BY THE SAME AUTHOR:

Statistical Methods for Research Workers
(1925–13th Ed. 1963)

Statistical Methods and Scientific Inference
(1956, 1959)

The Theory of Inbreeding
(1949, 1965)

WITH F. YATES:

Statistical Tables
(1938–6th Ed. 1963)

All published by Oliver and Boyd Ltd., Edinburgh

The Design of Experiments

By

Sir Ronald A. Fisher, Sc.D., F.R.S.

Honorary Research Fellow, Division of Mathematical Statistics, C.S.I.R.O., University of Adelaide; Foreign Associate, United States National Academy of Sciences, and Foreign Honorary Member, American Academy of Arts and Sciences; Foreign Member of the Swedish Royal Academy of Sciences, and the Royal Danish Academy of Sciences and Letters; Member of the Pontifical Academy; Member of the German Academy of Sciences (Leopoldina); formerly Galton Professor, University of London, and Arthur Balfour Professor of Genetics, University of Cambridge

© 1960.

8th Edition 1966

Oliver and Boyd
Edinburgh: Tweeddale Court
London: 39a Welbeck Street, W.1

First Published	1935
Second Edition	1937
Third Edition	1942
Fourth Edition	1947
Fifth Edition	1949
Sixth Edition	1951
Reprinted	1953
Seventh Edition	1960
Eighth Edition	1966

PRINTED AND PUBLISHED IN GREAT BRITAIN
BY OLIVER AND BOYD LTD., EDINBURGH

PREFACE TO FIRST EDITION

In 1925 the author wrote a book (*Statistical Methods for Research Workers*) with the object of supplying practical experimenters and, incidentally, teachers of mathematical statistics, with a connected account of the applications in laboratory work of some of the more recent advances in statistical theory. Some of the new methods, such as the analysis of variance, were found to be so intimately related with problems of experimental design that a considerable part of the eighth chapter was devoted to the technique of agricultural experimentation, and these sections have been progressively enlarged with subsequent editions, in response to frequent requests for a fuller treatment of the subject. The design of experiments is, however, too large a subject, and of too great importance to the general body of scientific workers, for any incidental treatment to be adequate. A clear grasp of simple and standardised statistical procedures will, as the reader may satisfy himself, go far to elucidate the principles of experimentation; but these procedures are themselves only the means to a more important end. Their part is to satisfy the requirements of sound and intelligible experimental design, and to supply the machinery for unambiguous interpretation. To attain a clear grasp of these requirements we need to study designs which have been widely successful in many fields, and to examine their structure in relation to the requirements of valid inference.

The examples chosen in this book are aimed at such an assessment would be inadmissible and irrelevant in judging the state of the scientific evidence; moreover, accurately assessable prior information is ordinarily known to be lacking. Such differences between the logical situations should be borne in mind whenever we see tests of significance spoken of as " Rules of Action ". A good deal of confusion has certainly been caused by the attempt to formalise the exposition of tests of significance in a logical framework different from that for which they were in fact first developed.

REFERENCES AND OTHER READING

R. A. Fisher (1925-1963). Statistical methods for research workers. Chap. III., §§ 15-19

R. A. Fisher (1926). The arrangement of field experiments. Journal of Ministry of Agriculture, xxxiii. 503-513.

III

A HISTORICAL EXPERIMENT ON GROWTH RATE

13. We have illustrated a psycho-physical experiment, the result of which depends upon judgments, scored " right " or " wrong," and may be appropriately interpreted by the method of the classical theory of probability. This method rests on the enumeration of the frequencies with which different combinations of right or wrong judgments will occur, on the hypothesis to be tested. We may now illustrate an experiment in which the results are expressed in quantitative measures, and which is appropriately interpreted by means of the theory of errors.

In the introductory remarks to his book on " The effects of cross and self-fertilisation in the vegetable kingdom," Charles Darwin gives an account of the considerations which guided him in the design of his experiments and in the presentation of his data, which will serve well to illustrate the principles on which biological experiments may be made conclusive. The passage is of especial interest in illustrating the extremely crude and unsatisfactory statistical methods available at the time, and the manner in which careful attention to commonsense considerations led to the adoption of an experimental design, in itself greatly superior to these methods of interpretation.

14. Darwin's Discussion of the Data

" I long doubted whether it was worth while to give the measurements of each separate plant, but have decided to do so, in order that it may be seen that the superiority of the crossed plants over the self-fertilised does not commonly depend on the presence of two or three extra fine plants on the one side, or of a few very poor plants on the other side. Although several observers have insisted in general terms on the offspring from intercrossed varieties being superior to either parent-form, no precise measurements have been given ; and I have met with no observations on the effects of crossing and self-fertilising the individuals of the same variety. Moreover, experiments of this kind require so much time—mine having been continued during eleven years—that they are not likely soon to be repeated.

" As only a moderate number of crossed and self-fertilised plants were measured, it was of great importance to me to learn how far the averages were trustworthy. I therefore asked Mr Galton, who has had much experience in statistical researches, to examine some of my tables of measurements, seven in number, namely those of *Ipomœa*, *Digitalis*, *Reseda lutea*, *Viola*, *Limnanthes*, *Petunia*, and *Zea*. I may premise that if we took by chance a dozen or score of men belonging to two nations and measured them, it would I presume be very rash to form any judgment from such small numbers on their average heights. But the case is somewhat different with my crossed and self-fertilised plants, as they were of exactly the same age, were subjected from first to last to the same conditions, and were descended from the same parents. When only from two to six pairs of plants were measured, the results are manifestly of little or no value, except in so far as they confirm and are confirmed by experiments made on a larger scale with other species. I will now give the report on the seven tables of measurements,

which Mr Galton has had the great kindness to draw up for me."

15. Galton's Method of Interpretation

" I have examined the measurements of the plants with care, and by many statistical methods, to find out how far the means of the several sets represent constant realities, such as would come out the same so long as the general conditions of growth remained unaltered. The principal methods that were adopted are easily explained by selecting one of the shorter series of plants, say of *Zea mays*, for an example.

" The observations as I received them are shown in columns II. and III., where they certainly have no *primâ facie* appearance of regularity. But as soon as we arrange them in the order of their magnitudes, as in columns IV. and V., the case is materially altered. We now see, with few exceptions, that the largest plant on the crossed side in each pot exceeds the largest plant on the self-fertilised side, that the second exceeds the second, the third the third, and so on. Out of the fifteen cases in the table, there are only two exceptions to this rule.* We may therefore confidently affirm that a crossed series will always be found to exceed a self-fertilised series, within the range of the conditions under which the present experiment has been made.

" Next as regards the numerical estimate of this excess. The mean values of the several groups are so discordant, as is shown in the table just given, that a fairly precise numerical estimate seems impossible. But the consideration arises, whether the difference between pot and pot may not be of much the same order of importance as that of the other conditions upon which the growth of the plants has been modified. If so, and only on that condition, it would follow that when all the measurements, either of the crossed or the self-fertilised plants, were combined into a single series, that series would be statistically regular. The experiment is tried in columns VII. and VIII., where the regularity is abundantly clear, and justifies us in considering its mean as perfectly reliable

* Galton evidently did not notice that this is true also before rearrangement.

TABLE I

Zea mays (*young plants*)

As recorded by Mr Darwin.			Arranged in Order of Magnitude.				
			In Separate Pots.		In a Single Series.		
Column I.	II.	III.	IV.	V.	VI.	VII.	VIII.
	Crossed.	Self-fert.	Crossed.	Self-fert.	Crossed.	Self-fert.	Difference.
	Inches.	Inches.	Inches.	Inches.	Inches.	Inches.	Inches.
Pot I.	23 4/8	17 3/8	23 4/8	20 3/8	23 4/8	20 3/8	−3 1/8
	12	20 3/8	21	20	23 2/8	20	−3 2/8
	21	20	12	17 3/8	23	20	−3
					22 1/8	18 5/8	−3 4/8
Pot II.	22	20	22	20	22 1/8	18 5/8	−3 4/8
	19 1/8	18 3/8	21 4/8	18 5/8	22	18 3/8	−3 5/8
	21 4/8	18 5/8	19 1/8	18 3/8	21 5/8	18	−3 5/8
					21 4/8	18	−3 4/8
Pot III.	22 1/8	18 5/8	23 2/8	18 5/8	21	18	−3
	20 3/8	15 2/8	22 1/8	18	21	17 3/8	−3 5/8
	18 2/8	16 4/8	21 5/8	16 4/8	20 3/8	16 4/8	−3 7/8
	21 5/8	18	20 3/8	16 2/8	19 1/8	16 2/8	−2 7/8
	23 2/8	16 2/8	18 2/8	15 2/8	18 2/8	15 4/8	−2 6/8
					12	15 2/8	+3 2/8
Pot IV.	21	18	23	18	12	12 6/8	+0 6/8
	22 1/8	12 6/8	22 1/8	18	...	...	...
	23	15 4/8	21	15 4/8	...	...	...
	12	18	12	12 6/8	...	...	...

I have protracted these measurements, and revised them in the usual way, by drawing a curve through them with a free hand, but the revision barely modifies the means derived from the original observations. In the present, and in nearly all the other cases, the difference between the original and revised means is under 2 per cent. of their value. It is a very remarkable coincidence that in the seven kinds of plants, whose measurements I have examined, the ratio between the heights of the crossed and of the self-fertilised ranges in five cases within very narrow limits. In *Zea mays* it is as 100 to 84, and in the others it ranges between 100 to 76 and 100 to 86.

TABLE 2

Pot.	Crossed.	Self-fert.	Difference.
I.	18 7/8	19 2/8	+0 3/8
II.	20 7/8	19	−1 7/8
III. . . .	21 1/8	16 7/8	−4 2/8
IV. . . .	19 6/8	16	−3 6/8

"The determination of the variability (measured by what is technically called the 'probable error') is a problem of more delicacy than that of determining the means, and I doubt, after making many trials, whether it is possible to derive useful conclusions from these few observations. We ought to have measurements of at least fifty plants in each case, in order to be in a position to deduce fair results. . . ."

"Mr Galton sent me at the same time graphical representations which he had made of the measurements, and they evidently form fairly regular curves. He appends the words 'very good' to those of *Zea* and *Limnanthes*. He also calculated the average height of the crossed and self-fertilised plants in the seven tables by a more correct method than that followed by me, namely by including the heights, as estimated in accordance with statistical rules, of a few plants which died before they were measured; whereas I merely added up the heights of the survivors, and divided the sum by their number. The difference in our results is in one way highly satisfactory, for the average heights of the self-fertilised plants, as deduced by Mr Galton, is less than mine in all the cases excepting one, in which our averages are the same; and this shows that I have by no means exaggerated the superiority of the crossed over the self-fertilised plants."

16. Pairing and Grouping

It is seen that the method of comparison adopted by Darwin is that of pitting each self-fertilised plant against a cross-fertilised one, in conditions made as equal as possible. The pairs so chosen for comparison had germinated at the same time, and the soil conditions in which they grew were largely equalised by planting in the same pot. Necessarily they were not of the same parentage, as it would be difficult in maize to self-fertilise two plants at the same time as raising a cross-fertilised progeny from the pair. However, the parents were presumably grown from the same batch of seed. The evident object of these precautions is to increase the sensitiveness of the experiment, by making such differences in growth rate as were to be observed as little as possible dependent from environmental circumstances, and as much as possible, therefore, from intrinsic differences due to their mode of origin.

The method of pairing, which is much used in modern biological work, illustrates well the way in which an appropriate experimental design is able to reconcile two desiderata, which sometimes appear to be in conflict. On the one hand we require the utmost uniformity in the biological material, which is the subject of experiment, in order to increase the sensitiveness of each individual observation; and, on the other, we require to multiply the observations so as to demonstrate so far as possible the reliability and consistency of the results. Thus an experimenter with field crops may desire to replicate his experiments upon a large number of plots, but be deterred by the consideration that his facilities allow him to sow only a limited area on the same day. An experimenter with small mammals may have only a limited supply of an inbred and highly uniform stock, which he believes to be particularly desirable for experimental purposes. Or, he may desire to carry out his experiments on members of the same litter, and feel that his experiment is limited by the size of the largest litter he can obtain. It has indeed frequently been argued that, beyond a certain moderate degree, further replication can give no further increase in precision, owing to the increasing heterogeneity with which, it is thought, it must be accompanied. In all these cases, however, and in the many analogous cases which constantly arise, there is no real dilemma. Uniformity is only requisite between the objects whose response is to be contrasted (that is, objects treated differently). It is not requisite that all the parallel plots under the same treatment shall be sown on the same day, but only that each such plot shall be sown so far as possible simultaneously with the differently treated plot or plots with which it is to be compared. If, therefore, only two kinds of treatments are under examination, pairs of plots may be chosen, one plot for each treatment; and the precision of the experiment will be given its highest value if the members of each pair are treated closely alike, but will gain nothing from similarity of treatment applied to different pairs, nor lose anything if the conditions in these are somewhat varied. In the same way, if the numbers of animals available from any inbred line are too few for adequate replication, the experimental contrasts in treatments may be applied to pairs of animals from different inbred lines, so long as each pair belongs to the same line. In these two cases it is evident that the principle of combining similarity between controls to be compared, with diversity between parallels, may be extended to cases where three or more treatments are under investigation. The requirement that animals to be contrasted must come from the same litter limits, not the amount of replication, but the number of different treatments that can be so tested. Thus we might test three, but not so easily four or five treatments, if it were necessary that each set of animals must be of the same sex and litter. Paucity of homogeneous material limits the number of different treatments in an experiment, not the number of replications. It may cramp the scope and comprehensiveness of an experimental enquiry, but sets no limit to its possible precision.

17. "Student's" *t* Test *

Owing to the historical accident that the theory of errors, by which quantitative data are to be interpreted, was developed without reference to experimental methods, the vital principle has often been overlooked that the actual and physical conduct of an experiment must govern the statistical procedure of its interpretation. In using the theory of errors we rely for our conclusion upon one or more estimates of error, derived from the data, and appropriate to the one or more sets

of comparisons which we wish to make. Whether these estimates are valid, for the purpose for which we intend them, depends on what has been actually done. It is possible, and indeed it is all too frequent, for an experiment to be so conducted that no valid estimate of error is available. In such a case the experiment cannot be said, strictly, to be capable of proving anything. Perhaps it should not, in this case, be called an *experiment* at all, but be added merely to the body of *experience* on which, for lack of anything better, we may have to base our opinions. All that we need to emphasise immediately is that, if an experiment does allow us to calculate a valid estimate of error, its structure must completely determine the statistical procedure by which this estimate is to be calculated. If this were not so, no interpretation of the data could ever be unambiguous; for we could never be sure that some other equally valid method of interpretation would not lead to a different result.

The object of the experiment is to determine whether the difference in origin between inbred and cross-bred plants influences their growth rate, as measured by height at a given date; in other words, if the numbers of the two sorts of plants were to be increased indefinitely, our object is to determine whether the average heights, to which these two aggregates of plants will tend, are equal or unequal. The most general statement of our null hypothesis is, therefore, that the limits to which these two averages tend are equal. The theory of errors enables us to test a somewhat more limited hypothesis, which, by wide experience, has been found to be appropriate to the metrical characters of experimental material in biology. The disturbing causes which introduce discrepancies in the means of measurements of similar material are found to produce quanti-

* A full account of this test in more varied applications, and the tables for its use, will be found in *Statistical Methods for Research Workers.* Its originator, who published anonymously under the pseudonym "Student," possesses the remarkable distinction that, without being a professed mathematician, but a research chemist, he made early in life this revolutionary refinement of the classical theory of errors.

tative effects which conform satisfactorily to a theoretical distribution known as the normal law of frequency of error. It is this circumstance that makes it appropriate to choose, as the null hypothesis to be tested, one for which an exact statistical criterion is available, namely that the two groups of measurements are samples drawn from the same normal population. On the basis of this hypothesis we may proceed to compare the average difference in height, between the cross-fertilised and the self-fertilised plants, with such differences as might be expected between these averages, in view of the observed discrepancies between the heights of plants of like origin.

We must now see how the adoption of the method of pairing determines the details of the arithmetical procedure, so as to lead to an unequivocal interpretation. The pairing procedure, as indeed was its purpose, has equalised any differences in soil conditions, illumination, air-currents, etc., in which the several pairs of individuals may differ. Such differences having been eliminated from the experimental comparisons, and contributing nothing to the real errors of our experiment, must, for this reason, be eliminated likewise from our estimate of error, upon which we are to judge what differences between the means are compatible with the null hypothesis, and what differences are so great as to be incompatible with it. We are therefore not concerned with the differences in height among plants of like origin, but only with differences in height between members of the same pair, and with the discrepancies among these differences observed in different pairs. Our first step, therefore, will be to subtract from the height of each cross-fertilised plant the height of the self-fertilised plant belonging to the same pair. The differences are shown below in eighths of an inch.

With respect to these differences our null hypothesis asserts that they are normally distributed about a mean value at zero, and we have to test whether our 15 observed differences are compatible with the supposition that they are a sample from such a population.

TABLE 3

Differences in eighths of an inch between cross- and self-fertilised plants of the same pair

49	23	56
−67	28	24
8	41	75
16	14	60
6	29	−48

The calculations needed to make a rigorous test of the null hypothesis stated above involve no more than the sum, and the sum of the squares, of these numbers. The sum is 314, and, since there are 15 plants, the mean difference is $20\frac{14}{15}$ in favour of the cross-fertilised plants. The sum of the squares is 26,518, and from this is deducted the product of the total and the mean, or 6573, leaving 19,945 for the sum of squares of deviations from the mean, representing discrepancies among the differences observed in the 15 pairs. The algebraic fact here used is that

$$S(x-\bar{x})^2 = S(x^2)-\bar{x}S(x)$$

where S stands for summation over the sample, and $\bar{x}$ for the mean value of the observed differences, x.

We may make from this measure of the discrepancies an estimate of a quantity known as the *variance* of an individual difference, by dividing by 14, one less than the number of pairs observed. Equally, and what is more immediately required, we may make an estimate of the variance of the mean of 15 such pairs, by dividing again by 15, a process which yields 94·976 as the estimate.

The square root of the variance is known as the standard error, and it is by the ratio which our observed mean difference bears to its standard error that we shall judge of its significance. Dividing our difference, 20·933, by *its* standard error 9·746, we find this ratio (which is usually denoted by *t*) to be 2·148.

The object of these calculations has been to obtain from the data a quantity measuring the average difference in height between the cross-fertilised and the self-fertilised plants, in terms of the observed discrepancies among these differences; and which, moreover, shall be distributed in a known manner when the null hypothesis is true. The mathematical distribution for our present problem was discovered by "Student" in 1908, and depends only upon the number of independent comparisons (or the number of degrees of freedom) available for calculating the estimate of error. With 15 observed differences we have among them 14 independent discrepancies, and our degrees of freedom are 14. The available tables of the distribution of *t* show that for 14 degrees of freedom the value 2·145 is exceeded by chance, either in the positive or negative direction, in exactly 5 per cent. of random trials. The observed value of *t*, 2·148, thus just exceeds the 5 per cent. point, and the experimental result may be judged significant, though barely so.

18. Fallacious Use of Statistics

We may now see that Darwin's judgment was perfectly sound, in judging that it was of importance to learn how far the averages were trustworthy, and that this could be done by a statistical examination of the tables of measurements of individual plants, though not of their averages. The example chosen, in fact, falls just on the border-line between those results which can suffice by themselves to establish the point at issue, and those which are of little value except in so far as they confirm or are confirmed by other experiments of a like nature. In particular, it is to be noted that Darwin recognised that the reliability of the result must be judged by the consistency of the superiority of the crossed plants over the self-fertilised, and not only on the difference of the averages, which might depend, as he says, on the presence of two or three extra-fine plants on the one side, or of a few very poor plants on the other side; and that therefore the presentation of the experimental evidence depended essentially on giving the measurements of each independent plant, and could not be assessed from the mere averages.

It may be noted also that Galton's scepticism of the value of the probable error, deduced from only 15 pairs of observations, though, as it turned out, somewhat excessive, was undoubtedly right in principle. The standard error (of which the probable error is only a conventional fraction) can only be estimated with considerable uncertainty from so small a sample, and, prior to "Student's" solution of the problem, it was by no means clear to what extent this uncertainty would invalidate the test of significance. From "Student's" work it is now known that the cause for anxiety was not so great as it might have seemed. Had the standard error been known with certainty, or derived from an effectively infinite number of observations, the 5 per cent. value of *t* would have been 1·960. When our estimate is based upon only 15 differences, the 5 per cent. value, as we have seen, is 2·145, or less than 10 per cent. greater. Even using the inexact theory available at the time, a calculation of the probable error would have provided a valuable guide to the interpretation of the results.

19. Manipulation of the Data

A much more serious fallacy appears to be involved in Galton's assumption that the value of the data, for the purpose for which they were intended, could be increased by rearranging the comparisons. Modern statisticians are familiar with the notions that any finite body of data contains only a limited amount of information, on any point under examination; that this limit is set by the nature of the data themselves, and cannot be increased by any amount of ingenuity expended in their statistical examination: that the statistician's task, in fact, is limited to the extraction of the whole of the available information on any particular issue. If the results of an experiment, as obtained, are in fact irregular, this evidently detracts from their value; and the statistician is not elucidating but falsifying the facts, who rearranges them so as to give an artificial appearance of regularity.

In rearranging the results of Darwin's experiment it appears that Galton thought that Darwin's experiment would be equivalent to one in which the heights of pairs of contrasted plants had been those given in his columns headed VI. and VII., and that the reliability of Darwin's average difference of about $2\frac{5}{8}$ inches could be fairly judged from the constancy of the 15 differences shown in column VIII.

How great an effect this procedure, if legitimate, would have had on the significance of the result, may be seen by treating these artificial differences as we have treated the actual differences given by Darwin. Applying the same arithmetical procedure as before, we now find t equals 5·171, a value which would be exceeded by chance only about once or twice in 10,000 trials, and is far beyond the level of significance ordinarily required. The falsification, inherent in this mode of procedure, will be appreciated if we consider that the tallest plant, of either the crossed or the self-fertilised series, will have become the tallest by reason of a number of favourable circumstances, including among them those which produce the discrepancies between those pairs of plants, which were actually grown together. By taking the difference between these two favoured plants we have largely eliminated real causes of error which have affected the value of our observed mean. We have, in doing this, grossly violated the principle that the estimate of error must be based on the effects of the very same causes of variation as have produced the real errors in our experiment. Through this fallacy Galton is led to speak of the mean as perfectly reliable, when, from its standard error, it appears that a repetition of the experiment would often give a mean quite 50 per cent. greater or less than that observed in this case.

20. Validity and Randomisation

Having decided that, when the structure of the experiment consists in a number of independent comparisons between pairs, our estimate of the error of the average difference must be based upon the discrepancies between the differences actually observed, we must next enquire what precautions are needed in the practical conduct of the experiment to guarantee that such an estimate shall be a valid one; that is to say that the very same causes that produce our real error shall also contribute the materials for computing an estimate of it. The logical necessity of this requirement is readily apparent, for, if causes of variation which do not influence our real error are allowed to affect our estimate of it, or equally, if causes of variation affect the real error in such a way as to make no contribution to our estimate, this estimate will be vitiated, and will be incapable of providing a correct statement as to the frequency with which our real error will exceed any assigned quantity ; and such a statement of frequency is the sole purpose for which the estimate is of any use. Nevertheless, though its logical necessity is easily apprehended, the question of the validity of the estimates of error used in tests of significance was for long ignored, and is still often overlooked in practice. One reason for this is that standardised methods of statistical analysis have been taken over ready-made from a mathematical theory, into which questions of experimental detail do not explicitly enter. In consequence the assumptions which enter implicitly into the bases of the theory have not been brought prominently under the notice of practical experimenters. A second reason is that it has not until recently been recognised that any simple precaution would supply an absolute guarantee of the validity of the calculations.

In the experiment under consideration, apart from chance differences in the selection of seeds, the sole source of the experimental error in the average of our fifteen differences lies in the differences in soil fertility, illumination, evaporation, etc., which make the site of each crossed plant more or less favourable to growth than the site assigned to the corresponding self-fertilised plant. It is for this reason that every precaution, such as mixing the soil, equalising the watering and orienting the pot so as to give equal illumination, may be expected to increase the precision of the experiment. If, now, when the fifteen pairs of sites have been chosen, and in so doing all the differences in environmental circumstances, to which the members of the different pairs will be exposed during the course of the experiment, have been predetermined, we then assign at random, as by tossing a coin, which site shall be occupied by the crossed and which by the self-fertilised plant, we shall be assigning by the same act whether this particular ingredient of error shall appear in our average with a positive or a negative sign. Since each particular error has thus an equal and independent chance of being positive or negative, the error of our average will necessarily be distributed in a sampling distribution, centred at zero, which will be symmetrical in the sense that to each possible positive error there corresponds an equal negative error, which, as our procedure guarantees, will in fact occur with equal probability.

Our estimate of error is easily seen to depend only on the same fifteen ingredients, and the arithmetical processes of summation, subtraction and division may be designed, and have in fact been designed, so as to provide the estimate appropriate to the system of chances which our method of choosing sites had imposed on the data. This is to say much more than merely that the experiment is unbiased, for we might still call the experiment unbiased if the whole of the cross-fertilised plants had been assigned to the west side of the pots, and the self-fertilised plants to the east side, by a single toss of the coin. That this would be insufficient to ensure the validity of our estimate may be easily seen ; for it might well be that some unknown circumstance, such as the incidence of different illumination at different times of the day, or the desiccating action of the air-currents prevalent in the greenhouse, might systematically favour all the plants on one side over those on the other. The effect of any such prevailing cause would then be confounded with the advantage, real or apparent, of cross-breeding over inbreeding, and would be eliminated from our estimate of error, which is based solely on the discrepancies between the differences shown by different pairs of plants. Randomisation properly carried out, in which each pair of plants are assigned their positions independently at random, ensures that the estimates of error will take proper care of all such causes of different growth rates, and relieves the experimenter from the anxiety of considering and estimating the magnitude of the innumerable causes by which his data may be disturbed. The one flaw in Darwin's procedure was the absence of randomisation.

Had the same measurements been obtained from pairs of plants properly randomised the experiment would, as we have shown, have fallen on the verge of significance. Galton was led greatly to overestimate its conclusiveness through the major error of attempting to estimate the reliability of the comparisons by rearranging the two series in order of magnitude. His discussion shows, in other respects, an over-confidence in the power of statistical methods to remedy the irregularities of the actual data. In particular, the attempt mentioned by Darwin to improve on the simple averages of the two series " by a more correct method . . . by including the heights, as estimated in accordance with statistical rules, of a few plants which died before they were measured," seems to go far beyond the limits of justifiable inference, and is one of many indications that the logic of statistical induction was in its infancy, even at a time when the technique of accurate experimentation had already been notably advanced.

21. Test of a Wider Hypothesis

It has been mentioned that " Student's " *t* test, in conformity with the classical theory of errors, is appropriate to the null hypothesis that the two groups of measurements are samples drawn from the same normally

distributed population. This is the type of null hypothesis which experimenters, rightly in the author's opinion, usually consider it appropriate to test, for reasons not only of practical convenience, but because the unique properties of the normal distribution make it alone suitable for general application. There has, however, in recent years, been a tendency for theoretical statisticians, not closely in touch with the requirements of experimental data, to stress the element of normality, in the hypothesis tested, as though it were a serious limitation to the test applied. It is, indeed, demonstrable that, as a test of this hypothesis, the exactitude of " Student's " *t* test is absolute. It may, nevertheless, be legitimately asked whether we should obtain a materially different result were it possible to test the wider hypothesis which merely asserts that the two series are drawn from the same population, without specifying that this is normally distributed.

In these discussions it seems to have escaped recognition that the physical act of randomisation, which, as has been shown, is necessary for the validity of any test of significance, affords the means, in respect of any particular body of data, of examining the wider hypothesis in which no normality of distribution is implied. The arithmetical procedure of such an examination is tedious, and we shall only give the results of its application in order to show the possibility of an independent check on the more expeditious methods in common use.

On the hypothesis that the two series of seeds are random samples from identical populations, and that their sites have been assigned to members of each pair independently at random, the 15 differences of Table 3 would each have occurred with equal frequency with a positive or with a negative sign. Their sum, taking account of the two negative signs which have actually occurred, is 314, and we may ask how many of the 2^{15} numbers, which may be formed by giving each component alternatively a positive and a negative sign, exceed this value. Since *ex hypothesi* each of these 2^{15} combinations will occur by chance with equal frequency, a knowledge of how many of them are equal to or greater than the value actually observed affords a direct arithmetical test of the significance of this value.

It is easy to see that if there were no negative signs, or only one, every possible combination would exceed 314, while if the negative signs are 7 or more, every possible combination will fall short of this value. The distribution of the cases, when there are from 2 to 6 negative values, is shown in the following table :—

TABLE 4

Number of combinations of differences, positive or negative, which exceed or fall short of the total observed

Number of negative values.	>314	= 314	<314	Total.
0 . . .	1	...	...	1
1 . . .	15	...	...	15
2 . . .	94	1	10	105
3 . . .	263	3	189	455
4 . . .	302	11	1,052	1,365
5 . . .	138	12	2,853	3,003
6 . . .	22	1	4,982	5,005
7 or more . .	...	...	22,819	22,819
Total . . .	835	28	31,905	32,768

In just 863 cases out of 32,768 the total deviation will have a positive value as great as or greater than that observed. In an equal number of cases it will have as great a negative value. The two groups together constitute 5·267 per cent. of the possibilities available, a result very nearly equivalent to that obtained using the *t* test with the hypothesis of a normally distributed population. Slight as it is, indeed, the difference between the tests of these two hypotheses is partly due to the continuity of the *t* distribution, which effectively counts only half of the 28 cases which give a total of exactly 314, as being as great as or greater than the observed value.

Both tests prove that, in about 5 per cent. of trials, samples from the same batch of seed would show differences just as great, and as regular, as those observed ; so that the experimental evidence is scarcely sufficient to stand alone. In conjunction with other experiments, however, showing a consistent advantage of cross-fertilised seed, the experiment has considerable weight ; since only once in 40 trials would a chance deviation have been observed both so large, and in the right direction.

How entirely appropriate to the present problem is the use of the distribution of *t*, based on the theory of errors, when accurately carried out, may be seen by inserting an adjustment, which effectively allows for the discontinuity of the measurements. This adjustment is not usually of practical importance, with the *t* test, and is only given here to show the close similarity of the results of testing the two hypotheses, in one of which the errors are distributed according to the normal law, whereas in the other they may be distributed in any conceivable manner. The adjustment * consists in calculating the value of *t* as though the total difference between the two sets of measurements were less than that actually observed by half a unit of grouping ;

* This adjustment is an extension to the distribution of *t* of Yates' adjustment for continuity, which is of greater importance in the distribution of χ^2, for which it was developed.

i.e. as if it were 313 instead of 314, since the possible values advance by steps of 2. The value of *t* is then found to be 2·139 instead of 2·148. The following table shows the effect of the adjustment on the test of significance, and its relation to the test of the more general hypothesis.

TABLE 5

		t.	Probability of a Positive Difference exceeding that observed.
Normal hypothesis	unadjusted	2·148	2·485 per cent.
	adjusted	2·139	2·529 ,,
General hypothesis			2·634 ,,

The difference between the two hypotheses is thus equivalent to little more than a probability of one in a thousand.

21·1. "Non-parametric" Tests

In recent years tests using the physical act of randomisation to supply (on the Null Hypothesis) a frequency distribution, have been largely advocated under the name of "Non-parametric" tests. Somewhat extravagant claims have often been made on their behalf. The example of this Chapter, published in 1935, was by many years the first of its class. The reader will realise that it was in no sense put forward to supersede the common and expeditious tests based on the Gaussian theory of errors. The utility of such non-parametric tests consists in their being able to supply confirmation whenever, rightly or, more often, wrongly, it is suspected that the simpler tests have been appreciably injured by departures from normality.

They assume less knowledge, or more ignorance, of the experimental material than do the standard tests, and this has been an attraction to some mathematicians who often discuss experimentation without personal knowledge of the material. In inductive logic, however, an erroneous assumption of ignorance is not innocuous; it often leads to manifest absurdities. Experimenters should remember that they and their colleagues usually know more about the kind of material they are dealing with than do the authors of text-books written without such personal experience, and that a more complex, or less intelligible, test is not likely to serve their purpose better, in any sense, than those of proved value in their own subject.

REFERENCES AND OTHER READING

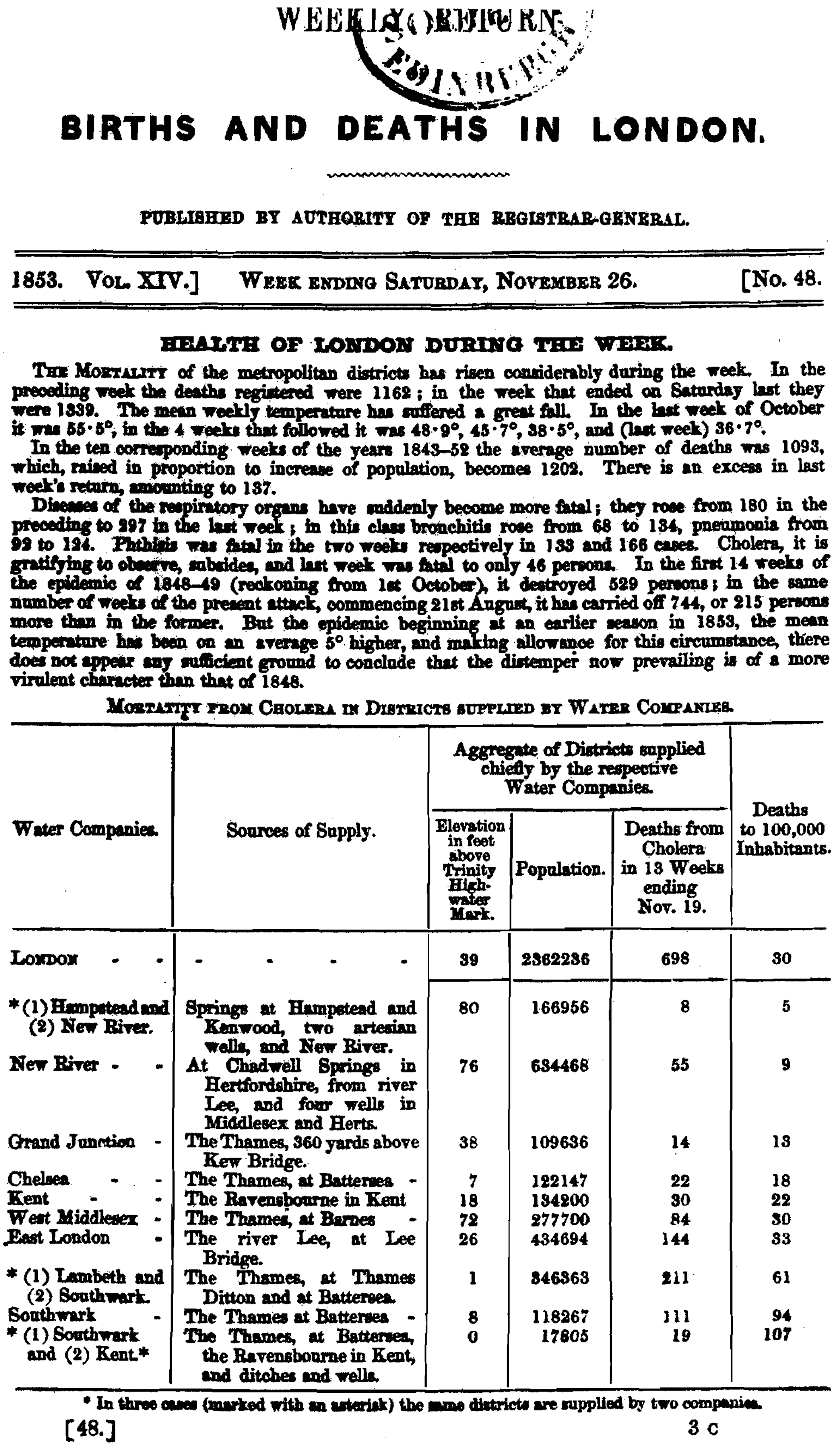
C. Darwin (1876). The effects of cross- and self-fertilisation in the vegetable kingdom. John Murray, London.

R. A. Fisher (1925). Applications of "Student's" distribution. Metron, v. 90-104.

R. A. Fisher (1925-1963). Statistical methods for research workers. Chap. V., §§ 23-24.

R. A. Fisher (1956, 1959). Statistical methods and scientific inference. Oliver and Boyd Ltd., Edinburgh.

"Student" (1908). The probable error of a mean. Biometrika, vi. 1-25.



WEEKLY RETURN

BIRTHS AND DEATHS IN LONDON.

PUBLISHED BY AUTHORITY OF THE REGISTRAR-GENERAL.

1853. VOL. XIV.] WEEK ENDING SATURDAY, NOVEMBER 26. [No. 48.

HEALTH OF LONDON DURING THE WEEK.

THE MORTALITY of the metropolitan districts has risen considerably during the week. In the preceding week the deaths registered were 1162; in the week that ended on Saturday last they were 1339. The mean weekly temperature has suffered a great fall. In the last week of October it was 55·5°, in the 4 weeks that followed it was 48·9°, 45·7°, 38·5°, and (last week) 36·7°.

In the ten corresponding weeks of the years 1843–52 the average number of deaths was 1093, which, raised in proportion to increase of population, becomes 1202. There is an excess in last week's return, amounting to 137.

Diseases of the respiratory organs have suddenly become more fatal; they rose from 180 in the preceding to 297 in the last week; in this class bronchitis rose from 68 to 134, pneumonia from 92 to 124. Phthisis was fatal in the two weeks respectively in 133 and 166 cases. Cholera, it is gratifying to observe, subsides, and last week was fatal to only 46 persons. In the first 14 weeks of the epidemic of 1848–49 (reckoning from 1st October), it destroyed 529 persons; in the same number of weeks of the present attack, commencing 21st August, it has carried off 744, or 215 persons more than in the former. But the epidemic beginning at an earlier season in 1853, the mean temperature has been on an average 5° higher, and making allowance for this circumstance, there does not appear any sufficient ground to conclude that the distemper now prevailing is of a more virulent character than that of 1848.

MORTATITY FROM CHOLERA IN DISTRICTS SUPPLIED BY WATER COMPANIES.

Water Companies.	Sources of Supply.	Aggregate of Districts supplied chiefly by the respective Water Companies.			Deaths to 100,000 Inhabitants.
		Elevation in feet above Trinity High-water Mark.	Population.	Deaths from Cholera in 13 Weeks ending Nov. 19.	
LONDON - -	- - - - -	39	2362236	698	30
*(1) Hampstead and (2) New River.	Springs at Hampstead and Kenwood, two artesian wells, and New River.	80	166956	8	5
New River - -	At Chadwell Springs in Hertfordshire, from river Lee, and four wells in Middlesex and Herts.	76	634468	55	9
Grand Junction -	The Thames, 360 yards above Kew Bridge.	38	109636	14	13
Chelsea - -	The Thames, at Battersea -	7	122147	22	18
Kent - -	The Ravensbourne in Kent	18	134200	30	22
West Middlesex -	The Thames, at Barnes -	72	277700	84	30
East London -	The river Lee, at Lee Bridge.	26	434694	144	33
* (1) Lambeth and (2) Southwark.	The Thames, at Thames Ditton and at Battersea.	1	346363	211	61
Southwark -	The Thames at Battersea -	8	118267	111	94
* (1) Southwark and (2) Kent.*	The Thames, at Battersea, the Ravensbourne in Kent, and ditches and wells.	0	17805	19	107

* In three cases (marked with an asterisk) the same districts are supplied by two companies.

[48.] 3 c

Figure 10.2. *Weekly Return of Births and Deaths* (26 November 1853).

were 1339. The mean weekly temperature has suffered a great fall. In the last week of October it was 55·5°, in the 4 weeks that followed it was 48·9°, 45·7°, 38·5°, and (last week) 36·7°.

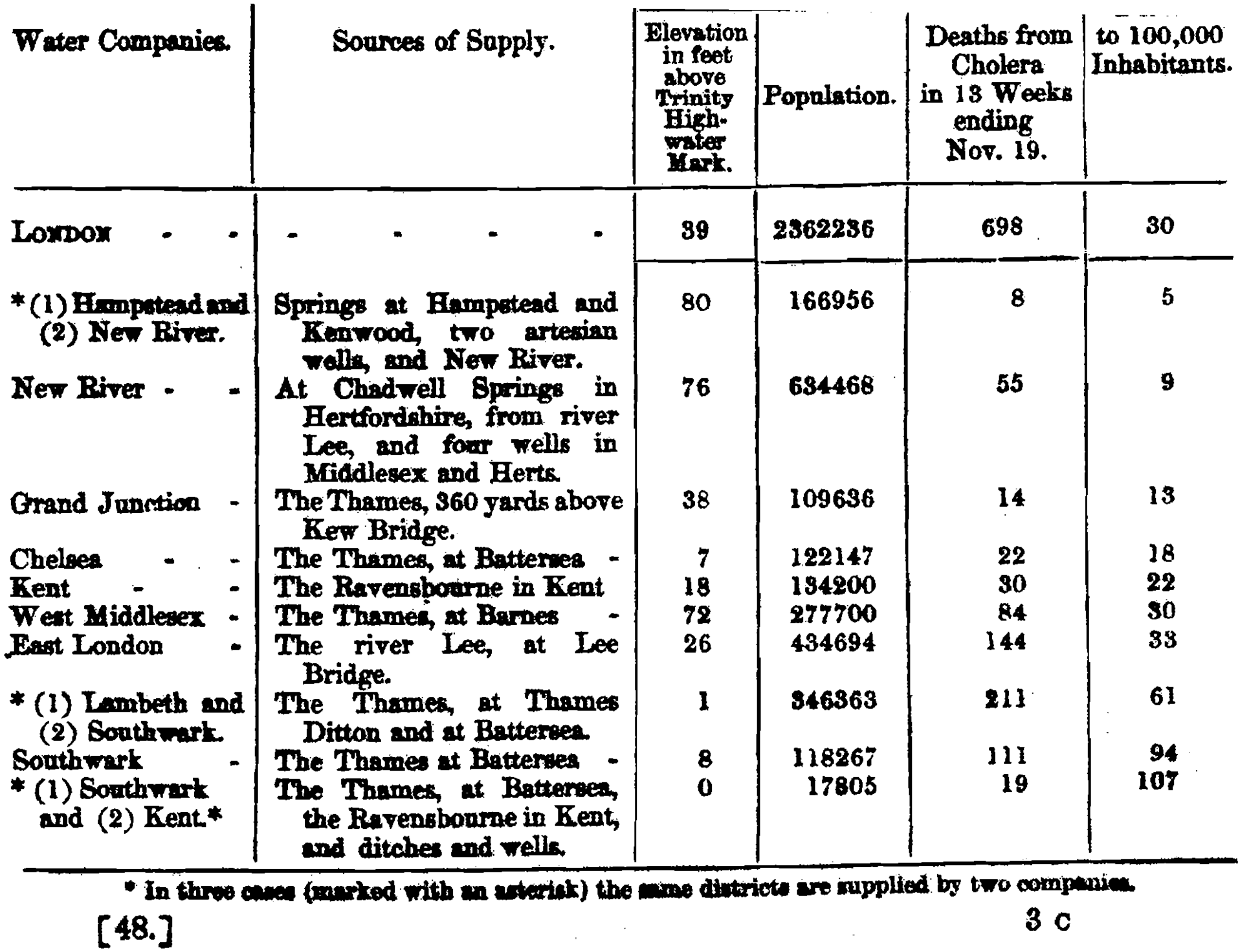
In the ten corresponding weeks of the years 1843–52 the average number of deaths was 1093, which, raised in proportion to increase of population, becomes 1202. There is an excess in last week's return, amounting to 137.

Diseases of the respiratory organs have suddenly become more fatal; they rose from 180 in the preceding to 297 in the last week; in this class bronchitis rose from 68 to 134, pneumonia from 92 to 124. Phthisis was fatal in the two weeks respectively in 133 and 166 cases. Cholera, it is gratifying to observe, subsides, and last week was fatal to only 46 persons. In the first 14 weeks of the epidemic of 1848–49 (reckoning from 1st October), it destroyed 529 persons; in the same number of weeks of the present attack, commencing 21st August, it has carried off 744, or 215 persons more than in the former. But the epidemic beginning at an earlier season in 1853, the mean temperature has been on an average 5° higher, and making allowance for this circumstance, there does not appear any sufficient ground to conclude that the distemper now prevailing is of a more virulent character than that of 1848.

MORTALITY FROM CHOLERA IN DISTRICTS SUPPLIED BY WATER COMPANIES.

Water Companies.	Sources of Supply.	Aggregate of Districts supplied chiefly by the respective Water Companies.			Deaths to 100,000 Inhabitants.
		Elevation in feet above Trinity High-water Mark.	Population.	Deaths from Cholera in 13 Weeks ending Nov. 19.	
LONDON - -	- - - - -	39	2362236	698	30
* (1) Hampstead and (2) New River.	Springs at Hampstead and Kenwood, two artesian wells, and New River.	80	166956	8	5
New River - -	At Chadwell Springs in Hertfordshire, from river Lee, and four wells in Middlesex and Herts.	76	634468	55	9
Grand Junction -	The Thames, 360 yards above Kew Bridge.	38	109636	14	13
Chelsea - -	The Thames, at Battersea -	7	122147	22	18
Kent - -	The Ravensbourne in Kent	18	134200	30	22
West Middlesex -	The Thames, at Barnes -	72	277700	84	30
East London -	The river Lee, at Lee Bridge.	26	434694	144	33
* (1) Lambeth and (2) Southwark.	The Thames, at Thames Ditton and at Battersea.	1	346363	211	61
Southwark -	The Thames at Battersea -	8	118267	111	94
* (1) Southwark and (2) Kent.*	The Thames, at Battersea, the Ravensbourne in Kent, and ditches and wells.	0	17805	19	107

* In three cases (marked with an asterisk) the same districts are supplied by two companies.



Water Companies.	Sources of Supply.	Elevation in feet above Trinity High-water Mark.	Population.	Deaths from Cholera in 13 Weeks ending Nov. 19.	to 100,000 Inhabitants.
London - -	- - - - -	39	2362236	698	30
*(1) Hampstead and (2) New River.	Springs at Hampstead and Kenwood, two artesian wells, and New River.	80	166956	8	5
New River - -	At Chadwell Springs in Hertfordshire, from river Lee, and four wells in Middlesex and Herts.	76	634468	55	9
Grand Junction -	The Thames, 360 yards above Kew Bridge.	38	109636	14	13
Chelsea - -	The Thames, at Battersea -	7	122147	22	18
Kent - -	The Ravensbourne in Kent	18	134200	30	22
West Middlesex -	The Thames, at Barnes -	72	277700	84	30
East London -	The river Lee, at Lee Bridge.	26	434694	144	33
* (1) Lambeth and (2) Southwark.	The Thames, at Thames Ditton and at Battersea.	1	346363	211	61
Southwark -	The Thames at Battersea -	8	118267	111	94
* (1) Southwark and (2) Kent.*	The Thames, at Battersea, the Ravensbourne in Kent, and ditches and wells.	0	17805	19	107

* In three cases (marked with an asterisk) the same districts are supplied by two companies.

[48.] 3 c

Figure 10.2. *Weekly Return of Births and Deaths* (26 November 1853).