"STUDENT'S" COLLECTED PAPERS

Edited by

E. S. PEARSON

AND

JOHN WISHART

With a Foreword

by

LAUNCE McMULLEN

Issued by the Biometrika Office
UNIVERSITY COLLEGE, LONDON

Printed by photo-lithography for the University Press, Cambridge,
by Bradford and Dickens, London, W.C. 1

First issued 1942
Reprinted 1947

PRINTED IN GREAT BRITAIN

CONTENTS

Frontispiece: WILLIAM SEALY GOSSET

Preface PAGE vii

Foreword PAGE ix

1. On the Error of Counting with a Haemacytometer. *Biometrika*, V (Cambridge University Press, 1907) PAGE 1

2. The Probable Error of a Mean. *Biometrika*, VI (Cambridge University Press, 1908) PAGE 11

3. Probable Error of a Correlation Coefficient. *Biometrika*, VI (Cambridge University Press, 1908) PAGE 35

4. The Distribution of the Means of Samples which are not drawn at Random. *Biometrika*, VII (Cambridge University Press, 1909) PAGE 43

5. Appendix to Mercer and Hall's paper on "The Experimental Error of Field Trials". *J. Agric. Sci.* IV (Cambridge University Press, 1911) PAGE 49

6. The Correction to be made to the Correlation Ratio for Grouping. *Biometrika*, IX (Cambridge University Press, 1913) PAGE 53

7. The Elimination of Spurious Correlation due to position in Time or Space. *Biometrika*, X (Cambridge University Press, 1914) PAGE 58

8. Tables for Estimating the Probability that the Mean of a Unique Sample of Observations lies between $-\infty$ and any given distance of the Mean of the Population from which the Sample is drawn. *Biometrika*, XI (Cambridge University Press, 1917) PAGE 61

9. An Explanation of Deviations from Poisson's Law in practice. *Biometrika*, XII (Cambridge University Press, 1919) PAGE 65

10. An Experimental Determination of the Probable Error of Dr Spearman's Correlation Coefficients. *Biometrika*, XIII (Cambridge University Press, 1921) PAGE 70

11. On Testing Varieties of Cereals. *Biometrika*, XV (Cambridge University Press, 1923) PAGE 90

12. New Tables for testing the Significance of Observations. *Metron*, V (Padova, 1925) PAGE 115

13. Mathematics and Agronomy. *J. Amer. Soc. Agron.* XVIII (Washington, 1926) PAGE 121

14. Errors of Routine Analysis. *Biometrika*, XIX (Cambridge University Press, 1927) PAGE 135

15. Yield Trials. *Baillière's Encyclopædia of Scientific Agriculture* (London, 1931) PAGE 150

16. The Lanarkshire Milk Experiment. *Biometrika*, XXIII (Cambridge University Press, 1931) PAGE 169

17. On the "z" Test. *Biometrika*, XXIII (Cambridge University Press, 1931) PAGE 179

18. Evolution by Selection. The Implications of Winter's Selection Experiment. *Eugenics Review*, XXIV (London, 1933) PAGE 181

19. A Calculation of the Minimum Number of Genes in Winter's Selection Experiment. *Annals of Eugenics*, VI (Cambridge University Press, 1934) PAGE 186

20. Co-operation in Large-Scale Experiments. Supplement to *J. Roy. Statist. Soc.* III (London, 1936) PAGE 192

21. Comparison between Balanced and Random Arrangements of Field Plots. *Biometrika*, XXIX (Cambridge University Press, 1938) PAGE 199

Miscellaneous Contributions:

A. Letters to *Nature* (London):

(i) Agricultural Field Experiments (CXXVI, 29 November 1930) PAGE 216

(ii) Agricultural Field Experiments (CXXVII, 14 March 1931) PAGE 217

(iii) The Half-Drill Strip System Agricultural Experiments (CXXXVIII, 5 December 1936) PAGE 218

B. Contributions to Discussions at Meetings of the Industrial and Agricultural Research Section of the Royal Statistical Society. Supplement to *J. Roy. Statist. Soc.* (London), I (1934), III (1936), IV (1937) PAGE 220

PREFACE

Early in 1938 a small group of Gosset's relatives and friends decided to examine the possibility of arranging for the re-issue in a single volume of all the scientific papers which he had published between 1907 and 1937 under the pseudonym of "Student". The project was a happy one, for a unity of purpose runs through the whole of his contributions. In nearly every case the origin of a paper lay in a problem or problems which required solution in connection with his or his colleagues' work at the Dublin brewery and, since the brewer is concerned with barley as well as with chemistry and engineering, the central theme of Student's contributions was the application of statistical method in the research and routine problems of both industry and agriculture.

The simplicity and directness of his methods of approach, his clear grasp of the practical issues, his appreciation of the limitation of mathematics when applied to the data of experience, his warning that statistical technique should be regarded as an aid to but not a substitute for common sense, have given to his writing a fundamental appeal which will last, although the precise mathematical methods by which he derived his results may have been superseded. He was a pioneer worker in a field which, during his later years, was rapidly expanding and his work is intimately related to the historical development of his subject. As such, it was inevitable that he made certain mistakes and that his proofs were not all correct, although it is surprising how right he was in general and how often he "got there first" by what was sometimes an inspired guess.

Under these circumstances we have regarded our editorial role as a minor one; we have not attempted to point out every place where later work may have modified certain of his methods of attack or simplified his mathematical proofs, for we do not expect the reader to regard this volume as a text book. Where numerical or algebraic slips have been discovered, some of them possibly misprints, we have corrected these without comment unless the alteration appeared seriously to modify the argument. Such few editorial comments as were considered necessary appear in footnotes enclosed in square brackets and followed by the abbreviation, Ed. In two instances the original paper contained contemporary editorial comment by Karl Pearson, and here we have inserted the letters K.P. to make the distinction clear. To assist the reader, Student's references in the text to his own contributions have been followed by the number with which the article is headed in this volume, e.g. [2, p. 29]. The main papers have been reprinted in the order of date of publication while a

few shorter miscellaneous contributions are added in a separate section at the end.

As a Foreword, an appreciation by Launce McMullen has been included which, with slight modifications, is the article headed "Student as a Man" that appeared in *Biometrika*, XXX (1938), pp. 205–10. For further appreciations of Student's personality and statistical work reference may be made to articles by R. A. Fisher, *Annals of Eugenics*, IX (1939), pp. 1–9; E. S. Pearson, *Biometrika*, XXX (1938), pp. 210–50; and to contributions by H. H., J. W. and E. M. E. in the *Journal of the Royal Statistical Society*, CI (1938), pp. 248–51.

The papers have been collected from a number of sources and we must thank the following authorities for freely granting permission for their re-issue in the present volume: the Editor of the *Annals of Eugenics*; Messrs Baillière, Tindall & Cox, the Publishers of *Baillière's Encyclopædia of Scientific Agriculture*; the Trustees of *Biometrika*; the Editor of the *Eugenics Review*; the Editor of the *Journal of Agricultural Science*; the Editor of the *Journal of the American Society of Agronomy*; the Director of *Metron*; the Proprietors of *Nature*; the Council of the *Royal Statistical Society*.

We are very grateful to Dr R. C. Geary and Mr E. Somerfield for assistance in proof reading.

Finally we should like to thank Mrs W. S. Gosset and her brother, Mr G. S. Phillpotts, for giving us this opportunity as joint editors of helping to commemorate a friend and teacher to whose inspiration in the past we have owed much. On their behalf we must also thank the Trustees of *Biometrika* for accepting responsibility for publication and Mr Walter Lewis of the Cambridge University Press for invaluable help in the arrangement and printing of the volume.

E. S. PEARSON
JOHN WISHART

September 1942

FOREWORD

WILLIAM SEALY GOSSET was born in 1876—the eldest of four sons and a daughter. His father was Colonel Frederic Gosset, R.E., who married Agnes Sealy Vidal in 1875. The Gossets were an old Huguenot family who left France at the Revocation of the Edict of Nantes.

He was a Scholar of Winchester, and wishing to join the Royal Engineers passed into Woolwich but was rejected in the subsequent medical examination (again in 1916 he wished to volunteer for the Army but was rejected for short sight). He then went as a Scholar to New College, Oxford, where he obtained First Classes in Mathematical Moderations and in Natural Science. In the autumn of 1899 he went as a Brewer to Messrs Guinness in Dublin.

In 1906 he married Marjory Surtees Phillpotts, youngest daughter of the late Headmaster of Bedford School. She was at about that time Captain of the English Ladies Hockey Team, and subsequently she played for, and captained, the Irish Team. They had one son and two daughters.

He died on 16 October 1937 and was survived by both his parents, his wife and children and one grandson.

It is not known exactly how or when "Student's" interest in statistics was first aroused, but at this period scientific methods and laboratory determinations were beginning to be seriously applied to brewing, and it is obvious that some knowledge of error functions would be necessary. A number of university men with science degrees had been taken on, and it is probable that "Student", who was the most mathematical of them, was appealed to by the others with various questions and so began to study the subject. It is known that he could calculate a probable error in 1903. The circumstances of brewing work, with its variable materials and susceptibility to temperature change and necessarily short series of experiments, are all such as to show up most rapidly the limitations of large sample theory and emphasize the necessity for a correct method of treating small samples. It was thus no accident, but the circumstances of his work, that directed "Student's" attention to this problem, and so led to his discovery of the distribution of the sample standard deviation, which gave rise to what in its modern form is known as the t-test. For a long time after its discovery and publication the use of this test hardly spread outside Guinness's brewery, where it has been very extensively used ever since. In the Biometric school at University College the problems investigated were almost all concerned with much larger samples than those in which "studentizing", as it was sometimes called, made any difference. Nevertheless, although their lines of research

diverged somewhat rapidly, the close statistical contact and personal friendship between Karl Pearson and "Student", which began during his year at University College, were only terminated by death.

The purpose of this note is not however to give an account of "Student's" statistical work, but to try to give a more general impression of the man himself. Although his public reputation was entirely as a statistician, and he was acknowledged to be one of the leading investigators in that subject, his time was never wholly and rarely even mainly occupied with statistical matters. For one who saw enough of him to know roughly how his time was spent both at work and at home, it was very difficult to understand how he managed to get so much activity into the day. At work he got through an enormous amount of the ordinary routine of the brewery, as well as his statistics. Until 1922 he had no regular statistical assistant, and did all the statistics and most of the arithmetic himself; later there was a definite department, of which he was in charge till 1934, but throughout he did a great deal of arithmetic and spade-work himself. It might be supposed from the amount he did in the time that he was unusually good at arithmetic and the arrangement of work; such, however, was not the case, for his arithmetic frequently contained minor errors. In one of his obituary notices a tendency to do work on the backs of envelopes in trains was mentioned, but this tendency was not confined to trains; even in his office much work was done on random scraps of paper. He also had a great dislike of the tabulation of results, and preferred to do everything from first principles whenever possible. This preference led in certain instances to waste of time in routine work, but was of assistance in maintaining that flexibility and speed of attack on new problems which was so characteristic of him. An actual example would need too much explanation of relevant circumstances, but I can vouch for the analogical truth of the following. If a body performs simple harmonic motion with acceleration μ per unit displacement, it may readily be shown that the period of a complete oscillation is $2\pi/\sqrt{\mu}$. Hence, in the case of a simple pendulum $t=2\pi\sqrt{(l/g)}$ and $l=gt^2/4\pi^2$, where l is the length of the pendulum and g the acceleration due to gravity. If it were necessary to calculate the lengths of pendulum corresponding to different periods as a routine matter, most people would evaluate $g/4\pi^2$ for their locality and always multiply t^2 by this numerical constant, which would be about 24·85. "Student" would probably have started from $2\pi/\sqrt{\mu}$ every time. If therefore he had suddenly wanted to calculate the period of oscillation of a weight on a stretched spring he could have done it, whereas the man who only remembered that $l=24{\cdot}85t^2$ for a pendulum would be unable to tackle the problem without much more preliminary work.

His method was, of course, not necessarily the most suitable for others not aspiring to the same degree of versatility. Perhaps it is not altogether fanciful

to compare the two methods with the organic evolution of, say, the human hand, the most versatile object known, and the construction of some highly efficient but absolutely specialized piece of machinery. I do not mean to imply that he gave this explanation, or was even altogether conscious of it. When he handed over to me a routine calculation which he had done for many years, I was astonished to find that he had written out every week an almost unvarying form of words with different figures. To my question, "Why ever don't you get a printed form?" he did not reply, "Doing it from first principles every time preserves mental flexibility". He would have considered such a remark unbearably pompous. He said, "Because I'm too lazy", to which I replied, "Well, I'm too lazy not to."

To many in the statistical world "Student" was regarded as a statistical adviser to Guinness's brewery; to others he appeared to be a brewer devoting his spare time to statistics. I have tried to show that though there is some truth in both of these ideas they miss the central point, which was the intimate connexion between his statistical research and the practical problems on which he was engaged. I can imagine that many think it wasteful that a man of his undoubted genius should have been engaged in industry, yet I am sure that it is just that association with immediate practical problems which gives "Student's" work its unique character and importance relative to its small volume. On at least one occasion he was offered an academic appointment, but it is almost certain that he would not have been a successful lecturer, though perhaps a good individual teacher; nor is it likely that his research work would have flourished in more academic circumstances; his mind worked in a different way.

The work in connexion with barley breeding carried out by the Department of Agriculture in Ireland, in which Messrs Guinness took a prominent part, enabled "Student" to get that first-hand experience of yield trials and agricultural experiments generally which contributed so largely to his great knowledge of the subject. He did not merely sit in his office and calculate the results, but discussed all the details and difficulties with the Department officials, and went round all the experiments before harvest, when a "grand tour" is annually carried out by the Department, the brewery, and sometimes statisticians or others interested from England or abroad. As well as the work carried out at the actual cereal station near Cork, three or four varieties of barley are grown in $\frac{3}{4}$ or 1 acre plots at ten farms representing all the principal barley-growing districts of Ireland, so a visit to all of them entails a fairly comprehensive inspection of the crops.

"Student" took a great deal of interest in this work from the beginning and correspondence shows that he discussed the results of these tests with Karl Pearson at great length when he went to study with him at University College in 1906.

In the last ten years or so of his time in Ireland he played a leading part in these investigations, and thus had a perhaps unique opportunity of following experimental varieties from sowing through growing and harvest to malting and brewing results, and also of carrying out or supervising all the relevant mathematical work. At one time he also made some barley crosses in his own garden, and accelerated their multiplication by having one generation grown in New Zealand during our winter. These crosses were known as Student I and II, and have now been discarded as failures, the inevitable fate of the large majority. With characteristic self-effacement he was the first to point out that they were not worth going on with.

He also made frequent visits to Dr E. S. Beaven, whose work on barley breeding is well known, and discussed every aspect of yield trials with him. These visits were undoubtedly very useful, and although Dr Beaven was never tired of protesting that he was no mathematician and did not understand "magic squares" or "birds of freedom", names which he preferred to the more orthodox expressions, he had a vast experience of agricultural trials and was very quick to see the weak point of any experiment.

In spite of the quantity of work "Student" did he was never in a hurry or fussed; this was largely due to the absence of lag when he turned his mind to a new subject; unfortunately others were not always equal to this. He would ring one up on the telephone and plunge straight into some subject which might have been discussed some days previously. The slower-witted listener would probably lose the thread of his discourse before realizing what it was about and would ignominiously have to ask him to begin again. I have many times seen him hard at it on a Monday morning, but at first meeting it was always "How did the sailing go?" "Well, did you catch any fish?", and he would recount any notable event of his own week-end before plunging into the very middle of some subject. I never heard him say "I'm busy".

"Student" had many correspondents, mostly agricultural and other experimenters, in different parts of the world. He took immense pains with these and often explained points to them at great length when he could easily have given a reference. His letters contain some of his clearest writing, and the more difficult points are often better elucidated than in his published papers.

Karl Pearson emphasized the fact that a statistician must advise others on their own subject, and so may incur the accusation of butting in without adequate knowledge. "Student" was particularly expert at avoiding any such disagreement; usually he was such an enthusiastic learner of the other's subject that the fact that he was giving advice escaped notice.

The reader will by now have realized that "Student" did a very large quantity of ordinary routine as well as his statistical work in the brewery, and all that in addition to consultative statistical work and to preparing his various published

papers. It might thus be thought that he could have done nothing else but eat and sleep when at home; this, however, was far from being the case, and he had a great many domestic and sporting interests. He was a keen fruit-grower and specialized in pears. He was also a good carpenter, and built a number of boats; the last, which was completed in 1932, and on whose maiden voyage I had the honour to be nearly frozen to death, was equipped with a rudder at each end by means of which the direction and speed of drift could be adjusted—an advantage which will be readily appreciated by fly-fishermen. This boat with its arrangement of rudders was described in the *Field* of 28 March 1936. In his carpentry he showed preferences analogous to his mathematical ones previously mentioned; he disliked complicated or specific tools, and liked to do anything possible with a pen-knife. On one occasion, seeing him countersinking screw-holes with a pocket-knife, I offered him a proper countersink bit which I had with me, but he declined it with some embarrassment, as he would not have liked to explain or perhaps could not have explained why he preferred using the pen-knife. Out of doors he was an energetic walker and also cycled extensively in the pre-war period. He did a lot of sailing and fishing. For his last boat he had a most unconventional sail, which cannot be exactly described under any of the usual categories; it was illustrated in the *Field* article referred to above.

In fishing he was an efficient performer; he used to hold that only the size and general lightness or darkness of a fly were important; the blue wings, red tails and so on being only to attract the fisherman to the shop. This view was more revolutionary when I first heard it than it is now. He was a sound though not spectacular shot, and was well above the average on skates. Until the accident to his leg in 1934 he was quite a regular golfer, and once went round a fairly difficult course in 85 strokes and $1\frac{1}{4}$ hours by himself. He used a remarkable collection of old clubs dating at least from the beginning of the century. In the last few years since his accident he took up bowls with great keenness, and induced many other people to play as well. One of his last visits to Ireland was with a team which he had organized at the new brewery at Park Royal.

On top of all this he knew as much as most people of the affairs of the world in general and of what was going on about him. It became very difficult to imagine how he found 24 hours in any way a sufficient length for the day. His wife certainly organized things so that the minimum amount of time was wasted, but even so few people could approach such activity in quantity or diversity.

In personal relationships he was very kindly and tolerant and absolutely devoid of malice. He rarely spoke about personal matters but when he did his opinion was well worth listening to and not in the least superficial.

In the summer of 1934 he had a motor accident and broke the neck of his femur. He had to lie up for three months, of course working at statistics, and was a semi-cripple for a year. This was particularly irksome for such an active man,

as was the sheer unnecessariness of the accident, for he ran into a lamp-post on a straight road, through looking down to adjust some stuff he was carrying; but with great hard work and persistence he eventually reduced the disability to a slight limp.

At the end of 1935 he left Ireland to take charge of the new Guinness brewery in London, and I saw comparatively little of him after that. The departure from Ireland of "Student" and his family was a great loss to many who had experienced their hospitality. His work in London was necessarily very hard and accompanied by all the vexations inevitably associated with a big undertaking in its first stages, before any settled routine has been established; nevertheless, he still found time to continue his statistical work and wrote several papers.

His death at the comparatively early age of 61 was not only a heavy blow to his family and friends, but a great loss to statistics, as his mind retained its full vigour, and he would undoubtedly have continued to work for many more years.

I am very conscious of the inadequacy of this sketch, which cannot hope to convey more than a faint impression of his unique personal quality to those who did not know him, but it will have served its purpose if it helps any readers to grasp the essential unity and directness of the personality which lay behind such widely varied manifestations.

LAUNCE McMULLEN

1

ON THE ERROR OF COUNTING WITH A HAEMACYTOMETER

[*Biometrika*, V (1907), p. 351]

When counting yeast cells or blood corpuscles with a haemacytometer there are two main sources of error: (1) the drop taken may not be representative of the bulk of the liquid; (2) the distribution of the cells or corpuscles over the area which is examined is never absolutely uniform, so that there is an "error of random sampling".

With the first source of error we are concerned only to this extent; that when the probable error of random sampling is known we can tell whether the various drops taken show significant differences. What follows is concerned with the distribution of particles throughout a liquid, as shown by spreading it in a thin layer over a measured surface and counting the particles per unit area.

Theoretical Consideration

Suppose the *whole* liquid to have been well mixed and spread out in a thin layer over N units of area (in the haemacytometer the usual thickness is 0·01 mm. and the unit area of $\frac{1}{400}$ sq. mm.).

Let the particles subside and let there be on an average m particles per unit area, that is Nm altogether. Then, assuming the liquid has been properly mixed, a given particle will have an equal chance of falling on any unit area:

i.e. the chance of its falling in a given unit area is $1/N$ and of its not doing so $1-1/N$.

Consequently, considering all the mN particles, the chances of 0, 1, 2, 3, ... particles falling on a given area are given by the terms of the binomial

$$\left\{\left(1-\frac{1}{N}\right)+\frac{1}{N}\right\}^{mN},$$

and if M unit areas be considered the distribution of unit areas containing 0, 1, 2, 3, ... particles is given by $M\left\{\left(1-\frac{1}{N}\right)+\frac{1}{N}\right\}^{mN}$.

Now in practice N is to be measured in millions and may be taken as infinite. Let us find the limit when N is infinite of the general term of this expansion.

The $(r+1)$th term is

$$\left(1-\frac{1}{N}\right)^{mN-r}\cdot\left(\frac{1}{N}\right)^{r}\frac{mN(mN-1)(mN-2)\dots(mN-r+1)}{r!}$$

$$=\left(1-\frac{1}{N}\right)^{mN-r}\frac{m\left(m-\frac{1}{N}\right)\left(m-\frac{2}{N}\right)\dots\left(m-\frac{r-1}{N}\right)}{r!}$$

$$=\left(1-\frac{mN-r}{N}+\frac{(mN-r)(mN-r-1)}{N^2.2!}-\dots\right.$$

$$\left.+(-1)^s\frac{(mN-r)\dots(mN-r-s+1)}{N^s.s!}+\dots\right)$$

$$\times m\frac{\left(m-\frac{1}{N}\right)\left(m-\frac{2}{N}\right)\dots\left(m-\frac{r-1}{N}\right)}{r!}.$$

But when we proceed to the limit $\frac{1}{N}, \frac{2}{N}, \dots \frac{r-1}{N}$ and $\frac{r}{N}, \frac{r+1}{N}, \dots \frac{r+s-1}{N}$ are all negligibly small compared to m, so that the expression reduces to

$$\left(1-m+\frac{m^2}{2!}-\dots+(-1)^s\frac{m^s}{s!}-\dots\right)\times\frac{m^r}{r!}=e^{-m}\times\frac{m^r}{r!}.$$

That is to say, the expansion is equal to

$$e^{-m}\left\{1+m+\frac{m^2}{2!}+\dots+\frac{m^r}{r!}+\dots\right\}.$$

Hence it is this distribution with which we are concerned.

The first moment about the origin, O, taken at zero number of particles is

$$e^{-m}\left\{m+\frac{2m^2}{2!}+\frac{3m^3}{3!}+\dots+\frac{rm^r}{r!}+\dots\right\}$$

$$=me^{-m}\left\{1+\frac{m}{1!}+\frac{m^2}{2!}+\dots+\frac{m^{r-1}}{(r-1)!}+\dots\right\}$$

$$=m\times\text{total frequency}.$$

Hence the mean is at m.

The second moment about the point O is

$$e^{-m}\left\{m+\frac{2^2m^2}{2!}+\frac{3^2m^3}{3!}+\dots+\frac{r^2m^r}{r!}+\dots\right\}$$

$$=e^{-m}\left\{m+\frac{2m^2}{1!}+\frac{3m^3}{2!}+\dots+\frac{rm^r}{(r-1)!}+\dots\right\}$$

$$=e^{-m}\left\{m+\frac{m^2}{1!}+\dots+\frac{m^r}{(r-1)!}+\dots+m^2+\frac{2m^3}{2!}+\dots+\frac{(r-1)m^r}{(r-1)!}+\dots\right\}$$

$$=(m+m^2)\times\text{total frequency}.$$

Hence the second moment coefficient about the mean

$$\mu_2 = m + m^2 - m^2 = m.$$

By similar* methods the moment coefficients up to μ_6 were obtained, as follows:

$$\mu_1' = m.$$
$$\mu_2 = m.$$
$$\mu_3 = m.$$
$$\mu_4 = 3m^2 + m.$$
$$\mu_5 = 10m^2 + m.$$
$$\mu_6 = 15m^3 + 25m^2 + m.$$

Hence
$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{1}{m},$$

and
$$\beta_2 = \frac{\mu_4}{\mu_2^2} = 3 + \frac{1}{m}.$$

It will be observed that the limit to which this distribution approaches as m becomes infinite is the normal curve with its β_1, β_3, β_5, etc. all equal to 0, and $\beta_2 = 3$, $\beta_4 = 15$, etc.

Further, any binomial $(p+q)^n$ can be put into the form $(p+q)^{nq/q}$, and if q be small and nq not large it approaches the distribution just given.

Thus if 1000 $(\frac{99}{100} + \frac{1}{100})^{500}$ be expanded, the greatest difference between any of its terms and the corresponding term of $1000\, e^{-5}\left(1 + 5 + \frac{5^2}{2!} + \ldots + \frac{5^r}{r!} + \ldots\right)$

* The evaluation of the moments about the point O will be found to depend on the expansion of r^n in the form

$$r^n = r\left\{\frac{(r-1)!}{(r-n-2)!} + a_1\frac{(r-1)!}{(r-n-1)!} + a_2\frac{(r-1)!}{(r-n)!} + \ldots + a_{n+1}\frac{(r-1)!}{(r-1)!}\right\}$$
$$= r\left\{\frac{1}{(r-n-2)!} + \frac{a_1}{(r-n-1)!} + \frac{a_2}{(r-n)!} + \ldots + \frac{a_{n+1}}{(r-1)!}\right\}(r-1)!.$$

Then if we form the series for $n+1$ from this it will be found that the following relations hold between a_1, a_2, a_3, etc. and the corresponding coefficients for $n+1$, A_1, A_2, A_3, etc.:

$$A_1 = a_1 + n,$$
$$A_2 = a_2 + (n-1)a_1,$$
$$A_p = a_p + (n-p+1)a_{p-1}.$$

From these equations we can write down any number of moments about the point O in turn, and from these may be found the moments about the mean by the ordinary formulae.

The moments may also be deduced from the point binomial $(p+q)^{nq/q}$ when q is small and n large and $nq = m$, i.e. $p = 1$, $q = 0$, $nq = m$. We have

$$\mu_1' = nq = m,$$
$$\mu_2 = npq = m,$$
$$\mu_3 = npq(p-q) = m,$$
$$\mu_4 = npq\{1 + 3(n-2)pq\} = m(1+3m) = 3m^2 + m.$$

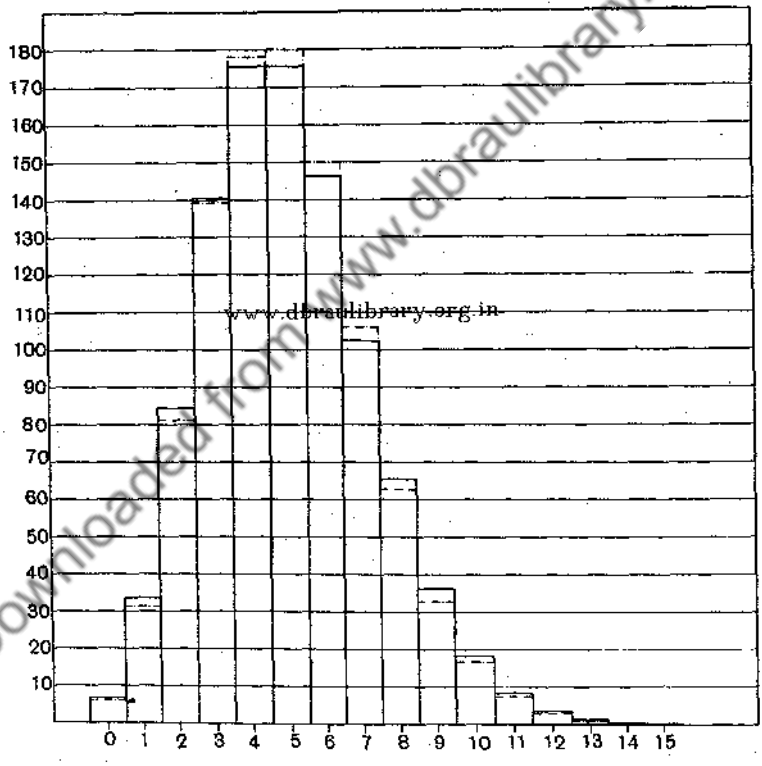
is never as much as 1, being about 0·8 for the term $1000\, e^{-5}\frac{5^5}{5!}$, which is 175·5 against 176·3 from the binomial.

Diagram I compares $1000\, e^{-5}\left(1+5+\frac{5^2}{2!}+\ldots+\frac{5^r}{r!}+\ldots\right)$ with the binomial $1000\,(\frac{19}{20}+\frac{1}{20})^{100}$, which of course differ, but not by very much.

DIAGRAM I. Comparison of the Exponential and Binomial Expansions

Firm line represents $1000\, e^{-5}\left(1+5+\ldots+\frac{5^r}{r!}+\ldots\right)$

Broken line represents $1000\left(\frac{19}{20}+\frac{1}{20}\right)^{100}$



In applying this to actual cases it must be noted that we have not taken into account any "interference" between the particles; there has been supposed the same chance of a particle falling on an area which already has several particles as on one altogether unoccupied. Clearly if m be large this will not be the case, but with the dilutions usually employed this is not of any importance.

It will be shown that the actual distributions which were tested do not diverge widely from this law, so we will consider the probable error of random sampling on the supposition that they follow it.

We have seen that $\mu_2 = m$.

Hence the standard deviation $= \sqrt{m}$.

So that if we have counted M unit areas the probable error of our mean (m) is $0{\cdot}67449\sqrt{\frac{m}{M}}$.

If we are working with a haemacytometer in which the volume over each square is $\frac{1}{40000}$ mm. there will be 40,000,000 m particles per c.c. and the probable error will be $40{,}000{,}000 \times 0{\cdot}67449 \times \sqrt{\frac{m}{M}}$.

Suppose now that we dilute the liquid to q times its bulk, we shall then have m/q particles per square, and if we count M squares as before, our probable error for the number of particles per c.c. in the original solution will be $40{,}000{,}000 \times 0{\cdot}67449 \times q\sqrt{\left(\frac{m}{q} \times \frac{1}{M}\right)}$. That is $40{,}000{,}000 \times 0{\cdot}67449\sqrt{\frac{mq}{M}}$.

That is, we shall have to count qM squares in order to be as accurate as before.

So that the same accuracy is obtained by counting the same number of particles whatever the dilution, or, to look at it from a slightly different point of view, whatever be the size of the unit of area adopted.

Hence the most accurate way is to dilute the solution to the point at which the particles may be counted most rapidly, and to count as many as time permits: then the probable error of the mean is $0{\cdot}67449\sqrt{\frac{m}{M}}$, where m is the mean and M is the number of unit areas counted over, squares, columns of squares, microscope fields, or whatever unit be selected.

But owing to the difficulty of obtaining a drop representative of the bulk of the liquid the larger errors will probably be due to this cause, and it is usual to take several drops: if two of these differ in their means by a significant amount compared with the probable error (which is $0{\cdot}67449\sqrt{\left(\frac{m_1+m_2}{M}\right)}$, where m_1, m_2 are the means and M the number of unit areas counted), it is probable that one at least of the drops does not represent the bulk of the solution.

Experimental Work

This theoretical work was tested on four distributions* which had been counted over the whole 400 squares of the haemacytometer. The particles counted were yeast cells which were killed by adding a little mercuric chloride to the water in which they had been shaken up. A small quantity of this was mixed with a 10 % solution of gelatine, and after being well stirred up drops were put on the haemacytometer. This was then put on a plate of glass kept at a temperature just above the setting point of gelatine and allowed to cool slowly till the gelatine had set. Four different concentrations were used.

* One of these is given in Table I.

In this way it was possible to count at leisure without fear of the cells straying from one square to another owing to accidental vibrations. A few cells stuck here and there to the cover glass, but, as they appeared to be fairly uniformly distributed and were very few compared with those that sank to the bottom, they were neglected: had the object of the experiment been to find the number of cells present they would have been counted by microscope fields, and correction made for them; but in our case they were considered to belong to a different "population" to those which sank.

Those cells which touched the bottom and right-hand lines of a square were considered to belong to the square; a convention of this kind is necessary as the cells have a tendency to settle on the lines.

There was some difficulty owing to the buds of some cells remaining undetached in spite of much shaking. In such cases an obvious bud was not counted, but sometimes, no doubt, a bud was counted as a separate cell, which slightly increases the number of squares with large numbers in them.

In order to test whether there was any local lack of homogeneity the correlation was determined between the number of cells on a square and the number of cells on each of the four squares nearest it; if from any cause there had been a tendency to lie closer together in some parts than in others this correlation would have been significantly positive.

Distributions 3 and 4 were tested in this way (Table II), with the result that the correlation coefficients were $0{\cdot}016 \pm 0{\cdot}037$ and $0{\cdot}015 \pm 0{\cdot}037$. This is satisfactory as showing that there is no very great difficulty in putting the drop on to the slide so as to be able to count at any point and in any order; as good a result may be expected from counting a column as from counting the same number of squares at random.

The actual distributions of cells are given below, and compared with those calculated on the supposition that they are random samples from a population following the law which we have investigated: the probability P of a worse fit occurring by chance is then found.

I. Mean $= 0{\cdot}6825$: $\mu_2 = 0{\cdot}8117$: $\mu_3 = 1{\cdot}0876$.

Containing	0	1	2	3	4	5 cells
Actual	213	128	37	18	3	1
Calculated	202	138	47	11	1·84	0·24
					2	

Whence $\chi^2 = 9{\cdot}92$ and $P = 0{\cdot}04$.

Best-fitting binomial $(1{\cdot}1893 - 0{\cdot}1893)^{-3{\cdot}6054} \times 400$ for which $P = 0{\cdot}52$.

II. Mean $= 1{\cdot}3225$: $\mu_2 = 1{\cdot}2835$: $\mu_3 = 1{\cdot}3574$.

Containing	0	1	2	3	4	5	6 cells
Actual	103	143	98	42	8	4	2
Calculated	106	141	93	41	14	4	1

Whence $\chi^2 = 3{\cdot}98$ and $P = 0{\cdot}68$.

Best-fitting binomial $(0{\cdot}97051 + 0{\cdot}02949)^{46{\cdot}2084} \times 400$ for which $P = 0{\cdot}72$.

III. Mean = 1·80: $\mu_2 = 1·96$: $\mu_3 = 2·529$.

Containing	0	1	2	3	4	5	6	7	8	9 cells
Actual	75	103	121	54	30	13	2	1	0	1
Calculated	66	119	107	64	29	10	3		1	

Whence $\chi^2 = 9·03$ and $P = 0·25$.
Best-fitting binomial $(1·0889 - 0·0889)^{-29·2473} \times 400$ for which $P = 0·37$.

IV. Mean = 4·68: $\mu_2 = 4·46$: $\mu_3 = 4·98$.

Containing	0	1	2	3	4	5	6	7	8	9	10	11	12 cells
Actual	0	20	43	53	86	70	54	37	18	10	5	2	2
Calculated	4	17	41	63	74	70	54	36	21	11	5	2	1

Whence $\chi^2 = 9·72$ and $P = 0·64$.
Best-fitting binomial $(0·9525 + 0·0475)^{93·53} \times 400$ for which $P = 0·68$.

These results are given graphically in Diagram II, on the next page.

It is possible to fit a point binomial from the mean and the second moment according to the two equations $\mu_1' = nq$, $\mu_2 = npq$, and these point binomials fit the observations better than the exponential series, but the constants have no physical meaning except that $nq = m$. And since the exponential series is a particular form of the point binomial and is fitted from one constant, while two are used for the *ad hoc* binomial, this better fit was only to be expected.

It will be noticed that in both I and III the second moment is greater than the mean, due to an excess over the calculated among the high numbers in the tail of the distribution. As was pointed out before, the budding of the yeast cell increases these high numbers, and there is also probably a tendency to stick together in groups which was not altogether abolished even by vigorous shaking.

In any case, the probabilities 0·04, 0·68, 0·25 and 0·64, though not particularly high, are not at all unlikely in four trials, supposing our theoretical law to hold, and we are not likely to be very far wrong in assuming it to do so.

Let us now apply it to a practical problem: for some purposes it is customary to estimate the concentration of cells and then dilute so that each two drops of liquid contain on an average one cell. Different flasks are then seeded with one drop of the liquid in each, and then "most of those flasks which show growths are pure cultures".

The exact distribution is given by

$$e^{-\frac{1}{2}}\left(1 + \frac{1}{2} + \frac{(\frac{1}{2})^2}{2!} + \frac{(\frac{1}{2})^3}{3!} + \ldots\right),$$

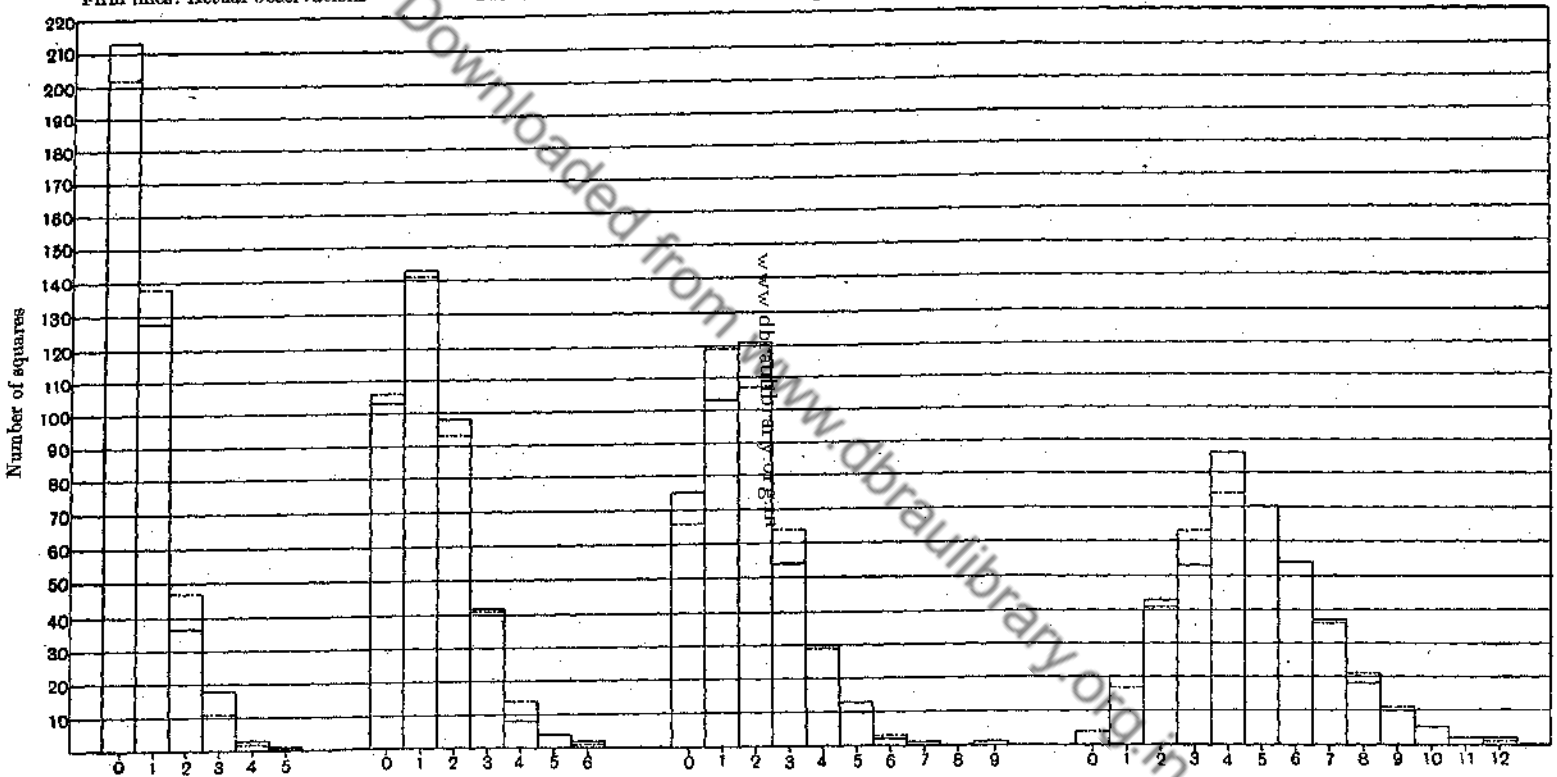
which is

No. of yeast cells	*0*	*1*	*2*	*3*	*4*
Percentage frequency	60·65	30·33	7·58	1·26	0·16

or approximately three-quarters of those which show growth are pure cultures.

DIAGRAM II. Distribution of 400 Squares

Firm lines: Actual observations Broken lines: Calculated from the exponential series Where they coincide the firm line alone is given



Conclusions

We have seen that the distribution of small particles in a liquid follows the law

$$e^{-m}\left\{1+m+\frac{m^2}{2!}+\ldots+\frac{m^r}{r!}+\ldots\right\};$$

where m is the mean number of particles per unit volume* and the various terms in the series give the chances that a given unit volume contains 0, 1, 2, ..., r, ... particles. We have also seen that this series represents the limit to which any point binomial $(p+q)^n$ approaches when q is small, insomuch that even $(\frac{19}{20}+\frac{1}{20})^{100}\times 1000$ is represented by $e^{-5}\left(1+5+\frac{5^2}{2!}+\ldots+\frac{5^r}{r!}+\ldots\right)\times 1000$ with a maximum error of about 4·5 in 180.

For the rough calculation of odds with n small compared to $1/q$ the exponential series may be used instead of the binomial as being less laborious.

Finally, we have found that the standard deviation of the mean number of particles per unit volume is $\sqrt{\frac{m}{M}}$, where m is the mean number and M the number of unit volumes counted, so that the criterion of whether two solutions contain different numbers of cells is whether m_1-m_2 is significant compared with $0{\cdot}67449\sqrt{\left(\frac{m_1}{M_1}+\frac{m_2}{M_2}\right)}$.

Table I

Distribution of Yeast Cells over 1 *sq. mm. divided into* 400 *squares*

2	2	4	4	4	5	2	4	7	7	4	7	5	2	8	6	7	4	3	4
3	3	2	4	2	5	4	2	8	6	3	6	6	10	8	3	5	6	4	4
7	9	5	2	7	4	4	2	4	4	4	3	5	6	5	4	1	4	2	6
4	1	4	7	3	2	3	5	8	2	9	5	3	9	5	5	2	4	3	4
4	1	5	9	3	4	4	6	6	5	4	6	5	5	4	3	5	9	6	4
4	4	5	10	4	4	3	8	3	2	1	4	1	5	6	4	2	3	3	3
3	7	4	5	1	8	5	7	9	5	8	9	5	6	6	4	3	7	4	4
7	5	6	3	6	7	4	5	8	6	3	3	4	3	7	4	4	4	5	3
8	10	6	3	3	6	5	2	5	3	11	3	7	4	7	3	5	5	3	4
1	3	7	2	5	5	5	3	3	4	6	5	6	1	6	4	4	4	6	4
4	2	5	4	8	6	3	4	6	5	2	6	6	1	2	2	2	5	2	2
5	9	3	5	6	4	6	5	7	1	3	6	5	4	2	8	9	5	4	3
2	2	11	4	6	6	4	6	2	5	3	5	7	2	6	5	5	1	2	7
5	12	5	8	2	4	2	1	6	4	5	1	2	9	1	3	4	7	3	6
5	6	5	4	4	5	2	7	6	2	7	3	5	4	4	5	4	7	5	4
8	4	6	6	5	3	3	5	7	4	5	5	5	6	10	2	3	8	3	5
6	6	4	2	6	6	7	5	4	5	8	6	7	6	4	2	6	1	1	4
7	2	5	7	4	6	4	5	1	5	10	8	7	5	4	6	4	4	7	5
4	3	1	6	2	5	3	3	3	7	4	3	7	8	4	7	3	1	4	4
7	6	7	2	4	5	1	3	12	4	2	2	8	7	6	7	6	3	5	4

* The prism standing on unit area.

It must be noted, however, that the probable error will always be greater than that calculated on this formula when for any reason the organisms occur as aggregates of varying size.

In conclusion, I should like to thank Prof. Adrian J. Brown, of Birmingham University, for his valuable advice and assistance in carrying out the experimental part of the inquiry.

TABLE II

"Centre" squares

"Adjacent" squares	1	2	3	4	5	6	7	8	9	10	11	12	Totals
1	6	6	9	15	15	9	4	3	2	—	—	—	69
2	6	14	17	31	24	17	10	5	6	2	1	1	134
3	8	15	25	32	37	20	15	7	7	1	4	—	171
4	18	34	33	45	48	41	22	7	5	4	1	—	258
5	15	24	37	47	39	37	18	12	11	4	1	2	247
6	9	17	25	39	34	32	14	8	2	4	1	1	186
7	5	12	14	21	19	16	9	7	3	—	—	—	106
8	3	5	7	8	12	8	6	1	3	4	—	—	57
9	2	6	7	5	10	2	2	3	—	1	—	—	38
10	—	1	1	4	4	4	—	3	—	1	—	—	18
11	—	1	4	1	1	1	—	—	—	—	—	—	8
12	—	1	1	—	1	1	—	—	—	—	—	—	4
Totals	72	136	180	248	244	188	100	56	40	20	8	4	1296

Mean of "Centre" squares, 4·8821; S.D. 2·139.
Mean of "Adjacent" squares, 4·7014; S.D. 2·116.
$r = +0{\cdot}016 \pm 0{\cdot}037$.

Correlation table between the number of cells in a square and the numbers of cells in the four adjacent squares taken all over Table I.

2

THE PROBABLE ERROR OF A MEAN

[*Biometrika*, VI (1908), p. 1]

Introduction

Any experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a greater number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information as to the value of the mean, but if our sample be small, we have two sources of uncertainty: (1) owing to the "error of random sampling" the mean of our series of experiments deviates more or less widely from the mean of the population, and (2) the sample is not sufficiently large to determine what is the law of distribution of individuals. It is usual, however, to assume a normal distribution, because, in a very large number of cases, this gives an approximation so close that a small sample will give no real information as to the manner in which the population deviates from normality: since some law of distribution must be assumed it is better to work with a curve whose area and ordinates are tabled, and whose properties are well known. This assumption is accordingly made in the present paper, so that its conclusions are not strictly applicable to populations known not to be normally distributed; yet it appears probable that the deviation from normality must be very extreme to lead to serious error. We are concerned here solely with the first of these two sources of uncertainty.

The usual method of determining the probability that the mean of the population lies within a given distance of the mean of the sample is to assume a normal distribution about the mean of the sample with a standard deviation equal to $s/\sqrt{n}$, where s is the standard deviation of the sample, and to use the tables of the probability integral.

But, as we decrease the number of experiments, the value of the standard deviation found from the sample of experiments becomes itself subject to an increasing error, until judgments reached in this way may become altogether misleading.

In routine work there are two ways of dealing with this difficulty: (1) an experiment may be repeated many times, until such a long series is obtained that the standard deviation is determined once and for all with sufficient accuracy. This value can then be used for subsequent shorter series of similar experiments. (2) Where experiments are done in duplicate in the natural course of the work, the mean square of the difference between corresponding pairs is equal to the standard deviation of the population multiplied by $\sqrt{2}$. We can thus combine together several series of experiments for the purpose of determining the standard deviation. Owing however to secular change, the value obtained is nearly always too low, successive experiments being positively correlated.

There are other experiments, however, which cannot easily be repeated very often; in such cases it is sometimes necessary to judge of the certainty of the results from a very small sample, which itself affords the only indication of the variability. Some chemical, many biological, and most agricultural and large-scale experiments belong to this class, which has hitherto been almost outside the range of statistical inquiry.

Again, although it is well known that the method of using the normal curve is only trustworthy when the sample is "large", no one has yet told us very clearly where the limit between "large" and "small" samples is to be drawn.

The aim of the present paper is to determine the point at which we may use the tables of the probability integral in judging of the significance of the mean of a series of experiments, and to furnish alternative tables for use when the number of experiments is too few.

The paper is divided into the following nine sections:

I. The equation is determined of the curve which represents the frequency distribution of standard deviations of samples drawn from a normal population.

II. There is shown to be no kind of correlation between the mean and the standard deviation of such a sample.

III. The equation is determined of the curve representing the frequency distribution of a quantity z, which is obtained by dividing the distance between the mean of a sample and the mean of the population by the standard deviation of the sample.

IV. The curve found in I is discussed.

V. The curve found in III is discussed.

VI. The two curves are compared with some actual distributions.

VII. Tables of the curves found in III are given for samples of different size.

VIII and IX. The tables are explained and some instances are given of their use.

X. Conclusions.

Section I

Samples of n individuals are drawn out of a population distributed normally, to find an equation which shall represent the frequency of the standard deviations of these samples.

If s be the standard deviation found from a sample $x_1 x_2 \ldots x_n$ (all these being measured from the mean of the population), then

$$s^2 = \frac{S(x_1^2)}{n} - \left(\frac{S(x_1)}{n}\right)^2 = \frac{S(x_1^2)}{n} - \frac{S(x_1^2)}{n^2} - \frac{2S(x_1 x_2)}{n^2}.$$

Summing for all samples and dividing by the number of samples we get the mean value of s^2, which we will write $\bar{s}^2$:

$$\bar{s}^2 = \frac{n\mu_2}{n} - \frac{n\mu_2}{n^2} = \frac{\mu_2(n-1)}{n},$$

where μ_2 is the second moment coefficient in the original normal distribution of x: since x_1, x_2, etc. are not correlated and the distribution is normal, products involving odd powers of x_1 vanish on summing, so that $\frac{2S(x_1 x_2)}{n^2}$ is equal to 0.

If M'_R represent the Rth moment coefficient of the distribution of s^2 about the end of the range where $s^2 = 0$,

$$M'_1 = \mu_2 \frac{(n-1)}{n}.$$

Again
$$\begin{aligned} s^4 &= \left\{\frac{S(x_1^2)}{n} - \left(\frac{S(x_1)}{n}\right)^2\right\}^2 \\ &= \left(\frac{S(x_1^2)}{n}\right)^2 - \frac{2S(x_1^2)}{n}\left(\frac{S(x_1)}{n}\right)^2 + \left(\frac{S(x_1)}{n}\right)^4 \\ &= \frac{S(x_1^4)}{n^2} + \frac{2S(x_1^2 x_2^2)}{n^2} - \frac{2S(x_1^4)}{n^3} - \frac{4S(x_1^2 x_2^2)}{n^3} + \frac{S(x_1^4)}{n^4} \\ &\quad + \frac{6S(x_1^2 x_2^2)}{n^4} + \text{other terms involving odd powers of } x_1, \text{ etc. which} \\ &\qquad \text{will vanish on summation.} \end{aligned}$$

Now $S(x_1^4)$ has n terms, but $S(x_1^2 x_2^2)$ has $\frac{1}{2}n(n-1)$, hence summing for all samples and dividing by the number of samples, we get

$$\begin{aligned} M'_2 &= \frac{\mu_4}{n} + \mu_2^2\frac{(n-1)}{n} - \frac{2\mu_4}{n^2} - 2\mu_2^2\frac{(n-1)}{n^2} + \frac{\mu_4}{n^3} + 3\mu_2^2\frac{(n-1)}{n^3} \\ &= \frac{\mu_4}{n^3}\{n^2 - 2n + 1\} + \frac{\mu_2^2}{n^3}(n-1)\{n^2 - 2n + 3\}. \end{aligned}$$

Now since the distribution of x is normal, $\mu_4 = 3\mu_2^2$, hence

$$M'_2 = \mu_2^2\frac{(n-1)}{n^3}\{3n - 3 + n^2 - 2n + 3\} = \mu_2^2\frac{(n-1)(n+1)}{n^2}.$$

In a similar tedious way I find

$$M_3' = \mu_2^3 \frac{(n-1)(n+1)(n+3)}{n^3}$$

and

$$M_4 = \mu_2^4 \frac{(n-1)(n+1)(n+3)(n+5)}{n^4}.$$

The law of formation of these moment coefficients appears to be a simple one, but I have not seen my way to a general proof.

If now M_R be the Rth moment coefficient of s^2 about its mean, we have

$$M_2 = \mu_2^2 \frac{(n-1)}{n^2}\{(n+1)-(n-1)\} = 2\mu_2^2 \frac{(n-1)}{n^2},$$

$$M_3 = \mu_2^3 \left\{\frac{(n-1)(n+1)(n+3)}{n^3} - \frac{3(n-1)}{n}\cdot\frac{2(n-1)}{n^2} - \frac{(n-1)^3}{n^3}\right\}$$

$$= \mu_2^3 \frac{(n-1)}{n^3}\{n^2+4n+3-6n+6-n^2+2n-1\} = 8\mu_2^3 \frac{(n-1)}{n^3},$$

$$M_4 = \frac{\mu_2^4}{n^4}\{(n-1)(n+1)(n+3)(n+5) - 32(n-1)^2 - 12(n-1)^3 - (n-1)^4\}$$

$$= \frac{\mu_2^4 (n-1)}{n^4}\{n^3+9n^2+23n+15-32n+32-12n^2+24n-12-n^3+3n^2-3n+1\}$$

$$= \frac{12\mu_2^4(n-1)(n+3)}{n^4}.$$

Hence

$$\beta_1 = \frac{M_3^2}{M_2^3} = \frac{8}{n-1}, \quad \beta_2 = \frac{M_4}{M_2^2} = \frac{3(n+3)}{n-1},$$

$$\therefore\ 2\beta_2 - 3\beta_1 - 6 = \frac{1}{n-1}\{6(n+3) - 24 - 6(n-1)\} = 0.$$

Consequently a curve of Prof. Pearson's Type III may be expected to fit the distribution of s^2.

The equation referred to an origin at the zero end of the curve will be

$$y = Cx^p e^{-\gamma x},$$

where

$$\gamma = 2\frac{M_2}{M_3} = \frac{4\mu_2^2(n-1)n^3}{8n^2\mu_2^3(n-1)} = \frac{n}{2\mu_2}$$

and

$$p = \frac{4}{\beta_1} - 1 = \frac{n-1}{2} - 1 = \frac{n-3}{2}.$$

Consequently the equation becomes

$$y = Cx^{\frac{n-3}{2}} e^{-\frac{nx}{2\mu_2}},$$

which will give the distribution of s^2.

The area of this curve is $C\int_0^\infty x^{\frac{n-3}{2}} e^{-\frac{nx}{2\mu_2}} dx = I$ (say).

The first moment coefficient about the end of the range will therefore be

$$\frac{C\int_0^\infty x^{\frac{n-1}{2}} e^{-\frac{nx}{2\mu_2}} dx}{I} = \frac{C\left[\frac{-2\mu_2}{n} x^{\frac{n-1}{2}} e^{-\frac{nx}{2\mu_2}}\right]_{x=0}^{x=\infty}}{I} + \frac{C\int_0^\infty \frac{n-1}{n}\mu_2 x^{\frac{n-3}{2}} e^{-\frac{nx}{2\mu_2}} dx}{I}.$$

The first part vanishes at each limit and the second is equal to

$$\frac{\frac{n-1}{n}\mu_2 I}{I} = \frac{n-1}{n}\mu_2,$$

and we see that the higher moment coefficients will be formed by multiplying successively by $\frac{n+1}{n}\mu_2$, $\frac{n+3}{n}\mu_2$, etc., just as appeared to be the law of formation of M_2', M_3', M_4', etc.

Hence it is probable that the curve found represents the theoretical distribution of s^2; so that although we have no actual proof we shall assume it to do so in what follows.

The distribution of s may be found from this, since the frequency of s is equal to that of s^2 and all that we must do is to compress the base line suitably.

Now if $y_1 = \phi(s^2)$ be the frequency curve of s^2

and $y_2 = \psi(s)$ the frequency curve of s

then $y_1 d(s^2) = y_2 ds,$

or $y_2 ds = 2y_1 s\, ds,$

$$\therefore y_2 = 2sy_1.$$

Hence $$y_2 = 2Cs(s^2)^{\frac{n-3}{2}} e^{-\frac{ns^2}{2\mu_2}}$$

is the distribution of s.

This reduces to $$y_2 = 2Cs^{n-2} e^{-\frac{ns^2}{2\mu_2}}.$$

Hence $y = Ax^{n-2}e^{-\frac{nx^2}{2\sigma^2}}$ will give the frequency distribution of standard deviations of samples of n, taken out of a population distributed normally with standard deviation σ. The constant A may be found by equating the area of the curve as follows:

$$\text{Area} = A\int_0^\infty x^{n-2} e^{-\frac{nx^2}{2\sigma^2}} dx. \quad \left(\text{Let } I_p \text{ represent } \int_0^\infty x^p e^{-\frac{nx^2}{2\sigma^2}} dx.\right)$$

Then
$$I_p = \frac{\sigma^2}{n}\int_0^\infty x^{p-1}\frac{d}{dx}\left(-e^{-\frac{nx^2}{2\sigma^2}}\right) dx$$
$$= \frac{\sigma^2}{n}\left[-x^{p-1}e^{-\frac{nx^2}{2\sigma^2}}\right]_{x=0}^{x=\infty} + \frac{\sigma^2}{n}(p-1)\int_0^\infty x^{p-2} e^{-\frac{nx^2}{2\sigma^2}} dx$$
$$= \frac{\sigma^2}{n}(p-1)\, I_{p-2},$$

since the first part vanishes at both limits.

By continuing this process we find

$$I_{n-2} = \left(\frac{\sigma^2}{n}\right)^{\frac{n-2}{2}} (n-3)(n-5)\dots 3.1\, I_0$$

or
$$= \left(\frac{\sigma^2}{n}\right)^{\frac{n-3}{2}} (n-3)(n-5)\dots 4.2\, I_1$$

according as n is even or odd.

But I_0 is
$$\int_0^\infty e^{-\frac{nx^2}{2\sigma^2}} dx = \sqrt{\left(\frac{\pi}{2n}\right)}\sigma,$$

and I_1 is
$$\int_0^\infty x e^{-\frac{nx^2}{2\sigma^2}} dx = \left[-\frac{\sigma^2}{n} e^{-\frac{nx^2}{2\sigma^2}}\right]_{x=0}^{x=\infty} = \frac{\sigma^2}{n}.$$

Hence if n be even,

$$A = \frac{\text{Area}}{(n-3)(n-5)\dots 3.1 \sqrt{\left(\frac{\pi}{2}\right)\left(\frac{\sigma^2}{n}\right)^{\frac{n-1}{2}}}},$$

and if n be odd,

$$A = \frac{\text{Area}}{(n-3)(n-5)\dots 4.2\left(\frac{\sigma^2}{n}\right)^{\frac{n-1}{2}}}.$$

Hence the equation may be written

$$y = \frac{N}{(n-3)(n-5)\dots 3.1\sqrt{\left(\frac{2}{\pi}\right)}\left(\frac{n}{\sigma^2}\right)^{\frac{n-1}{2}}} x^{n-2} e^{-\frac{nx^2}{2\sigma^2}} \quad (n \text{ even})$$

or
$$y = \frac{N}{(n-3)(n-5)\dots 4.2}\left(\frac{n}{\sigma^2}\right)^{\frac{n-1}{2}} x^{n-2} e^{-\frac{nx^2}{2\sigma^2}} \quad (n \text{ odd}),$$

where N as usual represents the total frequency.

Section II

To show that there is no correlation between (a) the distance of the mean of a sample from the mean of the population and (b) the standard deviation of a sample with normal distribution.

(1) Clearly positive and negative positions of the mean of the sample are equally likely, and hence there cannot be correlation between the absolute value of the distance of the mean from the mean of the population and the standard deviation, but (2) there might be correlation between the square of the distance and the square of the standard deviation.

Let
$$u^2 = \left(\frac{S(x_1)}{n}\right)^2 \quad \text{and} \quad s^2 = \frac{S(x_1^2)}{n} - \left(\frac{S(x_1)}{n}\right)^2.$$

Then if m_1', M_1' be the mean values of u^2 and s^2, we have by the preceding part

$$M_1' = \mu_2 \frac{(n-1)}{n} \quad \text{and} \quad m_1' = \frac{\mu_2}{n}.$$

Now $$u^2s^2 = \frac{S(x_1^2)}{n}\left(\frac{S(x_1)}{n}\right)^2 - \left(\frac{S(x_1)}{n}\right)^4$$

$$= \left(\frac{S(x_1^2)}{n}\right)^2 + 2\frac{S(x_1x_2).S(x_1^2)}{n^3} - \frac{S(x_1^4)}{n^4} - \frac{6S(x_1^2x_2^2)}{n^4}$$

$-$ other terms of odd order which will vanish on summation.

Summing for all values and dividing by the number of cases we get

$$R_{u^2s^2}\sigma_{u^2}\sigma_{s^2} + m_1M_1 = \frac{\mu_4}{n^2} + \mu_2^2\frac{(n-1)}{n^2} - \frac{\mu_4}{n^3} - 3\mu_2^2\frac{(n-1)}{n^3},$$

where $R_{u^2s^2}$ is the correlation between u^2 and s^2.

$$R_{u^2s^2}\sigma_{u^2}\sigma_{s^2} + \mu_2^2\frac{(n-1)}{n^2} = \mu_2^2\frac{(n-1)}{n^3}\{3+n-3\} = \mu_2^2\frac{(n-1)}{n^2}.$$

Hence $R_{u^2s^2}\sigma_{u^2}\sigma_{s^2} = 0$, or there is no correlation between u^2 and s^2.

Section III

To find the equation representing the frequency distribution of the means of samples of n drawn from a normal population, the mean being expressed in terms of the standard deviation of the sample.

We have $y = \frac{C}{\sigma^{n-1}}s^{n-2}e^{-\frac{ns^2}{2\sigma^2}}$ as the equation representing the distribution of s, the standard deviation of a sample of n, when the samples are drawn from a normal population with standard deviation σ.

Now the means of these samples of n are distributed according to the equation

$$y = \frac{\sqrt{(n)}N}{\sqrt{(2\pi)}\sigma}e^{-\frac{nx^2}{2\sigma^2}},*$$

and we have shown that there is no correlation between x, the distance of the mean of the sample, and s, the standard deviation of the sample.

Now let us suppose x measured in terms of s, i.e. let us find the distribution of $z = x/s$.

If we have $y_1 = \phi(x)$ and $y_2 = \psi(z)$ as the equations representing the frequency of x and of z respectively, then

$$y_1dx = y_2dz = y_2\frac{dx}{s},$$

$$\therefore\ y_2 = sy_1.$$

* Airy, *Theory of Errors of Observations*, Part II, § 6.

Hence $$y = \frac{N\sqrt{(n)}\,s}{\sqrt{(2\pi)}\,\sigma} e^{-\frac{ns^2z^2}{2\sigma^2}}$$

is the equation representing the distribution of z for samples of n with standard deviation s.

Now the chance that s lies between s and $s+ds$ is

$$\frac{\int_s^{s+ds} \frac{C}{\sigma^{n-1}} s^{n-2} e^{-\frac{ns^2}{2\sigma^2}} ds}{\int_0^{\infty} \frac{C}{\sigma^{n-1}} s^{n-2} e^{-\frac{ns^2}{2\sigma^2}} ds},$$

which represents the N in the above equation.

Hence the distribution of z due to values of s which lie between s and $s+ds$ is

$$y = \frac{\int_s^{s+ds} \frac{C}{\sigma^n}\sqrt{\left(\frac{n}{2\pi}\right)} s^{n-1} e^{-\frac{ns^2(1+z^2)}{2\sigma^2}} ds}{\int_0^{\infty} \frac{C}{\sigma^{n-1}} s^{n-2} e^{-\frac{ns^2}{2\sigma^2}} ds} = \frac{\sqrt{\left(\frac{n}{2\pi}\right)} \int_s^{s+ds} s^{n-1} e^{-\frac{ns^2(1+z^2)}{2\sigma^2}} ds}{\sigma \int_0^{\infty} s^{n-2} e^{-\frac{ns^2}{2\sigma^2}} ds},$$

and summing for all values of s we have as an equation giving the distribution of z

$$y = \frac{\sqrt{\left(\frac{n}{2\pi}\right)} \int_0^{\infty} s^{n-1} e^{-\frac{ns^2(1+z^2)}{2\sigma^2}} ds}{\sigma \int_0^{\infty} s^{n-2} e^{-\frac{ns^2}{2\sigma^2}} ds}.$$

By what we have already proved this reduces to

$$y = \frac{1}{2}\frac{n-2}{n-3}\cdot\frac{n-4}{n-5}\cdots\frac{5}{4}\cdot\frac{3}{2}(1+z^2)^{-\frac{1}{2}n}, \quad \text{if } n \text{ be odd,}$$

and to $$y = \frac{1}{\pi}\frac{n-2}{n-3}\cdot\frac{n-4}{n-5}\cdots\frac{4}{3}\cdot\frac{2}{1}(1+z^2)^{-\frac{1}{2}n}, \quad \text{if } n \text{ be even.}$$

Since this equation is independent of σ it will give the distribution of the distance of the mean of a sample from the mean of the population expressed in terms of the standard deviation of the sample for any normal population.

Section IV. Some Properties of the Standard Deviation Frequency Curve

By a similar method to that adopted for finding the constant we may find the mean and moments: thus the mean is at I_{n-1}/I_{n-2},

which is equal to $$\frac{(n-2)}{(n-3)}\frac{(n-4)}{(n-5)}\cdots\frac{2}{1}\sqrt{\left(\frac{2}{\pi}\right)}\frac{\sigma}{\sqrt{n}}, \quad \text{if } n \text{ be even,}$$

or $$\frac{(n-2)}{(n-3)}\frac{(n-4)}{(n-5)}\cdots\frac{3}{2}\sqrt{\left(\frac{\pi}{2}\right)}\frac{\sigma}{\sqrt{n}}, \quad \text{if } n \text{ be odd.}$$

The second moment about the end of the range is

$$\frac{I_n}{I_{n-2}} = \frac{(n-1)\,\sigma^2}{n}.$$

The third moment about the end of the range is equal to

$$\frac{I_{n+1}}{I_{n-2}} = \frac{I_{n+1}}{I_{n-1}} \cdot \frac{I_{n-1}}{I_{n-2}}$$
$$= \sigma^2 \times \text{the mean}.$$

The fourth moment about the end of the range is equal to

$$\frac{I_{n+2}}{I_{n-2}} = \frac{(n-1)\,(n+1)}{n^2}\,\sigma^4.$$

If we write the distance of the mean from the end of the range $D\sigma/\sqrt{n}$ and the moments about the end of the range ν_1, ν_2, etc.,

then $$\nu_1 = \frac{D\sigma}{\sqrt{n}}, \quad \nu_2 = \frac{n-1}{n}\sigma_2, \quad \nu_3 = \frac{D\sigma^3}{\sqrt{n}}, \quad \nu_4 = \frac{n^2-1}{n}\sigma^4.$$

From this we get the moments about the mean:

$$\mu_2 = \frac{\sigma^2}{n}(n-1-D^2),$$

$$\mu_3 = \frac{\sigma^3}{n\sqrt{n}}\{nD - 3(n-1)\,D + 2D^3\} = \frac{\sigma^3 D}{n\sqrt{n}}\{2D^2 - 2n + 3\},$$

$$\mu_4 = \frac{\sigma^2}{n^2}\{n^2 - 1 - 4D^2 n + 6(n-1)\,D^2 - 3D^4\} = \frac{\sigma^4}{n^2}\{n^2 - 1 - D^2(3D^2 - 2n + 6)\}.$$

It is of interest to find out what these become when n is large.

In order to do this we must find out what is the value of D.

Now Wallis's expression for π derived from the infinite product value of $\sin x$ is

$$\frac{\pi}{2}(2n+1) = \frac{2^2 . 4^2 . 6^2 \ldots (2n)^2}{1^2 . 3^2 . 5^2 \ldots (2n-1)^2}.$$

If we assume a quantity $\theta\left(= a_0 + \frac{a_1}{n} + \text{etc.}\right)$ which we may add to the $2n+1$ in order to make the expression approximate more rapidly to the truth, it is easy to show that $\theta = -\frac{1}{2} + \frac{1}{16n} - \text{etc.}$, and we get

$$\frac{\pi}{2}\left(2n + \frac{1}{2} + \frac{1}{16n}\right) = \frac{2^2 . 4^2 . 6^2 \ldots (2n)^2}{1^2 . 3^2 . 5^2 \ldots (2n-1)^2}.*$$

From this we find that whether n be even or odd D^2 approximates to $n - \frac{3}{2} + \frac{1}{8n}$ when n is large.

* This expression will be found to give a much closer approximation to π than Wallis's.

Substituting this value of D we get

$$\mu_2 = \frac{\sigma^2}{2n}\left(1-\frac{1}{4n}\right), \quad \mu_3 = \frac{\sigma^3\sqrt{\left(1-\frac{3}{2n}+\frac{1}{16n^2}\right)}}{4n^2}, \quad \mu_4 = \frac{3\sigma^4}{4n^2}\left(1+\frac{1}{2n}-\frac{1}{16n^2}\right).$$

Consequently the value of the standard deviation of a standard deviation which we have found $\left(\frac{\sigma}{\sqrt{(2n)}\sqrt{\{1-(1/4n)\}}}\right)$ becomes the same as that found for the normal curve by Prof. Pearson $\{\sigma/(2n)\}$ when n is large enough to neglect the $1/4n$ in comparison with 1.

Neglecting terms of lower order than $1/n$, we find

$$\beta_1 = \frac{2n-3}{n(4n-3)}, \quad \beta_2 = 3\left(1-\frac{1}{2n}\right)\left(1+\frac{1}{2n}\right).$$

Consequently, as n increases, β_2 very soon approaches the value 3 of the normal curve, but β_1 vanishes more slowly, so that the curve remains slightly skew.

DIAGRAM I. Frequency Curve giving the Distribution of Standard Deviations of samples of 10 taken from a Normal Population

Equation $y = \frac{N}{7.5.3}\frac{10^{\frac{9}{2}}}{\sigma^9}\sqrt{\left(\frac{2}{\pi}\right)}x^8 e^{-\frac{10x^2}{2\sigma^2}}$

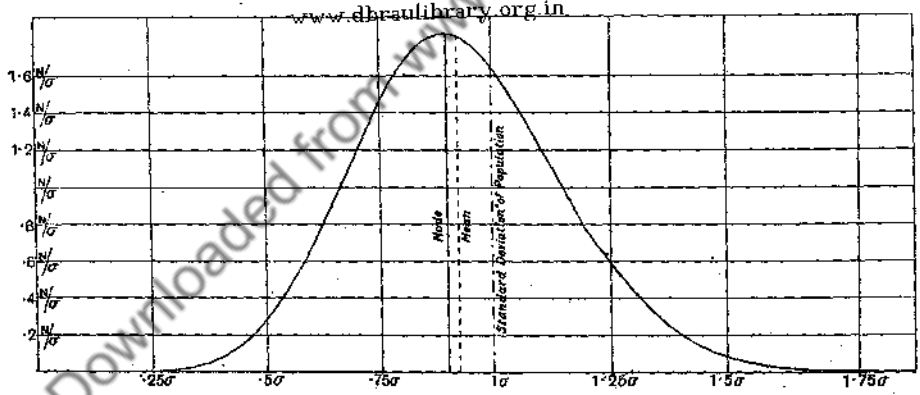


Diagram I shows the theoretical distribution of the standard deviations found from samples of 10.

$$y = \frac{N10^{\frac{9}{2}}}{7.5.3}\sqrt{\left(\frac{2}{\pi}\right)}\frac{x^8}{\sigma^9}e^{-\frac{10x^2}{2\sigma^2}}.$$

SECTION V. SOME PROPERTIES OF THE CURVE

$$y = \frac{n-2}{n-3}\cdot\frac{n-4}{n-5}\cdots\begin{pmatrix}\frac{4}{3}\cdot\frac{2}{\pi} \text{ if } n \text{ be even}\\ \frac{5}{4}\cdot\frac{3}{2}\cdot\frac{1}{2} \text{ if } n \text{ be odd}\end{pmatrix}(1+z^2)^{-\frac{1}{2}n}$$

Writing $z = \tan\theta$ the equation becomes $y = \frac{n-2}{n-3}\cdot\frac{n-4}{n-5}\ldots$ etc. $\times \cos^n\theta$, which affords an easy way of drawing the curve. Also $dz = d\theta/\cos^2\theta$.

Hence to find the area of the curve between any limits we must find

$$\frac{n-2}{n-3}\cdot\frac{n-4}{n-5}\ldots\text{etc.}\times\int\cos^{n-2}\theta\,d\theta$$

$$= \frac{n-2}{n-3}\cdot\frac{n-4}{n-5}\ldots\text{etc.}\left\{\frac{n-3}{n-2}\int\cos^{n-4}\theta\,d\theta + \left[\frac{\cos^{n-3}\theta\sin\theta}{n-2}\right]\right\}$$

$$= \frac{n-4}{n-5}\cdot\frac{n-6}{n-7}\ldots\text{etc.}\int\cos^{n-4}\theta\,d\theta + \frac{1}{n-3}\cdot\frac{n-4}{n-5}\ldots\text{etc.}\,[\cos^{n-3}\theta\sin\theta],$$

and by continuing the process the integral may be evaluated.

For example, if we wish to find the area between 0 and θ for $n = 8$ we have

$$\text{Area} = \frac{6}{5}\cdot\frac{4}{3}\cdot\frac{2}{1}\cdot\frac{1}{\pi}\int_0^\theta\cos^6\theta\,d\theta$$

$$= \frac{4}{3}\cdot\frac{2}{\pi}\cdot\int_0^\theta\cos^4\theta\,d\theta + \frac{1}{5}\cdot\frac{4}{3}\cdot\frac{2}{\pi}\cos^5\theta\sin\theta$$

$$= \frac{\theta}{\pi} + \frac{1}{\pi}\cos\theta\sin\theta + \frac{1}{3}\cdot\frac{2}{\pi}\cos^3\theta\sin\theta + \frac{1}{5}\cdot\frac{4}{3}\cdot\frac{2}{\pi}\cos^5\theta\sin\theta,$$

and it will be noticed that for $n = 10$ we shall merely have to add to this same expression the term $\frac{1}{7}\cdot\frac{6}{5}\cdot\frac{4}{3}\cdot\frac{2}{\pi}\cos^7\theta\sin\theta$.

The tables at the end of the paper give the area between $-\infty$ and z

$$\left(\text{or } \theta = -\frac{\pi}{2} \text{ and } \theta = \tan^{-1}z\right).$$

This is the same as 0·5 + the area between $\theta = 0$, and $\theta = \tan^{-1}z$, and as the whole area of the curve is equal to 1, the tables give the probability that the mean of the sample does not differ by more than z times the standard deviation of the sample from the mean of the population.

The whole area of the curve is equal to

$$\frac{n-2}{n-3}\cdot\frac{n-4}{n-5}\ldots\text{etc.}\times\int_{-\frac{1}{2}\pi}^{+\frac{1}{2}\pi}\cos^{n-2}\theta\,d\theta,$$

and since all the parts between the limits vanish at both limits this reduces to 1.

Similarly, the second moment coefficient is equal to

$$\frac{n-2}{n-3}\cdot\frac{n-4}{n-5}\ldots\text{etc.}\times\int_{-\frac{1}{2}\pi}^{+\frac{1}{2}\pi}\cos^{n-2}\theta\tan^2\theta\,d\theta$$

$$=\frac{n-2}{n-3}\cdot\frac{n-4}{n-5}\ldots\text{etc.}\times\int_{-\frac{1}{2}\pi}^{+\frac{1}{2}\pi}(\cos^{n-4}\theta-\cos^{n-2}\theta)\,d\theta$$

$$=\frac{n-2}{n-3}-1=\frac{1}{n-3}.$$

Hence the standard deviation of the curve is $1/\sqrt{(n-3)}$. The fourth moment coefficient is equal to

$$\frac{n-2}{n-3}\cdot\frac{n-4}{n-5}\ldots\text{etc.}\times\int_{-\frac{1}{2}\pi}^{+\frac{1}{2}\pi}\cos^{n-2}\theta\,\tan^4\theta\,d\theta$$

$$=\frac{n-2}{n-3}\cdot\frac{n-4}{n-5}\ldots\text{etc.}\times\int_{-\frac{1}{2}\pi}^{+\frac{1}{2}\pi}(\cos^{n-6}\theta-2\cos^{n-4}\theta+\cos^{n-2}\theta)\,d\theta$$

$$=\frac{n-2}{n-3}\cdot\frac{n-4}{n-5}-\frac{2(n-2)}{n-3}+1=\frac{3}{(n-3)(n-5)}.$$

The odd moments are of course zero, as the curve is symmetrical, so

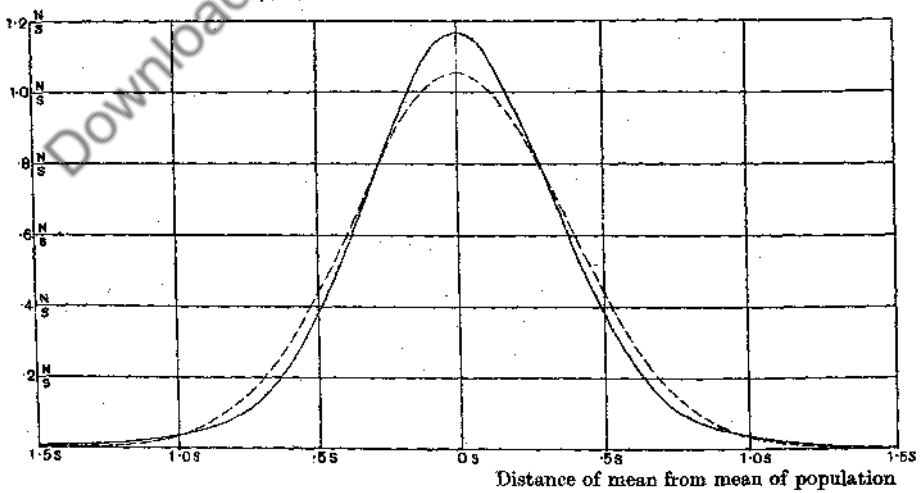
$$\beta_1=0,\quad \beta_2=\frac{3(n-3)}{n-5}=3+\frac{6}{n-5}.$$

Hence as n increases the curve approaches the normal curve whose standard deviation is $1/\sqrt{(n-3)}$.

β_2, however, is always greater than 3, indicating that large deviations are more common than in the normal curve.

DIAGRAM II. Solid curve $y=\frac{N}{s}\times\frac{8}{7}\cdot\frac{6}{5}\cdot\frac{4}{3}\cdot\frac{2}{\pi}\cos^{10}\theta,\quad x/s=\tan\theta$

Broken line curve $y=\frac{\sqrt{7}.N}{\sqrt{(2\pi)}.s}e^{-\frac{7x^2}{2s^2}}$, the normal curve with the same standard deviation



I have tabled the area for the normal curve with standard deviation $1/\sqrt{7}$ so as to compare with my curve for $n = 10$.* It will be seen that odds laid according to either table would not seriously differ till we reach $z = 0{\cdot}8$, where the odds are about 50 to 1 that the mean is within that limit: beyond that the normal curve gives a false feeling of security, for example, according to the normal curve it is 99,986 to 14 (say 7000 to 1) that the mean of the population lies between $-\infty$ and $+1{\cdot}3s$, whereas the real odds are only 99,819 to 181 (about 550 to 1).

Now 50 to 1 corresponds to three times the probable error in the normal curve and for most purposes would be considered significant; for this reason I have only tabled my curves for values of n not greater than 10, but have given the $n = 9$ and $n = 10$ tables to one further place of decimals. They can be used as foundations for finding values for larger samples.†

The table for $n = 2$ can be readily constructed by looking out $\theta = \tan^{-1}z$ in Chambers's tables and then $0{\cdot}5 + \theta/\pi$ gives the corresponding value.

Similarly $\frac{1}{2}\sin\theta + 0{\cdot}5$ gives the values when $n = 3$.

There are two points of interest in the $n = 2$ curve. Here s is equal to half the distance between the two observations. $\tan^{-1}\frac{s}{s} = \frac{\pi}{4}$, so that between $+s$ and $-s$ lies $2 \times \frac{\pi}{4} \times \frac{1}{\pi}$ or half the probability, i.e. if two observations have been made and we have no other information, it is an even chance that the mean of the (normal) population will lie between them. On the other hand the second moment coefficient is

$$\frac{1}{\pi}\int_{-\frac{1}{2}\pi}^{+\frac{1}{2}\pi} \tan^2\theta\, d\theta = \frac{1}{\pi}\Big[\tan\theta - \theta\Big]_{-\frac{1}{2}\pi}^{+\frac{1}{2}\pi} = \infty,$$

or the standard deviation is infinite while the probable error is finite.

Section VI. Practical Test of the foregoing Equations

Before I had succeeded in solving my problem analytically, I had endeavoured to do so empirically. The material used was a correlation table containing the height and left middle finger measurements of 3000 criminals, from a paper by W. R. Macdonell (*Biometrika*, I, p. 219). The measurements were written out on 3000 pieces of cardboard, which were then very thoroughly shuffled and drawn at random. As each card was drawn its numbers were written down in a book, which thus contains the measurements of 3000 criminals in a random order. Finally, each consecutive set of 4 was taken as a sample—750 in all—and the mean, standard deviation, and correlation‡ of each sample determined. The

* See p. 29.

† E.g. if $n = 11$, to the corresponding value for $n = 9$, we add $\frac{7}{8} \times \frac{5}{6} \times \frac{3}{4} \times \frac{1}{2} \times \frac{1}{2}\cos^8\theta \sin\theta$: if $n = 13$ we add as well $\frac{9}{10} \times \frac{7}{8} \times \frac{5}{6} \times \frac{3}{4} \times \frac{1}{2} \times \frac{1}{2}\cos^{10}\theta \sin\theta$, and so on.

‡ I hope to publish the results of the correlation work shortly. [See 3 below. Ed.]

difference between the mean of each sample and the mean of the population was then divided by the standard deviation of the sample, giving us the z of Section III.

This provides us with two sets of 750 standard deviations and two sets of 750 z's on which to test the theoretical results arrived at. The height and left middle finger correlation table was chosen because the distribution of both was approximately normal and the correlation was fairly high. Both frequency curves, however, deviate slightly from normality, the constants being for height $\beta_1 = 0{\cdot}0026$, $\beta_2 = 3{\cdot}175$, and for left middle finger lengths $\beta_1 = 0{\cdot}0030$, $\beta_2 = 3{\cdot}140$, and in consequence there is a tendency for a certain number of larger standard deviations to occur than if the distributions were normal. This, however, appears to make very little difference to the distribution of z.

Another thing which interferes with the comparison is the comparatively large groups in which the observations occur. The heights are arranged in 1 inch groups, the standard deviation being only 2·54 inches: while the finger lengths were originally grouped in millimetres, but unfortunately I did not at the time see the importance of having a smaller unit and condensed them into 2 millimetre groups, in terms of which the standard deviation is 2·74.

Several curious results follow from taking samples of 4 from material disposed in such wide groups. The following points may be noticed:

(1) The means only occur as multiples of 0·25.

(2) The standard deviations occur as the square roots of the following types of numbers: n, $n+0{\cdot}19$, $n+0{\cdot}25$, $n+0{\cdot}50$, $n+0{\cdot}69$, $2n+0{\cdot}75$.

(3) A standard deviation belonging to one of these groups can only be associated with a mean of a particular kind; thus a standard deviation of $\sqrt{2}$ can only occur if the mean differs by a whole number from the group we take as origin, while $\sqrt{1{\cdot}69}$ will only occur when the mean is at $n \pm 0{\cdot}25$.

(4) All the four individuals of the sample will occasionally come from the same group, giving a zero value for the standard deviation. Now this leads to an infinite value of z and is clearly due to too wide a grouping, for although two men may have the same height when measured by inches, yet the finer the measurements the more seldom will they be identical, till finally the chance that four men will have *exactly* the same height is infinitely small. If we had smaller grouping the zero values of the standard deviation might be expected to increase, and a similar consideration will show that the smaller values of the standard deviation would also be likely to increase, such as 0·436, when 3 fall in one group and 1 in an adjacent group, or 0·50 when 2 fall in two adjacent groups. On the other hand, when the individuals of the sample lie far apart, the argument of Sheppard's correction will apply, the real value of the standard deviation being more likely to be smaller than that found owing to the frequency in any group being greater on the side nearer the mode.

These two effects of grouping will tend to neutralize each other in their effect on the mean value of the standard deviation, but both will increase the variability.

Accordingly, we find that the mean value of the standard deviation is quite close to that calculated, while in each case the variability is sensibly greater. The fit of the curve is not good, both for this reason and because the frequency is not evenly distributed owing to effects (2) and (3) of grouping. On the other hand, the fit of the curve giving the frequency of z is very good, and as that is the only practical point the comparison may be considered satisfactory.

The following are the figures for height:

Mean value of standard deviations:	Calculated	$2{\cdot}027 \pm 0{\cdot}021$
	Observed	$2{\cdot}026$
	Difference	$= -0{\cdot}001$
Standard deviation of standard deviations:	Calculated	$0{\cdot}8556 \pm 0{\cdot}015$
	Observed	$0{\cdot}9066$
	Difference	$= +0{\cdot}0510$

Comparison of Fit. Theoretical Equation: $y = \dfrac{16 \times 750}{\sqrt{(2\pi)\sigma^3}} x^2 e^{-\frac{2x^2}{\sigma^2}}$

Scale in terms of standard deviation of population																
0 to ·1	*·1 to ·2*	*·2 to ·3*	*·3 to ·4*	*·4 to ·5*	*·5 to ·6*	*·6 to ·7*	*·7 to ·8*	*·8 to ·9*	*·9 to 1·0*	*1·0 to 1·1*	*1·1 to 1·2*	*1·2 to 1·3*	*1·3 to 1·4*	*1·4 to 1·5*	*1·5 to 1·6*	*1·6 to 1·7*	Greater than *1·7*
Calculated frequency																	
1½	10½	27	45½	64½	78½	87	88	81½	71	58	45	33	23	15	9½	5½	7
Observed frequency																	
3	14½	24½	37½	107	67	73	77	77½	64	52½	49½	35	28	12½	9	11½	7
Difference																	
+1½	+4	−2½	−8	+42½	−11½	−14	−11	−4	−7	−5½	+4½	+2	+5	−2½	−½	+6	0

Whence $\chi^2 = 48{\cdot}06$, $P = 0{\cdot}00006$ (about).

In tabling the observed frequency, values between 0·0125 and 0·0875 were included in one group, while between 0·0875 and 0·0125 they were divided over the two groups. As an instance of the irregularity due to grouping I may mention that there were 31 cases of standard deviations 1·30 (in terms of the grouping) which is 0·5117 in terms of the standard deviation of the population, and they were therefore divided over the groups 0·4 to 0·5 and 0·5 to 0·6. Had they all been counted in groups 0·5 to 0·6 χ^2 would have fallen to 29·85 and P would have risen to 0·03. The χ^2 test presupposes *random* sampling from a frequency following the given law, but this we have not got owing to the interference of the grouping.

When, however, we test the z's where the grouping has not had so much effect, we find a close correspondence between the theory and the actual result.

There were three cases of infinite values of z which, for the reasons given above, were given the next largest values which occurred, namely $+6$ or -6. The rest were divided into groups of 0·1; 0·04, 0·05 and 0·06, being divided between the two groups on either side.

The calculated value for the standard deviation of the frequency curve was 1 ($\pm$0·017), while the observed was 1·039. The value of the standard deviation is really infinite, as the fourth moment coefficient is infinite, but as we have arbitrarily limited the infinite cases we may take as an approximation $1/\sqrt{1500}$ from which the value of the probable error given above is obtained. The fit of the curve is as follows:

Comparison of Fit. Theoretical Equation: $y = \frac{2N}{\pi}\cos^4\theta,\ z = \tan\theta$

Scale of z	Less than −3·05	−3·05 to −2·05	−2·05 to −1·55	−1·55 to −1·05	−1·05 to −·75	·75 to −·45	·45 to −·15	·15 to +·15	+·15 to +·45	+·45 to +·75	+·75 to +1·05	+1·05 to +1·55	+1·55 to +2·05	+2·05 to +3·05	More than +3·05
Calculated frequency	5	9½	13½	34½	44½	78½	119	141	119	78½	44½	34½	13½	9½	5
Observed frequency	9	14½	11½	33	43½	70½	119½	151½	122	67½	49	26½	16	10	6
Difference	+4	+5	−2	−1½	−1	−8	+½	+10½	+3	−11	+4½	−8	+2½	+½	+1

Whence $\chi^8 = 12{\cdot}44$, $P = 0{\cdot}56$.

This is very satisfactory, especially when we consider that as a rule observations are tested against curves fitted from the mean and one or more other moments of the observations, so that considerable correspondence is only to be expected; while this curve is exposed to the full errors of random sampling, its constants having been calculated quite apart from the observations.

The left middle finger samples show much the same features as those of the height, but as the grouping is not so large compared to the variability the curves fit the observations more closely. Diagrams III* and IV give the standard deviations of the z's for this set of samples. The results are as follows:

Mean value of standard deviations: Calculated $2{\cdot}186 \pm 0{\cdot}023$

Observed $2{\cdot}179$

Difference $= -0{\cdot}007$

* There are three small mistakes in plotting the observed values in Diagram III, which make the fit appear worse than it really is.

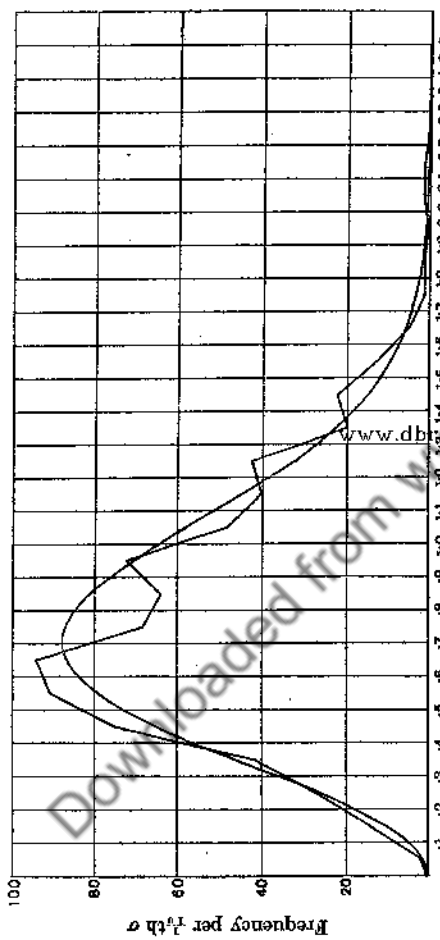


DIAGRAM III. Comparison of Calculated Standard Deviation Frequency Curve with 750 actual Standard Deviations

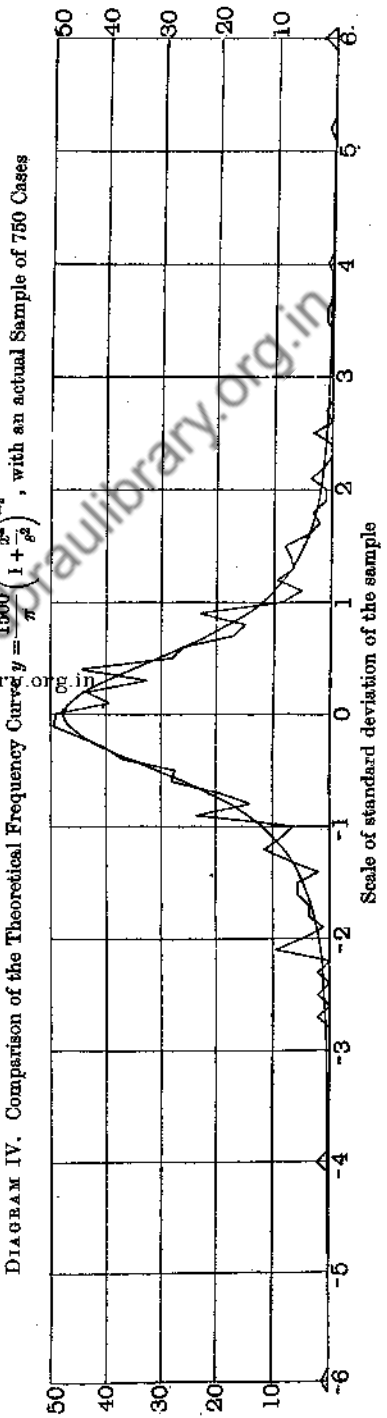


DIAGRAM IV. Comparison of the Theoretical Frequency Curve $y = \frac{1500}{\pi}\left(1+\frac{x^2}{\sigma^2}\right)^{-2}$, with an actual Sample of 750 Cases

Standard deviation of standard deviations: Calculated $0{\cdot}9224 \pm 0{\cdot}016$

Observed $0{\cdot}9802$

Difference $= +0{\cdot}0578$

Comparison of Fit. Theoretical Equation: $y = \dfrac{16 \times 750}{\sqrt{(2\pi)\sigma^3}} x^2 e^{-\frac{2x^2}{\sigma^2}}$

Scale in terms of standard deviation of population

0 to ·1	·1 to ·2	·2 to ·3	·3 to ·4	·4 to ·5	·5 to ·6	·6 to ·7	·7 to ·8	·8 to ·9	·9 to 1·0	1·0 to 1·1	1·1 to 1·2	1·2 to 1·3	1·3 to 1·4	1·4 to 1·5	1·5 to 1·6	1·6 to 1·7	Greater than 1·7
Calculated frequency																	
1½	10½	27	45½	64½	78½	87	88	81½	71	58	45	33	23	15	9½	5½	7
Observed frequency																	
2	14	27½	51	64½	91	94½	68½	65½	73	48½	40½	42½	20	22½	12	5	7½
Difference																	
+½	+3½	+½	+5½	—	+12½	+7½	−19½	−16	+2	−9½	−4½	+9½	−3	+7½	+2½	−½	+½

Whence $\chi^2 = 21{\cdot}80$, $P = 0{\cdot}19$.

Value of standard deviation: Calculated 1 ($\pm 0{\cdot}017$)

Observed 0·982

Difference $= -0{\cdot}018$

Comparison of Fit. Theoretical Equation: $y = \dfrac{2N}{\pi} \cos^4\theta,\ z = \tan\theta$

Scale of z

Less than −3·05	−3·05 to −2·05	−2·05 to −1·55	−1·55 to −1·05	−1·05 to −·75	−·75 to −·45	−·45 to −·15	−·15 to +·15	+·15 to +·45	+·45 to +·75	+·75 to +1·05	+1·05 to +1·55	+1·55 to +2·05	+2·05 to +3·05	More than +3·05
Calculated frequency														
5	9½	13½	34½	44½	78½	119	141	119	78½	44½	34½	13½	9½	5
Observed frequency														
4	15½	18	33½	44	75	122	138	120½	71	46½	36	11	9	6
Difference														
−1	+6	+4½	−1	−½	−3½	+3	−3	+1½	−7½	+2	+1½	−2½	−½	+1

Whence $\chi^2 = 7{\cdot}39$, $P = 0{\cdot}92$.

A very close fit.

We see then that if the distribution is approximately normal our theory gives us a satisfactory measure of the certainty to be derived from a small sample in both the cases we have tested; but we have an indication that a fine grouping is

of advantage. If the distribution is not normal, the mean and the standard deviation of a sample will be positively correlated, so that although both will have greater variability, yet they will tend to counteract each other, a mean deviating largely from the general mean tending to be divided by a larger standard deviation. Consequently, I believe that the table given in Section VII below may be used in estimating the degree of certainty arrived at by the mean of a few experiments, in the case of most laboratory or biological work where the distributions are as a rule of a "cocked hat" type and so sufficiently nearly normal.

Section VII. Tables of $\frac{n-2}{n-3}\frac{n-4}{n-5}\cdots\begin{pmatrix}\frac{3}{2}\cdot\frac{1}{2}\,n\text{ odd}\\ \frac{2}{1}\cdot\frac{1}{\pi}\,n\text{ even}\end{pmatrix}\int_{-\frac{1}{2}\pi}^{\tan^{-1}z}\cos^{n-2}\theta\,d\theta$

FOR VALUES OF n FROM 4 TO 10 INCLUSIVE

Together with $\frac{\sqrt{7}}{\sqrt{(2\pi)}}\int_{-\infty}^{x}e^{-\frac{7x^2}{2}}dx$ *for comparison when* $n = 10$

$z\left(=\frac{x}{s}\right)$	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$	$n=9$	$n=10$	For comparison $\left(\frac{\sqrt{7}}{\sqrt{(2\pi)}}\int_{-\infty}^{x}e^{-\frac{7x^2}{2}}dx\right)$
0·1	0·5633	0·5745	0·5841	0·5928	0·6006	0·60787	0·61462	0·60411
0·2	0·6241	0·6458	0·6634	0·6798	0·6936	0·70705	0·71846	0·70159
0·3	0·6804	0·7096	0·7340	0·7549	0·7733	0·78961	0·80423	0·78641
0·4	0·7309	0·7657	0·7939	0·8175	0·8376	0·85465	0·86970	0·85520
0·5	0·7749	0·8131	0·8428	0·8667	0·8863	0·90251	0·91609	0·90691
0·6	0·8125	0·8518	0·8813	0·9040	0·9218	0·93600	0·94732	0·94375
0·7	0·8440	0·8830	0·9109	0·9314	0·9468	0·95851	0·96747	0·96799
0·8	0·8701	0·9076	0·9332	0·9512	0·9640	0·97328	0·98007	0·98253
0·9	0·8915	0·9269	0·9498	0·9652	0·9756	0·98279	0·98780	0·99137
1·0	0·9092	0·9419	0·9622	0·9751	0·9834	0·98890	0·99252	0·99820
1·1	0·9236	0·9537	0·9714	0·9821	0·9887	0·99280	0·99539	0·99926
1·2	0·9354	0·9628	0·9782	0·9870	0·9922	0·99528	0·99713	0·99971
1·3	0·9451	0·9700	0·9832	0·9905	0·9946	0·99688	0·99819	0·99986
1·4	0·9531	0·9756	0·9870	0·9930	0·9962	0·99791	0·99885	0·99989
1·5	0·9598	0·9800	0·9899	0·9948	0·9973	0·99859	0·99926	0·99999
1·6	0·9653	0·9836	0·9920	0·9961	0·9981	0·99903	0·99951	
1·7	0·9699	0·9864	0·9937	0·9970	0·9986	0·99933	0·99968	
1·8	0·9737	0·9886	0·9950	0·9977	0·9990	0·99953	0·99978	
1·9	0·9770	0·9904	0·9959	0·9983	0·9992	0·99967	0·99985	
2·0	0·9797	0·9919	0·9967	0·9986	0·9994	0·99976	0·99990	
2·1	0·9821	0·9931	0·9973	0·9989	0·9996	0·99983	0·99993	
2·2	0·9841	0·9941	0·9978	0·9992	0·9997	0·99987	0·99995	
2·3	0·9858	0·9950	0·9982	0·9993	0·9998	0·99991	0·99996	
2·4	0·9873	0·9957	0·9985	0·9995	0·9998	0·99993	0·99997	
2·5	0·9886	0·9963	0·9987	0·9996	0·9998	0·99995	0·99998	
2·6	0·9898	0·9967	0·9989	0·9996	0·9999	0·99996	0·99999	
2·7	0·9908	0·9972	0·9991	0·9997	0·9999	0·99997	0·99999	
2·8	0·9916	0·9975	0·9992	0·9998	0·9999	0·99998	0·99999	
2·9	0·9924	0·9978	0·9993	0·9998	0·9999	0·99998	0·99999	
3·0	0·9931	0·9981	0·9994	0·9998	—	0·99999	—	

Section VIII. Explanation of Tables

The tables give the probability that the value of the mean, measured from the mean of the population, in terms of the standard deviation of the sample, will lie between $-\infty$ and z. Thus, to take the table for samples of 6, the probability of the mean of the population lying between $-\infty$ and once the standard deviation of the sample is 0·9622, or the odds are about 24 to 1 that the mean of the population lies between these limits.

The probability is therefore 0·0378 that it is greater than once the standard deviation and 0·0756 that it lies outside $\pm 1{\cdot}0$ times the standard deviation.

Section IX. Illustrations of Method

Illustration I. As an instance of the kind of use which may be made of the tables, I take the following figures from a table by A. R. Cushny and A. R. Peebles in the *Journal of Physiology* for 1904, showing the different effects of the optical isomers of hyoscyamine hydrobromide in producing sleep. The sleep of ten patients was measured without hypnotic and after treatment (1) with D. hyoscyamine hydrobromide, (2) with L. hyoscyamine hydrobromide. The average number of hours' sleep gained by the use of the drug is tabulated below.

The conclusion arrived at was that in the usual dose 2 was, but 1 was not, of value as a soporific.

Additional hours' sleep gained by the use of hyoscyamine hydrobromide

Patient	1 (Dextro-)	2 (Laevo-)	Difference (2–1)
1	+0·7	+1·9	+1·2
2	−1·6	+0·8	+2·4
3	−0·2	+1·1	+1·3
4	−1·2	+0·1	+1·3
5	−0·1	−0·1	0
6	+3·4	+4·4	+1·0
7	+3·7	+5·5	+1·8
8	+0·8	+1·6	+0·8
9	0	+4·6	+4·6
10	+2·0	+3·4	+1·4
	Mean +0·75	Mean +2·33	Mean +1·58
	S.D. 1·70	S.D. 1·90	S.D. 1·17

First let us see what is the probability that 1 will on the average give increase of sleep; i.e. what is the chance that the mean of the population of which these experiments are a sample is positive. $+0{\cdot}75/1{\cdot}70 = 0{\cdot}44$, and looking out $z = 0{\cdot}44$ in the table for ten experiments we find by interpolating between 0·8697 and 0·9161 that 0·44 corresponds to 0·8873, or the odds are 0·887 to 0·113 that the mean is positive.

That is about 8 to 1, and would correspond in the normal curve to about 1·8 times the probable error. It is then very likely that 1 gives an increase of sleep, but would occasion no surprise if the results were reversed by further experiments.

If now we consider the chance that 2 is actually a soporific we have the mean increase of sleep = 2·33/1·90 or 1·23 times the S.D. From the table the probability corresponding to this is 0·9974, i.e. the odds are nearly 400 to 1 that such is the case. This corresponds to about 4·15 times the probable error in the normal curve. But I take it the real point of the authors was that 2 is better than 1. This we must test by making a new series, subtracting 1 from 2. The mean values of this series is + 1·58, while the S.D. is 1·17, the mean value being + 1·35 times the S.D. From the table the probability is 0·9985, or the odds are about 666 to 1 that 2 is the better soporific. The low value of the S.D. is probably due to the different drugs reacting similarly on the same patient, so that there is correlation between the results.

Of course odds of this kind make it almost certain that 2 is the better soporific, and in practical life such a high probability is in most matters considered as a certainty.

Illustration II. Cases where the tables will be useful are not uncommon in agricultural work, and they would be more numerous if the advantages of being able to apply statistical reasoning were borne in mind when planning the experiments. I take the following instances from the accounts of the Woburn farming experiments published yearly by Dr Voelcker in the *Journal of the Agricultural Society.*

A short series of pot culture experiments were conducted in order to determine the causes which lead to the production of Hard (glutinous) wheat or Soft (starchy) wheat. In three successive years a bulk of seed corn of one variety was picked over by hand and two samples were selected, one consisting of "hard" grains and the other of "soft". Some of each of these were planted in both heavy and light soil and the resulting crops were weighed and examined for hard and soft corn.

The conclusion drawn was that the effect of selecting the seed was negligible compared with the influence of the soil.

This conclusion was thoroughly justified, the heavy soil producing in each case nearly 100 % of hard corn, but still the effect of selecting the seed could just be traced in each year.

But a curious point, to which Dr Voelcker draws attention in the second year's report, is that the soft seeds produced the higher yield of both corn and straw. In view of the well-known fact that the *varieties* which have a high yield tend to produce soft corn, it is interesting to see how much evidence the experiments afford as to the correlation between softness and fertility in the same *variety*.

Further, Mr Hooker* has shown that the yield of wheat in one year is largely

* *Journal of the Royal Statistical Society*, 1907.

determined by the weather during the preceding harvest. Dr Voelcker's results may afford a clue as to the way in which the seed is affected, and would almost justify the selection of particular soils for growing seed wheat.*

The figures are as follows, the yields being expressed in grammes per pot:

Year	1899		1900		1901		Average	Standard deviation	z
Soil	Light	Heavy	Light	Heavy	Light	Heavy			
Yield of corn from soft seed	7·85	8·89	14·81	13·55	7·48	15·39	11·328		
" " hard "	7·27	8·32	13·81	13·36	7·97	13·13	10·643		
Difference	+0·58	+0·57	+1·00	+0·19	−0·49	+2·26	+0·685	0·778	0·88
Yield of straw from soft seed	12·81	12·87	22·22	20·21	13·97	22·57	17·442		
" " hard "	10·71	12·48	21·64	20·26	11·71	18·96	15·927		
Difference	+2·10	+0·39	+0·78	−0·05	+2·66	+3·61	+1·515	1·261	1·20

If we wish to find the odds that soft seed will give a better yield of corn on the average, we divide the average difference by the standard deviation, giving us

$$z = 0{\cdot}88.$$

Looking this up in the table for $n = 6$ we find $p = 0{\cdot}9465$ or the odds are 0·9465 to 0·0535 about 18 to 1.

Similarly for straw $z = 1{\cdot}20$, $p = 0{\cdot}9782$, and the odds about 45 to 1.

In order to see whether such odds are sufficient for a practical man to draw a definite conclusion, I take another set of experiments in which Dr Voelcker compares the effects of different artificial manures used with potatoes on the large scale.

The figures represent the difference between the crops grown with the use of sulphate of potash and kainit respectively in both 1904 and 1905:

	cwt.	qr.	lb.		ton	cwt.	qr.	lb.
1904	+ 10	3	20	: +	1	10	1	26
1905	+ 6	0	3	: +		13	2	8

(two experiments in each year).

The average gain by the use of sulphate of potash was 15·25 cwt. and the S.D. 9 cwt., whence, if we want the odds that the conclusion given below is right, $z = 1{\cdot}7$, corresponding, when $n = 4$, to $p = 0{\cdot}9698$ or odds of 32 to 1; this is midway between the odds in the former example. Dr Voelcker says: "It may now fairly be concluded that for the potato crop on light land 1 cwt. per acre of sulphate of potash is a better dressing than kainit."

As an example of how the tables should be used with caution, I take the following pot culture experiments to test whether it made any difference whether large or small seeds were sown.

Illustration III. In 1899 and in 1903 "head corn" and "tail corn" were taken

* And perhaps a few experiments to see whether there is a correlation between yield and "mellowness" in barley.

from the same bulks of barley and sown in pots. The yields in grammes were as follows:

	1899	1903
Large seed ...	13·9	7·3
Small seed ...	14·4	8·7
	+0·5	+1·4

The average gain is thus 0·95 and the S.D. 0·45, giving $z = 2{\cdot}1$. Now the table for $n = 2$ is not given, but if we look up the angle whose tangent is 2·1 in Chambers's tables,

$$p = \frac{\tan^{-1} 2{\cdot}1}{180^\circ} + 0{\cdot}5 = \frac{64^\circ\,39'}{180^\circ} + 0{\cdot}5 = 0{\cdot}859,$$

so that the odds are about 6 to 1 that small corn gives a better yield than large. These odds* are those which would be laid, and laid rightly, by a man whose only knowledge of the matter was contained in the two experiments. Anyone conversant with pot culture would however know that the difference between the two results would generally be greater and would correspondingly moderate the certainty of his conclusion. In point of fact a large-scale experiment confirmed the result, the small corn yielding about 15 % more than the large.

I will conclude with an example which comes beyond the range of the tables, there being eleven experiments.

To test whether it is of advantage to kiln-dry barley seed before sowing, seven varieties of barley were sown (both kiln-dried and not kiln-dried) in 1899 and four in 1900; the results are given in the table.

It will be noticed that the kiln-dried seed gave on an average the larger yield

	Lb. head corn per acre			Price of head corn in shillings per quarter			Cwt. straw per acre			Value of crop per acre in shillings†		
	N. K. D.	K. D.	Diff.	N. K. D.	K. D.	Diff.	N. K. D.	K. D.	Diff.	N. K. D.	K. D.	Diff.
1899	1903	2009	+106	26½	26½	0	19½	25	+5½	140½	152	+11½
	1935	1915	− 20	28	26½	−1½	22½	24	+1½	152½	145	−7½
	1910	2011	+101	29½	28½	−1	23	24	+1	158½	161	+2½
	2496	2463	− 33	30	29	−1	23	28	+5	204½	199½	−5
	2108	2180	+ 72	27½	27	−½	22½	22½	0	162	164	+2
	1961	1925	− 36	26	26	0	19½	19½	−½	142	139½	−2½
	2060	2122	+ 62	29	26	−3	24½	22½	−2½	168	155	−13
1900	1444	1482	+ 38	29½	28½	−1	15½	16	+½	118	117½	−½
	1612	1542	− 70	28½	28	−½	18	17½	−½	128½	121	−7½
	1316	1443	+127	30	29	−1	14½	15½	+1½	109½	116½	+7
	1511	1535	+ 24	28½	28	−½	17	17½	+½	120	120½	+½
Average	1841·5	1875·2	+33·7	28·45	27·55	−0·91	19·95	21·05	+1·10	145·82	144·68	+1·14
Standard deviation	—	—	63·1	—	—	0·79	—	—	2·25	—	—	6·67
Standard deviation $\div\sqrt{8}$	—	—	22·3	—	—	0·28	—	—	0·80	—	—	2·40

† Straw being valued at 15*s.* per ton.

* [Through a numerical slip, now corrected, Student had given the odds as 33 to 1 and it is to this figure that the remarks in this paragraph relate. ED.]

of corn and straw, but that the quality was almost always inferior. At first sight this might be supposed to be due to superior germinating power in the kiln-dried seed, but my farming friends tell me that the effect of this would be that the kiln-dried seed would produce the better quality barley. Dr Voelcker draws the conclusion: "In such seasons as 1899 and 1900 there is no particular advantage in kiln-drying before sowing." Our examination completely justifies this and adds "and the quality of the resulting barley is inferior though the yield may be greater".

In this case I propose to use the approximation given by the normal curve with standard deviation $s/\sqrt{(n-3)}$ and therefore use Sheppard's tables, looking up the difference divided by $s/\sqrt{8}$. The probability in the case of yield of corn per acre is given by looking up $33{\cdot}7/22{\cdot}3 = 1{\cdot}51$ in Sheppard's tables. This gives $p = 0{\cdot}934$, or the odds are about 14 to 1 that kiln-dried corn gives the higher yield.

Similarly $0{\cdot}91/0{\cdot}28 = 3{\cdot}25$, corresponding to $p = 0{\cdot}9994$,* so that the odds are very great that kiln-dried seed gives barley of a worse quality than seed which has not been kiln-dried.

Similarly, it is about 11 to 1 that kiln-dried seed gives more straw and about 2 to 1 that the total value of the crop is less with kiln-dried seed.

Section X. Conclusions

1. A curve has been found representing the frequency distribution of standard deviations of samples drawn from a normal population.

2. A curve has been found representing the frequency distribution of values of the means of such samples, when these values are measured from the mean of the population in terms of the standard deviation of the sample.

3. It has been shown that this curve represents the facts fairly well even when the distribution of the population is not strictly normal.

4. Tables are given by which it can be judged whether a series of experiments, however short, have given a result which conforms to any required standard of accuracy or whether it is necessary to continue the investigation.

Finally I should like to express my thanks to Prof. Karl Pearson, without whose constant advice and criticism this paper could not have been written.

* As pointed out in Section V, the normal curve gives too large a value for p when the probability is large. I find the true value in this case to be $p = 0{\cdot}9976$. It matters little, however, to a conclusion of this kind whether the odds in its favour are 1660 to 1 or merely 416 to 1.

3

PROBABLE ERROR OF A CORRELATION COEFFICIENT

[*Biometrika*, VI (1908), p. 302]

At the discussion of Mr R. H. Hooker's recent paper "The correlation of the weather and crops" (*Journal of the Royal Statistical Society*, 1907) Dr Shaw made an inquiry as to the significance of correlation coefficients derived from small numbers of cases.

His question was answered by Messrs Yule and Hooker and Prof. Edgeworth, all of whom considered that Mr Hooker was probably safe in taking 0·50 as his limit of significance for a sample of 21. They did not, however, answer Dr Shaw's question in any more general way. Now Mr Hooker is not the only statistician who is forced to work with very small samples, and until Dr Shaw's question has been properly answered the results of such investigations lack the criterion which would enable us to make full use of them. The present paper, which is an account of some sampling experiments, has two objects: (1) to throw some light by empirical methods on the problem itself, (2) to endeavour to interest mathematicians who have both time and ability to solve it.

Before proceeding further, it may be as well to state the problem which occurs in practice, for it is often confused with other allied questions.

A random sample has been obtained from an indefinitely large* population and r† calculated between two variable characters of the individual composing the sample. We require the probability that R for the population from which the sample is drawn shall lie between any given limits.

It is clear that in order to solve this problem we must know two things: (1) the distribution of values of r derived from samples of a population which has a given R, and (2) the *a priori* probability that R for the population lies between any given limits. Now (2) can hardly ever be known, so that some arbitrary assumption must in general be made; when we know (1) it will be time enough to discuss

* Note that the indefinitely large population need not actually exist. In Mr Hooker's case his sample was 21 years of farming under modern conditions in England, and included all the years about which information was obtainable. Probably it could not actually have been made much larger without loss of homogeneity, due to the mixing with farming under conditions not modern; but one can imagine the population indefinitely increased and the 21 years to be a sample from this.

† Throughout the rest of this paper r is written for the correlation coefficient of a sample and R for the correlation coefficient of a population.

what will be the best assumption to make, but meanwhile I may suggest two more or less obvious distributions. The first is that any value is equally likely between $+1$ and -1, and the second that the probability that x is the value is proportional to $1-x^2$: this I think is more in accordance with ordinary experience: the distribution of *a priori* probability would then be expressed by the equation $y = \frac{3}{4}(1-x^2)$.

But whatever assumption be made, it will be necessary to know (1), so that the solution really turns on the distribution of r for samples drawn from the same population. Now this has been determined for *large* samples with as much accuracy as is required, for Pearson and Filon (*Phil. Trans.* A, CXCI, p. 229 *et seq.*) showed that the standard deviation is $(1-r^2)/\sqrt{n}$ and of course for large samples the distribution is sure to be practically normal unless r is very close to unity. But their method involves approximations which are not legitimate when the sample is small. Besides this the distribution is not then normal, so that even if we had the standard deviation a great deal would still remain unknown.

In order to throw some light on this question I took a correlation table* containing 3000 cases of stature and length of left middle finger of criminals, and proceeded to draw samples of four from this population.† This gave me 750 values of r for a population whose real correlation was 0·66. By taking the statures of one sample with the middle finger lengths of the next sample I was enabled to get 750 values of r for a population whose real correlation was zero. Next I combined each of the samples of four with the tenth sample before it and with the tenth sample after it, thus obtaining two sets of 750‡ values from samples of 8, with real correlation 0·66 and zero.

Besides this empirical work it is possible to calculate *a priori* the distribution for samples of two as follows.

For clearly the only values possible are $+1$ and -1, since two points must always lie on the regression line which joins them.§

Next consider the correlation between the difference between the values of one character in two successive individuals, and the difference between the values of the other character in the same individuals. It is well known to be the same as that between the values themselves, if the individuals be in random order.

Also, if an indefinitely large number of such differences be taken, it is clear that the means of the distributions will have the value zero. Hence, if the correlation be determined from a fourfold division through zero we can apply Mr Sheppard's‖

* *Biometrika*, I, p. 219: W. R. Macdonnell. † *Biometrika*, VI, p. 13. [2, p. 23.]

‡ Not strictly independent, but practically sufficiently nearly so. This method was adopted in order to save arithmetic.

§ There are of course indeterminate cases when the values are the same for one character, but they become rarer as we decrease the unit of grouping until, with an infinitesimal unit of grouping, the statement in the text is true.

‖ *Phil. Trans.* A, CXCII, p. 141.

result that if A and B be the numbers in the large and the small divisions of the table respectively $\cos\frac{\pi B}{A+B} = R$, where R is the correlation of the original system.

But if a pair of individuals whose difference falls in either of the small divisions be considered to be a random sample of 2, their r will be found to be -1, while that of a pair whose difference falls in one of the large divisions is $+1$. Hence the distribution of r for samples of 2 is AN at $+1$, and BN at -1, where $A+B=1$, and $B = \frac{\cos^{-1}R}{\pi}$.

When $R=0$, there is of course even division, half the values being $+1$, and half -1; when $R=0{\cdot}66$, $B = \frac{\cos^{-1}0{\cdot}66}{\pi} = 0{\cdot}271$, therefore $A = 0{\cdot}729$, and the mean is at $0{\cdot}729 - 0{\cdot}271 = 0{\cdot}458$. The S.D. $= \sqrt{\{1-(0{\cdot}458)^2\}} = 0{\cdot}889$. It is noteworthy that the mean value is considerably less than R.

I have dealt with the cases of samples of 2 at some length, because it is possible that this limiting value of the distribution with its mean of $(2/\pi)\sin^{-1}R$ and its second moment coefficient of $1-\{(2/\pi)\sin^{-1}R\}^2$ may furnish a clue to the distribution when n is greater than 2.

Besides these series, I have another shorter one of 100 values of r from samples of 30, when the real value is 0·66. The distributions of the various trials are given in the table below.

Several peculiarities will be noticed which are due to the effects of grouping, particularly in the samples of 4. Firstly, there is a lump at zero; with such small numbers zero is not an uncommon value of the product moment and then, whatever the values of the standard deviations, $r=0$.

Next there are five indeterminate cases in each of the distributions for samples of 4. These are due to the whole sample falling in the same group for one variable. In such a case, both the standard deviation and the product moment vanish and r is indeterminate.

Lastly, with such small samples one cannot use Sheppard's corrections for the standard deviations, as r often becomes greater than unity. So I did not use the corrections except in the case of the samples of 30, yet on the whole the values of the standard deviations are no doubt too large. This does not much affect the values of r in the neighbourhood of zero, but there is a tendency for larger values to come too low, so that there is a deficiency of cases towards 1 and -1. This introduces an error into the standard deviation of all the series to some extent, but of course the mean is unaltered when there is no correlation. The series for samples of 4 are affected more than those from samples of 8, as the mean standard deviation of samples of 4 is the smaller, so that the unit of grouping is comparatively larger.

Distribution of Correlation Coefficients

Scale	−1·00 — −·98	−·97 — −·93	−·92 — −·88	−·87 — −·83	−·82 — −·78	−·77 — −·73	−·72 — −·68	−·67 — −·63	−·62 — −·58	−·57 — −·53	−·52 — −·48	−·47 — −·43	−·42 — −·38	−·37 — −·33	−·32 — −·28	−·27 — −·23	−·22 — −·18	−·17 — −·13	−·12 — −·08	−·07 — −·03	−·02 — +·02	+·03 — +·07	+·08 — +·12	+·13 — +·17	+·18 — +·22	+·23 — +·27	+·28 — +·32	+·33 — +·37	+·38 — +·42	+·43 — +·47	+·48 — +·52	+·53 — +·57	+·58 — +·62	+·63 — +·67	+·68 — +·72	+·73 — +·77	+·78 — +·82	+·83 — +·87	+·88 — +·92	+·93 — +·97	+·98 — +1·00
No correlation, samples of 4*	8	17	16	23	13	23	9½	19½	19	17	23	18½	25½	15	22	22	19	19½	23½	15	30	18	14	22	22	22½	24	13½	17	23	24	24½	9½	17½	17½	11½	12½	9	22½	13½	9
No correlation, samples of 8	1	—	—	1	7	10	10	15	14	15	18	24	18	27	36	33	43	45	26½	28½	34	42½	27½	34	23½	36½	33½	28½	19	24	22	15	13½	9	2½	7	5	3	—	—	—
Correlation of 0·66, samples of 4†	3	3½	2½	2	2	8	3½	1½	5	9	8	7	3	1	4	10	5	5	3	6	16	4	7	9	11	8	14	11	20½	18	30	23½	27	33½	45½	39½	41	80½	64	91	59
Correlation of 0·66, samples of 8	—	—	—	—	—	—	—	—	2	—	2	—	3	—	3	6	4	3	4	5	7	7	5	11	17	20	17	22	34	46½	44½	61	59	66½	56½	80½	60	66	33½	4	

Scale	·26 to ·28	·29 to ·31	·32 to ·34	·35 to ·37	·38 to ·40	·41 to ·43	·44 to ·46	·47 to ·49	·50 to ·52	·53 to ·55	·56 to ·58	·59 to ·61	·62 to ·64	·65 to ·67	·68 to ·70	·71 to ·73	·74 to ·76	·77 to ·79	·80 to ·82	·83 to ·85
Distribution with samples of 30	1	—	—	1	—	1	—	1	6	4	5	9	12	10	15	12	7	10	5	1

* There are five indeterminate cases, so that the total is 745, while there are 750 in the other two distributions.

† The moment coefficients of this distribution were actually calculated from a different grouping as below:

−1·00 — −·95	−·94 — −·90	−·89 — −·85	−·84 — −·80	−·79 — −·75	−·74 — −·70	−·69 — −·65	−·64 — −·60	−·59 — −·55	−·54 — −·50	−·49 — −·45	−·44 — −·40	−·39 — −·35	−·34 — −·30	−·29 — −·25	−·24 — −·20	−·19 — −·15	−·14 — −·10	−·09 — −·05	−·04 — −·00	−·01 — +·05	+·06 — +·10	+·11 — +·15	+·16 — +·20	+·21 — +·25	+·26 — +·30	+·31 — +·35	+·36 — +·40	+·41 — +·45	+·46 — +·50	+·51 — +·55	+·56 — +·60	+·61 — +·65	+·66 — +·70	+·71 — +·75	+·76 — +·80	+·81 — +·85	+·86 — +·90	+·91 — +·95	+·96 — +1·00
5	4	2	2	6	4	3	4	7	9	7	5	1	2	6	7	7	5	4	18	2	5	11	8	12	12½	7½	17	21	24	20	35	31	35	50	34	58½	78	77	98½

The moment coefficients of the five distributions were determined, and the following values found:*

	Mean	S.D.	μ_2	μ_3	μ_4	β_1	β_2
Samples of 4 ($r=0$)	—	0·5512	0·3038	—	0·1768	—	1·918
Samples of 8 ($r=0$)	—	0·3731	0·1392	—	0·0454	—	2·336
Samples of 4 ($r=0·66$)	0·5609	0·4680	0·2190	−0·1570	0·2152	2·245	4·489
Samples of 8 ($r=0·66$)	0·6139	0·2684	0·07202	−0·02634	0·02714	1·857	5·232
Samples of 30 ($r=0·66$)	0·661	0·1001	0·01003	−0·000882	0·000461	0·7713	4·580

Considering first the "no correlation" distributions, I attempted to fit a Pearson curve to the first of them. As might be expected, the range proved limited and as symmetry had been assumed in calculating the moments, a Type II curve resulted. The equation was $y = y_0\left(1-\frac{x^2}{1·076}\right)^{0·272}$, the range of which is 2·074.

Now the real range is clearly 2, and only a very small alteration in β_2 is required to make the value of the index zero. Consequently the equation $y = y_0(1-x^2)^0$ was suggested. This means an even distribution of r between 1 and -1, with S.D. $= 0·5774 \pm 0·010$ vice 0·5512 actual, $\mu_2 = 0·3333 \pm 0·0116$ vice 0·3038, $\mu_4 = 0·2000 \pm 0·016$ vice 0·1768 and $\beta_2 = 1·800 \pm 0·12$ vice 1·918, all values as close as could perhaps be expected considering that the grouping must make both μ_2 and μ_4 too low.

Working from $y = y_0(1-x^2)^0$ for samples of 4, I guessed the formula $y = y_0(1-x^2)^{\frac{n-4}{2}}$ and proceeded to calculate the moments.

By using the transformation $x = \sin\theta$, we get $y = y_0\cos^{n-4}\theta$,

$$dx = \cos\theta\, d\theta,$$

$$2\int_0^1 y\,dx = 2y_0\int_0^{\frac{1}{2}\pi}\cos^{n-3}\theta\, d\theta,$$

$$2\int_0^1 x^2 y\,dx = 2y_0\int_0^{\frac{1}{2}\pi}\cos^{n-3}\theta\, d\theta - 2y_0\int_0^{\frac{1}{2}\pi}\cos^{n-1}\theta\, d\theta,$$

and so on.

Whence

$$\mu_2 = \frac{1}{n-1},\quad \mu_4 = \frac{3}{(n-1)(n+1)},\quad \beta_2 = \frac{3(n-1)}{n+1} = 3-\frac{6}{n+1}.$$

Putting $n = 8$ we get the equation $y = y_0(1-x^2)^2$ and

$$\mu_2 = \tfrac{1}{7} = 0·1429 \pm 0·0050 \text{ instead of actual } 0·1392,$$
$$\mu_4 = \tfrac{1}{21} = 0·0476 \pm 0·0038 \quad ,, \quad 0·0454,$$
$$\sigma = 0·3780 \pm 0·0066 \quad ,, \quad 0·3731,$$
$$\beta_2 = 3-\tfrac{6}{9} = 2·333 \pm 0·012 \quad ,, \quad 2·336.$$

* In the cases of no correlation the moments were taken about zero, the known centroid of the distribution.

The equation calculated from the actual moments is $y = y_0\left(1 - \frac{x^2}{0{\cdot}9802}\right)^{2{\cdot}021}$, whence the calculated range is 1·98, whereas it is known to be 2.

The following tables compare the actual distributions with those calculated from the equations.

Distribution of r from samples of 4 compared with the equation

$$y = \tfrac{7}{2}\tfrac{4}{5}(1-x^2)^0$$

	−1 to −·825	−·825 to −·675	−·675 to −·525	−·525 to −·375	−·375 to −·225	−·225 to −·075	−·075 to +·075	+·075 to +·225	+·225 to +·375	+·375 to +·525	+·525 to +·675	+·675 to +·825	+·825 to +1
Actual ...	64	45½	55½	67	59	62	63	58	60	64	51½	41½	54
Calculated	65	56	56	56	56	56	56	56	56	56	56	56	65
Difference	−1	−10½	−½	+11	+3	+6	+7	+2	+4	+8	−4½	−14½	−11

From this we get $\chi^2 = 13{\cdot}30$, $P = 0{\cdot}34$. It will however be noticed that the grouping has caused all the middle compartments to contain more than the calculated, as pointed out above.

Distribution of r from samples of 8 compared with the equation

$$y = \frac{750 \times 15}{16}(1-x^2)^2$$

	−1 to −·825	−·825 to −·675	−·675 to −·525	−·525 to −·375	−·375 to −·225	−·225 to −·075	−·075 to +·075	+·075 to +·225	+·225 to +·375	+·375 to +·525	+·525 to +·675	+·675 to +·825	+·825 to +1
Actual ...	2	27	44	60	96	114½	103	85	98½	65	37½	14½	3
Calculated	4½	20½	43	67	87	100½	105	100½	87	67	43	20½	4½
Difference	−2½	+6½	+1	−7	+9	+14	−2	−15½	+11½	−2	−5½	−6	−1½

Whence $\chi^2 = 13{\cdot}94$, $P = 0{\cdot}30$.

In this case the grouping has had less influence and the largest contributions to χ^2 (in the second, sixth, eighth and twelfth compartments) are due to differences of opposite sign on opposite sides, and may therefore be supposed to be entirely due to random sampling.

My equation then fits the two series of empirical results about as well as could be expected. I will now show that it is in accordance with the two theoretical cases n "large" and $n = 2$, for $\sigma = 1/\sqrt{(n-1)}$, which approximates sufficiently closely to Pearson and Filon's $(1-r^2)/\sqrt{n}$ when $r = 0$ and n is large. Also when n is large β_2 becomes 3 and the distribution is normal.

And if $n = 2$, the equation becomes $y = y_0(1-x_2)^{-1}$,* where

$$y_0 = \frac{N}{2\int_0^1 (1-x^2)^{-1}\,dx}.$$

Put $x = \sin\theta$. Then $dx = \cos\theta\,d\theta$,

$$y_0 = \frac{N}{2}\Big/\int_0^{\frac{1}{2}\pi} \sec\theta\,d\theta = \frac{N}{2}\Big/\infty = 0,$$

i.e. there is no frequency except where $(1-x^2)^{-1}$ is infinite, all the frequency is equally divided between $x = 1$ and $x = -1$, which we know to be actually the case.

Consequently, I believe that the equation $y = y_0(1-x^2)^{\frac{n-4}{2}}$ probably represents the theoretical distribution of r when samples of n are drawn from a normally distributed population with no correlation. Even if it does not do so, I am sure that it will give a close approximation to it.

Let us consider Mr Hooker's limit of 0·50 in the light of this equation. For 21 cases the equation becomes $\left.\begin{aligned} x &= \sin\theta \\ y &= y_0\cos^{17}\theta \end{aligned}\right\}$ and the proportion of the area lying beyond $x = \pm 0{\cdot}50$ will be

$$\frac{\int_{\theta=\sin^{-1}0{\cdot}50}^{\theta=\frac{1}{2}\pi} \cos^{18}\theta\,d\theta}{\int_0^{\frac{1}{2}\pi} \cos^{18}\theta\,d\theta}.$$

I find this to be 0·02099, or we may expect to find one case in 50 occurring outside the limits $\pm 0{\cdot}50$ when there is no correlation and the sample numbers 21.

When however there is correlation, I cannot suggest an equation which will accord with the facts; but as I have spent a good deal of time over the problem I will point out some of the necessities of the case.

(1) With small samples the value certainly lies nearer to zero than the real value of R, e.g.

Samples of 2:	Mean at $\frac{2}{\pi}\sin^{-1}R$,
Samples of 4 (real value 0·66):	0·561† ± 0·011,
Samples of 8 (real value 0·66):	0·614‡ ± 0·065.

* If a Pearson curve be fitted to the distribution whose moment coefficients are $\mu_2 = 1 = \mu_4$ and $\mu_3 = 0$, we have $\beta_2 = 1$, $\beta_1 = 0$, hence the curve must be of Type II and the equation is given by

$$y = y_0\left(1-\frac{x^2}{a^2}\right)^m, \text{ where } a^2 = \frac{2\mu_2\beta_2}{3-\beta_2} = 1 \text{ and } m = \frac{5\beta_2-9}{2(3-\beta_2)} \text{ or } y = y_0(1-x^2)^{-1},$$

agreeing with the general formula.

† The value must be slightly larger than this (perhaps even by 0·03) as Sheppard's corrections were not used.

‡ Again higher, but not by more than 0·02.

But with samples of 30 (real value 0·66), mean at 0·6609 ± 0·0067 shows that the mean value approaches the real value comparatively rapidly.

(2) The standard deviation is larger than accords with the formula

$$(1-r^2)/\sqrt{(n-1)}$$

even if we give the mean value of r for samples of the size taken, e.g. for samples of 2,

$$\text{S.D.} = \sqrt{\left\{1-\left(\frac{2}{\pi}\sin^{-1}R\right)^2\right\}}.$$

For samples of 4: calculated* 0·3957 ± 0·0069; actual 0·4680,

For samples of 8: calculated 0·2355 ± 0·0041; actual 0·2684.

But samples of 30, calculated 0·1046 ± 0·0018, actual 0·1001, again show that with samples as large as 30 the ordinary formula is justified.

(3) When there was no correlation the range found by fitting a Pearson curve to the distribution was accurately 2 in the theoretical case of samples of 2, and well within the probable error for empirical distributions of samples of 4 and 8. But when we have correlation this process does not give the range closely for the empirical distribution (samples of 4 give 2·137, samples of 8, 2·699, samples of 30, infinity) and the range calculated from samples of 2, which is

$$\frac{2\sqrt{(4+3\mu_2+18\mu_2^2-9\mu_2^3)}}{3-\mu_2}$$

(where $\mu_2 = 1-\{(2/\pi)\sin^{-1}R\}^2$), is always less than 2 except in the case where μ_2 is 1, i.e. when there is no correlation.

Hence the distribution probably cannot be represented by any of Prof. Pearson's types of frequency curve unless $R = 0$.

(4) The distribution is skew with a tail towards zero.

(5) To sum up: If $y = \phi(x, R, n)$ be the equation, it must satisfy the following requirements. If $R = 1$, 1 is the only value of x which gives the value of y other than zero. If $n = 2$, ± 1 are the only values of x to do so. If $R = 0$, the equation probably reduces to $y = y_0(1-x^2)^{\frac{n-4}{2}}$.

Conclusions

It has been shown that when there is no correlation between two normally distributed variables $y = y_0(1-x^2)^{\frac{n-4}{2}}$ gives fairly closely the distribution of r found from samples of n.

Next, the general problem has been stated and three distributions of r have been given which show the sort of variation which occurs. I hope they may serve as illustrations for the successful solver of the problem.

* $(1-r^2)/\sqrt{(n-1)}$, where r is taken as the mean value for the size of the sample. If we took the real value R, the difference would be even greater.

4

THE DISTRIBUTION OF THE MEANS OF SAMPLES WHICH ARE NOT DRAWN AT RANDOM

[*Biometrika*, VII (1909), p. 210]

It is one of the advantages of the normal curve that if samples are drawn at random from any population, no matter how distributed, the distributions of the statistical constants of the samples rapidly approach the Gaussian as the samples grow large.

This being so, the result of grouping 2000 in samples of 25 given in Drs Greenwood and White's very interesting paper in *Biometrika* (VI (1909), pp. 376–401) is surprising.

For it is easy to show that if B_1, B_2 be the constants of the distribution of the means of samples of n drawn at random, corresponding to β_1, β_2 in the original frequency distribution, then*

$$B_1 = \frac{\beta_1}{n} \quad \text{and} \quad B_2 - 3 = \frac{\beta_2 - 3}{n}$$

But in this case

$$\beta_1 = 1{\cdot}7977 \quad \text{and} \quad B_1 = 0{\cdot}4756, \quad \text{while} \quad \frac{\beta_1}{n} = 0{\cdot}0719,$$

$$\beta_2 - 3 = 2{\cdot}5790 \quad \text{and} \quad B_2 - 3 = 0{\cdot}3185, \quad \text{while} \quad \frac{\beta_2 - 3}{n} = 0{\cdot}1032.$$

Now neither of these can be considered significant with a sample of 80 means, but at the same time they are both sufficiently different to suggest that the conditions which led to the theoretical result have not been fulfilled.

The first thing which occurred to me was that as Sheppard's corrections had been used for the means but not for the original distribution it might be well to try applying them to both.

This however makes but little difference, for we get

$$\beta_1 = 1{\cdot}9898 \quad \text{so that} \quad \frac{\beta_1}{25} = 0{\cdot}0796,$$

$$\beta_2 - 3 = 2{\cdot}7725 \quad \text{so that} \quad \frac{\beta_2 - 3}{25} = 0{\cdot}1109.$$

I next considered the possibility that the samples were not strictly random but that there was some slight correlation between successive observations.

* Henderson, R., *J. Inst. Actu.* XLI, pp. 429–42.

I therefore assumed that the individuals composing the sample were more like each other than to the rest of the population, that in fact there was homotyposis, and working from this hypothesis I found that the slightest correlation produces a very marked retardation in the approach to normality with increase in the size of the sample.

It will be observed that this is essentially a "small sample" problem, for with increase in the size of the sample the correlation due to likeness between successive individuals diminishes except in exceptional cases, when it becomes manifest as a well-marked heterogeneity.

My results emphasize the necessity of avoiding anything which tends to produce secular variation and as far as possible to neutralize it by repeating observations only after some time has elapsed.

Thus repetitions of analyses in a technical laboratory should never follow one another but an interval of at least a day should occur between them. Otherwise a spurious accuracy will be obtained which greatly reduces the value of the analyses.

In the present case there is not sufficient evidence to show whether correlation was really present, but as in the course of a fairly extended practice I have not yet met with observations in which this tendency was altogether absent, I incline to the belief that it was.

In any case, being ignorant of the technique, I can only suggest as possibilities slight variations from point to point on the slide, differences in light or in the observer as the day went on.

The general problem is as follows:

Let samples of n be drawn from a population with constants $\mu_2, \mu_3, \mu_4, \beta_1, \beta_2$, and let the samples be drawn in such a manner that the individuals composing each sample are correlated with correlation coefficient r, then, assuming linear regression and homoscedastic arrays, the constants of the distribution of their means (M_2, M_3, M_4, B_1, B_2) are as follows:

$$M_2 = \frac{\mu_2}{n}\{1+(n-1)r\},$$

$$M_3 = \frac{\mu_3}{n^2}\{1+(n-1)r\}\{1+(2n-1)r\},$$

$$M_4 = \frac{\{1+(n-1)r\}}{n^3(1+2r)}[\mu_4\{1+(3n-1)r+3n(n-1)r^2\}+3(n-1)(1-r)(1+nr)\mu_2^2],$$

$$B_1 = \frac{\beta_1}{n}\frac{\{1+(2n-1)r\}^2}{(1+r)^2\{1+(n-1)r\}},$$

$$B_2 = \frac{\beta_2}{n}\frac{\{1+(3n-1)r+3n(n-1)r^2\}}{(1+2r)\{1+(n-1)r\}}+\frac{3(n-1)(1-r)(1+nr)}{n(1+2r)\{1+(n-1)r\}}.$$

As the method of determining the three moment coefficients is the same in each case and it is merely a question of reduction to obtain B_1 and B_2, it will be sufficient for me to give the proof for M_4.

Let $x_1 x_2 \ldots x_n$ be the values, measured from the mean of the population, of the individuals composing the typical sample, and let there be N such samples.

Then

$$M_4 = \frac{1}{N}\Sigma\left\{\frac{x_1+x_2+\ldots+x_n}{n}\right\}^4$$

$$= \frac{1}{N}\Sigma\frac{S(x_1^4)+4S(x_1^3x_2)+6S(x_1^2x_2^2)+12S(x_1^2x_2x_3)+24S(x_1x_2x_3x_4)}{n^4} \quad \ldots\ldots(\text{i})$$

Taking each of these six terms in turn we have

$$\frac{\Sigma\{S(x_1^4)\}}{Nn^4} = \frac{n\Sigma(n_{x_1}x_1^4)}{N.n^4} = \frac{\mu_4}{n^3}. \quad \ldots\ldots(\text{ii})$$

For $S(x_1^4)$ has n terms, and when they are taken over all the N samples which compose the population there will be $n.n_{x_1}$ of x_1^4, n_{x_1} being the number of x_1's in the population and $n_{x_1x_2}$ the number of x_1's associated with x_2's, and so on.

Again, there are $n(n-1)$ terms in $S(x_1^3x_2)$,

$$\therefore \frac{\Sigma\{4S(x_1^3x_2)\}}{N.n^4} = \frac{4(n-1)\Sigma(n_{x_1x_2}.x_1^3x_2)}{N.n^3}$$

$$= \frac{4(n-1)\Sigma(n_{x_1}.x_1^3.\text{mean value of } x_2)}{N.n^3}.$$

But the mean value of x_2 associated in the sample with x_1 will be $\frac{r\sigma_{x_2}}{\sigma_{x_1}}x_1$, or since $\sigma_{x_1} = \sigma_{x_2}$ it is rx_1,

$$\therefore \frac{\Sigma\{4S(x_1^3x_2)\}}{N.n^4} = \frac{4(n-1)\Sigma(n_{x_1}.x_1^4.r)}{N.n^3}$$

$$= \frac{4(n-1)r}{n^3}\mu_4. \quad \ldots\ldots(\text{iii})$$

Next, $$\frac{\Sigma\{6S(x_1^2x_2^2)\}}{N.n^4} = \frac{3(n-1)}{n^3}.\frac{\Sigma(n_{x_1x_2}.x_1^2x_2^2)}{N}$$

$$= \frac{3(n-1)}{n^3}\frac{\Sigma(n_{x_1}.x_1^2.\text{mean value of } x_2^2)}{N}$$

[Now the mean value of x_2^2 is equal to the square of the S.D. of the x_1 array of x_2's, $\{\mu_2(1-r^2)\}$, added to the square of the mean value of x_2, $(r^2x_1^2)$]

$$= \frac{3(n-1)}{n^3}\frac{\Sigma n_{x_1}\{r^2x_1^4+x_1^2\mu_2(1-r^2)\}}{N}$$

$$= \frac{3(n-1)}{n^3}\{r^2\mu_4+(1-r^2)\mu_2^2\}. \quad \ldots\ldots(\text{iv})$$

Again,

$$\frac{\Sigma\{12S(x_1^2x_2x_3)\}}{Nn^4} = \frac{6(n-1)(n-2)}{n^3}\frac{\Sigma(n_{x_1x_2x_3}.x_1^2x_2x_3)}{N}$$

$$= \frac{6(n-1)(n-2)}{n^3}\frac{\Sigma(n_{x_1x_2}.x_1^2x_2.\text{mean value of } x_3)}{N}.$$

The mean value of x_3 for values x_1 and x_2 of the other two variables is given by the equation

$$m_{x_3} = -\frac{\sigma_{x_3}}{R_{33}}\left\{\frac{R_{31}x_1}{\sigma_{x_1}} + \frac{R_{32}x_2}{\sigma_{x_2}}\right\},$$

where the R's are the minors of the determinant

$$\begin{vmatrix} 1, & r, & r \\ r, & 1, & r \\ r, & r, & 1 \end{vmatrix} \quad \text{or} \quad m_{x_3} = (x_1+x_2).\frac{(r-r^2)}{1-r^2} = (x_1+x_2).\frac{r}{1+r}.$$

Substituting, we get

$$\frac{\Sigma\{12S(x_1^2x_2x_3)\}}{N.n^4} = \frac{6(n-1)(n-2)}{n^3}.\frac{r}{1+r}.\frac{\Sigma(n_{x_1x_2}(x_1^3x_2+x_1^2x_2^2))}{N}.$$

By (iii) and (iv),

$$= \frac{6(n-1)(n-2)}{n^3}.\frac{r}{1+r}\{r\mu_4 + r^2\mu_4 + (1-r^2)\mu_2^2\}$$

$$= \frac{6(n-1)(n-2)}{n^3}.r\{r\mu_4 + (1-r)\mu_2^2\}. \qquad \ldots\ldots(v)$$

Lastly,

$$\frac{\Sigma\{24S(x_1x_2x_3x_4)\}}{Nn^4} = \frac{(n-1)(n-2)(n-3)}{n^3}\frac{\Sigma(n_{x_1x_2x_3x_4}x_1x_2x_3x_4)}{N}$$

$$= \frac{(n-1)(n-2)(n-3)}{n^3}\frac{\Sigma(n_{x_1x_2x_3}.x_1x_2x_3.\text{mean value of } x_4)}{N}.$$

As before the mean value of x_4 comes from the multiple regression equation

$$m_{x_4} = -\frac{\sigma_{x_4}}{R_{44}}\left\{x_1\frac{R_{41}}{\sigma_{x_1}} + x_2\frac{R_{42}}{\sigma_{x_2}} + x_3\frac{R_{43}}{\sigma_{x_3}}\right\},$$

where the R's are minors of

$$\begin{vmatrix} 1, & r, & r, & r \\ r, & 1, & r, & r \\ r, & r, & 1, & r \\ r, & r, & r, & 1 \end{vmatrix}.$$

$$\therefore\ m_{x_4} = (x_1+x_2+x_3)\frac{(r(1-r)^2)}{1-3r^2+2r^3} = (x_1+x_2+x_3).\frac{r}{1+2r}.$$

Substituting, we get

$$\frac{\Sigma\{24S(x_1x_2x_3x_4)\}}{Nn^4} = \frac{(n-1)(n-2)(n-3)}{n^3}.\frac{r}{1+2r}\frac{\Sigma\{n_{x_1x_2x_3}.x_1x_2x_3(x_1+x_2+x_3)\}}{N}$$

$$= \frac{(n-1)(n-2)(n-3)}{n^3}.\frac{r}{1+2r}.\frac{3\Sigma(n_{x_1x_2x_3}.x_1^2x_2x_3)}{N}.$$

Applying (v), $$= \frac{3(n-1)(n-2)(n-3)}{n^3}.\frac{r^2}{1+2r}\{r\mu_4+(1-r)\mu_2^2\}. \quad \text{.....(vi)}$$

Substituting (ii) ... (vi) in (i), we get

$$M_4 = \frac{1}{n^3}\Big\{\mu_4+4(n-1)r\mu_4+3(n-1)\{r^2\mu_4+(1-r^2)\mu_2^2\}$$
$$+6(n-1)(n-2).r.\{r\mu_4+(1-r)\mu_2^2\}$$
$$+3(n-1)(n-2)(n-3).\frac{r^2}{1+2r}\{r\mu_4+(1-r)\mu_2^2\}\Big\},$$

which reduces to the result given above, viz.

$$M_4 = \frac{\{1+(n-1)r\}}{n^3(1+2r)}[\{1+(3n-1)r+3n(n-1)r^2\}\mu_4+3(n-1)(1-r)(1+nr)\mu_2^2],$$

Using these equations it is possible to find values of r which would satisfy the conditions for the various constants.

Thus (using Sheppard's corrections for both sets of constants) I find that with the given values of μ_2 and M_2, $r = 0{\cdot}003$,

of β_1 and B_1, $r = 0{\cdot}063$,

of β_2 and B_2, $r = 0{\cdot}033$.

Now clearly if r were fitted by least squares or in any other way from these three values it must clearly come closest to the μ_2 value owing to the lower probable error of μ_2. As a proper fitting would clearly be very complicated owing to the intercorrelations of the constants, I have assumed a value $r = 0{\cdot}01$ as a nice round number; this gives a value of M_2 higher than that found in the sample before us, but not at all impossibly so.

This gives $M_2 = 0{\cdot}1101$, actual $0{\cdot}1074$,

$B_1 = 0{\cdot}1397$, ,, $0{\cdot}4756$,

$B_2 = 3{\cdot}2012$, ,, $3{\cdot}3185$.

These constants give a Type I curve

$$y = 97{\cdot}57\left(1+\frac{x}{1{\cdot}65}\right)^{24{\cdot}64}\left(1-\frac{x}{47{\cdot}82}\right)^{714{\cdot}2}.$$

If we assume no correlation I get a curve

$$y = 109{\cdot}0\left(1+\frac{x}{1{\cdot}92}\right)^{41{\cdot}7}\left(1-\frac{x}{58{\cdot}31}\right)^{1266{\cdot}8},$$

whence I get the following "fits".*

* The figures given are really mid-ordinates, but for such small numbers the difference between the mid-ordinate and the area on the base unit is negligible.

Below 1·10	1·10 to 1·22	1·22 to 1·34	1·34 to 1·46	1·46 to 1·58	1·58 to 1·70	1·70 to 1·82	1·82 to 1·94	1·94 to 2·06	2·06 to 2·18	2·18 to 2·30	2·30 to 2·42	2·42 to 2·54	2·54 to 2·66	2·66 to 2·78	Above 2·78
Actual															
—	4	8	7	14	12	12	5	7	5	2	1	2	0	1	—
Calculated: No correlation															
1·01	2·42	5·28	8·86	11·69	13·07	12·18	9·97	6·90	4·27	2·36	1·18	—	—	0·92	—
Calculated: Correlation 0·01															
1·85	3·27	6·02	8·92	11·01	11·71	10·84	8·95	6·64	4·52	2·82	1·64	0·90	—	0·85	—

These give $P = 0{\cdot}46$ and $P = 0{\cdot}86$ respectively, the first being a good deal helped by the convention that the tail should not be carried beyond the point at which a single unit may be expected and the second much less so.

As the empirical curve fitted from the actual moments has a P of 0·92, the second curve may be considered fairly good, depending as it does on a guess following on calculation. On the other hand a P of 0·46 with so few cases as 80 is not particularly good, and as Prof. Pearson has pointed out to me the graph distinctly gives an idea of greater skewness than is represented by the no correlation curve. I do not, however, wish to contend that the circumstances attending the production of the sample actually conformed to the arbitrary conditions which I found it necessary to assume in order to simplify the analysis. But seeing that the fit is good and that with such a small sample even the divergent B_1 is not altogether impossible, I think it likely that there was some sort of correlation, though probably not that particular kind which has been assumed in this note.

Conclusions

1. That the approach to normality of the distribution of means of samples drawn from a non-Gaussian population is delayed by the existence of correlation between the individuals composing the samples.

2. That on certain arbitrary assumptions the constants of the new distribution can be found given the constants of the old one and r according to formulae given above.

3. That using the above formulae and choosing a likely looking value of r, a curve can be drawn to represent the sample in Drs Greenwood and White's paper with fair likelihood.

5

APPENDIX TO MERCER AND HALL'S PAPER ON "THE EXPERIMENTAL ERROR OF FIELD TRIALS"

[*J. Agric. Sci.* IV (1911), p. 128]

Note on a Method of Arranging Plots so as to Utilize a given Area of Land to the Best Advantage in Testing Two Varieties

THE authors have shown that to reduce the error as low as possible it is necessary to "scatter" the plots. I propose to deal with this point in the special case when a comparison is to be made between only two kinds of plots, let us say two varieties of the same kind of cereal.

If we consider the causes of variation in the yield of a crop it seems that broadly speaking they are divisible into two kinds.

The first are random, occurring at haphazard all over the field. Such would be attacks by birds, the incidence of weeds or the presence of lumps of manure. The second occur with more regularity, increase from point to point or having centres from which they spread outwards; we may take as instances of this kind changes of soil, moist patches over springs or the presence of rabbit holes along a hedge.

Having made this distinction between random and regular causes of variation let me hasten to add that almost all causes of variation may belong to one or other or both of these classes according to the size of the plot in question.

In any case a consideration of what has been said above will show that any "regular" cause of variation will tend to affect the yield of adjacent plots in a similar manner; if the yield of one plot is reduced by rabbits from a bury near by, the plot next it will hardly escape without injury, while one some distance away may be quite untouched and so forth. And the smaller the plots the more are causes of variation "regular"; for example, with large plots a thistly patch may easily occur wholly within a single plot leaving adjacent plots nearly or altogether clean, but with quite small plots one which is overgrown with thistles is almost sure to have neighbours also affected.

Now if we are comparing two varieties it is clearly of advantage to arrange the plots in such a way that the yields of both varieties shall be affected as far as possible by the same causes to as nearly as possible an equal extent.

To do this it is necessary, from what has been said above, to compare together plots which lie side by side and also to make the plots as small as may be practicable and convenient.

There is a reason, apart from the difficulty of cultivating very small plots, why the plots should not be made too small, and that is, that when two different

varieties are sown next one another the outside drill of each is under abnormal conditions and if it be counted in the plot may introduce an error which in a small plot may be quite substantial, but if it is not counted the space wasted by rejecting the outside drills of small plots becomes considerable.

Let us suppose that the smallest practicable size of plot has been chosen and the land available for the comparison has been divided up into plots of this size and sown, chequer fashion, with seed of the two varieties.

Obviously nothing that we can do (supposing of course careful harvesting) can now alter the accuracy of the resulting comparison of yields, but we can easily make different estimates of the reliance which we can place on the figures.

For example, the simplest way of treating the figures would be to take the yields of the plots of each variety and determine the standard deviation of each kind. Then from published tables we can judge whether such a difference as we find between the total yields is likely to have arisen by chance.

An advance on this is to compare each plot with its neighbour and to determine the standard deviation of the differences between these pairs of adjacent plots.

From what has been said above as to the occurrence of "regular" sources of error it will be seen that such differences as these will be to a much larger extent dependent on the variety, and to a less extent on errors, than if the mere aggregates are compared.

The standard deviation will therefore be smaller and the confidence which can be placed in the result increased.

By a further device we can still further decrease the standard deviation and increase our certainty.

For if, instead of harvesting the whole of each plot together, we divide each plot into two before harvesting (and that this can be done is clear from the account of the work done with the mangolds and wheat), then we get twice the number of comparisons, and the plots being half the size are comparatively closer together and the error of their comparison is reduced.

But, it will be asked, why take all this trouble? The error of comparing plots of any given size has been found by the authors of the paper, and all that has to be done is to apply this knowledge to the particular set of experiments.

The answer to this point is that there is no such thing as the absolute error of a given size of plot. We may find out the order of it, be sure perhaps that it is not likely to be less than (say) 5 % nor more than 15 % without producing visible heterogeneity, but the error of a given size of plot must vary with all the external conditions as well as with the particular crops upon which the experiment is being conducted, and it is far better to determine the error from the figures of the experiment itself; only so can proper confidence be placed in the result of the experiment.

The diagram illustrates the proposed method of arranging the plots.

The different shading represents the two different varieties.

The firm lines represent the outside of the original plots.

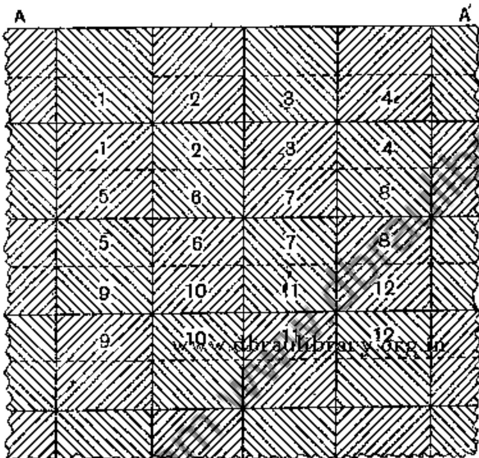
AA' is part of the boundary of the experimental ground, part of which is given in the diagram.

The dotted lines show the further division made at harvesting.

Then the yields of the half-plots 1, 1 : 2, 2 : ... etc. are compared together.

The outside half-plots are neglected as it is usual to discard the edge of the field.

I have determined the error of comparing plots of different sizes in this way both with the mangold and the wheat figures.



Considering first the mangolds:

The crop on half an acre in the present experiment was about 32,860 lb., and the standard deviation of a single one-two-hundredth acre was found to be 20·37 lb. Hence the standard deviation of half an acre made up at random from 100 such small plots would be $20{\cdot}37 \times \sqrt{100}$ or 203·7 lb., and the standard deviation of the comparison between two such half acres would be $203{\cdot}7 \times \sqrt{2}$ or 287 lb.

This would amount to 0·87 %, so that one could not begin to be sure that a difference between two varieties of mangolds compared in this way (one-two-hundredth plots arranged at random) until it amounted to say 2·6 %.

But now suppose that the plots were each originally one-hundredth acre, bisected at harvest and compared as suggested above.

Then the actual figures given by the authors enable us to determine the standard deviation of the difference between the half acre.

It amounts to no more than 223 lb. or 0·68 %. I.e. although working with plots twice the size up to harvest time we get the same accuracy with one acre of ground as would have been obtained with $(0{\cdot}87/0{\cdot}68)^2$ acres or 1·65 acres on the first plan.

Now suppose the plots to be one-fiftieth divided into one-hundredths at harvest.

Then I find the S.D. to be 274 lb. or 0·83 %.

Similarly $\frac{1}{20}$th acre plots harvested as $\frac{1}{10}$ths give a S.D. of comparison 289 lb. or 0·88 %
" $\frac{1}{10}$ " " square $\frac{1}{20}$ " " 374 " 1·14 %
" $\frac{1}{10}$ " " long $\frac{1}{20}$ " " 329 " 1·00 %

With such small numbers the difference between the last two cannot be taken as significant, but one would expect the square plot to give a worse comparison than the long plot.

We may summarize the above results in the table below:

Size of plot	Percentage S.D. of comparing $\frac{1}{2}$ acres	Total area required to give a S.D. of 1 % in the comparison
$\frac{1}{100}$th harvested as $\frac{1}{200}$ths	0·68	0·46 acres
$\frac{1}{50}$ " $\frac{1}{100}$ "	0·83	0·69 "
$\frac{1}{25}$ " $\frac{1}{50}$ "	0·88	0·77* "
$\frac{1}{10}$ " square $\frac{1}{20}$ "	1·14	1·30* "
$\frac{1}{10}$ " long $\frac{1}{20}$ "	1·00	1·00* "

The corresponding figures derived from the wheat results are set out in the second table:

Size of plot	S.D. in lb. of comparing two half acres	S.D. as a % of crop of a half acre	Total area required to give a S.D. of 1 % in the comparison
$\frac{1}{250}$th divided into $\frac{1}{500}$ths at harvest	7·02	0·71	0·50 acres
$\frac{1}{125}$ " $\frac{1}{250}$ "	8·54	0·86	0·74 "
$\frac{1}{125}$ " $\frac{1}{125}$ "	11·56	1·17	1·37 "
$\frac{1}{25}$ " $\frac{1}{50}$ "	10·40	1·05	1·10* "
$\frac{1}{50}$ " $\frac{1}{25}$ "	19·40	1·96	3·84* "
$\frac{1}{500}$ taken at random	10·28	1·04	1·08 "

* These samples are too small to give more than a rough indication of the S.D. and of the area required. I have elsewhere (*Biometrika*, VI, p. 19 [2, p. 29]) given special tables for dealing with such small numbers.

Both these tables show that in the actual fields which were measured, the area of land required to give a comparison between two varieties would increase rapidly as the size of plot increased if the same accuracy were required in the result.

Roughly speaking one-twentieth acre plots of mangolds would require at least twice as much land as one-two-hundredth acre plots in order that we may place as much confidence in the result, while one-fiftieth acre plots of wheat would probably require more than twice as much as one-five-hundredth acre plots.

Hence it is clearly of advantage to use the smallest practicable size of plot.

Also the advantage of comparing adjacent plots is apparent in these examples, since with the roots less than two-thirds of the land is required to give the same accuracy as random comparison and with the wheat less than half.

Of course the comparison of whole half-acre plots would be liable to give errors of quite a different order: thus the South half acre of mangolds is 4·7 % better than the North half acre, while the West half acre of wheat is 8·3 % better than the East half acre; such differences would be quite impossible if the half acres were subdivided into the smaller sizes of plots.

6

THE CORRECTION TO BE MADE TO THE CORRELATION RATIO FOR GROUPING*

[*Biometrika*, IX (1913), p. 316]

USING the ordinary notation, viz. n_{x_ρ} = the number in the x array of y's whose mean is at x_ρ, $\bar{y}_{x_\rho}$ = the mean of this array, N the total number in the sample, and $\bar{y}$ the general mean of y, we have η^2 defined by the relation

$$\eta^2 = \frac{S\{n_{x_\rho}(\bar{y}_{x_\rho} - \bar{y})^2\}}{N\sigma_y^2}. \qquad \text{.....(i)}$$

If η^2 is required to fit a regression curve to the actual observations as in Prof. Pearson's original memoir "On the general theory of skew correlation and non-linear regression" (*Drapers' Company Research Memoirs*, Biometric Series, II (1905)), no correction is necessary.

But if we require a ratio which shall remain constant under wide variations of grouping and of number in the sample and which shall consequently be more comparable from one sample to another, there are two corrections to be made.

The first of these has already been given by Prof. Pearson (*Biometrika*, VIII (1911), p. 256), and he has expressed it as follows: If $\overline{\eta}^2$ be the value of η^2 actually found by the use of (i), and η^2 be the value which would be found from an infinitely large sample, then if κ be the number of x arrays,

$$\eta^2 = \frac{\overline{\eta}^2 - (\kappa - 1)/N}{1 - (\kappa - 2)/N}. \qquad \text{.....(ii)}$$

But there is a further effect of grouping which has not hitherto been noted and which can be evaluated as follows:

Suppose the x_ρ array to be divided into elementary x arrays and let y_p be the mean of the x_p elementary array and n_p its frequency.

Then clearly the proper contribution of the x_ρ array to η^2 is

$$\frac{S\{n_p(y_p - \bar{y})^2\}}{N\sigma_y^2}.$$

This is equal to

$$\frac{S\{n_p(\bar{y}_{x_\rho} - \bar{y} + y_p - \bar{y}_{x_\rho})^2\}}{N\sigma_y^2} = \frac{1}{N\sigma_y^2}[S\{n_p(\bar{y}_{x_\rho} - \bar{y})^2\} + 2S\{n_p(\bar{y}_{x_\rho} - \bar{y})(y_p - \bar{y}_{x_\rho})\} + S\{n_p(y_p - \bar{y}_{x_\rho})^2\}].$$

Now $\bar{y}_{x_\rho} - \bar{y}$ is of course constant for this summation,

$$S(n_p) = n_{x_\rho} \quad \text{and} \quad S\{n_p(y_p - \bar{y}_{x_\rho})\} = 0,$$

* See "On the measurement of the influence of 'broad categories' on correlation" by Karl Pearson, *Biometrika*, IX (1913), p. 118.

therefore the contribution to η^2

$$= \frac{n_{x_\rho}(\bar{y}_{x_\rho} - \bar{y})^2}{N\sigma_y^2} + \frac{S\{n_p(y_p - \bar{y}_{x_\rho})\}}{N\sigma_y^2}. \qquad \text{.....(iii)}$$

The first of these two terms is that which is obtained in the ordinary way, so the contribution of each array should be corrected by the addition of the second term and η^2 itself by the addition of

$$S\left[\frac{S\{n_p(y_p - \bar{y}_{x_\rho})^2\}}{N\sigma_y^2}\right]. \qquad \text{.....(iv)}$$

Now if Prof. Pearson's correction (ii) has been made, we may take the point whose coordinates are (x_p, y_p) to lie on the regression line, and if further we assume the regression line to be linear throughout the x_ρ group and to be inclined at an angle of $\tan^{-1} r_\rho \frac{\sigma_y}{\sigma_x}$ to the horizontal, we have

$$y_p = x_p . r_\rho \frac{\sigma_y}{\sigma_x} \quad \text{and} \quad \bar{y}_{x_\rho} = x_\rho . r_\rho \frac{\sigma_y}{\sigma_x}.$$

Hence (iv) becomes

$$S\left[\frac{r_\rho^2 S\{n_p(x_p - \bar{x}_\rho)^2\}}{N\sigma_x^2}\right]. \qquad \text{.....(v)}$$

Now $S\{n_p(x_p - \bar{x}_\rho)^2\}$ is the second moment of the x_ρ group about its own mean and when the distribution is known can often be approximately evaluated. Similarly, when the distribution is known r_ρ can be estimated and the correction to η^2 calculated group by group.

But by making certain assumptions we can very much simplify the work, and a practical test, in which the assumptions are not justified, will show the sort of errors which are introduced.

The first assumptions are that the regression is linear and the arrays homoscedastic. In this case of course r_ρ is constant and equal to η; we are practically determining a value of r by the η method.

The correction then becomes

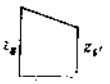
$$\frac{\eta^2}{N\sigma_x^2} S[S\{n_p(x_p - \bar{x}_\rho)^2\}],$$

or writing $N\sigma_x^2 \lambda^2 = S[S\{n_p(x_p - \bar{x}_\rho)^2\}]$ and H^2 for the raw value of η^2 after using Pearson's correction, we get from (iii) $\eta^2 = H^2 + \eta^2\lambda^2$ or

$$\eta^2 = \frac{H^2}{(1 - \lambda^2)}. \qquad \text{.....(vi)}$$

To obtain a value for λ^2 we still require to postulate something of the nature of the distribution, and I propose to treat (i) of the case where the unit of grouping is constant and small enough for the frequency in each group to be considered to be distributed as a trapezium, and (ii) of the case where the frequency distribution is normal.

(i) First to find the second moment of a trapezium about its mean.



Let z_s and $z_{s'}$ be the ordinates forming the "walls" of the trapezium and let the group unit be h.

Then $y = z_s + \left(\frac{z_{s'} - z_s}{h}\right) x$ is the equation to the "roof" referred to the "floor" and left-hand "wall" as axes. The area is clearly $\frac{(z_s + z_{s'})h}{2}$.

The mean is at

$$\frac{2}{h(z_s + z_{s'})} \int_0^h yx\,dx = \frac{2}{h(z_s + z_{s'})} \left\{ \frac{(z_{s'} - z_s)h^3}{3h} + \frac{z_s h^2}{2} \right\} = \frac{h}{3} \cdot \frac{2z_{s'} + z_s}{z_s + z_{s'}}.$$

The second moment coefficient about the axis of y is

$$\frac{2}{h(z_s + z_{s'})} \int_0^h yx^2\,dx = \frac{2}{h(z_s + z_{s'})} \left\{ \frac{(z_{s'} - z_s)h^4}{4h} + \frac{z_s h^3}{3} \right\} = \frac{h^2}{6} \cdot \frac{3z_{s'} + z_s}{z_s + z_{s'}}.$$

The second moment coefficient about the mean is

$$\frac{h^2}{6} \cdot \frac{3z_{s'} + z_s}{z_s + z_{s'}} - \frac{h^2}{9} \cdot \frac{(2z_{s'} + z_s)^2}{(z_s + z_{s'})^2} = \frac{h^2}{18} \left\{ \frac{z_s^2 + 4z_{s'}z_s + z_s^2}{(z_s + z_{s'})^2} \right\} = \frac{h^2}{12} \left\{ 1 - \frac{1}{3} \left(\frac{z_s - z_{s'}}{z_s + z_{s'}} \right)^2 \right\}.$$

Clearly when h is reasonably small $\left(\frac{z_s - z_{s'}}{z_s + z_{s'}}\right)^2$ is a quantity of the second order and in this case

$$\lambda^2 = \frac{h^2}{12\sigma_x^2}, \qquad \text{.....(vii)}$$

so that

$$\eta^2 = \frac{1}{\left(1 - \frac{h^2}{12\sigma_x^2}\right)} \left\{ \frac{\overline{\eta}^2 - \frac{\kappa - 1}{N}}{1 - \frac{\kappa - 2}{N}} \right\}, \qquad \text{.....(viii)}$$

when the unit of grouping is uniform and small.

(ii) When the unit of grouping is neither uniform nor small and there is no special knowledge of the nature of the distribution, we must needs fall back on the Gaussian curve to give us a first approximation to z_s and $z_{s'}$ for each group.

In this case

$$1 - \lambda^2 = NS \left\{ \frac{(z_s - z_{s'})^2}{n_{x_p}} \right\}, \qquad \text{.....(ix)*}$$

and it is necessary to determine it, after fitting the frequency by means of Sheppard's tables.

* The suggestion of this formula I owe to Prof. Pearson.

Finally, what correction, if any, is to be made for the grouping of y?

This will become more apparent from the alternative formula for η^2, namely

$$\eta^2 = 1 - \frac{S(y-\bar{y}_s)^2}{N\sigma_y^2}.$$

For the second moment of each array should be corrected by the subtraction of $n_s k^2/12$, where k is the unit of grouping of y, so that

$$\begin{aligned}
\eta^2 &= 1 - \frac{S(y-\bar{y}_s)^2 - Nk^2/12}{S(y-\bar{y})^2 - Nk^2/12} \\
&= \frac{S(y-\bar{y})^2 - S(y-\bar{y}_s)^2}{S(y-\bar{y})^2 - Nk^2/12} \\
&= \frac{S(y-\bar{y}_s+\bar{y}_s-\bar{y})^2 - S(y-\bar{y}_s)^2}{N\sigma_y^2} \\
&= \frac{S(y-\bar{y}_s)^2 + 2S(y-\bar{y}_s)(\bar{y}_s-\bar{y}) + S(\bar{y}_s-\bar{y})^2 - S(y-\bar{y}_s)^2}{N\sigma_y^2} \\
&= \frac{S\{n_s(y_s-\bar{y})^2\}}{N\sigma_y^2},
\end{aligned}$$

since $S(\bar{y}_s-\bar{y})^2$ when summed for each individual becomes $S\{n_s(\bar{y}_s-\bar{y})^2\}$ when summed for each array, and $S(y-\bar{y}_s)(\bar{y}_s-\bar{y})$ vanishes for each array.

Hence there is no correction to be made for the y grouping except Sheppard's correction for the standard deviation of y.

I have tested the results on an instance given in Prof. Pearson's original *Drapers' Company Research Memoir*, namely the age and auricular height in girls, correlation table pp. 34 and 54. The means of the arrays in the full table are as follows:

Even grouping Number of grouping			Age	Mean auricular height	Number of cases	Uneven grouping number			
III	II	I				IV	V	VI	VII
		—	*3– 4*	115·25	1				
		—	*4– 5*	116·9643	7				
		—	*5– 6*	117·4722	18				
		—	*6– 7*	119·1000	40				
		—	*7– 8*	120·3026	76				
		—	*8– 9*	121·6340	125				
		—	*9–10*	121·7246	177				
		—	*10–11*	122·8160	235	—			
		—	*11–12*	123·1427	261	—			
		—	*12–13*	123·8908	309	—	—		
		—	*13–14*	124·8622	263	—	—		
		—	*14–15*	125·7146	198	—			
		—	*15–16*	126·1565	214	—			
		—	*16–17*	126·5340	162				
		—	*17–18*	126·9132	95				
		—	*18–19*	127·0205	61				
		—	*19–20*	129·5577	13				
		—	*20–21*	123·8214	7				
		—	*21–22*	126·5000	8				
		—	*22–23*	125·25	2				

These were grouped in seven ways, in three of which the groups were of equal width, and the other four give an attempt at equal frequency: the method of grouping is set out by means of columns headed in Roman numerals. The age distribution differs significantly from the normal, the constants being $\beta_1 = 0{\cdot}0013$, $\beta_2 = 2{\cdot}7101$, but it would perhaps have been better to have selected a less normal distribution: still it represents the ordinary "cocked hat" statistics that tend to occur.

The regression is certainly not very linear, the growth apparently ceasing at about 18–19.

The values of $\overline{\eta}^2$ (the raw value), H^2 (the value after using Prof. Pearson's correction) and η^2 (the value after attempting to use the λ^2 correction) are given in the following table:

Number of grouping	Number of groups	$\overline{\eta}^2$	$\overline{\eta}$	H^2	$\left(\lambda^2 = \frac{h^2}{12\sigma_x^2}\right)$		$1-\lambda^2$ from normal curve	
					η^2	η	η^2	η
I	20	0·09183	0·303	0·08414	0·08489	0·291	0·08494	0·291
II	10	0·08657	0·294	0·08290	0·08595	0·293	0·08510	0·292
III	5	0·07701	0·278	0·07535	0·08786	0·296	0·08635	0·294
IV	9	0·08836	0·297	0·08510	—	—	0·08953	0·299
V	6	0·08342	0·289	0·08136	—	—	0·08913	0·299
VI	5	0·08218	0·287	0·08052	—	—	0·08885	0·298
VII	2	0·06203	0·249	0·06159	—	—	0·09739	0·312

It will be seen that the first three, with even grouping, are very close together, though the number of groups has been reduced from 20 to 5. Similarly, the next three are close together, and the last is again by itself.

An examination of the way in which the groups are taken shows that the more the tail is bunched together the higher is the value found, and this is what would be expected in this particular case, since there is practically no increase of head height with age at the "old" end of the scale, whereas for purpose of calculation we have assumed a constant angle for the regression line. But it may be pointed out that η varies (to the second place of decimals) only from 0·29 to 0·31 even if we reduce the twenty groups to two, an extreme proceeding which is never done in practice.

At the same time the ordinary six or eight groups may be expected to give results a little too high when, as is usual, the regression line is curved.

7

THE ELIMINATION OF SPURIOUS CORRELATION DUE TO POSITION IN TIME OR SPACE

[*Biometrika*, X (1914), p. 179]

In the *Journal of the Royal Statistical Society* for 1905,* p. 696, appeared a paper by R. H. Hooker giving a method of determining the correlation of variations from the "instantaneous mean" by correlating corresponding differences between successive values. This method was invented to deal with the many statistics which give the successive annual values of vital or commercial variables; these values are generally subject to large secular variations, sometimes periodic, sometimes uniform, sometimes accelerated, which would lead to altogether misleading values were the correlation to be taken between the figures as they stand.

Since Mr Hooker published his paper, the method has been in constant use among those who have to deal statistically with economic or social problems, and helps to show whether, for example, there really is a close connexion between the female cancer death rate and the quantity of imported apples consumed per head!

Prof. Pearson, however, has pointed out to me that the method is only valid when the connexion between the variables and time is linear, and the following note is an effort to extend Mr Hooker's method so as to make it applicable in a rather more general way.

If x_1, x_2, x_3, etc., y_1, y_2, y_3, etc. be corresponding values of the variables x and y, then if x_1, x_2, x_3, etc., y_1, y_2, y_3, etc. are randomly distributed in time and space, it is easy to show that the correlation between the corresponding nth differences is the same as that between x and y.

Let ${}_nD_x$ be the nth difference.

For $$ {}_1D_x = x_1 - x_2, \quad \therefore\ {}_1D_x^2 = x_1^2 - 2x_1x_2 + x_2^2. $$

Summing for all values and dividing by N and remembering that since x_1 and x_2 are mutually random $S(x_1x_2) = 0$, we get†

$$ \sigma^2_{{}_1D_x} = 2\sigma_x^2. $$

* The method had been used by Miss Cave in *Proc. Roy. Soc.* LXXIV, pp. 407 *et seq.*, that is in 1904, but being used incidentally in the course of a paper it attracted less attention than Hooker's paper which was devoted to describing the method. The papers were no doubt quite independent.

† The assumption made is that n is sufficiently large to justify the relations

$S_1^{n-1}(x)/(n-1) = S_2^n(x)/(n-1) = S_1^n(x)/n$ and $S_1^{n-1}(x^2)/(n-1) = S_2^n(x^2)/(n-1) = S_1^n(x^2)/n$ being taken to hold.

Again, $\quad {}_1D_y = y_1 - y_2, \quad \therefore\ {}_1D_{x_1}{}_1D_y = x_1y_1 - x_2y_1 - x_1y_2 + x_2y_2.$

Summing for all values and dividing by N, and remembering that x_1 and y_2 and x_2 and y_1 are mutually random,

$$r_{{}_1D_{x_1}{}_1D_y}\sigma_{{}_1D_x}.\sigma_{{}_1D_y} = 2r_{xy}\sigma_x\sigma_y,$$

$$\therefore\ r_{{}_1D_{x_1}{}_1D_y} = r_{xy}.$$

Proceeding successively,

$$r_{{}_nD_{x_n}{}_nD_y} = r_{{}_{n-1}D_{x_{n-1}}{}_{n-1}D_y} = \ldots = r_{xy}. \qquad \ldots\ldots(1)$$

Now suppose x_1, x_2, x_3, etc. are not random in space or time; the problems arising from correlation due to successive positions in space are exactly similar to those due to successive occurrence in time, but as they are to some extent complicated by the second dimension, it is perhaps simpler to consider correlation due to time.

Suppose then

$$x_1 = X_1 + bt_1 + ct_1^2 + dt_1^3 + \text{etc.}, \quad x_2 = X_2 + bt_2 + ct_2^2 + dt_2^3 + \text{etc.},$$

where X_1, X_2, etc. are independent of time and t_1, t_2, t_3 are successive values of time, so that $t_n - t_{n-1} = T$, and suppose $y_1 = Y_1 + b't_1 + c't_1^2 + \text{etc.}$ as before.

Then $\quad {}_1D_x = {}_1D_X - bT - cT(t_1 + t_2) - dT(t_1^2 + t_1t_2 + t_2^2) - \text{etc.}$

$${}_1D_x = {}_1D_X - \{bT + cT^2 + dT^3 + \text{etc.}\} - t_1\{2cT + 3dT^2 + 4eT^3 + \text{etc.}\}$$

$$- t_1^2\{3dT + 6eT^2 + \text{etc.}\} - \text{etc.}$$

In this series the coefficients of t_1, t_2, etc. are all constants and the highest power of t_1 is one lower than before, so that by repeating the process again and again we can eliminate t from the variable on the right-hand side, provided of course that the series ends at some power of t.

When this has been done, we get

$${}_nD_x = {}_nD_X + \text{a constant},$$

$${}_nD_y = {}_nD_Y + \text{a constant},$$

so $\quad r_{{}_nD_{x_n}{}_nD_y} = r_{{}_nD_X{}_nD_Y} = r_{XY},$

and of course $r_{{}_{n+1}D_{x_{n+1}}{}_{n+1}D_y} = r_{{}_nD_{x_n}{}_nD_y}$, for ${}_nD_x$ and ${}_nD_y$ are now random variables independent of time.

Hence if we wish to eliminate variability due to position in time or space and to determine whether there is any correlation between the residual variations, all that has to be done is to correlate the 1st, 2nd, 3rd, ... nth differences between successive values of our variable with the 1st, 2nd, 3rd, ... nth differences between successive values of the other variable. When the correlation between the two nth differences is equal to that between the two $(n+1)$th differences, this value gives the correlation required.

This process is tedious in the extreme, but that it may sometimes be necessary is illustrated by the following examples: the figures from which the first two are taken were very kindly supplied to me by Mr E. G. Peake, who had been using

them in preparing his paper "The application of the statistical method to the bankers' problem" in *The Bankers' Magazine* (July–August 1912). The material for the next is taken from a paper in *The Journal of Agricultural Science* (IV, 1911) by Mercer and Hall, on the error of field trials, and are the yields of wheat and straw on five hundred $\frac{1}{500}$ acre plots into which an acre of wheat was divided at harvest. The remainder are from the three of the Registrar-General's Returns.

	I	II	III	IV	V	VI
Correlation between ...	Sauerbeck's index numbers	Marriage rate	Yield of grain	Tuberculosis death rate		
and	Bankers' clearing house returns per head	Wages	Yield of straw	Infantile mortality		
				Ireland	England	Scotland
Raw figures	−0·33	−0·52	+0·753	+0·63	+0·35	+0·02
First difference	+0·51	+0·67	+0·590	+0·75	+0·69	+0·51
Second difference	+0·30	+0·58	+0·539	+0·74	+0·74	+0·65
Third difference	+0·07	+0·52	+0·530	—	—	—
Fourth difference	+0·11	+0·55	+0·524	—	—	—
Fifth difference	+0·05	+0·58	—	—	—	—
Sixth difference	—	+0·55	—	—	—	—
Number of cases	41 years	57 years	500 plots	42 years		

The difference between I and II is very marked, and would seem to indicate that the causal connexion between index numbers and Bankers' clearing house rates is not altogether of the same kind as that between marriage rate and wages, though all four variables are commonly taken as indications of the short period trade wave. I had hoped to investigate this subject more thoroughly before publishing this note, but lack of time has made this impossible.

8

TABLES FOR ESTIMATING THE PROBABILITY THAT THE MEAN OF A UNIQUE SAMPLE OF OBSERVATIONS LIES BETWEEN $-\infty$ AND ANY GIVEN DISTANCE OF THE MEAN OF THE POPULATION FROM WHICH THE SAMPLE IS DRAWN

[*Biometrika*, XI (1917), p. 414]

In the last number of *Biometrika* (XI (1916), p. 277) Mr Young completes the table given in vol. X, p. 522 of the standard deviation frequency curves for small samples by working out the cases where the numbers in the sample are as low as two and three.

In the course of his note he writes: "The smallest sample considered is that of $n = 4$ but samples of two and three are of occasional occurrence, especially in physical work, and now and again a value of the probable error of an experimental result is deduced from a set of two or of three observations."

Further on he states: "It is evident that the probable error determined from a set of three observations is very untrustworthy and that when there are only two observations it is very much worse."

Now in my original paper (*Biometrika*, VI, p. 1 [2]) I stopped at $n = 4$ because I had not realized that anyone would be foolish enough to work with probable errors deduced from a smaller number of observations, but now I too will complete my tables, which will I think emphasize the moderation of the second quotation from Mr Young's note.

Generally speaking there are two objects in determining the standard deviation of a set of observations, namely (1) to compare it with the standard deviation of similar sets of observations, and (2) to estimate the accuracy with which the mean of the observations represents the mean of the population from which the sample is drawn.

The former purpose is served by the table which Mr Young was engaged in completing, the latter, which is by far the most common use of the S.D., by the table which I gave in my original paper and which I now propose to complete downwards by including $n = 2$ and $n = 3$ and to extend upwards as far as $n = 30$.

In the tables the probability is given (to four places of decimals) that the mean of a unique sample shall lie between $-\infty$ and a distance z from the mean of the population, z being measured in terms of the S.D. (s) of the sample.

[By unique I mean to say that all the information which we have (or at all events intend to use) about the distribution of the population is given by the sample in question.]

Table of the Probability Integral of $z(x/s)$.

$z(=x/s)$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$	$n=9$	$n=10$	$n=11$	$n=12$	$n=13$	$n=14$	$n=15$	$n=16$	$n=17$	$n=18$	$n=19$
0·1	0·5317	0·5498	0·5633	0·5745	0·5841	0·5928	0·6006	0·6079	0·6146	0·6210	0·6269	0·6326	0·6380	0·6432	0·6481	0·6529	0·6575	0·6619
0·2	0·5628	0·5981	0·6241	0·6453	0·6634	0·6795	0·6936	0·7066	0·7185	0·7295	0·7397	0·7493	0·7583	0·7668	0·7748	0·7824	0·7896	0·7965
0·3	0·5928	0·6437	0·6804	0·7096	0·7340	0·7549	0·7733	0·7896	0·8042	0·8174	0·8294	0·8404	0·8505	0·8598	0·8683	0·8762	0·8836	0·8904
0·4	0·6211	0·6857	0·7309	0·7657	0·7940	0·8175	0·8375	0·8547	0·8697	0·8827	0·8942	0·9045	0·9135	0·9217	0·9289	0·9354	0·9413	0·9466
0·5	0·6476	0·7236	0·7749	0·8131	0·8428	0·8667	0·8863	0·9025	0·9161	0·9275	0·9373	0·9455	0·9527	0·9588	0·9640	0·9686	0·9725	0·9760
0·6	0·6720	0·7573	0·8125	0·8518	0·8813	0·9040	0·9218	0·9360	0·9473	0·9565	0·9640	0·9701	0·9752	0·9793	0·9827	0·9856	0·9879	0·9899
0·7	0·6944	0·7867	0·8440	0·8830	0·9109	0·9314	0·9468	0·9585	0·9675	0·9744	0·9798	0·9840	0·9873	0·9900	0·9920	0·9936	0·9949	0·9960
0·8	0·7148	0·8123	0·8701	0·9076	0·9332	0·9512	0·9640	0·9733	0·9801	0·9851	0·9888	0·9916	0·9937	0·9952	0·9964	0·9972	0·9979	0·9984
0·9	0·7333	0·8345	0·8915	0·9269	0·9498	0·9652	0·9756	0·9828	0·9878	0·9913	0·9938	0·9956	0·9968	0·9977	0·9983	0·9988	0·9991	0·9994
1·0	0·7500	0·8536	0·9092	0·9419	0·9622	0·9751	0·9834	0·9889	0·9925	0·9949	0·9966	0·9977	0·9984	0·9989	0·9993	0·9995	0·9997	0·9998
1·1	0·7651	0·8700	0·9236	0·9537	0·9714	0·9821	0·9887	0·9928	0·9954	0·9970	0·9981	0·9988	0·9992	0·9995	0·9997	0·9998	0·9999	0·9999
1·2	0·7789	0·8841	0·9354	0·9628	0·9782	0·9870	0·9922	0·9953	0·9971	0·9982	0·9989	0·9993	0·9996	0·9997	0·9998	0·9999		
1·3	0·7913	0·8963	0·9451	0·9700	0·9832	0·9905	0·9946	0·9969	0·9982	0·9990	0·9994	0·9996	0·9998	0·9999	0·9999			
1·4	0·8026	0·9069	0·9531	0·9756	0·9870	0·9930	0·9962	0·9979	0·9989	0·9994	0·9997	0·9998	0·9999					
1·5	0·8128	0·9160	0·9598	0·9800	0·9899	0·9948	0·9973	0·9986	0·9993	0·9996	0·9998	0·9999						
1·6	0·8222	0·9240	0·9652	0·9836	0·9920	0·9961	0·9981	0·9990	0·9995	0·9997	0·9999							
1·7	0·8308	0·9310	0·9699	0·9864	0·9937	0·9970	0·9986	0·9993	0·9997	0·9998								
1·8	0·8386	0·9371	0·9737	0·9886	0·9950	0·9977	0·9990	0·9995	0·9998	0·9999								
1·9	0·8458	0·9425	0·9770	0·9904	0·9959	0·9983	0·9992	0·9997	0·9999									
2·0	0·8524	0·9472	0·9797	0·9919	0·9967	0·9986	0·9994	0·9998										
2·1	0·8585	0·9514	0·9821	0·9931	0·9973	0·9989	0·9996	0·9998										
2·2	0·8642	0·9552	0·9841	0·9941	0·9978	0·9992	0·9997	0·9999										
2·3	0·8695	0·9585	0·9858	0·9950	0·9982	0·9993	0·9998											
2·4	0·8743	0·9615	0·9873	0·9957	0·9985	0·9995	0·9998											
2·5	0·8789	0·9642	0·9886	0·9963	0·9987	0·9996	0·9998											
2·6	0·8831	0·9667	0·9898	0·9967	0·9989	0·9996	0·9999											
2·7	0·8871	0·9689	0·9908	0·9972	0·9991	0·9997												
2·8	0·8908	0·9709	0·9916	0·9975	0·9992	0·9998												
2·9	0·8943	0·9727	0·9924	0·9978	0·9993	0·9998												
3·0	0·8976	0·9743	0·9931	0·9981	0·9994	0·9998												
3·1	0·9007	0·9759	0·9937	0·9983	0·9995	0·9999												
3·2	0·9036	0·9772	0·9942	0·9984	0·9996													
3·3	0·9063	0·9785	0·9947	0·9986	0·9996													
3·4	0·9089	0·9797	0·9951	0·9988	0·9997													
3·5	0·9114	0·9808	0·9955	0·9989	0·9997													

4·0	0·9220	0·9851	0·9969	0·9994	0·9998
4·5	0·9303	0·9881	0·9977	0·9996	0·9998
5·0	0·9372	0·9903	0·9984	0·9997	0·9999
5·5	0·9428	0·9919	0·9988	0·9998	
6·0	0·9474	0·9932	0·9990	0·9999	
6·5	0·9514	0·9942	0·9992		
7·0	0·9548	0·9950	0·9994		
7·5	0·9578	0·9956	0·9995		
8·0	0·9604	0·9961	0·9996		
8·5	0·9627	0·9966	0·9996		
9·0	0·9648	0·9969	0·9997		
9·5	0·9666	0·9973	0·9997		
10·0	0·9683	0·9975	0·9998		
15·0	0·9788	0·9989	0·9999		
20·0	0·9841	0·9994			
25·0	0·9873	0·9996			
30·0	0·9894	0·9997			
35·0	0·9909	0·9998			
40·0	0·9920	0·9998			
45·0	0·9929	0·9999			
50·0	0·9935				
60·0	0·9947				
70·0	0·9955				
80·0	0·9960				
90·0	0·9965				
100·0	0·9968				
120·0	0·9973				
140·0	0·9977				
150·0	0·9979				
160·0	0·9980				
180·0	0·9982				
200·0	0·9984				
250·0	0·9987				
300·0	0·9989				
350·0	0·9991				
400·0	0·9992				
450·0	0·9993				
500·0	0·9994				
600·0	0·9995				
700·0	0·9996				
1000·0	0·9997				
1500·0	0·9998				
2000·0	0·9998				
3000·0	0·9999				

$z\left(=\frac{x}{s}\right)$	$n=20$	$n=21$	$n=22$	$n=23$	$n=24$	$n=25$	$n=26$	$n=27$	$n=28$	$n=29$	$n=30$	For comparison $\left(\frac{\sqrt{(27)}}{\sqrt{(2\pi)}}\int_{-\infty}^{x} e^{-\frac{27x^2}{2}}dx\right)$
0·1	0·6662	0·6703	0·6744	0·6783	0·6821	0·6858	0·6894	0·6929	0·6964	0·6997	0·7030	0·6983
0·2	0·8030	0·8093	0·8152	0·8209	0·8264	0·8316	0·8367	0·8415	0·8462	0·8506	0·8550	0·8507
0·3	0·8967	0·9026	0·9082	0·9133	0·9182	0·9227	0·9270	0·9309	0·9347	0·9383	0·9415	0·9405
0·4	0·9613	0·9556	0·9595	0·9630	0·9662	0·9691	0·9717	0·9742	0·9763	0·9784	0·9801	0·9802
0·5	0·9789	0·9815	0·9838	0·9858	0·9875	0·9890	0·9903	0·9915	0·9925	0·9934	0·9942	0·9953
0·6	0·9915	0·9929	0·9940	0·9950	0·9958	0·9964	0·9970	0·9975	0·9979	0·9982	0·9985	0·9991
0·7	0·9967	0·9974	0·9979	0·9984	0·9987	0·9990	0·9992	0·9994	0·9995	0·9996	0·9997	0·9999
0·8	0·9988	0·9991	0·9993	0·9995	0·9996	0·9997	0·9998	0·9999	0·9999	0·9999	0·9999	
0·9	0·9995	0·9996	0·9997	0·9998	0·9999	0·9999	0·9999					
1·0	0·9999	0·9999	0·9999	0·9999								

To compare with the last column of the table ($n = 30$) I have given the corresponding probability calculated from the nearest normal curve, namely the one with S.D. $s/\sqrt{(n-3)}$ (not $s/\sqrt{(n-1)}$ as is usually given), and this shows I think that for ordinary purposes Sheppard's tables may be used with $n > 30$.

With regard to samples of two it will be seen that odds of 9 to 1 are reached at a little more than three times the S.D., of 99 to 1 at a little more than thirty times, of 999 to 1 at a little more than 300 times, while 9999 to 1 is reached at in or about 3000 times the S.D.!

Perhaps I may be permitted to restate my opinion as to the best way of judging the accuracy of physical or chemical determinations.

After considerable experience I have not encountered any determination which is not influenced by the date on which it is made; from this it follows that a number of determinations of the same thing made on the same day are likely to lie more closely together than if the repetitions had been made on different days.

It also follows that if the probable error is calculated from a number of observations made close together in point of time much of the secular error will be left out and for general use the probable error will be too small.

Where then the materials are sufficiently stable it is well to run a number of determinations on the same material through any series of routine determinations which have to be made, spreading them over the whole period.

Thus an analyst may be determining the percentage of nitrogen in different samples of seed corn and wish to know the probable error of the determination, i.e. how accurately his figures give the percentage of nitrogen in a bulk of corn.

Let us suppose that he makes ten determinations a day for sixty days and that it is of some real importance to him to get a clear idea of his error; he will do well to get sixty *different* samples from the same bulk of corn and analyse one of these on each of the sixty days; unless I am much mistaken he will have a more modest idea of his infallibility than he had before he compared the sixty results together. He will also, in so far as his repeated sample is representative, get a close approximation to the probable error of a single determination.

In some cases it is not possible to obtain a sufficient bulk of material, and then it may be better to determine each result in duplicate, the repetitions being separated as widely as possible in point of time. Then the square root of the mean of the squares of the differences between corresponding pairs gives twice the standard deviation of the average of a pair, and if enough pairs can be taken and the determinations made on different samples this is a better method than the other, as the error of the sampling is better sampled.

In the preparation of the tables a slight mistake was discovered in the second row of the odd numbers in the original table by Mr W. L. Bowie, to whom I am indebted for the calculation of the new figures.

9

AN EXPLANATION OF DEVIATIONS FROM POISSON'S LAW IN PRACTICE

[*Biometrika*, XII (1919), p. 211]

In her paper on the Poisson law of small numbers, *Biometrika*, x, pp. 36 *et seq.*, Miss Whitaker after a very interesting analysis of the various attempts which have been made to test Poisson's law on actual statistics concludes that "A general interpretation based on a very simple conception seems needed for those demographic cases in which the law of small numbers appears far more often to correspond to a negative than to a positive binomial".

The following is an attempt to explore the general question of what effect various departures from the conditions which lead to Poisson's law have on the resulting statistics, and especially which conditions lead to positive and which to negative binomials when the exponential might at first sight be expected.

Poisson's law has been applied to the occurrence of different numbers of individuals in divisions of space or time: thus of yeast cells in squares of a haemacytometer, of deaths from the kick of a horse in Prussian Army Corps which may be taken as individuals occurring in divisions of space, or of suicides of children per year in Prussia which are individuals occurring in divisions of time. In such cases it has been asserted that if the chance of an individual being found in a given division is so small that when multiplied by the very large number of individuals the product is still a reasonably small number, then the frequency of divisions containing 0, 1, 2, ... r individuals will be given by the terms of the exponential

$$Ne^{-m}\left\{1+m+\frac{m^2}{2!}+\ldots+\frac{m^r}{r!}+\ldots\right\},$$

where N is the number of divisions and m the mean number of individuals occurring in a division.

For the above to be true it is necessary

(1) That the chance of falling in a division is the same for each individual.

(2) That the chance of an individual falling in it is the same for each division.

(3) That the fact that an individual has fallen in a division does not affect the chance of other individuals falling therein.

As to these three conditions (1) is seldom or never true. I propose to show that this is generally unimportant; unless the chances of some individuals falling in a particular division are relatively high the Poisson law holds; the tendency however is towards a positive binomial.

Next (2) is comparatively seldom true except in the case of artificial divisions. The result of this, as Pearson has shown, is that a negative binomial fits the results better than the exponential.

Lastly (3) is often untrue. It will be shown that if the presence of an individual makes another less likely to fall into a division the positive binomial, but if more likely, the negative binomial, will fit the figures best.

We may start from the fact that if the chance of an event happening be q and of its not happening p, then the chances of its happening 0, 1, 2, etc. times in n trials are given by the terms of the expansion of $(p+q)^n$, viz.

$$p^n : np^{n-1}q : \frac{n(n-1)}{2!}p^{n-2}q^2 : \text{etc.}$$

As the moment coefficients of this series about the zero end of the range are

$$\nu_1 = nq,$$
$$\nu_2 = npq + n^2q^2, \text{ whence, } \mu_2 = npq,$$

the binomial is completely determined if we know ν_1 and μ_2 for

$$p = \frac{\mu_2}{\nu_1}, q = 1-p = 1-\frac{\mu_2}{\nu_1} \text{ and } n = \frac{\nu_1}{q} = \frac{\nu_1^2}{\nu_1-\mu_2},$$

and in particular the binomial is positive (i.e. n and q are positive) if $\mu_2/\nu_1 < 1$ and negative if $\mu_2/\nu_1 > 1$. In the particular case when $\mu_2/\nu_1 = 1$ the binomial becomes the Poisson exponential.

It is therefore unnecessary to deal with higher moments than the second for the purpose in hand.

Let us first consider the result of each individual having a different chance of falling in a given division.

Let the chances of n individuals falling in a given division be $q_1, q_2, q_3, \ldots q_n$. The chances of their not doing so are therefore $(1-q_1), (1-q_2), (1-q_3), \ldots (1-q_n)$, and the chances that 0, 1, 2, ..., n of them will fall in that division are given by the various terms of the expansion of

$$\{(1-q_1)+q_1\}\{(1-q_2)+q_2\}\{(1-q_3)+q_3\}\ldots\{(1-q_n)+q_n\},$$

i.e. by

$$(1-q_1)(1-q_2)\ldots(1-q_n) + S\{q_1(1-q_2)\ldots(1-q_n)\}$$
$$+ S\{q_1q_2(1-q_3)\ldots(1-q_n)\} + \ldots + S\{q_1q_2q_3\ldots q_r(1-q_{r+1})\ldots(1-q_n)\} + \ldots$$
$$+ q_1q_2q_3\ldots q_n,$$

the term $S\{q_1q_2q_3\ldots q_r(1-q_{r+1})\ldots(1-q_n)\}$ giving the chance that exactly r individuals will fall in the division.

The sum of the above series is clearly unity, so that the first and second moment coefficients about the zero end of the series are given by two series of which the rth terms are

$$rS\{q_1q_2\ldots q_r(1-q_{r+1})\ldots(1-q_n)\} \quad \text{and} \quad r^2S\{q_1q_2\ldots q_r(1-q_{r+1})\ldots(1-q_n)\}$$

respectively.

These series may be summed by rearranging them in the ascending order of the q products thus:

$$S\{q_1(1-q_2)(1-q_3)\dots(1-q_n)\}=S(q_1)-2S(q_1q_2)+\dots+(-1)^{r-1}r.S(q_1q_2\dots q_r)+\dots$$
$$2S\{q_1q_2(1-q_3)(1-q_4)\dots(1-q_n)\}=2S(q_1q_2)+\dots+(-1)^{r-2}r(r-1)\,S(q_1q_2\dots q_r)+\dots$$

....

$$tS\{q_1q_2\dots q_t(1-q_{t+1})\dots(1-q_n)\}=tS(q_1q_2\dots q_t)+\dots$$
$$+(-1)^{r-t}\frac{r.(r-1)!}{(t-1)!\,(r-t)!}S(q_1q_2\dots q_r)+\dots$$

....

$$rS\{(q_1q_2\dots q_r(1-q_{r+1})\dots(1-q_n)\}=\dots\dots\dots\dots r.S(q_1q_2\dots q_r)+\dots.$$

Adding these we get on the left ν_1 and on the right $S(q_1)+$ a number of terms of the form $r(1-1)^{r-1}S(q_1q_2\dots q_r)$ which accordingly vanish and we get

$$\nu_1=S(q_1).$$

In a similar manner it can be shown that

$$\nu_2=S(q_1)+2S(q_1q_2),$$

and other moment coefficients about zero can be found in the same way, but we are not here concerned with them.*

If $\bar{q}$, $\overline{q^2}$ are the mean values of q and q^2, obviously

$$\nu_1=S(q_1)=n\bar{q}, \qquad \dots\dots(1)$$

and

$$\nu_2=S(q_1)+2S(q_1q_2)=S(q_1)+\{S(q_1)\}^2-S(q_1^2)$$
$$=n\bar{q}+n^2\bar{q}^2-n\overline{q^2}, \qquad \dots\dots(2)$$
$$=n\bar{q}+n^2\bar{q}^2-n\bar{q}^2-n\sigma_q^2, \qquad \dots\dots(3)$$
$$\therefore\ \mu_2=n\bar{q}-n\bar{q}^2-n\sigma_q^2$$
$$=n\bar{q}\left(1-\bar{q}-\frac{\sigma_q^2}{\bar{q}}\right). \qquad \dots\dots(4)$$

If now the distribution of chances is to be represented by the binomial $(P+Q)^N$, then

$$Q=1-\frac{\mu_2}{\nu_1}=1-\frac{n\bar{q}(1-\bar{q}-\sigma_q^2/\bar{q})}{n\bar{q}}$$
$$=\bar{q}+\frac{\sigma_q^2}{\bar{q}}. \qquad \dots\dots(5)$$

* The moment coefficients are:

$\mu_2=n\overline{pq}-n\,{}_q\mu_2,$

$\mu_3=n\overline{pq}(\bar{p}-\bar{q})-3n(\bar{p}-\bar{q})\,{}_q\mu_2+2n\,{}_q\mu_3,$

$\mu_4=n\overline{pq}\{1+3(n-2)\,\overline{pq}\}-n\{7+6(n-6)\,\overline{pq}\}\,{}_q\mu_2+12n(\bar{p}-\bar{q})\,{}_q\mu_3-6n\,{}_q\mu_4+3n^2\,{}_q\mu_2^2,$

where ${}_q\mu_2$ etc. are the moment coefficients of the q distribution and $\bar{p}=1-\bar{q}$.

Since the original q's are the chances of events happening they are always positive, so that the above expression must be positive and the binomial positive.

If now we introduce the Poisson condition that $\bar{q}$ though positive is negligibly small (5) becomes in general zero, for σ_q is usually of the same order as $\bar{q}$, and in that case Poisson's law holds in spite of the inequality of the original q's. If however $\sigma_q^2/\bar{q}$ is appreciably greater than zero (as in the extreme case

$$q_1 = \tfrac{1}{2}, \quad q_2 = q_3 = \ldots = q_n = 0 \quad \text{when} \quad \frac{\sigma_q^2}{\bar{q}} = \frac{n-1}{2n} = \tfrac{1}{2}),$$

the distribution of chances is to be represented by a positive binomial.

Next we have to consider the effect of disregarding condition (2), namely that the chance of an individual falling into it must be the same for each division.

Let us suppose then that the q's are all different for each division, so that $n\bar{q}$ is also different.

Then writing m for $n\bar{q}$ and $\overline{m}$, $\overline{m^2}$, $\overline{nq^2}$ for the means of m, m^2 and nq^2 taken over all the divisions, we get from (1)

$$\nu_1 = \overline{m}, \qquad \ldots\ldots(6)$$

from (2)
$$\nu_2 = \overline{m} + \overline{m^2} - \overline{nq^2}$$
$$= \overline{m} + \overline{m}^2 + \sigma_m^2 - \overline{nq^2}, \qquad \ldots\ldots(7)$$
$$\therefore\ \mu_2 = \overline{m} + \sigma_m^2 - \overline{nq^2}.* \qquad \ldots\ldots(8)$$

As before, if $(P+Q)^N$ is the best-fitting binomial,

$$Q = 1 - \frac{\mu_2}{\nu_1} = \frac{\overline{nq^2} - \sigma_m^2}{\overline{m}}.$$

Hence if $\sigma_m^2 > \overline{nq^2}$, which if there is any appreciable variation in m is probable, since as explained above nq^2 is generally negligible, a negative binomial will be found to fit better than the exponential.

Clearly condition (2) is usually not fulfilled in the vital and demographic statistics; divisions either of space or time are generally governed by different environments which will vary the chances of an individual falling into them, and so we may expect that as a rule negative binomials will occur in place of the exponential.

Finally, suppose that the presence of an individual in a division influences the chance of other individuals falling in that division.

Clearly it may do so either by way of increasing the chance or diminishing it.

* If we suppose that q does not vary with the individual but that $nq\ (=m)$ varies with the division, the moment coefficients of the m distribution being written ${}_m\mu$, then the moment coefficients of the resulting distribution of divisions are as follows:

$$\mu_2 = \overline{m} + {}_m\mu_2,$$
$$\mu_3 = \overline{m} + 3\,{}_m\mu_2 + {}_m\mu_3,$$
$$\mu_4 = \overline{m} + 3\overline{m}^2 + (7 + 6\overline{m})\,{}_m\mu_2 + 6\,{}_m\mu_3 + {}_m\mu_4.$$

If the chance be increased it is clear that we shall get for the same mean number of individuals per division a larger number of divisions containing high numbers of individuals and a larger number of zero divisions. In other words, for the same mean we shall get a larger standard deviation, so that μ_2/ν_1 will be greater than 1 and a negative binomial will fit better than the exponential. On the other hand, if the chance of other individuals is decreased by the presence of one already in a division μ_2/ν_1 will become less than unity and the best-fitting binomial will be positive. The first of these two cases includes linking or clumping of events or bacteria, the second such a thing as the counting of large cells on a haemacytometer whose divisions are comparable in size with them.

We have now shown that a population which might be expected at first sight to follow Poisson's law

(1) Will do so if the only deviation from the ideal conditions is that the chances of different individuals falling into the same division are not equal, as long as these chances are all small.

(2) If in addition to this the chances of some individuals are large a positive binomial will fit the results better than the exponential.

(3) If the different divisions have different chances of containing individuals, as is usual, a negative binomial will fit the results better than the exponential, except in so far as (2) may interfere.

(4) If the presence of one individual in a division increases the chance of other individuals falling into that division, a negative binomial will fit best, but if it decreases the chance a positive binomial.

Generally speaking (3) is the operating deviation from Poisson's conditions and accordingly most statistics give negative binomials.

Finally, I should like to point out that the object of my original paper (*Biometrika*, vol. v [1]) was to give the user of the haemacytometer a guide to the error which he may expect from its use, and that the net result was that the probable error of his count was $0{\cdot}6745\sqrt{N}$, where N was the total number counted,* and that if N be a reasonably large number tables of the probability integral may be used, otherwise the exponential (or better still go on counting). This result is not affected by slight deviations from the Poisson law, any more than slight deviations from the normal law affect our use of the probability integral tables.

* *Biometrika*, v, p. 355. The probable error of mean is $0{\cdot}6745\sqrt{(m/M)}$, where m is the mean and M the number of unit areas counted. If in this we put $M = 1$, then $m = N$ and the total count is $N \pm 0{\cdot}6745\sqrt{N}$ as above.

10

AN EXPERIMENTAL DETERMINATION OF THE PROBABLE ERROR OF DR SPEARMAN'S CORRELATION COEFFICIENTS

[Being a paper read to the Society of Biometricians and Mathematical Statisticians, 13 December 1920]

[*Biometrika*, XIII (1921), p. 263]

In the *British Journal of Psychology*, II, p. 96,* Dr Spearman suggested two methods of determining correlation, based on replacing actual measurements by ranks.

As an illustration we may take the following purely imaginary example:

TABLE I

Individual	Height	Length of middle finger mm.	Rank in height	Rank in length of finger
A	6′ 0″	12·8	2	1
B	5′ 3″	11·5	4	3
C	5′ 7″	10·0	3	4
D	6′ 1″	12·4	1	2

Instead of correlating the figures in the second and third columns of the above table Dr Spearman proposed to use the figures in the fourth and fifth columns, and to determine one or other of two coefficients: of these the first (ρ) gives the ordinary correlation coefficient between the figures representing the ranks, and the second (R) was described as a "footrule" for correlation, i.e. a rough instrument which could be used by the unskilled. Dr Spearman also proposed to use R in cases where it was thought advisable to weight mediocre observations more heavily than extremes.

The method of determining ρ and R was to take the difference D between the numbers representing the ranks, e.g. for A in Table I

$$D = 2 - 1 = 1.$$

* [Dr Spearman's results were first given in a paper entitled "The proof and measurement of association between two things" in the *American Journal of Psychology*, XV, pp. 72–101. The dogmatic statements as to the accuracy of his methods in that paper are, I think, erroneous, and he does not lay adequate stress on the fact that correlation of ranks is not a correlation of variates and may differ very considerably from it. The suggestion of considering the correlation of ranks is due to A. Binet and V. Henri: see *La Fatigue Intellectuelle* (Paris, 1898), p. 252, also *L'Année Psychologique* (Paris, 1898), IV, p. 155. Their process is very obscure and they also do not appear to have realized that the correlation of variates is not that of ranks. K.P.]

Then
$$\rho = 1 - \frac{S(D^2)}{\frac{n(n^2-1)}{6}} \qquad \text{.....(i)}$$

and
$$R = 1 - \frac{S(D)}{\frac{n^2-1}{6}}, \qquad \text{.....(ii)}$$

where n is the number in the sample: in the case of R, $S(D)$ denotes the summation of positive differences only.

Dr Spearman gave an empirical formula connecting R and ρ, viz. $\rho = \sin(\frac{1}{2}\pi R)$, but I do not suppose that he attached any very great importance to this.

He further gave the probable errors of ρ and R for the case of no correlation as $0{\cdot}6745/\sqrt{n}$ and $0{\cdot}4266/\sqrt{n}$.

In his memoir "On further methods of determining correlation"* Prof. Pearson investigated these coefficients for the case of the normal correlation surface and found the relations between ρ and R and r the ordinary correlation coefficient to be

$$r = 2\sin\left(\frac{\pi}{6}\rho\right) \qquad \text{.....(iii)}$$

and
$$r = 2\cos\frac{\pi}{3}(1-R) - 1. \qquad \text{.....(iv)}$$

Pearson further found the standard error of ρ to be for large samples

$$\frac{1-\rho^2}{\sqrt{n}}\{1 + 0{\cdot}086\rho^2 + 0{\cdot}013\rho^4 + 0{\cdot}002\rho^6 + \ldots\}, \qquad \text{.....(v)}$$

and of r_ρ, i.e. r determined from ρ by (iii), to be

$$1{\cdot}0472\frac{1-r^2}{\sqrt{n}}\{1 + 0{\cdot}042r^2 + 0{\cdot}008r^4 + 0{\cdot}002r^6 + \ldots\}. \qquad \text{.....(vi)}$$

He did not succeed in evaluating the error of R or of r_R (i.e. of r determined by (iv)), but pointed out that just as in the case of r the $\sqrt{n}$ in the denominator is really $\sqrt{(n-1)}$. He also pointed out that R can only take values between $+1$ and $-0{\cdot}5$ and that Spearman's $0{\cdot}4226/\sqrt{(n-1)}$ does not imply that R is more accurate than ρ or r with their probable error of $0{\cdot}6745/(\sqrt{(n-1)}$, since R itself is smaller than ρ or r in about the same proportion.

Since that time the use of ρ and R has become general among psychologists, especially in America, where they are preferred to r on account of the ease and speed with which they can be determined for small samples.

For example, in a note on correlation in *Employment Psychology*, by H. C. Link,† a book written to urge the claims of Psychology on the devotees of "Scientific Management", the author mentioned three methods of determining correla-

* *Drapers' Company Research Memoirs*, Biometric Series, IV, 1907.

† Macmillan, 1919.

tion, ρ which is to be used for samples smaller than 30, R for samples over 30 and r which, though acknowledged to be rather more accurate, is not to be used at all since it takes four times as long to calculate as the others.

Now to save time at the expense of accuracy is justifiable when, and only when, the time saved can be devoted to increasing the number of observations so as to obtain greater accuracy on the whole series, otherwise it will take longer to get equally trustworthy conclusions, and it seems to be of interest to investigate the probable errors of ρ and R for samples of the size that the employment psychologist is contemplating. And here we may note that the saving of time only occurs when the sample is comparatively small; as it increases, the labour of grading becomes more and more severe till at some point in the neighbourhood of 40 it becomes quicker to use the ordinary product moment r if that be possible.

It should perhaps be pointed out that there are many cases where it is possible to grade a sample for some character which is not capable of being measured on a scale, and it might be thought that in this case large samples could profitably be dealt with by the ρ or R method, but in fact it is just these scaleless characters which present the greatest difficulty in grading.

We have then to consider the variability of ρ and R and of the derivatives r_ρ and r_R, determined from small samples, and it seemed worth while to use the material of a former sampling experiment so as to get an idea of how small samples depart from the results obtained by Prof. Pearson for ideally large samples. The material in question consists of 750 samples of four drawn from a population of 3000 criminals whose height and left middle finger length give an approximately normal correlation surface with correlation 0·66.

These are capable of being combined easily to give 375 samples of 8 and in addition there are 100 samples of 30, which may be taken to be a size of sample which is no longer quite "small".

Accounts of the former results were given in *Biometrika*, VI, p. 1 [2] and p. 302 [3], since which time the frequency distributions of the correlation coefficients of small samples drawn from normally correlated populations have been very thoroughly investigated by Soper, Fisher and the authors of the co-operative paper in vol. XI, p. 328 of *Biometrika*: it is hoped that some mathematician may be interested in the general solution of the problems raised in the present paper, which may then afford material for checking his results.

When I came to apply the methods to my samples I found that, owing to the rather coarse grouping, there were a large number of ties, so that it became necessary to find out the right correction for ties.

Prof. Pearson had discussed the question of ties and had suggested two ways of dealing with them. One way was to rank them all as if they were the highest number of the tie, which he called the bracket-rank method, and the other was to rank them all half-way down the tie, which he called the mid-rank method.

Thus the first way would rank 1, 2, 2, 4, while the second would rank 1, $2\frac{1}{2}$, $2\frac{1}{2}$, 4 if the second and third of four individuals constituted a tie.

Now the first would give different results according as we read the scale forwards or backwards and also alter the mean of the set of numbers, so I have only tried to use the mid-rank method, for which I have found the correction which follows.

Correction of ρ for Ties

If $D = x - y$, when x and y are any two variables measured from their means, then

$$D^2 = x^2 + y^2 - 2xy.$$

Summing for all n samples and dividing by n,

$$\frac{\Sigma(D^2)}{n} = \sigma_x^2 + \sigma_y^2 - 2r_{xy}\sigma_x\sigma_y,$$

$$\therefore\ r_{xy} = \left(\sigma_x^2 + \sigma_y^2 - \frac{\Sigma(D^2)}{n}\right) \Big/ 2\sigma_x\sigma_y. \qquad \text{.....(vii)}$$

If now x and y are the first n numbers, then

$$\sigma_x^2 = \sigma_y^2 = \frac{1}{n} \times \text{sum of squares of first } n \text{ numbers} - \left(\frac{\text{sum of first } n \text{ numbers}}{n}\right)^2$$

$$= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}$$

$$= \frac{n^2-1}{12}. \qquad \text{.....(viii)}$$

Substituting in (vii), we find

$$\rho = r_{xy} = \left(\frac{n^2-1}{6} - \frac{\Sigma(D^2)}{n}\right) \Big/ \frac{n^2-1}{6}$$

$$= \frac{\frac{n(n^2-1)}{6} - \Sigma(D^2)}{\frac{n(n^2-1)}{6}}. \qquad \text{.....(ix)}$$

Now suppose that there is on the x side a tie of t in number from q to $q+t-1$. Using the mid-rank method we substitute for each of the numbers

$$q,\ q+1,\ \ldots\ q+t-1 \text{ their mean } \frac{2q+t-1}{2}.$$

Hence in finding σ_x^2 the mean is unaltered, but in the sum of the squares $q^2 + (q+1)^2 + \ldots + (q+t-1)^2$ is replaced by $\frac{t(2q+t-1)^2}{4}$.

Hence σ_x^2 is smaller by

$$\frac{1}{n}\left\{q^2 + (q+1)^2 + \ldots + (q+t-1)^2 - \frac{t(2q+t-1)^2}{4}\right\}.$$

This is equal to

$$\frac{1}{n}\left\{tq^2+2q\{1+2+\ldots+(t-1)\}+\{1^2+2^2+\ldots+(t-1)^2\}-tq^2-qt(t-1)-\frac{t(t-1)^2}{4}\right\}$$

$$=\frac{1}{n}\left\{\frac{(t-1)t(2t-1)}{6}-\frac{t(t-1)^2}{4}\right\}$$

$$=\frac{t(t^2-1)}{12n},$$

$$\therefore\ \sigma_x^2=\frac{n^2-1}{12}-\frac{t(t^2-1)}{12n}.$$

This is clearly additive for any number of ties, so that if $T_x=\Sigma\left(\frac{t(t^2-1)}{12}\right)$ summing for all the ties on the x side and similarly T_y for the y side

$$\sigma_x^2=\frac{n^2-1}{12}-\frac{T_x}{n}\quad\text{and}\quad\sigma_y^2=\frac{n^2-1}{12}-\frac{T_y}{n},$$

and substituting in (vii),

$$\rho=r_{xy}=\frac{\left\{\frac{n^2-1}{6}-\frac{1}{n}(T_x+T_y)-\frac{\Sigma(D^2)}{n}\right\}}{\sqrt{\left\{\left(\frac{n^2-1}{6}-\frac{2T_x}{n}\right)\left(\frac{n^2-1}{6}-\frac{2T_y}{n}\right)\right\}}}$$

$$=\frac{\frac{n(n^2-1)}{6}-(T_x+T_y)-\Sigma(D^2)}{\sqrt{\left\{\left(\frac{n(n^2-1)}{6}-2T_x\right)\left(\frac{n(n^2-1)}{6}-2T_y\right)\right\}}}\qquad\ldots\ldots(x)$$

$$=\frac{\frac{n(n^2-1)}{6}-(T_x+T_y)-\Sigma(D^2)}{\left\{\frac{n(n^2-1)}{6}-(T_x+T_y)\right\}\sqrt{\left\{1-\frac{(T_x-T_y)^2}{\left\{\frac{n(n^2-1)}{6}-(T_x+T_y)\right\}^2}\right\}}}.$$

So that if T_x and T_y do not differ appreciably

$$\rho=1-\frac{\Sigma(D^2)}{\frac{n(n^2-1)}{6}-(T_x+T_y)}.\qquad\ldots\ldots(xi)$$

In estimating T_x or T_y each pair contributes $\frac{1}{2}$,

triplet ,, 2,
quartet ,, 5,
quintet ,, 10,

and so on. For example, if the x ranks for a sample of 10 were

$$1,\ 2\tfrac{1}{2},\ 2\tfrac{1}{2},\ 5,\ 5,\ 5,\ 8\tfrac{1}{2},\ 8\tfrac{1}{2},\ 8\tfrac{1}{2},\ 8\tfrac{1}{2},$$

T_x would be $\frac{1}{2}+2+5=7\frac{1}{2}$ and if there were no ties in the y ranks ρ would be

$$\frac{165-7\frac{1}{2}-S(D^2)}{\sqrt{\{(165-15)165\}}}=\frac{157\frac{1}{2}-S(D^2)}{\sqrt{(150.165)}},$$

and if we were to take it as $1-\dfrac{S(D^2)}{157\frac{1}{2}}$ the error would come in the third significant place of decimals.

In determining ρ for my 375 samples of 8 I found that much-tied samples usually gave low values of ρ, and it occurred to me that although undoubtedly equation (x) gives the true value of the correlation of ranks, yet it might be that the loss of precision due to ties would give low values for the correlation. To test this I doubled the width of my unit of grouping first for one variable and then for the other, so that I got three values of ρ for each sample:

(i) Converting the original figures into ranks.

(ii) Using coarser grouping on one side and the original grouping on the other before converting into ranks.

(iii) Using coarser grouping on both sides.

An example will make my meaning clearer.

(1) Original figures		(2) x grouped coarsely		(3) Both grouped coarsely		Ranks					
						(1)		(2)		(3)	
x	y	Putting +1 and 0 as $\frac{1}{2}$, etc.	y	x	y	x	y	x	y	x	y
0	+3	+ $\frac{1}{2}$	+3	+ $\frac{1}{2}$	+2½	5½	3	4½	3	4½	3½
−2	0	−1½	0	−1½	+ $\frac{1}{2}$	8	7	7½	7	7½	6½
+3	+3	+1½	+3	+2½	+2½	1½	3	1½	3	1½	3½
−1	−2	−1½	−2	−1½	−1½	7	8	7½	8	7½	8
+1	+3	+ $\frac{1}{2}$	+3	+ $\frac{1}{2}$	+2½	3½	3	4½	3	4½	3½
+1	+2	+ $\frac{1}{2}$	+2	+ $\frac{1}{2}$	+2½	3½	5	4½	5	4½	3½
+3	+4	+1½	+4	+2½	+4½	1½	1	1½	1	1½	1
0	+1	+ $\frac{1}{2}$	+1	+ $\frac{1}{2}$	+ $\frac{1}{2}$	5½	6	4½	6	4½	6½
				Pairs ...	...	3	—	2	—	2	1
				Triplets	...	—	1	—	1	—	—
				Quartets	...	—	—	1	—	1	1

Here originally $T_x=1\frac{1}{2}$ and $T_y=2$ and $\dfrac{n(n^2-1)}{6}-(T_x+T_y)=80\frac{1}{2}$.

After grouping x coarsely $T_x=6$ and $T_y=2$,, $=76$.

After grouping both coarsely $T_x=6$ and $T_y=5\frac{1}{2}$,, $=72\frac{1}{2}$,

and ρ will be found to take the values 0·832, 0·869 and 0·828 in succession. Working in this way I determined three values of ρ for each of the 375 samples and determined the mean, standard deviation and mean (T_x+T_y) for each of the three series of 375. These results are given in Table II.

Here an increase in the correction to be made for ties from 3·82 to 9·04 has

made a difference of 0·01 in the mean value of ρ, the probable error being about 0·015, and a still less appreciable difference in the standard deviation. It is,

TABLE II

	Mean ρ	σ	Mean T_x	Mean T_y	Mean (T_x+T_y)
Original series	0·5798	0·2887	1·92	1·90	3·82
x grouped coarsely ...	0·5798	0·2903	4·67	1·90	6·57
x and y grouped coarsely	0·5696	0·2874	4·67	4·37	9·04

I think, a fair inference that the correction is applicable to the series in question, and the reason for the observed low values of ρ in much-tied samples is to be sought elsewhere.* But it will be asked "what if no correction be made for ties?" The answer is that the mean value of ρ will rise as the ties become more numerous and the S.D. will fall. Thus Table II would become Table III if no corrections were made.

TABLE III

	Mean ρ	σ	Mean (T_x+T_y)
Original series	0·602	0·2677	3·82
x grouped coarsely ...	0·616	0·2887	6·57
x and y grouped coarsely	0·622	0·2414	9·04

At first sight this may appear to be highly advantageous, since the mean value approximates more nearly to the value which would be obtained from a large sample and the S.D. is smaller. A little reflexion will show, however, that the means of the ρ's of all populations would be subject to the same rise and that in fact the ρ of one population is no more differentiated from the ρ of another population than it is when corrected, while the mean value when corrected is constant over a fairly wide range of ties. If the correction is not made ρ can be cooked up to any required value by increasing the ties.

The fact is that as soon as there is a single tie, uncorrected ρ can no longer take all values between $+1$ and -1 and if one of the scales be reversed the correlation instead of being $-\rho$ becomes $-\rho+\frac{(T_x+T_y)}{\frac{n(n^2-1)}{6}}$. We are therefore forced to use the

* The low value of ρ for much-tied samples is due to the fact that a much-tied sample is as a rule one in which the S.D. of the original variables is low.

Now as a matter of experience I find that of samples drawn from a normally distributed population those with S.D. above the average tend to give high and stable values of the correlation coefficient, while those with S.D. below the average tend to give low and variable values.

The form of the correlation surface for variables σ_x and r_{xy} is of considerable interest to those who have to deal with small samples and merits the attention of mathematicians. I hope to deal with the experience obtained from my samples at some later time.

correction which after all gives us the distribution of ρ that we should get from ideal material containing no ties.

To see what happens when ties are carried to an extreme I determined ρ from the original table of 3000 entries (*Biometrika*, I, p. 216) and from the same table condensed to six groups each way by using a 4 in. scale of height and 0·8 mm. scale of finger lengths.

In the first case $\rho = 0{\cdot}637$ giving r_ρ 0·655 and in the second 0·557 with r_ρ 0·575. There seems therefore in extreme cases to be a tendency for the correction to give too low a value of ρ.

Correction of R for Grouping

In Dr Spearman's original paper R is defined as $1 - \dfrac{S(D)}{\dfrac{n^2-1}{6}}$ when $\dfrac{n^3-1}{6}$ is taken as the average value which $S(D)$ assumes.

The simplest way to see that this is the average value is to write down all the possible D's thus:

1	1					
2	2	1				
3	3	2	1			
4	4	3	2	1		
⋮	⋮	⋮	⋮	⋮		
$(n-1)$	$(n-1)$	$(n-2)$	$(n-3)$	$(n-4)$	...	1
n	n	$(n-1)$	$(n-2)$	$(n-3)$	...	2 1

Here the two columns on the left are composed of the first n numbers. The third column is formed by subtracting the top number of the first column from all the numbers in the second column in turn, the fourth by subtracting the second number from all numbers which give a positive remainder, and so on.

Thus the numbers in the second column could be arranged opposite the numbers of the first column in $n!$ ways.

And in $(n-1)!$ of these arrangements any given pair will occur.

Hence the average value of $S(D)$ will be

$$\frac{(n-1)!}{n!}\{[1+2+\ldots+(n-1)]+[1+2+\ldots+(n-2)]+\ldots+(1+2)+1\},$$

$$\therefore \text{ Average value of } S(D) = \frac{1}{n}\left\{\frac{n(n-1)}{2}+\frac{(n-1)(n-2)}{2}+\ldots+\frac{2.3}{2}+\frac{1.2}{2}\right\}$$

$$= \frac{1}{2n}S_1^n(n^2-n) = \frac{1}{2n}\left\{\frac{n(n+1)(2n+1)}{6}-\frac{n(n+1)}{2}\right\}$$

$$= \frac{(n+1)}{12}\{2n+1-3\} = \frac{n^2-1}{6}. \qquad \ldots\ldots\text{(xii)}$$

If we now substitute in the second column ties instead of consecutive numbers we can find out what effect ties will have on the average value of $S(D)$. As I can see no general way of proving the results I propose merely to state my results as follows:

(1) A tie of t on one side which is opposed by no ties on the other side will diminish $\frac{n^2-1}{6}$ by $\frac{t(t^2-1)}{24n}$ if t be odd and by $\frac{t(t^2-4)}{24n}$ if t be even.

(2) Overlapping ties on opposite sides interfere with the above simple rule, the total to be subtracted from $\frac{n^2-1}{6}$ being increased or decreased according to Table IV.

TABLE IV

		Distance between centres of ties											
x tie	y tie	0	½	1	1½	2	2½	3	3½	4	4½	5	5½	6
2	2	+ 1		0→										
3	2		0→											
3	3	+ 2		−1		0→								
4	2	+ 2		0→										
4	3		+ 1		−1		0→							
4	4	+ 6		0		−1		0→						
5	2		0→											
5	3	+ 4		−1		−1								
5	4		+ 3		−2		−1		0→					
5	5	+10		0		−4		−1		0→				
6	2	+ 3		0→										
6	3		+ 2		−1		−1		0→					
6	4	+10		+1		−2		−1		0→				
6	5		+ 7		−2		−4		−1		0→			
6	6	+19		+4		−4		−4		−1		0→		
7	2		0→											
7	3	+ 6		−1		−1		−1		0→				
7	4		+ 5		−2		−2		−1		0→			
7	5	+16		+2		−5		−4		−1		0→		
7	6		+13		−1		−7		−4		−1		0→	
7	7	+28		+7		−6		−10		−4		−1		0→

etc., etc.

As an example of the use of Table IV, suppose a set of eight ranks to contain on the x side a tie of 5 centred at 3, i.e. let the x ranks be 3, 3, 3, 3, 3, 6, 7, 8, and let the y ranks have a tie of 4 centred at 2½, i.e. let the y ranks be 2½, 2½, 2½, 2½, 5, 6, 7, 8. Then the amount to be subtracted from $\frac{8^2-1}{6}$ is firstly $\frac{5}{8}$ (for the 5 tie) $+\frac{2}{8}$ (for the 4 tie) $+\frac{3}{8}$ (from Table IV) $= 1\frac{1}{4}$. Had the y ranks been 1, 3½, 3½, 3½, 3½, 6, 7, 8, the correction would be the same, but if the y ranks were 1, 2, 4½, 4½, 4½, 4½, 7, 8, the correction would be $\frac{5}{8}+\frac{2}{8}-\frac{2}{8}=\frac{5}{8}$,

and if 1, 2, 3, 5½, 5½, 5½, 5½, 8, $\frac{5}{8}+\frac{2}{8}-\frac{1}{8}=\frac{3}{4}$,

and if 1, 2, 3, 4, 6½, 6½, 6½, 6½, $\frac{5}{8}+\frac{2}{8}=\frac{7}{8}$.

It is only with very small and much-tied samples that the correction is appreciable.

TABLE V

Giving Frequency Distributions of Various Correlation Coefficients from 375 samples of 8

	−·8005 to −·7505	−·7505 to −·7005	−·7005 to −·6505	−·6505 to −·6005	−·6005 to −·5505	−·5505 to −·5005	−·5005 to −·4505	−·4505 to −·4005	−·4005 to −·3505	−·3505 to −·3005	−·3005 to −·2505	−·2505 to −·2005	−·2005 to −·1505	−·1505 to −·1005	−·1005 to −·0505	−·0505 to −·0005	−·0005 to ·0495	·0495 to ·0995	·0995 to ·1495	·1495 to ·1995	·1995 to ·2495	·2495 to ·2995	·2995 to ·3495	·3495 to ·3995	·3995 to ·4495	·4495 to ·4995	·4995 to ·5495	·5495 to ·5995	·5995 to ·6495	·6495 to ·6995	·6995 to ·7495	·7495 to ·7995	·7995 to ·8495	·8495 to ·8995	·8995 to ·9495	·9495 to ·9995
(1) Calculated from the co-operative paper	—	—	0·1	0·1	0·1	0·2	0·3	0·3	0·4	0·5	0·7	0·9	1·1	1·4	1·7	2·2	2·7	3·3	4·0	5·0	6·1	7·5	9·2	11·2	13·7	16·6	20·0	23·7	28·2	32·8	37·0	40·3	40·5	35·3	22·1	5·9
(2) r actual using Sheppard's correction	—	—	1	0	0	1	0	1	2	0	0	2	2	1	2	0	3	4	5	4	8	7	13	9	5	20	14	19	37	28	39	33	36	46	23	10
(3) r actual using no corrections	—	—	—	1	0	1	0	0	2	1	0	1	3	1	2	0	3	4	5	5	7	7	13	11	9	18	15	22	39	28	38	31	39	40	22	7
(4) r_ρ actual ...	—	—	—	—	—	—	1	2	2	2	0	2	0	0	5	10	2	2	4	6	8	6	9	16	17	13	20	19	26	25	37	39	34	34	21	11
(5) r_R actual ...	1	0	0	0	1	0	3	1	0	1	0	3	3	4	2	2	4	3	7	11	8	15	20	7	20	16	26	19	27	21	55	34	19	29	6	
(6) ρ actual ...	—	—	—	—	—	—	—	3	1	3	0	1	1	0	5	10	2	4	3	8	8	8	11	16	18	15	22	21	25	24	39	36	35	29	21	8
(7) R actual ...	—	—	—	—	—	—	—	—	1	3	0	1	1	4	6	4	11	9	12	5	8	11	16	18	15	22	21	25	24	39	36	35	29	21	8	
																				22	21	27	16	28	24	29	43	31	27	19	19	10	2	3	1	—

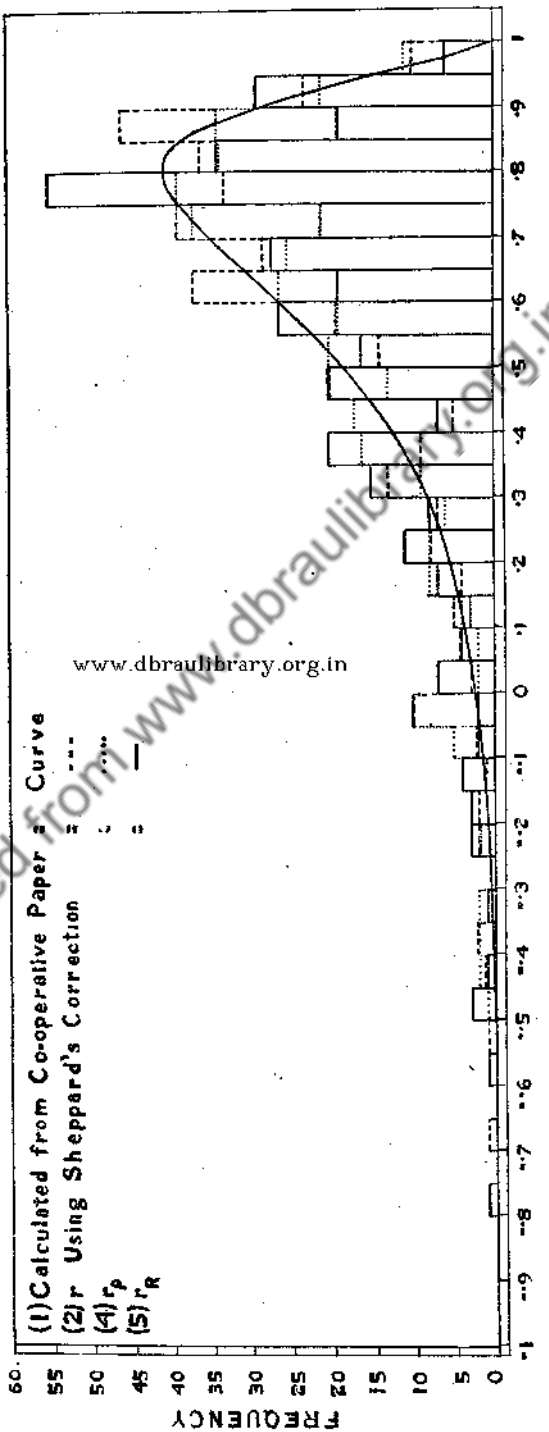
TABLE VI

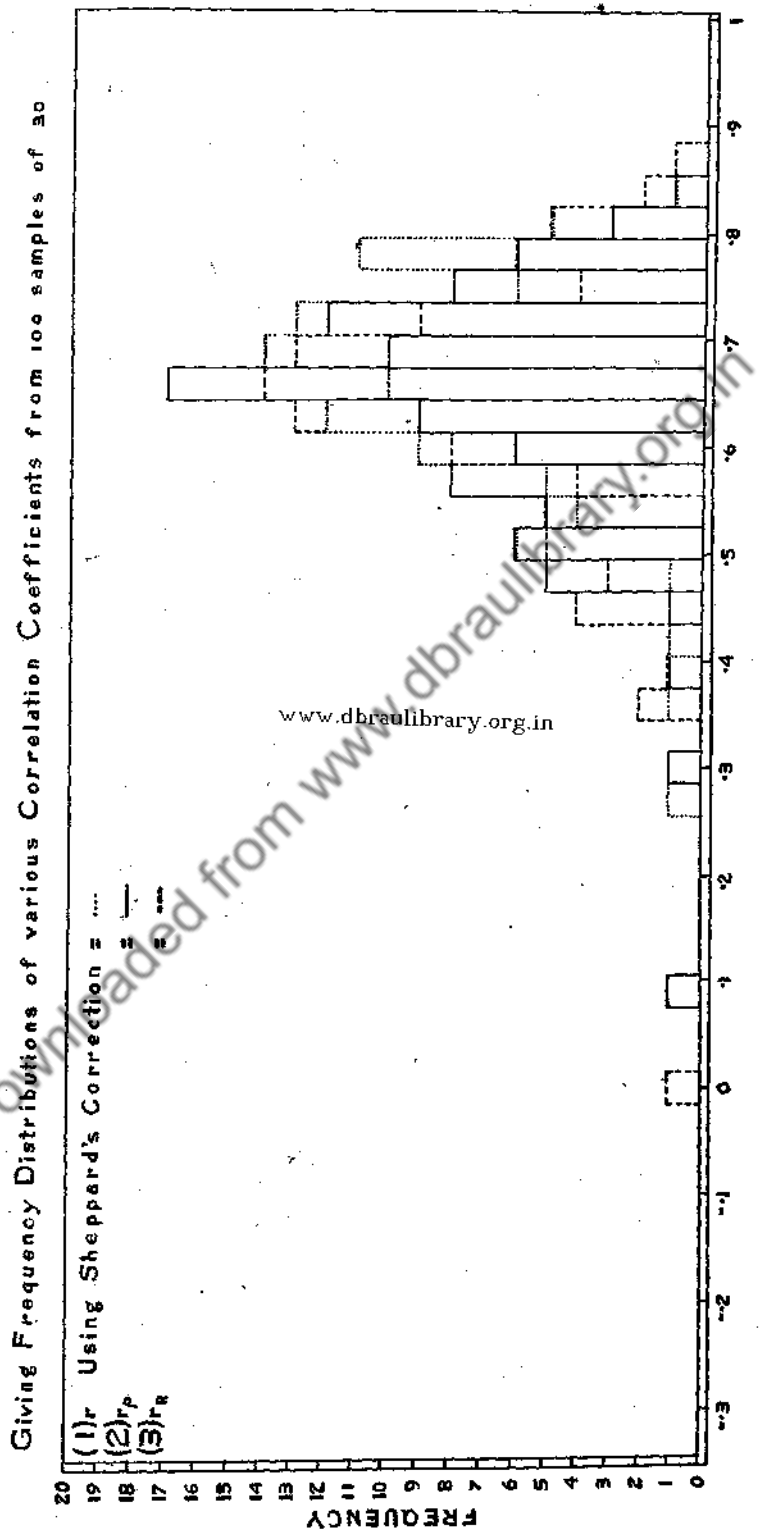
Giving Frequency Distributions of Various Correlation Coefficients from 100 samples of 30

	−·165 to −·135	−·135 to −·105	−·105 to −·075	−·075 to −·045	−·045 to −·015	−·015 to ·015	·015 to ·045	·045 to ·075	·075 to ·105	·105 to ·135	·135 to ·165	·165 to ·195	·195 to ·225	·225 to ·255	·255 to ·285	·285 to ·315	·315 to ·345	·345 to ·375	·375 to ·405	·405 to ·435	·435 to ·465	·465 to ·495	·495 to ·525	·525 to ·555	·555 to ·585	·585 to ·615	·615 to ·645	·645 to ·675	·675 to ·705	·705 to ·735	·735 to ·765	·765 to ·795	·795 to ·825	·825 to ·855	·855 to ·885	·885 to ·915
(1) r actual using Sheppard's correction	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	0	0	1	0	1	0	1	6	4	5	9	12	10	14	13	6	11	5	1	—	—
(2) r_ρ actual	—	—	—	—	—	—	—	—	1	0	0	0	0	0	0	1	0	0	1	0	1	5	6	5	8	6	9	17	10	12	8	6	3	1	—	—
(3) r_R actual	—	—	—	—	—	1	0	0	0	0	0	0	0	0	0	0	0	2	1	0	4	3	5	5	4	8	13	14	13	9	4	6	5	2	1	—
(4) r actual by median fourfold division $= \cos \frac{\pi B}{A+B}$	1	0	0	0	0	0	0	0	1	1	0	0	1	1	0	1	3	0	0	3	7	0	10	5	5	5	9	8	6	9	5	1	9	6	1	2
(5) ρ actual	—	—	—	—	—	—	—	1	0	0	0	0	0	0	0	1	0	1	0	1	2	7	4	8	5	10	13	13	8	12	6	5	2	1	—	—
(6) R actual	—	—	—	—	—	1	0	0	0	0	0	0	2	1	3	6	5	7	9	15	17	12	8	4	6	3	0	0	1	—	—	—	—	—	—	—

GRAPH FROM TABLE V.

Giving Frequency Distributions of various Correlation Coefficients from 375 samples of 8





Discussion of the Frequency Distributions obtained

Tables V and VI give the frequency distributions of r, determined with Sheppard's corrections for grouping, of ρ and of R and their derivatives (from equations (iii) and (iv)) r_ρ and r_R.

In addition we have, in Table V, r determined without Sheppard's corrections and the theoretical distribution of r calculated from the table (*Biometrika*, XI, p. 384), of the co-operative paper by interpolating between $\rho = 0{\cdot}65$ and $\rho = 0{\cdot}70$, drawing the frequency curve and estimating the areas by counting the squares: it is probably not very accurate, but fairly close to the truth.

In Table VI is included r calculated from the fourfold table taken through the medians by Sheppard's formula $r = \cos\frac{\pi B}{A+B}$, where B is the frequency in the "small" cells. This probably suffers a good deal from the coarse grouping which makes it necessary to divide the centre groups in an arbitrary manner.

The most remarkable thing about these tables is the very wide spread of all the distributions. There is of course nothing new in this, but I cannot help thinking that an examination of these tables may be beneficial for all who try to work with very small samples.

Besides this there is not very much to be found in these tables which is not seen to greater advantage in Tables VIII and IX of the means and standard deviations, but as a matter of interest I have compared lines 2–4 of Table V with line 1 by the χ^2 test with the following results:

Table VII

	25 groups		16 groups	
	χ^2	P	χ^2	P
r with Sheppard's corrections	30·10	0·18	20·49	0·17
r without Sheppard's corrections	20·43	0·67	12·39	0·66
r_ρ actual	60·25	0·000,064	30·63	0·01
r_R actual	74·10	0·000,002	55·65	(say) 0·000,002

The 25 groups were the 24 groups on the right of the table and the tail, which includes all groups which are less than 1·0 in line 1: the 16 groups were taken so that no group in line 1 was less than 10.

With such a small sample as 375 the χ^2 test is only decisive for considerable departures from the theoretical and the regular excess over the theoretical for all groups less than 0·40 avoids detection.

At the same time it is interesting to note that, judged by the χ^2 test, Sheppard's corrections do not seem to have improved the calculation of r.

Tables VIII and IX give the means, standard deviations and coefficients of

TABLE VIII

Certain Constants of the Frequency Distributions of Various Correlation Coefficients derived from 375 *samples of* 8

	Mean	S.D.	Coefficient of variation	Number of samples required to give as great accuracy as 100 samples of (1)	Number of samples required to give as great accuracy as 100 samples of (2)
(1) r calculated from co-operative paper	0·631	0·250	39·6	100	—
(2) r actual using Sheppard's corrections	0·624 ±0·010	0·274 ±0·007	43·9 ±1·3	120 ±5·9	100
(3) r actual using no correction for grouping	0·614 ±0·010	0·271 ±0·007	44·1 ±1·3	117 ±5·8	98
(4) r_ρ actual	0·586 ±0·010	0·291 ±0·007	49·7 ±1·5	135 ±6·7	113
(5) r_R actual	0·566 ±0·011	0·309 ±0·008	54·6 ±1·7	153 ±7·5	127
(6) ρ actual	0·580 ±0·011	0·289 ±0·007	49·8 ±1·9	—	—
(7) R actual	0·407 ±0·008	0·237 ±0·006	58·2 ±1·9	—	—

TABLE IX

Certain Constants of the Frequency Distributions of Various Correlation Coefficients derived from 100 *samples of* 30

	Mean	S.D.	Coefficient of variation	Number of samples required to give as great accuracy as 100 samples of (1)	Number of samples required to give as great accuracy as 100 samples of (2)
(1) r calculated from co-operative paper	0·653	0·109	16·7	100	—
(2) r actual using Sheppard's corrections	0·661 ±0·007	0·101 ±0·005	15·3 ±0·7	86 ± 8·2	100
(3) r_ρ actual	0·639 ±0·008	0·113 ±0·005	17·7 ±0·9	108 ±10·3	125
(4) r_R actual	0·638 ±0·008	0·122 ±0·006	19·1 ±0·9	125 ±11·9	146
(5) r actual from median fourfold division $= \cos \frac{\pi B}{A+B}$	0·609 ±0·012	0·183 ±0·009	30·1 ±1·6	282 ±25·1	328
(6) ρ actual	0·624 ±0·008	0·116 ±0·006	18·6 ±0·9	—	—
(7) R actual	0·428 ±0·007	0·100 ±0·005	23·4 ±1·2	—	—

variation of the frequency distributions in Tables V and VI and in addition the calculated constants for the samples of 30.

As well as this I have calculated the number of samples which would be required to give as great accuracy by the less accurate methods as 100 samples determined (1) on the theoretical basis of normal correlation, and (2) on the actual samples by the product-moment method using Sheppard's corrections.

The object of this is to get an idea of how much time must be saved in order to gain by using the rank methods. First, however, we may note in Table VIII the marked difference between the theoretical S.D. and that actually obtained by the product-moment method.

I attribute this almost entirely to the grouping, which was unfortunately rather coarse and which cannot be corrected by Sheppard's corrections in small samples. The slight divergency of the population from normal correlation may have helped to a very small extent, but for the most part the excess in the lower values of r which cause the mean to be low and the S.D. to be high is due to those samples which have low S.D.'s, and I incline to believe that if the grouping can be chosen so that the S.D.'s are not less than 3 the actual distribution of r will be found to be very close to the calculated for samples drawn from normally correlated material.

In Table IX, on the other hand, the actual has a higher mean and lower S.D. than the calculated, but as the differences are in each case less than twice the probable error, I think we may put them down to the error of random sampling, which is of course large in such a small sample as 100.

Next we may note that Prof. Pearson's formulae, no doubt because they are correct for grades, do not enable us to correct rank correlations for small samples. The means of both r_ρ and r_R are too low for samples of 8, and for samples of 30 probably so.

As for the S.D.'s of ρ and ρ_r, the values found are in the case of samples of 8 much higher than those calculated from equations (v) and (vi), which are 0·258 and 0·243 respectively. The samples of 30, however, give values which agree sufficiently well, for the calculated S.D. is in each case 0·114, well within the probable error.

Line 5 in Table IX shows that as determined in this investigation Sheppard's median division formula gives a mean value of r well below the population value and a very high S.D.* While this is not unlikely to be the case for small samples the arbitrary division of the central groups makes it impossible to say that this is not due to the fact that we have only used an approximation to median division in this case.

The chief point of interest however in Tables VIII and IX lies in column 4, showing the number of samples which we must have to get the same accuracy by the various methods as that given by 100 samples in which r is determined with sufficiently fine grouping by the product-moment method.

Column 5 is put in in case there are any who do not accept my explanation of the difference between the calculated and actual distribution of r, namely that it is due to the coarse grouping. I have not been able to estimate the probable errors of the figures in column 5 as they are complicated by correlation between

* The S.D. calculated from the formula $\sigma_r = \frac{2\pi\sqrt{(1-r^2)}}{\sqrt{N}}\left\{\frac{ab}{N^2}\right\}^{\frac{1}{4}}$ is, however, rather higher, being 0·191 if r be taken as 0·66 and 0·207 if r be taken as 0·609. Miss Elderton kindly looked up this formula for me, but I cannot find that it has been published. [See, however, *Biometrika*, IX, p. 23. It is also involved in the early paper by Sheppard, *Phil. Trans.* A, CXCII, pp. 147 *et seq.* K.P.]

the numerator and denominator of the fractions from which the figures are calculated. They must, however, be larger than the probable errors in column 4.

In any case there is a strong indication that with samples of 8 the loss of accuracy due to the use of r_ρ instead of r will practically always more than counterbalance the gain of time in calculation. Either method is, however, so little to be depended upon for a single sample of very small size, except as the merest indication, that very little is lost by the use of r_ρ. If, however, a number of small samples can be averaged so as to obtain a coefficient of some value, the product-moment method should be used when possible.

With samples of 30 the 8 % more samples required compares fairly with Prof. Pearson's 10 % more for large samples, but seeing that the particular sample of 100 gave too low a value for σ_r, the value of σ_{r_ρ} which must be correlated with it is likely to be low also and the 8 % may easily be 18 % or more.

In any case it would very seldom pay to have to collect 8 % more samples of 30 even if one could save 8 % of the time on samples of that size.

In both tables there is a considerable loss from the use of r_R instead of r_ρ, since from 13 to 16 % more samples would be required of the former to give the same accuracy as the latter. The gain in calculation is not very appreciable, since most of the time is spent in ranking the samples. Dr Spearman prefers R to ρ at times because less importance attaches to outlying samples, but as the extremes of small samples tend to be outliers even in normally correlated material owing to the phenomenon to which attention was drawn in Galton's Difference problem,* it seems to me that as much weight as possible should be given to them.

To Combine Two Methods of Determination

At an early stage in the investigation I hoped to be able to combine r and r_ρ to get a value less subject to error than either. Curiously enough Prof. Pearson in his editorial in the last number of *Biometrika* gives the equations which I proposed to use for the purpose (p. 7 (29)).

As they are perfectly general I will state them in a slightly more general form.

If x and y be two estimates of any quantity obtained in different ways, then a quantity z can always be found which will have a lower error than either of them, unless x and y are perfectly correlated.

Thus
$$z = \frac{\sigma_y^2 - r_{xy}\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2-2r_{xy}\sigma_x\sigma_y}x + \frac{\sigma_x^2 - r_{xy}\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2-2r_{xy}\sigma_x\sigma_y}y, \qquad \text{.....(xiii)}$$

and
$$\sigma_z^2 = \frac{\sigma_x^2\sigma_y^2(1-r_{xy}^2)}{\sigma_x^2+\sigma_y^2-2r_{xy}\sigma_x\sigma_y}. \qquad \text{.....(xiv)}$$

* *Biometrika*, I, pp. 385–99. In this connexion it is of interest to note that the correlation surface of ranks is not an elliptical hill as is the normal correlation surface but two comparatively steep ridges joined by a saddle, the ridges having a skew section.

Giving Correlation (0·885) *between r (no corrections) and* r_ρ *for samples of 8*

r \ r_ρ	−·5005 to −·4505	−·4505 to −·4005	−·4005 to −·3505	−·3505 to −·3005	−·3005 to −·2505	−·2505 to −·2005	−·2005 to −·1505	−·1505 to −·1005	−·1005 to −·0505	−·0505 to −·0005	−·0005 to −·0495	·0495 to ·0995	·0995 to ·1495	·1495 to ·1995	·1995 to ·2495	·2495 to ·2995	·2995 to ·3495	·3495 to ·3995	·3995 to ·4495	·4495 to ·4995	·4995 to ·5495	·5495 to ·5995	·5995 to ·6495	·6495 to ·6995	·6995 to ·7495	·7495 to ·7995	·7995 to ·8495	·8495 to ·8995	·8995 to ·9495	·9495 to ·9995	Totals
−·6505 to −·6005	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
−·6005 to −·5505	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0
−·5505 to −·5005	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
−·5005 to −·4505	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0
−·4505 to −·4005	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0
−·4005 to −·3505	—	1	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2
−·3505 to −·3005	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
−·3005 to −·2505	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0
−·2505 to −·2005	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
−·2005 to −·1505	—	—	1	—	—	—	—	—	2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3
−·1505 to −·1005	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
−·1005 to −·0505	—	—	—	1	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2
−·0505 to −·0005	—	—	—	—	—	—	—	—	—	—	2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0
−·0005 to ·0495	—	—	—	—	—	—	—	—	1	—	—	1	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3
·0495 to ·0995	—	—	1	—	—	—	—	—	1	2	—	—	—	—	—	—	—	2	—	—	—	—	—	—	—	—	—	—	—	—	4
·0995 to ·1495	—	—	—	—	—	1	—	—	—	1	—	—	1	1	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	5
·1495 to ·1995	—	—	—	—	—	—	—	—	1	1	—	—	1	1	—	2	2	1	—	—	—	—	—	—	—	—	—	—	—	—	5
·1995 to ·2495	—	—	—	—	—	—	—	—	—	—	—	1	—	1	2	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	7
·2495 to ·2995	—	—	—	—	—	—	—	—	—	2	—	—	1	2	2	3	4	—	1	2	—	1	—	—	—	—	—	—	—	—	13
·2995 to ·3495	—	—	—	—	—	—	—	—	—	—	—	—	—	1	1	1	2	4	2	—	2	—	1	—	—	—	—	—	—	—	11
·3495 to ·3995	—	—	—	—	—	—	—	—	—	1	—	—	—	1	1	—	—	1	1	2	1	—	—	—	—	—	—	1	—	—	9
·3995 to ·4495	—	—	—	—	—	—	—	—	—	—	—	—	—	1	3	1	—	3	1	3	1	5	1	1	1	—	—	—	—	—	18
·4495 to ·4995	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	1	1	3	4	1	2	2	3	—	—	—	—	—	15
·4995 to ·5495	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	1	1	4	5	5	2	1	—	1	—	—	22
·5495 to ·5995	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	2	—	3	—	5	4	8	7	5	2	—	—	—	39
·5995 to ·6495	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	2	4	2	2	9	5	2	1	—	28
·6495 to ·6995	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	1	3	1	—	3	4	9	7	7	—	—	38
·6995 to ·7495	—	—	—	—	—	—	—	—	1	1	—	—	—	—	—	—	—	—	—	1	—	2	—	2	2	4	8	5	5	2	31
·7495 to ·7995	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2	7	12	9	5	4	39
·7995 to ·8495	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2	2	1	1	5	7	15	5	2	40
·8495 to ·8995	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	1	—	1	6	9	4	22
·8995 to ·9495	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	1	5	7
·9495 to ·9995	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Totals	1	2	2	2	0	2	0	0	5	10	2	2	4	8	8	6	9	16	17	13	20	19	26	25	37	39	34	34	21	11	375

TABLE XI

*Giving Correlation (0·903) between ρ Calculated in Original Grouping (ρ_1) and ρ Calculated in Groups twice as coarse (ρ_2)**

ρ_2 \ ρ_1	−·4995 to −·4495	−·4495 to −·3995	−·3995 to −·3495	−·3495 to −·2995	−·2995 to −·2495	−·2495 to −·1995	−·1995 to −·1495	−·1495 to −·0995	−·0995 to −·0495	−·0495 to +·0005	+·0005 to ·0505	·0505 to ·1005	·1005 to ·1505	·1505 to ·2005	·2005 to ·2505	·2505 to ·3005	·3005 to ·3505	·3505 to ·4005	·4005 to ·4505	·4505 to ·5005	·5005 to ·5505	·5505 to ·6005	·6005 to ·6505	·6505 to ·7005	·7005 to ·7505	·7505 to ·8005	·8005 to ·8505	·8505 to ·9005	·9005 to ·9505	·9505 to 1·0005	Totals
−·4995 to −·4495	1									1					1															1
−·4495 to −·3995									1																					3
−·3995 to −·3495																														0
−·3495 to −·2995				1																										1
−·2995 to −·2495		1																												1
−·2495 to −·1995				2						1																				3
−·1995 to −·1495		1				1				1																				3
−·1495 to −·0995										1																				1
−·0995 to −·0495									1		1																			2
−·0495 to +·0005									1	1		1	2			1		1												8
+·0005 to ·0505			1						1	2				1					1											5
·0505 to ·1005																														3
·1005 to ·1505																	1													1
·1505 to ·2005									1	1		1	1			2	2	1	1											9
·2005 to ·2505									1				1	5		2	2	3												12
·2505 to ·3005														1	1			2	1	1										10
·3005 to ·3505						1				1			2	2	1	2	1	2	4	1	1	2								11
·3505 to ·4005												1					1	4	2	2	1	2	1							14
·4005 to ·4505															1		2	2	3	2	3	2	1							18
·4505 to ·5005																	1	2	3	4	3	2	2	2	1					19
·5005 to ·5505																	1	2	1	5	6	6	2	2		1				24
·5505 to ·6005																				1	1	2	8	5	5	2	1			29
·6005 to ·6505														1					1		1	3	5	9	4	1	2			27
·6505 to ·7005																	1				2	6	4	2	8	6	2	2		33
·7005 to ·7505																						1	1	3	9	6	6	1	1	29
·7505 to ·8005																			1				4	3	3	10	6		4	31
·8005 to ·8505																									3	6	6	4	3	1	23
·8505 to ·9005																								3	2	3	9	14	4		35
·9005 to ·9505																										1	3	4	2	1	11
·9505 to 1·0005																												2	5	1	8
Totals	1	2	1	3	0	2	0	0	6	10	1	3	6	10	9	5	12	16	18	15	23	24	26	27	35	35	36	27	19	3	375

* The reader is requested to note that the subranges here are not the same as in Tables V, VI and X. This is the only table which contains a perfect correlation coefficient which occurred in the ρ_2 series.

In the case of the samples of 8, x may be taken as r without Sheppard's corrections and y as r_ρ, when we have

$$\sigma_x^2 = (0{\cdot}271)^2 = 0{\cdot}073{,}441, \quad \sigma_x\sigma_y = 0{\cdot}078{,}861,$$
$$\sigma_y^2 = (0{\cdot}291)^2 = 0{\cdot}084{,}681,$$
$$r_{xy} = 0{\cdot}885,$$

and hence from (xiii),

$$z = 0{\cdot}804r + 0{\cdot}196r_\rho$$

and

$$\sigma_z = 0{\cdot}270,$$

i.e. there is no appreciable gain in our case since σ_r is 0·271. It may be that with a lower value of the population correlation the gain would be greater, but on the other hand if r had been determined for very fine grouping σ_r^2 would have been 0·0625, the contribution of r_ρ to z would have been practically negligible, and the gain in accuracy by the use of z less than that found. There is, however, another case where the above formulae might be applied, namely to the values of ρ obtained from the original grouping and those from coarse grouping.

These are given in Table II from the first and third lines of which it appears that σ_{ρ_1} and σ_{ρ_3} may both be taken as 0·288.

In this case σ_z^2 reduces to $\dfrac{\sigma_\rho^2(1+r_{\rho_1\rho_3})}{2}$,

and as

$$r_{\rho_1\rho_3} = 0{\cdot}903, \quad \sigma_z = 0{\cdot}281.$$

This is somewhat more encouraging, but the process is rather troublesome and could only be applied to cases where there is a proper scale. If, however, there is a proper scale greater accuracy could be obtained by the product-moment method with very little more trouble (since we have now to make two calculations to find ρ).

We may therefore conclude that as far as this sampling experiment may be taken as typical:

(1) Where the unit of grouping is small (say $< \frac{1}{3}$ the S.D.) the product-moment method should be used if the most is to be made of the time and statistics at our disposal, however small the sample.

(2) Where a coarse grouping has to be used, the mean value of r will fall below that calculated from the co-operative paper (*Biometrika*, XI, pp. 328 *et seq.*) and the S.D. will rise. For small samples Sheppard's corrections will approximately correct the former but will increase the latter still further. Indeed it is possible that for very coarse grouping ρ might vary less than r.

(3) For this, or any other, purpose ties should be dealt with by one or other of the formulae in equations (x) and (xi) of this paper.

(4) Where one or both variables can be ranked but not scaled, as frequently happens in some kinds of work, or for what Prof. Pearson has called "purposes

of assay", ρ can be determined with advantage and may be considered the natural method to adopt.

(5) In such cases it should be borne in mind that for small samples the distribution of ρ is similar to the distribution of r, but that the mean, even of r_ρ, is lower than that of r and the S.D. greater, by amounts which doubtless depend on the population correlation.

(6) R and r_R are not worth determining in serious work; their use should therefore be confined to the elementary statistics for which its author intended R.

(7) It is interesting to observe that Sheppard's median division fourfold table has given for small samples a mean value very much below the population value. While this is only what one might have expected, it may in this case be due to the coarse grouping which prevented me from making an accurate median division.

(8) The following problems might be of interest to mathematicians:

(*a*) The determination of the form of the rank correlation surface.

(*b*) The determination of the frequency distribution of ρ for small samples drawn from a normally distributed population.

(*c*) The determination of the nature of the correlation surface when a standard deviation is taken as one variable and the correlation coefficient as the other, both being determined from small samples drawn from a normally distributed population.

11

ON TESTING VARIETIES OF CEREALS

[Being a Paper read to the Society of Biometricians and Mathematical Statisticians, 28 May, 1923]

[*Biometrika*, XV (1923), p. 271]

Object of Experiments

The object of testing varieties of cereals is to find out which will pay the farmer best. This may depend on quality, but in general it is an increase of yield which is profitable, and since yield is very variable from year to year and from farm to farm it is a difficult matter upon which to obtain conclusive evidence.

Yet it is certain that very considerable improvements in yield have been made as the result of replacing the native cereals by improved varieties; as an example of this I may cite the case of Ireland, where varieties of barley have been introduced which were shown by experiment to have an average yield of 15 to 20 % above those which they replaced. This represents, probably, a gain to the country of not less than £250,000 per year. As the cost of experiments from the commencement to the present time cannot have reached £40,000 the money has been well spent.

Origin of Varieties

In the first place the ordinary cereals, wheat, barley, oats, and so on (maize is not here considered), are all self-fertilized and occur in races broadly distinguished by different botanical characters—Potato Oats, Rivett Wheat, Chevalier Barley, and so forth.

Besides these botanically distinguishable races, it is possible to pick out strains from commercial seed which differ from one another in all kinds of ways: time of ripening, percentage of nitrogen, yield, etc., although botanically the same. Many of these strains have been selected from time to time, certainly from the end of the eighteenth century up to the present time.

Finally there are hybrids, the result of deliberate crossing, and the selection of the best individuals out of the many thousands which may be grown in three generations is one of the more difficult problems with which the plant breeder has to deal, but it is only after he has made his preliminary selection that his hybrids concern the experimenter who is testing varieties.

Owing to the fact of self-fertilization, the various races, strains, and even to a large extent the hybrids, remain practically constant from year to year if once pure seed has been obtained.

Chief Sources of Error

The peculiar difficulties of the problem lie in the fact that the soil in which the experiments are to be carried out is nowhere really uniform; however little it may vary to the eye, it is found to vary not only from acre to acre but from yard to yard, and even from inch to inch. This variation is anything but random, so that the ordinary formulae for combining errors of observation which are based on randomness are even less applicable than usual.

Next, of course, is the weather: that will hardly affect experiments carried out in the same field in the same year, but experiments carried out in different districts and seasons meet with variations of weather which may produce results quite inconsistent with the experimental error determined at either place. Obviously, the weather needs to be well sampled before drawing general conclusions.

The effects of soil and weather on the yields are far greater than the differences which we have to investigate, and it is because the planning of experiments and their interpretation when completed are not quite straightforward that this paper has been written.

Methods of Operating

There are, broadly speaking, two methods of operating:

(i) On a large enough scale to use the ordinary agricultural implements, ploughs, seed drills, reaping machines, etc.

(ii) On quite a small scale with spades and dibblers, and scissors, under a wire net to keep out birds and rabbits.

Taking first the large scale, it has the advantage that the farmer, who always has a healthy contempt for gardening, may pay some attention to the results; he is to this extent right, that large-scale conditions cannot be accurately reproduced in a wire cage, and in fact some varieties which have come out well on the small scale have not done as well in the field, though this is not at all common. Large-scale work, then, is necessary as a final demonstration, and historically, it was on the large scale that variety experiments were first carried out.

Large-Scale Work

As an instance of large-scale work we may take a series of experiments carried out by the Department of Agriculture in Ireland to find out the best variety of barley to grow in that country.

The experiments lasted six years, *vide* Table I, and during that time seven varieties were tested; only two, however, Archer and Goldthorpe, were carried through from start to finish, as the others were either dropped when they were found to be inferior, or were not among those chosen in the first place. The original seed was ordinary commercial seed, and the plots were two acres in extent. This is very large even for a large-scale plot, but it was intended that the produce

should form the raw material for further manufacturing experiments. This was a wise precaution, as has been found recently when a barley in other ways among the best was found to be quite unsuitable as malting material.

The produce of the plots was all valued (in those days—1901–6—values were fairly steady from year to year), and this gives a method of combining yield and quality, but although the quality varied very much from one farm to another, there was generally only a small difference between the quality of different varieties grown on the same farm in the same season. The value of the crop per acre depended chiefly on the yield.

During the six years 193 plots were grown and at different times eighteen farms provided the land. These farms were scattered up and down the barley-growing districts in Ireland. Here, however, we shall deal only with the 51 plots of Archer, and the corresponding 51 plots of Goldthorpe.

The value per acre, then, of the 51 Archer plots varied between 90*s*. and 234*s*. with a mean of 178*s*. and a standard deviation of 33·6*s*. The value per acre of the Goldthorpe plots varied between 99*s*. and 230*s*. with a mean of 166*s*. and a standard deviation of 33*s*. The difference, therefore, was 12*s*., and at first sight this hardly appears significant, for had the Archer and Goldthorpe plots been independent, the standard deviation of their difference would have been about 6·5.

This brings us to the first principle of all agricultural experiments, viz. that only comparative values are of any use. If we are told that on a certain farm a new variety of barley produced 30 cwt. to the acre, we admit that the crop is good, but are not much interested. If, in addition, we hear that Archer gave 25 cwt. to the acre on the same farm, we begin to take notice; for it is some evidence as to the value of the new variety, and it is the difference of 5 cwt. to the acre which appeals to us and not the actual yields themselves. In point of fact, of course, the yields in these experiments were not independent. Each Archer has a corresponding Goldthorpe, and by considering the 51 differences, we find that the mean difference between Archer and Goldthorpe has a standard deviation of 3·3*s*.

This reduction of the standard deviation of the mean difference from 6·5 to 3·3*s*., by considering the individual differences between corresponding pairs, depends of course on the fact that corresponding pairs are highly correlated, so that the last term in the formula

$$\sigma^2_{A-B} = \sigma^2_A + \sigma^2_B - 2r_{AB}\sigma_A\sigma_B$$

is by no means negligible. The art of designing all experiments lies even more in arranging matters so that r_{AB} is as large as possible, than in reducing σ^2_A and σ^2_B.

That the conclusion that Archer was better than Goldthorpe was fully justified is shown by the fact that taking the yearly averages Archer beat Goldthorpe every year, while in the individual farms Archer beat Goldthorpe in all but three out of eighteen, and of these one farm was only used one season, and the other two in two seasons. Further, it was discovered during the course of the experiments

that the Archer was practically identical with a barley which the Danes called Prentice, which had beaten all others in their long series of experiments. Both Archer and Goldthorpe were, practically speaking, new to Ireland, and they—or some improvement* on them—have now almost entirely driven out the other inferior barleys from most parts of the country.

Such, then, is the sort of error which attaches to large experimental plots, that is to say a standard deviation of about 10–15 % for a *single* comparison, and this is found to be the order of the error in all ordinary large-scale work—it does not vary very closely with the size of the plot, provided that the plot be above say one-tenth of an acre, though there may be a slight decrease of error with increase of size.

It follows that although it is quite within the power of any individual farmer to carry out a large-scale experiment (and the larger the easier to carry out), it is only by co-operation that enough evidence can be obtained to be of any value. This co-operation can in practice only be arranged by a government department, a large agricultural company, or a farmers' association, and it is government departments that have had most success.

Small-Scale Work

We may next discuss small-scale work, leaving to the end a modification introduced by Dr E. S. Beaven, which combines the advantages of the ordinary large scale with a considerably smaller error. The considerations which led to this modification were derived from experience of small-scale technique.

Preliminary Considerations. Before coming to any actual comparison of varieties on the small scale, attention is directed to some preliminary experiments carried out by three different sets of investigators: Stratton and Wood at Cambridge,† Mercer and Hall at Rothamsted,‡ and Montgomery at Nebraska Agricultural Experimental Station.§

The first harvested $\frac{9}{10}$th acre of mangolds in $\frac{1}{1000}$-acre plots: the second, one acre of wheat in $\frac{1}{500}$-acre plots, and an acre of mangolds in $\frac{1}{200}$-acre plots: the third two years in succession harvested the same $\frac{7}{45}$th acre of wheat in $\frac{1}{1440}$-acre plots, and all weighed the produce of each plot; Montgomery determined the percentages of nitrogen as well. All three experiments showed the same thing: that

* In particular a hybrid of Archer with Spratt made by Capt. Hunter, Spratt-Archer 37/6, which proved its superiority to Archer and other varieties in "chessboard" trials similar to that detailed below.

† *J. Agric. Sci.* III, p. 417, "The interpretation of experimental results".

‡ *J. Agric. Sci.* IV, p. 107, "The experimental error of field trials".

§ *Nebr. Agric. Expt. Sta. 25th Ann. Report*, 1910–11, pp. 164–80, "Variation in yield and methods of arranging plots to secure comparative results"; and *U.S. Dept. Agric. Bur. Plant Indust. Bul.* 269, "Experiments in wheat breeding: experimental error in the nursery and variation in nitrogen and yield".

the variation is not random; the yield varies from point to point with an irregular regularity; there is consequently correlation between one plot and its neighbours, and generally there is a tendency for one end of a field to yield more than the other.

This is only what is to be expected from a priori considerations; naturally the nearer two plots are together the more likely is the soil and its condition to be similar on each of them, and the obvious conclusion may be drawn that the smaller the plots the more exactly can the yield of adjacent plots be compared.

Taking the investigation of Mercer and Hall on the 500 "plots" of wheat, it should be noted that they were only taken as plots at harvest and before cutting formed an unusually uniform area of one acre, part of a much larger field of wheat. The mean yield of grain per plot was 3·95 lb. with a range of 2·75–5·14, and a standard deviation of 0·46 lb., or 11·6 % of the mean weight of a plot.

If two adjacent plots were taken as $\frac{1}{250}$-acre plots the S.D. fell to 10 % instead of the 8·2 % of random sampling.

If four adjacent plots were taken as $\frac{1}{125}$-acre plots the S.D. fell to 8·9 % instead of the 5·8 % of random sampling.

If ten adjacent plots were taken as $\frac{1}{50}$-acre plots the S.D. fell to *6·3 % instead of the 3·7 % of random sampling.

If twenty adjacent plots were taken as $\frac{1}{25}$-acre plots the S.D. fell to *5·7 % instead of the 2·6 % of random sampling.

If fifty adjacent plots were taken as $\frac{1}{10}$-acre plots the S.D. fell to *5·1 % instead of the 1·6 % of random sampling.

The high value of the standard deviation of the larger plots compared with that which would have been expected had the aggregation been carried out randomly is due to a similar cause to that which decreased the error of the comparison of Archer and Goldthorpe. There is correlation between the neighbouring small plots which make up the larger plots, so that the last term in the formula

$$\sigma^2_{A+B} = \sigma^2_A + \sigma^2_B + 2r_{AB}\sigma_A\sigma_B$$

is not negligible. This last term is in fact the bridge over a pitfall which has trapped many, including—as will be shown later—the present writer.

In an appendix to Mercer and Hall's paper I pointed out that advantage may be taken of this correlation if we consider the difference between adjacent plots.

Thus we have

Size of plot (acres)	S.D. of single plot as percentage	Calculated S.D. of difference between random pairs	Actual S.D. of difference between adjacent pairs	Total acreage required to reduce S.D. of a comparison to 1 %
1/500	11·6	16·4	11·2	0·50 acre
1/250	10·0	14·1	9·7	0·74 ,,
1/125	8·9	12·6	9·3*	1·37 ,,
1/50	6·3	8·9	3·7*	1·10 ,,
1/25	5·7	8·1	3·9*	3·84 ,,

* The numbers are too few to do much more than indicate the tendency.

Except in the case of the $\frac{1}{125}$-acre plots we actually find that the standard deviation of a difference between two plots is less than the standard deviation of a single plot, and that working with $\frac{1}{500}$-acre plots, the standard deviation of a comparison between the varieties grown on a total area of half an acre is as low as 1%. On the lines of the 2-acre plots more than half a square mile would have been required. Further, there is every indication that smaller plots would be still more economical of ground.

These have been termed preliminary experiments, and so they are for the purpose of this paper; but in point of fact they followed the practical application of the principle which has just been outlined, and a further step in advance had already been made.

Carrying the principle of maximum contiguity, which he had deduced a priori, to its extreme logical limit, Beaven had compared two varieties in his cage by sowing alternate rows. He used a pure line of Archer barley, and one of a variety called "Plumage", which is allied to the Goldthorpe of the Irish experiments. He also grew $\frac{1}{10}$th acre of each outside the cage and found that whereas the Archer gave slightly the better yield outside the cage, the cage work gave the yield of Plumage some 20 % better than the Archer.

He sent me the figures to look at, and I found that so far from the correlation between the yields of adjacent drills being positive, it was significantly negative.

This was quite unexpected at the time (1905), but the explanation was simple, viz. that when a plant of one variety is grown next to one of another variety it is abnormally situated, and is subject to abnormal competition.

In this case the Plumage was a taller barley and shaded the Archer; probably, also, it started growth more quickly undergound and so annexed more of the soil than its competitor. Anyhow, it was clear that a comparison of adjacent rows, with the possibility of interference of this kind was useless.

The Square Yard Plot

To avoid this difficulty, Beaven invented in 1909 the "square yard" plot, which is formed by sowing eight rows 6 inches apart, 4 feet long, and with seed 2 inches apart in the row. This gives in the first place a plot 4 feet by 4 feet; but at harvest the outside rows are rejected and the outside 6 inches at each end of all the other rows, thus leaving the inside square* yard for the measurement of yield free from the competition of other varieties.

* There has been some controversy in America as to the advisability of testing varieties in alternate rows, but lately T. A. Kiesselbach (*J. Amer. Soc. Agron.* (1919), No. 6, pp. 235–41, "Experimental error in field trials"; pp. 242–7, "Plant competition as a source of error in field plots") has come to much the same conclusion as Beaven, viz. that although certain varieties *may* not under some circumstances interfere with one another, yet it is dangerous to allow any chance of the experiment being subject to this source of error, and that the only safe thing to do is to surround each experimental area with a border of the variety grown upon it, and to discard this border at harvest.

TABLE I

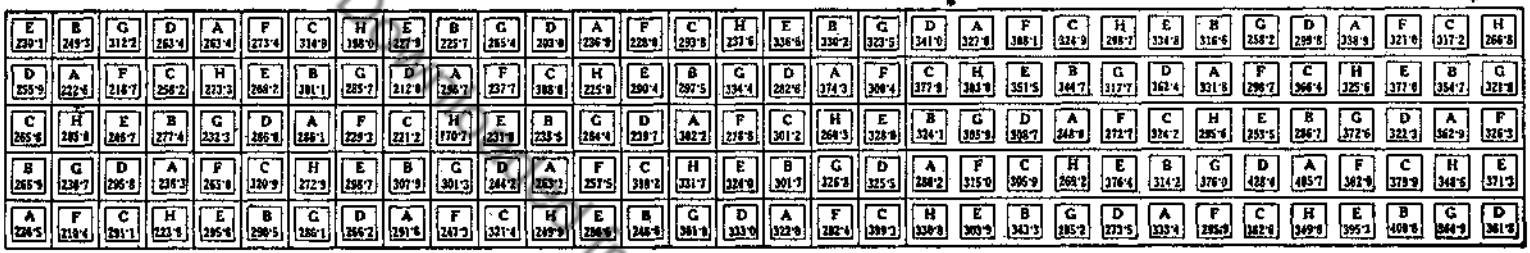
Irish Experimental Barley Plots. Yield and Money Value Per Acre of Archer and Goldthorpe 1901–6

Farmer	Place	District	Archer					Goldthorpe				
			Yield		V. P. A.			Yield		V. P. A.		
			Barrels	Stones	£	*s.*	*d.*	Barrels	Stones	£	*s.*	*d.*
1901:												
McCarthy	Ballinacurra	Cork	11	4	9	0	0	7	0	5	2	0
Hawkins	Whitegate	,,	10	3	8	3	0	7	12	6	3	0
Dwan	Thurles	Central Plain	15	2	11	13	0	13	14	11	0	0
Wolfe	Nenagh	,,	11	0	8	13	0	10	0	8	3	0
1902:												
McCarthy	Ballinacurra	Cork	12	6	8	13	0	11	14	8	11	0
Hawkins	Whitegate	,,	14	0	10	12	0	13	0	10	3	0
Wolfe	Nenagh	Central Plain	12	2	9	4	0	13	6	10	2	0
Willington	Birr	,,	12	6	9	16	0	9	3	7	6	0
Gorman	Enniscorthy	Wexford	11	5	9	2	0	11	14	9	2	0
Nunn	Castlebridge	,,	11	3	8	18	0	11	4	9	0	0
1903:												
McCarthy	Ballinacurra	Cork	6	10	4	13	0	7	4	5	5	0
Hawkins	Whitegate	,,	8	12	7	1	0	7	5	5	19	0
Wolfe	Nenagh	Central Plain	8	2	5	9	0	8	7	6	11	0
Willington	Birr	,,	9	13	7	9	0	8	0	6	6	0
Gorman	Arnestown	Wexford	5	5	4	10	0	7	11	5	15	0
Nunn	Castlebridge	,,	12	7	9	16	0	9	15	7	16	0
Quinn	Carlingford	Louth	11	12	8	4	0	9	3	7	0	0
Kearney	Greenore	,,	11	3	8	12	0	7	13	5	19	0
1904:												
McCarthy	Ballinacurra	Cork	10	4	7	15	0	11	14	9	8	0
Hawkins	Whitegate	,,	10	11	8	7	0	10	4	8	4	0
Wolfe	Nenagh	Central Plain	13	3	10	9	0	11	8	9	7	0
Willington	Birr	,,	11	3	8	17	0	11	14	9	4	0
Kelly	Portarlington	,,	12	1	9	12	0	11	3	9	0	0
Allardyce	Monasterevan	,,	10	7	8	1	0	10	5	8	7	0
Roche	New Ross	Wexford	8	2	5	16	0	7	0	5	6	0
Nunn	Castlebridge	,,	9	2	7	8	0	6	4	4	19	0
Kearney	Carlingford	Louth	8	0	6	7	0	9	7	7	9	0
Segrave	Dunleer	,,	12	1	9	9	0	11	7	9	7	0
1905:												
McCarthy	Ballinacurra	Cork	12	8	9	8	0	13	1	9	16	0
Hawkins	Whitegate	,,	11	11	9	8	0	11	5	8	14	0
Wolfe	Nenagh	Central Plain	14	6	10	14	0	15	10	10	3	0
Willington	Birr	,,	14	11	11	14	0	13	8	10	11	0
Luttrell	Monasterevan	,,	14	8	11	0	0	12	13	9	14	0
Kelly	Portarlington	,,	12	1	8	19	0	10	8	7	17	0
Matthews	Tullamore	,,	13	12	10	18	0	10	10	8	1	0
Nunn	Castlebridge	Wexford	11	6	7	15	0	11	6	8	0	0
Dooley	New Ross	,,	13	0	10	0	0	13	10	10	12	0
Kearney	Carlingford	Louth	14	6	9	12	0	11	4	7	12	0
Segrave	Dunleer	,,	14	7	11	11	0	12	8	9	19	0
1906:												
McCarthy	Ballinacurra	Cork	9	14	7	6	0	9	11	7	2	0
Hawkins	Whitegate	,,	10	9	7	12	0	8	14	6	9	0
Wolfe	Nenagh	Central Plain	11	12	8	14	0	8	13	6	11	0
Willington	Birr	,,	10	15	8	0	0	9	15	7	6	0
Luttrell	Monasterevan	,,	9	10	7	3	0	10	9	7	17	0
Mulhall	,,	,,	12	8	9	6	0	13	14	10	7	0
Matthews	Tullamore	,,	8	14	6	16	0	8	11	6	11	0
Tennant	Bagnalstown	,,	15	4	11	7	0	13	14	10	9	0
Nunn	Castlebridge	Wexford	11	10	8	19	0	10	9	7	18	0
Dooley	New Ross	,,	14	3	10	7	0	12	5	9	0	0
Kearney	Carlingford	Louth	11	11	8	9	0	12	12	9	8	0
Segrave	Dunleer	,,	14	6	10	8	0	13	6	9	16	0

NOTE. The Irish barrel of barley contains 16 stones.

Direction described as "vertical"

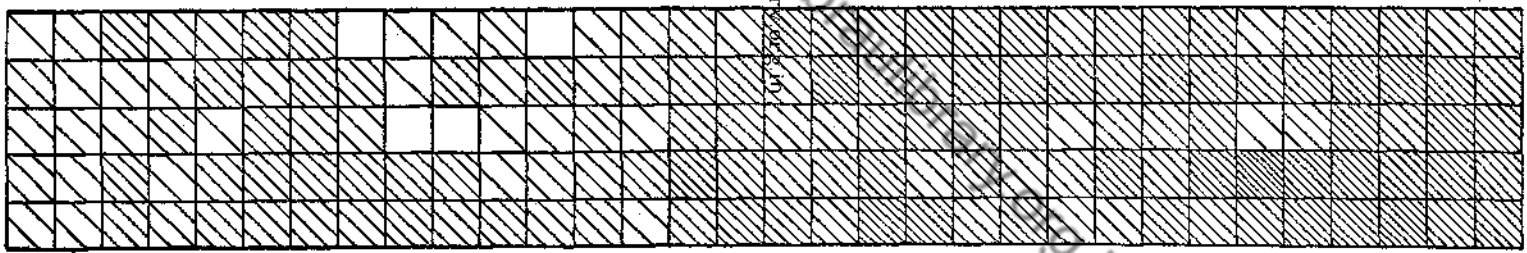
DIAGRAM I. Giving the Yields in Grammes of the Individual Plots of Dr Beaven's No. 1 Yield Experiment of 1913



——Direction described as right ——→ and left ←——

A. English Archer. B. 145. C. 145/46. D. Plumage.
E. Early Archer. F. Irish Archer. G. 7 A. H. Biffen.

DIAGRAM II. Plan of Experimental Area, showing Variation in Yield in Different Parts. Each line represents 20 grammes more yield compared with the mean of the race grown on that plot



* Owing to a slip of the draughtsman the vertical rules between these asterisks have been omitted; the omission has no significance.

So far as I am aware, no one has made any further inquiry as to the most economical size of plot; the square-yard plot only utilizes for yield determination $\frac{9}{16}$th of the experimental area, and to make it smaller would waste still more ground, while the larger the plot the more we depart from the principle of maximum contiguity.

There are probably not enough data to discover by the calculus the size of plot which will give the minimum probable error per acre, and no one seems to have faced the labour of an experimental determination. At all events, without any further investigation the square yard plot has been adopted as the unit in some six or seven experimental cages in the British Isles.

Comparison on a "Chessboard"

Having adopted the unit, it was a comparatively simple matter to set units of two varieties in a "chess" or "chequer" board: subsequently it was found that more than two varieties could be economically compared at the same time.

To illustrate the problems which arise when we come to compare several varieties grown together on a "chessboard", we may take Beaven's No 1 Yield Experiment of 1913.*

In this, 20 plots of each of eight races of barley were grown on a regular system of repetition, and the following observations were made for each plot:

Number of plants,
Number of ears,
Weight of ears,
Weight of straw.

For the purpose of this illustration we need only consider yield of corn, i.e. weight of ears.

The eight races consisted of

Four strains of Archer	English Archer 7*A*	Selection made by Beaven.
	Irish or Early Archer / Irish Archer, No. 5	Selection made by Capt. H. Hunter, B.Sc., of the Irish Department of Agriculture.
Plumage		A selection made by Beaven which originated in Denmark. Wide-eared barley somewhat like Goldthorpe.

Each of these was, of course, descended from a single seed a few generations back, and

Three hybrids	145 and 145/46	From a Plumage-Archer cross made by Beaven, the second being a re-selection from the first.
	"Biffen"	Selected by the Professor of that name from a Plumage-Archer cross of his own.

In order to simplify the comparison of errors it is best to work as long as possible, not with the standard error but the "variance", or square of the standard error. It has two advantages: (i) that variance can be added or subtracted without

* *Vide* Diagram I, p. 97.

the preliminary squaring and subsequent extraction of the square root, and (ii) that the area required to give any required accuracy varies directly with it; in order to give the same error a comparison with a variance of 60 only requires half as much ground as a comparison with a variance of 120.

Further, the variance taken in each case will be the variance of the average of 20 plots or differences between plots, or whatever it may be, and to get this we divide by 19, and not by 20, to correct for the small number.

The following table gives the means and variances of the average of 20 plots for the eight races as follows:

TABLE II

	Mean weight per plot, grammes	Variance of the average of 20 plots
145/46	318·7	94·7
Early Archer	306·5	138·9
7*A*	304·6	80·7
145	300·7	94·9
English Archer	297·8	128·8
Plumage	295·2	150·8
Irish Archer, No. 5	276·5	81·7
Biffen	270·8	142·0
Average	296·4	114·1

CORRECTION FOR POSITION*

There is a great disadvantage in correcting any figures for position, inasmuch as it savours of cooking, and besides the corrected figures do not represent anything real. It is better to arrange in the first place so that no correction is needed.

In the present case the "vertical" arrangement is satisfactory, but as to right and left it is not so. English Archer averages 0·2 rows to the left of 145, 0·4 to the left of 145/46 and so on, 1·4 rows to the left of Biffen. As the average value per plot of a row is about 3·3 grammes higher than that of the row on its left, it might be thought right to make the following corrections:

145/46	318·7 + 1·0 = 319·7
Early *A*	306·5 − 0·3 = 306·2
7*A*	304·6 − 1·7 = 302·9
145	300·7 + 1·7 = 302·4
English *A*	297·8 + 2·3 = 300·1
Plumage	295·2 + 0·3 = 295·5
Irish *A*, No. 5 ...	276·5 − 1·0 = 275·5
Biffen	270·8 − 2·3 = 268·5

* For an elaborate method of Correction for Inequality of Soil, see Pearl, "A method of correcting for soil heterogeneity in variety tests", *J. Agric. Res.* v, p. 1039.

In this paper Dr Pearl has corrected yield on the analogy of a contingency table. The method, which is probably as good a way as any of correcting for position, seems to me to be open to serious objections. A blot on the paper is the publishing of a "probable error" calculated from four cases without either correcting for the very small number or calling attention to the fact that they are appreciably too low.

The error of a comparison would no doubt be reduced very slightly as it generally is by any operation of this kind.

In any case the order is not altered, and I do not think the correction is worth making; the proper course would have been to reverse the order of the plots half way through so as to compensate for a possible tendency to improve from one end of the experimental area to the other.

Variance in Table II

With the small numbers in question the variance figures do not differ significantly, but incidentally there is no indication that the hybrids are more variable in yield than the pure lines.

In order to get a clear idea of what these figures mean, let us suppose that a standard error of 1 % is desired, say 3 grammes, a variance of 9. That would require an area 114·1/9, or 12·7 times as large as the present 20 plots.

If now the plots had been randomly placed, the variance of a comparison between two of the races would have been approximately 228, and about 25 times as much ground as was used would have been required to reduce the standard error of a comparison to 1 %.

In order to give a general idea of the nature of the variability, chiefly due to soil, which has to be regarded as error when we consider the yield of varieties, Diagram II has been prepared in which each 20 grammes of yield above 100 grammes below the average yield of the variety is represented by a diagonal line drawn across the square representing the plot. It will be noticed that the shading grows heavier towards the right of the diagram, and that while it is by no means regular, the correlation between the shading of neighbouring plots is obvious to the eye.

The arrangement of the different races in a chessboard is of course designed to take advantage of this correlation by comparing always neighbouring plots as in the following example which concerns the first pair of races in the table.

Beginning at the left hand of Diagram I, 145/46 is in the middle of the first vertical line, and Early Archer at the top—the former being indicated by the letter C, and the latter by E. The yield of the first is 265·6, and of the second, 230·1. That gives a positive difference of 35·5. The next appearance is in the third line, again a positive difference, this time of 44·4. In the third occurrence the 145/46 is in the fourth line, and the Early Archer in the fifth line, and the difference this time is negative and 37·4, and so on.

The variance of the average of the 20 differences thus obtained is 124·0, very much less than the 233·6, which is the sum of the variances of the averages of the two races.

Now, if there were only two races in the chessboard it would be comparatively

straightforward—the standard deviation would be found from the variance, and Sheppard's tables (or preferably with such small numbers, "Student's") would be used to judge the significance of the mean difference. In point of fact, however, the two races do not stand alone, and the question arises whether it would not be better to take the average variance of all the 28 differences between all the possible pairs of eight races.

Of course, it is not likely that all our races would have the same variance, but with our small numbers such differences as there may be are almost certainly swamped by the error of random sampling, which, as pointed out above, will account for the observed values. From that point of view then it is better to average.

Again, all the comparisons are not of equal value: Irish Archer No. 5 is always found exactly on the right of English Archer, while Plumage is either three squares above English Archer or two below and one row to the right, and as will be shown later, there are indications that this is enough to affect the variance. Still it is not a very big thing, and the advantages of using a single figure far outweigh the slight loss of accuracy. I have calculated the 28 variances and they range from 44·1 (English Archer–Irish Archer No. 5) to 192·9 (Early Archer–Plumage), with a mean of 107·9. This is slightly lower than the 114·1, the average variance of the races. In other words, we have gained by chessboarding to the extent that we are as accurate as if we had devoted twice the area to plots randomly arranged.

The calculation of these 28 variances is tedious, but fortunately there is a short cut which gives an identical result.

In the following proof capital subscripts indicate variance directly measurable, which is taken as the mean value of such variance, while small subscripts indicate variance deducible from the observations.

If we suppose the total variance σ_l^2 of mn plots (i.e. n groups of one of each of m races subject to the error of random sampling) to be divided into three parts:

(i) that due to the m races if measured without error: σ_r^2;

(ii) that due to the position of the n groups of m races from left to right of the diagram (in this case 20 groups of eight) also measured without error: σ_g^2;

(iii) the casual error, which is the only part subject to random sampling: σ_e^2; these three parts may be assumed to be independent so that

$$\sigma_l^2 = \sigma_r^2 + \sigma_g^2 + \sigma_e^2;$$

also the variance of the means of the races as we measure them is

$$\sigma_R^2 = \sigma_r^2 + \frac{\sigma_e^2}{n} - \frac{\sigma_e^2}{mn},$$

the last term being due to the fact that we have only mn cases to give us the mean.

Similarly, the variance of the means of the groups as we measure them is

$$\sigma_G^2 = \sigma_g^2 + \frac{\sigma_e^2}{m} - \frac{\sigma_e^2}{mn},$$

and the total variance as we measure it is

$$\sigma_T^2 = \sigma_l^2 - \frac{\sigma_e^2}{mn};$$

from which eliminating σ_l^2, σ_r^2, σ_g^2, we find

$$\sigma_e^2 = \frac{mn(\sigma_T^2 - \sigma_R^2 - \sigma_G^2)}{(m-1)(n-1)},$$

and consequently

$$\frac{2\sigma_e^2}{n},$$

which is the variance of a comparison between n groups of two races, is

$$\frac{2m(\sigma_T^2 - \sigma_R^2 - \sigma_G^2)}{(m-1)(n-1)}.*$$

* In my first attempt to obtain this formula, I overlooked the $-\sigma_e^2/mn$ in the three equations for σ_R^2, σ_G^2 and σ_T^2. It was only after receiving a letter from Mr R. A. Fisher, who had independently arrived at the correct formula, that I found my mistake. Mr Fisher sent me two proofs, one of which was purely algebraical, proving in his notation the identity

$$\frac{2}{m(m-1)} S \frac{1}{2(n-1)} \left\{ \sum_1^n (X_{pq} - X_{p'q})^2 - n(X_p - X_{p'})^2 \right\}$$

$$= \frac{\sum_1^n \sum_1^m (X - \bar{X})^2 - m \sum_1^n (X_q - \bar{X})^2 - n \sum_1^m (\bar{X}_p - \bar{X})^2}{(m-1)(n-1)};$$

and the other, which he himself prefers, I append below:

"Let there be n trials indicated by suffices $1 \ldots, q \ldots, n$ of each of m varieties similarly indicated by suffices $1 \ldots, p \ldots, m$.

Recognizing that not only differences of variety but differences in the conditions of the trials may have affected the yields, we may obtain an estimate of what the variability would be if the conditions of any one trial could be replicated in a number of experiments with the same variety, provided the following simple assumptions hold good. The yield obtained in any experiment is the sum of three quantities, one depending only on the variety; a second, depending only on the 'trial'; and a third, which may be regarded as the 'experimental error' varying independently of variety and trial in a normal distribution about zero with a standard deviation which it is desired to estimate.

To obtain such an estimate we may fit the system of yields X_{pq} with a system of values $A_p + B_q$, choosing the latter so that

$$S(X_{pq} - A_p - B_q)^2 \qquad \ldots\ldots(1)$$

is a minimum. Any one of the $m+n$ quantities A_p, B_q may be assigned an arbitrary value, and the remaining $m+n-1$ are then determinate: the observed values may therefore differ from those fitted in $(m-1)(n-1)$ degrees of freedom, and the corresponding estimate of the standard deviation ascribable to experimental error will be found by dividing the minimum value of (1) by $(m-1)(n-1)$. Evidently (1) will be a minimum if

$$A_p + B_q = \bar{X}_p + \bar{X}_q - \bar{X},$$

To obtain the variance by this formula is a comparatively simple operation. In this case owing to the fact that I grouped the 160 observations in 10-gramme groups I got 109·3 by the short cut instead of 107·9, but it really should give an identical value.

Taking the square root we get a standard deviation of 10·4 grammes or thereabout for the standard error of a comparison, i.e. a probable error of about 2·4 %. This is probably as near as it is worth while going in any one season, for the experiment must be repeated several times to sample the weather properly, and cage area is too valuable to expend more than is absolutely necessary on a single experiment.

Before leaving this subject of chessboards, I would like to show in rather more detail that even with such small plots as these, slight differences in the arrangement within the group tend to increase the variance over that due to the ideal juxtaposition.

I have, therefore (see Diagram III, p. 104), separated the various kinds of comparisons and averaged the variance, in each case as that of the average of 20 differences.

The figures are not of course worth a great deal, but there is a marked tendency for the comparisons between the more distant plots to be the less accurate.

For purposes of illustration, I have correlated the distances with the variance for the 13 positions by the Spearman method, and get $\rho = +0{\cdot}41$.

where $\overline{X}_p$ is the mean of the values obtained with variety p, $\overline{X}_q$ the mean of the values obtained with trial q, and $\overline{X}$ is the general mean.

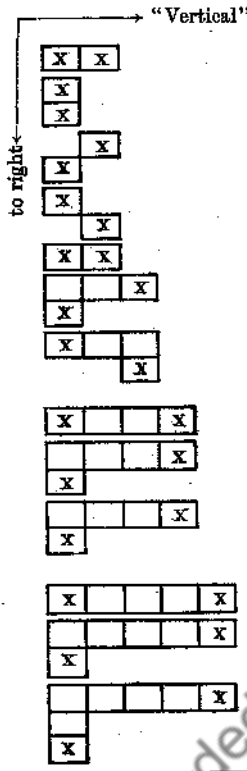
The actual evaluation is most conveniently carried out in the following form of the analysis of variance:

Variance	Degrees of freedom	Sum of squares
(*a*) Due to variety	$m-1$	$n\sum_{1}^{m}(\overline{X}_p-\overline{X})^2$
(*b*) Due to trial	$n-1$	$m\sum_{1}^{n}(\overline{X}_q-\overline{X})^2$
(*c*) Random variation	$(m-1)(n-1)$	$\sum_{1}^{m}\sum_{1}^{n}(X_{pq}-\overline{X}_p-\overline{X}_q+\overline{X})^2$
(*d*) Total	$mn-1$	$\sum_{1}^{m}\sum_{1}^{n}(X-\overline{X})^2$

The sum of squares in line (*c*) being calculated by subtracting the values of lines (*a*) and (*b*) from the total. If either variety or 'trial' were without significant effect on the yield, the corresponding mean square would not differ significantly from that of line (*c*). To test the significance of such a difference we may use the fact that the estimates of variance in (*a*), (*b*) and (*c*) are all independent, and when m and n are fairly large the natural logarithm of the mean square has standard deviation $\sqrt{(2/n_1)}$, where n_1 is the number of degrees of freedom. In comparing two such independent estimates of the mean square, we therefore obtain the difference of their natural logarithms, and assign to it a standard deviation

$$\sqrt{\left(\frac{2}{n_1}+\frac{2}{n_2}\right)}.$$"

DIAGRAM III

Position	Distance between centres	Variance	Number of differences
	4′	113·9	112
	4′	66·5	60
	5·7′	92·9	64
	5·7′	125·5	32
	8′	91·2	72
	9′	101·2	60
	9′	167·7	12
	12′	146·5	40
	12·7′	114·6	48
	14·4′	146·5	8
	16′	132·0	16
	16·5′	131·1	28
	17·9′	94·5	8

THE HALF-DRILL STRIP METHOD*

The small-scale work with which I have just dealt affords a means of picking out good varieties which can be tested in field trials. The whole eight varieties were tested on about $\frac{1}{17}$ acre, sowing about a quarter of a pound of seed for each race. We now proceed to the most accurate method yet devised for field trials by which two varieties are compared on a total area of 5200 square yards, just over an acre, with, in the case which I shall give you, a standard error of 0·63 %. Of course, it will not necessarily be as low as this always.

The field is cultivated as usual up to the time of sowing, except that particular care is taken to clean the ground of weeds.

* For a full account *vide* "Trials of new varieties of cereals", by E. S. Beaven, *J. Minist. Agric.* XXIX, Nos. 4 and 5 (1922).

When sowing, the seed box of the drill is divided into two across the middle, and the middle coulter put out of action. The seed of the two varieties is put in the seed box, one on each side of the division. Thus when sowing a drill strip, one half (i.e. 6 or 7 rows) is sown with one variety and the other half with the other. On turning the drill at the end, the next strip is sown so that two half strips of the same variety are next each other, but care is taken to leave an interval between the two drill strips exactly equal to the gap in the middle of each drill strip between the two varieties. It requires careful steering but it can be done.

When the experimental field is sown, we get first a single half-drill strip of one variety, then two of the other, then two of the first and so forth, ending with a half-drill strip of the first. This ending is necessary in order to discount any fertility slope from one end to the other of the field. The space outside the experimental area should be sown all round with a similar grain, as the outside is naturally abnormal and is more liable to attacks from all kinds of enemies.

At harvest the outside row of each half-drill strip next to the other variety is pulled up by hand and discarded to eliminate the "border" effect, and also to facilitate the use of the ordinary reaping machine. If the two varieties do not ripen together one must be cut by hand when ripe, but if there is so little difference that both can be cut on the same day the reaping machine can be used on both. In either case each half-drill strip is cut in such a way that the produce of each $\frac{1}{500}$-acre can be tied up in two sheaves separately. In Beaven's case ten such $\frac{1}{500}$-acre plots went to each half-drill strip.

These sheaves can be weighed on the field, and so we can get the total produce of the field in plots of $\frac{1}{500}$-acre and can compare each $\frac{1}{500}$-acre with an adjoining one of the other variety.

Two things are to be noted at this point: (1) That without a very great deal of trouble the plots cannot be threshed out separately, but, fortunately, it has so far always been found where the matter has been put to the test that the variability of the yield of grain expressed as a percentage of the grain is less than the variability of the total yield expressed as percentage of total yield. In the Mercer and Hall experiment, the standard errors were 11·6 and 11·9 %, and Beaven's experience has been similar. Thus the figure which we obtain for the standard error is likely to be in excess of the truth. (2) From a practical point of view it is easier to work with a few half-drill strips than a larger number of short ones, but if we depend on the weights of a few drill strips, there is considerable uncertainty about the standard error of the result. It was hoped that by determining the standard error of the difference between adjacent $\frac{1}{500}$-acre plots, we could deduce the standard error of the average of n such differences by the formula $\sigma_a = \sigma/\sqrt{n}$, so that it would be immaterial whether the drill strips were long and few or short and many, as long as altogether there were n pairs of adjacent subplots. Indeed up to the time when I came to write this section, it was believed that this could be done.

Beaven showed me his figures before publication, and I did not at the time observe that the formula cannot be used without further investigation, nor, so far as I am aware, has anyone else drawn attention to it. Nevertheless, I think it will be clear from the general considerations which have been advanced throughout the paper that there is a danger that the differences between corresponding constituent plots of a drill strip, even when they are as narrow as these, will tend to be correlated, and the formula $\sigma_a = \sigma/\sqrt{n}$, which required independence of the individuals which are to be averaged, cannot be used without correction.* That this is so in the particular case which we are considering is made highly probable from the fact that the variance, expressed in terms of the percentage of the total weight of C, of the difference between the total produce from A and C is 0·664 of the total weight of C when calculated from the 27 differences between adjacent half-drill strips, while it is only 0·301 when calculated from the 270 differences between adjacent subplots. The two figures should be the same within the error of random sampling, but differ probably by more than twice their standard deviation.

The results of the 1921 Trial are shown in Tables III and IV, which are taken, with his kind permission and that of the Ministry of Agriculture, from the Supplement to Beaven's paper, and give the weights of the sheaves, on the individual half-drill strips, and on 243 of the 270 "plots", which go to make up the half-drill strips respectively.

It will be seen that by taking the differences between adjoining half-drill strips (or plots) a large part of the error is, as usual, eliminated.

* A fallacy arising from a similar neglect of correlation has come under my notice in some American work, but there the absurdity is more easily demonstrated. In the *J. Amer. Soc. Agron.* IX, p. 138, A. G. McCall proposed that in order to save the trouble of harvesting and weighing $\frac{1}{10}$th acre plots a number of square yards should be cut out and harvested separately, the square yards being taken systematically through the $\frac{1}{10}$th acre plot, and the yield per acre calculated from these square yards. So far, so good, by taking enough square yards the slight loss of accuracy may perhaps be made up by gain in time or feasibility of operating. But in 1919, Arny and Steinmetz, *J. Amer. Soc. Agron.* XI, pp. 88, 89, applying this method, compared the error of the yield calculated from a few square yards cut from each of a number of $\frac{1}{10}$th acre plots with that calculated from the $\frac{1}{10}$th acre plots themselves. They found it substantially greater, but, say they, by increasing the number of square yards cut from each $\frac{1}{10}$th acre plot to n, we can decrease the error in the proportion $1/\sqrt{n}$, and so we can actually determine the yield more accurately by weighing up 10 or 20 square yards than by weighing up the whole half acre. It is rather surprising that they did not realize that there are 484 square yards in $\frac{1}{10}$th acre, so that by taking 484 square yards they would be likely to be more accurate than if they took any lesser number and *a fortiori* tremendously more accurate than they would be if they took the same 484 square yards and called it $\frac{1}{10}$th acre! Of course their formula also should be $\sigma\sqrt{\left(\frac{1+(n-1)r}{n}\right)}$, where r is the correlation between the yields on the square yards composing $\frac{1}{10}$th acre plots, and not $\sigma/\sqrt{n}$.

The same fallacy has been used to extol the "rod row" method of determining yield, i.e. the method of cutting along the drill a row one rod in length to represent the yield of the plot from which it is cut.

TABLE III. *Warminster Field Variety Trial*, 1921. *Half-Drill Strip Weights, comparing: Two races of barley, viz. "C" and "A". Area of each half-drill strip = 100 sq. yd. Total area = 1700 sq. yd. = 0·56 acre for each race. Showing total weight of sheaves on each half-drill strip*

Half-drill strip		Weight of sheaves on half-drill strip		Difference between "A" and "C"	Half-drill strip		Weight of sheaves on half-drill strip		Difference between "A" and "C"
Number		lb.		lb.	Number		lb.		lb.
"C"	"A"	"C"	"A"	"A"–"C"	"C"	"A"	"C"	"A"	"A"–"C"
1	2	165·4	164·6	– 0·8	29	30	160·9	160·2	– 0·7
4	3	159·5	173·4	+13·9	32	31	153·2	164·3	+11·1
5	6	169·3	169·3	—	33	34	144·9	154·3	+ 9·4
8	7	179·8	174·9	– 4·9	36	35	147·7	158·6	+10·9
9	10	172·5	177·6	+ 5·1	37	38	142·4	143·0	+ 0·6
12	11	170·7	182·9	+12·2	40	39	138·7	143·6	+ 4·9
13	14	173·3	167·5	– 5·8	41	42	131·1	143·2	+12·1
16	15	166·1	178·5	+12·4	44	43	141·6	145·3	+ 3·7
17	18	174·5	170·3	– 4·2	45	46	145·0	150·1	+ 5·1
20	19	163·3	176·0	+12·7	48	47	155·4	154·0	– 1·4
21	22	166·0	159·1	– 6·9	49	50	151·1	149·3	– 1·8
24	23	161·2	168·7	+ 7·5	52	51	145·6	149·7	+ 4·1
25	26	169·3	164·2	– 5·1	53	54	146·3	158·5	+12·2
28	27	156·5	167·0	+10·5	Total		4251·3	4368·1	—
					Average per cent.		157·5 100	161·8 102·7	+ 4·3 + 2·7

TABLE IV. *Showing weights of each "plot" (two consecutive sheaves), 9 plots (= 18 sheaves) on each half-drill strip. Total number of plots on 27 half-drill strips = 27 by 9 = 243 plots of each race (excluding end sheaves which are left out of account in this case in computing "probable error")*

No. of half-drill strips "C"	"A"	"Plot" weights 2 sheaves lb. "C"	"A"	A − C lb.
1	2	17·0	17·6	+ 0·6
		15·8	16·9	+ 1·1
		14·2	15·0	+ 0·8
		14·2	13·8	− 0·4
		16·2	15·8	− 0·4
		17·3	17·0	− 0·3
		18·0	17·1	− 0·9
		17·8	17·2	− 0·6
		16·6	17·8	+ 1·2
		147·1	148·2	
		18·3*	16·4*	
		165·4	164·6	− 0·8
9	10	17·6	16·7	− 0·9
		17·2	16·9	− 0·3
		16·0	15·8	− 0·2
		17·0	15·9	− 1·1
		17·3	17·4	+ 0·1
		16·0	17·2	+ 1·2
		20·1	19·0	− 1·1
		18·6	18·5	− 0·1
		20·5	19·0	− 1·5
		160·3	156·4	
		12·2*	21·2*	
		172·5	177·6	+ 5·1

No. of half-drill strips "A"	"C"	"Plot" weights 2 sheaves lb. "A"	"C"	A − C lb.
3	4	17·4	16·3	+ 1·1
		16·6	16·0	+ 0·6
		16·5	16·1	+ 0·4
		15·4	13·2	+ 2·2
		15·6	13·7	+ 1·9
		18·9	16·7	+ 2·2
		18·6	15·6	+ 3·0
		17·8	16·9	+ 0·9
		18·6	17·3	+ 1·3
		155·4	141·8	
		18·0*	17·7*	
		173·4	159·5	+13·9
11	12	18·0	16·1	+ 1·9
		16·1	15·6	+ 0·5
		15·9	14·2	+ 1·7
		16·9	15·6	+ 1·3
		18·1	16·5	+ 1·6
		17·6	15·7	+ 1·9
		19·4	18·5	+ 0·9
		18·2	17·9	+ 0·3
		19·2	19·9	− 0·7
		159·4	150·0	
		23·5*	20·7*	
		182·9	170·7	+12·2

No. of half-drill strips "C"	"A"	"Plot" weights 2 sheaves lb. "C"	"A"	A − C lb.
5	6	16·4	16·2	− 0·2
		17·7	16·8	− 0·9
		13·0	14·8	+ 1·8
		14·5	14·4	− 0·1
		15·2	15·8	+ 0·6
		17·7	17·3	− 0·4
		18·7	17·9	− 0·8
		17·7	16·5	− 1·2
		19·1	18·9	− 0·2
		150·0	148·6	
		19·3*	20·7*	
		169·3	169·3	—
13	14	16·4	15·7	− 0·9
		16·5	17·1	+ 0·6
		14·4	13·7	− 0·9
		14·4	15·0	+ 0·6
		17·2	15·8	− 1·4
		16·5	17·3	+ 0·8
		19·6	16·8	− 2·8
		18·5	18·6	+ 0·1
		21·8	17·3	− 4·5
		155·3	147·3	
		18·0*	20·2*	
		173·3	167·5	− 5·8

No. of half-drill strips "A"	"C"	"Plot" weights 2 sheaves lb. "A"	"C"	A − C lb.
7	8	18·1	17·9	+ 0·2
		16·8	17·1	− 0·3
		14·8	15·3	− 0·5
		16·0	15·9	+ 0·1
		16·3	16·6	− 0·3
		17·0	18·3	− 1·3
		18·5	19·7	− 1·2
		17·9	18·6	− 0·7
		19·3	19·7	− 0·4
		154·7	159·1	
		20·2*	20·7*	
		174·9	179·8	− 4·9
15	16	18·5	16·4	+ 2·1
		15·9	16·1	− 0·2
		15·9	16·0	− 0·1
		16·0	16·0	—
		15·6	16·1	− 0·5
		17·4	16·1	+ 1·3
		19·6	16·7	+ 2·9
		20·3	17·4	+ 2·9
		20·1	17·1	+ 3·0
		159·3	147·9	
		19·2*	18·2*	
		178·5	166·1	+12·4

17	18	18·1	17·1	− 1·0	19	20	16·2	15·7	+ 0·5	21	22	16·3	15·9	− 0·4	23	24	14·4	15·2	− 0·8
		16·6	17·6	+ 1·0			17·7	15·2	+ 2·5			16·5	14·6	− 1·9			16·5	14·7	+ 1·8
		17·1	16·1	− 1·0			16·8	15·4	+ 1·4			16·4	13·6	− 2·8			15·3	16·2	− 0·9
		16·8	16·4	− 0·4			16·6	14·1	+ 2·5			14·7	14·0	− 0·7			14·6	14·2	+ 0·4
		15·8	14·6	− 1·2			16·8	14·6	+ 2·2			13·3	13·5	+ 0·2			14·3	14·6	− 0·3
		16·0	15·3	− 0·7			16·4	17·3	− 0·9			17·1	15·4	− 1·7			16·9	15·3	+ 1·6
		18·1	18·1	—			19·1	17·7	+ 1·4			18·2	18·1	− 0·1			18·7	17·3	+ 1·4
		17·3	18·0	+ 0·7			18·0	16·3	+ 1·7			16·8	17·2	+ 0·4			18·2	16·6	+ 1·6
		18·5	17·3	− 1·2			19·1	16·4	+ 2·7			16·9	16·7	− 0·2			18·7	17·0	+ 1·7
		154·3	150·5				156·7	142·7				146·2	139·0				147·6	141·1	
		20·2*	19·8*				19·3*	20·6*				19·8*	20·1*				21·1*	20·1*	
		174·5	170·3	− 4·2			176·0	163·3	+12·7			166·0	159·1	−6·9			168·7	161·2	+ 7·5
25	26	15·5	14·1	− 1·4	27	28	14·7	14·6	+ 0·1	29	30	14·7	14·0	− 0·7	31	32	14·2	13·5	+ 0·7
		13·7	16·7	+ 3·0			15·6	15·3	+ 0·3			14·2	13·6	− 0·6			15·3	13·6	+ 1·7
		16·0	14·5	− 1·5			14·8	15·3	− 0·5			14·8	14·5	− 0·3			14·4	13·5	+ 0·9
		14·1	15·7	+ 1·6			14·2	13·7	+ 1·5			15·7	14·8	− 0·9			13·3	12·0	+ 1·3
		16·6	16·9	+ 0·3			15·4	15·8	− 0·4			16·9	15·9	− 1·0			17·1	17·3	− 0·2
		17·2	15·2	− 2·0			18·8	15·8	+ 3·0			16·4	16·5	+ 0·1			16·8	14·7	+ 2·1
		19·4	18·3	− 1·1			18·3	17·0	+ 1·3			17·0	16·8	− 0·2			18·9	17·3	+ 1·6
		18·0	17·1	− 0·9			19·2	15·7	+ 3·5			16·7	18·0	+ 1·3			19·0	15·9	+ 3·1
		18·6	16·8	− 1·8			17·5	16·6	+ 0·9			16·5	17·5	+ 1·0			17·2	17·8	− 0·6
		149·1	145·3				148·5	139·8				142·9	141·6				146·2	135·6	
		20·2*	18·9*				18·5*	16·7*				18·0*	18·6*				18·1*	17·6*	
		169·3	164·2	− 5·1			167·0	156·5	+10·5			160·9	160·2	− 0·7			164·3	153·2	+11·1
33	34	13·5	14·8	+ 1·3	35	36	15·0	13·1	+ 1·9	37	38	12·6	11·4	− 1·2	39	40	11·8	11·4	+ 0·4
		10·7	12·5	+ 1·8			13·7	11·3	+ 2·4			11·4	12·9	+ 1·5			11·9	11·1	+ 0·8
		12·8	12·7	− 0·1			12·6	13·5	− 0·9			11·9	11·9	—			12·3	12·9	− 0·6
		12·6	14·0	+ 1·4			14·3	14·5	− 0·2			15·4	14·3	− 1·1			12·8	13·6	− 0·8
		15·3	16·3	+ 1·0			16·4	16·8	− 0·4			15·1	16·7	+ 1·6			16·6	13·1	+ 3·5
		15·3	15·9	+ 0·6			17·4	12·6	+ 4·8			14·0	15·8	+ 1·8			14·2	15·3	− 1·1
		17·5	18·1	+ 0·6			15·3	17·4	− 2·1			15·1	13·8	− 1·3			16·8	16·0	+ 0·8
		15·9	17·5	+ 1·6			16·1	14·6	+ 1·5			13·8	14·7	+ 0·9			15·3	14·7	+ 0·6
		14·8	15·1	+ 0·3			17·9	16·3	+ 1·6			16·0	14·7	− 1·3			16·2	13·8	+ 2·4
		128·4	136·9				138·7	130·1				125·3	126·2				127·9	121·9	
		16·5*	17·4*				19·9*	17·6*				17·1*	16·8*				15·7*	16·8*	
		144·9	154·3	+ 9·4			158·6	147·7	+10·9			142·4	143·0	+ 0·6			143·6	138·7	+ 4·9

* These figures represent weights of the first and last sheaves on each half-drill strip added together, and are excluded in calculating the average weights and also in calculating the "probable error".

TABLE IV (*continued*)

No. of half-drill strips "C"	No. of half-drill strips "A"	"Plot" weights 2 sheaves lb. "C"	"Plot" weights 2 sheaves lb. "A"	A − C lb.
41	42	10·6	11·4	+ 0·8
		9·7	11·1	+ 1·4
		11·4	10·1	− 1·3
		11·9	10·7	− 1·2
		12·9	13·3	+ 0·4
		13·8	15·9	+ 2·1
		16·5	19·5	+ 3·0
		14·4	16·6	+ 2·2
		14·1	16·2	+ 2·1
		115·3	124·8	
		15·8*	18·4*	
		131·1	143·2	+12·1
49	50	13·6	13·8	+ 0·2
		12·5	13·0	+ 0·5
		13·6	13·8	+ 0·2
		14·5	13·0	− 1·5
		15·3	13·9	− 1·4
		15·9	15·6	− 0·3
		16·6	15·6	− 1·0
		16·8	17·1	+ 0·3
		16·4	16·0	− 0·4
		135·2	131·8	
		15·9*	17·5*	
		151·1	149·3	− 1·8

No. of half-drill strips "A"	No. of half-drill strips "C"	"Plot" weights 2 sheaves lb. "A"	"Plot" weights 2 sheaves lb. "C"	A − C lb.
43	44	11·1	11·8	− 0·7
		10·9	12·0	− 1·1
		12·3	11·0	+ 1·3
		12·0	13·6	− 1·6
		14·4	15·6	− 1·2
		15·5	15·0	+ 0·5
		17·3	16·2	+ 1·1
		15·7	15·8	− 0·1
		18·3	14·3	+ 4·0
		127·5	125·3	
		17·8*	16·3*	
		145·3	141·6	+ 3·7
51	52	14·2	13·5	+ 0·7
		12·8	11·5	+ 1·3
		14·3	13·7	+ 0·6
		13·5	14·2	− 0·7
		14·3	14·1	+ 0·2
		15·3	14·1	+ 1·2
		14·5	16·0	− 1·5
		16·7	15·7	+ 1·0
		16·4	15·6	+ 0·8
		132·0	128·4	
		17·7*	17·2*	
		149·7	145·6	+ 4·1

No. of half-drill strips "C"	No. of half-drill strips "A"	"Plot" weights 2 sheaves lb. "C"	"Plot" weights 2 sheaves lb. "A"	A − C lb.
45	46	11·6	13·2	+ 1·6
		11·7	12·5	+ 0·8
		10·7	12·2	+ 1·5
		13·9	14·5	+ 0·6
		15·8	15·6	+ 0·2
		14·8	16·2	+ 1·4
		16·8	16·1	− 0·7
		17·4	16·0	− 1·4
		16·8	16·2	− 0·6
		129·5	132·5	
		15·5*	17·6*	
		145·0	150·1	+ 5·1
53	54	13·8	13·1	− 0·7
		12·3	12·2	− 0·1
		11·8	14·8	+ 3·0
		13·9	13·6	− 0·3
		16·2	15·6	− 0·6
		15·1	16·5	+ 1·4
		15·5	16·0	+ 0·5
		15·4	19·0	+ 3·6
		17·2	16·5	− 0·7
		131·2	137·3	
		15·1*	21·2*	
		146·3	158·5	+12·2

No. of half-drill strips "A"	No. of half-drill strips "C"	"Plot" weights 2 sheaves lb. "A"	"Plot" weights 2 sheaves lb. "C"	A − C lb.
47	48	11·8	12·9	− 1·1
		13·0	13·4	− 0·4
		13·9	13·9	—
		16·7	14·9	+ 1·8
		15·6	16·3	− 0·7
		16·4	17·0	− 0·6
		15·6	17·0	− 1·4
		17·6	15·7	+ 1·9
		16·3	17·1	− 0·8
		136·9	138·2	
		17·1*	17·2*	
		154·0	155·4	− 1·4
Average weight per "plot" (end sheaves excluded)		15·56	15·88	+0·32
%		100·0	102·1	+2·10

* These figures represent weights of the first and last sheaves on each half-drill strip added together, and are excluded in calculating the average weights and also in calculating the "probable error".

Further, it is obvious that there is a general decrease in fertility as we go from drill strips with low numbers to drill strips with high numbers. It follows that the difference $A - C$ will tend to be greater when C follows A than when A follows C, and since this is always possible, experiments of this nature should always be planned so that there shall be an even number of differences, the series should begin and end with half-drill strips of the same variety: in this case we may simply leave out the last drill strip and finish at half-drill strip 52.

There is also a curious feature about these figures which can only be put down to some systematic error in technique; namely that when we compare together the adjacent half-drill strips of A, that with the higher number always yields higher, although the general fertility runs the other way, and the same is true with regard to C in eight cases out of thirteen.

Both these kinds of error (that due to the general fertility slope and that due to the different fertility of odd and even half-drill strips) are largely eliminated by Beaven's arrangement by which in alternate comparisons A follows C and C follows A, and this can be made evident by adopting as unit not the difference between adjacent half-drill strips but that between the sum of the two contiguous half-drill strips of A and the sum of the two half-drill strips of C which enclose them.

This may be described as a "sandwich", and it may be noted that just as there are subplots composing a half-drill strip, so there are "subsandwiches" which will also tend to eliminate the same errors as the "sandwiches".

The following table gives the differences $A - C$ for the thirteen "sandwiches" composed of half-drill strips 1 to 52:

TABLE V

Half-drill strip numbers	$A - C$	Half-drill strip numbers	$A - C$
1 to 4	+13·1	29 to 32	+10·4
5 „ 8	− 4·9	33 „ 36	+20·3
9 „ 12	+17·3	37 „ 40	+ 5·5
13 „ 16	+ 6·6	41 „ 44	+15·8
17 „ 20	+ 8·5	45 „ 48	+ 3·7
21 „ 24	+ 0·6	49 „ 52	+ 2·3
25 „ 28	+ 5·4		

The mean $A - C$ for sandwiches is +8·05 and the variance, making allowance for the pitifully small number, is 51·41. This leads to a variance of the difference between the total produce of A and of C expressed in terms of the total weight of C of 0·398, intermediate between the 0·664 calculated from the half-drill strip differences and the 0·301 calculated from the subplot differences.

It should be noted at this point that the "sandwich" is a perfectly legitimate device for eliminating errors common to both variants whose difference is to be measured, and that it is only by using it that we can get the true value of the error

of the comparison, whereas the subplot difference would really lead to a larger value than 0·301 if we had sufficient knowledge to be able to apply the true formula

$$\frac{\sigma^2(1+(n-1)r)}{n}.$$

A similar calculation based on the "subsandwiches", i.e. sandwiches one plot in depth, gives a value of the variance 0·248 corresponding to the 0·398 from the whole sandwiches. The standard deviation of these to some extent correlated figures is not easy to determine, but the difference between them must be of the order of once the standard deviation. This is not significant, but with our small numbers it is not inconsistent with the expected correlation between the "subsandwiches," composing a sandwich. Until a number of experiments have been carried out in several places and the results submitted to analysis, it would be wise to keep the number of drill strips as large as possible and economize in length in spite of the practical difficulties of doing so.

Since the variance calculated from the drill-strip sandwiches is subject to a large error of random sampling owing to the necessary paucity of numbers, it is well to calculate also from the "subsandwiches" and take the larger of the two in determining the standard error.

It is possible that some of my readers may devise some better method of utilizing the weights of the "subplots" than I have been able to do, and I commend the problem to them.

In the present case it is probably better with only thirteen sandwiches to take the standard error of a single sandwich and use "Student's" tables, when the probability that such a large positive difference should occur by chance is found to be 0·001. The difference is therefore quite significant. If, however, it is required to compare the standard error with other experiments, we can say that the most probable value is only 0·63 % on a total area of about 1 acre.

Other precautions, such as correction for moisture, etc., are taken as a matter of course.

Conclusions

The chief difficulty of comparing varieties consists in the fact that the differences to be measured are quite small compared with the variations due to soil and weather. While the latter is not within our control, the errors due to the soil may be reduced to reasonable proportions in any one of three ways:

(1) Large plots may be repeated many times. An instance is given of this when in the Irish 2-acre experimental plots a difference of 7 % in the value per acre was proved with a standard deviation of about 2 % in 51 trials, extending over six years.

Undertakings of this magnitude are hardly to be put in hand by any but Departments of State.

(2) Quite small plots of one square yard, surrounded by a border of the same variety as in the square yard, may be grown under a wire cage on a regular system, technically called a "chessboard". An instance of this is given when, in Beaven's No. 1 Yield Experiment of 1913, eight varieties were compared on a total area of about $\frac{1}{17}$th acre using about 5 oz. of seed of each variety, with a standard deviation of a comparison in a single year of about $3\frac{1}{2}\%$.

The large number of varieties which may be compared at once, and the small area which is required, make this an ideal method of testing new varieties. On the other hand, a wire cage is not a cornfield, and the varieties found to be best in the cage will always require further testing on the large scale. The method is, however, within the powers of anyone who can build a cage, and has the necessary skill and patience to conduct the experiments.

(3) By means of Beaven's "half-drill strip" method, two varieties may be compared on a total area of about one acre in one year with a standard deviation of a comparison of less than 1%. This combines the advantage of growing corn on the large scale with an accuracy almost as great as that of small-scale work; and is within the powers of anyone who can combine the necessary knowledge and patience with the control of skilled agricultural labour.

It is shown that methods (2) and (3) depend for their accuracy on the fact that the nearer two plots of ground are situated, the more highly are the yields correlated, so that we are able to increase the effect of the last term of the equation

$$\sigma^2_{A-B} = \sigma^2_A + \sigma^2_B - 2r_{AB}\sigma_A\sigma_B$$

(where A and B are the varieties to be compared) by placing the plots to be compared with one another as near together as possible.

A formula, due to Mr R. A. Fisher, is given for calculating the error of a comparison in a "chessboard" experiment, which may perhaps be found useful elsewhere.

Finally I have to thank Dr Beaven both for allowing me to use his experimental material and for much invaluable assistance in the preparation of the paper.

Addendum

Since writing the above I have had the advantage of witnessing the harvesting of Dr Beaven's 1923 experiment and of discussing the whole question with him very thoroughly.

He thinks it probable that the whole or a part of the correlation between the yields of the "plots" which together formed a drill strip in the 1921 experiment may have been due to slight differences in area consequent on irregular steering of the seed drill, such as would have been caused by the horses pulling unequally.

Measurements which we made on the stubble of the similar 1923 experiment showed not only that such inaccuracies occur, but also that they can favour one of the varieties.

It is, however, a fairly easy matter after harvest to measure the total width from the outside drill of one half-drill strip to the outside drill of the same variety. This measurement includes the space between the drill strips, which is variable owing to the difficulty of steering and is now made in practice across each drill strip in several places.

It is thus possible to estimate accurately the total area occupied by each variety and to make the necessary correction to the total yields.

As, however, it would hardly be possible to correct the individual drill strips or "plots" which are used for the purpose of calculating the error, that calculated error will be in excess of the truth.

In Dr Beaven's opinion the operation of taking differences has for all practical purposes eliminated the correlation due to the position of the "plots", and in view of the other causes of variation in the differences, numerous and diverse as they are, he still considers it legitimate to treat the differences between the "plots" as if they were random, and to use the formula $\sigma/\sqrt{n}$ in calculating the error of his mean difference. I feel, however, that a single operation of this nature is hardly likely to eliminate all the correlation and that there is need for further inquiry: if as the result of a number of experiments it is found that the error of the mean difference calculated from the weights of the half-drill strips is not significantly greater than that calculated from the "plots", then the latter undoubtedly provide the more accurate data for the calculation of that error, and it will be a matter of indifference whether the drill strips be few and long or short and many.

Meanwhile they should be made as numerous as is consistent with the successful carrying out of the various agricultural operations, which are of course made infinitely more difficult and tedious by the necessity of turning horses and machines at the end of each short length.

But whether we use few long or many short strips is not a question of the first importance: in either case the method is without doubt the best that has hitherto been devised for large-scale experiments.

Later Note

The following note relating to a paragraph on pp. 98–9 above was included by Student in the next volume of *Biometrika* (xvi (1924), p. 411):

I wish to apologize to the readers of *Biometrika* for having allowed it to appear that I was the author of the term "Variance" defined as the square of the Standard Deviation. It was first used by Mr R. A. Fisher in 1918 in a paper entitled "The Correlation between relatives on the Supposition of the Mendelian Inheritance", *Trans. Roy. Soc. Edin.* lii, 2, pp. 399–433; and he has published many papers since in which the word has been used.

12

NEW TABLES FOR TESTING THE SIGNIFICANCE OF OBSERVATIONS

[*Metron*, V (1925), p. 105]

In *Biometrika*, VI, pp. 1–25 [2] it was suggested if $z = \bar{x}/s$, where x is the distance of the mean of a sample of n from the true mean of a normally distributed population, and s is the standard deviation of the same sample, i.e.

$$\sqrt{\left(\frac{S(x-\bar{x})^2}{n}\right)},$$

then the frequency of z is given by the frequency curve

$$y = \frac{\Gamma(\frac{1}{2}n)}{\Gamma\{\frac{1}{2}(n-1)\}}(1+z^2)^{-\frac{1}{2}n},$$

and that consequently the integral

$$p = \frac{1}{2} + \frac{\Gamma(\frac{1}{2}n)}{\Gamma\{\frac{1}{2}(n-1)\}}\int_0^z (1+z^2)^{-\frac{1}{2}n}\,dz$$

gives the probability that the mean of a sample of n drawn from a normally distributed population, measured in terms of the standard deviation of the sample, shall exceed the value z.

Tables were constructed for values of n from 4 to 10 [2, p. 29], and subsequently, in *Biometrika*, XI, p. 416 [8, pp. 62–3], from 2 to 30.

It has since been shown, as in the preceding paper by Mr Fisher (*Metron*, V (1925), pp. 90–104), that the suggestion was in fact justified, and that the integral has a much wider application than was originally supposed.

The tables hitherto published suffer however from two defects: (i) that as n increases the z scale becomes very coarse, and (ii) that except in the case for which it was designed, n, the number in the sample, is not the best number under which to enter the table, but $n-1$, the number of degrees of freedom.

The present tables have, therefore, at Mr Fisher's suggestion been constructed with argument $t = z\sqrt{n}$, where n is now one less than the number in the sample, which we may call n'. They correspond to Sheppard's table, when that is used to test the significance of the mean of a large number of observations.

Table I extends from $t = 0$ to $t = 6$, at intervals of 0·1, from $n = 1$ to $n = 20$, inclusive; in each column in which values of more than 0·99995 occur, the first of these is written 1·0000, and further values are not given.

Table II gives values beyond $t = 6$, to six places of decimals, from which values accurate to four places of decimals can be calculated by proportional interpolation. The intervals are, therefore, unequal, and increase as t becomes larger. In this table no values are given under $n = 1$ and $n = 2$, as these can be easily calculated from the ordinary trigonometrical tables by the formulae

$$n = 1, \quad p = \frac{1}{2} + \frac{\theta}{\pi} \text{ (where } \tan\theta = t),$$

$$n = 2, \quad p = \tfrac{1}{2} + \tfrac{1}{2}\sin\theta \left(\text{where } \tan\theta = \frac{t}{\sqrt{2}}\right).$$

Table III gives coefficients for calculating the difference between the value for $n = \infty$, i.e. Sheppard's table, and that for n, where p is arrived at by the formula

$$p = p_{\infty} - \frac{C_1}{n} - \frac{C_2}{n^2} - \frac{C_3}{n^3} - \frac{C_4}{n^4}.$$

This gives values of p, estimated to be accurate to 0·000005, when n is greater than 20, and, in fact, at 20 and 24 the following differences were found:

alues of t	0·5	1·5	2·0	2·5	3·0	3·5	4·0	4·5	5·0	5·5	6·0	6·5	7·0
ifferences $=20$	0	0	0	0	+23	+17	−33	−46	−30	+41	+19	+9	+4
ifferences $=24$	—	—	—	—	+ 8	+ 9	−14	−18	− 4	+20	+ 7	—	—

The above differences are in the seventh place of decimals, and are between values of p given by the approximation and those derived from the cosine formula using seven-place tables. Mr Fisher's note (*Metron*, v (1925), pp. 109–12) explains the basis on which the coefficients were calculated.

The methods of calculating and checking the tables were as follows:

1. Values of p for

$$t = 0{\cdot}5,\ 1{\cdot}0,\ 1{\cdot}5,\ 2{\cdot}0,\ 2{\cdot}5,\ 3{\cdot}0,\ 3{\cdot}5,\ 4{\cdot}0,\ 4{\cdot}5,\ 5{\cdot}0,\ 5{\cdot}5,\ 6{\cdot}0,\ 6{\cdot}5,\ 7{\cdot}0,$$

and

$$n = 1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 10,\ 12,\ 15,\ 20,\ 24,$$

were calculated from the cosine formula (*Biometrika*, VI, p. 10 [2, p. 21]), using seven-figure tables; these values, though they are the sum of $\frac{1}{2}n$ terms, appear to be accurate within about 0·0000003, and were checked by recalculation. They were also compared with the values obtained by the use of Table III, which both served as a further check, and also to show within what limits Table III could be used for constructional purposes.

2. From the values thus calculated under $n = 6$, 8, 12, 24, together with $n = \infty$ (i.e. Sheppard), the remaining frame values under $n = 7$, 9, 11, 13, 14, 15, 16, 17, 18, 19 were interpolated by coefficients calculated by Mr Fisher for asymptotic interpolation. These were checked by recalculation and cross-

differencing, i.e. by comparing the difference $p_n - p_{n-1}$ with $p_{n+1} - p_n$ for the same values of t, and any doubtful values were recalculated by the cosine formula, as also were any values in which the fourth place of decimals was doubtful, i.e. whenever the fifth place of decimals was 4 or 5.

3. Having thus obtained a frame, this was filled in to five places of decimals in three ways: (*a*) by interpolation, using where necessary both four- and six-point central interpolation. It was found that over the greater part of the table the true values lie between the four-point and the six-point interpolation, but for high values of t it was usually sufficient to use four-point, six-point being required only to locate doubtful values. (*b*) For very low values of $n (= 1, 2$ and $3)$ the frame was found not to be sufficiently close with low values of t, and alternate values had to be calculated from the cosine formula, the remaining odd values being interpolated by four- or six-point central interpolation. (*c*) As n increased it was found possible to make more and more use of Table III, beginning at $n = 4$ with values of t less than 1 and ending at $n = 20$ with the whole table. These values were recalculated as a check. Second differences were then taken down the columns and any doubtful figures checked from the cosine formula; as before, this was done whenever the fourth figure was in any doubt.

Finally, the whole table was cross-differenced, and a very large number of values were recalculated from the cosine formula. Very few alterations were, however, found to be necessary.

Table II was altogether calculated from the cosine formula; as it is designed to give an accuracy of four figures by proportional interpolation, it was possible to increase the interval between the t entries as t increases.

Table III was calculated from Mr Fisher's formulae, and I have to thank Miss W. A. Mackenzie, M.Sc., of the Rothamsted Statistical Laboratory for kindly checking this part of the work.

TABLE I. *The Probability Integral of t*

t	$n=1$ $n'=2$	2 3	3 4	4 5	5 6	6 7	7 8	8 9	9 10	10 11
0·0	·500,0	·500,0	·500,0	·500,0	·500,0	·500,0	·500,0	·500,0	·500,0	·500,0
0·1	·531,7	·535,3	·536,7	·537,4	·537,9	·538,2	·538,4	·538,6	·538,7	·538,8
0·2	·562,8	·570,0	·572,9	·574,4	·575,3	·576,0	·576,4	·576,8	·577,0	·517,3
0·3	·592,8	·603,8	·608,1	·610,4	·611,9	·612,9	·613,6	·614,1	·614,5	·614,8
0·4	·621,1	·636,1	·642,0	·645,2	·647,2	·648,5	·649,5	·650,2	·650,8	·651,2
0·5	·647,6	·666,7	·674,3	·678,3	·680,9	·682,6	·683,8	·684,7	·685,5	·686,1
0·6	·672,0	·695,3	·704,6	·709,6	·712,7	·714,8	·716,3	·717,4	·718,3	·719,1
0·7	·694,4	·721,8	·732,8	·738,7	·742,4	·744,9	·746,7	·748,1	·749,2	·750,1
0·8	·714,8	·746,2	·758,9	·765,7	·770,0	·772,9	·775,0	·776,6	·777,8	·778,8
0·9	·733,3	·768,4	·782,8	·790,5	·795,3	·798,6	·801,0	·802,8	·804,2	·805,4
1·0	·750,0	·788,7	·804,5	·813,0	·818,4	·822,0	·824,7	·826,7	·828,3	·829,6
1·1	·765,1	·807,0	·824,2	·833,5	·839,3	·843,3	·846,1	·848,3	·850,1	·851,4
1·2	·778,9	·823,5	·841,9	·851,8	·858,1	·862,3	·865,4	·867,8	·869,6	·871,1
1·3	·791,3	·838,4	·857,8	·868,3	·874,8	·879,3	·882,6	·885,1	·887,0	·888,6
1·4	·802,6	·851,8	·872,0	·882,9	·889,8	·894,5	·897,9	·900,5	·902,5	·904,1
1·5	·812,8	·863,8	·884,7	·896,0	·903,0	·907,9	·911,4	·914,0	·916,1	·917,7
1·6	·822,2	·874,6	·896,0	·907,6	·914,8	·919,6	·923,2	·925,9	·928,0	·929,7
1·7	·830,7	·884,4	·906,2	·917,8	·925,1	·930,0	·933,5	·936,2	·938,3	·940,0
1·8	·838,6	·893,2	·915,2	·926,9	·934,1	·939,0	·942,6	·945,2	·947,3	·949,0
1·9	·845,8	·901,1	·923,2	·934,9	·942,1	·946,9	·950,4	·953,0	·955,1	·956,7
2·0	·852,4	·908,2	·930,3	·941,9	·949,0	·953,8	·957,2	·959,7	·961,7	·963,3
2·1	·858,5	·914,7	·936,7	·948,2	·955,1	·959,8	·963,1	·965,5	·967,4	·969,0
2·2	·864,2	·920,6	·942,4	·953,7	·960,5	·964,9	·968,1	·970,5	·972,3	·973,8
2·3	·869,5	·925,9	·947,5	·958,5	·965,1	·969,4	·972,5	·974,8	·976,5	·977,9
2·4	·874,3	·930,8	·952,1	·962,8	·969,2	·973,4	·976,3	·978,4	·980,1	·981,3
2·5	·878,9	·935,2	·956,1	·966,6	·972,8	·976,7	·979,5	·981,5	·983,1	·984,3
2·6	·883,1	·939,2	·959,8	·970,0	·975,9	·979,7	·982,3	·984,2	·985,6	·986,8
2·7	·887,1	·942,9	·963,1	·973,0	·978,6	·982,2	·984,7	·986,5	·987,8	·988,8
2·8	·890,8	·946,3	·966,1	·975,6	·981,0	·984,4	·986,7	·988,4	·989,6	·990,6
2·9	·894,3	·949,4	·968,7	·977,9	·983,1	·986,3	·988,5	·990,1	·991,2	·992,1
3·0	·897,6	·952,3	·971,2	·980,0	·985,0	·988,0	·990,0	·991,5	·992,5	·993,3
3·1	·900,7	·954,9	·973,4	·981,9	·986,6	·989,4	·991,3	·992,7	·993,6	·994,4
3·2	·903,6	·957,3	·975,3	·983,5	·988,0	·990,7	·992,5	·993,7	·994,6	·995,3
3·3	·906,3	·959,6	·977,1	·985,0	·989,3	·991,8	·993,4	·994,6	·995,4	·996,0
3·4	·908,9	·961,7	·978,8	·986,4	·990,4	·992,8	·994,3	·995,3	·996,1	·996,6
3·5	·911,4	·963,6	·980,3	·987,6	·991,4	·993,6	·995,0	·996,0	·996,6	·997,1
3·6	·913,8	·965,4	·981,6	·988,6	·992,2	·994,3	·995,6	·996,5	·997,1	·997,6
3·7	·916,0	·967,0	·982,9	·989,6	·993,0	·995,0	·996,2	·997,0	·997,5	·997,9
3·8	·918,1	·968,6	·984,0	·990,4	·993,7	·995,5	·996,6	·997,4	·997,9	·998,3
3·9	·920,1	·970,1	·985,0	·991,2	·994,3	·996,0	·997,1	·997,7	·998,2	·998,5
4·0	·922,0	·971,4	·986,0	·991,9	·994,8	·996,4	·997,4	·998,0	·998,4	·998,7
4·1	·923,9	·972,7	·986,9	·992,6	·995,3	·996,8	·997,7	·998,3	·998,7	·998,9
4·2	·925,6	·973,9	·987,7	·993,2	·995,8	·997,2	·998,0	·998,5	·998,8	·999,1
4·3	·927,3	·975,0	·988,4	·993,7	·996,1	·997,5	·998,2	·998,7	·999,0	·999,2
4·4	·928,9	·976,0	·989,1	·994,2	·996,5	·997,7	·998,4	·998,9	·999,1	·999,3
4·5	·930,4	·977,0	·989,8	·994,6	·996,8	·997,9	·998,6	·999,0	·999,3	·999,4
4·6	·931,9	·977,9	·990,3	·995,0	·997,1	·998,2	·998,8	·999,1	·999,4	·999,5
4·7	·933,3	·978,8	·990,9	·995,3	·997,3	·998,3	·998,9	·999,2	·999,4	·999,6
4·8	·934,6	·979,6	·991,4	·995,7	·997,6	·998,5	·999,0	·999,3	·999,5	·999,6
4·9	·935,9	·980,4	·991,9	·996,0	·997,8	·998,6	·999,1	·999,4	·999,6	·999,7
5·0	·937,2	·981,1	·992,3	·996,3	·997,9	·998,8	·999,2	·999,5	·999,6	·999,7
5·1	·938,4	·981,8	·992,7	·996,5	·998,1	·998,9	·999,3	·999,5	·999,7	·999,8
5·2	·939,5	·982,5	·993,1	·996,7	·998,3	·999,0	·999,4	·999,6	·999,7	·999,8
5·3	·940,6	·983,1	·993,4	·997,0	·998,4	·999,1	·999,4	·999,6	·999,8	·999,8
5·4	·941,7	·983,7	·993,8	·997,2	·998,5	·999,2	·999,5	·999,7	·999,8	·999,8
5·5	·942,8	·984,2	·994,1	·997,3	·998,6	·999,2	·999,5	·999,7	·999,8	·999,9
5·6	·943,8	·984,8	·994,4	·997,5	·998,7	·999,3	·999,6	·999,7	·999,8	·999,9
5·7	·944,7	·985,3	·994,6	·997,7	·998,8	·999,4	·999,6	·999,8	·999,9	·999,9
5·8	·945,7	·985,8	·994,9	·997,8	·998,9	·999,4	·999,7	·999,8	·999,9	·999,9
5·9	·946,6	·986,2	·995,1	·997,9	·999,0	·999,5	·999,7	·999,8	·999,9	·999,9
6·0	·947,4	·986,7	·995,4	·998,1	·999,1	·999,5	·999,7	·999,8	·999,9	·999,9

$n=11$ $n'=12$	12 13	13 14	14 15	15 16	16 17	17 18	18 19	19 20	20 21	∞	t
·500,0	·500,0	·500,0	·500,0	·500,0	·500,0	·500,0	·500,0	·500,0	·500,0	·500,000,0	0·0
·538,9	·539,0	·539,1	·539,1	·539,2	·539,2	·539,2	·539,3	·539,3	·539,3	·539,827,8	0·1
·577,4	·577,6	·577,7	·577,8	·577,9	·578,0	·578,1	·578,1	·578,2	·578,2	·579,259,7	0·2
·615,1	·615,3	·615,5	·615,7	·615,9	·616,0	·616,1	·616,2	·616,3	·616,4	·617,911,4	0·3
·651,6	·651,9	·652,2	·652,4	·652,6	·652,8	·652,9	·653,1	·653,2	·653,3	·655,421,7	0·4
·686,5	·686,9	·687,3	·687,6	·687,8	·688,1	·688,3	·688,4	·688,6	·688,7	·691,462,5	0·5
·719,7	·720,2	·720,6	·721,0	·721,3	·721,5	·721,8	·722,0	·722,2	·722,4	·725,746,9	0·6
·750,8	·751,4	·751,9	·752,3	·752,7	·753,0	·753,3	·753,6	·753,8	·754,0	·758,036,3	0·7
·779,7	·780,4	·781,0	·781,5	·781,9	·782,3	·782,6	·782,9	·783,2	·783,4	·788,144,6	0·8
·806,3	·807,1	·807,8	·808,3	·808,8	·809,3	·809,7	·810,0	·810,3	·810,6	·815,939,9	0·9
·830,6	·831,5	·832,2	·832,9	·833,4	·833,9	·834,3	·834,7	·835,1	·835,4	·841,344,7	1·0
·852,6	·853,5	·854,4	·855,1	·855,7	·856,2	·856,7	·857,1	·857,5	·857,8	·864,333,9	1·1
·872,3	·873,4	·874,2	·875,0	·875,6	·876,2	·876,7	·877,2	·877,6	·877,9	·884,930,3	1·2
·889,9	·891,0	·891,9	·892,7	·893,4	·894,0	·894,5	·895,0	·895,4	·895,8	·903,199,5	1·3
·905,5	·906,6	·907,5	·908,4	·909,1	·909,7	·910,3	·910,7	·911,2	·911,6	·919,243,3	1·4
·919,1	·920,3	·921,2	·922,1	·922,8	·923,5	·924,0	·924,5	·925,0	·925,4	·933,192,8	1·5
·931,0	·932,2	·933,2	·934,0	·934,8	·935,4	·936,0	·936,5	·937,0	·937,4	·945,200,7	1·6
·941,4	·942,6	·943,5	·944,4	·945,1	·945,8	·946,3	·946,8	·947,3	·947,7	·955,434,6	1·7
·950,3	·951,5	·952,5	·953,3	·954,0	·954,6	·955,2	·955,7	·956,1	·956,5	·964,069,7	1·8
·958,0	·959,1	·960,1	·960,9	·961,6	·962,2	·962,7	·963,2	·963,6	·964,0	·971,283,4	1·9
·964,6	·965,7	·966,6	·967,4	·968,0	·968,6	·969,1	·969,6	·970,0	·970,4	·977,249,9	2·0
·970,2	·971,2	·972,1	·972,8	·973,5	·974,0	·974,5	·975,0	·975,3	·975,7	·982,135,6	2·1
·975,0	·975,9	·976,8	·977,4	·978,1	·978,6	·979,0	·979,4	·979,8	·980,1	·986,096,6	2·2
·979,0	·979,9	·980,7	·981,3	·981,9	·982,4	·982,8	·983,2	·983,5	·983,8	·989,275,9	2·3
·982,4	·983,2	·984,0	·984,6	·985,1	·985,5	·985,9	·986,3	·986,6	·986,9	·991,802,5	2·4
·985,2	·986,0	·986,7	·987,3	·987,7	·988,2	·988,5	·988,8	·989,1	·989,4	·993,790,3	2·5
·987,7	·988,4	·989,0	·989,5	·990,0	·990,3	·990,7	·991,0	·991,2	·991,4	·995,338,8	2·6
·989,7	·990,3	·990,9	·991,4	·991,8	·992,1	·992,4	·992,7	·992,9	·993,1	·996,533,0	2·7
·991,4	·992,0	·992,5	·992,9	·993,3	·993,6	·993,8	·994,1	·994,3	·994,5	·997,444,9	2·8
·992,8	·993,3	·993,8	·994.2	·994,5	·994,5	·995,0	·995,2	·995,4	·995,6	·998,134,2	2·9
·994,0	·994,5	·994,9	·995,2	·995,5	·995,8	·996,0	·996,2	·996,3	·996,5	·998,650,1	3·0
·994,9	·995,4	·995,8	·996,1	·996,3	·996,6	·996,7	·996,9	·997,1	·997,2	·999,032,4	3·1
·995,8	·996,2	·996,5	·996,8	·997,0	·997,2	·997,4	·997,5	·997,6	·997,8	·999,312,9	3·2
·996,5	·996,8	·997,1	·997,4	·997,6	·997,7	·997,9	·998,0	·998,1	·998,2	·999,516,6	3·3
·997,0	·997,4	·997,6	·997,8	·998,0	·998,2	·998,3	·998,4	·998,5	·998,6	·999,663,1	3·4
·997,5	·997,8	·998,0	·998,2	·998,4	·998,5	·998,6	·998,7	·998,8	·998,9	·999,767,4	3·5
·997,9	·998,2	·998,4	·998,6	·998,7	·998,8	·998,9	·999,0	·999,0	·999,1	·999,840,9	3·6
·998,2	·998,5	·998,7	·998,8	·998,9	·999,0	·999,1	·999,2	·999,2	·999,3	·999,892,2	3·7
·998,5	·998,7	·998,9	·999,0	·999,1	·999,2	·999,3	·999,3	·999,4	·999,4	·999,927,7	3·8
·998,8	·998,9	·999,1	·999,2	·999,3	·999,4	·999,4	·999,5	·999,5	·999,6	·999,951,9	3·9
·999,0	·999,1	999,2	999,3	999,4	·999,5	999,5	999,6	·999,6	·999,6	·999,968,3	4·0
·999,1	·999,3	·999,4	·999,5	·999,5	·999,6	·999,6	·999,7	·999,7	·999,7	·999,979,3	4·1
·999,3	·999,4	·999,5	·999,6	·999,6	·999,7	·999,7	·999,7	·999,8	·999,8	·999,986,7	4·2
·999,4	·999,5	·999,6	·999,6	·999,7	·999,7	·999,8	·999,8	·999,8	·999,8	·999,991,5	4·3
·999,5	·999,6	·999,6	·999,7	·999,7	·999,8	·999,8	·999,8	·999,8	·999,9	·999,994,6	4·4
·999,5	·999,6	·999,7	·999,8	·999,8	·999,8	·999,8	·999,9	·999,9	·999,9	·999,996,6	4·5
·999,6	·999,7	·999,8	·999,8	·999,8	·999,9	·999,9	·999,9	·999,9	·999,9	·999,997,9	4·6
·999,7	·999,7	·999,8	·999,8	·999,9	·999,9	·999,9	·999,9	·999,9	·999,9	·999,998,7	4·7
·999,7	·999,8	·999,8	·999,9	·999,9	·999,9	·999,9	·999,9	·999,9	·999,9	·999,999,2	4·8
·999,8	·999,8	·999,9	·999,9	·999,9	·999,9	·999,9	·999,9	1·000,0	1·000,0	·999,999,5	4·9
·999,8	·999,8	·999,9	·999,9	·999,9	·999,9	·999,9	1·000,0			·999,999,7	5·0
·999,8	·999,9	·999,9	·999,9	·999,9	·999,9	1·000,0				·999,999,8	5·1
·999,9	·999,9	·999,9	·999,9	·999,9	1·000,0					·999,999,9	5·2
·999,9	·999,9	·999,9	·999,9	1·000,0						·999,999,9	5·3
·999,9	·999,9	·999,9	1·000,0							1·000,000,0	5·4
·999,9	·999,9	·999,9									5·5
·999,9	·999,9	1·000,0									5·6
·999,9	1·000,0										5·7
·999,9											5·8
·999,9											5·9
1·000,0											6·0

NOTE. $n=n'-1$ is the number of degrees of freedom used in the estimate of variance.

TABLE II. *Supplementary table for high values of t*

t	$n=3$ $n'=4$	$n=4$ $n'=5$	$n=5$ $n'=6$	$n=6$ $n'=7$	$n=7$ $n'=8$	$n=8$ $n'=9$	$n=9$ $n'=10$	$n=10$ $n'=11$
6·0	·995,364	·998,059	·999,077	·999,518	·999,729	·999,838	·999,899	·999,934
6·5	·996,303	·998,555	·999,357	·999,684	·999,833	·999,906		·999,966
7·0	·997,007	·998,904	·999,542	·999,788	·999,894	·999,944	·999,968	
7·5	·997,544	·999,155				·999,965		
8·0	·997,962	·999,338	·999,754	·999,898	·999,954			
8·5	·998,290							
9·0	·998,552	·999,578	·999,859	·999,947				
10·0	·998,936	·999,719	·999,915	·999,971				
11·0	·999,196							
12·0	·999,377	·999,862	·999,965					
14·0	·999,605	·999,924						
16·0	·999,735	·999,955						
20·0	·999,863							
24·0	·999,921							
28·0	·999,950							

Linear interpolation between adjacent entries will give four figure accuracy.

TABLE III

t	C_1	C_2	C_3	C_4	t	C_1	C_2	C_3	C_4
0·1	·010,023,1	−·001,261	−·001,55	+ ·000,4	2·1	·026,862,2	+·207,289	+·028,49	−1·368,5
0·2	·020,334,2	−·002,616	−·003,08	·000,8	3·2	·021,437,7	·193,351	·144,23	−1·610,8
0·3	·031,178,5	−·004,177	−·004,53	·001,3	3·3	·016,897,1	·176,859	·254,58	−1·837,2
0·4	·042,719,3	−·006,087	−·005,86	·001,7	3·4	·013,155,2	·158,774	·353,27	−2·034,8
0·5	·055,010,2	−·008,509	−·006,97	·002,2	3·5	·010,117,7	·139,969	·435,35	−2·732,8
0·6	·067,977,8	−·011,595	−·007,72	·002,6	3·6	·007,687,9	·121,313	·497,49	−2·503,1
0·7	·081,420,2	−·015,432	−·007,96	·002,8	3·7	·005,772,2	·103,371	·538,13	−2·123,7
0·8	·095,018,8	−·019,991	−·007,74	·002,6	3·8	·004,282,3	·086,649	·557,33	−1·625,7
0·9	·108,363,2	−·025,066	−·006,51	·002,1	3·9	·003,139,7	·071,486	·556,67	−1·050,1
1·0	·120,985,4	−·030,246	−·005,04	·001,1	4·0	·002,275,1	·058,066	·538,80	− ·441,7
1·1	·132,399,7	−·034,907	−·003,76	·000,4	4·1	·001,629,5	·046,453	·507,08	+ ·155,5
1·2	·142,144,2	−·038,248	−·003,75	·001,1	4·2	·001,153,8	·036,613	·465,19	·702,6
1·3	·149,819,0	−·039,363	−·006,56	·005,8	4·3	·000,807,4	·028,438	·416,81	1·169,2
1·4	·155,117,7	−·037,344	−·014,10	·018,2	4·4	·000,558,6	·021,773	·365,31	1·535,3
1·5	·157,849,6	−·031,399	−·028,41	·043,0	4·5	·000,382,1	·016,436	·313,56	1·791,6
1·6	·157,951,2	−·020,971	−·051,29	·084,9	4·6	·000,258,4	·012,235	·263,86	1·938,7
1·7	·155,486,7	−·005,832	−·083,84	·161,2	4·7	·000,172,8	·008,984	·217,87	1·985,3
1·8	·150,637,0	+·013,846	−·126,09	·232,3	4·8	·000,114,3	·006,508	·176,63	1·946,2
1·9	·143,682,2	·037,483	−·176,64	·335,1	4·9	·000,074,6	·004,652	·140,70	1·839,4
2·0	·134,977,5	·064,114	−·232,56	·446,6	5·0	·000,048,3	·003,281	·110,17	1·683,9
2·1	·124,924,4	·092,473	−·289,45	·551,2	5·1	·000,030,9	·002,284	·084,85	1·498,6
2·2	·113,944,4	·121,099	−·341,85	·627,7	5·2	·000,019,5	·001,570	·064,29	1·299,5
2·3	·102,451,8	·148,472	−·383,79	·652,2	5·3	·000,012,2	·001,065	·047,96	1·100,4
2·4	·090,832,2	·173,147	−·409,50	·601,0	5·4	·000,007,6	·000,713	·035,27	·910,7
2·5	·079,425,1	·193,877	−·414,17	·455,3	5·5	·000,004,6	·000,472	·025,47	·379,9
2·6	·068,512,4	·209,710	−·394,59	+ ·205,6	5·6	·000,002,8	·000,308	·018,15	·506,5
2·7	·058,312,9	·220,052	−·349,67	− ·145,4	5·7	·000,001,7	·000,199	·012,74	·437,0
2·8	·048,980,8	·224,693	−·280,61	− ·580,9	5·8	·000,001,0	·000,127	·008,82	·349,5
2·9	·040,609,7	·223,749	−·190,82	−1·070,4	5·9	·000,000,6	·000,080	·006,02	·262,6
3·0	·033,238,9	·217,715	−·085,59	−1·572,7	6·0	·000,000,3	·000,050	·004,05	·193,9

13

MATHEMATICS AND AGRONOMY*

[*J. Amer. Soc. Agron.* XVIII (1926), p. 703]

THE nature of pure mathematics is such that the conclusions follow inevitably from the premises and may be said to be contained in them. Consequently, if in applying mathematics to affairs we reach absurd conclusions, we may be sure either that a blunder has been made or that in some essential point the data of the mathematical problem did not correspond to the facts.

For it must be remembered that mathematical analysis deals with abstractions and that commonly the abstractions chosen are very much more simple than the facts, either in order to secure a generalized result, or because the analysis would otherwise become too difficult.

Thus, even in the ordinary textbook problem we may have to deal with weightless ropes or frictionless pulleys, with basins which empty through the waste at a uniform speed regardless of the depth, or bricklayers who work at the same rate, however closely they may be crowded together.

GENERAL CONSIDERATIONS

It may be assumed then that if mathematical analysis applied to the interpretations of agronomic experiments has given absurd or inconsistent results, it is probably because the facts were not correctly represented by the abstractions with which the mathematics dealt. It may, therefore, be worth while to consider what limitations are imposed by the imperfect correspondence between the conditions of our experiments and the mathematical abstractions from which are constructed the tables which are used to interpret their results. It may also be possible to find means of designing experiments so that they may be interpreted with as little error as possible.

I shall begin by setting out, as far as may be, in non-mathematical language, the reasons which lead us to use certain tables in interpreting our experiments, and then examine the conditions under which we are justified in doing so. But however much it may be desired to avoid mathematical language, it is necessary to define a certain number of terms accurately, and in what follows the following words will be used in the sense given below:

1. *Variable.* A quantity that can present more than one numerical value, e.g. height, birth-rate, the yield of a plot.

* Personal contribution. Received for publication 26 March 1926. I should like to thank Prof. J. H. Parker and Dr H. Hunter, who kindly suggested that I should write this paper, Dr R. A. Fisher, "Mathetes", and several other friends who have helped to clear up the obscurities of the original manuscript.

2. *Variate.* An individual value of a variable, e.g. 5 ft. 10 in.; 19·63 per 1000; 19½ lb.

3. *Population.* All the individuals under discussion. It should be noted that all these individuals need not exist. We may be dealing with a population of all individuals which could have existed under certain conditions. A population may, and generally does, vary in more than one character. It is necessary to be quite sure exactly what is the population with which we are dealing, and to remember that our conclusions cannot necessarily be extended to other populations. If, for example, we have a series of plots from which we deduce that one variety of oats will give a higher yield than a second, and all the experiments were carried out in an exceptionally dry summer, our population would be "comparisons of yields in an exceptionally dry summer", and without further work it is obviously impossible to draw general conclusions applicable to comparisons of yields in all summers.

4. *Sample.* A number of individuals selected to represent a population.

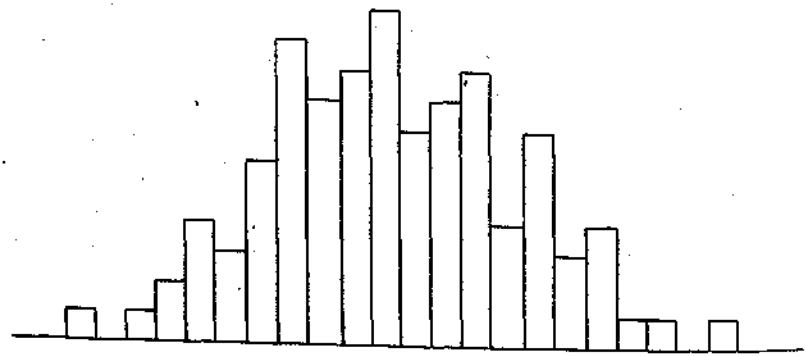
5. *Random Sample.* A sample selected in such a way that any individual in the population has an equal chance of being included in the sample. It is always difficult, and often impossible, to discover anything definite about a population from a sample which is not random.

6. *Frequency.* The number of variates occurring between any limiting values of a variable.

Clearly, it is possible to give a geometrical representation of the frequencies occurring in any sample by setting out the scale of its variable horizontally along a base line and measuring vertically the frequency on each unit of the scale. This gives the familiar figure consisting of columns of equal width but of unequal height which is known as a histogram.

If the sample is small we may have such a figure as this:

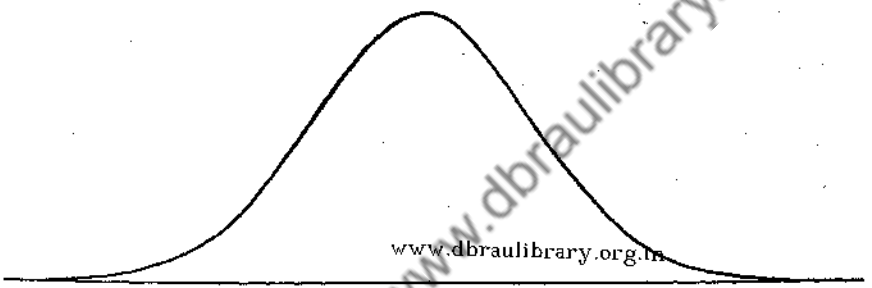
Each square represents one variate; but if the sample is larger, it will take a more continuous form, such as this:



As the numbers increase, the tendency is for the outline to become more and more regular.

It should be noted that, in practical affairs, the columns of the histogram must necessarily have a definite width, that of the unit of measurement; this must be at least the width of the smallest measurable unit of the variable and is usually much wider. Thus, although weight can be measured in fractions of a gramme, the yields of plots are given to the nearest pound or cental.

Nevertheless, we can imagine that if the unit of measurement were to be decreased indefinitely and the sample increased without limit so as to become an infinitely large population, the histogram with its irregular steps would be replaced by the smooth continuous curve which is known as a frequency curve. These frequency curves are necessarily abstractions; nobody ever reached one



by plotting out the frequency of a sample, but it is often comparatively easy by following the instructions of mathematicians who have studied the subject to find the equation and draw the graph of a frequency curve which describes a population such that a given sample might have been drawn from it by random selection. While frequency curves are of many types, the only one to which attention need here be drawn is that discovered by Gauss and La Place, and known variously by their names and as the "Probability Curve" or "Normal Curve of Error". This curve was reached by supposing that the error of an observation is the sum total of an infinite number of infinitely small components each of which may be either positive or negative, and it purports to give the frequency with which errors of any given magnitude occur. The following properties are of interest:

1. It is symmetrical about a middle vertical line—the mean.

2. The curve is completely determined if we know the total frequency which it represents, i.e. the area between the curve and the base line, the mean, and either (*a*) the average of the squares of the distances of the errors from the mean, the mean square of error—called by R. A. Fisher the Variance—or (*b*) the average distance of the errors from the mean—the mean error.*

* In view of the fact that some American writers have stated that it does not much matter whether the probable error be calculated from the mean square (Bessel's formula) or the

3. The square root of the mean square of error is called the "Standard Error" or "the Standard Deviation", S.D.

4. 0·6745 times the standard deviation is called the "Probable Error", and is such that *in this special type of curve* one-half of the errors lie within a distance of once the probable error on either side of the mean. It should be noted that apart from this normal curve of error, the probable error has no exact meaning.

5. Tables giving the area of the curve lying between any given error, x, and either the mean or $-\infty$ have been constructed. In these "x" is measured either in terms of the standard deviation or of the "Modulus" c (c = the S.D. $\times\sqrt{2}$) and the area as the fraction of the total area of the curve.

Since an unknown observation may fall with equal probability in any equal area of the curve, these tables can be used to calculate the odds against an observation falling beyond any required distance from the mean.

6. Many naturally occurring populations may be described very closely by a normal curve of frequency, and can then be determined by the total frequency, the mean, and the S.D.

7. Although many populations exist which cannot be described by this curve, the samples which we are able to obtain in agronomic work are generally too small for us to be sure that the population they represent is not normal.

8. Even in the case of samples drawn from a population admittedly not normal, the *means* of such samples belong to a population (of means) which becomes more and more nearly normal the larger the samples.

It is, therefore, usual to assume, and *in the case of large samples* the assumption can be made without appreciable error, that the published tables of the normal curve can be used to calculate the odds against the mean of the sample differing by more than any required amount from the mean of the population.

It should here be remarked that in order to be able to use the tables in this way there must be a unit of measurement of the variation (standard deviation, probable error, or modulus of error), and there are two ways in which this can be arrived at.

The first way is that used by astronomers, routine analysts, and such people as can repeat observations many times in a standard manner. Working in this way, they can find a value of the S.D. from some hundreds of determinations of the same quantity, and they can then use this figure for smaller numbers of determinations in subsequent experiments.

The second way is more usual. It is to calculate the S.D. of the sample and use this value instead of the S.D. of the population. This has the advantage that

mean error (Peter's formula), it may be as well to state categorically that it does matter. R. A. Fisher (in *Monthly Notices, Roy. Astron. Soc.* June 1920) has shown that the latter method is equivalent to wasting 12 % of the observations, since 100 cases treated by the first give as accurate a measure of the probable error as 114 cases treated by the second.

at all events there can be no possibility of using the S.D. of the wrong population, which, conceivably, might otherwise happen. But, on the other hand, very few series of experiments are sufficiently long to allow of an exact estimate of the S.D. being made.

For the S.D. determined from a sample is just as much subject to error as is the mean, and consequently, if x is to be measured in terms of the S.D. of the sample, the uncertainty of the conclusion is necessarily increased. Further, it does not follow that the frequency curve of means of samples when measured in this new unit, which is different for each sample, will any longer be found to approximate to the "normal" curve. In fact, it has been shown *not* to do so for small samples, and Student's tables* have been constructed to meet the particular case of small samples drawn from a population which is itself normally distributed.

This survey of the foundations on which the application of probability to affairs are based has doubtless seemed long and, I fear, tedious; yet even so, it cannot be regarded as more than the merest sketch, and I shall be fortunate if it is even considered accurate by those entitled to an opinion on the subject. Nevertheless, we are now in a position to judge how far it is appropriate to use the two sets of tables, i.e. those of the normal curve, typified by Sheppard's, and Student's, in the interpretation of agronomic work.

Application to Agronomy

It may be assumed that the object of all agronomic experiments is to find out whether some change of practice is likely to benefit farmers who follow it. The change is commonly of manure or of seed, but sometimes of method of operation. In order to simplify matters, it is proposed in the first place to deal with a change typified by the replacement of one variety of cereal seed by another.

Taking this simple case, the following points must be borne in mind when using the tables in order to judge of the significance of conclusions:

1. The population to which the conclusions are to be applied is one of yields of cereals grown in fields on the large scale. That being so, the population of which the experiments are to be a sample must not differ in any essential point from this, and in particular must be coextensive with the possible large-scale population.

Thus, if it is desired to estimate the result of replacing variety A by variety B over an area part of which is affected by drought and part not, the experiments must be spread over land subject to both sets of conditions, and even then it is best to regard them as belonging to two separate populations.

Similarly, in a variable climate (and where does the climate not vary?), the experiments must be carried over a series of years to correspond to that population of large-scale practice which is spread over the future.

* [See pp. 29, 62–3 and 118–20 above. Ed.]

Again, there is a disproportionate amount of border in any reasonable size of experimental plot. This border must either be in contact with another variety or with ground unoccupied by crop. In either case the yield of the border strip is liable to be different from that of the interior, so that if the results are to be applied to the large-scale population of which the border forms a negligible fraction, it must be rejected.

Lastly, as far as may be, large-scale methods of agriculture should be used. Granted that it is often not possible, there is a danger that results may not be applicable to the farmer's case every time this principle is departed from, and every result obtained by small-scale methods should be rigorously checked on the large scale before making recommendations to the farmer.

2. Generally speaking, but not necessarily, the population of large-scale yields with which we are concerned is a population of "differences", i.e. some such question as the following is asked: "By how much may we expect the yield of variety B to exceed that of variety A if they were sown alternatively on the same soil in the same season?"

That being so, it is clear that the observed differences will not represent the true differences even in the sample plots as two crops cannot occupy the same place at the same time. Observed differences will miss the mark not only because the experimental soil and the weather experienced by the experiment may not be random samples of the soil and weather to be explored, but also because the actual plots laid out for the two varieties will usually differ in fertility. This is one of the largest sources of errors in field experiments.

Nevertheless, we are still dealing with a sample of differences and it is clearly advantageous in this simple case to do all calculations in terms of differences.*

This is not to say that percentages should never be used; that is another method of substituting one figure for two which has its uses, but percentages should be used with the greatest care, they are fertile mothers of fallacy.

3. In using either of the tables we assume that the experimental results are a sample drawn from a population distributed normally. This is doubtless very often nearly true, but the limited number of experiments usually prevents us from being sure of it. What, then, is the extent of the uncertainty arising from

* Note that the formula connecting the S.D. of a difference with those of its components is $\sigma_{A-B}^2 = \sigma_A^2 + \sigma_B^2 - 2r_{AB}\sigma_A\sigma_B$, where σ_{A-B} is the S.D. of $A-B$ and so on, and r_{AB} is the correlation between A and B. Only if $r_{AB} = 0$ does this degenerate into what I may call the astronomer's formula:

$$\sigma_{A-B}^2 = \sigma_A^2 + \sigma_B^2.$$

In any well-planned experiment r_{AB} is high, and there is considerable advantage in calculating the odds according to the correct formula. By considering the *differences* at once, we avoid all this difficulty of correcting for correlation.

In some American work the taking of differences seems to be considered the essential point of what they are kind enough to call "Student's Method", but this old artifice must at least date back to Noah, who doubtless had occasion to estimate the comparative appetites of his male and female passengers.

this cause? The answer is, that if *we have enough data* no appreciable error is introduced, since even if the population is not normal the distribution of the means of large samples is very nearly so, but with very few repetitions we have to fall back on the general experience that such frequencies as those of yields are generally not badly represented by the normal curve, and hope for the best. Fortunately, the approach to normality of the distribution of the means of samples is very rapid, and appreciable errors are not likely to arise from *this* assumption, if we are dealing with the mean of more than a dozen repetitions.

4. Even supposing that an assumption of normality is justifiable, Student's tables must be used in calculating from *small* samples the probability that the results could have occurred by chance. To use the other method is definitely wrong, especially as it gives too high an estimate of the reliability of the results.

5. That being so, the only object in calculating the probable error in such cases is to compare with other experiments. Even for this purpose it is necessary with small samples to divide by $n-1$ and not by n to reach the mean square. But indeed "probable errors" derived from only two or three cases are so subject to chance that it is somewhat doubtful whether any useful purpose is served by calculating them. For example, if 10 were the value of the "probable error" of a population and values were to be found from samples of two or three, only 49 % of the values would lie between 5 and 15 in the case of samples of two and but 68 % in the case of samples of three.

The use of $n-1$ as divisor is also necessary in calculating a standard deviation from a large number of small samples of size n, which H. K. Hayes* proposes to call the "Deviation from the Mean Method".

It is necessary to remember that the correct formula to use is

$$\sigma = \sqrt{\left(\frac{S(d^2) \times n}{N(n-1)}\right)},$$

where d is the deviation from the mean of the sample and N is the total number of deviations. When n is quite small this correction makes an appreciable difference.

6. Frequency curves are reached by assuming an infinitely large population and an infinitely small unit of measurement, and there is no trouble in *imagining* an infinitely large population though we have only to deal with the finite sample before us. But the unit of measurement must be the same for both, and, therefore, not only not infinitely small but as large as is convenient or customary. This is another of the discrepancies between the facts and the mathematics which does not matter very much *as long as the samples are large*, but may make a good deal of difference when they are small. With small samples the unit of measurement should be quite small compared with the difference which is being measured.

* "Control of soil heterogeneity and use of the probable error concept in plant breeding studies": *Minn. Agric. Expt. Sta. Tech. Bul.* 30 (1925).

For example, Student's table has been made to give absurd results by supposing that all the values happened by chance to coincide, when the odds became infinite. The probability that results should have the same value is not negligible when the unit of measurement is large but becomes vanishingly small as the unit of measurement is decreased, until, in the limit, the infinite odds only occur infinitely seldom. Nevertheless, when the repetitions are few and very high odds are obtained by the use of Student's tables, it is well to consider whether the result is not due to a value of the S.D. having occurred which is much smaller than usual, and if this seems likely, to discount the apparent certainty accordingly. The tables are calculated to give the odds correctly if *all* the available information is contained in the sample. If additional information is available, such as that the S.D. of similar experiments is usually larger, we are quite entitled to draw attention to it, even though it may not be possible to introduce it into the calculations. In fact, tables can only be an aid to, and not a substitute for, common sense.

7. The experiments must be capable of being considered to be a *random* sample of the population to which the conclusions are to be applied. Neglect of this rule has led to the estimate of the value of statistics which is expressed in the crescendo "lies, damned lies, statistics".

Well-conducted experiments can often be supposed to give results which are random samples of the population of possible differences between the yields of plots sown with varieties *A* and *B* which could be grown on the experimental area under climatic conditions similar to those of the season in which the experiments were carried out, but it must be confessed that in some cases it is only by courtesy that experiments can be considered to be a random sample of *any* population. In such cases the greatest care must be exercised in drawing conclusions.

Nevertheless, we need not go as far as S. C. Salmon,* who says: "It is with this source of error (soil heterogeneity) that Student's method may entirely fail", and proceeds to illustrate this by a comparison of yields in a tillage experiment carried out on two plots over a period of ten years. With all respect, I do not think Salmon credits the user of the method with common sense. For he supposes that as the result of this comparison it will be concluded that the fact that one of the plots gave significantly higher yields than the other will be put down to the tillage treatment.

A moment's consideration shows, however, that the population from which the sample was drawn is a sample of differences of yield between *these two plots* in all possible seasons; whereas, considered as a sample of difference in yield due to tillage treatment in all soils similar to that experimented in, it is only *one* case from which, of course, *no definite conclusions can be drawn* either by Student's

* *J. Amer. Soc. Agron.* XVI (1924), pp. 717–21.

method or any other. If, however, ten repetitions had been made *with an arrangement of the plots which could be considered random*, the population sampled would have been that of "all similar soils", and the error introduced by soil heterogeneity would have been weighed and allowed for by the use of the tables.*

8. To sum up, *the experiments must be conducted in such a way that their results may be capable of being considered to be a random sample of the population to which the conclusions are to apply.* The unit of measurement must be small compared with the differences likely to be found, and the replications must be sufficient (*a*) to give significance to the mean difference, and (*b*) to give a sufficiently close estimate of the variability to enable us to measure that significance with accuracy.

And here it may be pointed out that in some cases, could we but know the variability accurately, very few experiments would be required to demonstrate significance. If, to take an extreme example, a difference of ten units is found between a single pair of experiments and it is known from other work that in this case the S.D. of a single difference is likely to be in the neighbourhood of two units, a considerable, though somewhat indefinite, degree of confidence could be reposed in the result. This leads me to suggest that a careful tabulation and examination of S.D.'s of experiments conducted at each station might be very valuable as showing within what limits the S.D. of a new experiment might be expected to lie, and what sort of weight might be given to a result which would otherwise lack significance owing to want of knowledge of the variability. Useful though this might be, it is clearly better to arrange the experiments so that we shall have sufficient replications to lead to significance without going beyond the experiment itself.

Elsewhere† I have drawn attention to Beaven's half-drill strip method of comparing two varieties of cereals—a method which seems to me to fulfil the necessary requirements when but two varieties are in question. Here I propose to deal shortly with R. A. Fisher's "Latin square" arrangement of experimental plots. This arrangement is calculated to reduce and allow for the error introduced by soil heterogeneity and is suitable for work on any scale from rod rows or small rectangular plots up to large-scale plots of all sizes, provided always that the borders of small plots are discarded or that there is room enough for large plots.

Fisher outlines the method of the Latin square on pp. 229 to 232 of his book

* As Hayes (*loc. cit.*) has also complained that when comparing different seasons' yields Student's method does not allow for soil heterogeneity, I should like to emphasize that it may be used to estimate the uncertainty due to the season or to the soil heterogeneity, or even to both, provided we are satisfied that the experiments may be considered to belong to a single population. To compare mere average yields in different seasons and then to complain that no account has been taken of soil heterogeneity is as if a man were to feed wheat into a mill and then complain that the resulting meal "had entirely failed to make oaten porridge".

† *Biometrika*, XV (1923), pp. 271 *et seq.* [11].

on *Statistical Methods for Research Workers* (Messrs Oliver and Boyd, Edinburgh), and bases it on the following principles:

1. If there are contributory sources of variation *which are all independent*, the variance of the whole will be the simple sum of the variance contributed by all the sources. As mentioned above, Fisher defines variance as the square of the standard deviation, or in the case of errors, as the mean square of error. We may therefore, for example, be able to analyse a total variance into (*a*) that part contributed by the varieties (of seed or culture) having different yields; (*b*) that part contributed by say an East to West heterogeneity of soil; (*c*) that part contributed by say a North to South heterogeneity of soil; and (*d*) a random effect of soil heterogeneity not included in (*b*) and (*c*), which are not random.

2. It is possible to arrange n plots of each of n different varieties in a square* so that each row and each column of the square contains one plot of each variety, but that otherwise the arrangement is "random". Having done so, we can estimate variances (*a*), (*b*) and (*c*), whence by subtracting their sum from the total variance of the n^2 plots, we can estimate variance (*d*), which is now the only one which affects the comparison between the varieties.

Fisher's justification of his method might perhaps be considered to come under the head of mathematics, which we have agreed to avoid, so assuming its correctness we may proceed to illuminate the subject by the consideration of a simple example.

Let us suppose that we are to test four varieties (*A*, *B*, *C* and *D*) of a cereal by sowing four plots of each in a Latin square. We have to arrange the 16 plots so that each row and column of the square contains one of each of the varieties, and yet the arrangement is otherwise to be random. We first proceed to draw a diagram with four rows and four columns to represent the 16 experimental plots. By suitably allocating four faces of a die, we can throw to find out which variety shall occupy the left top corner. Let us suppose *B*. We then proceed along the top row, throwing a die each time, and get *C* and *A*. The fourth must be *D*. Next, the left-hand column is suitably filled in by *D*, *C*, *A*. Note when there are only three possibilities two faces can be allocated to each variety and when only two, three faces.

The intersection of the second row and column can now only be filled by *A* or *B*, and a throw of the die makes it *A*. The intersection of the second row and third column may be *B* or *C*, and we find *C*, the last of the second row is therefore

$$\begin{matrix} & & D \\ B, & \text{and the last column must be} & B \\ & & A \\ & & C \end{matrix}$$

or there would be two *C*'s in the third row.

* The actual shape of the Latin square will be similar to that of one of its constituent plots and may therefore be only diagrammatically square. This is quite immaterial to the argument.

Finally, the intersection of the third row and second column may be B or D, and a throw of the die makes it B, which fixes the remaining three places.

B	*C*	*A*	*D*
D	*A*	*C*	*B*
C	*B*	*D*	*A*
A	*D*	*B*	*C*

This, which was actually arrived at by die throwing, is one of the 288 possible arrangements, and we may further use it for purposes of illustration by supposing that the yields were those of the S.E. corner of Montgomery's* diagram of plots of Turkey wheat given on p. 37 of his classical "Experiments in Wheat Breeding".

The yields in grammes are as follows:

				Sum of rows	Means of rows
B 617	*C* 683	*A* 726	*D* 835	2,861	715·25
D 602	*A* 662	*C* 640	*B* 700	2,604	651
C 665	*B* 736	*D* 630	*A* 598	2,629	657·25
A 609	*D* 706	*B* 790	*C* 678	2,783	695·75
Sums of cols. 2,493	2,787	2,786	2,811	10,877	
623·25	696·75	696·5	702·75		General mean 679·81

									Average
A =	726	+	662	+	598	+	609	=	2,595	648·75
B =	617	+	700	+	736	+	790	=	2,843	710·75
C =	683	+	640	+	665	+	678	=	2,666	666·50
D =	835	+	602	+	630	+	706	=	2,773	693·25

(*a*) *Variance of Columns*

Taking 2600 as a working mean, the deviations of *sums* of columns from this are

	−107	and the squares	11,449
	+187		34,969
	+186		34,596
	+211		44,521
Total	477		125,535
Deduct	$\frac{1}{4}(477)^2$	=	56,882·25
			68,652·75 ÷ 4* = 17,163·1875

* Divide by 4 because we have worked with *totals* and we want to change to means.

(*b*) *Variance of rows*

As above, deviations of sums of rows from 2600 are

	+261	and squares	68,121
	+ 4		16
	+ 29		841
	+183		33,489
Total	+477		102,467
Deduct	$\frac{1}{4}(477)^2$		56,882·25
			45,584·75 ÷ 4 = 11,396·1875

* *U.S. Dept. Agric. Bur. Plant Indust. Bul.* 269.

(c) *Variance of varieties*

Deviations of sums of varieties from 2600 are

	− 5	and squares	25
	+243		59,049
	+ 66		4,356
	+173		29,929
Total	477		93,359
Deduct	$\frac{1}{4}(477)^2$		56,882·25
			36,476·75 ÷ 4 = 9,119·1875

(d) *Total Variance*

Taking 680 as a working mean the deviations of the yields of the individual plots are as follows:

Yield of plot − 680	Squared
− 63	3,969
+ 3	9
+ 46	2,116
+155	24,025
− 76	6,084
− 18	324
− 40	1,600
+ 20	400
− 15	225
+ 56	3,136
− 50	2,500
− 82	6,724
− 71	5,041
+ 26	676
+110	12,100
− 2	4
− 3	68,933
Deduct $\frac{1}{16}(3)^2$	·5625
	68,932·4375

We have next to perform an operation analogous to that of multiplying by $\sqrt{\{n/(n-1)\}}$ in the case of finding the S.D. from a sample of n. Fisher's way of doing this is given below in the table of the analysis of the variance:

Variance due to	Degrees of freedom	Sum of squares	Variance	Standard deviation
Varieties	3	9,119·19		
Columns	3	17,163·19		
Rows	3	11,396·19		
Remainder	6	31,253·87	5,208·94	72·2
Total	15	68,932·44		

In the above table the first column is descriptive of the variance arising from different sources.

We are chiefly concerned with that entitled "Remainder", which enables us to arrive at an estimate of the random errors which are not associated either with variety or with that part of soil heterogeneity common to whole rows or columns. The second column gives the "Degrees of freedom". In the first, second, third and fifth rows of the table the degrees of freedom merely represent one less than the number in the sample (4 varieties, 4 columns, and 16 plots altogether) and are strictly analogous to the $n-1$ quoted above. The number

in the fourth row is obtained by making the first four rows add up to the total of the fifth.

The principle of degrees of freedom is widely applied by Fisher, and the idea behind it is that if there are a number n of variates of which the mean is used in the calculation, all but one of them can take any possible value; but when $n-1$ values have been chosen the last one is fixed by the mean, so that only $n-1$ variates are free to vary. If, in addition, some other statistic is used, such as the S.D., only $n-2$ of them can be varied, and so on.

In this case there are fifteen degrees of freedom in the total and three sets of three degrees of freedom are taken up by the varieties, the rows, and the columns, leaving six for the determination of the variance of the random error.

In the third column the first, second, third and fifth rows are the sums of squares calculated in (*c*), (*a*), (*b*) and (*d*) above, and the fourth is found by making the first four rows add up to the last.

In the fourth column the required variance is given by dividing the figure in the third column by that in the second and from this is obtained the S.D. by extracting the square root. If this had been found from enough degrees of freedom, we could find the S.D. of the difference between two varieties appropriate to use with tables of the normal curve by dividing by $\sqrt{\frac{4}{2}}$ (the 2 in the denominator being due to the fact that we are to judge of the significance of a *difference*, and the 4 to the number of replications), which would give a S.D. of about 50, while the greatest difference between "varieties" is only about 60. Obviously, this would not be significant, as indeed in this example it should not be, the "varieties" being all the same—Turkey Red wheat. In fact the significance is even less, as with only six degrees of freedom Student's table must be used.*

Unfortunately, Student's tables were constructed some time before the Latin square was thought of, and it requires some care to enter the table aright.

In the first place we have to enter the table under the heading $n = 7$, one more than the degree of freedom, since if Student's table had been headed with the degrees of freedom, the headings would have been one less.

Secondly, to obtain z, we divided the difference (say $B-A$ which is 62) by the S.D. $\times \frac{\sqrt{2} \times \sqrt{7}}{\sqrt{4}}$ in which the $\sqrt{2}$ corresponds to the fact that we are considering a *difference*, the $\sqrt{7}/\sqrt{4}$ to the fact that the original table was constructed so as to give the probability for means of 7, while we only have means of 4. z is here, therefore, $\frac{62}{130}$ or just under 0·5, which if looked out in this table under $n = 7$ gives $P = 0{\cdot}86$, a satisfactorily non-significant result.

* This applies to the tables given in *Biometrika*, VI, p. 19 and XI, p. 416; and those in the new edition of *Tables for Statisticians and Biometricians*. The tables in Fisher's *Statistical Methods for Research Workers* and those which are to appear in the next number of *Metron* are given under the headings of the degrees of freedom. [The *Biometrika* and *Metron* tables are those printed on pp. 29, 62–3 and 118–20 of this volume. ED.]

Looked at from another point of view we should require a difference between varieties of not less than 100 grammes, or some 15 %, for it to be worth while testing under such conditions as Montgomery's with only four plots of each variety.

I have illustrated the method on a Latin square of four plots per side, choosing a small number so as to make it easy to follow the arithmetic, but in point of fact *four replications are decidedly too few* and much larger squares are recommended. One of the disadvantages of this particular illustration has been that whereas usually the variance is much reduced by the subtraction of that associated with rows and columns, there has by chance been very little reduction in this case.

The following table gives other possibilities:

Number of varieties	Number of replications	Total number of plots	Number of degrees of freedom for calculation of error	Heading of column to be used in Student's tables	Factor to multiply S.D. by in calculating z*
4	4	16	6	7	$\sqrt{\left(\frac{2\times 7}{4\dagger}\right)}$
5	5	25	12	13	$\sqrt{\left(\frac{2\times 13}{5\dagger}\right)}$
6	6	36	20	21	$\sqrt{\left(\frac{2\times 21}{6}\right)}$
7	7	49	30	—	Use normal curve with S.D. $\times \sqrt{\left(\frac{2}{16-3\ddagger}\right)}$
4	16	64	46	—	

† Number of replications. ‡ Three less than the number of replications.

In the last case there will be two replications in each row and column, and care must be taken that the arrangement is really random, e.g. if one plot of A has been fixed in a row the chance of filling the next with A must be only half that of filling with one of the other letters not yet represented in the row.

Conclusion

To sum up, in planning agronomic experiments *use plenty of replications* and make quite sure that your results are capable of being considered to be a random sample of the population about which you wish to draw conclusions.

* [It should be remembered that this is the z of Student's original notation [2, p. 17 above] and not the quantity defined by R. A. Fisher and now generally used in the analysis of variance. Ed.]

14

ERRORS OF ROUTINE ANALYSIS

[*Biometrika*, XIX (1927), p. 151]

Introduction. Dr E. S. Pearson, *Biometrika*, XVIII, p. 192, has given the moment coefficients of the distributions of range in small samples drawn from the normal population when the number in the sample lies between 2 and 6. Mr L. H. C. Tippett, *Biometrika*, XVII, pp. 364–87, had already provided similar data for samples of 10, 20 and 60, but Dr Pearson gives improved values in his Table VIII which I have used.

These constants provide a means of drawing curves which approximate closely to the actual frequency curves of the distribution of ranges, apparently sufficiently closely for us to use their integrals as probability integrals for the occurrence of ranges of fairly large size.

Thus the real frequency curve for range in samples of two is known to be a half normal curve of standard deviation $\sqrt{2}.\sigma$, whereas the Pearson curve found from the moments is a Type I with equation

$$y = 543{\cdot}062\left(1+\frac{x}{0{\cdot}574}\right)^{0{\cdot}569}\left(1-\frac{x}{6{\cdot}623}\right)^{6{\cdot}569}.$$

If they be drawn on the same scale, Fig. 1, we see that for the greater part of the way the two curves are practically identical.

Assuming then, as seems likely, that the approximation in the case of the larger samples is even closer than in the case of samples of two, we have here a means of determining the probability of occurrence of ranges of given size in the case of quite small samples, assuming as always a normal population. Now it is just in the case of these small samples that most of the tests which have been proposed for the rejection of observations fail; there is no possibility of finding the true mean of the population. Mr J. O. Irwin, *Biometrika*, XVII, pp. 238–50, has, it is true, proposed to use Galton's differences for this purpose, but on the other hand, there are cases in which the true standard deviation of the population is known with some approach to accuracy, and it seemed in such cases Dr Pearson's work should enable us to reject determinations so widely spread as to render the occurrence of the observed range unlikely to any specified degree.

Happening to mention to Dr Pearson that I proposed to apply his work to the rejection and repetition of analytical results, he suggested that the readers of

Biometrika might be interested both in the application and, indeed, in a description of the errors of routine analysis from which the necessity of rejection arises.

In endeavouring to fall in with this suggestion, I propose to set out, firstly, what routine analyses are, and to what sort of errors they are liable; secondly, the advantages that accrue from a statistical examination of these errors; and, lastly, the bearing of Dr Pearson's paper on the vexed question of the repetition and rejection of results.

At the outset I may state that, though no analyst, I have been in close touch for some years with a routine laboratory, the authorities of which have very kindly supplied me with some of their results for the purpose of the present paper.

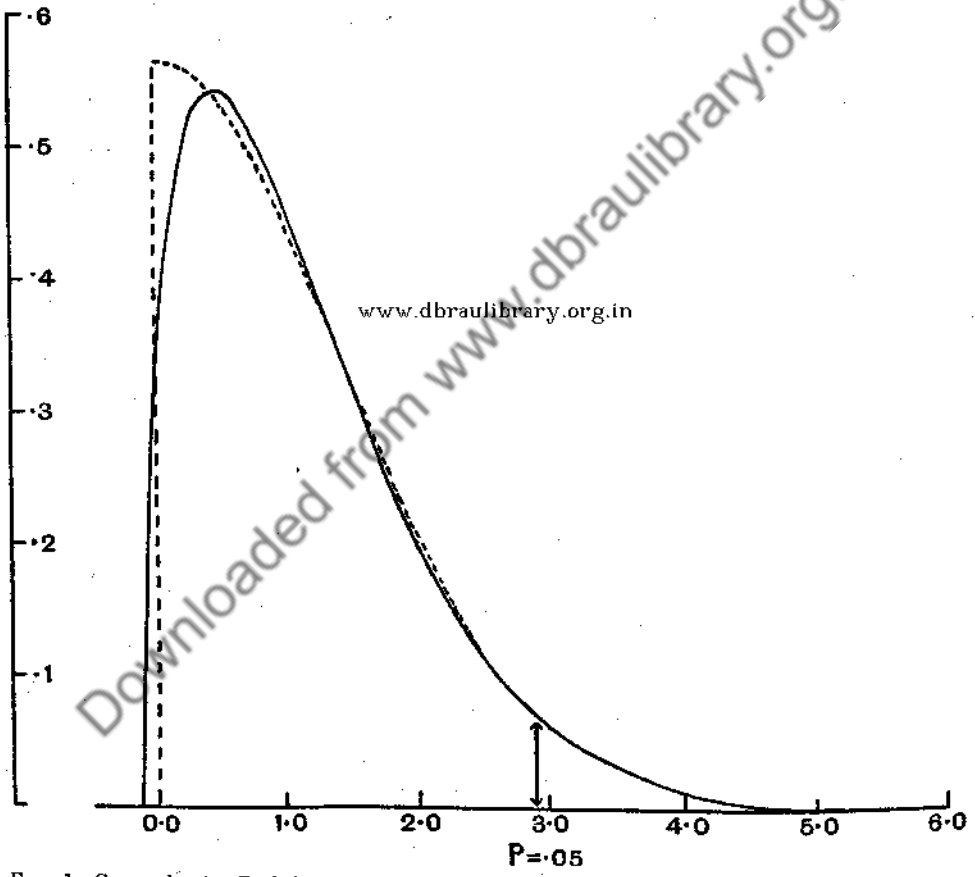


FIG. 1. Curve showing Probability of Occurrence of Various Ranges in Samples of two drawn from a series Distributed Normally with S.D. $=\sigma$.

——— Pearson's Approximation: Type I curve

$$y = 0{\cdot}543\left(1+\frac{x}{0{\cdot}574}\right)^{0{\cdot}569}\left(1-\frac{x}{6{\cdot}623}\right)^{6{\cdot}569}.$$

----- Actual Curve for Comparison: Half of normal curve

$$y = \frac{1}{2\sigma\sqrt{\pi}}\, e^{-\frac{x^2}{4\sigma^2}}$$

Routine Analysis. The difference between research and routine is fundamental to the scope of the present paper, and it lies in the relation between the analyst and his work rather than in the actual process of the analysis. Thus, what is at one time a research involving concentrated thought and watchfulness on the part of the analyst may, later on, become the merest routine; every step known and prepared for in advance, and requiring not the resourcefulness of the chemist of high degree but the machine-like accuracy of the well-trained assistant.

This is not to say either that research may not make use of routine processes, as it frequently does, or that routine processes may not form the subject of research, as they constantly should; yet, broadly speaking, we are not concerned with the distinguished chemist who determines the atomic weight of an element to n places of decimals, and has theories about the value of the $(n+1)$th—it would be impertinence to talk of errors in such a connexion.

No, we are going to deal with the chemist who has to make similar analyses day after day and year after year; with, for example, the public analyst who provides evidence to convict the milkman of watering his milk, and the grocer of sanding his sugar; with the works chemist, who maybe spends his whole life in determining the acidity or alkalinity of solutions; or again, with the assayer, who must find out which of innumerable samples of ore are payable.

There is often enough little or no scientific interest in such determinations, yet their practical value is in the aggregate enormous; the application of science to industry would without them be all but impossible.

These people are not so much troubling themselves about the nth place of decimals; their problem is to get results as quickly and as cheaply as possible; quickly, because events may be waiting upon them, and cheaply for reasons that need hardly be elaborated.

They must, however, attain sufficient accuracy for the purpose in hand, which is generally concerned with the third figure rather than the fourth, and is often enough satisfied with the second. Nevertheless without this minimum of accuracy the analysis is worthless, so that the chemist in charge of the laboratory has to make himself very sure that it is reached.

Obviously he cannot be sure, unless he has made some determinations of the error, and he can only reduce his error if he has a working knowledge of the sources of error.

Sources of Error. The first of these, very often the chief of them, is not strictly a laboratory error; it arises from the difficulty of obtaining a sample in a bottle which shall represent perhaps some tons, or even hundreds of tons, of material. This difficulty of sampling provides a convenient excuse for discordant results, but the wise chemist will see to it that the sample is drawn in a manner which will rob this excuse of any appreciable validity. And that is by no means easy: but the errors of commercial sampling do not fall within the scope of this paper, so I do not propose to say more about them here.

Nor do I propose to deal with the allied problem of subsampling the sample which has been received for analysis. This may be in the case of solids quite a difficult matter, and can lead to appreciable error unless a suitable technique is employed.

After this, each operation of the analysis contributes its error; I am told that the standard error of weighing on a balance is about one in two thousand; all analyses involve at least two weighings and there are often more. Then we have such things as titration, generally contributing quite a small error; transfer of material from one vessel to another; digestion at a uniform temperature; filtration, incineration, and so forth; all these add their quota.

These errors are not necessarily symmetrical, some of them involve loss of material, and for this reason a chemist will sometimes prefer the higher of two results.

Perhaps a description of a very simple analysis may illustrate the kind of thing that happens. Let us suppose that it is required to estimate the percentage of moisture in a sample of grain, not as part of a research but for the commercial valuation of a large bulk in a ship or warehouse; it will very likely be one of a number of analyses the results of which will be required by the next day.

First, the sample is subsampled and a weighed portion of the ground-up material is put into an oven on a small tray. The oven is kept at a constant temperature for a fixed number of hours, the tray is then removed, cooled over concentrated sulphuric acid and quickly weighed. The loss of weight is taken to be the moisture present in the weighed quantity which was put into the oven.

Here we have the errors of subsampling, grinding, two weighings, and of driving off moisture by heat; hardly any one of these operations is as simple as it sounds. The grinding, for example, whether done in a mill or with a pestle and mortar, leaves material on the grinding surfaces; this material is not the same as the bulk but is composed of the finer or more adhesive part of it. It is, therefore, necessary to grind and throw away a small quantity before dealing with the portion which is to be weighed. Then we have the fact that organic matter exposed to the atmosphere, generally if not always, tends to get into equilibrium with the moisture in the air, hence both grinding and weighing must be done rapidly.

When in the oven the loss of weight will depend not only on the exact temperature and time, but on the ventilation of the oven and the number of samples in it. Nor is all the loss necessarily moisture, carbon dioxide may either be formed and lost by oxidation, or be lost by splitting off from some already oxidized compound. We may even get the estimation too low owing to an increase of weight due to absorption of oxygen.

Of course, in a research one would work in an atmosphere of nitrogen and weigh the moisture absorbed by phosphorus pentoxide, determining one sample at a time and weighing at intervals until the weight became constant, but routine

analysis has neither time nor money for this: it has to rely on keeping the conditions constant. The most it can do is to check an occasional result by the more lengthy method.

All this sounds as if the results would be very inaccurate, yet it is not so. The moisture of grain, lying between 10–20 %, can be determined with a standard deviation measured in percentage moisture of about 0·2, or 1 part in 500. Naturally, different laboratories, using different ovens set up under different conditions, do not necessarily agree with one another, but they will probably agree to this order of accuracy in their relative estimates when comparing different samples, and that is usually what is required.

We now come to a phenomenon which will be familiar to those who have had astronomical experience, namely that analyses made alongside one another tend to have similar errors; not only so but such errors, which I may call semi-constant, tend to persist throughout the day and some of them throughout the week or the month.

Why this is so is often quite obscure, though a statistical examination may enable the head of the laboratory to clear up large sources of error of this kind: it is not likely that he will eliminate all such errors.

The chemist who wishes to impress his clients will therefore arrange to do repetition analyses as nearly as possible at the same time, but if he wishes to diminish his real error he will separate them by as wide an interval of time as possible. Here are some examples:

In 1905 a quantity of material was taken, mixed as well as possible and stored in Winchester bottles. Samples were taken from these and analysed daily between the beginning of April and the end of August—100 in all. This, though statistically speaking a small sample, represents an amount of work which a routine chemist will not easily be persuaded to undertake.

At each analysis seven items were determined and of these I have now examined five: all are troubled to a greater or less extent by semi-constant errors, as is most easily shown by a comparison of twice the variance of a single analysis with once that of the difference between consecutive observations: if the arrangement were random they would of course be the same within the error of random sampling.

TABLE I

Item	Twice variance	Variance of difference	Correlation between consecutive analyses
1	2·20	1·60	+0·27
2	0·625	0·434	+0·31
3	0·0748	0·0606	+0·19
4	0·171	0·157	+0·09
5	5·42	4·68	+0·09

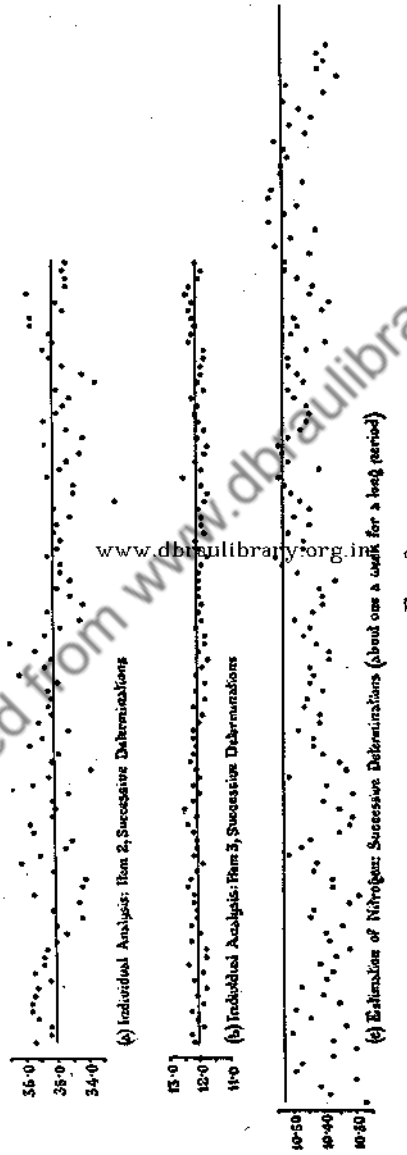


FIG. 2

Of course, not all of these correlation coefficients are individually significant, but they are illustrative of a general phenomenon. I do not recollect having met with a case where the correlation was negative.

The two top lines of dots in Fig. 2 give the individual analyses of items 2 and 3, the latter of which gives the percentage of moisture in the samples. The lines across the diagrams show the mean values of these.

I will now give another case of a routine analysis repeated in a time series. Here, as a check on the accuracy of the estimation of nitrogen by the Kjeldahl method, a determination of the nitrogen in pure crystalline aspartic acid was made about once a week from 1903 to the present time. The method is a standard one for the determination of "amino" nitrogen in organic matter. A weighed quantity of the substance to be analysed is digested in strong sulphuric acid which destroys the organic matter and converts the nitrogen into ammonium sulphate. Excess of alkali is then added and the nitrogen distils over in the form of ammonia and is caught in a measured quantity of acid, where it is estimated by titration of the excessive acid with deci-normal soda.

Of course the amount of nitrogen in a crystalline substance can be calculated within narrow limits and the third row in Fig. 2 gives the calculated (as a straight line) and the actual (as spots) since 29 April 1924 up to the end of 1926.

At first the results were all too low, but the details of the process were under examination and the later estimates have risen and the variance has decreased owing to improvements which have been effected: simultaneously, the time taken has been reduced by half. For about six months before the beginning of last November the results were remarkably good; one could have calculated the atomic weight of nitrogen from the mean with an accuracy which would hardly have disgraced research, but there has since been a falling off, the average of the last seven being rather over 1% too low. This illustrates the sort of difficulty which arises in routine analysis, for no one is conscious of any alteration in method, nor has a close search revealed the cause of the change.

The error statistics which I have cited up to the present have all been obtained by the laboratory in the course of investigation into, and control of, its error. I am now going to give some figures taken from some published analyses which seem to me to show that similar "semi-constant" errors probably exist in another laboratory; it would surprise me to find any laboratory without them, but it is only by chance that they become apparent unless they are deliberately sought for.

The analyses are published in the *Report on the Sugar Beet Experiments* 1925, issued and distributed without charge by the Department of Agriculture of the Irish Free State.

These experiments were conducted at 424 farms, all the twenty-six counties being represented, and the complete programme, consisting of two plots of each of four varieties, one top dressed with nitrate of soda and one not, was successfully

carried out in 163 cases, in another 190 it was found necessary to top dress all the plots, and the remaining 71 cases fell through for one reason or another. It is the 163 complete results with which I propose to deal.

It will be seen that each farm produced eight different lots of beet and as *each* of these was analysed to find the percentage of sugar we can average the figures to get the percentage of sugar for the farm. Further, the date on which the analyses of each farm were carried out is given in the report, and in Fig. 3 are given the averages of analyses made on the same day as central points with lines extending upwards and downwards showing the extent of twice the standard deviation of the mean of the number of analyses, ranging from 8 (one farm) to 96 (twelve farms), which were made on that day.

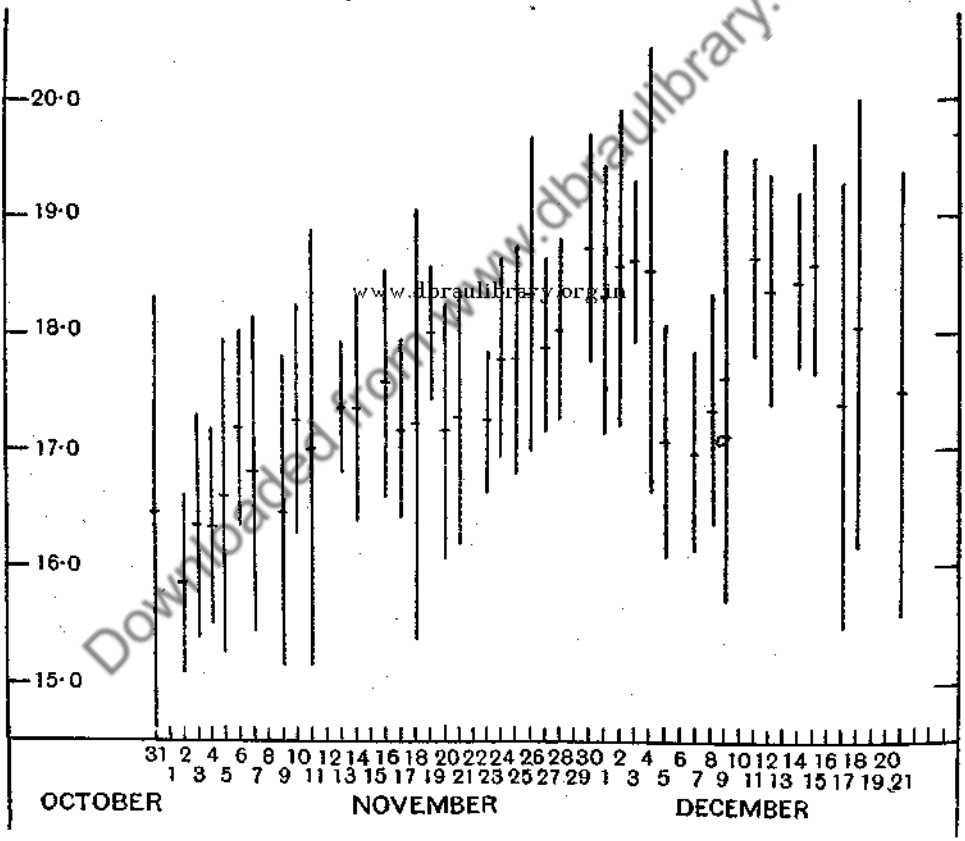


FIG. 3. Means of Daily Analyses with lines showing on each side of the Mean twice the S.D. appropriate to the Number of Analyses made on any given day. The S.D. is derived from the total observations by the formula

$$\sigma = \frac{1}{\sqrt{n}} \sqrt{\left(\frac{S(a-\bar{a})^2}{S(n-1)}\right)},$$

where a = average of a farm,
$\bar{a}$ = mean of a day's analyses,
n = number of farms analysed in the day.

It is obvious from an inspection of the figure that there was a distinct rise of sugar between the beginning and end of November, which is doubtless due to the gradual maturing of the roots, but it is not easy to account for the marked dip shown by the analyses carried out on the 5th, 7th and 8th of December, on any other supposition except that of laboratory error.

The thirteen farms, the produce of which was analysed on those dates, were in five counties, so that the roots were sent up by five different men and may be considered a random sample of the material to be analysed in the early part of December. It has been suggested that the loss of sugar was due to the action of frost on the roots before they were drawn from the ground or whilst in transit from the farms to the State Laboratory. From inquiries which I have made I am satisfied that the lower sugar content is not attributable to such action for such frosts as were experienced did not apparently affect the leaves, let alone the roots, and the packing of the beet to ensure its arrival in fresh condition at the laboratory obviated any possibility of freezing in transit.

I have also been informed that beetroots lose sugar when they are clamped. I am assured, however, that none of the samples to which the report relates was pitted or clamped but that each sample of roots was washed, topped or crowned and dispatched to the laboratory immediately after being taken out of the ground. The roots were forwarded by passenger train so as to secure quick transit, were unpacked immediately before the analysis was commenced and, as a rule, the analysis of a sample was completed within twenty-four hours of its receipt in the laboratory. It seems likely, therefore, that the low results were due to errors of a similar nature to those which were observed in the other laboratory.

To embark on a long series of analyses in order to determine error is always a considerable undertaking and is often impossible owing to the tendency of organic substances to change with time: added to this, unless special precautions are taken, such as were taken in 1905, the operators may, in spite of themselves, be more careful when analysing special samples of this kind, so that the series may not represent a random sample of analytical errors.

It is convenient, therefore, to take advantage of the fact that important analyses are often repeated as part of the routine and to calculate the standard deviation of the error from the differences between pairs by simply dividing the variance of the differences by 2 and taking the square root.

I give in Table II the standard deviations of errors of the items 1 to 5, the variance of which I gave before, but having in addition further determinations made from the differences between 100 pairs analysed in 1925 and in 1926.

The standard error arrived at in this way is that of analyses made within a comparatively short period of time and does not take account of the variation of the "instantaneous mean" which we have just been observing. It is therefore the correct measure of the error if we wish to compare such analyses with each other

TABLE II

Item	S.D. 1905	Error differences between pairs		
		1905	1925	1926
1	1·048	0·895	0·731	0·660
2	0·559	0·466	0·386	0·523
3	0·193	0·174	0·138	0·152
4	0·293	0·280	0·326	0·272
5	1·640	1·570	2·810	2·120

but is too small if the analyses were separated by a wide interval. On the other hand, the standard error derived from 100 analyses spread over three months is too large when we are dealing with the differences between consecutive analyses. The difficulty can only be removed by reducing the secular variation to negligible limits.

Perhaps it would be well to illustrate this point in some further detail. Suppose a merchant to be offered two samples of grain at the same price: as far as he can judge they are of equal value but he is uncertain whether the moisture is the same. He gets them analysed and is returned the figure 14 % for sample *A*, and 15 % for sample *B*. If the standard deviation of the error is 0·2 % clearly he should purchase *A*; if the error were 4 % it would not much matter which he bought. But observe in this case, as in many others, he is really only concerned with the difference between *A* and *B*, and if he controls the analysis he will get them done alongside each other so as to avoid their being affected by semi-constant error, and the error of the analysis will be about that found from 100 differences.

On the other hand, suppose he has bought a cargo of grain and an analysis tells him that the moisture is 17 % while it is common knowledge that 17·5 % is the highest moisture at which grain will keep. Here he is not concerned with relative but with absolute values and the error now includes the semi-constant error, so that the value deduced from the 100 analyses spread over a long time is the better.

As a sort of corollary of the existence of semi-constant errors in the same laboratory, we find that different laboratories have different constant errors, and a wise man will always consult the same analyst and not be troubled overmuch if a second analyst does not exactly agree with him.

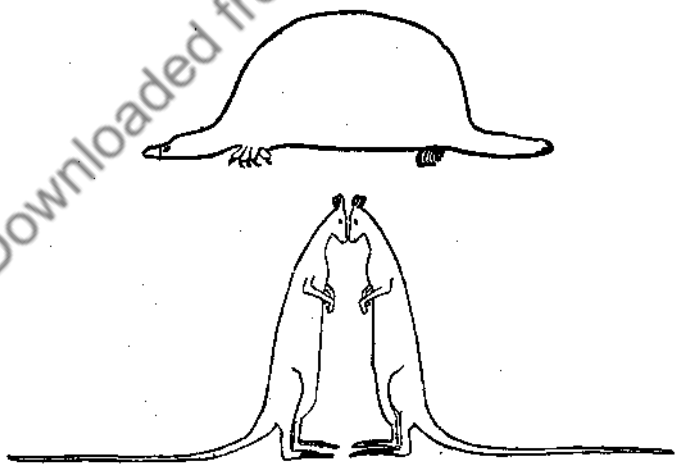
I have now, I hope, shown that routine analyses are subject to errors of which it behoves the head of the laboratory to be well aware. He may then judge whether his analyses are sufficiently accurate to bear the weight of any actions which it may be proposed to base upon them, and if not, how many repetitions will suffice to make them so; he will realize that an analysis made elsewhere is not necessarily less valuable than his own because it does not agree absolutely with it, and he will be in a better position to set about improving the details of his methods than if he were ignorant of the magnitude of his errors.

I now turn to the particular point raised by Dr Pearson's paper. It will be realized from what has gone before that important analyses may have to be repeated and the same applies of course to those which have given results at variance with *a priori* expectation. Very important results may even have to be repeated more than once, and it is only natural to regard these pairs—triplets or quartets—with suspicion if the results are not "concordant", i.e. have a wide range.

The result is that there is a tendency to make further repetition in such cases, to reject discordant results, and to accept the mean of the remaining observations: all the same this instinctive distrust of width of range needs some justification, and, if justified, some rules for repetitions.

For if the error were normally distributed there would be no advantage in rejection; this follows from the fact that in normal distributions there is no correlation between the square of the mean and the variance: similarly, in platykurtic* distributions those samples with large variance even tend to have the *more* accurate means. Actually, however, many if not most routine analyses have a leptokurtic error system, possibly because the standard deviation as well as the mean is subject to variation with time, and in such cases rejection of outlying observations improves the accuracy of the mean; apart from this we are all fallible and the procedure takes account of blunders.

* In case any of my readers may be unfamiliar with the term "kurtosis" we may define mesokurtic as "having β_2 equal to 3", while platykurtic curves have $\beta_2 < 3$ and leptokurtic



> 3. The important property which follows from this is that platykurtic curves have shorter "tails" than the normal curve of error and leptokurtic longer "tails". I myself bear in mind the meaning of the words by the above *memoria technica*, where the first figure represents platypus, and the second kangaroos, noted for "lepping", though, perhaps, with equal reason they should be hares!

The following table gives the values of β_2 for samples of the five items of analysis which I have given before:

TABLE III

Item	100 analyses of 1905	Differences between consecutive analyses of 1905	Differences between 100 pairs in 1925	Differences between 100 pairs in 1926
1	3·1	2·7	4·8	5·1
2	3·5	2·9	8·2	7·4
3	2·3	2·6	2·9	3·2
4	2·9	2·7	10·4	16·2
5	10·0	5·5	5·0	7·1

In this table the differences between the β_2's of twenty years ago and those of the present day are rather remarkable, and though with small samples such as these the standard deviation of β_2 is enormous I should hesitate to assert that they are due to random sampling; I am inclined to think that there has probably been a twofold change, (1) that the error of the great majority has decreased, and (2) that possibly owing to work being carried on at higher pressure there is a rather greater liability to blunders. In this way the standard deviation remained much the same but the kurtosis has increased. Be that as it may, the tendency to leptokurtosis is apparent and repetitions justified except in the case of No. 3, which, as I mentioned before, indicates moisture. Here the kurtosis of the difference between pairs is approximately "meso" while that of the 100 analyses appears to be distinctly platykurtic; this is in accordance with another distribution of moisture determinations which I have examined.

Why this should be I have no idea, but obviously if a normal error were superposed on an instantaneous mean which moves to and fro on, let us say, a sine curve, the resulting distribution would be platykurtic: something of this sort may have happened.

Assuming, however, that discordant observations are to be repeated and if necessary rejected, it is obviously of advantage to work on a regular system, and since we do not know where the mean is I propose to use the range as follows:

Let W_n be the limit at which with a sample of n, the chance of obtaining a greater range than W_n is p (say 0·05), then if w_n the actual range of a sample be greater than W_n repetition should be made. Let w_{n+1} be the range of the new sample including the repetition, then if $w_{n+1} < W_{n+1}$ the mean of the $n+1$ results should be accepted. If, on the other hand, $w_{n+1} > W_{n+1}$ the most outlying observation should be rejected, and if then the resulting $w_n < W_n$ the mean of these n should be accepted, but if not, a further repetition should be made and the whole $n+2$ observations examined afresh, and so on until a sample of at least n is

obtained lying within the required limits. For example, we may have a quartet of analyses

$$\left.\begin{matrix}22{\cdot}8\\23{\cdot}5\\26{\cdot}0\\26{\cdot}6\end{matrix}\right\}$$ the values of W_n for this analysis (S.D. 0·675) being as follows $$\begin{cases}W_4 = 2{\cdot}4,\\W_5 = 2{\cdot}6,\\W_6 = 2{\cdot}7,\\W_7 = 2{\cdot}8.\end{cases}$$

Here $w_4 = 3{\cdot}8$, so we repeat and get 23·9. Then $w_5 = 3{\cdot}8$ and we reject 22·8 leaving $w_4 = 3{\cdot}1$. We therefore repeat again getting 25·5. Then we have $w_6 = 3{\cdot}8$, $w_5 = 3{\cdot}1$ (rejecting 22·8) and $w_4 = 2{\cdot}5$ (rejecting 26·6). Still another repetition gives 25·0 and we reject in turn 22·8, 26·6 and 26·0, leaving 23·5, 23·9, 25·0 and 25·5, with a range of 2·0 and an average of 24·5 which we accept.

To obtain W_n, the curves giving the frequency distributions of range for samples taken from a normal population were drawn from Pearson's constants and the limits at which p is 0·1, 0·5 and 0·02 were determined. This gives us limits for samples of 2, 3, 4, 5, 6, and between 6–10 we can interpolate with the aid of Tippett's values for 10, 20 and 60. W_n is of course given in terms of the standard error calculated from samples of analyses such as I have instanced above.

TABLE IV

	$p = 0{\cdot}1$	$p = 0{\cdot}05$	$p = 0{\cdot}02$
W_2	2·3	2·9	3·3
W_3	2·9	3·4	3·8
W_4	3·2	3·6	4·1
W_5	3·4	3·8	4·3
W_6	3·7	4·0	4·5
W_7	3·7	4·1	4·5
W_8	3·8	4·2	4·6
W_9	3·9	4·3	4·7
W_{10}	4·1	4·5	4·9

Fig. 4 gives a comparison of the distribution of range in samples of 4 calculated from Pearson's constants with the actual distribution in samples of 4 which occurred in the ordinary course of business when an important series of analyses was being made: the item was that which I have indicated as (1) in the tables of this paper.

It will be seen that while the general shape of the curve gives a fairly good fit ($P = 0{\cdot}13$ for 5 groups) there is excess at the tail end, showing the leptokurtic nature of the distribution and the advantage of repetition.

To recapitulate, routine analyses are subject to errors of which an estimate can be made either by a special analysis of a comparatively large number of samples of the same material, or by considering the differences between pairs which occur in the ordinary course of business. Owing to the fact that there is usually a secular variation in the error these will not in general give the same result and care must be exercised in the use of the standard deviation obtained. From such

determinations of error combined with certain factors obtained from Dr Pearson's paper on the range of small samples, we have derived limits at which repetitions

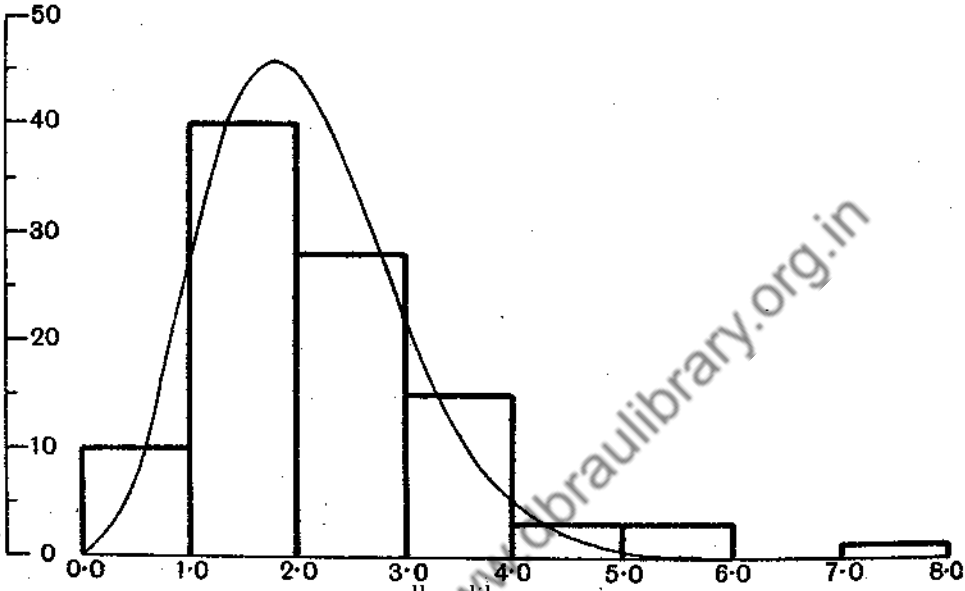


FIG. 4. Frequency Curve showing Expected and Histogram showing Actual Number of Ranges of given size in samples of 4 for 100 trials.

Equation to the curve is

$$y = 45{\cdot}694\left(1+\frac{x}{2{\cdot}101}\right)^{4{\cdot}499}\left(1-\frac{x}{9{\cdot}055}\right)^{19{\cdot}395}.$$

Equations to the Curves for Distribution of Ranges in Samples of Various Sizes

n	Type of curve	Equation (to give $N=100$)
2	I	$y = 543{\cdot}062\left(1+\frac{x}{0{\cdot}574}\right)^{0{\cdot}569}\left(1-\frac{x}{6{\cdot}623}\right)^{6{\cdot}569}.$
3	I	$y = 462{\cdot}017\left(1+\frac{x}{1{\cdot}547}\right)^{2{\cdot}456}\left(1-\frac{x}{7{\cdot}590}\right)^{12{\cdot}050}.$
4	I	$y = 456{\cdot}941\left(1+\frac{x}{2{\cdot}101}\right)^{4{\cdot}499}\left(1-\frac{x}{9{\cdot}055}\right)^{19{\cdot}395}.$
5	I	$y = 465{\cdot}255\left(1+\frac{x}{2{\cdot}478}\right)^{6{\cdot}601}\left(1-\frac{x}{10{\cdot}892}\right)^{29{\cdot}009}.$
6	I	$y = 475{\cdot}268\left(1+\frac{x}{2{\cdot}754}\right)^{8{\cdot}688}\left(1-\frac{x}{13{\cdot}134}\right)^{41{\cdot}433}.$
10	I	$y = 508{\cdot}412\left(1+\frac{x}{3{\cdot}659}\right)^{20{\cdot}772}\left(1-\frac{x}{73{\cdot}905}\right)^{419{\cdot}514}.$
20	VI	$y = y_0(x-41{\cdot}567)^{31{\cdot}440}\,x^{-373{\cdot}818}$ $[\log y_0 = 603{\cdot}835\ 959\ 3].$
60	VI	$y = y_0(x-11{\cdot}673)^{39{\cdot}324}\,x^{-174{\cdot}252}$ $[\log y_0 = 187{\cdot}208\ 715\ 5].$

should be made and beyond which outlying observations should be rejected. A rule is given for the application of this procedure, but it should always be remembered that such rules are to be regarded as aids to and not as substitutes for common sense.

I should like to thank the authorities in charge of the laboratory who have allowed me to use their figures and several friends who have helped in the preparation of the paper, particularly "Mathetes", who has computed the equations and drawn the figures for me. I am further indebted to Dr Hinchcliff of the Free State Department of Agriculture, who supplied me with information about the sugar beet.

15

YIELD TRIALS

[*Baillière's Encyclopædia of Scientific Agriculture* (1931), p. 1342]

It is quite easy to produce new hybrids; from a single cross fertilization between two varieties of a cereal one can select thousands of strains, differing more or less in some character from all the others. Whether any particular strain is worth preserving will depend upon many things, but among them one is indispensable: the yield *must be sufficiently high to make the crop profitable.*

Similarly new manures or combinations of manure are continually being proposed, and the one condition of their use is that the increase in yield which they provoke must be such as to pay for the cost of applying them.

As improvements are nowadays not likely to be very great, it becomes necessary to estimate comparative yields very closely, and this is not as simple a matter as it may appear at first sight.

In the case of selection of strains of high yield the difficulties are of two kinds: (1) That similar environmental conditions of weather, soil, etc., may evoke different responses in strains, even though nearly related, of the same race; one may be better suited than another by a light soil, or a dry summer, and so on. (2) Quite apart from characteristics of this kind, the soil on which the plants are grown is never uniform, so that differences in yield arise which have nothing to do with the strains which are being tested. Similar considerations apply to manurial and other trials.

Clearly, difficulties of the first kind can only be surmounted by repeated trials in many seasons and in all relevant types of soil and situation; but until we have arrived at some method of estimating what error is introduced into our conclusions by difficulties of the second kind and of reducing this error to manageable dimensions, we are not in a position to say whether observed differences in yield are due to the different strains, to their differential response to their environment, or merely to chance variation in the soil on which they have been grown.

In what follows it is proposed first of all to give a brief account of some of the methods of arranging yield trials which have been introduced during the past twenty-five years, then to indicate the general reasoning which enables us to estimate the degree of reliance which we can place on our results, and, finally, to work out two examples of the actual calculation of such estimates.

Although it must have been recognized long ago that experiments to determine comparative yields were not quite straightforward, the science of planning such

experiments was not investigated up to twenty-five years ago, and the practice of the art is only now becoming general.

It is true that sound results had been obtained by long continued trials carried out over a wide area with comparatively large plots, notably by the Danish Royal Agricultural Society and by the Irish Department of Agriculture (*Slutningsberetning om Maltbyg- og Hvedeudvalgets Virksomhed Vedrorende Byg- og Hvedeavlen*, Chr. Sonne, Foredrag i Det Kgl. Danske Landhusholdnings-Selskab den, 1 April 1903; H. Hunter, *The Barley Crop*, Ernest Benn, London), but, on the other hand, there was a tremendous amount of energy wasted on experiments from which, as we now know, it was impossible to have reached reliable conclusions.

To obtain decisive results in this way it is not only necessary to work on a very large scale (in the Irish work Archer and Goldthorpe barleys were compared fifty-one times over a period of six years), but the differences to be determined have to be comparatively large, for a single one-acre plot must exceed another by at least 25 % if it is to be considered significantly better.

It was not, however, until 1910–11, with the publication of papers by Stratton and Wood, Mercer and Hall, and Montgomery (T. B. Wood and F. J. M. Stratton, "The interpretation of experimental results", *J. Agric. Sci.* vol. III, No. 4; W. Mercer and A. D. Hall, "Experimental error of field trials", *J. Agric. Sci.* vol. IV, No. 2; E. G. Montgomery, "Variation in yield and methods of arranging plots to secure comparative results", *Nebr. Agric. Expt. Stat. 25th Ann. Report*, and "Experiments in wheat breeding", *U.S. Dept. Agric. Bur. Plant Indust. Bul.* 269), that the real difficulties of the problem became fully apparent. Each of these papers dealt with the yields on the component parts of an area of land. Stratton and Wood dealt with $\frac{9}{10}$ acre of mangolds in plots of $\frac{1}{1000}$ acre; Mercer and Hall with 1 acre of wheat in $\frac{1}{500}$-acre plots, and also with 1 acre of mangolds in $\frac{1}{200}$-acre plots; Montgomery, for two years in succession, with wheat grown on the same $\frac{7}{45}$ acre and harvested in $\frac{1}{1440}$-acre plots.

In each case the area was chosen as being particularly uniform in appearance, in each case the yields showed unexpected variability. Further, this variability *was not random* (see section on Randomness), nor, on the other hand, was it, except in the very slightest degree, regular. There was, it is true, a general tendency for plots at one end or side of the area to give higher yields than those at the other, but the "contours of fertility", though they existed, showed no exact parallelism (unpublished work by R. A. Fisher).

This suggested at once, firstly, that great accuracy would be obtained if plots whose yields are to be compared were sited closely together so that chance variations in the soil fertility should be shared as equally as possible; and, secondly, that to obtain this close siting the plots must be kept as small as it is convenient to work with, especially if many variants are being tested.

But, besides convenience in working, there is another limit to the smallness of experimental plots.

This is due to the fact that the outside of a plot does not represent a field crop, since it is in contact with plants of some other variety, or subjected to another method of treatment, and experience has shown (T. A. Kiesselbach, "Plot competition as a source of error in crop tests", *J. Amer. Soc. Agron.* vol. XI; E. S. Beaven, "Pedigree seed corn", *J.R.A.S.E.* vol. LXX, 1909) that plants growing alongside one another are in strong competition for both food and light. Nor can this difficulty be overcome by leaving unoccupied space between the plots, for even in this case the outside plants are only representative of the outside of a field where the plants are able to get excessive nourishment, and, of course, the outside forms a small part of a large field, but a very sensible proportion of a small plot. It is, therefore, necessary either to have the plots so large that the "border effect" is negligible, or to discard the outside rows and plants from the portion which is to be weighed.

The first system of yield trials based on a realization of the foregoing facts was Dr Beaven's "chessboard" system of square yard plots (E. S. Beaven, *ibid.*; "Student", "On testing varieties of cereals", *Biometrika*, XV, pp. 271–93 [11]), which was, in fact, in use at Warminster in 1909 before the publication of the three papers which have been cited.

This system, which has become the standard method of comparing the yields of varieties of cereals under wire cages, derives its name from the fact that the plots are square. Each square measures 4 ft. along the side, and in it are sown eight rows of seeds at 6 in. between the rows, the seeds being planted 2 in. apart in the rows. At harvest, however, the two outside rows are rejected, and also plants in the 6 in. at both ends of the other rows. Thus interference with neighbouring varieties is reduced to a very small amount.

The arrangement of the plots in the experimental area merits attention, and should fulfil the following conditions:

(1) The mean position of each strain tested should be the same, to counteract the effects of a possible "fertility slope".

(2) Different plots of the same strain should be spaced so as not to be needlessly close to one another; each strain then shares as far as possible in the casual vicissitudes of the experimental area.

Beaven's own arrangement was as follows:

A	*F*	*C*	*H*	*E*	*B*	*G*	*D*
B	*G*	*D*	*A*	*F*	*C*	*H*	*E*
C	*H*	*E*	*B*	*G*	*D*	*A*	*F*
D	*A*	*F*	*C*	*H*	*E*	*B*	*G*
E	*B*	*G*	*D*	*A*	*F*	*C*	*H*

the panel of forty plots being repeated as often as is considered necessary, generally, in his case, four times.

It will be seen that this evens up a fertility slope across the panel, but that the earlier letters are more to the left than the later by a small amount—*A* averages $\frac{7}{10}$ of a square to the left of the centre line, *H* the same amount to the right of it—and to correct for this, the present writer has suggested that alternate panels should be reversed, *H* being written for *A*, and so forth. This contravenes condition (2) at the places where the panels join, but not to a very serious extent. Even so, this arrangement has been criticized on the ground that a regular pattern makes it impossible to assume that the error of estimation is random.

Be this as it may, such an arrangement has two solid advantages for this particular purpose over partly random or controlled random arrangements, such as Fisher's randomized blocks or Latin squares which are described below. These advantages are: (1) that the chances of mistake are lessened by a regular system, and such mistakes have been known to occur even with the most experienced workers; (2) that the use of such plots for observation purposes is very much facilitated by the ease with which a particular strain may be picked out.

The chessboard arrangement has been of great practical service in testing barley hybrids, and, in fact, the two varieties of barley most popular at the present time in the British Isles, Plumage-Archer and Spratt-Archer (see Barley), were both tried out in wire cages in this way, and found to be superior to their competitors before proceeding to trial on a larger scale.

On the other hand, selections which have proved successful in the cage have not always succeeded in the field, though this is probably due to the fact that there is a difference between horticultural methods such as are there used and the ordinary procedure of agriculture. It may also be due to the wire covering, and experiments which Mr M. Caffrey is conducting at the Royal Albert Agricultural College at Glasnevin, Dublin, in the open, may throw light on this point.

Before considering large-scale work, it is necessary to refer to "rod rows", i.e. plots consisting of a single row of plants one rod in length. These have been used largely in American work (H. K. Hayes and A. C. Arny, "Experiments in field technique in rod row tests", *J. Agric. Res.* XI, p. 399), but in their original form, though, of course, very convenient for purposes of observation, they are nearly useless for the determination of yield even when replicated many times. This follows from the fact noticed above, that the yield is due not only to the inherent quality of the seed, but also to the vigour or lack of vigour of its neighbours.

In a modified form, where three or more rows of the same variety are grown consecutively, and the outer rows rejected when determining yields, but retained if necessary for use as seed, the rod row system can give useful results if sufficient replications are made. Even so the area wasted by rejected border amounts to

a large proportion of the whole (67 % with three consecutive rows, 50 % with four, as compared with 44 % in the chessboard), so that the method is not recommended, except as a rough test at the stage where a large number of strains or hybrids is to be cut down by wholesale discards, while at the same time as much seed as possible is wanted for those selected for further trial (F. W. Hilgendorf, "Plant breeding methods results", *N.Z. J. Agric.* March 1928).

Large-Scale Work

Beaven's Half-Drill Strip Method. We now come to methods of carrying out experiments on an agricultural scale, and here, again, Beaven has introduced a method which takes full advantage of the light thrown on the problem by the papers cited above (E. S. Beaven, "Trials of new varieties of cereals", *J. Minist. Agric.* vol. xxix, Nos. 4 and 5 (1922); "Student", *loc. cit.*).

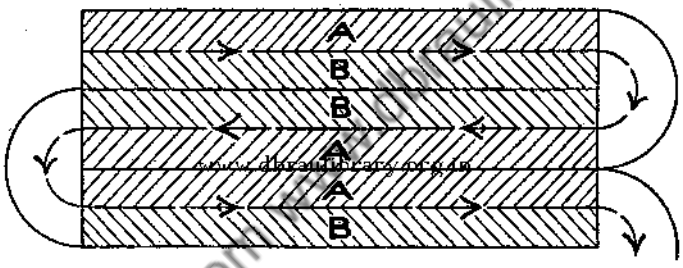


Fig. 1

In order to compare yields grown on areas as contiguous as possible, he took an ordinary seed drill of which he put the middle coulter out of action and divided the seed box into two halves. Seed of different strains having been put in the two halves, the drill is driven down the field, and wheeling at each end, it sows the strains as in Fig. 1. At harvest the outside drills of each half-drill strip—those in contact with the other strain—were pulled up by hand and discarded to avoid the "border effect", and each half-drill strip cut separately. If the two strains ripened simultaneously they were cut by a reaping machine, but, if not, one had to be cut by hand.

Originally a machine was used which delivered a separate sheaf off each $\frac{1}{500}$ acre, but this procedure was criticized on the ground of lack of randomness, and was afterwards found not to be necessary, the gain in the apparent accuracy over the plan of weighing only the totals of each half-drill strip not being found worth the additional trouble.

The weight of each half-drill strip is then compared with its neighbour of the other strain, and the layout of long narrow plots placed closely together ensures that the error of the comparison shall be small.

To compensate for the probable fertility slope the series should begin and end with the same strain, and the following precautions be taken:

(1) The ground chosen for the experiment must be free from periodic changes of level, such as those left by having been laid down to grass in "lands", or, if present, the seed must be drilled across these "lands".

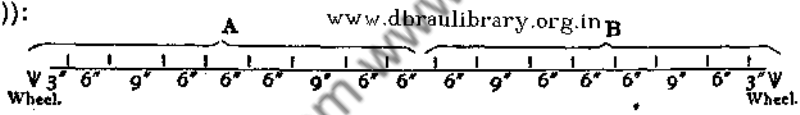
(2) The drills should run across those of the previous cultivation.

(3) The experimental area should be surrounded by at least one drill of the same kind of crop as is being experimented with.

(4) Great care must be taken when sowing to drive quite straight, so that the "inside" wheel of the machine may run as nearly as possible in the same track as it made when it was the "outside" wheel in the last journey in the other direction.

(5) After harvest the drilling must be checked by measuring the distances between the outside rows of corn, which may be appreciably different in the case of the two varieties owing to the horses pulling unequally, and this will favour that variety which has the wider gap. As, however, the gain will be approximately proportional to the gain in area, allowance can be made for this.

To avoid this difficulty Beaven now uses a special drill as follows (W. H. Parker, "Report on trials of four new barleys", *J. Nat. Inst. Agric. Bot.* No. 14 (1925)):



As before, the seed box is divided so as to take two strains, but the coulters are spaced so that between each four rows, 6 in. apart, there is a wider gap of 9 in., enabling each four rows to be cut separately.

Half of the sets of four consist of two rows of each strain, and these are discarded at harvest; the others are thus arranged in the same manner as before (*A B B A A ... B A*), and not only have the advantage that they are sown on equal areas whether the drill be driven straight or no, but they are also flanked by two discarded rows of their own kind, thus reducing interference to a minimum.

It will be seen that the half-drill strip can only compare two strains at a time, and if several are to be tested it is necessary to compare each with a "control". This is a serious limitation; nevertheless, the shape of the plots enables the comparison to be made with great accuracy.

This arrangement of plots also has been criticized on the ground that it is not random.

The application of the principle of long and very narrow strips thus introduced by Beaven for cereals can, of course, be applied to root crops, but for manurial experiments there would be danger of the benefit of the manure straying to the neighbouring plot. For these, the plots must be wider, and the method of their

arrangement has been made a special study by Dr Fisher, at Rothamsted. As a result he has evolved (1) "Randomized blocks", and (2) the "Latin square" (R. A. Fisher, *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh).

In the former he divides up the experimental area into blocks, which will contain one of each of the variants to be tested (varieties, manurial treatment, or whatever it may be). Within the block the arrangement is random, determined by some such method as dice throwing. The advantage is that since all the area of any one block is likely to be more uniform than the whole area, the trial within that block will be affected to a less extent by variations in soil fertility than if the plots were scattered about over the whole area, and yet the arrangement is random. The disadvantage is that in practice it may happen that the particular random arrangement adopted may result in one strain (or treatment) having a more favourable mean position than another: in a majority of the random blocks it may be on the north, and the north end more fertile than the south. Such a possibility is allowed for in the subsequent calculation, but the general effect is to introduce an unnecessary increase of uncertainty into the result: the error is larger than need be.

To meet this Fisher evolved the Latin square, where the mean position of each variety is situated in the same place by making it occur in each row and in each column of a "square", repeating the "squares" as often as may be required. Thus four strains (*A*, *B*, *C* and *D*) might be arranged in a square thus:

A	*D*	*C*	*B*
B	*A*	*D*	*C*
D	*C*	*B*	*A*
C	*B*	*A*	*D*

FIG. 2

the actual position being obtained by dice throwing with increasing limitation, as, for example, after the top row has been fixed no further plot in the first column can be A_1 and so on.

This most ingenious arrangement is ideal from the point of view of interpretation, but care should be taken not to have the plots so large that they do not lie closely together, or the error of the results, accurately though it is estimated, may be so large as to make the experiments inconclusive.

The Latin square need not, of course, be square; it will be of the same shape as any of its constituent plots. Thus, in the case of potatoes, where there is some evidence (R. N. Salaman, "The determination of the best method for estimating potato yields", etc., *J. Agric. Sci.* XIII (1923), p. 361) that the "border" effect is negligible, the width of each "plot" might be a single drill, and the length (for

four strains) one-quarter the length of the field; a number of such "squares" side by side would doubtless give results subject to a very small error.

At first sight it might seem that cereals might be tested by a combination of Beaven's half-drill strip with the Latin square, but the practical difficulty of the time taken to clear out the seed boxes after every strip, or alternatively of driving the drill straight enough to be able to drill in one variety at a time, and then fill in with the others afterwards, seems to be insuperable, unless a drill with spare boxes and arrangements for changing them quickly could be devised for the purpose. In any case, room for turning the horses would have to be left between the ends of the plots.

But besides the technical difficulties, it may be impossible to use the Latin square for lack of room, since the repetitions must equal the number of varieties or treatments to be tested. In such cases it is often possible to use equalized random blocks,* which enable the principle of the Latin square to be used without the very large number of repetitions. For example, the writer was able to suggest an arrangement to a horticultural experimenter who wished to compare ten treatments with five plots of each. He was anxious to use the Latin square, but realized that if he used two Latin squares the two sets of five treatments would not be properly comparable.

The proposed arrangement included five randomized blocks, but, whereas the first was completely random, each further successive block had its randomness more and more controlled, just as each successive row in a Latin square.

It will be seen that each column can equally be considered a "block", and that with one small exception it is as "equalized" as a Latin square: a fertility slope, therefore, either in the direction of the rows or of the columns, does not introduce errors, and the error of a comparison will be correspondingly reduced. The exception is that owing to there being an odd number of blocks, *A*, *D*, *E*, *F* and *J*

G *F*	*H* *D*	*E* *J*	*C* *B*	*A* *I*	Block I
H *B*	*J* *G*	*D* *I*	*F* *A*	*E* *C*	Block II
E *J*	*I* *B*	*A* *C*	*G* *H*	*D* *F*	Block III
C *A*	*F* *E*	*B* *G*	*I* *D*	*J* *H*	Block IV
D *I*	*A* *C*	*F* *H*	*J* *E*	*B* *G*	Block V
Block 1	Block 2	Block 3	Block 4	Block 5	

FIG. 3

* I have seen no account of work planned in this way, but it is an obvious application of Fisher's methods.

occur in the top row of their block three times and in the lower row twice, and vice versa with the others.

To sum up, the following methods of testing yields have been described:

On the small scale:

(*a*) Beaven's chessboard.

(*b*) Rod rows.

On the large scale:

(*c*) Beaven's half-drill strip.

Applicable to either large or small scale:

(*d*) Fisher's randomized blocks.

(*e*) Fisher's Latin square and its modification, equalized random blocks.

In the foregoing all reference to two methods—namely, the use of "control" plots, and the estimation of yield from samples taken from the plots instead of by harvesting the whole plots—has been omitted.

The former method was never very satisfactory (R. Summerby, "Accuracy in field experiments", *J. Amer. Soc. Agron.* vol. XVII, No. 3), and it was quite usual for "corrections" based on the "control" plots to increase the error of comparisons: it has now been superseded by the methods outlined above.

The latter method, on the other hand, has not yet been fully worked out, though it appears likely that in some cases it will become the ordinary way of estimating yield (F. L. Engledow, "A census of an acre of corn", *J. Agric. Sci.* XVI (1926), p. 191; A. R. Clapham, "The estimation of yield in cereal crops by sampling methods", *J. Agric. Sci.* XIX, p. 214; J. Wishart and A. R. Clapham, "A study in sampling technique: the effect of artificial fertilisers on the yield of potatoes", *J. Agric. Sci.* XIX, p. 600). Care should, of course, be taken to discount any sources of error, such as loss of corn from shattered ears, which may take place in one variety on the large scale, but not in samples cut by hand.

Statistical Interpretation of Results

For an adequate exposition of the methods of statistical analysis the reader is referred to treatises on the subject (R. A. Fisher, *loc. cit.*), but an indication of how it comes about that we must invoke the aid of mathematics to make the most of our experiments may not be amiss.

To take a very simple case: Suppose *A* and *B* are compared in 1927 and 1928, also *C* and *D*, and the following results obtained:

Year	Yield per acre in cwt.			
	A	*B*	*C*	*D*
1927	20	23	20	29
1928	19	24	21	20

In each case there is an average difference of 4 cwt. between the pairs, yet most people would probably conclude instinctively that more reliance could be placed on the comparatively concordant differences between *A* and *B* than on the discordant results for *C* and *D*. We therefore tend to give weight to *concordance*.

Again, suppose a further experiment in 1929 gave $\frac{A}{22}\frac{B}{26}$, we should probably feel satisfied that our conclusion that *B* is better yielding than *A* is strengthened; we therefore put more reliance on *increased repetition of experiments*. But suppose that instead we had repeated the *C D* comparison and obtained $\frac{C}{18}\frac{D}{22}$, would the three *C D* comparisons which include the discordant 1928 result be better than the two *A B* comparisons, concordant though they were?

Clearly we cannot answer questions of this kind by unaided common sense, but fortunately mathematicians have dealt with the evidential value of events of this nature, and we can use the methods and tables of the theory of chance, provided always we make certain that the fundamental condition of applying the theory—namely, that *the events with which we deal are random*—is adequately satisfied.

Randomness. In view of this proviso it is important to have a clear idea of what constitutes randomness, and this is by no means easy.

In our particular case a series of yields would be random if the value of each of them in relation to those of the others were quite independent of its position in time and place relative to them.

As mentioned above, this does not happen in practice; yields of plots situated close to one another are more alike than those far apart, and in particular there is a general tendency for yields to increase or decrease as we go from one end or side of the experimental area to the other. It is therefore necessary to arrange the positions in such a way as to superpose randomness upon the biased fertility of the soil. This may be done in two ways: either the positions may be assigned by one of the recognized methods of invoking chance—dice throwing, coin tossing, card drawing, and the like—or a regular pattern may be devised which will equalize the more probable variations in fertility, but which will yet be sufficiently complicated for it to be a matter of chance how the residual variations may affect any particular comparison.

Thus in Fisher's randomized blocks each treatment is repeated once per block, ensuring that each shall be equally affected by such variation as is common to the block, but the position within the block is determined by chance; similarly, the variation in fertility common to the plots which make up any row or column of a Latin square is equally shared, but the positions in the row or column are determined by chance.

Beaven's chessboard system, on the other hand, depends on a regular arrange-

ment, but one which is sufficiently complicated for the claim to be made that, by imposing it upon the ordinary variation in the soil, we get in fact a randomness in the residual variation from the mean of each small group of consecutive plots, which enables us to take advantage of mathematical analysis. Nevertheless, we must be careful to avoid arrangements in which the mean position of the different strains is not the same for all, and also such an order as:

A	*E*	*D*	*C*	*B*
B	*A*	*E*	*D*	*C*
C	*B*	*A*	*E*	*D*
D	*C*	*B*	*A*	*E*
E	*D*	*C*	*B*	*A*

where a possible crest or trough of fertility parallel to the diagonal of *A*'s might improve or depress the yield of one variety, without any warning being given by the calculation of a large error of observation from the observations.

Beaven's half-drill strip is, essentially, in a rather different position. Its pattern is of the simplest, it can only vary between *A B B ... B B A* and *A B B ... A A B*, of which the former must be chosen in case there may be, as is probable, a "fertility slope" across the drills. The extreme length and narrowness, however, ensure that the difference between adjacent half-drill strips is otherwise random on ordinary soils, but it is necessary to avoid possible periodic variations in fertility parallel to the drills. Thus, if the field had been laid up in "lands", and the drills were of such a width that the bottom of the land was always occupied by the same strain, a non-random system of error would be included which would vitiate the result. Similarly, a periodic variation in fertility might be left by previous cultivation, and to avoid this, Dr Hilgendorf of the Canterbury Agricultural College, New Zealand, drills diagonally across previous cultivation.

Analysis of Variance. With suitable precautions, then, all these arrangements give results which can be treated by the methods of the Theory of Chance, and, as it happens, the particular method ("The Analysis of Variance") introduced by Fisher, primarily for this purpose (R. A. Fisher, *loc. cit.*), can be applied to all of them, and I have limited my discussion of the Theory of Chance to such considerations as seem to me to be necessary to an understanding of this method.

If, then, an experiment is repeated several times, we get as a rule as many different results, though by chance some may be identical. If these results are random, we may attach more weight to their mean value the more numerous and the more concordant they are.

The measure of the weight to be given to a Mean is the Standard Deviation (S.D. or σ), which is derived as a rule from the results themselves by the following procedure:

Taking the difference between each result and the Mean, it is squared, and the sum of these squared differences is divided by the number of experiments less one. The quotient is called the "Variance" of the results, and the square root of the Variance is the Standard Deviation.

The Variance of the Mean is obtained by dividing the Variance of the results by their number, and, as before, the Standard Deviation of the Mean is the square root of its Variance.

In algebraic notation, if $x_1, x_2, \ldots x_n$ be n experimental results, and $\bar{x}$ their mean, then

the Variance $= \dfrac{S(x-\bar{x})^2}{n-1}$,

the Standard Deviation, σ, $= \sqrt{\left(\dfrac{S(x-\bar{x})^2}{n-1}\right)}$,

the Variance of the Mean $= \dfrac{S(x-\bar{x})^2}{n(n-1)}$,

and the Standard Deviation of the Mean

$$= \sqrt{\left(\frac{S(x-\bar{x})^2}{(n(n-1)}\right)}.$$

Having obtained the S.D., we can, by referring to tables constructed for this purpose, find the chance that the mean of an infinite number of repetitions under the same conditions would differ by less than any given amount from the mean of the few experiments which we have made. Thus a difference as large as, or larger than, once the S.D. of the mean of a large number of results occurs 16 times in 100 such series of experiments; as large as, or larger than, twice the S.D. about 2·3 in 100; as large as, or larger than, three times the S.D. only about 0·13 times in 100, and thereafter its rarity increases very rapidly.

We can thus judge of the value of our evidence, and as in other matters, the degree of accuracy which we demand will depend on the importance of the action to be taken relative to the cost of repeating the experiments.

For many purposes a probability of twenty to one is considered sufficient to justify drawing a conclusion, and a result which leads to such a probability is often conventionally called "significant". This corresponds to a quantity 1·65 times the S.D., when the S.D. is known accurately.

Now it can be shown that, provided randomness has been observed, variances are additive. If one set of causes, say the innate differences in fertility between the strains which are being tested (variance V_s), act simultaneously with another set of causes, say random errors of the plots which are being tested (variance V_e), then if V_t be the total variance of the yields,

$$V_t = V_s + V_e.$$

Similarly, if we can arrange the plots as in randomized blocks, or the sets of Beaven's chessboard, so that part of the variation is common to the blocks or sets (variance V_b) and part random (V_e as before), then

$$V_t = V_s + V_b + V_e.$$

Or in the Latin square:

$$V_t = V_s + V_{\text{rows}} + V_{\text{columns}} + V_e.$$

In any of these cases it is the differences between the strains which make up V_s about which we have to form a judgment, and it is $\sigma_s = \sqrt{V_s}$ with which we have to measure the certainty.

Now V_t is calculable, it is the total variance of the yields, and V_s, V_b, etc., are the variances of the means of the strains, of the means of the blocks, etc., so that we can find V_s by difference.

Degrees of Freedom. But before giving an example of the determination of V_s by this method, it is necessary to introduce the reader to one other technicality.

It will have been noticed that in calculating the variance the sum of the squares was divided not by n the number of results, but by one less than that number. The reason for this is that since we do not know what would be the mean value of an infinite number of results obtained under like conditions, we are driven to use the mean of the n results which we have, and it can be shown that this would necessarily give too low a result were we to divide by n. We are on the average right if we diminish that number by one.

Now for any given mean it is only possible to vary $n-1$ of the results, the last one is fixed by the mean and the other $n-1$; hence, there are said to be $n-1$ *degrees of freedom.* If in addition to the mean of the whole we also calculate from the mean of a group, say the yield of the plots of a given strain, the number of results which can vary is again diminished by one, there are now but $n-2$ degrees of freedom, and similarly for each such mean. But it should be noticed that if the general mean and the means of all the strains but one are calculated, the remaining one is now fixed, i.e. the means of the strains, too, have a degree less freedom than the number of the strains.

In this way the original $n-1$ degrees of freedom may be allotted to the different variances with which we are dealing, each variance accounting for one less than the number of categories from which it is calculated, and the balance is left for the calculation of the random variation. Thus, in a Latin square in which five strains are tested in twenty-five plots there are

5 strains	taking up	4	degrees of	freedom.
5 rows	,,	4	,,	,,
5 columns	,,	4	,,	,,

So that of the original 24 degrees of freedom only 12 are left for the calculation of the variance due to random error.

We are now in a position to calculate a very simple numerical example.

Suppose we had arranged plots of four strains (A, B, C and D) in a Latin square as follows:

A	*B*	*C*	*D*
C	*D*	*A*	*B*
D	*C*	*B*	*A*
B	*A*	*D*	*C*

and that the following yields had been obtained:

5	8	4	2
5	4	2	5
3	6	3	5
5	2	1	4

Can we say that the yield of B under the conditions of the experiment is significantly better than that of D?

For simplicity the above yields have been chosen so that the general mean is a whole number, 4, and in working we may rewrite the yields as differences from that number, thus:

	Sums of rows	Averaged	Average
		Also:	
+1 +4 0 −2	+3	$+\frac{3}{4}$	$A = +1-2+1-2 = -2$	$-\frac{1}{2}$
+1 0 −2 +1	0	0	$B = +4+1-1+1 = +5$	$+1\frac{1}{4}$
−1 +2 −1 +1	+1	$+\frac{1}{4}$	$C = 0+1+2+0 = +3$	$+\frac{3}{4}$
+1 −2 −3 0	−4	−1	$D = -2+0-1-3 = -6$	$-1\frac{1}{2}$
Sums of Columns +2 +4 −6 0	0			
Average $+\frac{1}{2}$ +1 $-\frac{3}{2}$ 0				

The *total variance* then is the sum of the squares divided by the 15 degrees of freedom; or fifteen times

$$V_t = (1+16+0+4+1+0+4+1+1+4+1+1+1+4+9+0) = 48.$$

The contribution made to this by the variance of the strains is $\frac{1}{4}$ for each A, $\frac{25}{16}$ for each B, and so forth, or

$$4\left\{\frac{1}{4}+\frac{25}{16}+\frac{9}{16}+\frac{9}{4}\right\} = 18\tfrac{1}{2}.$$

It will be seen that this result can be arrived at more easily by taking $\frac{1}{4}$ of the squares of the sums, i.e.

$$\tfrac{1}{4}(4+25+9+36) = 18\tfrac{1}{2}.$$

Similarly, the contribution made by the variance of the rows is

$$\tfrac{1}{4}(9+0+1+16) = 6\tfrac{1}{2}.$$

And the contribution made by the variance of the columns is

$$\tfrac{1}{4}(4+16+36+0) = 14.$$

These facts are then set out in a table as follows:

Variance due to	Degrees of freedom	Sums of squares	Variance	Standard deviation
Strains	3	$18\frac{1}{2}$	6·17	—
Rows	3	$6\frac{1}{2}$	2·17	—
Columns	3	14	4·67	—
Random error	6	9	1·5	1·225
Total	15	48	3·2	—

the degrees of freedom and the sum of the squares due to random error being obtained by differences between the total and the sum of the other three; the variance is obtained by dividing the third column by the second, and the standard deviation by taking the square root of the variance.

Now the average difference between B and D is $2\frac{3}{4}$, and the random variance of each mean is 1·5/4, but since we are dealing with the difference between the two, the variance of this difference is twice this or 0·75, and the standard deviation is $\sqrt{0{\cdot}75}$, or 0·866. The difference between B and D is, therefore, 2·75/0·866 = 3·17 times the standard deviation, and is to be looked out opposite $t = 3{\cdot}1/3{\cdot}2$ in the column headed by 6 degrees of freedom.

With standard deviations calculated from so few degrees of freedom, "Student's tables must be used: these tables are given in full in *Metron*, vol. III (1925) [**12**]; an abstract is given in Fisher's *Statistical Methods for Research Workers*, 1st ed. p. 137, and they are also given in a somewhat less convenient form in *Biometrika*, XI, p. 416 [**8**], and *Tables for Statisticians and Biometricians*, 2nd ed. p. 63. If using the last two, the 3·17 must be divided by the square root of one more than the degrees of freedom, and looked out under the column headed by the same number, i.e. we look out $3{\cdot}17/\sqrt{7}$.

In any case we find the probability of obtaining a smaller difference by chance to be about 0·99—i.e. it is 99 to 1 against getting such a large one—and we may therefore suppose that the difference between B and D would again come out in favour of B, if B and D were grown again under similar conditions.

On the other hand, the difference between C and A is only 1·25/0·866 times its standard deviation, say 1·45, and on looking this up we find the probability of obtaining a smaller result by chance is only 0·9, i.e. the odds are only 9 to 1 against getting such a large difference, and we cannot conclude that the difference between C and A is due to the strains and not to the positions of the plots in which they are grown.

It should be noticed, however, that the formula which we have used in comparing B and D is that which it is correct to use when there are but two means to compare: it would be right to use it, for example, if we have a number of trials

of B and D in different places or seasons, and we wish to examine the whole series of comparisons of B and D.

If, however, there is no particular reason why we should compare these two rather than any other pair, it is clear that the chance of obtaining a large difference between some pair or other is greater the larger the number of possible pairs. To meet this Fisher suggests that the mean of each strain should be compared with the general mean, and that the strains should be divided in this way into

(*a*) those significantly greater than the mean;
(*b*) those not differing significantly from the mean; and
(*c*) those significantly less than the mean.

The appropriate standard deviation to use, if σ be the standard deviation of the random error of a single plot, and there are n repetitions and m strains, is

$$\frac{\sigma\sqrt{(m-1)}}{\sqrt{(n\,m)}}.$$

In this case $n = m = 4$, and the standard deviation is therefore $1{\cdot}225 \times \frac{\sqrt{3}}{4} = 0{\cdot}53$.

Referring to the table, we find that the 20 : 1 limit corresponds to twice the standard deviation ($t = 2{\cdot}0$) for 6 degrees of freedom, and accordingly we may divide the four strains as follows:

B, significantly better than the mean;
A and *C*, not different significantly from the mean; and
D, significantly worse than the mean.

The above example, though "made up", illustrates what commonly happens in practice—namely, that the variance of the rows and of the columns which we have neutralized in the arrangement of the Latin square is in each case greater than the Random Error, and we have therefore increased the precision of our experiment by this arrangement.

It is not usual, however, to have an exact figure as the mean, and the following example of a half-drill strip experiment which was actually carried out at Ballinacurra, Co. Cork, in 1929, gives the procedure when measuring not from the mean, but, to avoid working with fractions, from some arbitrarily chosen origin. This trial compared a selection from Dr Hunter's Spratt-Archer barley with his selected Archer in twenty-two half-drill strips each. It will be noticed that when Spratt-Archer was on the *north*, there was practically no difference, but that it was markedly better when on the *south*; thus there was a fertility slope across the strips, and the two sets should be averaged separately at the loss of a degree of freedom.

Naturally with only two varieties to compare we do not concern ourselves with anything but the differences between corresponding strips, and to avoid fractions, these differences are measured in $\frac{1}{4}$-lb. units. The mean is here fractional, and to save arithmetic, an arbitrary point is chosen as origin. Now any arbitrary

origin may be chosen, but since the largest difference is +40 and the smallest −24, the obvious origin is zero. [A useful exercise for the beginner would be to take another origin (say +10), measure each difference from this, and work out the example again; the same result should be obtained, and to facilitate an exercise of this kind the figures which should be identical in any such comparison are given in italics.

Thus, if +10 were chosen the differences would run:

+4 +25 −34 −6 −16, and so on.]

Table of Yields

	Yields		Difference, SA.-A.	
	Spratt-Archer 37, No. 3	Archer	Spratt-Archer on North	Spratt-Archer on South
	lb.	lb.	¼ lb.	¼ lb.
	36½	33	+14	
	41½	32¾		+35
	34	40	−24	
	40	39		+ 4
	37	38½	− 6	
	37½	36½		+ 5
	38½	35	+14	
	42¼	38½		+15
	41	42½	− 6	
	43	40		+12
	42¼	42	+ 1	
	42	38½		+14
	41	39¼	+ 7	
	45	38½		+26
	44½	42	+ 9	
	42¼	40⅜		+ 7
	40	39½	+ 2	
	41	39⅜		+ 6
	41	39	+ 8	
	39½	41¼		− 7
	36	39¼	−14	
	43	33		+40
Sum	888½	848	+ 5	+157
			+162	
Average	40·4	38·5	+ 7·364	

I owe these figures to the courtesy of the Irish Free State Department of Agriculture.

We have first the sum of the squares of all differences:

196 + 1225 + 576 + 16 + 36 + 25 + 196 + 225 + 36 + 144 + 1 + 196
+ 49 + 676 + 81 + 49 + 4 + 36 + 64 + 49 + 196 + 1600 ... = 5676

To correct for the arbitrary origin we subtract 22 times the square of the mean distance from the origin, i.e. $22 \times 7{\cdot}364^2$ = 1193

4483

The sum of the squares due to the North/South fertility slope is:

$\frac{1}{11}(5^2 + 157^2)$ ... ... = 2243

Subtracting the same correction for the arbitrary mean ... = 1193

We get *1050*

We now arrange as before:

Variance due to	Degrees of freedom	Sums of squares	Variance	Standard deviation
North/South fertility slope	1	*1050*	*1050*	—
Random error	20	*3433*	*171·65*	*13·1*
Total	21	*4483*	*213·5*	—

It will be observed that all the figures in this table are the same no matter what arbitrary origin is chosen, and this method is the same no matter how many different sources of variance are accounted for; the same correction (the square of the mean multiplied by the *total* number) is subtracted from the sum of squares due to each source and from the total.

The mean difference is then 7·364 in favour of Spratt-Archer 37, No. 3, and the variance of the comparison is $\frac{171{\cdot}65}{22}$, giving a standard deviation of 2·79 (about 1·7 %), and we have

$$t = \frac{7{\cdot}364}{2{\cdot}79} = 2{\cdot}64.$$

Looking this out in the table under 20 degrees of freedom, we find $P = 0{\cdot}9921$; i.e. we should get a result in favour of Spratt-Archer 37, No. 3, as large as this by chance if there were really no difference only 79 times in 10,000 trials, and we may conclude that under the conditions of the trial Spratt-Archer is definitely the higher yielding barley.

The following table gives the variances which are removed in finding the random error in the various methods described when calculating the degree of significance of the results by the analysis of variance:

Chessboard (balanced panels as recommended)	Half-drill strips	Randomized blocks	Latin square	Equalized randomized blocks
Variance of strains	Variance of strains	Variance of strains	Variance of strains	Variance of strains
Variance of sets		Variance of blocks	Variance of rows	Variance of blocks "across"
Variance of fertility slope in balanced direction	Fertility slope across strips		Variance of columns	Variance of blocks "down"

General Observations

While it is obvious that there is much to be gained by planning yield trials in such a way as both to reduce the experimental error and to obtain an accurate estimate of it, *it is important to remember that conclusions can only be drawn applicable to the particular conditions under which the trials were carried out.*

For this reason, trials should be repeated season after season, and in so many different places as to cover the probable variations in soil and climate in the districts in which practical application is to be made.

When this has been done, the accuracy which our technique has enabled us to obtain will enable us to analyse the results, so as to find out whether some of our strains/treatments suit certain soils/seasons/climates better than others.

Moreover, it is not enough to show that one strain/treatment is superior to another in yield; to achieve lasting success we must subject our product to tests for quality.

Thus, the Irish barley trials culminated in malting and brewing tests, as also those now being conducted by the National Institute of Agricultural Botany; Biffen's wheats were chosen for their baking strength, as well as yield, and so on. On the other hand, accounts of yield trials of potatoes rarely conclude with a table of moisture percentages and an estimate of relative palatability (F. Johnson and O'C. Boyle, "The industrial and nutritive value of the potato in Ireland", *J. Dept. Agric. for Ireland*, vol. XVIII, No. 4). This may help to account for the modern potato.

General References

G. Udny Yule, *Introduction to the Theory of Statistics*, Griffin and Co., London; F. L. Engledow and G. Udny Yule, "The principles and practice of yield trials", *The Empire Cotton Growing Review*, vol. III, Nos. 2 and 3, Empire Cotton Growing Corporation, Millbank, London; Fisher and Wishart, *Imp. Bur. Soil Science, Tech. Comm.* No. 10, "The arrangement of field experiments and the statistical reduction of the results".

An excellent bibliography of papers, etc. is contained in W. Horton Beckett's "Methods of field experimentation", 1928 *Year Book of the Dept. of Agric.*, Gold Coast.

16

THE LANARKSHIRE MILK EXPERIMENT

[*Biometrika*, XXIII (1931), p. 398]

In the spring of 1930* a nutritional experiment on a very large scale was carried out in the schools of Lanarkshire.

For four months 10,000 school children received $\frac{3}{4}$ pint of milk per day, 5000 of these got raw milk and 5000 pasteurized milk, in both cases Grade A (Tuberculin tested); another 10,000 children were selected as controls and the whole 20,000 children were weighed and their height was measured at the beginning and end of the experiment.

It need hardly be said that to carry out an experiment of this magnitude successfully requires organization of no mean order and the whole business of distribution of milk and of measurement of growth reflects great credit on all those concerned.

It may therefore seem ungracious to be wise after the event and to suggest that had the arrangement of the experiment been slightly different the results would have carried greater weight, but what follows is written not so much in criticism of what was done in 1930 as in the hope that in any further work full advantage may be taken of the light which may be thrown on the best methods of arrangement by the defects as well as by the merits of the Lanarkshire experiment.

The 20,000 children were chosen in 67 schools, not more than 400 nor less than 200 being chosen in any one school, and of these half were assigned as "feeders" and half as "controls", some schools were provided with raw milk and the others with pasteurized milk, no school getting both.

This was probably necessary for administrative reasons, owing to the difficulty of being sure that each of as many as 200 children gets the right kind of milk every day if there were a possibility of their getting either of the two. Nevertheless, as I shall point out later, this does introduce the possibility that the raw and pasteurized milks were tested on groups of children which were not strictly comparable.

Secondly, the selection of the children was left to the head teacher of the school and was made on the principle that both "controls" and "feeders" should be representative of the average children between 5 and 12 years of age: the actual method of selection being important I quote from Drs Leighton and McKinlay's*

* Department of Health for Scotland: *Milk Consumption and the Growth of Schoolchildren*, by Dr Gerald Leighton and Dr Peter L. McKinlay (Edinburgh and London: H.M. Stationery Office, 1930).

Report: "The teachers selected the two classes of pupils, those getting milk and those acting as 'controls', in two different ways. In certain cases they selected them by ballot and in others on an alphabetical system." So far so good, but after invoking the goddess of chance they unfortunately wavered in their adherence to her for we read: "In any particular school where there was any group to which these methods had given an undue proportion of well-fed or ill-nourished children, others were substituted in order to obtain a more level selection." This is just the sort of after-thought that most of us have now and again and which is apt to spoil the best laid plans. In this case it was a fatal mistake, for in consequence the "controls" were, as pointed out in the Report,* definitely superior both in weight and height to the "feeders" by an amount equivalent to about 3 months' growth in weight and 4 months' growth in height.

Presumably this discrimination in height and weight was not made deliberately, but it would seem probable that the teachers, swayed by the very human feeling that the poorer children needed the milk more than the comparatively well to do, must have unconsciously made too large a substitution of the ill-nourished among the "feeders" and too few among the "controls" and that this unconscious selection affected, secondarily, both measurements.

Thirdly, it was clearly impossible to weigh such large numbers of children without impedimenta. They were weighed in their indoor clothes, with certain obvious precautions, and the difference in weight between their February garb and their somewhat lighter clothing in June is thus necessarily subtracted from their actual increase in weight between the beginning and end of the experiment. Had the selection of "controls" and "feeders" been a random one, this fact, as pointed out in the Report,* would have mattered little, both classes would have been affected equally, but since the selection was probably affected by poverty it is reasonable to suppose that the "feeders" would lose less weight from this cause than the "controls". It is therefore not surprising to find that the gain in weight of "feeders" over "controls", which includes this constant error, was more marked, relatively to their growth rate, than was their gain in height, which was fortunately not similarly affected.

Fourthly, the "controls" from those schools which took raw milk were bulked with those from the schools which took pasteurized milk.

Now with only 67 schools, at best 33 against 34, in a district so heterogeneous both racially and socially, it is quite possible that there was a difference between the averages of the pupils at 33 schools and those of the pupils at another 34 schools both in the original measurements and in the rate of growth during the experiment.

In that case the average "control" could not be used appropriately to compare with either the "raw" group or the "pasteurized" group.

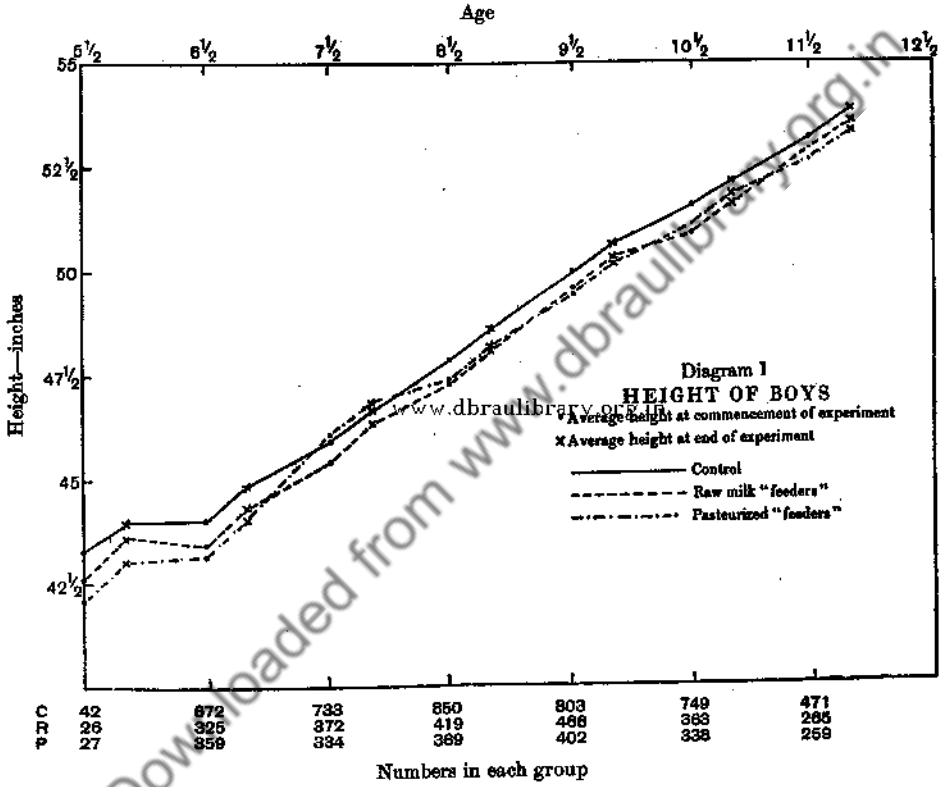
This possibility is enhanced by the aforementioned selection of "controls"

* See footnote on p. 169.

which can hardly have been carried out in a uniform manner in different schools.

Fortunately it would still be possible to correct this, for the figures for the different schools must still be available in the archives.

Diagrams 1 and 2 give the average heights of "controls", raw milk "feeders" and pasteurized milk "feeders" for boys and girls respectively. The heights at the beginning of the experiments are set out against a uniform age scale centring

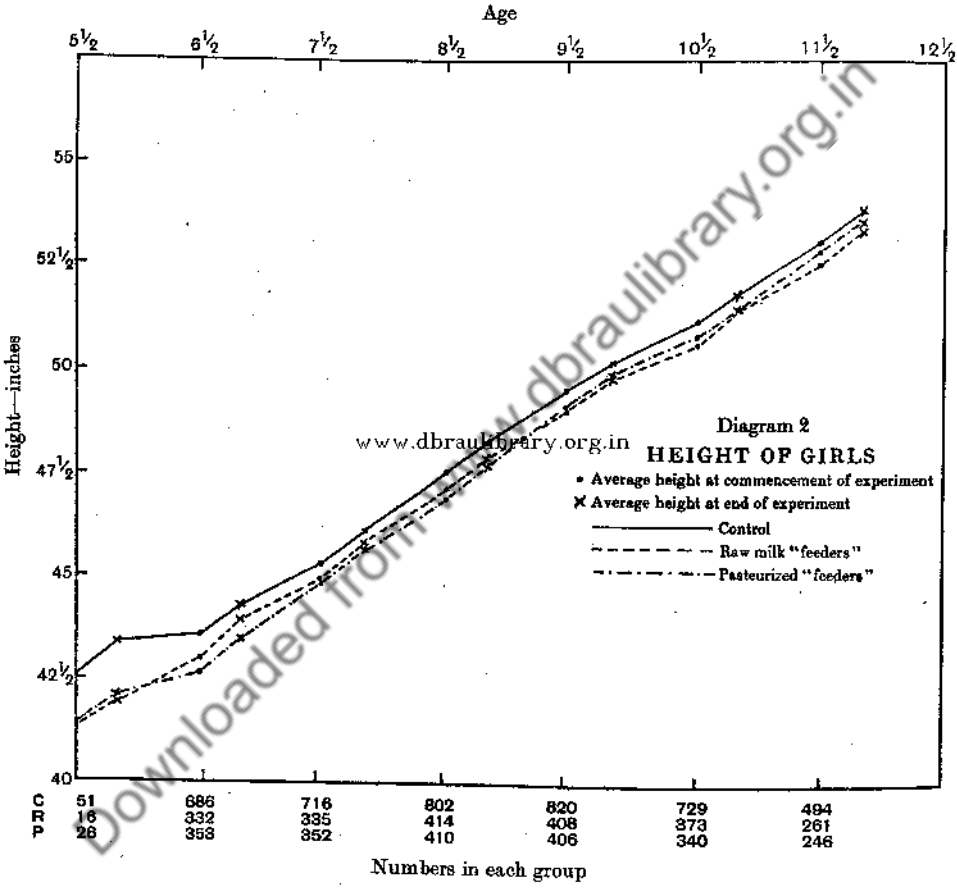


each group at the half year above the whole number. This is doubtless accurate enough except for the first group aged "5 and less than 6", which was very much smaller in numbers than the other groups, either because only the older (or larger) children are sent to school between 5 and 6 or because the teachers did not think that the smaller children would be able to play their part. For this reason they should probably be centred more to the right compared to the others. A similar argument might lead us to centre the "11 and over" group a little more to the left.

The average heights at the end of the experiment are of course set out four months to the right of those at the beginning and it will be noticed that except for the first group, which is clearly out of place, not any of the points diverge very

much from their appropriate line of growth whether "controls", "raws" or "pasteurized".

The case is very different in Diagrams 3 and 4 which show the corresponding average weights. Here there is, after the first two ages, a very decided dip, especially in the later ages. The weights at the end of the experiment are too low. This might be accounted for by a tendency in older children to grow normally in

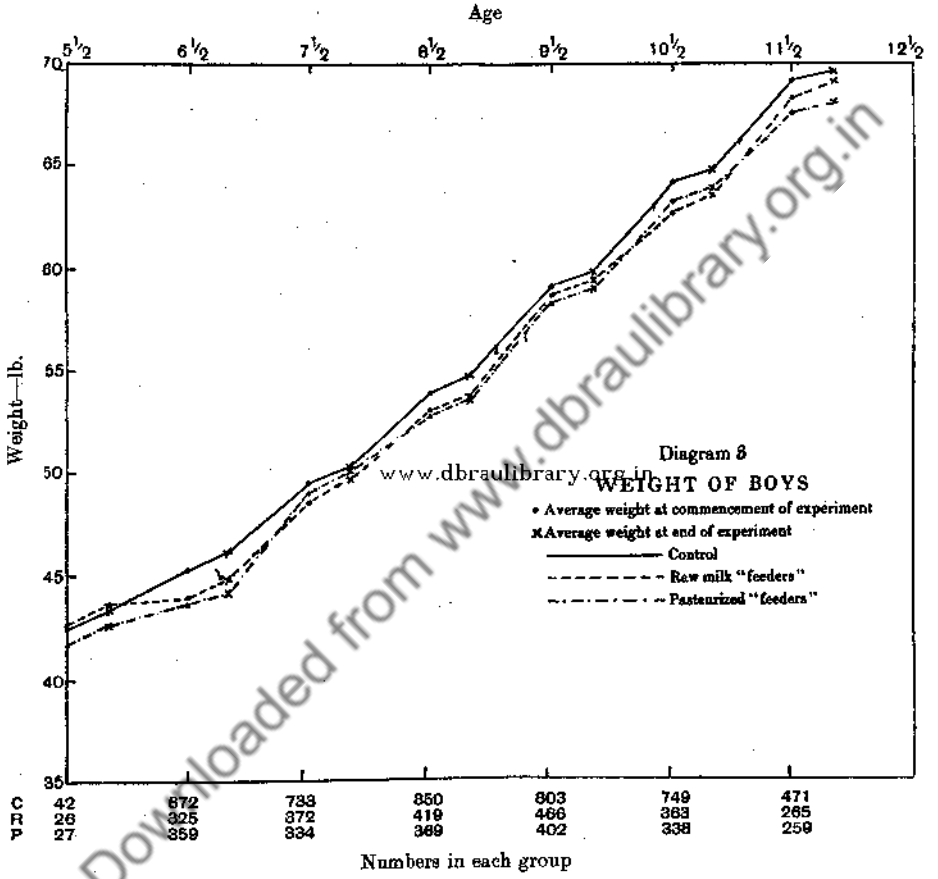


height and subnormally in weight during the spring, but I think it much more likely that older children wear about 1 lb. more clothes in February than they do in June, while in the case of younger children a more limited wardrobe permits of fewer discards.

The authors have tried to show that the selection of the "controls" has not affected the validity of the comparison, by computing the correlation coefficients between the original heights (and weights) and the growth during the experiment for each of the 42 age groups into which the measurements were divided. These

they find to be quite small even though they are here and there significant, and they argue that the additional height and weight of the "controls" was without effect on the comparison of subsequent growth.

Now this might have been a perfectly good argument had the height and weight been selected directly, but if, as I have indicated was very likely the case, the



selection was made according to some unconscious scale of well-being, then it is surely natural to suppose that the relatively ill-nourished "feeders" would benefit more than their more fortunate school mates, the "controls", would have done by the extra ¾ pint of milk per day.

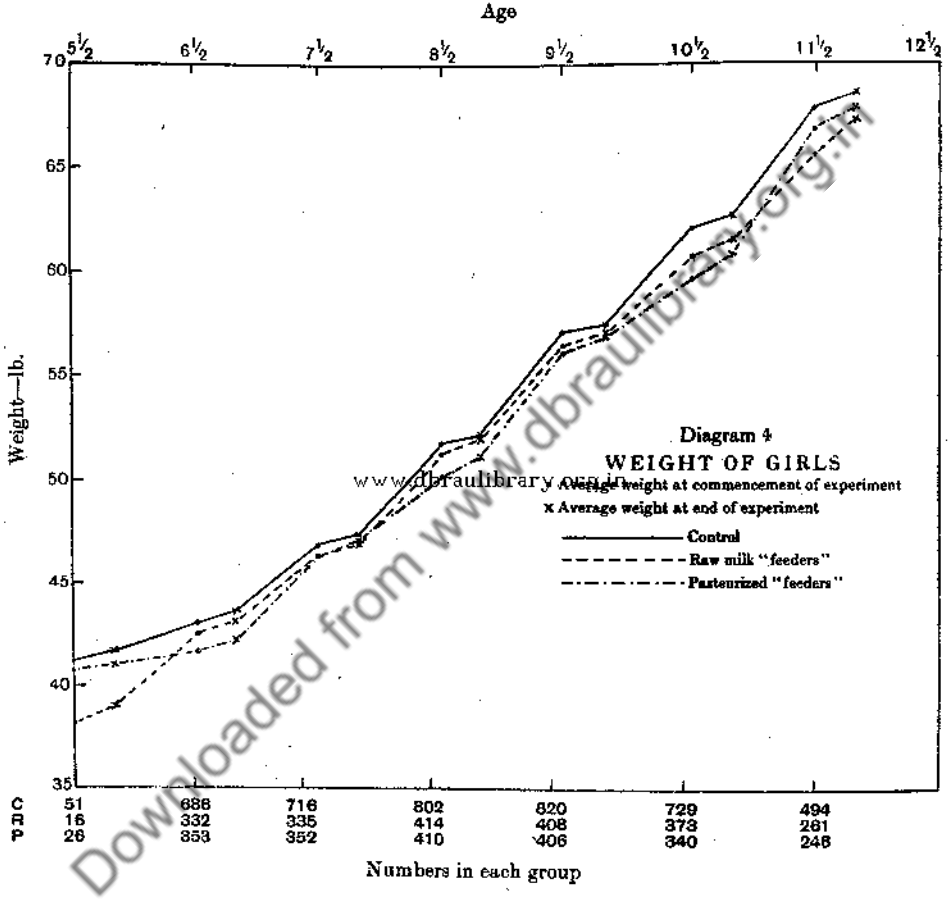
That being so, how are we to regard the conclusions of the Report:*

(1) "The influence of the addition of milk to the diet of school children is reflected in a definite increase in the rate of growth both in height and weight."

This conclusion was probably true; the average increase for boys' and girls'

* See footnote on p. 169.

heights was 8 % and 10 % over "controls" and for boys' and girls' weights was 30 % and 45 % respectively, and though, as pointed out, the figures for weights were wholly unreliable it is likely enough that a substantial part of the difference in height and a small part of that in weight were really due to the good effect of the milk. The conclusion is, however, shifted from the sure ground of scientific



inference to the less satisfactory foundation of mere authority and guesswork by the fact that the "controls" and "feeders" were not randomly selected.

(2) "There is no obvious or constant difference in this respect between boys and girls and there is little evidence of definite relation between the age of the children and the amount of improvement. The results do not support the belief that the younger derived more benefit than the older children. As manifested merely by growth in weight and height the increase found in younger children through the addition of milk to the usual diet is certainly not greater than, and is probably not even as great as, that found in older children."

Now from the authors' point of view, believing in the validity of their comparisons in weight, this is much understating the case, as the following table derived from Capt. Bartlett's condensed tables* shows:

Age in years	Gain in weight in ounces by feeders over controls		Gain in height in inches by feeders over controls		As % of control			
					Weight		Height	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
5, 6 and 7	1·13 ± 0·73	1·24 ± 0·72	0·083 ± 0·011	0·059 ± 0·011	9	13	11	8
8 and 9	3·15 ± 0·68	4·47 ± 0·67	0·071 ± 0·011	0·098 ± 0·010	30	51	10	14
10 and 11	5·21 ± 0·85	7·88 ± 0·79	0·037 ± 0·012	0·055 ± 0·012	78	73	5	8

Note that the P.E.'s are calculated from Capt. Bartlett's tables and are subject, as his are, to his having interpreted the methods of the original Report correctly.

From this they might have concluded:

(*a*) That in the matter of weight older children, both boys and girls, derived more benefit than younger, while

(*b*) In height the younger boys did better than the older, though the difference is not quite significant, but that there was no regular tendency in the matter of girls' height.

In the light of previous criticism, however, we must be content to say that apparently the differential shedding of clothes between the "feeders" and the more fortunate "controls" is more marked with older children (and possibly with girls than with boys), and that there is some probability that younger boys gain in height more than older.

Finally, conclusion (3) runs: "In so far as the conditions of this investigation are concerned the effects of raw and pasteurized milk on growth in weight and height are, so far as we can judge, equal."

This conclusion has been challenged by Capt. Bartlett,* and by Dr Fisher and Capt. Bartlett,† who conclude that there is definite evidence of the superiority of raw over pasteurized milk in both height and weight.

Even they, however, point out that the raw and pasteurized milk were not supplied to the same schools, and their conclusion amounts to saying: "If the groups of children taking raw and pasteurized milk respectively were random samples from the same population, the observed differences would be decisively in favour of the raw milk."

Unfortunately they were not random samples from the same population: they were selected samples from populations which may have been different, and moreover the "controls" with which they were compared were not appropriate to

* "Nutritional value of raw and pasteurized milk", by Stephen Bartlett, M.C., B.Sc. (*J. Min. Agric*, April 1931).

† *Nature*, 18 April 1931, p. 591, "Pasteurized and raw milk".

either group; and so—again it is a matter of guess and authority—I would be very chary of drawing any conclusion from these small biased differences.

That is not to say that there is no difference between the effect of raw and pasteurized milk—personally I believe that there is and that it is in favour of raw milk—but that this experiment, in spite of all the good work which was put into it, just lacked the essential condition of randomness which would have enabled us to prove the fact.

This note would be incomplete without some constructive proposals in case it should be considered necessary to do further work upon the subject, and accordingly I suggest the following:

(1) If it should be proposed to repeat the experiment on the same spectacular scale,

(*a*) The "controls" and "feeders" should be chosen by the teachers in pairs of the same age group and sex, and as similar in height, weight and especially physical condition (i.e. well- or ill-nourished) as possible, and divided into "controls" and "feeders" by tossing a coin for each pair. Then each pair should be considered to be a unit and the gain in weight and height by the "feeder" over his own "control" should also be considered as a unit for the purpose of determining the error of the gain in weight or height.

In this way the error will almost certainly be smaller, perhaps very much smaller, than if calculated from the means of "feeders" and "controls".

If in addition the social status of each pair be noted (well-to-do, medium, poorly nourished or some such scale) further useful information will be available for comparing pasteurized and raw "feeders".

If this is found to be too difficult a perfectly good comparison can be made by adhering to the original plan of the 1930 experiment and drawing lots to decide which should be "controls" and which "feeders" (this is better than an alphabetical arrangement), but the error of the comparison is likely to be larger than in the plan outlined above.

(*b*) If it is at all possible each school should supply an equal number of raw and pasteurized "feeders", again by selection of similar children followed by coin tossing, but I fear that this is a counsel of perfection.

(*c*) Some effort should be made to estimate the weight of clothes worn by the children at the beginning and end of the experiment: possibly the time of year could be chosen so that there would be little change in this respect.

(2) If it be agreed that milk is an advantageous addition to children's diet—and I doubt whether any one will combat that view—and that the difference between raw and pasteurized milk is the matter to be investigated, it would be possible to obtain much greater certainty at an expenditure of perhaps 1–2 % of the money* and less than 5 % of the trouble.

* This is a serious consideration: the Lanarkshire experiment cost about £7500.

For among 20,000 children there will be numerous pairs of twins; exactly how many it is not easy to say owing to the differential death-rate, but, since there is about one pair of twins in 90 births, one might hope to get at least 160 pairs in 20,000 children. But as a matter of fact the 20,000 children were not all the Lanarkshire schools population, and I feel pretty certain that some 200–300 pairs of twins would be available for the purpose of the experiment.

Of 200 pairs some 50 would be "identicals" and of course of the same sex, while half the remainder would be non-identical twins of the same sex.

Now identical twins are probably better experimental material than is available for feeding experiments carried out on any other mammals, and the error of the comparison between them may be relied upon to be so small that 50 pairs of these would give more reliable results than the 20,000 with which we have been dealing.

The proposal is then to experiment on all pairs of twins of the same sex available, noting whether each pair is so similar that they are probably "identicals" or whether they are dissimilar.

"Feed" one of each pair on raw and the other on pasteurized milk, deciding in each case which is to take raw milk by the toss of a coin.

Take weekly measurements and weigh without clothes.

Some way of distinguishing the children from each other is necessary or the mischievous ones will play tricks. The obvious method is to take finger-prints, but as this is identified with crime in some people's minds, it may be necessary to make a different indelible mark on a fingernail of each, which will grow off after the experiment is over.

With such comparatively small numbers further information about the dietetic habits and social position of the children could be collected and would doubtless prove invaluable.

The comparative variation in the effect in "identical" twins and in "unlike" twins should furnish useful information on the relative importance of "Nature and Nurture".

To sum up: The Lanarkshire experiment devised to find out the value of giving a regular supply of milk to children, though planned on the grand scale, organized in a thoroughly business-like manner and carried through with the devoted assistance of a large team of teachers, nurses and doctors, failed to produce a valid estimate of the advantage of giving milk to children and of the difference between raw and pasteurized milk.

This was due to an attempt to improve on a random selection of the "controls" which in fact selected as "controls" children who were on the average taller and heavier than those who were given milk.

The hypothesis is advanced that this was due not to a selection of the shorter, lighter children as such to take the milk, but to an unconscious bias leading the

teachers to pick out for this purpose the needier children whom the milk would be most likely to benefit.

This hypothesis is supported by the fact that while the advantage derived from the milk was only 8–10 % of the gain in height, without much variation for age, it was 30–45 % of the gain in weight, varying from 9 to 13 % in the younger children (who do not seem to have shed much clothing in the summer) up to 73–78 % in the older children—who obviously did.

Suggestions are made for the arrangement:

(1) Of a similar large-scale experiment on random lines, and

(2) Of a much smaller and cheaper experiment carried out on pairs of twins of like sex.

The second is likely to provide a much more accurate determination of the point at issue, owing to the possibility of balancing both nature and nurture in the material of the experiment.

17

ON THE "z" TEST

[*Biometrika*, XXIII (1931), p. 407]

In the last number of *Biometrika* Prof. Pearson sounds a warning note against the use of "Student's" z to determine the significance of an average difference between two sets of correlated variables.

As this use is one to which I attach considerable importance, and as Prof. Pearson's criticism does not all seem to me to be concerned with my own method of using z, I should like to present the case from the point of view of the experimenter who for some reason or another has to work with small samples.

In the great majority of experiments of this kind we are concerned with a *difference*: e.g. of yield between two varieties of a cereal; of weight between pigs fed on complete food and others on food deficient in vitamins; of size of loaf between breads baked from different flours; of reaction times between alcoholized and non-alcoholized persons, and so on.

Now it is an elementary principle in all such experiments to reduce the error of such differences by arranging that they should lie between variates which, apart from the experiment, are as similar as possible.

Thus each pair of cereal plots should be grown not only on the same field but as near as possible to each other in that field; the pigs should be of the same sex, from the same litter and as nearly as possible the same weight at the start; the loaves should be mixed at the same time and put next to one another in the oven, and the alcoholized and non-alcoholized persons should alternate their roles so as to compare each person with himself.

In other words, every care should be taken when planning the experiment to get the correlation between corresponding variates as high as possible, with the object of reducing the error and so of obtaining significant results from the small number of experiments which it is possible to carry out.

Now Prof. Pearson's criticism may be summarized under three heads:

(1) That, assuming the advisability of the "z" Test, it is only when the value of z which is obtained is high that we can draw any useful conclusion: if it is low it cannot detect samples which may be abnormal in other ways.

Agreed, but this inability to detect abnormalities extraneous to the test itself is shared with all single tests of significance and the result is that the wise man will never go further in the direction of asserting similarity than to say, "The sample affords no evidence that, etc."

(2) That, since we cannot deduce with accuracy the correlation of the population at large from the small sample before us, we are debarred from making use of that correlation to reduce our error.

But in fact we do *not* use the correlation in testing the significance. In obtaining our sample of difference, yes, but once we have the differences they are merely a sample from the indefinitely large population of differences, which might have been produced under similar conditions and I may say with the same ρ (not r) and which we, rightly or wrongly, assume to be normal. At all events Prof. Pearson is not here attacking z on the grounds of lack of normality. The correlation will vary from sample to sample, just as does the mean or standard deviation, but these variations in correlation do not affect the fact, which Prof. Pearson admits, that in a normal population z can be used to test the significance of the mean of small samples.

What we actually ask ourselves is the following question:

If the average difference between A and B in the population were zero, what would be the probability of obtaining a sample of differences giving a value of z as high as that observed? and if this probability is sufficiently small we say that the difference is significant.

(3) Prof. Pearson warns us to be careful to draw our conclusion from the experiment we have carried out and according to the particular set of differences which we have tested for significance.

In this of course I agree with him, yet I do not feel that the warning is very much needed. In the hyoscyamine experiment which he quotes, we are able to deduce significance from a consideration of the effects of these drugs on the same individuals while we could not do so from groups of different individuals. But surely no one is much interested in the latter point; if I am to take one of the drugs I will pay a good deal of attention to the probability that laevo will make me personally sleep longer than dextro and very little to the fact that the experiments give no satisfactory answer to the question of what will happen if I take laevo and my wife dextro.

To sum up, in properly planned experiments errors should be reduced as much as possible by the selection of highly correlated individuals to compare with one another. This correlation should to a greater or lesser extent reduce the variation in the differences between these individuals but does not prevent them being considered to be a sample drawn from a population of differences to which the "z" Test may be applied. Finally, care must be taken in planning the experiment that the differences to be examined for significance shall be those which furnish an answer to the question which we are asking.

18

EVOLUTION BY SELECTION

THE IMPLICATIONS OF WINTER'S SELECTION EXPERIMENT*

[*Eugenics Review*, XXIV (1933), p. 293]

FOR some time after the publication of the *Origin of Species* it was generally held by those who accepted Darwin's reasoning that species originated by the accumulation of small variations in the same direction under the influence of natural selection; and the occurrence of large "mutations", such as the Ancon sheep, was perhaps rather overlooked. The rediscovery of Mendelism, however, has tended to emphasize the latter portion of Darwin's work, rather to the exclusion of the former, until it is actually held in certain quarters that the selection of small differences can only lead to small, or at all events strictly limited, changes of type.

Yet it cannot be denied that, apart from colours and other "fancy" points, the actual improvement of domestic animals has usually proceeded by just this accumulation of small differences.

If I am not mistaken, the view that selection is limited can be traced back to Johannsen's work, where he showed that from an ordinary stock of beans there could be isolated a number of "pure lines", which differed from each other in the mean weight of their seed, but within each of which no appreciable genetic variation in seed weight could be detected.

His work has led to a considerable advance in the selection of cereal seed, since it is quite certain that for practical purposes "pure line" seed will behave in much the same way as if the plants were propagated vegetatively; they will start growing together and will ripen together, and their seed will be uniform and behave uniformly in its turn. Yet Johannsen, working of course with self-fertilizing material, found pure *lines*, not *a* pure line. Obviously, therefore, mutations had occurred with sufficient frequency to produce them; and, given time, it may be supposed that even in self-fertilized organisms progress could be made merely by selecting the extreme pure line, waiting for a mutation, selecting again, and so on. Tedious work—but for the Origin of Species there is now plenty of time.

From a practical point of view, however, the plant breeder cannot afford to wait for favourable mutations; he cross-fertilizes—and so in most cases does

* Winter, Floyd L., "Continuous selection for composition in corn", *J. Agric. Res.* July-December 1929, pp. 451–75.

nature. Now until experience has been accumulated, the results of cross-fertilization are unpredictable; but very soon certain facts begin to emerge—"Tall is dominant to dwarf", "Two rowed is dominant to six rowed", and so forth, and such things attract attention and rather obscure other equally important facts. Cross a "dense" and a "lax" variety, and among the ultimate progeny may be found plants "denser" than the "dense" parent and "laxer" than the "lax" and almost anything between. Cross high and low protein, and the same overlapping will be found when the first mix-up has sorted itself out.

Try to explain this on Mendelian lines and it will soon become obvious that even in self-fertilized plants there must be a tremendous variety of genetical make-up; one or two relevant genes will be quite inadequate to explain the facts, ten or twenty will complicate the calculation, but will be none too many. Perhaps it would be better to postulate 200–300 and reduce the problem to mathematics.

Since characters which do not affect the survival of the organism are not encountering selection, an ordinary cross-fertilizing population must be expected to accumulate among all its members very large numbers of genes corresponding to such unessential characters. In ordinary times these would roughly *neutralize one another*, each individual carrying a mixture of genes which would produce variation in opposite directions, so that only a limited genetic variation would result; but with a change of environment this reservoir of genes would serve a very useful purpose as raw material for selection: some characters, formerly neutral, would then affect survival and all those genes which produce favourable somatic variation would tend to be preserved while their opposite numbers would be eliminated. Thus the accumulation of small variations in the same direction could proceed far beyond the original range.*

* Perhaps this argument may be clarified by an illustration. Suppose during a period when height is of no particular importance to an organism two hundred small mutations have succeeded in establishing themselves in equilibrium, each of which affects height to an equal extent, say, 1 mm. We may represent the first gene as either a_1, present, or b_1, absent, the second as a_2 or b_2, and so on. Then any individual will contain either a_1a_1, a_1b_1, or b_1b_1 and the proportions in which these possibilities occur will be assumed for the sake of illustration to be 1.2.1; similarly with the other subscripts, so that the distribution of individuals according to the numbers of "a" genes which they contain will be in proportion to the coefficients of the binomial $(a^2+2ab+b^2)^{200}$ or of $(a+b)^{400}$.

The standard deviation of this binomial distribution is 10, so that although it would be possible for an individual to contain the "a" genes in any number from 0 to 400, yet in practice even a population of 100,000,000 would be very unlikely to outrange 140–260 corresponding to 120 mm. of height between the highest and the lowest individual, less than one-third the possible range.

If now we imagine only the highest half of the population to mate (at random) we should get a rise in "a" content of 8 in the mean value, to 208, while the standard deviation and range would hardly be altered, so that the process could be repeated, a further rise of 8 mm. obtained, and so on until the mean would rise well beyond the value of the original extreme individual: and all this without fresh mutations. Of course this illustration has been simplified to the point of absurdity, but it may serve to exhibit the possibility of such potential variation.

That such a state of things does indeed exist seems to be indicated by Winter's paper, to which I am now drawing attention. This describes a very determined experiment carried out on "corn", i.e. maize. Now maize is commonly cross-fertilized; unless cross-fertilization takes place, the stock is apt to die out—which makes pure line selection very difficult. Nevertheless much may be done by mass selection, and it is with mass selection that Winter was concerned.

Premising that he selected continuously for twenty-eight years, from 1896 to 1924, it is perhaps best to quote his description of the procedure verbatim:

> One hundred and sixty-three ears of a variety known as "Burr's White" were used as foundation stock from which selections were made in four different directions, namely for high oil, low oil, high protein and low protein.
>
> These four strains were carried on in the same way. In the high protein, for example, twenty-four ears highest in protein were selected for seed and planted in an isolated plot, each ear in a separate row. These ears were harvested separately and the seed for the next crop selected from the ears which were found to be highest in protein. Nine years later the system was modified somewhat in an attempt to prevent loss of vigour by inbreeding. Alternate rows were detasselled and seed was selected only from the highest yielding detasselled rows. In 1921 this system was again modified to reduce the amount of inbreeding. Two seed ears were taken from each of the detasselled rows regardless of yield.
>
> The high oil, low oil and low protein tests were similarly conducted, selection being made each year of ears highest in oil, lowest in oil and lowest in protein, respectively.

For a proper appreciation of the work the original paper should be consulted, but only a few figures will be necessary to display the interest of the results:

I will deal with the figures giving the percentage of oil, which are the more striking, but the facts are similar in the case of the protein.

(1) Two strains have been selected, one which has a mean percentage of oil about *twelve* times the standard deviation of the original population above the original mean, and the other about *seven* times below. As illustrating this, the minimum value in the high race during the last five years is considerably higher than the maximum value found during the first four years and, on the other hand, the maximum value in the low race is even more markedly below the lowest in the first four years.

(2) Although the standard deviation of the high race has risen and that of the low has fallen during the experiment, it would be hard to say whether on the whole there has been a decrease or an increase in variability owing to the selection.*

We may assume the variance to be composed of two parts, one inherent and therefore subject to selection, and the other environmental, or "fluctuating", and therefore a hindrance to selection. Just what proportion we should allot to

* Dr Rasmussen, of Svalof, has pointed out to me that this might perhaps be explained by an exaggeration in the environmental effect when acting on plants enfeebled by inbreeding. But the steady rise in oil percentage right up to the end of experiment seems to require an almost undiminished genetic variability.

PROTEIN

Year	Mean value %		Standard deviation		Lowest variate		Highest variate	
1896	10·93		1·04		8·3		13·9	
	High	Low	High	Low	High	Low	High	Low
1897	10·99	10·63	1·16	0·90	8·3	8·2	13·6	14·0
1898	10·98	10·49	1·22	1·32	7·7	7·5	14·9	13·4
1899	11·62	9·59	1·28	1·01	8·4	6·7	14·8	13·1
1920	14·01	7·54	1·79	0·89	9·5	6·0	17·4	10·5
1921	16·66	9·14	1·84	1·35	9·4	6·6	18·8	13·4
1922	17·34	7·42	1·24	0·70	12·6	6·1	20·6	9·6
1923	16·53	6·48	1·41	0·73	13·1	5·0	19·7	9·4
1924	16·60	8·38	1·19	1·17	14·6	6·1	19·2	11·8

OIL

Year	Mean value % of oil		Standard deviation		Lowest variate		Highest variate	
1896	4·68		0·41		3·9		6·0	
	High	Low	High	Low	High	Low	High	Low
1897	4·79	4·10	0·38	0·29	3·6	3·4	5·7	4·7
1898	5·10	3·59	0·48	0·32	4·1	3·2	6·7	4·8
1899	5·65	3·85	0·42	0·32	4·3	2·8	6·5	4·6
1920	9·28	1·80	0·52	0·21	7·8	1·0	10·6	2·4
1921	9·94	1·71	0·66	0·15	8·4	1·0	11·7	2·3
1922	9·86	1·68	0·54	0·19	8·7	0·9	11·3	2·2
1923	10·08	1·58	0·65	0·24	8·3	1·1	11·8	2·1
1924	9·86	1·51	0·61	0·22	8·4	0·9	11·7	2·2

each of these we have no sure means of judging, but in both cases the latter is, I believe, likely to be very large. Incidentally it may perhaps account in both cases for the obvious correlation* between the mean and the standard deviation.

In any case the inherent part of the variation had of course a *smaller* standard deviation than that observed for the whole, perhaps much smaller, so that the movements of the means were, respectively, *more* than twelve and seven times this "inherent" standard deviation. Hence either the possibilities of variation latent in the original material were enormous or a steady stream of favourable mutations was maintained to carry the means along.

In any case, these results cannot be explained on the basis of a few easily detected genes. But by reducing the problem to the simplest possible basis—starting from the intensity of selection, the rate of movement of the means at first, and the difference between the initial and final values of the mean—it is possible to make some sort of calculation of the minimum number of genes which might allow of so large a change by repeated selection. And I find that the order of these numbers is 100–300. There is little indication, however, that selection

* It is reasonable to suppose that a given variation in environment would produce greater variation in a high genetic stock than in a low one.

had yet reached its limit after twenty-eight years, and we should probably be within the mark if we assumed that the number of genes affecting oil (or protein) content in Burr's White Maize may run up to thousands.

But if we have thousands of genes, continuous selection in one direction may, in fact must, result in progress almost without limit (at all events until the progress itself induces counter-selection as perhaps it does in the case of low oil content) for although the selection will reduce the number of genes there will be time for fresh mutations to occur to keep up the possibility of further selection.

Summary and Conclusion

To sum up: Winter has in this experiment succeeded, by continuous mass selection, in producing two races of maize, one of which has more than twice, and the other less than one-third, the normal oil content.

In a character so influenced by environment the progress has, of course, not been uniform in its manifestation; but it appears to have been comparatively so genetically, and shows little or no indication that it has reached its limit in either direction.

It does not appear that such steady progress could be obtained with less than hundreds of genes affecting oil content and it seems not unlikely that there are thousands. In any case it is clear that the possibilities of continuous selection of small variations for the formation of new species are likely to be very much greater than would appear merely from a consideration of Johannsen's work on pure lines, which was carried out on a self-fertilizing organism.

And so we reach the conception of species patiently accumulating a store of genes, of no value under existing conditions and for the most part neutralized by other genes of opposite sign. When, however, conditions change, unless too suddenly or drastically, the species finds in this store genes which give rise to just the variation which will enable it to adapt itself to the change.

It follows that the change appears to have produced the variation which it has merely selected from among those potentially present. Thus we can reconcile the view held, amongst other people, by the late Walter Heape, that the environment produces the required variation, with the older Darwinian selection of random variations, to which it appears at first sight to be diametrically opposed.

19

A CALCULATION OF THE MINIMUM NUMBER OF GENES IN WINTER'S SELECTION EXPERIMENT

[*Annals of Eugenics*, VI (1934), p. 77]

In a note on Winter's selection experiment* published in the *Eugenics Review*† I made the following claim:

> By reducing the problem to the simplest possible basis . . . it is possible to make some sort of calculation of the minimum number of genes which might allow of so large a change by repeated selection. And I find that the order of these numbers is 100–300.

Prof. Fisher, however, pointed out in *Nature*‡ that I had in fact over-simplified my problem and that no such conclusion could be drawn from my "sort of calculation".

This did not in fact invalidate my main thesis, which was that species tend to accumulate a sufficient store of genes of no particular value until they meet with a change of environment, when the store provides material for selection far beyond the normal range.

But although the calculation was based on over-simplified data and was superfluous to my argument, it is of some interest in itself, and the present note is an attempt to "mend my hand" by making more reasonable assumptions.

I shall start by giving a very short account of Winter's experiment with an abbreviated table, hoping that my readers may be sufficiently interested to study Winter's paper for themselves.

Then I shall make an estimate of the standard deviation of that part of the variation in oil content of Winter's maize which was due to genetic constitution, and measure the difference between the mean oil content of his "high" and "low" races in terms of this standard deviation.

I shall next make an estimate of the minimum numbers of genes which would suffice to account for so large a ratio between the possible range and the standard deviation.

Finally, I shall discuss the various assumptions which have been made, pointing out which of them are in my opinion reasonable, which have reduced the minimum

* "The mean and variability as affected by continuous selection for composition in corn", *J. Agric. Res.* XXXIX (1929), pp. 451–75.

† "Evolution by selection. The implications of Winter's selection experiment", *Eugen. Rev.* XXIV (4 Nov. 1933) [18].

‡ "Number of Mendelian factors in quantitative inheritance", *Nature*, CXXI (18 March 1933), p. 400.

number of genes to a figure below that which is probable, and which are merely the best assumptions we can make.

Winter's experiment, then, was concerned with a continuous selection of maize in the directions of high and low protein and high and low oil content, and I am only concerned here with the latter.

The experiment was begun in 1896 and has continued to the present day,* but only the first 28 years were reported on in his paper, i.e. till 1924. The following is his description of the procedure, which I have only altered by instancing the oil content part of the experiment, whereas he quoted the similar case of the protein:

One hundred and sixty-three ears of a variety known as "Burr's White" were used as foundation stock from which selections were made in four different directions, namely for high oil, low oil, high protein and low protein.

These four strains were carried on in the same way. In the high oil, for example, twenty-four ears highest in oil were selected for seed and planted in an isolated plot, each ear in a separate row. These ears were harvested separately and the seed for the next crop selected from the ears which were found to be highest in oil. Nine years later the system was modified somewhat in an attempt to prevent loss of vigour by inbreeding. Alternate rows were detasselled and seed was selected only from the highest yielding detasselled rows. In 1921 this system was again modified to reduce the amount of inbreeding. Two seed ears were taken from each of the detasselled rows regardless of yield.

The high protein, low protein and low oil tests were similarly conducted, selection being made each year of ears highest in protein, lowest in protein and lowest in oil, respectively.

Year	No. of ears analysed		Mean value percentage of oil		Standard deviation		Lowest variate		Highest variate	
1896	163		4·68		0·41		3·9		6·0	
	High	Low	High	Low	High	Low	High	Low	High	Low
1897	80	50	4·79	4·10	0·38	0·29	3·6	3·4	5·7	4·7
1898	216	108	5·10	3·59	0·48	0·32	4·1	3·2	6·7	4·8
1899	108	144	5·65	3·85	0·42	0·32	4·3	2·8	6·5	4·6
1900	108	144	6·10	3·57	0·44	0·36	4·6	2·6	7·4	4·5
1901	126	126	6·24	3·45	0·45	0·26	4·9	2·8	7·1	4·1
1920	120	120	9·28	1·80	0·52	0·21	7·8	1·0	10·6	2·4
1921	120	120	9·94	1·71	0·66	0·15	8·4	1·0	11·7	2·3
1922	120	120	9·86	1·68	0·54	0·19	8·7	0·9	11·3	2·2
1923	120	120	10·08	1·58	0·65	0·24	8·3	1·1	11·8	2·1
1924	120	120	9·86	1·51	0·61	0·22	8·4	0·9	11·7	2·2

The above table gives certain figures for the first six and the last five years of the experiment, and it will be seen that, although the original maize only varied

* Mr Winter in correspondence a year or two ago told me that both these experiments and one on height were still being continued and still showed a continued, if less marked, effect of selection. In the latter case he had arrived at mean heights of 8 ft. and 8 in. in two races derived from a 4 ft. maize.

in oil content from 3·9 % to 6·0 %, the lowest variate of the high race after 28 years of selection was 8·4 % in oil content, while the highest variate of the low race was only 2·2 %; in each case they were clean outside the original range, a fact which seems difficult to explain except on the hypothesis that the oil content of the original race was due to a number of genes which largely neutralized one another, some raising and some lowering it, thus allowing selection far outside the original range.

It will be noticed that the standard deviation of the percentage of oil in the original race was 0·41 and that as time went on the high race became more variable and the low less so: this was presumably due to the interaction of the environmental variation with the genetic, an individual tending to produce high oil giving more scope to changes of environment than one which tends to produce low oil.

Nevertheless, on the whole the variation has not decreased, and we shall probably not be far wrong in assuming that there was no appreciable change in variability for the first three generations of selection. So that we may take the original standard deviation as the root mean square of the seven values 0·41, 0·38, 0·29, 0·48, 0·32, 0·42 and 0·32, which is 0·38.

After three selections in each direction the mean of the high race was 5·65 and that of the low 3·85, a difference of 1·80, and this difference may be taken as genetic.

Now we are told that in the first generation 24 ears were selected in each direction out of 163 and, on the assumption of normal distribution of oil content, the mean of these selected ears would have an oil content $1{\cdot}56 \times \sigma_v$ above (or below) the mean, σ_v being the standard deviation of the oil distribution. It is further stated that there were 80 ears analysed of the high race and 50 of the low in the next generation, and it is, I think, a fair inference that 24 of each of these were taken in the next selection. This is confirmed by the fact that in the later generations 120 ears (5×24) were invariably analysed.

The mean of 24 ears selected from 80 (on the normal assumption) is $1{\cdot}16\sigma_v$ above the mean and that of 24 from 50, $0{\cdot}83\sigma_v$ below the mean, and the corresponding figures for the next selection ($\frac{24}{216}$ and $\frac{24}{108}$) are $1{\cdot}71\sigma_v$ and $1{\cdot}34\sigma_v$, so that the total shift of the mean of the high race was $(1{\cdot}56 + 1{\cdot}16 + 1{\cdot}71)\,\sigma_v = 4{\cdot}43\sigma_v$ and that of the low race

$$(1{\cdot}56 + 0{\cdot}83 + 1{\cdot}34)\,\sigma_v = 3{\cdot}73\sigma_v,$$

or altogether the races were shifted apart $8{\cdot}16\sigma_v$, of which 1·80 appears to have been genetic, as shown by the distance apart after the six selections.

Now if σ_v be the standard deviation of total variation and σ_g of that part which is genetic, then, on the supposition of independence between the environmental and genetic parts of the variation σ_g/σ_v is the correlation between the

genetic and the total variation, so that σ_g^2/σ_v^2 is the regression factor reducing the mean of the selected portion to the mean of the next generation.

Hence
$$8{\cdot}16\sigma_v \times \frac{\sigma_g^2}{\sigma_v^2} = 1{\cdot}80,$$
$$\sigma_g = \sqrt{\left(\frac{1{\cdot}80}{8{\cdot}16} \times 0{\cdot}38\right)} = 0{\cdot}29.$$

Since the differences between the means of the high and low races in the last five generations were 7·48, 8·23, 8·18, 8·50, 8·35, we shall not be far wrong if we estimate the genetic range at not less than 29 times the genetic standard deviation ($29 \times 0{\cdot}29 = 8{\cdot}41$).

We have now to estimate the minimum number of genes which will give as large a ratio as 29 between the maximum range and the standard deviation.

In the first place it is clear that less genes will be required if the effect of each on the oil content is the same, and we shall assume that each gene if homozygous produces an effect $2k$ and if heterozygous k. Further, let us suppose n genes, the rth to be present in P_r of the possible loci and absent in Q_r, and let

$$p_r = \frac{P_r}{P_r + Q_r} \text{ and } q_r = \frac{Q_r}{P_r + Q_r}.$$

Then p_r^2 individuals will be $2k$ higher owing to that gene, $2p_rq_r$ individuals will be k higher owing to that gene, and q_r^2 individuals will have no effect from that gene. (I have taken the rth gene as increasing the oil, but clearly, the same effect is produced in the case of a gene which decreases the oil, but the convention in this case is that p represents the absence of such a gene and q its presence.)

Then the distribution of all the n genes will be given by the various terms of the expansion
$$(p_1+q_1)^2(p_2+q_2)^2 \dots (p_r+q_r)^2 \dots (p_n+q_n)^2,$$
and the extreme individuals (in genetic constitution) will be present in the proportions
$$p_1^2p_2^2p_3^2 \dots p_r^2 \dots p_n^2 \quad \text{and} \quad q_1^2q_2^2 \dots q_r^2 \dots q_n^2,$$
and they will differ by $2nk$, whereas the standard deviation of the expansion is*
$$\sqrt{\{2n(\overline{pq} - \sigma_p^2)\}}\, k,$$
where $\overline{p}$ is the mean of the p's and $\overline{q}$ of the q's (which we may take as $\frac{1}{2}$ each), while σ_p is the standard deviation of the p's (which also $= \sigma_q$).

Now according to Prof. Fisher† the frequency distribution of the p's is given by the equation $\Delta f = C/pq$ and, after some tedious algebra, I find that
$$\sigma_p^2 = \tfrac{1}{4} - \frac{N-1}{2NS_{N-1}(1/r)},$$
where N is the number of loci (here $2 \times 163 = 326$), and this reduces to $\frac{1}{4} - 0{\cdot}0783$.

* "An explanation of deviations from Poisson's Law in practice", *Biometrika*, XII (1919), p. 213 footnote [9, p. 67].

† *Genetical Theory of Natural Selection* (1930), p. 91.

Hence we have the standard deviation of the expansion

$$\sqrt{\{2n \times 0{\cdot}0783\}}\, k,$$

and the ratio of the extreme range to the standard deviation

$$\frac{2nk}{\sqrt{\{2n \times 0{\cdot}0783\}}\, k} \quad \text{or} \quad \sqrt{(25{\cdot}5n)}.$$

Hence to determine n we equate $\sqrt{(25{\cdot}5n)} = 29$, $n = 33$.

In this calculation the following assumptions have been made, which seem to me to be reasonable, and small departures from them will not seriously affect the result:

(*a*) The distribution of the percentage of oil has been taken as normal.

(*b*) I have taken the genetic standard deviation as being appreciably constant for the first three generations and have assumed that the difference between the high and low races at this point will be sufficiently accurate to give the genetic part of the variation.

To test this I have calculated the number of genes on the basis of

taking the first	pair of selections	giving	33	genes
up to the second	,,	,,	,,	25	,,
,, ,, third	,,	,,	,,	33	,,
,, ,, fourth	,,	,,	,,	31	,,
,, ,, fifth	,,	,,	,,	36	,,

All numbers of much the same order.

(*c*) I have assumed linear regression of genetic on total variation and independence between the genetic and environmental variation.

(*d*) I have assumed that the mean value of the p's and q's is $\frac{1}{2}$.

(*e*) Following Fisher, I have assumed an equal distribution of the logarithm of the gene ratio. This should follow whether the gene is absolutely neutral or has a small selective advantage. My own feeling is that there must be a large class of variations which, if they occur in an individual at one end of the range, are favourable, but are unfavourable at the other. As the general distribution in the species tends to be broken up into local races with means more or less different from the general mean, genes will introduce themselves by mutation into such local races as are favourable to their retention and, when firmly established, into the main body of the species.

The following assumptions are such as to give a minimum value of n:

(*f*) I have taken the minimum range as 8·41, i.e. I have not allowed for any genetic variation beyond the means of the last generations, whereas Winter actually found during the next eight years that the means were still moving apart.

If, for example, I had added even as little as three times the standard deviation outwards at each end, making 35 times the standard deviation, n would have risen to 48.

(*g*) I have already mentioned that the assumption of equal effects from all the genes minimizes n.

(*h*) I have assumed absence of dominance. Clearly dominance would increase the standard deviation for the same range and so increase n.

(*i*) I have, naturally, only been able to deal with such genes as were included in Winter's sample of 163 heads, tracing back to, at most, 326 loci. Hence only quite a small proportion of the rarer genes can have been included, and, according to Fisher, far the greater number of genes consists of those which individually occur but seldom. Further, even of these genes included in the original sample, many must have been lost at random in the first few selections and so not have been taken into account by the calculation.

Lastly, the remaining assumptions cast an element of doubt on the whole calculation:

(*j*) Although the standard deviation is correlated with the mean, so that we seem to be measuring variation at the low end of the distribution in smaller units than at the high end, I have taken the difference between the means of the high and low units as if it was uniform, and divided by the standard deviation determined at the middle of the scale. I suspect that this tends to exaggerate the difference and so n.

(*k*) I have assumed that the effect of the genes is additive, whereas they may really obey some quite other law.

Nevertheless, though I do not feel that the above calculation can be altogether absolved from the charge of "playing with figures", I think that it does really afford some evidence that the oil percentage of Winter's maize was conditioned by the presence, or absence, of a number of genes, at least of the order 20–40, possibly of 200–400, and not at all likely to be of the order 5–10.

The 100–300 minimum of genes of the former paper has therefore been reduced to 20–40, but however few or many genes may have been present, the fact remains that Winter was able to select his maize races far outside the range of his original material. This seems to me to justify*

the conception of species patiently accumulating a store of genes, of no value under existing conditions and for the most part neutralized by other genes of opposite sign. When, however, conditions change, unless too suddenly or drastically, the species finds in this store genes which give rise to just the variation which will enable it to adapt itself to the change.

It follows that the change appears to have produced the variation which it has merely selected from among those potentially present. Thus we can reconcile the view that the environment produces the required variation, with the older Darwinian selection of random variations, to which it appears at first sight to be diametrically opposed.

* "Evolution by selection. The implication of Winter's selection experiments", *Eugen. Rev.* XXIV (1933), [18, p. 185].

20

CO-OPERATION IN LARGE-SCALE EXPERIMENTS

[A Discussion, opened by Mr W. S. Gosset, at the meeting of the Industrial and Agricultural Research Section of the Royal Statistical Society, 26 March 1936. Sir Daniel Hall, K.C.B., F.R.S., in the Chair.]

[Supplement to *J. Roy. Statist. Soc.* III (1936), p. 115]

At the outset I must confess that the title is to some extent misleading: co-operation is, I am quite sure, advantageous in all large-scale experiments whether industrial or agricultural, but it happens that, though no farmer, I have only had first-hand experience of co-operation in agriculture and my paper must, therefore, deal with that. On the other hand, there are several Fellows present who will doubtless be able to draw analogies from agriculture to industry as the general principles of experimentation are common to both.

Forty years ago agricultural experiments were mainly carried out in fairly large plots, generally without replication, and in consequence the soil differences between two plots which were to be compared were often so large as to obscure the issue.

Then about thirty years ago, several different investigators harvested apparently uniform fields by small plots, and it at once became obvious that the variation in fertility from point to point in a field is so distributed that to obtain the best experimental results it is necessary to work with a number of small plots. These should be arranged so that comparable plots lie close together, and it appeared further that this replication of plots enabled us to make an estimate of the error of our results in a single experiment; before this it had only been possible to estimate the error of a series of experiments carried out at a number of stations or in a number of years.

Finally, about fifteen years ago, Prof. Fisher introduced the principle of randomizing the position of the plots in the various systems of randomized blocks and Latin squares with which many of you are familiar. This enabled us to obtain a certainly valid estimate of the variability of our results, though usually at the expense of increasing that variability when compared with balanced arrangements.

Nevertheless, it must not be supposed that valuable results could not be obtained by the primitive methods of forty years ago; for example, in the 1880's and 1890's the Danes, working with comparatively large plots, with few replications, but at several co-operating stations and in a number of successive seasons,

were able to establish that Prentice was the most suitable barley to grow in Denmark.

On the other hand, Mr Yates has pointed out that it is not uncommon, when using the most modern methods in manurial experiments, to obtain a significant result on one occasion, but, on repeating the experiment in another year or in another field, to get an equally significant result in the opposite direction.

Nor is the reason of this far to seek; among the many causes which influence the result of an experiment, we can only control by the arrangement of our plots those connected with the variation in fertility of the experimental area; apart from these we have the wide differences in soil and climate over the districts in which we wish to apply the conclusions which we draw from our experiments.

Hence the old work, if repeated on a representative scale and sufficiently often, was able to give results which were applicable over a wide area, while the very accuracy of Mr Yates's methods enables him to reach significance for results of merely local value.

Nevertheless, it would be a mistake to reduce the accuracy by insufficient replication, for only by repeating such work at different times and places can the causes of such apparent anomalies be traced, and for that the more we can eliminate mere soil errors the better.

But such repetitions can only be carried out co-operatively, and I propose to give some instances of such co-operation, beginning with the simplest technique.

Just before the beginning of this century the Irish Agricultural Organization Society, which later became the Department of Agriculture, began a research into the most suitable variety of barley to grow in Ireland, and this research has been continued to the present day. During this time three varieties of barley have been introduced into Ireland, after adequate evidence had been obtained that each was better than the barley which it succeeded, and the methods of seed distribution are such that after a very few years the new barley has replaced the old in practically all the barley-growing districts in Ireland.

It is interesting to note that the first of the three barleys to be introduced was found to be identical with that which the Danes had proved to be most suitable for Denmark; the other two were obtained from it by cross-fertilization by Dr Hunter.

The resulting gain in yield has been remarkable, and though it would be easy to attach too much importance to evidence supplied by the official estimates, they tally fairly well with the claim which has been made, on the basis of the experimental plots, that there has been a gain of from 20 to 25 %.

During the last ten years the official yield has dropped below 5 qr. only once, while only twice in the previous sixty years did it rise above that figure.

The low yields between 1916 and 1925 were partly due to unfavourable weather, but also to the extension of arable land during the war, with consequent inclusion

TABLE I

Yield of Barley in Ireland in Quarters per Acre

Before experimenting		After experimenting	
1866–1870	4·0	1901–1905	4·5
1871–1875	4·1	1906–1910	4·7
1876–1880	3·9	1911–1915	4·8
1881–1885	3·9	1916–1920	4·1
1886–1890	3·9	1921–1925	4·1
1891–1895	4·3	1926–1930	5·3
1896–1900	4·3	1931–1935	5·1

of less suitable land, and to the subsequent decline in farming technique owing to wages being high compared with prices.

The experiments are carried out at about ten centres where three varieties are tested against the standard variety in one-acre plots. This somewhat primitive arrangement has been carried on up to the present day in order to provide plenty of barley for quality tests.

In any case, after some years the weather and the barley-growing land of the country were sampled in a way which would be impossible at a single station. The number of farms should, of course, be larger, and doubtless it would be but for the fact that only one official is available for supervision, and ten farms at distances of, in some cases, over 100 miles is as much as he can manage even when the experiment is of this very simple type.

The error of a comparison between two one-acre plots is large, and quite a number of seasons pass before enough repetitions are available to reduce the error to a figure which will show that a new variety really yields better than the standard. As, however, it is as necessary to sample weather as districts, this is of no great disadvantage.

The order of this error is of interest, and I have examined two series to determine it; the first was carried out between 1901 and 1906, when 51 comparisons between Archer and Goldthorpe gave an average advantage to Archer of 7·7 % with a standard error of a single comparison of 15·5 %. This tallies well enough with the traditional 10 % for the error of a comparison of a pair of plots at one station, having regard to the further real variation due to the differential response of the varieties to soil, climate and farming technique.

The second series was carried out between 1925 and 1935, when two selections of the Spratt-Archer cross were compared: they differed by 0·27 % in 103 trials with a standard error of 9·3 %.

These two estimates of the error of a comparison, 15·5 % and 9·3 %, differ significantly, and it is noteworthy that the smaller figure was found with barleys which might be expected to react in much the same way to differences in soil and weather.

A second set of experiments has been carried out by the National Institute of Agricultural Botany, and I instance it to give an idea of the advantage of using a method which reduces the error at each station—namely, Beaven's half-drill strip.

It has been said that from an experiment conducted by this method no valid conclusion can be drawn, but even if this were so, it would not affect a series of such experiments. Each is independent of all the others, and it is not necessary to randomize a series which is already random, for, as Lincoln said, "you can't unscramble an egg". Hence, since the tendency of deliberate randomizing is to increase the error, a balanced arrangement like the half-drill strip is best if otherwise convenient.

From this work I have taken two series, one of 22 comparisons between Spratt-Archer and Plumage-Archer barleys carried out from 1925 to 1928 when the former yielded 6·1 % more and the standard error of a comparison was 8·1 %.

There was, however, one experiment in which the method was not followed in several particulars, and if that be omitted the standard error falls to 5·6 %.

The second series of N.I.A.B. experiments was a comparison between Spratt-Archer and a selection from Plumage-Archer which was carried out at six stations and for three years. It is thus possible to analyse the variance, and though the numbers are too small to give a significant difference in variance, there is an indication that the greater part was connected with the stations. The average superiority in yield of Spratt-Archer was 8·2 % and the S.D. of a comparison was 8·4 %; this is significant for 18 comparisons, so that the main object of the experiment was attained provided that the stations could be assumed to be a representative sample.

The analysis of variance is as follows:

Degrees of freedom		Sum of squares	Mean squares
Seasons	2	22·25	11·13
Stations	5	815·34	163·07
Remainder	10	352·26	35·23
Total	17	1189·85	69·99

The remainder, of course, includes not only the error due to soil differences, but also those due to the local differences in climate within each season and to the difference between the fields used at each station.

I have drawn attention to this small series because it indicates the possibility, had there been sufficient stations, of connecting the peculiarities of the soil and weather at the stations with the relative yields of the varieties. Thus there was an indication that Spratt-Archer was less superior to Plumage-Archer when the yields were high, but it was by no means significant.

Assuming, then, that the error of the one-acre plot experiment is of the order 12 % and that of the half-drill strip 8 %, the advantage of the latter is not so much that fewer experiments would be needed to evaluate a given difference in yield, for in any case it is necessary to spread one's net widely both in time and space; nor is the smaller area occupied a clear gain, for it is offset by the necessity for closer supervision; but it does make it possible to contract the limits of significance so that more series of experiments give definite answers to the questions asked.

I have instanced the half-drill strip, but obviously any method of reducing the error is of advantage, whether it is by replication (including, for instance, multiple Latin squares), reduction of the size of plot, or regular balanced arrangement.

The instances given above have been fairly simple, inasmuch as the differential response of barleys to variations of soil and climate is small; but even in these cases it would have been of advantage to have spread the net wider: the next experiment to which I am going to refer is of a more complicated nature, and is concerned with the response of sugar-beet to artificial manures.

This has been described in the Rothamsted Report for 1934, and though I do not propose to try to add to the full analysis given therein, a short account of it may be instructive.

The experiment was carried out in two seasons, at 13 stations in 1933 and 15 in 1934; all combinations of three manures at three rates per acre were tried, and measurements of the weights of roots and tops, and of percentage of sugar and purity, were made, and various conclusions were drawn as to the effects of the manures. Among others, it appeared that some of these effects differed significantly at different farms.

The next thing, clearly, is to connect up these differences with the character of the soil and weather at the various farms, but though mechanical and chemical analyses of the soil were carried out, there is no mention in the report of any attempt to do this. Presumably there was no marked connexion, and further results are awaited, for if "8 of the 15 centres gave significant increases in yield of roots with sulphate of ammonia, while the remaining 7 centres showed no appreciable increases", the value of the result to the individual farmer will be much increased by some indication of whether his land is to be classed with the 8 or the 7. I call attention to this in no spirit of criticism, but in order to bring out the full possibilities of co-operation on a still larger scale.

Both Dr Beaven and the Rothamsted school have maintained that their methods can be carried out by the ordinary farmer; and if for ordinary you substitute exceptional, I agree; but the business, even of the exceptional farmer, is to farm, and he cannot afford the time to weigh up small experimental plots when he ought to be getting on with his work.

And so, while a co-operative series of experiments should always include a majority carried out on ordinary farms, there must be trained supervision and cultivation money, and this can only come from the Government, working through institutions like the National Institute of Agricultural Botany or Rothamsted.

Furthermore, the more complicated the method, the more supervision is required; one man can just look after ten experiments with acre plots, with half-drill strips you probably want at least three, and for more complicated experiments even more; but farming is a large industry, and a gain, even a small gain, per acre on 100,000 acres soon pays for the cost of making experiments.

APPENDIX

The Error of Half-Drill Strip Experiments

The half-drill strip technique has been criticized on the ground that no valid conclusion can be drawn from experiments carried out by it, and it may be well to examine what truth there is in the assertion.

Essentially the method consists in sowing long narrow strips of two varieties of cereals in alternation. By an ingenious arrangement at sowing, these strips can be split longitudinally at harvest, and each half strip of one variety is compared with the half strip of the other adjacent to it; to balance the linear term of the fertility slope, the series begins and ends with a half strip of the same variety. The series is therefore of the form $ABBAABBA \ldots ABBA$, and to calculate the error of the difference $(A-B)$ a degree of freedom is allocated to the fertility slope. This is determined by the difference $(S(\overline{AB})-S(\overline{BA}))\ 1/n$, where $S(\overline{AB})$ is taken to be the sum of $A-B$ for all the comparisons AB, $S(\overline{BA})$ for all the comparisons BA and n is the number of pairs.

Thus the analysis of the variance is given in a table of the form*

	Degrees of freedom	Sum of squares
Fertility slope	1	$(S(\overline{AB})-S(\overline{BA}))^2 1/n$
Random error	$n-2$	$S(A-B)^2-(S(\overline{AB})-S(\overline{BA}))^2 1/n$
Total	$n-1$	$S(A-B)^2$

If, then, the variation in fertility consisted of random deviations superposed on a uniform fertility slope, the procedure would be beyond criticism; it remains to be seen how departures from such an ideal system invalidate the argument.

The almost universal departure is that the fertility slope is not uniform, there are, ideally speaking, parabolic terms, so that the position $\overline{AB}$ represents a different advantage to A at different points in the series. This will have the effect of increasing the apparent error, since the sum of the squares of the differences,

* [In this table it seems necessary to read $S(A-B)^2-n(\overline{A-B})^2$ for $S(A-B)^2$. Ed.]

$S(A-B)^2$, includes just as large a component due to the fertility slope, while the component calculated, $(S(\overline{AB})-S(\overline{BA}))^2 1/n$, is smaller; this is because the sign of $\overline{AB}$ (and of $\overline{BA}$) changes on passing from a falling to a rising part of the curve. On the other hand, there is a corresponding increase in the real error owing to the fertility slope not being accurately balanced, this error amounting at most to $2/n$ of the fertility slope between a pair for each change of direction.

Furthermore unless the fertility slope is of a periodic nature, a case to be considered later, the incidence of these changes of curvature will be random, so that the general tendency will be slightly to over-estimate the error, a fault on the right side for most of us, and one which is compensated by the smaliness even of the apparent error.

Periodic fertility slopes may undoubtedly occur, but apart from those due to the works of man, they must be so rare as to add a negligible risk; where, however, they are due to such causes as old ploughman's "lands", it should be possible to avoid them by inspection; even if they have been overlooked, the chance of their affecting the mean difference is small, for to do so the period must very nearly coincide with an odd multiple of the width of a whole strip; in general, it is the apparent error that would be increased.

We may therefore conclude that there is a slight tendency for the error of a half-drill strip experiment to be over-estimated, so that somewhat fewer significant results are obtained than if the real error could be accurately determined; this is more than made up for by the smallness of the error itself as compared with that of most other arrangements.

There remain two other criticisms; firstly, that the system of drilling is such that half the coulters of the drill are allocated to one of the varieties and the rest to the other; if, then, the coulters on one side are badly set or stopped up, the other may have a constant advantage. This, though a real possibility, and one to be guarded against by careful inspection, is not as serious as it sounds, at all events with barley; for barley automatically fills up gaps to such an extent that the alteration in yield by large changes in seeding rates is almost inappreciable, so that within wide limits of faulty seeding it is the area devoted to the variety which counts, and not the exact distribution of seed within it.

The other criticism has more substance; by the half-drill strip method only two varieties are directly compared. This is just what is wanted where a standard variety or rate of manuring is to be compared with a competitor for the rank of standard; but if two or more varieties are to be compared with the standard, their inter-comparison is, of course, subject to a much greater error.

Up to the present, the half-drill strip method has, as far as I know, only been used for cereals in these Islands and in New Zealand, but it should be equally useful for such manures as can be drilled, and a modification has even been suggested for a forest experiment.

21

COMPARISON BETWEEN BALANCED AND RANDOM ARRANGEMENTS OF FIELD PLOTS

[*Biometrika*, XXIX (1938), p. 363]

[The following editorial note was printed at the head of this paper: With very deep regret the Editorial Committee has to report the death, on 16 October 1937, of Mr W. S. Gosset, whose scientific contributions under the pseudonym of "Student" are well known to all statisticians. It is hoped to include some account of his life and work in the next issue of the *Journal*.

Mr Gosset had been working at the following paper during the past summer, and a fortnight before his death had discussed the draft, which is printed below, with Dr J. Neyman and Prof. E. S. Pearson. It was then agreed that certain points in sections 2 and 3 needed clarification and Mr Gosset proposed to undertake this work himself; unfortunately this final revision was never completed. Dr Neyman and Prof. Pearson have therefore added in a separate Note (*Biometrika*, XXIX (1938), pp. 380–88) some comments, for which they take full responsibility, regarding the points on which they know Mr Gosset had intended to enlarge.]

IN a paper read before the agricultural and industrial section of the Royal Statistical Society* I ventured to point out that the advantages of artificial randomization are usually offset by an increased error when compared with balanced arrangements. Prof. Fisher does not agree and has written a paper to test the difference of opinion that there is between us.†

In this paper I propose to set out as clearly as I can just what is this difference of opinion.

Next I propose to show that the conclusions of Prof. Fisher's paper all follow firstly from his having made use of a method of calculating the error of the "systematic" arrangements which I showed fourteen years ago would lead to just the misleading conclusions which he has found, and secondly to his not having compared like with like.

Thirdly, I will show that if he had not fallen into these pitfalls he would have been able to show that in the case which he took, a balanced arrangement does in fact give a slightly smaller error than his randomized one.

Fourthly, I will describe just what is to be expected when balanced arrangements are compared with random,‡ viz. that when the variance due to treatment

* W. S. Gosset, "Co-operation in large-scale experiments", Supplement to *J. Roy. Statist. Soc.* III (1936), pp. 115–22, [20].

† Barbacki and Fisher, "A test of the supposed precision of systematic arrangements", *Ann. Eugen.* VII (1936), pp. 189–93.

‡ Note that an arrangement can be both balanced and random and where this is practicable the aims of Prof. Fisher and myself are both satisfied.

is low compared with the error of the experiment, fewer significant results are obtained than with random arrangements, but when the variance due to treatment is high more significant results are obtained with balanced arrangements.

Lastly, I will give in an appendix the results of some testing of balanced versus random arrangements on uniformity trials by Mr A. W. Hudson of Massey College, N.Z.

§ 1. The Effect of Lack of Randomness on Bias

It is almost invariably necessary, when applying mathematics to practical affairs, to replace the actual conditions by a set of simpler approximations with which the mathematics are capable of dealing, and mathematical statistics are no exception to this rule.

For example, the analysis of variance which is generally used to determine the error of agricultural experiments requires three assumptions to be made before we can apply the method strictly:

(1) The systems concerned are to have normal variation.

(2) The variances of like things should be equal.

(3) The sampling should be random.

(1) If, as is usual, the variation is not normal our argument will not be impaired unless the number of replications is very small, when departure from normality introduces an added uncertainty to the estimation both of mean and perhaps even more of variance.

(2) If, as often happens, the variances are not equal, as for example when we are pooling the variances of the yields of barleys which react differently to soils of different fertility, we shall not in general invalidate our conclusions appreciably, though in extreme cases attention should be paid to this source of error.

(3) If, however, the sampling be not random, there are such possibilities of drawing false conclusions that Prof. Fisher has introduced a system of artificial randomizing to ensure that the third condition is satisfied and brands all other systems invalid.

Nevertheless, it is possible, by balancing sources of error which would otherwise lead to bias, to obtain arrangements of greater precison which are nevertheless effectively random, by which I mean that the departure from randomness is only liable to affect our conclusions to the same sort of extent as do departures from normality or inequality of variances.

Lack of randomness can affect either the mean or the variance, and it is the first of these which is apt to lead to invalid conclusions. Thus Mr Yates has shown that it is practically impossible for anyone to select shoots of corn of average length by eye, and in fact none of the senses can be trusted to behave without bias. Those of taste or smell are peculiarly liable, and if comparisons are to be made it is necessary to avoid giving the least inkling of the order in which the samples

are to be presented, in fact it is better to let it be known that it is a random order. In some cases the only way of avoiding bias is to withhold all knowledge of the object of the investigation from those taking part, though unfortunately this engenders a lack of interest in the proceedings.

Again, a promising experiment in nutrition was ruined by departure from randomness when the schoolmasters were allowed to adjust the supposed uneven effects of a chance selection of subjects for the Lanarkshire Milk Experiment, and in doing so managed to select, doubtless from the most humane motives, 10,000 children to receive milk who were significantly lighter and shorter than the 10,000 "controls" who did not.

In agricultural experiments there are obvious possibilities of bias affecting the mean in badly arranged experiments, for it is usual to find "fertility slopes" in most "uniformity" experiments, i.e. when an apparently uniform field is harvested in small-sized plots it is usual to find that the yield is higher in some parts than in others and tends to change more or less gradually from one place to another. Hence if plots of one variety are sited, whether systematically or by chance, nearer to one end of the experimental area than to the other, the mean is likely to be biased.

To take the simplest case of two varieties or treatments, the layouts

A B A B A B A B (systematic)

and *A B A A B A B B* (random)

will both favour *B* if the field is more fertile on the right than on the left hand, the second rather more than the first.

On the other hand the layout

A B B A A B B A

is balanced with regard to a simple "linear" fertility slope, and the mean of neither *A* nor *B* will be biased except by departure from linearity.

It is, of course, possible to imagine particular variations in soil fertility which will bias the means of plots arranged in this manner, but with one exception they are of the same nature and lead to the same sort of bias—but usually to a smaller extent—as occurs with artificially randomized layouts.

The one exception is a periodic wave of fertility due to previous cultivations which happens to coincide in period with the width of an odd integral number of quartets, a not particularly likely occurrence.

Such layouts as *ABBA* are termed balanced, and any number of treatments may be set in a balanced layout, as, for example, in the Latin square which is not only balanced but random as well, "thus conforming to all the principles of allowed witchcraft".

It is reasonable to expect that balanced layouts will on the whole be successful and that the mean will be less biased than in random, and this expectation is

illustrated by some experimental sampling carried out by Mr A. W. Hudson of Massey College, N.Z., who tested balanced and random blocks against one another on three different uniformity trials. His results are given in the Appendix, and all that need be said here is that in fifteen experiments the balanced layouts showed slightly more bias in three and less in twelve, the reduction of bias being very considerable in some of the twelve.*

And this brings me to a question which has often interested me. Suppose there are two treatments to be randomized—I take two for simplicity only—and suppose that by the luck of the draw they come to be arranged in a very *un*balanced manner, say *AAAABBBB*: is it seriously contended that the risk should be accepted of spoiling the experiment owing to the bias which will affect the mean if there is the usual fertility slope? For, as will be shown later, not only will the mean be biased, but the apparent precision will tend to be high, and misleading conclusions drawn much more often than the 1 or 5 % of the tables. It is of course perfectly true that *in the long run*, taking all possible arrangements, exactly as many misleading conclusions will be drawn as are allowed for in the tables, and anyone prepared to spend a blameless life in repeating an experiment would doubtless confirm this; nevertheless it would be pedantic to continue with an arrangement of plots known beforehand to be likely to lead to a misleading conclusion.

Let us suppose therefore—as indeed it is rumoured—that common sense prevails and chance is invoked a second time and that such an arrangement as *BBABBAAA* is offered; is this to be accepted? It is more likely to give a biased mean than *BABABABA*, but then of course it is random!

And if this is not to be used, how about *BBABABAA*? In short, there is a dilemma—either you must occasionally make experiments which you know beforehand are likely to give misleading results or you must give up the strict applicability of the tables; assuming the latter choice, why not avoid as many misleading results as possible by balancing the arrangements? And this, to do Prof. Fisher justice, is the direction towards which he is tending; in his paper with Dr Barbacki he treats for the first time of "randomized sandwiches" to which the objection is, not an appreciable increase of error, but the practical difficulty of working them.

To sum up, lack of randomness may be a source of serious blunders to careless or ignorant experimenters, but when, as is usual, there is a fertility slope, balanced arrangements tend to give mean values of higher precision compared with artificial arrangements.

Next, what is the effect of lack of randomness on the variance?

In a later section I will show that since in the "null" case, i.e. when no real treatment differences exist, the aggregate variance due to "treatments" and

* Mr Borden, of Hawaiian Sugar Planters' Association, Hawaii, has obtained similar results in similar experiments, and I have no doubt that this will always tend to happen.

residual error is constant for all arrangements of treatments in the blocks, those with low actual error necessarily give high calculated values for the error and vice versa, the calculated error, however, varying much less than the actual in ordinary experiments owing to the larger number of degrees of freedom of the residual error.

This, of course, has nothing to do with the origin of the experiment whether randomized or not.

If, however, the arrangement is "randomized" one can—*before the draw*—state accurately, subject to normality, etc., what the chance of getting any particular partition of variance between "treatment" and "residual error" will be in the "null" case. After the draw, when one particular arrangement has been chosen, it is often possible to be sure that the chance has changed in one direction or another without, however, being able to define exactly what it is.* In particular, balanced arrangements tend to have lower actual errors and higher calculated errors than would be expected by chance before a random selection is made, and this is so even if a degree of freedom is allocated to fertility slope, owing to the departure of the "slope" from linearity.

The consequence is that balanced arrangements more often fail to describe small departures from the "null" hypothesis as significant than do random, though they make up for this by ascribing significance more often when the differences are large.

Thus such departures from the "null" hypothesis as are found to be significant by balanced are likely to be larger than those found by randomized arrangements, and in particular those discovered in the "null" case itself—5 or 1 % as the case may be—tend to disappear altogether with balanced arrangements.

It will be seen then that the difference between Prof. Fisher and myself is not a matter of mathematics—heaven forbid—but of opinion. He holds that balanced arrangements may or may not lead to biased means according to the lie of the ground, but that in any case the value obtained for the error is so misleading that conclusions drawn are not valid, while I maintain that these arrangements tend to reduce the bias due to soil heterogeneity and that so far from the conclusions not being valid they are actually less likely to be erroneous than those drawn from artificially randomized arrangements. Further, that in the really important agricultural experiments which are carried out at more than one centre—and it was of these that I was speaking—the very slight disadvantage that an occasional result at an individual station may not be recognized as significant owing to over-estimation of the error at that station is more than offset by the greater precision of the experiment as a whole.

* This is analogous to the use of a life table to give the expectation of life. Thus the expectation of life of an Englishman of 40 can be referred to an appropriate table, but when we particularize the Englishman of 40 as a tin-miner or an agricultural labourer we know that the expectation is lower or higher than that given in the table without perhaps knowing very exactly by how much.

§ 2. Barbacki and Fisher

Such being our opinions, based in each case on *a priori* argument, Prof. Fisher rightly decided to put the matter to the test by assigning imaginary treatments to plots of which the yield had been determined in a uniformity experiment both on a random and on a balanced system, and published a paper,* of which he gives the following summary:

"1. This inquiry was carried out to test the truth of the opinion expressed by 'Student' that randomization achieves its object 'usually at the expense of increasing the variability when compared with balanced arrangements', and that one of the means available to experimenters of reducing the error is by adopting a 'regular balanced arrangement'.

"2. Using an extensive uniformity test it is found that the arrangements randomizing either pairs or sandwiches of half-drill strips give smaller errors than the systematic arrangement advocated as more precise.

"3. As a consequence experimenters using the systematic arrangements systematically underestimate their errors.

"4. The error estimated from a systematic arrangement is ambiguous, and the experimenter has an arbitrary choice between several widely different estimates.

"5. Owing to the failure to furnish a valid estimate of error, 'Student's' test of significance is not approximately correct for systematic arrangements."

The particular arrangement which Prof. Fisher intended to test was the Half-Drill Strip† introduced by Dr Beaven some fourteen years ago and widely used since then, but unfortunately half-drill strips are too large to lend themselves easily to testing on ordinary uniformity trials, and although Prof. Fisher has laid out eight pairs of half-drill strips on his uniformity trials he has not in fact compared them with a corresponding random arrangement but has cut them up transversely into 5-yard lengths and has compared the actual error of the large half-drill strips with that calculated from the randomized‡ sheaf weights of which they are composed.

Now it happens that Dr Beaven had originally proposed to calculate the error of the half-drill strip from sheaf weights of this kind, and that I pointed out in this *Journal* thirteen years ago§ that since such "sheaf weights" may be positively correlated such a method of calculating the error is fallacious.

* Barbacki and Fisher, "A test of the supposed precision of systematic arrangements", *Ann. Eugen.* VII, pp. 189–93.

† Prof. Fisher prefers to call this the "Split Drill" Method, but though I agree that the name is more descriptive it is a pity to confuse the matter by a change of name after all these years. More particularly is it confusing to transfer the name "Half-Drill Strip" to small portions of the original half-drill strip as he has done, and I have called them by Dr Beaven's name of "Sheaf Weights".

‡ *Not very much randomized;* he compares corresponding pairs just as anyone else would.

§ "On testing varieties of cereals", *Biometrika,* XV (1923), pp. 271–93, [11].

This method of calculating the error has, of course, nothing to do with balanced arrangements, except that it was proposed by Dr Beaven, the author of the half-drill strip; it might just as well be applied to random arrangements, as, for example, the "randomized pairs" of Prof. Fisher's experiment, each of which was actually harvested in six separate drills from which the error could have been equally erroneously calculated.

Prof. Fisher has therefore calculated the error of the half-drill strip by a method which I showed thirteen years ago would be likely to give a fallaciously low value, and quite rightly has not used this method to calculate the error of his "randomized pairs": it is entirely due to this that he can draw conclusion (2) of his summary.

From this single fallacious conclusion he boldly generalizes to reach conclusion (3) which, as was shown by O. Tedin whom he quotes, is directly at variance with the facts. Conclusion (5) also follows solely from Prof. Fisher's faulty method and not from the balanced arrangement.

When the paper appeared I wrote a letter to *Nature** pointing this out, and that the actual error of the half-drill strip aggregate was in good conformity with that calculated from the weights of the whole strips.

In answering me Prof. Fisher replied that in that case the error of the "randomized sheaf weights" was so much smaller than that of half-drill strips that eleven times the area would have to be used to reduce the error of half-drill strips to that of "randomized sheaf weights" and further repeating his conclusion (4) with which I shall deal later.

Now one of the things that was noticed when uniformity trials first began was that the same piece of land laid out in large plots gave a very much larger error than if subdivided into small plots, and since half-drill strips were in this trial twelve times as large as "sheaf weights", Prof. Fisher's conclusion naturally follows since he is not comparing like with like.

Yet even so, those who have actually had to carry out agricultural experiments might very well prefer to work eleven times the area with ordinary agricultural methods and tools than have to sow and harvest 192 "randomized sheaf weights", if indeed that could be done at all under ordinary weather conditions.

Nevertheless, it is a fact that the error of this particular set of half-drill strips is unusually large. This arises partly because the number of repetitions is low but chiefly from the fact that the uniformity trial which Prof. Fisher chose to illustrate his argument showed a rather unusual feature due to faulty technique.

An examination of the original drills which were condensed to form the half-drill strips shows a periodicity, the averages of each eighth drill being for fifteen repetitions: 6739 7200 7839 6795 6689 7478 6897 6697

These variations are obviously not due to chance (for instance, the third drill

* [See pp. 218–19 below. Ed.]

gave the highest yield in twelve of the sets of eight and second highest in the other three) and are doubtless connected with some defect in the seed drill, probably the tines were not evenly spaced, and this could possibly have been detected had it occurred to Mr Wiebe to examine the working of the drill before sowing.

The result is that since six of the eight drills were added up to form a "half-drill strip", then one drill omitted, and then another six, and so on, there was a periodic variation in fertility not coinciding in period with the width of the half-drill strip, and this, as I pointed out in the Appendix to my Royal Statistical Society paper, increases the calculated error but does not bias the mean.

For the same reason the correlation between the corresponding sheaf weights is very much higher than would usually be the case and full scope is thereby given to Prof. Fisher's faulty method of calculating the error.

Let us now deal with Prof. Fisher's fourth conclusion: "The error estimated from a systematic arrangement is ambiguous and the experimenter has an arbitrary choice between several widely different estimates."

We may observe in passing that this is another instance of Prof. Fisher's passion for generalizing on somewhat narrow foundations, for the possibility which he refers to is peculiar to the half-drill strip arrangement.

In the half-drill strip, however, it is possible either to calculate the error from such aggregates as *ABBA* which I termed sandwiches in my paper to this *Journal* or from the separate parts of such aggregates, *AB* and *BA*, termed "pairs" by Prof. Fisher.

Of these the former is clearly the better if only there is a sufficient number of replications to give a good estimate of the error. As this is unusual it is generally best to give a degree of freedom to the fertility slope and calculate the error from "pairs".

Admittedly this tends to overestimate the error with the sort of results obtained in § 4. Faced with this choice, I personally choose the method which is most likely to be profitable when designing the experiment rather than use Prof. Fisher's system of *a posteriori* choice* which has always seemed to me to savour rather too much of "heads I win, tails you lose".

§ 3. A Properly Balanced Arrangement

It appears then that Prof. Fisher's paper is altogether irrelevant to the question at issue, but in order that Dr Barbacki's work may not be wholly wasted we can make a calculation of the error of a properly balanced arrangement of plots of the same size as the "randomized sandwiches" of which he has calculated the error.

For it will be noticed that Prof. Fisher's "systematic" arrangement, though "balanced" as "half-drill strips", is not so when regarded as a number of "sheaf weights": lateral balance is necessary.

* *Statistical Methods for Research Workers*, § 24.1 (5th ed.), p. 125.

The obvious layout is therefore to have the *ABBA* arrangement in both directions.

Thus:

A	*B*	*B*	*A*	*A*	*B*	*B*	*A*	*A*	*B*		*A*	*A*	*A*	*A*	*A*	*A*	*A*	*A*	
B	*A*	*A*	*B*	*B*	*A*	*A*	*B*	*B*	*A*		*B*	*B*	*B*	*B*	*B*	*B*	*B*	*B*	
B	*A*	*A*	*B*	*B*	*A*	*A*	*B*	*B*	*A*	etc. instead of:	*B*	*B*	*B*	*B*	*B*	*B*	*B*	*B*	etc.
A	*B*	*B*	*A*	*A*	*B*	*B*	*A*	*A*	*B*		*A*	*A*	*A*	*A*	*A*	*A*	*A*	*A*	
A	*B*	*B*	*A*	*A*	*B*	*B*	*A*	*A*	*B*		*A*	*A*	*A*	*A*	*A*	*A*	*A*	*A*	
B	*A*	*A*	*B*	*B*	*A*	*A*	*B*	*B*	*A*		*B*	*B*	*B*	*B*	*B*	*B*	*B*	*B*	
etc.											etc.								

This is merely a chessboard with fringes, each square being divided at harvest into four. The "squares" should be long and narrow, to gain the advantage of contiguity, and the comparisons should be made between adjacent long subplots of the different varieties. I have not seen this rather obvious arrangement mentioned before; it is admittedly no more suited for agricultural work than "randomized sandwiches", but it might be used in horticultural work, where the reduced "borders" would be of advantage, or for pot culture.

In this case we can start from Dr Fisher's Table II by reversing the signs of columns (ii), (iii), (vi), (vii), (x) and (xi) and calculate the error from an analysis of variance as follows:*

Variance due to	Degrees of freedom	Sum of squares of "split drill" differences
Longitudinal fertility slopes	12	887,171
Lateral fertility slopes	8	4,508,506
Varietal difference	1	2,741
Residual errors	75	3,988,681
Total	96	9,387,099

The difference between *A* and *B* is thus 513 g. and the S.D. of this difference 2259, as compared with 2353 calculated from "random sandwiches".

Thus, as we should expect, the difference is comfortably within the S.D., and the S.D. a little below that calculated from "randomized sandwiches", itself a partially balanced arrangement though random.

We see then that if a properly balanced arrangement is put down on the uniformity experiment of Dr Fisher's choice the error is found to be, as usual, less than his random arrangement, though not by much since "sandwiches" are themselves balanced.

* [The basis of this analysis does not seem quite clear. It was a point on which "Student" had promised to enlarge before the final presentation of the paper; see the editorial note on p. 199 at the head of this article. Ed.]

§ 4. The Effect of "Balancing" on the "Validity" of Conclusions

From *a priori* considerations—and Mr Hudson's and Mr Borden's experiments are in accordance with this expectation—it seems fairly certain (i) that "balancing" has no tendency to bias the mean, and (ii) that when there is a "fertility slope"—or anything corresponding to it, e.g. a time effect—the result will be to increase the apparent error but to decrease the real error. What effect has this on the "validity" of conclusions drawn from balanced experiments?

(i) *The case of blocks, randomized or balanced, judged by the z test*

Let us take the case of four treatments in six blocks giving fifteen degrees of freedom to the residual error and three for treatments, and let us suppose the arrangement put down on a uniformity trial.

Then, once the plots and blocks are marked out, the "total sum of squares" and the "sum of squares due to blocks" are fixed; the difference between these represents in all cases the eighteen degrees of freedom due to treatments and residual error, but will be divided between the two in different proportions according to the chosen arrangement of the treatments in the blocks. If the arrangement is random the frequency of any particular ratio is known to follow the z distribution, and owing to the skewness of this there will more often than not be a lower variance of the treatments with three degrees of freedom than of the residuals with fifteen.

If the arrangement is not random the frequencies will not follow the z distribution, e.g. with regular unbalanced arrangements the variance "due to treatment" will tend to be high compared with that of "residual error", while with regular balanced arrangements the reverse is the case. It will therefore be of interest to see what happens when a real "variance due to treatment" is imposed on uniformity trials which give ratios at different points of the z scale.

Thus it may be convenient to take as norm those uniformity trials which have the same variance for "means of treatments" as that calculated from the residuals and let this variance be σ^2. Then another set of trials may be considered of which the means have a variance of $0{\cdot}5\sigma^2$ and consequently a variance of "residual error" of $1{\cdot}1\sigma^2$, since $15 \times 1{\cdot}1 + 3 \times 0{\cdot}5 = 18$. This set may be taken to represent the tendency of balanced arrangements to produce low variance "due to treatment". A third set representing "unbalanced" arrangements may be taken with a means variance $1{\cdot}5\sigma^2$ and a variance calculated from residuals of $0{\cdot}9\sigma^2$.

All three of these occur, of course, in their proper proportions in random trials and are none of them uncommon. They are merely taken here as types.

In what follows I shall for convenience term the variance of means the *actual* variance of error, σ_e^2, and the variance calculated from residuals the *calculated* variance of error.

Now suppose that a real variance due to treatment—measured without error, σ_T^2—be superposed upon the uniformity experiment. Then the calculated variance of error will be unaffected and the observed variance due to treatments will be $\sigma_T^2 + \sigma_e^2 + 2r_{eT}\sigma_T\sigma_e$ and, since T and e are independent, the distribution of the observed variance can be calculated from the known distribution of r when there is no correlation, which in this case of four treatments is uniform between $+1$ and -1.

From this we can determine the probability that any given σ_T^2, superposed on any particular arrangement, will be deemed "significant" when compared with the corresponding "calculated variance of error".

The results of such calculations are given in the following table, which gives the probability of exceeding the 5 % limit of significance, or if preferred can be read as the percentages of "significant" results.

Value of σ_T^2/σ^2	Probability of obtaining significant result		
	Actual variance of error		
	$1{\cdot}5\sigma^2$	$1{\cdot}0\sigma^2$	$0{\cdot}5\sigma^2$
	Limit of significance		
	$2{\cdot}96\sigma^2$	$3{\cdot}29\sigma^2$	$3{\cdot}63\sigma^2$
0·5	0·22	0	0
1·0	0·41	0·18	0
1·5	0·51	0·34	0·03
2·0	0·58	0·45	0·22
2·5	0·63	0·53	0·36
3·0	0·68	0·60	0·48
3·5	0·72	0·66	0·57
4·0	0·76	0·71	0·66
4·5	0·79	0·76	0·73
5·0	0·82	0·80	0·80
5·5	0·85	0·84	0·86
6·0	0·88	0·88	0·92
6·5	0·90	0·91	0·97
7·0	0·93	0·94	1·00
7·5	0·95	0·98	—
8·0	0·97	1·00	—
8·5	0·99	—	—
9·0	1·00	—	—

This table illustrates the fact that arrangements which give an actual error less than the calculated fail to give as many "significant" results as those which give larger actual errors up to a real treatment variance of about five times the average residual variance, at which point about 20 % of the experiments still fail to show significance in each case. When the real treatment variance rises above this point, the smaller the actual error the *more* are the significant results.

It is perhaps rather invidious to decide below what value of the real treatment variance "significant" results are misleading, but in any case it is clear that the fault of the arrangements with low actual variance is not lack of validity. On the contrary, conclusions drawn from experiments giving significant results by such arrangements are *more* valid in the ordinary sense of that word.

These arrangements have so far been considered as having arisen in a random manner, but by using balanced arrangements the proportion of arrangements having actual low errors is increased, and hence conclusions arrived at from balanced arrangements are more, not less, valid.

Nevertheless, it is clear that if it is required to calculate the error from an experiment carried out at a single station it is advisable not only to balance the experiment but to allow for the error eliminated by allocating a degree of freedom to the fertility slope. Even so it is likely that the actual error will be less than the calculated and the conclusions more valid than they appear to be.

(ii) *The case of half-drill strips judged by the t test*

I showed in the Appendix to my paper on Co-operative Experiments that it is usually advantageous to allot one degree of freedom to the fertility slope, and that since fertility slopes are not usually strictly linear there is a tendency for the calculated error to be larger than the actual error. Let us illustrate this in the case of experiments carried out on the scale adopted by the N.I.A.B., namely, with ten pairs of comparisons; this is of course rather a small scale, and of the nine degrees of freedom one is allocated to the fertility slope and eight to the residual error of comparing the two varieties.

In this case we are to vary, not the position of treatments on a given piece of ground, but the pieces of ground on which a half-drill strip of ten pairs is set and the "norm" which we shall take is the case where, owing to a particular uniform fertility slope, the calculated and the actual error exactly correspond with the standard error σ.

With this we can compare a case where the variance of actual error is $0{\cdot}5\sigma^2$ and the calculated error therefore $\left(1+\frac{0{\cdot}5}{8}\right)\sigma^2 = 1{\cdot}062\sigma^2$, i.e. standard errors $0{\cdot}71\sigma$ and $1{\cdot}03\sigma$. A tendency in this direction is, as noted above, common, since fertility slopes are naturally not uniform; on the other hand, when the fertility slope is small, random sampling may give us a case where the actual error is larger than the calculated, let us say standard errors of $1{\cdot}22\sigma$ and $0{\cdot}97\sigma$.

Then in the three cases we find from the t table that the 5 % significance point is for the "norm" $2{\cdot}30\sigma$, for the low actual error $2{\cdot}37\sigma$, and for the high actual error $2{\cdot}23\sigma$, while the actual errors are distributed normally with S.E.'s. σ, $0{\cdot}707\sigma$ and $1{\cdot}22\sigma$ and the percentage of "significant" results, i.e. those above the

significant point calculated above, can be readily determined for values of the real (i.e. measured without error) differences between the two "varieties", say $A - B$.

These are given in the following table.

Variance of calculated error	$0{\cdot}94\sigma^2$	$1{\cdot}0\sigma^2$	$1{\cdot}06\sigma^2$
Variance of actual error	$1{\cdot}5\sigma^2$	$1{\cdot}0\sigma^2$	$0{\cdot}5\sigma^2$
S.E. calculated	$0{\cdot}97\sigma$	$1{\cdot}0\sigma$	$1{\cdot}03\sigma$
S.E. actual	$1{\cdot}22\sigma$	$1{\cdot}0\sigma$	$0{\cdot}707\sigma$
Limit of significance	$2{\cdot}23\sigma$	$2{\cdot}30\sigma$	$2{\cdot}37\sigma$
Value of $\frac{A-B}{\sigma}$	Probability of significant results		
0	0·07	0·02	0
0·5	0·01 0·08	0·04	0
1·0	0·16	0·10	0·03
1·5	0·27	0·21	0·11
2·0	0·42	0·38	0·30
2·5	0·59	0·58	0·58
3·0	0·74	0·76	0·81
3·5	0·85	0·88	0·95
4·0	0·93	0·96	0·99
4·5	0·97	0·99	1·00
5·0	0·99	1·00	—
5·5	1·00	—	—

It will be noticed that in the left-hand column there are two probabilities given opposite 0·5, 0·01 that a negative significant result and 0·08 that a positive significant result will be obtained. Fortunately such a case is almost impossible unless of course "randomized pairs" were used instead of a half-drill strip. What we are concerned with in practice is something which tends towards the right-hand column which, as in the case of the balanced blocks, errs by failing to give significant results when the difference to be measured is small, but from a value of about 2·55—at which all produce significant results in 60 % of trials—gives a higher percentage than when the calculated and actual errors are equal.

It is clear, therefore, that in this case too, conclusions drawn from a balanced arrangement are not less but more valid than if the arrangement had been random.

The above tables rather emphasize the well-known paradox that it is just when the experimenter is congratulating himself on the unusual smallness of his experimental error—unusual, that is, for the type of experiment and number of replications—that he is most likely to be betrayed into drawing false conclusions: for the small calculated error indicates a large actual error, and this whether the arrangement be random or balanced, though it is likely to occur more frequently in the random.

In conclusion, I should like to emphasize the fact that when using the phrase criticized by Prof. Fisher I was concerned with co-operative experiments carried out at a number of different places.

Such experiments, as indeed all agricultural experiments, are only of value in so far as the venue is representative of the conditions under which the results of the experiment are to be applied, and so the result at any single station is not of any particular importance in itself but only in its interaction with the results obtained at the other stations, for only so can its representative nature be established.

To take a simple case a variety trial may indicate that one wheat will do better than another in heavy but not in light soils; such a conclusion is more likely to follow from an experiment carried out with a low real error and a correspondingly high calculated error at the individual stations than if a low calculated error gave "significant" results sporadically.

It is therefore important that the results should be determined with as little real error as possible, and the calculated error at each station is superseded by the error of the experiment as a whole.

APPENDIX GIVING MR A. W. HUDSON'S COMPARISONS OF RANDOM AND REGULAR ARRANGEMENTS IN UNIFORMITY TRIALS

Mr Hudson's account of his procedure is as follows:

"(i) Four, five or six imaginary treatments were allocated according to which was the most suitable to the full utilization of the data.

"(ii) These were allocated to blocks in a regular-balanced fashion and then to the same blocks randomwise, using various numbers of 'units' per individual plot.

"The regular arrangements were balanced by using two or four series in which the treatments in the second and fourth series were in opposite order to those in the first and third, thus:

1, 2, 3, 4, 1, 2, 3, 4, etc.
4, 3, 2, 1, 4, 3, 2, 1, etc.
2, 1, 4, 3, 2, 1, 4, 3, etc.
3, 4, 1, 2, 3, 4, 1, 2, etc.

or alternatively, where the shape of the individual plot permitted, only a single series, thus:

etc., 2, 1, 4, 3, 2, 1 Middle 1, 2, 3, 4, 1, 2, 3, etc."

Mr Hudson's experimental work must not be taken as an attempt at a proof that balanced arrangements are likely to give a lower error than random unbalanced arrangements; that seems to me obvious, and it is for those who wish to disprove the obvious to obtain evidence in support of their eccentric opinions, but it does give an interesting illustration of what is likely to happen in practice, and I print it in the hope that it will help to clarify other people's ideas as it has mine.

TABLE I

Data from Journal of Agricultural Science, Vol. IV, Part 2, 1911
Mercer and Hall. Mangold Plots

Number of rows ... 20
Units per row ... 10
Total number of units 200, but only 160 used in first three.

B./Tr.	R. × U.	G.M.	Random			Balanced		
			Calculated S.E.	Dev. of T.M. from G.M.	Actual S.E.	Calculated S.E.	Dev. of T.M. from G.M.	Actual S.E.
20/4	1 × 2	656·4	6·63	− 3·3 − 5·7 + 1·6 + 7·5	5·84	6·73	+ 4·4 − 1·0 − 1·7 − 1·7	2·95
10/4	2 × 2	1312·8	14·16	+10·2 − 4·1 −12·2 + 6·3	10·15	14·42	− 0·8 + 8·3 − 1·9 − 5·4	5·54
10/4	1 × 4	1312·8	16·40	+12·7 −18·0 + 7·5 − 2·0	13·48	16·61	+16·3 − 5·8 − 6·8 − 3·5	10·92
8/5	1 × 5	1642·9	21·62	−32·8 −20·0 +27·8 + 4·3 +20·5	25·9	22·92	+15·0 −16·2 − 5·5 +19·7 −13·2	16·4
4/5	2 × 5	3285·7	50·78	+55·0 −25·0 − 4·2 +43·0 −68·7	50·6	54·62	+15·3 −15·7 −53·2 + 9·3 +44·5	36·7

Table headings: B./Tr. Blocks (replications) and treatments.
R. × U. Size of plot, rows × units.
G.M. General mean of all plots.
Calculated S.E., i.e. of means of treatments by analysis of variance.
Dev. of T.M. from G.M. Deviation of treatment means from general means.
Actual S.E., i.e. calculated from previous column.

TABLE II

Data from Journal of Agricultural Research, Vol. XLIV, *No.* 8, *April* 1932
F. R. Immer. Yields of sugar beet

Number of rows 60
Units per row 10
Total number of units ... 600

B./Tr.	R. × U.	G.M.	Random			Balanced		
			Calculated S.E.	Dev. of T.M. from G.M.	Actual S.E.	Calculated S.E.	Dev. of T.M. from G.M.	Actual S.E.
20/6	1 × 5	255·9	3·28	+ 1·1 + 3·3 + 0·6 + 0·4 − 6·2 + 0·8	3·22	3·27	− 1·1 − 3·7 − 2·1 + 6·1 + 1·1 0	3·40
10/6	2 × 5	511·9	8·42	+11·0 +13·0 − 1·6 − 7·2 +13·0 − 6·2	10·5	8·52	+ 7·5 + 5·1 − 5·8 −17·3 + 2·6 + 7·8	9·8
10/6	1 × 10	511·9	8·04	− 6·1 + 7·4 − 4·4 + 9·5 − 0·4 − 6·0	6·90	8·11	− 6·2 − 1·7 +10·4 − 2·7 − 3·7 + 3·7	6·08
4/6	5 × 5	1279·7	23·68	+68·2 +40·5 −41·6 − 6·3 −73·3 +12·5	52·1*	37·87	+11·4 + 5·4 −17·0 + 2·1 − 5·4 + 3·6	10·0

* This is a "significant" result—beyond the 1% level—and it is perhaps a little unfortunate that it should have occurred in a mere sample of 21. It has, however, been checked both by Mr Hudson and myself.

TABLE III

Data from Journal of Agricultural Science, Vol. XXII, Part 2, April 1932
Kalankar. Potatoes

Number of rows ... 96 Units per row ... 6 Total number of units ... 576

B./Tr.	R. × U.	G.M.	Random A			Random B			Balanced		
			Calc. S.E.	Dev. of T.M. from G.M.	Actual S.E.	Calc. S.E.	Dev. of T.M. from G.M.	Actual S.E.	Calc. S.E.	Dev. of T.M. from G.M.	Actual S.E.
32/6	1 × 3	69·8	0·74	− 0·2 + 0·4 − 0·1 − 0·6 − 0·3 + 0·8	0·51	0·74	− 0·6 − 0·1 − 0·3 + 0·2 + 0·6 + 0·3	0·44	0·74	− 0·5 + 0·3 + 0·9 − 0·1 + 0·3 − 1·0	0·67
16/6	1 × 6	139·6	1·52	− 1·9 0 + 2·7 − 1·9 + 0·7 + 0·4	1·74	1·49	+ 1·1 + 1·8 − 3·1 + 1·2 + 1·1 − 2·1	2·05	1·55	− 1·4 + 1·3 + 0·9 + 0·7 0 − 1·5	1·20
16/6	2 × 3	139·6	2·16	+ 0·2 − 1·1 − 2·9 − 0·1 + 0·4 + 3·5	2·10	2·19	− 2·8 + 1·5 + 0·8 + 0·2 + 0·1 + 0·2	1·52	2·20	+ 0·2 − 0·3 − 2·1 − 0·5 + 2·0 + 0·8	1·38
8/6	2 × 6	279·2	5·35	+ 8·8 + 0·9 − 5·4 − 1·4 − 2·8 − 0·1	4·84	5·47	+ 0·8 + 6·0 + 2·1 − 3·8 − 1·1 − 4·1	3·83	5·56	+ 3·0 + 3·3 − 2·0 − 3·3 0 − 1·0	2·68
8/6	4 × 3	279·2	5·67	+ 1·3 + 6·1 + 1·1 − 3·9 − 3·5 − 1·1	3·71	5·58	+ 8·0 − 1·3 − 3·8 − 0·1 − 4·4 + 1·6	4·52	5·60	+ 2·4 − 4·4 − 2·2 − 2·7 − 0·9 + 7·7	4·42
4/6	8 × 3	558·4	33·3	− 12·1 − 7·2 + 41·0 + 35·8 − 19·0 − 38·7	31·7	35·85	− 5·9 + 39·0 + 10·8 − 16·9 − 12·5 − 14·6	21·6	37·76	− 3·3 − 10·8 + 5·6 − 0·8 + 1·1 + 7·8	6·7

MISCELLANEOUS CONTRIBUTIONS

A. LETTERS TO *NATURE*

(i) Agricultural Field Experiments

[*Nature*, cxxvi (29 November 1930), p. 843]

In the article with the above title which appears in *Nature* of 25 October last, p. 667, it is stated:

"Beaven's half-drill strip method is described, but without pointing out its two serious but remediable defects: that the continued use of one half of the drill for one variety, and of the other half for the variety with which it is to be compared, may introduce a constant difference the magnitude of which cannot be estimated; and that the regular alternation of strips of the two varieties does not permit of a valid estimate of experimental error."

I submit that these defects are more theoretical than practical, and that any modification of practice in the application of the method, such as changing over seed boxes, would be a retrograde step.

To take the first, there are three possible ways in which one half of a drill may differ from the other:

(1) It may cover a wider breadth of ground; this would doubtless have an appreciable effect, but it would be detected and allowed for by the routine measurements taken across the stubble.

(2) The coulters may be less evenly spaced than those of the other, and

(3) Less seed may be drilled from it than from the other.

Now, cereal crops are wonderfully independent of the amount of seed sown. I have in mind two chessboard experiments, in one of which half the area was sown with seed 1 in. apart instead of the usual 2 in., and in the other, the rows in half the experiment were 3 in. apart instead of 6 in. In each case the heavier seeding only resulted in a gain of about 3 %, and it is not to be expected that such slight irregularities as occur between the two halves of a drill would have any measurable effect.

The second defect, owing to the peculiar shape of the half-drill strip, would only exist if the experiment were to be sited so that some periodic variation existed across the breadth of the drills: otherwise randomness is supplied by the soil. By taking care that the experiment is drilled across ploughman's "lands" if they exist, and by bearing in mind the history of the last few crops, this danger can be avoided.

The pairs of strips fall naturally into two sets according as one or other variety is on the right hand, and in an analysis of the variance of the difference between varieties, one degree of freedom is taken up by these two sets. The estimate of the experimental error arrived at in this way is perfectly valid, provided the above precautions have been taken in siting the experiment.

It would be a pity to interfere unnecessarily with the simplicity of this very efficient method of conducting field trials.

(ii) Agricultural Field Experiments

[*Nature*, cxxvii (14 March 1931), p. 404]

Mr Howard's letter in *Nature* of 31 January last (p. 166) gives interesting confirmation of the reviewer's opinion in *Nature* of 29 November, 1930, p. 843, that depth of sowing influences the yield of wheat, yet I venture to suggest that such an extreme case as he quotes scarcely bears upon the point at issue. When seeds do not germinate, it is equivalent to a light seeding rate, which, as I pointed out, makes wonderfully little effect on the yield. Whether such differences as one may expect to occur between the depths of coulters in the same drill make any appreciable effect on the yields of the different rows is still, I think, an open question, and I suggest that the differences which the reviewer has observed between the yields of his rows may have been due to their being unevenly spaced. The yield which is comparatively unaffected by seeding rate, is that per areal and not that per linear unit. The reviewer quotes "an apparently uniform field" at Aarslev as upsetting my view that for practical purposes randomness can be obtained from the half-drill strip "provided care is taken to drill across ploughman's 'lands'", if they exist; yet Dr Sanders in his account of that experiment makes no mention of an "apparently uniform field" (*J. Agric. Sci.* xx, p. 65), but writes, "This oscillation apparently arose as a legacy of the old practice of ploughing in high ridges", and so on.

Even if the unsuitability of the field had been overlooked, the Aarslev plots were probably a good deal wider than drill width, and half-drill strips would have been extremely unlikely to coincide both in breadth and phase with the periodicity in question, while any partial coincidence would have betrayed the existence of the snare.

Finally, there is a fallacy in Mr Howard's last sentence—"It is obvious in such questions that nothing can be gained by the application of formulae and figures to the results obtained by poor agriculture." There is no question, of course, of connecting the half-drill strip method of experimenting with poor agriculture; its great merit lies in the fact that in its present form it is ordinary farming practice: if, however, that practice were poor agriculture, it would be

a mistake to carry out trials by methods conforming to better standards: field trials must be capable of being considered a random sample of the practice, not of the theory, of agriculture.

This may seem a hard saying, but an example will make my meaning clear. After a long series of experiments the Irish Department of Agriculture decided to introduce Dr Hunter's Spratt-Archer barley as being the best suited for the country. This was almost everywhere a great and outstanding success; yet in one district, which shall be nameless, the farmers refused to grow it, alleging that their own native race of barley was superior to it. After some time the Department, to demonstrate Spratt-Archer's superiority, produced a single-line culture of the native barley and tested it against the Spratt-Archer in the district in question. To their surprise, they found the farmers were perfectly right: the native barley gave the higher yield. At the same time the reason became plain: the barley in question starts more quickly and is able to smother the weeds, which flourish in that not too well farmed land; Spratt-Archer, growing less strongly at first, is, however, the victim and not the conqueror of the weeds, and the original experiments, carried out on well-farmed land, were definitely misleading when their conclusions were applied elsewhere.

Taught by experience, the Department is now engaged in breeding a barley to meet their conditions; and this barley, when obtained, will rightly be tested by "results obtained by poor agriculture".

(iii) The Half-Drill Strip System Agricultural Experiments

[*Nature*, cxxxviii (5 December 1936), p. 971]

Prof. R. A. Fisher and Dr Barbacki have recently published a paper in the *Annals of Eugenics* entitled "A test of the supposed precision of systematic arrangements".* There is a good deal in the paper with which I am not in agreement and with which I hope to deal elsewhere, but a letter from a friend of mine in Australia, who has heard at second-hand that Fisher's "results showed not only that the half-drill strip failed to give a valid estimate of error but was less accurate", shows that it would be better not to let such rumours get a start, for they are quite unfounded.

In the paper, the crop on a uniformly treated field was assigned to two imagined treatments A and B on a systematic plan in which eight strips of the width of a half drill were assigned to A, and eight to B, in the usual arrangement of an eight comparison half-drill strip experiment. Apart from the fact that one

* Barbacki and Fisher, "A test of the supposed precision of systematic arrangements", *Ann. Eug.* vii, Part 2 (1936).

should have at least ten comparisons—in Beaven's original paper* there were 26—the representation is a fair one.

The authors, for the purpose of ascertaining the degree of precision which is obtainable from the systematic arrangement in question, have taken the weights of grain, not from the total area of each of the 16 strips, but from 12 sections of each strip, and have treated these 192 sections as if they were independent half-drill strips—in fact they have called them half-drill strips—and from the 96 comparisons they have calculated a standard error to represent the precision which they suppose an advocate of systematic arrangements would attribute to the method. But, of course, the sections of a half-drill strip are not in fact independent, and in this case are markedly correlated, so that the figure which they obtain is much too small to account for the observed difference between the *A*'s and the *B*'s—and they draw conclusions adverse to the systematic arrangement and not to their own method of calculation.

The procedure adopted, of dividing up the long strips, is that which Dr Beaven* originally proposed in 1922, namely, weighing the sheaves off equal segments of his half-drill strips and calculating the error from these weights; but so early as 1923, I pointed out† that this method would probably give a fallaciously small value, and since then it has been customary to regard the whole length of the strip as the unit in the calculation.

Had Prof. Fisher and Dr Barbacki calculated the error on that basis,‡ they would have found a standard error of 2·37 % of the average yield, while the actual difference between the *A*'s and *B*'s amounts to 1·75 %; that is, the difference between two things which should be the same within the error of random sampling is in fact no more than 0·75 times the standard error.

The authors' practical demonstration of the correctness of my *a priori* reasoning is, of course, very gratifying to me, but I must nevertheless insist that their paper has no bearing whatever on the error of present-day half-drill strip experiments.

* Beaven, "Trials of new varieties of cereals", *J. Minist. Agric.* XXIX, Nos. 4 and 5 (1922).

† "Student", "On testing varieties of cereals", *Biometrika*, XV (1923), pp. 286, 287, [11, p. 106].

‡ "Student", "Yield trials", Baillère's *Encyc. Sci. Agric.* II (1931), [15]. "Co-operation in large-scale experiments", *J. Roy. Statist. Soc.* Supplement (1936), [20].

B. CONTRIBUTIONS TO DISCUSSIONS AT MEETINGS OF THE INDUSTRIAL AND AGRICULTURAL RESEARCH SECTION OF THE ROYAL STATISTICAL SOCIETY

(i)

[Supplement to *J. Roy. Statist. Soc.* (1934), I, p. 18]

MR GOSSET said that Dr Pickard had given such a wide and comprehensive survey of the application of statistical methods to industry that, in spite of having had a considerable experience himself, there was practically nothing that he could add on the subject. He had started with the raw material in the field, and ended with the finished cloth, and in his own particular industry he had had similar problems from one end to the other.

He would like to refer to the question raised at the end of the paper—the selection of the statistician for industry. In his firm, a man who had had some experience of the industry had been sent out and taught statistics. That had happened some time ago—in fact it was twenty-eight years since he had ridden across the Berkshire Downs on a bicycle to interview Prof. Karl Pearson in 1905. On the whole, this had been found to be a good method, and perhaps because they had been working at it for so long, they did not experience the difficulty of the horrible jargon referred to by Dr Pearson, and it did not appear to produce quite such terrors even among the senior members of the firm. They more or less understood, and if they did not understand they were quite polite about it.

If a man were sent out from the industry and put to school again, he was apt to forget what he had learned, and it was most important that such people should be in constant touch with their Professors. As Dr Pearson had pointed out, one reason was that the mathematical tools which the Professors provided would hardly be exactly what were wanted unless they knew how they were to be used.

Another point arose from the peculiar nature of statistics. It was impossible to apply statistical methods to industry or anything else unless one had a certain amount of intelligent experience as a background. That worked both ways. The practical man had to go and talk to his Professors partly in order that the Professor himself should share his experience. In actual fact all statistical methods were strictly inapplicable to practical affairs; they all depended upon random samples and, as everyone knew, there were no such things. That, of course, was an exaggeration; there were two random phenomena, one of which was the disintegration of radioactive elements, and the other was Tippett's numbers. The whole art of statistical inference lay in the reconciliation of random mathematics with biased samples. Every new problem had some fresh kind of bias

and might contain some new pitfall. The only way not to fall into these pitfalls was to talk over the problem with some intelligent critic; and so the practical man, if he were not entirely foolish, talked over his problems with the Professor, and the Professor would not consider himself to be a competent critic unless he had had some experience of applying the statistics to industry, and had learned the difficulties of that application.

(ii)

[Supplement to *J. Roy. Statist. Soc.* (1936), III, p. 173]

MR GOSSET said he would like to confirm a remark of Prof. Pearson's about the difficulty of working with large-scale results. In most cases the whole object —or one of the principal objects—of manufacture is to keep the product as uniform as possible. In addition to that and in order to obtain that, it is necessary to keep the raw materials as constant as possible; consequently, when one looks at large-scale results, there is no variation to work upon, and the statisticians are helpless, at any rate until something has gone wrong.

Mr Gosset said that up to the present he had been interested in spectacle glass only as a consumer, and his excuse for intervening in this discussion was that he could illustrate the use of a simple statistical method on the tables which were given at the end of the paper.

In an investigation such as this, where one wished to throw light on the behaviour of a large-scale process, the method of correlation was very often useful, but at first sight the tables did not look very promising, split up as they were into very small samples, both by the small numbers of journeys per pot and the different kinds of glass. In this connexion he would say to Mr Jennett that there were two uses in correlation, one was the use of the regression line, and that was doubtless the best, and the other its use merely as a measure of the relation between the two things.

There was a method of correlation used largely by psychologists, known as "Spearman's method"; it was not an efficient method—that is, it did not utilize all the information supplied by the samples, so that about 20 % larger samples must be collected to give as accurate a result as the ordinary correlation coefficient, yet, owing to an artful method of calculation, it was so simple that when playing with *other* people's figures, for instance on a railway journey, it was the obvious one to use. It consisted of replacing each variate by the figure representing its numerical order, and correlating these numbers.

By this method, Mr Gosset said he had obtained weighted average correlation coefficients between the number of veins and the order of the journey, which put it beyond all question that the later the journey the worse the veins. This weighted average was derived from all the 98 samples discoverable in the tables:

the mean size of sample was just over 4, the greatest was 8 and the smallest 2. The average results were as follows:

For *A*	0·31	2·6	times its	standard	deviation
„ *B*	0·27	3·1	„	„	„
„ *C*	0·44	2·2	„	„	„
„ *D*	0·11	0·8	„	„	„
„ *E*	0·19	1·3	„	„	„
Total	0·26	4·5	„	„	„

A, *B*, and *C* were all significant. *D* and *E* were not so, but there was no evidence that any glass behaved differently from the others. When he said "standard deviation" it was calculated on the supposition that there was no correlation at all. It meant the standard deviation of correlation coefficients of samples of the appropriate degrees of freedom drawn from uncorrelated material, and the mean 0·26 corresponded to a correlation coefficient of about 0·30 if large samples had been obtainable.

This did not confirm the authors' conclusion, and Mr Gosset could offer no opinion as to the disagreement unless it was the custom to stop using, at an early stage, pots which had given poor results. A similar investigation into seeds showed that there was no evidence of correlation between seeds and order of journey except in the case of glass *A*, where the correlation was 0·27, 2·3 times the standard deviation.

He had also tested the correlation between refractive index and both seeds and veins, the former without any success, but there was a distinct indication that the higher the refractive index, the worse the veins; perhaps the veins themselves had a low refractive index. The evidence was not significant, since the correlation coefficient 0·16 was but 1·6 times its standard deviation, but if the matter was of any importance, this might give a line for further investigation.

Mr Gosset again expressed his great interest in the subject-matter of the paper.

(iii)

[Supplement to *J. Roy. Statist. Soc.* (1937), IV, p. 89]

Mr Gosset wished to say a word for the control chart. It had been talked about as a sort of wall ornament, but in point of fact it was a very useful thing. He had had control charts in the laboratory which had led up to nearly halving a laboratory error, because they gave a hint as to what to look for.

And in this discussion, although the method of testing the strength had been aspersed, it was clear from the control chart that the method was good enough to show secular changes, unless indeed, as was unlikely, the secular changes were due to the testing machine itself.

(iv)

[Supplement to *J. Roy. Statist. Soc.* (1937), IV, p. 170]

ON reading Mr Bartlett's paper, I saw that I could add little or nothing to his treatment of the statistical principles involved, but it occurred to me that other people besides myself might have had their curiosity aroused by certain matters of less interest perhaps statistically but yet of some practical importance. I refer of course to the results of the experiments. I therefore wrote to Mr Bartlett, who very kindly sent me his copies of the four papers in the list of references, with which Dr Crowther's name is associated, and I am going to give an account, necessarily inadequate, of the fine piece of work which they describe.

The four papers deal primarily with the cotton crop in Egypt, particularly in the Delta. Cotton is grown in Egypt as an annual and not, as might be expected, as a perennial, because of the pests by which it is afflicted, especially the Pink Boll Worm. This has so much increased of late years that the methods of cultivation have during the last ten years been modified throughout the country. At the same time, new varieties have been introduced, the tendency being to produce larger yields of cotton of shorter staple. That being so, it became necessary to examine how far these changes have altered the old standards of manuring and, particularly, what profit was to be derived from nitrogenous manures.

The experiments directed by Dr Crowther were concerned mainly with the elucidation of this question and, as you have heard from Mr Bartlett, were carried out at several stations, where the effects of various levels of nitrogenous manuring were compared under different conditions of spacing, watering and phosphate manuring, and with different varieties of cotton.

The actual gain from the use of nitrogen varied with the spacing adopted, with the different varieties and, naturally enough, between the different stations, but the average profit from the use of nitrogenous manures was over £3 per acre, and at only one out of eight stations was the profit not appreciable. Had the optimum quantity of nitrogen been used, the gain would have been considerably more. Furthermore, an experiment with wheat following cotton at a single station showed that in that case the increased yield of wheat more than paid for the nitrogen applied to the cotton. I think that is a very good instance of what large gain can be made: £3 an acre on all the cotton of Egypt would produce an enormous amount of money.

These results may not seem to be very surprising until you learn (*a*) that previously it was generally believed that nitrogen was of little or no value to

the cotton crop, and (*b*) that in Egypt nitrogenous residues were supposed to be leached out by the irrigation water.

An investigation into the relation between the supply of nitrogen and the development of cotton leads Dr Crowther to the opinion that it is largely owing to the closer spacing of modern practice that the plant can make good use of added nitrogen, but I should like to ask whether the substitution of the modern nitro-chalk for nitrate of soda or ammonium sulphate may not also have had a beneficial effect of its own.

I have now much pleasure in moving a very hearty vote of thanks to Mr Bartlett for his paper, and if I have rather strayed from the straight and narrow path which he has himself followed, I have done so in the confident expectation that my lapse will be atoned for by the speakers who will follow.