	0.7	2.2	3.1	2.8	1.7	1.4	11.8	27.3
	8	6.4	11.5	2.5	1.4	1.7	1.4	1.7	1.1	0.8	28.5
	9	7.3	5.6	2.8	6.2	7.6	1.7	3.7	4.8	6.3	46.0
	Totals	14.8	19.6	6.0	9.8	12.4	5.9	7.1	7.3	18.9	101.8
Grand totals		45.3	34.0	22.2	25.6	26.2	16.0	17.5	15.7	26.9	229.4

Table XXVI presents the results of a fairly complex form of experiment that has features often encountered in technical experimentation. The general aim was to find the effect on the warp breakage rate in cotton weaving of certain loom settings. There were three types of setting: types I, II, and III; and for each type three values: a, b, and c. Any value of any one type can be combined with any value of each other type, so that the possible combinations gave $3 \times 3 \times 3 = 27$ settings altogether.

Type I settings could not be changed easily; a change involved a structural modification of the loom, and so there were 3 looms, one each for settings Ia, Ib, and Ic. It would have been preferable to have had at least 2 looms for each setting of type I, so that loom differences other than those due to the setting changes could have been brought into the estimate of error. Resources were not available for this, but

care was taken to standardise the looms, and the technicians had confidence that any differences in results between them could be attributed to the differences in the type I setting. An important possible source of error was considered to be the "gaiting" or setting-up of the warp in the loom. This must necessarily be done independently for each loom, and replication to measure this error was secured by having 3 independent gaitings for each loom. The arrangement of this part of the experiment is shown in the first two columns of Table XXVI.

Type II and type III settings could be changed quickly and easily, and without disturbing the gaiting, so that it was convenient to divide the weaving period for each gaiting into 9 parts and to assign them at random to the 9 settings formed by combining types II and III. The full results are given (not in chronological order) in Table XXVI.

Before analysing the results, let us inspect them. They are given to one decimal place and some figures are repeated exactly several times (e.g., 0.6, 1.4, and 1.7), because only a few warp breaks were observed for each result; but the time factor with which they were divided to deduce a rate was often the same. Had this factor been the same for all periods, the raw numbers of warp breaks could have been used in the analysis. Furthermore, the results vary enormously, from 0.3 to 11.8, but there are no grounds for rejecting extreme values as untypical. The highest values in order are 11.8, 11.5, 8.4, 7.8, 7.6; and the lowest are 0.3, 0.6, 0.7, 0.8, etc.; there is no dividing line which would set off one or two extreme values as untypical, and we must accept them all as belonging to the experiment. A considerable variation from a value near a "floor" (which is a breakage rate of zero) often indicates a possible heterogeneity of the error variability; one would expect it to be greater for those settings with a higher mean breakage rate. Knowledge gained in other experiments on warp breaks suggests that such an effect would be eliminated by finding the square root of each result in Table XXVI and performing the analysis in terms of values so transformed. Such a transformation facilitates the statistical interpretation of the data but it complicates the technical interpretation, and we shall proceed to the less laborious task of analysing the data of Table XXVI in their present form.

First let us deal with the gaiting totals in the last column of Table XXVI. They are unaffected by the changes in settings II and III, since all those settings are equally represented in each; and the variations between the 3 gaiting totals for each setting of type I include all the errors that affect the type I comparisons. The analysis of variance,

which is of the one-factor form, is in section A of Table XXVII, the sums of squares being

$$36.67 = \frac{60.1^2 + \cdots}{27} - \frac{229.4^2}{81}$$

$$36.41 = \frac{20.8^2 + \cdots}{9} - \frac{60.1^2 + \cdots}{27}$$

The value of F for testing the type I setting effect is $18.3/6.2 = 3.0$ and is well below the 0.05 level of significance for 2 and 6 degrees of freedom. There is no evidence that the variations in the settings of type I affect warp breakage rate.

The full results in Table XXVI for each setting of type I form a block in the two-factor form, and the sources of variation are settings II and III (8 degrees), gaitings (2 degrees), and a residual (16 degrees). The sums of squares and degrees of freedom may be added for the three blocks to give the results to the left of the columns of section B of Table XXVII. The sums of squares are

$$158.04 = \frac{16.8^2 + \cdots + 13.7^2 + \cdots}{3} - \frac{60.1^2 + \cdots}{27}$$

$$36.41 = \frac{20.8^2 + \cdots + 24.4^2 + \cdots}{9} - \frac{60.1^2 + \cdots}{27}$$

$$402.65 = (5.6^2 + \cdots + 2.2^2 + \cdots) - \frac{60.1^2 + \cdots}{27}$$

Variations associated with gaiting changes do not affect the comparisons between the settings of types II and III and are "taken out" in row (6) of Table XXVII, which is the same as row (2). The excess of residual G over residual W suggests at first sight that the disturbance due to re-gaitings does add to the errors of the comparisons for type I settings, but the difference is not significant ($F = 6.2/4.3 = 1.4$ lies well below the 0.05 level). Nevertheless, in view of the remarks on page 126 about the pooling of estimates of variance, it is well to use residual W for testing the effect of settings II and III. We have $F = 6.6/4.3 = 1.5$, and this is below the 0.05 level of significance.

We may, however, proceed further. The three rows of totals and that of grand totals in Table XXVI form a table in the two-factor form, with sources of variation as follows: settings II and III (the main effect common to all blocks, with 8 degrees), settings I (2 de-

TABLE XXVII

ANALYSIS OF VARIANCE OF WARP BREAKAGE RATES OF TABLE XXVI

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
A. *Between gaitings*			
(1) Settings I	36.67	2	18.3
(2) Residual G	36.41	6	6.1
(3) Total	73.08	8	
B. *Within gaitings*			
(4) Settings II and III — Settings II and III (main effect)	80.88	8	10.1
(5) and III — Interaction I × II and III	158.04 { 77.16	24 { 16	6.6 { 4.8
(6) Gaitings ($=$ residual G)	36.41	6	6.1
(7) Residual W	208.20	48	4.3
(8) Total	402.65	78	
C. *Between settings II and III*			
(9) Settings II	35.92	2	18.0
(10) Settings III	10.07	2	5.0
(11) Interaction II × III	34.89	4	8.7
(12) Total	80.88	8	

grees), and the interaction between settings I on the one hand and settings II and III on the other (16 degrees). The figures for the first and third of these are in rows (4) and (5) of Table XXVII. The virtual equality of 4.8 with 4.3 shows that the corresponding interaction has no statistical significance, but the main effect of settings II and III is now probably real ($F = 10.1/4.3 = 2.4$, and it lies between the 0.05 and 0.01 points for 8 and 48 degrees). This result is an example of a not infrequent experience, where the significance of an effect is, so to speak, "diluted" by being combined with one that is insignificant.

Finally, we may put the nine grand totals of Table XXVI in the two-factor form with sources of variation: settings II (2 degrees), settings III (2 degrees), and the interaction between settings II and

III. The results are in section C of Table XXVII, and each effect must be tested against residual W. The only effect that is significant is that of settings II ($F = 4.2$ lies between the 0.05 and 0.01 levels).

Thus the results of the experiment "boil down" to a statement that the settings of type II probably have an effect on the warp breakage rate, the means being

$$\text{Setting II}a \qquad \frac{101.5}{27} = 3.8$$

$$\text{Setting II}b \qquad \frac{67.8}{27} = 2.5$$

$$\text{Setting II}c \qquad \frac{60.1}{27} = 2.2$$

Since there are 27 readings per mean, the standard error of the difference between any 2 is $\sqrt{2 \times 4.3/27} = 0.56$. The difference between the means for settings IIb and IIc is small compared with this, and the effect is due to setting IIa giving a higher breakage rate than the other two.

Had the interaction in row (5) of Table XXVII been significant, it would have been worth while carrying the analysis further in order to separate out the first-order interactions, settings I $\times$ II and settings I $\times$ III, and the second-order interaction settings I $\times$ II $\times$ III. Even with an insignificant value for this interaction there is a possibility of one of these effects being significant, but it is remote, and it is not worth while carrying the analysis further.

When we look back over the results we see that there are as many results for each of settings I as for each of settings II, and that the differences between the three means for the two types of setting are about the same. [See the mean squares in rows (1) and (9) of Table XXVII.] Nevertheless the main effect of settings II is statistically significant and that of settings I is not, because the comparisons of settings I are possibly subject to a greater error than those of settings II, and moreover the greater error is estimated on fewer degrees of freedom. The difference between the two errors (residual G and residual W) is not statistically significant, but we are not justified in assuming that there is no difference.

From a technical point of view, the results of the experiment are meagre and disappointing. The experiment can be regarded as little more than exploratory, giving information on which a further one can

be planned. But the results illustrate a number of statistical points, and that is the reason for treating them here.

Incomplete Forms

All the basic forms except that with one factor are squares, in the sense that the number of representatives of each factor is the same, and this circumstance simplifies the analysis enormously. It is possible, however, though difficult, to deal with some incomplete forms such as the so-called quasi-Latin squares and Youden squares, which have more rows than columns. Sometimes, too, something goes wrong with an experiment arranged in some basic form, and one or two results are missing; the analysis can nevertheless be performed. The existence of these possibilities can only be mentioned here; you should refer to the text-books for full information.

Chapter 12. APPLICATIONS OF CORRELATION ANALYSIS

Correlation analysis is applied when two or more qualities are measured on each individual or unit and it is required to take account of the relationships between them. In their full development the methods form a considerable subject, and their applications are varied; the subject can only be introduced here and a few of the most typical applications illustrated.

Correlation of Two Variables

In Table XXVIII (taken from *Statistical Methods in Industry*) are given for 100 casts of steel the percentage of iron in the form of pig iron and the lime consumption in hundredweight per cast; the order in which the casts were made is regarded as having no significance, and the results are given in ascending order of percentage of pig iron. First we plot the two variables, as in Fig. 18; we notice a general tendency for the lime consumption to increase with the percentage of pig iron (a tendency well known to steel makers) and a considerable scatter, so that for any one percentage there is a wide variation in the lime consumption. The tendency is described as a *correlation* between the two variables, and a diagram like Fig. 18 is termed a *correlation* or *scatter diagram*.

One way of describing the "pig iron effect" on the lime consumption is to divide the casts into groups, each with one percentage of pig iron or a narrow range of percentages, and to regard the variation in lime consumption within each group as substantially random; the results of Table XXVIII are divided into 14 such groups by horizontal lines. Then we may find the mean lime consumption for each group, study the variations in the means, and perform an analysis of variance in order to test the significance of the effect and estimate the residual variance. The mean lime consumption is plotted against the mean percentage of pig iron for each group as a circle in Fig. 18, and the circles show the same kind of trend as the individual results except that, because of the effect of the averaging, the scatter of the circles is less than that of the points.

The analysis of variance is in the left-hand part of the columns of Table XXIX. Since the number of casts is not the same for all groups,

145
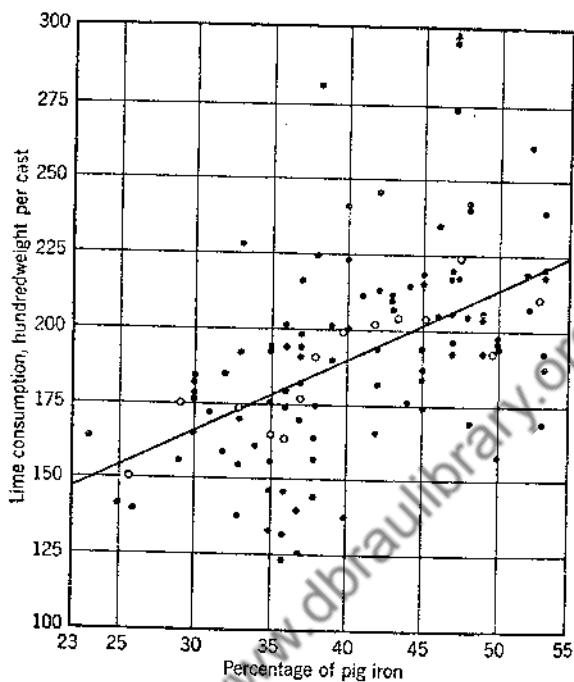
FIG. 18.

the calculations must be performed somewhat in the same way as those for the analysis of Table XI (p. 113). The value of F for testing the significance of the between-group effect is $3222/847 = 3.8$, and it lies a little above the 0.001 level of significance for 13 and 86 degrees of freedom. It will be noted that in this analysis the values of the percentage of pig iron merely provide a basis for dividing the casts into groups; we make no use of the fact that the percentage values order and space the groups.

When we examine the percentage pig iron effect as disclosed by the group means plotted in Fig. 18, the only feature that seems to have any technical meaning is the trend which, as far as the data go, may be represented by a straight line. The deviations of the points from such a line seem to be purely sporadic, and we are prepared to believe that they are primarily due to the residual variation of the individual values of lime consumption about the group means and to the fact that each point is the mean of only a few values. Not only does the representation of the pig iron effect by a straight line simplify the

TABLE XXVIII

PERCENTAGE OF PIG IRON AND LIME CONSUMPTION (HUNDREDWEIGHT PER CAST) IN STEEL MAKING

Pig	Lime	Pig	Lime	Pig	Lime	Pig	Lime	Pig	Lime
23	164	35	156	38	145	43	212	48	205
25	141	35	165	38	157	44	176	48	241
26	140	35	176	38	164	44	215	48	242
29	156	35	193	38	175				
		35	194	38	225	45	174	49	193
30	165			38	281	45	184	49	204
30	177	36	124			45	187	49	206
30	178	36	132	39	190	45	194	50	158
30	182	36	146	39	201	45	216	50	195
30	184	36	174	40	138	45	219	50	196
31	172	36	180	40	200	45	219	50	198
32	159	36	195	40	223	45	205		
32	185	36	201	40	241	46	235	52	208
								52	219
33	138	37	126	41	212	47	193	52	262
33	155	37	140	42	166	47	197	53	170
33	170	37	170	42	182	47	206	53	188
33	192	37	176	42	194	47	218	53	193
33	228	37	182	42	213	47	218	53	219
34	161	37	191	42	246	47	220	53	240
		37	194			47	274		
35	133	37	198	43	207	47	310		
35	146	37	216	43	210	48	170		

TABLE XXIX

ANALYSIS OF VARIANCE OF LIME CONSUMPTION

Source of Variation		Sum of Squares		Degrees of Freedom		Mean Square	
Between groups	Regression line	41,890.79	29,255.20	13	1	3,222	29,255
	Deviations from regression line		12,635.59		12		1,053
Residual		72,799.05		86		847	
Total		114,689.84		99		1,159	

presentation of the results, but it comes nearer, we believe, to what would be the result were values for many more casts available.

The most suitable straight line is that determined by the so-called *method of least squares* and may be written

$$(\hat{Y} - \overline{Y}) = a(X - \overline{X}) \tag{16}$$

where X is any given value of the independent variable (percentage pig iron in this case)

$\overline{X}$ is the mean of the actual values of X

$\hat{Y}$ is the value of the dependent variable (lime consumption), given by the straight line for the given value of X, and is to be distinguished from any actual value of Y

$\overline{Y}$ is the mean of the actual values of Y

a is a constant known as the *regression coefficient* and is calculated from the formula

$$a = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{\Sigma(X - \overline{X})^2} \tag{17}$$

where Σ is the same sign of summation as was used in equation (1) (p. 7). The regression coefficient may be calculated by finding the deviation of every value of X and Y from its corresponding mean and summing the squares and products; but there are short-cut methods for which you should refer to the text-books. The regression line goes through the point $(\overline{X}, \overline{Y})$ and has the slope a.

For the data of Table XXVIII,

$$\overline{X} = 40.54 \qquad\qquad \overline{Y} = 191.04$$
$$\Sigma(X - \overline{X})(Y - \overline{Y}) = 12,191.84 \qquad \Sigma(X - \overline{X})^2 = 5080.84$$
$$a = 2.39957 \quad (\text{say } 2.40)$$

Now we are in a position to find the straight-line value $\hat{Y}$ for every value of X, and hence the deviation $(\hat{Y} - \overline{Y})$, and on squaring and summing these we find the sum of squares in Table XXIX attributed to the regression line—29255.20. Again there are short-cut methods for obtaining this value. The sum of squares attributed to the deviations of the group means from the corresponding regression line values, 12635.59, may be obtained by difference, or by squaring and summing $(Y - \hat{Y})$. The corresponding degrees of freedom are given in Table XXIX, and they have been calculated according to statistical theory so that the mean squares may be interpreted as follows. If the variations in lime consumption are purely random and unrelated to the percentage of pig

iron, all four mean squares in Table XXIX will be statistically equal (in the sense defined on p. 130). We have already seen that there is a significant pig iron effect. If the pig iron effect is adequately described by a linear trend, so that the deviations from the regression line may reasonably be attributed to the residual variations, the corresponding mean squares, 1053 and 847, will be statistically equal, as indeed they are. In such circumstances, the pig iron effect is entirely associated with the 1 degree of freedom belonging to the regression line, instead of being divided over 13 degrees when groups are used, and is much more significant. The ratio F is $29255/847 = 34.5$, and is about three times the value on the 0.001 level for 1 and 86 degrees. Thus we have tested the trend for linearity, and, by using the values of percentage pig iron and taking account of the fact that the effect is a linear trend, we have enhanced its statistical significance. The analysis of variance based on grouping would give the same sums of squares, and hence significance, whatever the order of the groups; it is in accordance with commonsense that a set of means following a simple trend should be more significant than the same means varying randomly. The correlation procedure may be specially useful where the grouped results fail to give statistical significance.

The regression equation itself may be useful. It states what, on the average, will be the lime consumption for a given percentage of pig iron. It does not give a very good prediction for a single cast, but the average predicted consumption for a number of casts might, for example, be compared with the average actual consumption in order to provide an index of operating efficiency for a furnace, or source of pig iron, scrap iron, or lime, or whatever unit of operation the technician thinks it worth studying.

Another use is to correct a comparison of mean lime consumption between two sets of results for differences in the pig iron. For example, the 100 casts of Table XXVIII were produced "without slag control" and the mean lime consumption is 191.04 hundredweight per cast for a mean percentage of pig iron of 40.54. Another set of 100 casts was produced "with slag control," and the corresponding means are 152.4 hundredweight at 42.13 per cent of pig iron. In order to measure the effect of slag control alone we need to correct the lime consumptions to the same percentage of pig iron; let it be 42.13. Then the corrected mean lime consumption without slag control, obtained with the aid of the regression equation, is

$$191.04 + 2.40(42.13 - 40.54) = 194.86$$

The correction in this instance is quite unimportant. When the two series have different regression slopes, the correction will depend on the value of the independent variable at which it is chosen to make the comparison.

The regression coefficient may also have other uses. For example, if scrap iron (the alternative to pig iron) is dearer than pig iron, the cost of the extra lime is partly offset by a cheapening of the total iron when the percentage of pig iron is increased; from the regression coefficient and the costs it is easy to calculate at what relative costs for pig and scrap iron the two effects balance, and there is a change in preference from pig to scrap iron. (In this example it is assumed that, within limits at least, the relative costs of the two types of iron and lime alone matter; whether or not this is so does not matter: the aim has been merely to illustrate a use of the regression coefficient.)

The regression coefficient and values predicted from the regression line are subject to sampling errors, of which account must be taken if consistent conclusions are to be reached. These can not be dealt with here, but the regression coefficient is subject to a large standard error unless either the residual mean square is small or the number of observations is large.

In Table XXVIII it is easy for the technician to choose the percentage of pig iron as the independent, and the lime consumption as the dependent, variable; variations in the former cause variations in the latter. Sometimes, however, the choice is not so clear; and there are two regression lines according to which variable is regarded as independent. This leads to a difficult kind of situation which can not be treated here.

So far we have focussed attention on the relationship between lime consumption and the percentage of pig iron as expressed by the regression line; now let us consider the relative importance of this relationship. The points in Fig. 18 are widely scattered, and, although their general drift is quite apparent, it is clear that there are factors other than the percentage of pig iron that are having an important effect on the variation of lime consumption. When there is a general tendency for two factors to be related in this way, with variations additional to that contained in the relationship, the factors are said to be *correlated*. Scatter diagrams for different pairs of factors can vary from those in which the points fall almost on a line and there is little scatter—the sort of diagram one would get by measuring fairly accurately the diameters and circumferences of a number of circular rods

and plotting the two measurements—through those like Fig. 18 for which both the relationship and the scatter are clearly apparent, to those in which the scatter is substantially random and no relationship or correlation is apparent. Such diagrams show differences in the *degree* or *strength* of the correlation, in the importance of the independent variable as compared with all the other factors producing the scatter. This is measured by the *correlation coefficient*, denoted by the symbol *r*, and calculated from the formula

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2}} \tag{18}$$

The correlation coefficient is a dimensionless ratio. It is zero if there is no relationship between the two variables and unity if there is no scatter from the straight line. It can be positive or negative, but the sign merely shows whether the two variables tend to increase together, or whether one increases as the other decreases. For Fig. 18 the correlation coefficient is +0.505.

The correlation coefficient is useful as a shorthand description of the strength of correlation, but experience is required to appreciate it. The strength of correlation and its statistical significance are quite different things. In Fig. 18 the correlation is not very strong, but it is overwhelmingly significant.

Another instructive way of looking at a correlation analysis is to consider the variability. In arriving at Table XXIX the casts were divided into groups, partly to help the exposition and partly to test the assumption of the straight-line relationship. This is not usually done; linear correlation is taken for granted, and then the analysis of variance divides the sum of squares into two parts, one associated with the regression line (1 degree of freedom) and the other with the residual deviations from the regression (98 degrees in Table XXIX); the second gives a mean square of 872. The total variability of the lime consumption is measured by a mean square of 1159, giving a standard deviation of 34.1 hundredweight per cast; if the percentage of pig iron is kept constant, the variability is reduced to the residual value, with a standard deviation of 29.5 hundredweight per cast. This reduction in variability is a measure of the importance of the effect of the variations in the percentage of pig iron. Looked at in this way, the effect of the variations in the percentage of pig iron is not very important.

L

The correlation coefficient ties up with this reduction in variability by the approximate formula

$$\frac{\text{Residual standard deviation}}{\text{Total standard deviation}} = \sqrt{1 - r^2} \tag{19}$$

The quantity $\sqrt{1 - r^2}$ is probably a more readily appreciated measure of strength of correlation than r.

The results of a correlation analysis may be useful when comparing two series for variability. Thus the two sets of steel made with and without slag control gave the results of Table XXX. From the first row of Table XXX it appears that slag control has reduced the variability of lime consumption. But the second row shows that there was less variability in percentage pig iron for the casts with slag control than for those without, and, since lime consumption is correlated with percentage pig iron, this would produce some reduction in variability of lime consumption. We may eliminate this effect by using the correlation coefficients and equation (19) to calculate the standard deviation of lime consumption for constant percentage of pig iron; the results in the last row of Table XXX show that slag control has in fact reduced the variability.

TABLE XXX

RESULTS OF TESTS ON STEEL CASTS WITH AND WITHOUT SLAG CONTROL

	Without Slag Control	With Slag Control
Standard deviation of lime consumption, hundredweight per cast	34.1	21.7
Standard deviation of percentage pig iron	7.16	4.42
Correlation coefficient	0.505	0.62
Standard deviation of lime consumption (pig iron constant)	29.5	17.0

Multiple Regression

The methods of correlation analysis can be extended to cover the case of more than two variables; the following example illustrates a practical application and also shows what the methods can do.

In the weaving of cloth it is important that the weft (or filling)

packages should not disintegrate unduly under the forces of weaving, and in order to control this a measure of the tendency to disintegrate is required. The direct measure is the fraction of packages that disintegrate under standard weaving conditions, but this is very laborious and uneconomical to determine, and it was desired to develop more convenient laboratory tests as a substitute. From a knowledge of the physics of the situation, three different tests were devised which give measures we shall term X_1, X_2 and X_3 respectively. The problem was to discover (a) whether there was any advantage in using all three or two of the measures rather than two or one, (b) if two or one are sufficient, which two or one, (c) for the chosen combination how the values should be combined to give a composite measure, and (d) how closely the resulting measure is related to the direct measure of the tendency to disintegrate.

An experiment was made by preparing 18 lots of packages under such various conditions as are likely to obtain in practice, except that the variations were somewhat exaggerated in order to enhance the various effects and make them more easily apparent. From each lot 30 packages were woven under standard conditions (somewhat more stringent than those obtaining in practice), and the fraction that disintegrated was recorded. From other packages in the same lots determinations were made of the measures X_1, X_2, and X_3; these are given in the last three columns of Table XXXI. When the fraction disintegrated was plotted separately against each of the other variables, the scatter diagram was obviously non-linear and the points were not uniformly scattered. Statistical theory suggested that it might be better to use, instead of the fraction, the transformed variable Y, where

$$Y = \sin^{-1}\sqrt{\text{fraction disintegrated}} \quad \text{in degrees}$$

The values of Y are given in Table XXXI and plotted against X_1, X_2, and X_3 in Fig. 19. In the scatter diagrams the points for Y against X_2 and X_3 are fairly uniformly distributed about imaginary straight lines; those for Y against X_1 show signs of a slight curvature in the relationship, but on the whole the transformation is reasonably satisfactory.

First we must choose the best single variable. The three regression equations are calculated to be

$$(\bar{Y} - 36.3) = -8.057(X_1 - 6.32)$$
$$(\bar{Y} - 36.3) = -4.057(X_2 - 90.59)$$
$$(\bar{Y} - 36.3) = -3.779(X_3 - 47.17)$$

TABLE XXXI

MEASUREMENTS MADE ON WEFT (OR FILLING) PACKAGES

Y	X_1	X_2	X_3
31	6.2	93.5	50.8
31	6.2	93.1	41.2
21	10.1	95.3	55.4
21	8.4	96.3	53.0
57	2.9	82.9	43.0
80	2.9	80.3	41.5
35	7.4	92.9	47.8
10	7.3	92.6	49.0
0	11.1	96.5	51.4
0	10.7	96.4	53.3
35	4.1	87.2	40.7
63	3.5	82.2	42.5
10	5.0	93.5	45.4
51	4.5	93.9	44.6
24	9.5	95.1	52.9
15	8.5	96.2	55.2
80	2.6	83.6	42.0
90	2.9	79.1	39.4

and the sums of squares of $(\hat{Y} - \overline{Y})$ associated with these three regressions are respectively 9383, 10,642, and 7384. Clearly X_2 is the best single variable. You can judge whether this conclusion would have been reached from a visual examination of Fig. 19.

Now we find which is the better combination of two variables: X_2 with X_1 or X_2 with X_3. We can calculate a *multiple regression equation*
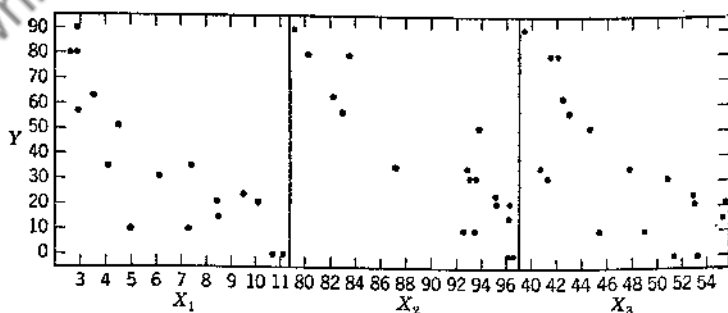
FIG. 19.

for each by the method of least squares. You should refer to the text-books for the details; here are the results:

$$(\bar{Y} - 36.3) = -2.981(X_2 - 90.59) - 2.656(X_1 - 6.32)$$

$$(\bar{Y} - 36.3) = -3.825(X_2 - 90.59) - 0.320(X_3 - 47.17)$$

The sums of squares associated with these two regressions are respectively 10,913 and 10,660, so that the regression on X_2 and X_1 is slightly preferable.

Finally we calculate the multiple regression equation giving $\bar{Y}$ in terms of X_2, X_1, and X_3; it is

$$(\bar{Y} - 36.3) = -3.111(X_2 - 90.59) - 3.777(X_1 - 6.32)$$
$$+ 0.8073(X_3 - 47.17)$$

and the sum of squares associated with the regression is 10,981. The total sum of squares of the deviations of the actual values of Y from their mean is 13,212. The sums of squares are entered in Table XXXII.

TABLE XXXII

ANALYSIS OF VARIANCE OF VALUES OF Y (TABLE XXXI)

Source of Variation	Sum of Squares			Degrees of Freedom	Mean Square
Regression on X_2	10,981	10,913	10,642	1	10,642
X_1			271	1	271
X_3			68	1	68
Residual			2,231	14	159
Total			13,212	17	777

The sum of squares of 10,981 associated with the regression on X_2, X_1, and X_3 arises from 3 degrees of freedom, and the residual of 2231 is obtained by difference. It is the sum of $(Y - \bar{Y})^2$ if the values of $\bar{Y}$ are calculated from the above equation in terms of X_2, X_1, and X_3. The corresponding sum of squares for the regression on X_2 and X_1 is 10,913, based on 2 degrees of freedom, and the difference of 68 represents the additional sum of squares taken into account by including also X_3 in

the equation. Similarly 10,642 is the sum of squares associated with the regression on X_2, and in a sense the difference of 271 may be associated with the variable X_1. The sums of squares to the right of the second column of Table XXXII show step by step by how much the residual is reduced by including additional variables in the equation, and the calculation is so arranged that, if any additional variable effects no real improvement in the formula, the associated mean square is statistically equal to the residual mean square. The mean square for the regression on X_2 is greater, and significantly greater, than the residual, and X_2 at least provides some index of the value of Y for any lot. The mean square for the inclusion of X_1 (271) is larger than the residual, but not significantly so, and there is thus no evidence that the equation involving X_2 and X_1 is any better than that involving X_2 alone. In view of this it is very unlikely that the further inclusion of X_3 would give any improvement, and the fact that the mean square for X_3 is not greater than the residual (it is less, but not significantly so) supports this view. Thus we reach the conclusion that the best single variable for predicting the value of Y is X_2, and that no improved prediction is obtained by using values of X_1 and X_3 in addition. The correlation coefficient between Y and X_2 is 0.90, showing that X_2 is a *good* index of Y. Indeed it is likely that the values of $(Y - \bar{Y})$ obtained with this equation are no greater than can be explained by errors in the experimental determination of Y. Problem (c) on page 153 does not arise, but, had we decided to use two or three variables, they would be combined by means of the appropriate regression equation.

It should be noted that the sums of squares and mean squares for the separate variables X_1, X_2, and X_3 are not independent in the ways that those for the factors dealt with in Chapters 10 and 11 are; they depend on the order in which the variables are taken. Had we, for example, started with X_3, we would have found that an equation including X_3 and X_2 was better than one based on X_3 alone. The variables have been taken in the order which reduces the residual sum of squares the most at each stage; there is no theoretical justification for doing this—it merely seems to be a reasonable thing to do.

Assumptions and Interpretation

The mathematical assumptions behind correlation analysis are that the regression is linear, and that the variability of the residual deviations from the regression line is homogeneous. If the regression is only slightly curved, the linear correlation analysis may be applied as an approximation; if it is markedly curved, the methods of analysis can

be extended accordingly. If the residual variation is not homogeneous, the only possible way out is to make some transformation of the variables.

The correlation analysis is purely mathematical and is not necessarily descriptive of the causal system behind the results. Strictly, the interpretation on page 149 of the regression coefficient of 2.40 relating lime consumption to percentage of pig iron applies only when casts of different percentages of pig iron are selected from the 100 casts of Table XXVIII; it applies to other and subsequent casts only if the same system of causes continues to operate, and it is the technician's rather than the statistician's job to decide whether that is likely to be the case. The steel technologist would probably be able to state under what conditions the regression coefficient of 2.40 is likely to apply; he would specify the steel-making process and possibly the quality of lime and scrap iron used.

Likewise the regression equation for the weft packages (p. 153) would not apply to packages made under conditions not included in those under which the lots of Table XXXI were made.

When several variables are exactly related by a straight-line law, but are subject to errors of measurement, the measurements form scatter diagrams when plotted, but the regression coefficients do not estimate the true relationships. This is a complicated question which can not be dealt with here; you are merely warned that statistics does not often help us to go behind errors of measurement and estimate true physical relationships.

In earlier years the correlation coefficient was much used as evidence of causal relationships, but it is now used in this way only with extreme caution. Very often two quantities are correlated, not because one directly causes another, but because they are both effects of a third cause or complex of causes. Attempts are then made to bring other factors into consideration and to separate their effects by calculating *partial correlation coefficients,* which measure the correlation between pairs of factors when all the other factors included in the analysis are constant. Even these do not describe causal relationships if all relevant factors have not been included in the analysis; and they may mislead because the straight-line law is far too simple a description of a complicated mathematical situation. These are the reasons why correlation analysis has not proved a very powerful tool for disentangling causes and effects in a complex situation. It is chiefly useful for giving quantitative descriptions when the causal pattern underlying the observations is fairly well understood. On the other

hand, occasionally the only available information is data observed under works conditions without any control, and correlation analysis may have to be used *faut de mieux* in order to give preliminary hints as to the important factors.

The methods described in this chapter are based on the method of least squares, and they must be used if significances are to be tested and standard errors of the various estimates specified. It is always a good thing, however, to plot the results in a scatter diagram, and very often an experienced man can do nearly all that is required by relatively simple graphical methods. The more elaborate statistical methods have their uses, but they should not be applied automatically and in all circumstances.

Chapter 13. PLANNING AN INVESTIGATION

In previous chapters we have considered the analysis of various forms of data without giving more than a passing thought to how the data "got that way." In this chapter we shall consider the art of planning an investigation so that the results lead to valid conclusions, reliably and economically.

This planning involves first choosing the field of investigation: the number of factors and the range over which they are to be investigated. Such choice is governed largely by technical rather than statistical considerations, but it will receive some attention later in this chapter. When the field has been chosen, the investigation is designed so that the results fall into a form easily susceptible of statistical analysis and so that the assumptions underlying the analysis are satisfied. Associated with this is a decision on the size of investigation necessary to achieve the required precision, and the choice of a design that will give this precision economically. The word *plan* will be used to refer to the whole procedure of arranging an investigation and the word *design* to refer to the latter more statistical part.

Investigations and *experiments* will be referred to in a specific way. In experiments there is some degree of experimental control of the factors; investigations include experiments and statistical studies of uncontrolled variations. Most of the chapter applies to experiments.

Until comparatively recently, it was common for an experimenter to complete his work without considering the requirements of statistical analysis, and for him to turn to statistics only when the results were in such a mess that he could make little of them. Then he would often find that the statistician also could make little of them. You may say, if you like, that the statistician is not clever enough to analyse data in any form whatever. There are fields of investigation, notably in agricultural experimentation, where virtually no conclusions can be reached without statistical analysis; and the commerce of ideas between experimenters and statisticians in these fields has led to the development of a statistical science of experimental design, and to a regular habit of consulting the principles of this science before embarking on any investigation. The science is most highly developed for application to agricultural and related experimentation, and the

application to technological investigation is still at a fairly rudimentary stage. The same general principles are valid whatever the field of application, but the details must vary according to the circumstances. Most of the literature on the subject still applies to agriculture and related fields; it is important to distinguish in it general principles from specific details and to be cautious in applying the latter to the field of industry and technology.

You may think that the claim of the statistician to have a say in the planning of experiments is an arrogant one, but to admit it is merely to extend slightly the general idea of the experimental scientific method. The experimenter aims to set up a relatively simple system of causes and effects so that observations on it provide data susceptible of analysis by the ordinary logic of the scientific method. The statistician merely adds the requirement that the data shall also be susceptible of analysis by known statistical methods. If you can reach conclusions without statistical analysis, there is no need to consider it in planning the investigation; but experience shows that the statistical approach is often helpful even in circumstances where it is not essential.

We shall first illustrate the principles of design, by referring to simple investigations in which a comparison is made between two "treatments" and the standard error forms the basis of the test of significance. The extension to more complex investigations will be considered in a separate section.

Randomisation

When applying the t test to the difference between two means, the "head in the sand" statistician may be content to give a verdict on the statistical significance of the difference, but the scientist wants to know whether or not the factors related to the two series have an effect—whether in Table VI (p. 76) the source of the nitrogen has an effect on its density, or whether the grinding wheels of maker A are more or less economical than those of maker B (p. 94). The assumptions on which such an inference is based are fully discussed in Chapter 9, and most of those of any importance are satisfied if the arrangement satisfies the principle of *randomisation*.

In an investigation where there is no experimental control this principle is satisfied by the adoption of a good sampling technique as discussed on pages 34 to 37, ensuring that no bias affects the difference between the two means, and choosing as statistical individuals clusters that are statistically independent.

Most experiments are done with material that shows some pattern

of variation. There are time trends, observer idiosyncrasies, differences between pieces of apparatus, patterns of variation in the substances used, and so on. The element of randomness necessary to satisfy the statistical assumptions can be supplied by distributing the experimental variations at random over the field. Suppose, for example, that there are to be two experimental treatments, A and B, 10 determinations for each, and 3 observers; and that observer I will make 6 and observers II and III, 7 each. Then the plan of the determinations will be as follows:

Observer I	1,	2,	3,	4,	5,	6	
Observer II	1,	2,	3,	4,	5,	6,	7
Observer III	1,	2,	3,	4,	5,	6,	7

where for each observer the determinations are made in the order 1, 2, 3, etc. The principle of randomisation may be satisfied by having 20 similar cards, writing A on 10 and B on 10, shuffling them, drawing them one at a time, and assigning the treatments A and B in the above plan in the order in which the corresponding cards are drawn. Such an arrangement might be

Observer I	A	B	A	A	B	B	
Observer II	B	A	A	B	B	A	B
Observer III	B	B	A	A	A	B	A

This satisfies the condition of randomness whether or not there are observer differences or trends for the observers, whether they use the same or different batches of raw material, or whether or not some observers change the apparatus or raw material part way through the experiment. Whatever the pattern of any uncontrolled variation, its effect on the comparison between A and B is broken up by the randomising process, and the t test leads to a valid inference about the relative effect of the two treatments, provided that it is merely a simple difference added to the uncontrolled variations. It is a matter partly of technical knowledge to decide for what divisions of the experimental field the experimental conditions can be changed from A to B. Sometimes the change can be made only by using different machines; sometimes the same machine can be changed at different times from A to B and vice-versa (see, for example, the weaving experiment described near Table XXVI, p. 139).

The need for randomisation is fairly easy to understand, but it is sometimes overlooked in the stress of practical life. Suppose, for example, that treatments A and B represent two settings of a machine,

that it takes considerable time and trouble to change the setting, and that the determinations are to be made on one machine consecutively in time. The experimenter will probably almost instinctively reject the arrangement

A A A A A A A A A A B B B B B B B B B B

for the 20 determinations, but he will need more strength of mind to reject the following non-random arrangements:

A A A A A B B B B B B B B B A A A A A A
A A B B A A B B A A B B A A B B A A B B

Such should be adopted only if one can be sure that the variation along the series is itself random. We shall see later, however, that designs of these types can be randomised.

The other desirable feature, which makes the ordinary inference from the t test substantially valid even if the error variation differs for the two series, is that there should be the same number of observations in each. It is a matter of minor convenience in calculating averages that the number should be a multiple of 5 or 10.

Economy

Economy can sometimes be achieved by arranging for some of the uncontrolled variation to affect the two treatments A and B equally, so that it has no effect on the comparison. This is exemplified by the determinations of fat in meat in Table VII (p. 90), where the standard error of the difference between the results of two methods of fat analysis for 20 pairs of determinations was reduced from 2.89 to 0.146 by eliminating the variation between meats from the comparison. The condition of randomness is satisfied by allowing chance (e.g., by the toss of a coin) to decide which portion of each meat shall be allocated to each kind of fat analysis.

It is convenient to express the relative economy of two arrangements by the ratio of the squares of the standard errors of the mean difference for a given number of observations, perhaps multiplied by 100 to give a percentage. It follows from the formula for standard error that this ratio is the same as the ratio of the numbers of observations necessary for a given standard error, and, if the cost of an investigation is proportional to the number of observations, this ratio is the ratio of the costs associated with the two arrangements for a given precision. Thus, for the meats this ratio is $2.89^2 : 0.146^2 = 100 : 0.25$, and 25 pairs of tests

by the paired arrangement give as good precision as 10,000 by the random arrangement.

This is true, however, only when the number of observations is large, say more than 20 per series, and the magnitude of the standard error alone matters. When the number is small, the degrees of freedom on which the error variance is estimated also affects the precision. The "paired" arrangement of an experiment halves these degrees of freedom, and, unless the corresponding reduction in error variance is large enough at least to off-set this, the unrestricted random arrangement is more profitable. Suppose, for example, that there are 8 observations per series; then for the random arrangement there are 14 degrees of freedom and $t = 2.14$ lies on the 0.05 level of significance; for the paired arrangement there are 7 degrees and $t = 2.36$ lies on the 0.05 level. In order to make the same difference between means significant at this level, therefore, the paired arrangement would need to reduce the standard error in the ratio of 2.36 to 2.14, or 1.1 to 1. Only if the reduction were greater would the paired arrangement be preferable. Other numbers of observations and other levels of significance give other results for this kind of calculation.

Sometimes there are variations that can either be treated as random and "averaged out" by doing many tests, or be eliminated by making some "control" determination; and then we want to know which is the more economical. The following experience illustrates this.

It was desired to obtain comparative measures of the "openness" of different lots of cotton fibre, originating from the same bale but having been subjected to different processing treatments. The aim was to assess the effect of the processing treatment. The measurement involved taking from the lot 20 grams of cotton (referred to as a "handful"), putting it into an apparatus, and measuring the resistance to air flow. Each result in the first and fourth columns of Table XXXIII gives the resistance of one handful, in arbitrary units, and each set of 5, 6, or 7 results printed in a group in the table refers to one lot, the seven groups referring to seven lots of cotton.

For any one lot the resistance varies because of variations from one handful to another in openness and other properties of the cotton, and of errors of determination. The variations in resistance may be regarded as random, and their effect may be reduced by taking a mean for each lot, the precision with which this mean estimates the true resistance for the lot being specified by the standard error. A pooled estimate of the error variance for this standard error is the within-lot

variance of the results in the first and fourth columns of Table XXXIII; it is 49.4 units (based on $40 - 7 = 33$ degrees of freedom).

TABLE XXXIII

RESISTANCE OF HANDFULS OF COTTON FIBRE TO AIR FLOW

Resistance	Control	Difference	Resistance	Control	Difference
30	43	13	38	49	11
33	47	14	25	35	10
25	36	11	31	40	9
16	26	10	40	50	10
23	34	11	21	31	10
30	43	13			
27	39	12	26	31	5
			22	31	9
50	59	9	24	34	10
40	45	5	30	39	9
26	32	6	28	37	9
32	38	6			
20	27	7	24	32	8
26	33	7	25	32	7
30	39	9	25	34	9
			40	50	10
41	57	16	22	30	8
30	44	14			
26	40	14	37	40	3
26	41	15	32	40	8
22	37	15	39	44	5
30	47	17	34	46	12
			26	33	7

It is possible, after testing the resistance of a handful, to reduce it to a state of virtually perfect openness by carding and to measure its resistance in that state; results for the handfuls already referred to are in the columns of Table XXXIII headed "Control." Variations in the control values for any one lot are due to errors and to the properties other than openness, and the "Differences" in Table XXXIII are a measure of openness, freed from the effects of the other properties, since these properties are common to the two determinations on each handful, but not from the effects of errors. The mean difference for any lot is a measure of the true openness, and the within-lot variance of the differences may be used to compute the standard error. The pooled

estimate of this variance is 3.07 units. It is much less than that for the crude resistance because variations in "other properties" no longer contribute to the error in the determination of openness. Roughly, in order to obtain the same precision, it requires 16 times as many handfuls if determinations of crude resistance alone are made as if control values are also determined and differences are measured; and since the use of controls in this instance only increases the cost four- or five-fold, it is clearly economical. Had the cost with controls been equal to or slightly more than 16 times that without, the physicist would probably still prefer to use controls; he would feel instinctively that it is better to eliminate the effect of "other properties" than to reduce it by statistical averaging, and I would not attempt to persuade him otherwise. But, had the method with controls been much less economical than that without, its use would be hard to justify.

In Table XXXIII, the simple difference between the two determinations for each handful rather than the percentage or some other index has been used. When the control and resistance values are plotted against each other, they form a chart that does not differ appreciably from the type of Fig. 17(a) (p. 100), and the simple difference is adequate. The adoption of pooled estimates of the error variances is justified on the grounds that there is no obvious objection to it from either an examination of the data or a consideration of the physics of the test, and that slight heterogeneity of variability would not invalidate the conclusions. If there were any interest in investigating the significance of the lot differences in Table XXXIII, a full analysis of variance could be done on the differences. The economy of using controls was made possible only because the first resistance determination and the control test could be done on the same handful.

When sampling is in clusters (e.g., when there are several leas of yarn per cop and many cops, Table V, p. 29), there arises the question of the most economical number of individuals per cluster. If the cost is the same for a given number of individuals irrespective of whether they are in many or few clusters, it is best to have one individual per cluster. But often this is not the case, as the following example will show.

It has been found * that in certain weaving experiments made to determine the mean warp breakage rate of warps of cotton yarn prepared under various conditions, when the total length woven is divided

---

* *Shirley Institute Memoirs*, Vol. 18, 1941, p. 109, or *Journal of the Textile Institute*, Vol. 32, 1941, p. T209.

into "pieces" and "sub-pieces," the error variance of the mean breakage rate per piece is

$$V = \frac{\sigma^2}{n} + 0.13\sigma^2$$

where σ^2 is the variance between sub-pieces within a piece, n is the number of sub-pieces per piece, and $0.13\sigma^2$ is the corrected between-piece variance (see p. 112). This corrected between-piece variance is independent of the number of sub-pieces per piece; the ratio 0.13 was determined empirically. If there are m pieces, and hence mn sub-pieces, per warp, the error variance of the mean breakage rate per warp is

$$\frac{V}{m} = \left(\frac{1}{mn} + \frac{0.13}{m}\right)\sigma^2$$

For a given number mn this error variance is least when m is greatest (i.e., when $n = 1$), and, if there is no difference in cost between sub-pieces from the same piece and those from different pieces, this is the most economical arrangement. Between each piece, however, certain machine changes have to be made such that, if w is the cost of weaving a sub-piece, the cost of making the changes is iw. Then the total cost of weaving the mn sub-pieces and making m changes (counting the setting-up for the first piece as equivalent to a change) is

$$C = mnw + miw$$

The problem now is to determine the value of n which makes V/m a minimum for a given C, or C a minimum for a given V/m, or, in other words, which minimises

$$\frac{VC}{m} = w\sigma^2(n + i)\left(\frac{1}{n} + 0.13\right)$$

This may be found by differentiating the above expression with respect to n and equating the result to zero, and thus it is found that

$$n^2 = \frac{i}{0.13}$$

For one particular set of warps, i happens to be about 20 so that the most economical value of n is $\sqrt{154} = 12$, approximately.

This example can easily be generalised. The particular result depends on the ratio of the corrected between-cluster variance to that within clusters (0.13 in the example), and the cost of moving from one

cluster to another compared with the cost of taking an extra individual from the same cluster (i in the example). If VC/m is calculated for various values of n in the neighbourhood of the most economical value, it will usually be found to increase only a little from the minimum, so that it is not important to keep very close to the optimum value. Calculations of the kind illustrated serve to show only roughly what the arrangement should be.

The following is a fictitious example that illustrates some interesting points. Let us suppose that the data in Table XXXIV, taken in order when reading along the rows, represent consecutive measurements of the quality of a product made during a trial run. They may be some quality of consecutive articles made by a machine or the average quality of consecutive batches; they may be the quality of consecutive casts of steel or some other manufactured substance (as opposed to articles) that is divided into lots; they may be consecutive intermittent readings of some instrument that indicates some state (e.g., the temperature) of a continuous process; or they may be readings of quality that are contiguous in space, as when taken along a wire or rod. If they are plotted against the measurement number, trends will be ob-

TABLE XXXIV

27	15	16	15	14	22	21	26	20	30	38	34
30	41	21	18	30	27	31	28	17	22	24	25
21	32	29	28	28	31	33	40	28	17	24	29
26	28	37	29	23	24	25	23	29	24	19	23
29	23	28	20	19	22	29	21	14	22	33	31
22	34	18	28	33	25	22	23	27	20	21	20
18	13	30	38	20	21	18	40	29	27	17	16
21	23	25	29	34	30	22	19	17	25	24	30

vious, showing that the process is out of control. Let us, for the sake of argument, accept this situation.

Now suppose that we wish to compare the effects of two experimental treatments, A and B, which alter only the average quality, and that the trial data are to be used to decide the best arrangement. We can not perform an actual experiment with the results of Table XXXIV, but we can form some idea of what would happen were an experiment done with another run of a similar production.

The calculations will be described at some length, and you will find it helpful to follow them through in full; but, if you find any difficulty

M

in doing this, you may skip the details and follow the arguments based
on the results of the calculations assembled in Table XXXV.

The most elementary arrangement, which we shall term arrange-
ment 0, is the purely random one. Then, despite the trends, if the
treatments do not differ in their effects the difference between the mean
for A and that for B will be no greater than that the randomising
process—chance—can produce, and the test of significance based on
the standard error leads to correct conclusions. The variance of the
96 readings of Table XXXIV is 39.73, and the error variance (the
square of the standard error) of a difference between two means based
on N_0 results for each treatment is

$$39.73 \left(\frac{1}{N_0} + \frac{1}{N_0} \right) = \frac{79.5}{N_0}$$

Another possible arrangement, arrangement 1, may be arrived at by
dividing the readings into consecutive pairs: 27, 15 | 16, 15 | 14, 22 |,
etc., and ensuring that both treatments appear once in each pair.
The two patterns A B | A B | A B |, etc., and B A | B A |
B A |, etc., do not satisfy the condition of randomness. The effect
of the trends may be to make the mean of the first readings of the pairs
higher or lower than that of the second, and such a bias (which in an-
other run of production would be unknown) would invalidate the test
of significance. The arrangement A B | B A | A B | B A |,
etc., might coincide with some pattern in the trends, but if, for each
pair, chance (perhaps through the toss of a coin) decides whether the
order shall be A B or B A within each pair, the condition of ran-
domness is satisfied for the purpose of testing the significance of the
mean difference. The design for an experiment might turn out to be
A B | A B | B A | A B | B A | B A |, etc.

The significance of the mean difference would be tested by taking
the differences between the members of the pairs, as was done for the
fat analyses of Table VII (p. 90), and the error variance may be
calculated from the differences between consecutive readings in Table
XXXIV. In practice the division into pairs would be either 27, 15 |
16, 15 |, etc., starting with the first reading, or 15, 16 | 15, 14 |, etc.,
starting with the second; but since the data of Table XXXIV are in
fact the result of the same treatment we may calculate the error
variance from all differences: $27 - 15 = 12, 15 - 16 = -1$, etc. There
are 95 differences, and, regarding their true mean as zero, we calculate
the variance by squaring them, adding the squares, and dividing the

sum by 95; the result is 54.12, and the error variance of a mean difference based on N_1 pairs of readings is $54.1/N_1$. For the same number of observations arrangement I gives a substantially lower error variance than arrangement 0 because the large variations between consecutive pairs of readings do not contribute to the error of the comparison made by arrangement I.

But arrangement I involves a change of treatment for every reading, and, if the change is troublesome or costly, the arrangement may be uneconomical. We could divide the series into groups of 4 and within each group distribute the treatments according to either the pattern A A B B or B B A A, thus reducing the proportion of changes to results. To adopt either pattern throughout or to alternate them would violate the condition of randomness, but to allow the toss of a coin to decide which shall be the pattern for each set of 4 would be to arrive at a satisfactory arrangement; let us call it arrangement II.

In order to estimate the error variance we might calculate the means of consecutive pairs: $\frac{1}{2}(27 + 15)$, $\frac{1}{2}(16 + 15)$, etc., and calculate the variance of the consecutive differences between the series of means; but it is more convenient to deal with the totals: $27 + 15 = 42$, $16 + 15 = 31, 36, 47$, etc., and to square the differences: $42 - 31 = 11$, $31 - 36 = -5$, $36 - 47 = -11$, etc., making an adjustment in the final division. A second series of consecutive pairs can be obtained from Table XXXIV: $15 + 16 = 31$, $15 + 14 = 29$, etc., and this leads to a second series of differences. This second series is not independent of the first, and the precision with which the error variance is estimated is not much improved by using both instead of one; but there is some improvement, and since no bias is introduced it is better to use both. There are 93 differences between totals of pairs, their mean square is 186.99, and this must be divided by $2^2 = 4$, because we have dealt with totals instead of pairs, giving 46.75. If in an experiment there are N_2 results for each treatment there are $N_2/2$ pairs of means, and the error variance of the mean difference is $(46.75 \times 2)/N_2 = 93.5/N_2$. This is rather larger than the error variance for arrangement I for the same number of readings because more of the trend enters into the error of the comparisons; indeed it is larger than the variance for arrangement 0; but in some circumstances it may be more economical than either.

In arrangement 0 the probability of a change of treatment between two consecutive readings is $\frac{1}{2}$, and for N_0 readings per treatment ($2N_0$ altogether) there are therefore N_0 changes on the average (counting the initial setting up as half a change). If it costs k times as much to make a change as to obtain a reading once the treatment is established,

the total relative cost of obtaining N_0 readings on each treatment is $(2 + k)N_0$.

In arrangement I there are N_1 pairs, and the probability of a change between the second reading of one pair and the first of the next is $\frac{1}{2}$; within each pair there is the certainty of a change. Thus there are $3N_1/2$ changes on the average (counting the first setting-up as half a change), and the relative cost is $[2 + (3k/2)]N_1$.

In arrangement II there are $N_2/2$ sets of 4 with, on the average, $N_2/4$ changes between sets and $N_2/2$ within, giving $3N_2/4$ changes and a total cost of $[2 + (3k/4)]N_2$.

For comparing the economy of the three arrangements we need to calculate the relative error variances for the same cost; let it be unit cost in terms of the cost to obtain a reading after a treatment has been set up. For any one arrangement the error variance increases in proportion as the number of readings (and hence cost) decreases, so the relative error variance for unit cost is the error variance previously calculated multiplied by the relative cost; these products are in the last column of Table XXXV.

TABLE XXXV

Arrangement	Error Variance	Relative Cost	Relative Error Variance per Unit Cost
0 Random	$\dfrac{79.5}{N_0}$	$(2 + k)N_0$	$79.5(2 + k)$
I $AB \mid BA$ (randomised)	$\dfrac{54.1}{N_1}$	$\left(2 + \dfrac{3k}{2}\right)N_1$	$54.1\left(2 + \dfrac{3k}{2}\right)$
II $AABB \mid BBAA$ (randomised)	$\dfrac{93.5}{N_2}$	$\left(2 + \dfrac{3k}{4}\right)N_2$	$93.5\left(2 + \dfrac{3k}{4}\right)$
III $AAABBB \mid BBBAAA$ (randomised)	$\dfrac{104.2}{N_3}$	$\left(2 + \dfrac{3k}{6}\right)N_3$	$104.2\left(2 + \dfrac{3k}{6}\right)$
IV $AAAABBBB \mid BBBBAAAA$ (randomised)	$\dfrac{105.9}{N_4}$	$\left(2 + \dfrac{3k}{8}\right)N_4$	$105.9\left(2 + \dfrac{3k}{8}\right)$

Now it can be seen that the most economical arrangement depends on the value of k. If $k = 0$ (i.e., if the change of treatment costs nothing), arrangement I is more economical than arrangement 0, 68 readings on I giving as good precision as 100 readings on 0. When $k = 31.8$,

arrangements 0 and I have the same relative error, and for larger values of k arrangement 0 is the more economical. If k is less than 7.2, arrangement I is more economical than arrangement II; if k is greater than 7.2, arrangement II is the more economical. Suppose, for example, that $k = 12$. Then the relative error variances for arrangements 0, I, and II respectively are 1113, 1082, and 1028; and the precision given by 100 readings on the purely random arrangement is given by 97 readings on arrangement I and 92 on II. In such a case, if there are only a few readings, arrangement 0 or I may be preferable as giving more degrees of freedom for the determination of the error variance from the actual experimental data.

In Table XXXV are given results for two further arrangements extended from the lines of those already discussed. It may be interesting and instructive to check the results given there, and to work out relative error variances for different values of k. The changes in the error variance with the pattern of the arrangement in the second column of Table XXXV depend on the form of the trends in the original data of Table XXXIV; the relative costs in the third column are independent of the data and depend only on k.

Another aspect of economy arises when the severity of a test is susceptible of adjustment. For example, the stability of the weft or filling package in weaving is important (i.e., its ability to withstand the forces set up by the accelerations and retardations it undergoes—cf. p. 153), and one measure is the proportion of packages that disintegrate under a given force. At what force should this be measured in order to determine whether two treatments of the package affect stability? If in a preliminary trial with two typical treatments the test is made with several values of the force, it will be found that as the force changes so do (a) the difference in mean stability for the two treatments and (b) the standard error of that difference for a given number of observations. The most economical force is that for which the ratio of (a) to (b) is the largest. It is technically important, of course, that the force shall be in the region of those operating in practice so that the practically important phenomenon is being studied.

Two kinds of knowledge are required to arrive at the most economical arrangements: statistical and technical. The statistical knowledge is the error variance for different arrangements, which may be obtained from preliminary experiments or observations under uniform conditions, although sometimes an actual experiment with two treatments

can give information that provides a guide for arranging further experiments. Technical knowledge can help in suggesting where variation is likely to lie, and what arrangements are worth investigating. The purely technical information is the relative costs of performing different parts of an experiment: of introducing a "control" in Table XXXIII, of changing pieces in the weaving example of page 165, and of changing from treatment A to B in the example connected with Table XXXIV. This information need not be very accurate. Its use is to find the most economical arrangement, and against any economy that is achieved by good design must be set the cost of obtaining the necessary information—to give a complete calculation of cost that includes both factors defeats me! Moreover, the most economical arrangement is not usually much better than those that are near it; little is gained by hitting the mark exactly. In any particular instance it is worth while going into the question thoroughly only if much experimental work is to be done in the same field; but, at the least, a knowledge of the principles of experimental design, plus some experience, plus general knowledge of the field of investigation, usually helps the experimenter at least to avoid adopting the least economical arrangements.

The Number of Observations

The statistics behind the determination of the number of observations or scale of experimentation necessary for a given precision has been dealt with in Chapter 8 (p. 87). Since we can not hope to do better than give rough ideas of the necessary scale, there is no point in attempting to be precise. For most purposes the number of observations should reduce the standard error of the difference to about the maximum error that can be tolerated, divided by 2 or 2.5. If this maximum tolerable error is chosen, and the error variance per observation for the arrangement decided upon is known, it is easy to calculate the number of observations. Thus, for the experiment on the process of Table XXXIV we might decide that the maximum tolerable error in the difference between the means for two treatments is 4.0, so that the standard error should not be more than $4 \div 2.5 = 1.6$. If we choose to follow arrangement II (Table XXXV), the value of N_2 should be $93.5/1.6^2$ (i.e., between 36 and 37). It would be wise to have 40 pairs of observations.

Attempts have been made to make this kind of determination more precise by using a ratio based on the t distribution instead of the simple ratio of 2 or 2.5, and even by using confidence limits for the error

variance when that is estimated from a limited amount of data, but I think that these attempts have not led to improvements of practical significance.

The error variance per observation must be known in advance. A value may be known from previous general experience in the field of enquiry, or one may be obtained from a special investigation. A good method is to do the experiment on a sizeable, but inadequate, scale and use the results to determine the error variance, and hence the necessary number of observations. Then the experiment can be continued until the required number of readings is available in total, and all of them can be used in obtaining the final result. When extended, this practice leads to the application of the "sequential idea" to experimentation, according to which the experiment is done in a number of stages, and at each stage all the data so far obtained are used to determine whether sufficient precision is attained; at the required point the experiment ceases. Theoretical work is being done on this subject, and it is likely that systematic procedures will be developed which can be applied by any experimenter even without complete understanding of the underlying statistical theory.

Multi-treatment Experiments

The following example, which is taken from *Statistical Methods in Industry*, illustrates a series of standard experimental arrangements that are much used when there are more than two treatments.

Table XXXVI gives the yield points of specimens taken from 36 steel discs. There were 6 ingots, and from each ingot the first 6 discs were taken in order, giving the orders 1 to 6 in Table XXXVI. The data are in the two-factor basic form for the analysis of variance; the row of order means shows a tendency for the later orders to give higher yield points, and there are even greater variations between the ingot means, although they follow no pattern. An analysis of the variance shows the order and ingot effects to be statistically significant.

Now let us suppose that we have in mind an experiment on a similar set of discs in which there are six treatments: I, II, III, IV, V, and VI, and that it is technically possible to allocate the treatments to the discs without restriction (the treatments would have to be varied after the discs had been cut from the ingots). We may investigate the errors associated with various arrangements by superimposing six "dummy" treatments on the data of Table XXXVI; the differences between the dummy treatment means will then show the errors of the comparisons.

TABLE XXXVI

Yield Point (Tons per Square Inch)

Ingot	Order of Disc						Mean
	1	2	3	4	5	6	
A	*21.2	21.0	20.0	20.8	20.2	21.2	20.73
B	20.4	20.6	22.0	21.6	22.6	22.8	21.67
C	20.6	20.4	21.2	21.2	21.4	23.6	21.40
D	22.8	22.8	22.0	22.8	23.2	22.8	22.73
E	20.8	20.8	22.8	23.2	22.6	23.2	22.23
F	21.4	20.4	20.6	20.8	22.2	22.2	21.27
Mean	21.20	21.00	21.43	21.73	22.03	22.63	21.67

Randomisation is as important for six treatments as for two, and a purely random arrangement may be made by having 6 similar tickets for each treatment, writing the corresponding treatment number on each ticket, mixing the tickets in a bag and withdrawing them one at a time, entering the treatment numbers in Table XXXVI above the readings in order. One such arrangement is in Table XXXVII, which

TABLE XXXVII

Random Arrangement of Treatments I to VI, and Mean Yield Points

Ingot	Order						Treat-ment	Mean
	1	2	3	4	5	6		
A	I	III	IV	VI	II	VI	I	21.17
B	V	VI	IV	III	III	III	II	22.43
C	IV	I	I	V	VI	II	III	22.40
D	VI	IV	IV	II	II	I	IV	21.37
E	IV	I	V	III	II	III	V	21.40
F	V	V	I	VI	V	II	VI	21.27

should be superimposed on Table XXXVI. Thus, the first disc of ingot A has assigned to it treatment I and the yield point is 21.2; the second

disc from ingot A has treatment III and the yield point is 21.0; and so on. If the yield points for the treatments are collected and the average for each is calculated, the results are as shown in the last two columns of Table XXXVII. The randomising process has ensured that, despite the order and ingot effects, the differences between the treatment means are due to chance, and the usual test based on a single-factor basic analysis of variance (treatments, 5 degrees of freedom; error, 30 degrees) would show that they are statistically insignificant. These differences are an indication of the errors with which any real treatment effects would be estimated, and as a rough measure we may note that the means range from 21.17 to 22.43, giving a largest difference of 1.26 due to errors alone.

All the sources of the variations in Table XXXVI contribute to these errors, but we can eliminate the ingot effect from the comparisons by ensuring that each treatment is associated equally with each ingot. Within each ingot the discs must be distributed at random. This time we have 6 tickets, one for each treatment, and after shuffling them we draw them one at a time and assign the treatments in the order of drawing to the six orders of ingot A; one such draw gave III IV II VI I V. Then the process may be repeated with the same tickets successively for the other ingots. The result of such a set of draws is in Table XXXVIII, and, if this is superimposed on Table XXXVI, the dummy treatment means are as given in the last column of Table XXXVIII. The errors now produce less variation in the dummy treat-

TABLE XXXVIII

RANDOMISED BLOCK ARRANGEMENT OF TREATMENTS I TO VI, AND MEAN YIELD POINTS (TONS PER SQUARE INCH)

Ingot	Order of Disc						Treatment	Mean
	1	2	3	4	5	6		
A	III	IV	II	VI	I	V	I	21.67
B	III	I	II	VI	V	IV	II	21.50
C	VI	V	II	III	I	IV	III	21.30
D	I	III	VI	II	V	IV	IV	22.37
E	III	VI	I	IV	II	V	V	21.87
F	III	II	V	IV	I	VI	VI	21.33

N

ment means, the range being from 21.30 to 22.37, with a largest difference of 1.07. This is called the *randomised block* arrangement, because the experimental material is divided into blocks (in this case sets of 6 discs from the same ingot), and the treatments are distributed at random within the block.

It is possible also to eliminate from the errors the order effect by adopting the *Latin square* arrangement, according to which each treatment occurs once in each row and once in each column, but otherwise at random. This may be achieved by having 6 tickets as before and entering each treatment number as it is drawn, filling up the rows in order, but moving a number on one place (or two if necessary) if it comes into a column already occupied by that treatment. If this procedure is followed, it will usually be found impossible in the later rows to avoid having the same treatment twice in a column, and a little "juggling" may be necessary; but it does not matter much how this is done, and very few adjustments suffice. There are standard ways of arriving at Latin squares, but it is hardly worth while mastering them unless a lot of experiments are to be made. It is easier, and in most industrial experiments good enough, to start with a systematic Latin square in which all the treatments of any one number occur along the diagonals:

$$
\begin{array}{cccccc}
\text{I} & \text{II} & \text{III} & \text{IV} & \text{V} & \text{VI} \\
\text{VI} & \text{I} & \text{II} & \text{III} & \text{IV} & \text{V} \\
\text{V} & \text{VI} & \text{I} & \text{II} & \text{III} & \text{IV} \\
\text{IV} & \text{V} & \text{VI} & \text{I} & \text{II} & \text{III} \\
\text{III} & \text{IV} & \text{V} & \text{VI} & \text{I} & \text{II} \\
\text{II} & \text{III} & \text{IV} & \text{V} & \text{VI} & \text{I}
\end{array}
$$

and to distribute the order numbers randomly among the columns (or rows) and the ingot letters randomly among the rows (or columns). For example, the columns might be allocated to orders 1, 3, 5, 2, 6, and 4 and the rows to ingots B, A, F, C, E, and D, in those sequences; and, when the above arrangement is correspondingly rearranged, the result is as shown in Table XXXIX.

When Table XXXIX is superimposed on Table XXXVI, the treatment means are as shown in the last column of Table XXXIX. It is obvious that their differences are unaffected by order or ingot variations, and the errors of the comparisons are further reduced, the largest difference being $22.03 - 21.40 = 0.63$. In an actual experiment the significance of the treatment means would be tested by a three-factor basic analysis of variance, and such an analysis performed on Tables

TABLE XXXIX

Latin Square Arrangement of Treatments I to VI, and Mean Yield Points
(Tons per Square Inch)

Ingot	Order of Disc						Treat-ment	Mean
	1	2	3	4	5	6		
A	VI	III	I	V	II	IV	1	21.50
B	I	IV	II	VI	III	V	II	22.03
C	IV	I	V	III	VI	II	III	21.63
D	II	V	III	I	IV	VI	IV	21.53
E	III	VI	IV	II	V	I	V	21.93
F	V	II	VI	IV	I	III	VI	21.40

XXXVI and XXXIX would lead to the verdict "not statistically significant."

The above investigation of dummy treatment means may be enlightening to the beginner, but it is not the best statistical way of investigating the relative advantage of the methods. The maximum difference is a poor measure of error, and the particular results depend on the particular arrangements the randomisation happens to have produced for Tables XXXVII to XXXIX. Indeed, some particular Latin square arrangements appear to be worse than particular randomised block or random arrangements; the relative merits of the three types appear only on the average when many experiments are done in the field. We obtain better measures by analysing the variance of Table XXXVI. This is done in Table XL.

For the purely random arrangement the error variance per test is the mean square in the "Total" row of Table XL (here, for once, we are interested in the "Total" mean square); it is 1.10. If the ingot effect is eliminated, the error variance per test is the residual of a single-factor analysis and is the mean square corresponding to residual (1) in Table XL (viz.: 0.78). The residual of a two-factor analysis, residual (2) in Table XL, estimates the error variance per test of a comparison between treatments for the Latin square arrangement of an experiment; it is 0.50. The relative economy of the three arrangements can be expressed by saying that, in order to achieve the precision of comparison given by 100 "repeats" on the purely random arrangement, there need to be 71 "repeats" in randomised blocks and

TABLE XL

ANALYSIS OF VARIANCE OF DATA OF TABLE

Source of Variation		Sum of Squares		Degrees of Freedom		Mean Square	
Ingots		15.365		5		3.07	
Residual (?)	Orders	23.287	10.739	30	5	0.78	2.15
	Residual (2)		12.548		25		0.50
Total		38.652		35		1.10	

45 in Latin squares. These conclusions do not, of course, apply exactly to any new experiments in this field because the estimates of variance in Table XL are themselves subject to error.

It is not essential to do a preliminary trial under uniform conditions in order to arrive at the estimates of error. Many weaving experiments have been done according to the Latin square arrangement of Table XVII (p. 123), and the analyses of variance combined to estimate the loom, period, and residual variances. The combined results showed that the loom effect was worth eliminating but that the period effect was not, and that the randomised block arrangement was preferable. The point is that this piece of information on experimental technique was obtained during the course of actual experiments.

Usually it is not advisable to eliminate an effect unless it is substantial. In an actual experiment, the error variance in the random arrangement of six treatments in 36 places is estimated on 30 degrees of freedom, in the randomised block arrangement on 25 degrees, and in the Latin square arrangement on 20 degrees; and these reductions in degrees of freedom need to be accompanied by sufficient reductions in the error variance to justify the later arrangements. Apart from this, there is usually no objection to these arrangements where they are technically feasible; the Latin square is as easy to organize as the purely random arrangement, unless there are many treatments.

In the random arrangement the number of replicates need not be the same for each treatment. In the randomised block arrangement it may be any number provided it is the same for each treatment; but the block must be large enough to contain one of each treatment. In the Latin square arrangement the number of replicates per treatment

must equal the number of treatments. These limitations sometimes make the randomised block arrangement unsuitable and usually make the Latin square arrangement unsuitable when there are many treatments, say more than six or seven.

For experimentation in the agricultural and related fields, there are additional arrangements that are useful in a variety of special circumstances, especially where there are more treatments than are suitable for the above arrangements. As I have not had experience, direct or indirect, of their application to technical experiments I do not propose to deal with them here. There is scope for some pioneer work in developing further methods for technical experimentation.

One fairly complicated design, which in agriculture would be termed a *split-plot* design, is exemplified in Table XXVI (p. 139). In this experiment, which should be studied again, some of the treatments (the settings of type I) were varied only between large sections (the looms), and the comparisons were subjected to larger errors than those for the other treatments (settings of types II and III), which were varied within the large sections. The appropriateness of this kind of arrangement depends on the technical possibilities, the relative importance of the different sources of variation, and the desirability of making some types of comparison more accurately than others.

All the foregoing discussion concerns the extension to more complicated situations of the simple idea of securing economy in comparing the means of two series by associating the readings in related pairs, as was done in Table VII (p. 90) for the comparison of the methods of analysing the fat in meats. It is very difficult to deal generally with the more complicated matter of taking into account such things as the costs of changing the treatments when there are more than two.

For deciding the number of observations or scale of experimentation necessary for a given precision I can not recommend any other procedure than that previously outlined for two treatments. The design must be chosen and the error variance associated with it be estimated; the largest error that can be tolerated in the comparison between any two treatments must be decided on; and then the number of observations can be calculated in the way described.

Economy in the Analysis of Variance

When in an investigation of naturally occurring variation corrected variances are estimated, a new question of economy arises. For example, in estimating the corrected variances of lea weights of cotton

yarn between and within cops (Table X, p. 111), is it better to have many cops with few leas per cop or vice-versa? The question becomes more complicated when there are two or more factors.

This problem has not been solved, at least not in a form to give an answer of direct practical application. Apart from the possibly greater cost of increasing the number of cops than of increasing the leas per cop, the best arrangement will depend on the relative magnitude of the two variances. Commonsense suggests it as a good rough working rule to have as many cops as practicable, reducing the tests per cop to 2. Likewise, in a two-factor analysis, it is usually better to estimate the variances from a large number of tables with few rows and columns in each.

Choice of Variable

The choice of the variable for expressing the results is largely a matter of technical convenience, for ultimately the results have to find technical application; but statistical questions do arise. Sometimes, when the variable that would be chosen by a technician is used, the important assumption of the uniformity of the error variance does not apply, but it may apply if some mathematical transformation of the data is used. Thus in weaving experiments for measuring warp break-age rates the error variance is proportional to the mean, but, if the square roots of the observed breakage rates are taken and the analysis is performed entirely on them, the error variance is substantially the same for all values of the mean $\sqrt{\text{breakage rate}}$. Other transformations have this effect in other circumstances, and the appropriate trans-formation, if there is one to "do the trick," may be found either em-pirically or from quasi-theoretical considerations.

Where the measured variable is merely an index of some underlying property and is not used quantitatively in technical calculations, a transformation that makes the statistical analysis easier can be used without question. Sometimes the technician does, on reflection, find it easier and more profitable to interpret the results in terms of the trans-formed variable than of the raw figures. But this may not always be so, and then the adoption of the transformation, which involves subordi-nating technical to statistical convenience, is inadvisable. In other words, use transformed variables only with understanding and care.

Planning an Experiment

When an experiment is planned in the fullest sense, a number of questions arise that are not strictly or purely statistical, but on which the statistician usually has to give advice.

One feature of many technical experiments is the number of experimental factors that have to be taken into account. Suppose, for example, that the object is to discover the best quantity of a given size (a size is a mixture of an adhesive, a yarn lubricant, and other ingredients) to put on a cotton warp yarn for minimising the warp breakage rate in weaving. Almost certainly the best quantity will depend on the type of cloth woven, but there will be profit in obtaining a result for even one type that is widely woven in industry, and we may limit the experiment accordingly. But the quality of yarn, the loom settings, the relative humidity of the atmosphere in which the weaving is done are also factors that vary from factory to factory in the industry and may have important effects. The best quantity of size may not be the same at all values of the other factors, and it is not enough to control the other factors at one value (or *level*, as it is generally called) and vary only the quantity of size in order to find the corresponding best quantity. The effect of quantity must be investigated over a range of yarns, loom settings, and relative humidities, so that, for any factory working at given levels of these three, the best quantity can be specified or, perhaps preferably, the best combination of quantity, yarn, loom settings, and relative humidity. According to what may be termed the classical method of experimentation, only one experimental factor would be varied at a time; the practical needs of technical experimentation call for the comprehensive study of a number of experimental factors at the same time in the so-called *factorial experiment*. Until the full investigation is complete, no answer can be given to the general question, "What is the best quantity of size?"

The words *experimental factor* denote something that is related to but not quite the same as that denoted by the word factor in Chapters 9 and 10. There factors correspond to the parts into which variance is analysed, and in an experiment the treatments en bloc are one factor. Experimental factors represent the parts into which the treatments are divided. Thus the six treatments of Table VIII (p. 106) representing different yarns are divided in Table XII (p. 115) into two experimental factors, cottons and twists, the former being at two levels and the latter at three. The recognition of experimental factors affects the make-up of the treatments; it does not directly determine whether the design is random, in randomised blocks, and so on.

Difficulties arise in making factorial experiments. In many fields of enquiry the number of relevant experimental factors tends to be large, especially when the quantities of different ingredients of some mixture are involved; and the number of treatments multiplies accordingly. If

there are only two levels per experimental factor, the number of treatments is 2^f, where f is the number of factors. But two levels are not always enough. They suffice if there is a straight-line relationship between the value of the level and the value of the effect, but if there is a minimum or maximum on the curve at least three or four levels are necessary. In these circumstances the number of treatments may become impossibly large, and it may be necessary to make a choice between investigating several experimental factors, each at two levels on the one hand, and investigating few factors, each at several levels on the other.

A complex factorial experiment raises difficulties of interpretation. In the analysis of variance the sum of squares for treatments is split up into parts associated with the main effects and various first- and higher-order interactions. There can easily be a list of between 12 and 20 mean squares, each based on 1 degree of freedom, and it is not easy to say which are truly significant. It is as difficult to test the significance of a long list of mean squares as of a long list of mean differences (see p. 106). Moreover, to all but those experienced in analysing the results of complex experiments, it is not easy to comprehend fully the meaning of a second- or higher-order interaction; and it is dangerous for the interpretation of results to go beyond the easy comprehension of the chemists, physicists, and engineers concerned with the technical aspects of an experiment.

The experimental plan finally adopted will depend on the circumstances and on the judgment (and even prejudices) of the people involved. Where there are many equally important experimental factors about which little or nothing is known, and the experimenters can work in one field long enough to master the mysteries of the analysis of complex results, elaborate factorial experiments may be undertaken with success; and some people have had the opportunity to develop an expertise in this direction that is very valuable. Where the experimenters have not the time or the will to develop this expertise they will seek some other way out. They will, on the basis of existing knowledge, be prepared to regard some experimental factors as having only secondary importance and to control them at one level. If the remaining factors are many, they will divide the field up, investigating, for example, the effect of quantity of size and yarn at one level of loom settings in one experiment, at another level in another experiment, and so on, thus advancing knowledge certainly on small sectors rather than attempting to advance on a broad front. And if they have to apply the results obtained at one level of settings to a factory working at a

second level before the second level has been investigated, they will do so knowing that life can not be lived without taking risks.

The situation often calls for compromise between the classical one-thing-at-a-time experiment and the elaborate factorial experiment. In arriving at the compromise I place the full comprehension of the experimental results by the technicians very high in the list of *desiderata*.

A related question, on which we have already touched, is how many levels of each factor there should be. No more than are necessary to define the curve of relationship between the measured quantities—two for a straight line, three for a parabola without a maximum or minimum in the range covered, four for one with a maximum or minimum (unless one can be sure that the curve is accurately a second-order parabola), and so on. The fewer the number of levels for a given total number of observations, the larger is the number of replicates per level, and the easier is it to test the statistical significance of the results.

Another question is: Over what range should the experimental factors be varied? The short answer is: Over the range of practical interest, having regard to the future as well as the present. Suppose the aim is to investigate the effect of temperature in some process on, say, the strength of a product. To investigate the effect of temperature over a range of, say, 100°F to 180°F and use the results to predict the strength at 80°F or 250°F is to run all the risks of extrapolation. These are usually well-appreciated. But if the range of interest is 100°F to 180°F, some people would suggest experimenting over a range of, say, 50°F to 300°F on the ground that the effect of temperature is thereby exaggerated and the relationship within the narrower range of interest more accurately defined. On this argument one would plot strength against temperature, fit some form of smooth curve, and use this curve within the range 100°F to 180°F. But the results obtained outside this range add information on what is happening within the range only if the temperature effect is of the same kind, differing only in degree over the full range investigated, and if the form of curve used is suitable. These conditions can not often be known to obtain, and the suggested procedure is not often advisable. Its justification is based on the same arguments as the justification for extrapolation from results obtained over a restricted range, but it is not so dangerous as extrapolation. The preferable course, largely on the grounds of economy, is to take as many observations as possible just covering the range of practical interest.

Conclusion

The application of statistical methods to investigations in the technological (and indeed any other) field is based on assumptions, is subject to limitations, and often leads to uncertain inferences. It is necessary to realise all this but not to be baffled by it. Investigators can go wrong by tacitly making assumptions that are false or misapplying statistical methods, but if commonsense and experience are also applied mistakes will not often be serious. Many useful conclusions have been reached with the aid of assumptions that, strictly, do not apply, and indeed rarely do all the assumptions and conditions that lie behind the relatively simple mathematical models apply exactly in practice. Statistical methods and ideas are a help and a guide, and they have enormously increased our powers of gaining knowledge, but they are only part of the mental tools we use. Very seldom does a particular investigation stand on its own, and we do not often have to rely only on the results of the statistical analysis in making inferences; the background of previous knowledge and general scientific insight should not be at a discount because statistics is used—rather they should guide the application of statistics. So I would counsel the investigator who is only beginning to use statistics not to be discouraged by the complexity and limitations of the subject. Apply the simpler designs first and go to the more complex ones only as experience and comprehension grow; perform the statistical analysis without being unduly concerned about the assumptions (although the more you understand them the better), give your general scientific knowledge and experience full weight in reaching conclusions; and always remember that you may be wrong.

BIBLIOGRAPHY

PART I

Background Reading

Economic Control of Quality of Manufactured Products, by W. A. Shewhart (Van Nostrand, New York, and Macmillan, London).

Manual on Presentation of Data (American Society for Testing Materials).

The Application of Statistical Methods to Industrial Standardisation and Quality Control, by E. S. Pearson (B.S. 600:1935, British Standards Institution).

An Engineer's Manual of Statistical Methods, by L. E. Simon (Wiley).

Short Elementary Manuals on Quality Control

Quality Control (Standards Z1.1-1941 and Z1.2-1941, American Standards Association).

Fraction-Defective Charts for Quality Control (B.S. 1313:1947, British Standards Institution).

Quality Control Charts, by B. P. Dudding and W. J. Jennett (B.S. 600R:1942, British Standards Institution).

Quality Control Chart Technique when Manufacturing to a Specification, by B. P. Dudding and W. J. Jennett (General Electric Co., Ltd., of England).

Full Text-Books on Quality Control

Statistical Quality Control, by E. L. Grant (McGraw-Hill, New York and London).

Control Charts in Factory Management, by W. B. Rice (Wiley, New York).

Control Charts, by E. S. Smith (McGraw-Hill, New York and London).

Quality Control in Production, by H. Rissik (Pitman, London).

An Introduction to Industrial Statistics and Quality Control, by Paul Peach (Edwards Broughton Co., Raleigh, N. C.).

Books on Sampling Inspection

Sampling Inspection Tables, by H. F. Dodge and H. G. Romig (Wiley, New York).

Sampling Inspection (Principles, Procedures and Tables for Single, Double and Sequential Sampling in Acceptance Inspection and Quality Control Based on Percent Defective), by Statistical Research Group, Columbia University (McGraw-Hill, New York and London).

PART II

Industrial Statistics, by H. A. Freeman (Wiley, New York).

The Methods of Statistics, by L. H. C. Tippett (Williams & Norgate, London).

Statistical Methods in Research and Production, edited by O. L. Davies (Oliver & Boyd, London and Edinburgh, for the Imperial Chemical Industries, Ltd.).

185

Industrial Experimentation, by K. A. Brownlee (Chemical Publishing Co., Brooklyn, N. Y.).

The Design of Experiments, by R. A. Fisher (Oliver & Boyd, London and Edinburgh).

Experimental Designs, by W. G. Cochran and G. M. Cox (Wiley, New York).

Index