Fisher to F.R. Immer: 13 *May* 1935

I have not happened to see Treloar and Wilder's paper,[1] nor have I heard it mentioned in England. I presume it must be a re-hash of the objections based on departure from normality of the parent population which were much fostered in the Statistical Department of this College,[2] some years ago, but which seem now to have subsided, perhaps as a result of my dealing sometimes with these questions in lectures.

Logically, the complete answer to this objection, in respect of any particular sample of data, is to show that the sample will yield a test of significance very close to that given by t or z, without introducing any normal distribution at all, but using merely the known arithmetical distribution produced by randomisation.

In the case of the z test, you may perhaps remember Eden and Yates[3] taking a set of 32 values for 4 varieties in 8 blocks, and finding the z distribution for 1000 randomly chosen reassignments within the blocks. Their object was to examine the validity of the z distribution in material which was rather more skew, and more unequally variable than most observational material, and in this their test was completely successful. It has, however, a rather deeper interest logically since, on the hypothesis to be tested, that the 4 varieties are without effect, the experimenter could know that the 1000 values of z obtained by sampling were, in fact, all equally probable, whatever the nature of the parent population. And if the value of z actually observed falls among the top 50 of the sample, he knows, as an arithmetical fact, and without any statistical theory, that values so high as this are sufficiently rare to be regarded as significant. Anyone, therefore, who wishes to criticise a conclusion as to significance drawn by the t or z tests, on the ground that the population sampled is not normal, or not equally variable, should be invited to see for himself, by direct, though tedious arithmetic, whether the conclusion indicated by the statistical test is valid or not, when no assumption whatever as to normality is made.

I have explained and illustrated the process in detail in a book now in the press, called *The Design of Experiments*, which I think may interest you also in other ways.

[P.S] I have now seen Treloar's paper, which does not seem to be so intelligent as I had guessed. Of course t is a ratio, that is what it is for. I do not think it has ever been overlooked. He is merely comparing a test of significance t with an *estimate* $\bar{x}$ which appears in the numerator.

[1] Treloar, A.E. and Wilder, M.A. (1934). The adequacy of 'Student's' criterion of deviations in small sample means. *Ann. Math. Stat.* **5**, 324–41. Immer had written seeking Fisher's opinion of this paper. (See also Fisher's letter of 20 May 1935 to E.B. Wilson, p. 237.)

[2] University College London.

[3] Eden, T. and Yates, F. (1933). On the validity of Fisher's z-test when applied to an actual example of non-normal data. *J. Agric. Sci.* **23**, 6–16.

Fisher to J.O. Irwin: 28 *May* 1941

. . . About χ^2, the distribution is certainly unaffected by making a sub-selection at random of the contents of each cell (plate).[1] I recall giving a good deal of attention at the time to this subnormal variance, which I think we never cleared up experimentally. The kind of factor which would produce such an effect is competition. The distribution of passengers in compartments of a train is notably subnormal, because late-comers try to avoid the more crowded carriages. A bacterial species producing a toxin injurious, at least at germination, to the same species would tend to occupy points on the plate on a lattice, the scale of which is independent of the size of the plate. Consequently, sufficiently large plates and sufficiently dense suspensions would give subnormal χ^2.

[1] Irwin had enquired about the occurrence of subnormal χ^2 in bacterial plate counts and the suggestion by Fisher, Thornton, and Mackenzie (1922) that this was 'indicative of some defect in the composition of the medium' (*CP* 22, p.358). Irwin said that whilst χ^2 would be reduced if a set of frequencies were cut down proportionately, he could not see that χ^2 would be affected if the 'mortality' were itself subject to sampling error.

H. Jeffreys to Fisher: 19 *May* 1938

. . . I take it that we should all agree that however we make a selection, there will occasionally be cases where the selection will be correlated with something that we are trying to find, and either the estimate or its standard error will be biased. General rules for assigning at random presumably express a belief that they will introduce this trouble as seldom as possible. Tippett's numbers cannot be random in the sense that I should understand in my theory, because once assigned they are known and we have information about them that is relevant to future trials. They can only be random once. But there may be valid grounds for expecting that they are uncorrelated with anything that we may reasonably expect to want to test. Thus a test of their usefulness really requires consideration of the kind of additional functions that we may want to test in practice and of their correlation with the Tippett numbers. Now there are, I gather, about 40 000 of the numbers. If a selection of 100 in turn are made, and analysed for a linear variation, this will exceed the 1 per cent limit about 4 times. Bad luck if one picks one of these sets of 100, but it could be avoided if these ranges are specially indicated in the list. Again, it will happen fairly often that the means of two sets of 100 differ by more than the 5 per cent limit. In either case a real linear variation in the thing we are measuring will lead to a bias in the mean of the whole. So I think that there may be some point in testing what Kendall calls local randomness. But the danger is distinctly subtle. As far as the means are concerned, it requires both a correlation between the numbers and something genuine in the population sampled, and a non-orthogonality in the normal equations that will push the bias of selection into a further parameter. As to the standard

errors, the correlation may lead to the treatment of something genuine as error, and there will be a second-order bias in the error; but this could presumably be eliminated by estimating the genuine effect and allowing for it. My general feeling is that the things Kendall is looking for are likely to enter in practice in such an indirect way that departures from randomness in the numbers can be tolerated at much higher levels of significance than we should use in testing anything that interests us directly.

Are you saying anything in reply to 'Student's' last paper?[1] There was a point that had rather bothered me about Latin squares, which may be related to his. Suppose you take Cartesian coordinates about the centre of the square. Then as you eliminate rows and columns completely you are virtually supposing that the fertility can be expressed in the form

$$F = a_0 + a_1 x + \ldots + a_4 x^4 + b_1 y + \ldots + b_4 y^4$$

with no cross terms like xy, such a form being capable of being chosen in a 5×5 square to fit every row and column total exactly. Now just what are you asserting about the higher terms? A linear gradient is presumably a known danger. But if the x^2 and y^2 terms are genuine, then except for special orientations there will in general be a term in xy, which will be treated as error. In some conditions this might be serious. Thus it might turn out that xy and the treatment means together account for the whole variation in the square, and the error is 0; but if xy is left in the error it may be important and real differences in the treatments means may be hidden.

The question seems to me to be one of fact, and that rules cannot be laid down from theory. I should say that if the x^2 and y^2 terms are worth eliminating, then xy is worth eliminating and x^3 and y^3 possibly are, possibly not; if they are not worth it, then xy, x^3 and y^3 can all be treated as error. The question seems to be whether analysis of row and column totals does in fact often lead to significant coefficients for x^2 and y^2. If it doesn't, your method will sacrifice some information about the error, though I suppose that you have enough anyhow for this not to matter much.

I haven't the remotest idea how important these considerations are in practical cases, but I should be interested to know.

[1] 'Student' (1937). Comparison between balanced and random arrangements of field plots. *Biometrika* **29**, 363–79.

Fisher to H. Jeffreys: 30 *May* 1938

. . . It [randomization] is, as it seems to me, a tribute to our ignorance of the nature of the errors to which our results will be liable. Thus, if I want to test the capacity of the human race for telepathically perceiving a playing card, I might choose the Queen of Diamonds, and get thousands of radio listeners to send in guesses. I should then find that considerably more than one in 52 guessed the card right; also that of those guessing wrong more than half got the colour right, and probably a number of such favourable indications would be obtained. On the other hand, if I choose the 8 of Spades, I should expect to get just the opposite result.

Experimentally this sort of thing arises because we are in the habit of making tacit hypotheses, e.g. 'Good guesses are at random except for a possible telepathic influence'. But in reality it appears that Red cards are always guessed more frequently than Black.

For years agricultural experimenters made the similar unconscious hypothesis, 'Errors on different plots are distributed independently in the normal curve'. Actually the normal curve is good enough, but the errors are very far from independent; consequently any systematic arrangement *may* contain factors in common with the actual pattern of natural fertility. This difficulty is so fundamental that one has to consider the problem in a extreme form. Let the Devil choose the yields of the plots to his liking; his only restriction is that he may not change his mind after I have chosen where the different treatments are to fall. If, now, I assign treatments to plots on any system which allows any two plots which *may* be treated alike an equal chance of being treated differently, in the different ways in which this is possible, then it can be shown that both the experiment is unbiassed by the Devil's machinations, and that my test of significance is valid.

Things are not really so bad as in this game. We know nothing in detail about the errors, but experience does indicate certain components which are very often important, and such components one does not leave to chance but completely eliminates. Thus, if one had equal areas in two fields, one might legitimately assign pairs of plots, one from each field, and toss up between treatments A and B; then it would be annoying if chance (or the sequence of random numbers) put 7 A's running in the same field, which is the sort of thing liable to happen. If the same sequence of numbers were used for adjacent plots, its locally systematic character would not matter, a run of alternate treatments might be quite good. Though, again, if we suspected alternation in field fertility, such as is known to occur sometimes, or even a steady gradient, one might prefer to randomise entire sandwiches $A\,B\,B\,A$, or $B\,A\,A\,B$ by a single act of randomisation.

I think you hit the nail on the head in saying that a sequence can only be random once. Hence one must insist on a fresh randomisation each time when a set of identical experiments is laid down. For the same reason it is desirable that different centres should, as is already found convenient, each use their own tables for those purposes for which tables save time.

In principle I agree with you about the Latin square. Rows and Columns usually take out a useful lot of error, but I should not claim that the elimination of the same number of degrees of freedom, 12 in a 7×7 square, could not take out more. If, however, you impose the limitation that the components eliminated shall be orthogonal to those used for treatments, the combinatorial problem becomes greatly involved. I should, however, have no

objection in theory to the claim that better arrangements can be found. One fact which makes the Latin Square work well in practice is that on any given field agricultural operations, at least for centuries, have followed one of two directions, which are usually those of the rows and columns; consequently streaks of fertility, weed infestation, etc., do, in fact, occur predominantly in these two directions. Streaks in other directions introduce bias in systematic squares, but validly estimated errors in random squares.

H. Jeffreys to Fisher: 22 *September* 1938

. . . I like your note on 'Student'.[1] Curiously, when I rediscovered his rule in *Sci. Inf.*[2] he was only a name to me; and when you went for me in the *Proc.Roy.Soc.* [*CP* 102] I thought that you were attacking this rule! But there are a few points that I don't quite agree with. . . . If randomization was necessary I think you would consider assigning all 25 plots at random in a 5×5 square as the ideal procedure, getting two of a treatment in one row and none in another. 'Student' misses the point of the Latin square, I think, when he says that it is both balanced and random. I should say that it is balanced for ground effects known to be often large (possibly indeed more balanced than is really necessary) and random for others that might mount up if the design was repeated in different squares; the balance and the randomness refer to different features of the ground variation. The real question, it seems to me, is, what ground effects matter in practice? If there is no correlation at all between plots it doesn't matter what you do. If there are correlations and systematic variations it is legitimate to arrange the work so as to estimate them and allow for them, also allowing of course for the degrees of freedom lost on the way. If they are doubtful, deliberate randomization will provide an alternative treatment and easier arithmetic. I am not sure but I think I have seen discussions on the point that don't notice that if there is anything to be dealt with by a systematic design, its proper analysis will involve the separation of additional degrees of freedom, the variance associated with which is neither treatment nor error variance. The row and column variance in a Latin square is a particular case; so would be the fertility gradient along the row in a randomized block experiment, which 'Student' suggests would be worth taking into account. That is, I think that in the Latin square design you admit his main contention; but I have not gone into the matter enough to have any opinion about whether he has got the best way of analysing the results of other systematic designs. It is a matter of stating explicitly what terms in the fertility are to be considered, and working out the maximum likelihood way of eliminating them. Neither 'Student's' paper nor that of E.S. Pearson is sufficiently explicit on this point to convince me that they have got to the root of the matter. I should say that any design will give a valid estimate of error if the results are analysed correctly.

When I sent 'Student' a copy of my paper 'The relation between direct and inverse methods', I got a short note in reply, calling attention to the fact that in the title of his table the words 'a unique sample' occurred. I should like to acknowledge this, because it shows that the condition that I said he had assumed without mentioning it, that the sample constituted the only relevant information, was in fact in his mind. If there are several samples from the same population everything in the direct argument still holds but nobody would use such a set separately, so that the extra condition comes in in the transition from the argument to the use that is made of it. . . .

[1] *CP* 165. See Fisher's letter of 19 September 1938 (p.169).
[2] Jeffreys, H. (1931). *Scientific inference*. Cambridge University Press.

Fisher to H. Jeffreys: 26 *September* 1938

So far as I can judge, 'Student' and I would have differed quite inappreciably on randomisation if we had seen enough of each other to know exactly what the other meant, and if he had not felt in duty bound, not only to extol the merits, but also to deny the defects of Beaven's half drill strip system.

For example, randomisation was never intended from the first moment it was advocated to exclude the elimination from the error of components of error which could be completely eliminated, as in the case of differences between blocks in a randomised block system. It only requires that these components shall equally be eliminated from the estimation of error, as is now usual, though it was not appreciated at the time I first wrote on the subject. I often put this by saying that it is only the components which contribute to the actual error of the experiment which need to be randomised to provide an estimate of that error. In the Latin Square, for example, these are the components of variation that would remain after elimination of the best formal additive in rows and columns.

The second point on which there has been some misapprehension is that I should take for granted that the experimenter can choose anything he likes to be regarded as a single plot. For example, in sampling agricultural crops, a sampling unit, as it is there called, may consist of 50 roots scattered over the whole sampling area, with a rigid structural relationship *inter se*, so that the whole unit is determined by a single act of randomisation. For example, one might choose roots 7, 27, 47, etc. walking up and down a line, until the whole sampling area has been covered. The number 7 has been chosen at random from the numbers from 1 to 20, and to provide an estimate of error, a second sampling unit of the same kind is always taken from the same sampling area, e.g. the series based on number 12. Comparison of the results from numerous such pairs of sampling units from the different sampling areas will then provide a valid estimate of the sampling error of the samples used.

This is actually parallel to putting down a systematic comparison of two barley varieties at a number of centres and using the discrepancies among the comparisons made at different centres to estimate the precision of the

aggregate comparison. In doing this, of course, we should make no pretence of knowing how precisely the comparison has been made at each particular point, i.e. in relation to local circumstances, which might favour one variety rather than another. The assertions made are valid only for the aggregate of all such tests, and if the difference between the two varieties is different on different farms, we shall have no valid estimate of error on which to judge the significance of such differences. One of the practical points on which, so far as I can judge, I differed from 'Student', is that he was willing to ignore such local differences, possibly because, in any case, Guinness wanted Ireland to grow only one barley variety, whereas I felt that we owed it to the farmer to encourage him to grow whatever variety seems most profitable to *him*.

I fancy also that Gosset never realised that a fertility gradient when, as in my experience is not very frequent, it is important enough to bother about, can easily be eliminated from a randomised experiment. It is, I think, my fault that I have not made this clear earlier, but until the last two years I had really thought that 'Student' accepted all that I had put forward on behalf of randomisation. In the next edition of *Statistical Methods* I am exhibiting the procedure of eliminating a fertility gradient, as one lot of uniformity trial data used in that book happens to show a gradient which it is profitable to eliminate in this way.

Thanks for sending me your paper on sampling a mixed population. I am glad to see how close to my position you come in the parts referred to. I suppose you will send it to Cambridge Phil. Soc., or somewhere of the kind.

Fisher to D.A. Kislovsky: 16 *February* 1929

I was much interested in your reprint from the *Journal of Heredity*[1] and especially in your conclusions respecting the effects of mass selection on extensive livestock. The method of Heincke appears to be sounder that that of Pearson in that the coefficient of racial likeness depends essentially on the sizes of the samples examined, and is therefore a crude test of significance of racial difference, rather than a measure of it.

Both methods I believe ignore the correlations between characters, and this becomes especially important where they are numerous, partly because it is reasonable to attach the greatest weight to distinctions in characters which are within each group least variable, partly because if the number of measurements is greatly increased the prediction of any one from the others becomes so accurate that they differ only by errors of measurement, and then a number of the apparent variates will represent not real biometrical variates, but errors of measurement only.

Logically, though this might be tedious, I can see no more proper way than to analyse the sum of squares of each variate, and the series of products of each pair of variates, into the two portions, *within races, between races*, taking

exact account of the degrees of freedom. The examination of the relationships between the two quadratic forms so obtained would be quite new work, but I think a straightforward process could be evolved, and if you would care to send over your data, either after or before the preliminary analysis, I should be pleased to try to see what could be done with it.

[1] Kislovsky, D. (1927). Types in animal breeding and their analytical study. *J. Hered.* **18**, 447–55.

Fisher to P.A. MacMahon: 3 *July* 1924

Many thanks for your letter. I have sent to the R.A.S. library for your book, but have not yet seen it. It is possible that your method of solution will solve for me at least the first of the outstanding questions.

(i) What experimental technique of filling up the square will give each solution an equal chance of appearing?

One might fill the first two at random in $n!$ ways, and then fill the first column at random in $(n–1)!$ ways. Then things become more difficult. Possibly the following would suffice. Choose one of the remaining rows at random, and proceed to fill up the spaces in order rejecting any entry which conflicts with the first row. For example, if there are five spaces, one might shuffle 100 numbered cards thoroughly and take a card numbered $5m + p$ ($p = 0, 1, \ldots 4$) to indicate the particular letter assigned to p; the first card dealt will then specify which letter is to occupy the first vacant space, unless this conflicts with the first row, in which case the first card is rejected, and the second card examined. In this way each space may be filled in succession. Questions which arise are:

(a) is it necessary to choose at random the row or column to be filled in next, or can they be filled in in order as they stand?

(b) is it necessary to choose at random the particular space to be filled in next, or can the spaces in a given row or column be filled in in order as they stand?

(c) on rejection of a card should the pack be reshuffled?

(ii) Is such a technique necessary or sufficient for the statistical validity of an agricultural experiment? This question involves points which it is not easy to reduce to a purely mathematical form.

An example will show the kind of lines along which I have been working.

The following are the actual yields of 25 areas in a uniformity trial (Mercer and Hall, 1910)

2713	2738	2673	2698	2822
2734	2657	2614	2559	2651
2753	2574	2623	2762	2765
2656	2562	2490	2605	2504
2586	2532	2489	2496	2459

The deviations of these from uniformity may be taken as giving the errors to which an experiment of this sort is subject. The sum of the squares of the 25 deviations from the mean is 253 106; dividing this by 24 we have a variance of 10 546. As an estimate of the standard error of the mean of 5 such plots taken at random, we divide this by 5 giving 2109.2 and take out the square root, 45.92.

If we had used five different treatments or varieties, each replicated five times, we should not actually reproduce this estimate; for example, assign plots to treatments wholly at random, as follows:

		Total deviation	Mean deviation
ACCED	*A*	− 68	−13.6
CECAA	*B*	−166	−33.2
BBEDE	*C*	+105	+21.0
ADDBB	*D*	− 3	− 0.6
BECAD	*E*	+132	+26.4

The arrangement has evidently happened to make the errors less than usual; in consequence of this the estimate of error found by comparing plots treated alike will be more than usual. The relation between the two things is best shown by what I call an analysis of variance:

Variance between	Degrees of freedom	Sum of Squares	Mean Square
Different treatments	4	12 127.6	3 031.90
Plots treated alike	20	240 978.4	12 048.92
Total	24	253 106	10 546

The 24 degrees of freedom between the 25 plots may be divided into (i) 4 degrees of freedom between the five different treatments, and (ii) 20 degrees of freedom representing 4 comparisons between the 5 parallels for each of the 5 treatments. Corresponding to this division the sum of the squares of the deviations, 253 106, may also be divided into two parts, the first of which may be found by squaring the total deviations for $A,B,C,D,E,$ summing and dividing by 5, and the second (in practice found by subtraction) is the sum of the 25 squares of the deviations of the individual values from the mean values for the corresponding treatments. The 'Mean Square' in the last column is found by dividing the Sum of Squares by the degrees of freedom. On the average of a large number of trials it is obvious that the mean square will take the same value in each line, apart from any real difference between the treatments. In fact the significance of any apparent difference between the treatments must be judged by the frequency with which a ratio between the 'Mean Square' values, as high as or higher than that observed, will occur under uniform treatment for given numbers (4,20) of degrees of freedom.

This distribution I have been able to obtain and sufficiently to tabulate assuming the original deviations to have been a sample from a normal distribution; this assumption is not, I think, a cause of trouble, since I find that moderate deviations from normality make wonderfully little difference.

The Latin square comes in at the next stage when we want to improve the experiment by eliminating part of the variability due to heterogeneity of different parts of the field. Treating the rows and columns as I have treated the 'treatments' above, we have an analysis of variance as follows:

Variance between	Degrees of freedom	Sum of Squares	Mean Square	S.D. of mean of 5
Rows	4	162 315.6		
Columns	4	32 842		
Remainder	16	57 948.4	3 621.775	26.92
Total	24	253 106	10 546	45.92

The variance is cut down to nearly a third of its previous value, if we can eliminate from the errors of the comparisons between treatments, differences due to the different fertility of the rows and columns. That is what a solution of the Latin Square does; if we take an arrangement such as

		Total deviation	Mean deviation
DECBA	*A*	−112	−22.4
BDEAC	*B*	− 67	−13.4
CABDE	*C*	+ 71	+14.2
EBACD	*D*	− 18	− 3.6
ACDEB	*E*	+126	+25.2

the deviations are only a little smaller than before, though again the arrangement is rather a fortunate one, none of the deviations exceeding the S.D. The full analysis of the experiment is now:

Variance between	Degrees of freedom	Sum of Squares	Mean Square
Rows	4	162 315.6	
Columns	4	32 842	
Treatments	4	7 654.8	1913.7
Remainder	12	50 293.6	4191.1
Total	24	253 106	

The advantage of the arrangement lies not only in the fact that the errors will generally be somewhat smaller, but in the fact that the estimate of error, obtained from the 'remainder' will generally be smaller. The advantage is somewhat counter-balanced by the fact that the estimate is based on fewer degrees of freedom, but from all that I have seen from uniformity trials, it is very well worth while. Everything depends on the validity of comparison of the mean square for the 'treatments' and the 'remainder', and I am fairly sure that it is valid provided every solution of the Latin square has an equal chance of occurrence.

P.A. MacMahon to Fisher: 30 *July* 1924

The question of the number of arrangements of letters of n different kinds in a square $n \times n$ so that each of the letters shall appear precisely once in each row

and in each column is known as the problem of the Latin Square. I have given the mathematical solution and you will find it in my *Combinatory Analysis*, Vol. 1, p.250.

For $n = 2$, no. of arrangements is	2	1^1
3 " " "	12	1^1
4 " " "	576	4^1
5 " " "	149 760	52^1

and I have not calculated the numbers any further.

I have had some difficulty in reading your handwriting and also in understanding what your further point is to which you allude in the last paragraph of your letter.[2]

If you will put a definite question to me in connection with it I will do my best.

[1] The corresponding numbers of reduced squares, as given in MacMahon's book, were written in by Fisher.

[2] Fisher's letters at this time were usually handwritten but the earliest of his letters to MacMahon of which we have a copy — bearing the date 3 July 1924 and reproduced above — was typewritten. MacMahon had evidently received an earlier handwritten letter from Fisher. To judge from the content of the Fisher–MacMahon correspondence as shown here from July 1924, it seems likely that the date given on one of the letters was incorrect, with Fisher's typed letter dated 3 July 1924 being his *reply* to MacMahon's letter dated 30 July 1924 and possibly incorporating material from the earlier handwritten letter that MacMahon had difficulty in reading.

P.A. MacMahon to Fisher: 19 *September* 1924

I have lost my Toronto note. Is the correct number of Latin Squares (reduced) now 56 instead of my number 52?[1]

[1] See *CP* 110, p.493.

Fisher to P.A. MacMahon: 21 *January* 1926

I think I have enumerated the 6 × 6 Latin Squares, making a total of 6008, of which 456 are symmetrical.[1]

The method of enumeration is to write down all possible diagonals, (types of diagonal), as on the enclosed sheet. The 10 arrangements in the right hand column give, on trial, no solution, the remaining 24 all give solutions, usually only a few each. For example, the diagonal *ACDBAB* gives two conjugate pairs of solutions. In this case all the letters are distinctly characterised,

B is the letter that occurs twice,
C " " " " takes *B*'s place,
D " " " " " *C*'s",
E " " " whose place is taken by *A*,

consequently by permuting *BCDEF* in any order and arranging rows and columns to make the first row and column come in the standard order, each solution gives 120 solutions or 480 in all.

With the more degenerate groups I have no resources but to make a number of trial permutations, until I am more or less satisfied as to the groups to which they belong, but this should not, I fancy, affect the number of solutions in the total. For example there are 60 different diagonals of the type *AAABBC*, and one of them on trial is found to give one pair of conjugate solutions; the total is therefore 120, irrespective of the fact that the substitution $D - E$ changes one solution into the other.

I hope you will think the method a valid one, although it is scarcely applicable to 7 × 7 as there must be about 10 000 groups in that case.

[P.S.] You will be interested to know that the Latin square has been a great success agriculturally.

[1] See Fisher, R.A. and Yates, F. (1934). The 6 × 6 Latin Squares. *Proc. Camb. Phil. Soc.* **30**, 492–507, (*CP* 110), where the correct enumeration was finally obtained.

P.A. MacMahon to Fisher: 24 *January* 1926

I congratulate [you] on a 'tour de force' in enumerating (actually) the Latin Squares of order 6. You obtain 5552 unsymmetrical and 456 symmetrical, total 6008. If you halve the number 5552 you get 2776 squares to be read both by rows and by columns. Thus $2776 + 456 = 3232 = 2^5 \cdot 101$ are necessary to exhibit the whole of the 6008.

In some recent work I have had to examine certain square formations connected with determinant theory and found that the grouping proceeded by diagonals — so that I am quite prepared to find that the diagonal method suits your enumeration best. I have no reason to doubt the validity of your method. As you say, the number increases with the order with great rapidity.

Fisher to K. Mather: 8 *May* 1950

After looking at some recent American books on statistics I have felt constrained to draft the enclosed as an addition to Section 65 of *The Design of Experiments*. I do not know whether you have been troubled by finding otherwise suitable textbooks showing the kind of misapprehension at which I aim. If not, of course, I don't see why you should be, but you may already have the question on your mind.

K. Mather to Fisher: 18 *May* 1950

Thank you for letting me see your draft for inclusion in *The Design of Experiments*. The distinction between definite and indefinite interactions is

most illuminating and I am sure that it will be most valuable in clarifying what to most people is an obscure and difficult subject.

I suppose that most biological interaction will be indefinite; but the plant physiologist, for example, might well run up against the sort of situation you illustrate from industry. . . .

Fisher to K. Mather: 20 *May* 1950

I think categories such as established varieties and inbred lines will be definite enough, because reproducible at will in future work. On the other hand, genotypic differences among F_2 plants as shown by F_3 progenies would be fairly typical indefinite categories.

What had shocked me, however, was a number of recent American books with practical aims specialising in industrial experimentation getting themselves into great difficulties through assuming that main effects were always to be compared with interactions with whatever factors were included in the experiment. Logically this makes them think that an experiment without e.g. variation of temperature gave definite information of a kind not available if the whole experiment were repeated at five different temperatures. This is manifest nonsense, and my new paragraphs aim at clarifying the confusion of such paradoxes. . . .

D. Michie to Fisher: 30 *October* 1956

. . . I would also be glad of your opinion on a statistical problem which my friend Dr. Murdoch Mitchison has brought to me.

He has been measuring the dry mass of individual living yeast cells by an ingenious optical method which allows successive observations to be made on the same cell without disturbing it. He is interested in the form which the increase of dry mass takes during the period between fissions. Is the increase linear with time, or is there some degree of curvature?

He has a number of independent series of measurements, each series on a separate cell (perhaps 20 series in all). The various series differ among themselves in the number of measurements and the time intervals between them. He is not able to assume that the different series agree with each other either in the magnitude of observer error or in the culture conditions — either their initial state or their rate of subsequent deterioration. . . .

Neither the slope nor the degree of curvature can be assumed the same from one series to the next. Dr. Mitchison is quite prepared to envisage cases where an intrinsic tendency of the cell to give a concave-upwards curve is actually reversed by rapid deterioration of the conditions in a particular culture so that curvature of opposite sign results.

He asks the question 'Is there a significant overall tendency towards curvilinearity as such, without regard either to the degree or to the sign of curvature from one series to the next?' I rephrase his question as follows: 'After fitting a linear regression line to each series, is a significantly large fraction of the residual variation between observations within series attributable to some degree of quadratic curvature in some of the series?'

The recipe which I have given him is to set out an analysis of variance for each series independently, after fitting to each a regression equation of the form $Y = a+bx+cx^2$, as follows:

Source of variation	*S.S*	*D.F*	*M.S.*
Linear term of regression equation		1	*A*
Quadratic term		1	*B*
Residual		$n - 3$	*C*
Total		$n - 1$	

n = no. of observations in the given series.

Tabulate the mean square ratio *B*/*C* for each series, together with the corresponding value of *P* derived by interpolation in a table of the Variance Ratio. On the set of *P* values obtained in this way, perform a combined test of significance according to the method described in section 21.1 of *Statistical Methods*.

Of course, if the combined test gave a significant answer he would still be faced with a problem of interpretation. But if it did not he would be able to say that his data did not contradict the idea that the dry mass of a yeast cell increases linearly with time.

I would very much value your comments on the approach which I have suggested to him. He had already calculated the required analysis of variance before consulting me, but since on crude inspection they showed great heterogeneity between series in almost every respect he was uncertain as to how to bring the relevant information from all series to bear upon the specific hypothesis of linearity.

Fisher to D. Michie: 31 *October* 1956

. . . On Mitchison's problem your solution is highly ingenious, but if I understand it correctly it does not take account of the *sign* of each quadratic term, whereas probably what Mitchison is interested in is just whether they are largely of the same sign representing curvature upward or downward.

You could take the square root of *B* and give it a weight inversely to its variance. This would be the classical method of combination, which is mildly defective in not taking account of the precision of these weights.

Simplicissime, and as an additional check, the regression line gives an expected midway value, and the observed points if odd in number give one realized, and if even in number I should not hesitate to take the mean of the

middle two; so you have a number of deviations from expectation, some positive and some negative; and ignoring their variable precision, on which the data tell one rather little, you could give them a simple t test. If the two methods do not agree substantially, do not trust either of them.

Fisher to J.A Nelder: 6 *November* 1956

I have just seen your review (*Journal of the Royal Statistical Society*, volume 119, page 340) of Federer's book on *Experimental Design, Theory and Application*. I notice that you reprove the author for quoting without comment my expression for the loss of information in experimental work due to the use of estimated, in place of true, variances. Your suggestion, if I read it aright, is that Federer ought to have used some 'other intuitively reasonable' measure of this effect.

I am, of course, aware that a number of people have found difficulty in understanding the cogency of the fiducial argument, though a good deal of progress seems to have been made already in this respect. I cannot expect it to be obvious, therefore, that the amount of information supplied by a test with an empirically estimated variance actually is the amount which I have calculated. I should, however, be glad to know what method is in your mind as an alternative, remembering that the planner of an experiment does not know in advance what value of t may emerge for any particular comparison, and would be foolish to suppose, as has sometimes been urged, that the purposes of his experiment are fulfilled only by deciding whether t does, or does not, exceed some value chosen in advance. Generally, a number of different comparisons are tested simultaneously in the same design.

Of course I have no objection to others discussing problems different from those which I have at some time set myself, but you will perhaps see reason in an objection to their solutions being claimed as solutions of the same problem as that which I have treated.

J.A. Nelder to Fisher: 8 *November* 1956

In reply to your letter of 6 Nov. about a point I raised in a review of Federer's book, I think I can make clear my attitude fairly briefly. It seems to me that there are two ways in which a text book can be written. It can be written from a strongly 'personal' view-point, in which the ideas of the author form a unifying background to the discussion. I would regard your books as coming into this category; I personally prefer books of this kind, though I may not always understand or agree with all that is in them. Alternatively an author can attempt to construct an encyclopaedia of his subject in which all the diverse work on related problems is set out, but in which no strongly expressed opinions of the author's own are expressed. My remark in the

review was prompted by the remembrance of Section 2.3 of Cochran and Cox's *Experimental Designs*. I imagine that you would argue that the alternative criteria considered there were all irrelevant to the problem. Personally I could not accept that, and my quarrel with the author was that since he does not appear to adopt the fiducial standpoint in his book, and appears to be attempting the encyclopaedic type of work, he should consider other approaches. It may be that these other approaches are faulty, the future alone will show that, but in my opinion if the ideas have been held by reputable workers then they should be discussed, particularly in a book containing such an immense amount of detail in other directions.

I hope these remarks of mine will make it clear that it was not my intention to suggest that your measure of relative accuracy was faulty, but merely to maintain that an encyclopaedia should be an encyclopaedia.

Fisher to J.A. Nelder: 10 *November* 1956

Thank you for your courteous letter. Looking at Cochran and Cox I suppose Table 2.3 is the relevant point. What I meant by other authors having tried to solve a different problem, namely one in which the level of significance, or the t value, is known in advance, is illustrated by Cochran and Cox's Table 2.3. What I wanted you to see was that this is a different problem, which indeed had already been solved by 'Student', from the one with which I was concerned, namely that to a man planning an experiment, who has a choice of the numbers of degrees of freedom which will be available but no knowledge of how large the several deviations may be, it is useful to know how much information is lost by the circumstance that the number of degrees of freedom will be finite, and how much more is lost when it is small than when it is large. Of course there are plenty of people to act as gramophone records of Neyman's views. I suppose if Federer had been confronted in the literature with alternative solutions of the same problem, it would be proper to expect him to mention both and perhaps to adjudicate between them. What perhaps you have overlooked is that the problem, when the experimenter has no knowledge of what levels of significance he may expect to find, and of course many different levels may appear for the same estimate of error even in the same experiment, and especially if he does not (and this is still the view of the majority of experimenters) think that his experiment exists solely to make an acceptance decision, is a different problem from any that Neyman and his followers have discussed.

On the following pages indeed, 28 and 29, the authors advise, 'For general purposes it is suggested that this table' (namely that of Fisher's correction) 'be used to take account of the differences in degrees of freedom for error in two designs that are being compared'. It would seem, therefore, that after considering Neyman's contribution these authors do not suggest that it supplies an alternative method.

J.A. Nelder to Fisher: 28 *November* 1956

Thank you for your letter of 10 November. I agree with you that Cochran and Cox suggest the use of the quantity $(n+1)/(n+3)s^2$ for the comparison of designs with different nos. of degrees of freedom. I take this to mean that they believe that if the experimenter is offered a choice of two designs, one having error variance 100 and a large no. of d.f. and the other 50 and 1 d.f. he is to regard them as equivalent. There are two difficulties here for me; first, as an experimenter I would intuitively reject the equivalence, and secondly I am not sure that it is a correct interpretation of your original derivation in Sec. 74 of *The Design of Experiments*. My own belief in the matter is that this is a problem with no unique solution, so that the discrepancies in Cochran and Cox's Table 2.3 are to be expected and are not resolvable without further detail in the statement of the problem. Thus in the kind of work we do here, where we have a general long-term aim to improve the practice of vegetable growing, an experimenter will always be finding in the course of his work certain indications, as a side-product of his experiments, that might be followed up. He will often do a rather small experiment to see whether such a possible line of work would be worth following up. His decision on the results of such an experiment will approximate rather closely to an acceptance procedure. In deciding, he will have to assess the sort of treatment difference that would be practically interesting and to assess a reasonable significance level. In practice, I admit, this will be difficult to do rationally, but nevertheless different assessments would lead to different assessments of relative accuracy after the manner of Table 2.3. There is one point concerning Sec. 74 of *The Design of Experiments* which I should be glad if you would clarify for me. In the other examples of the chapter the information is derived from the likelihood function and expressed in terms of the parameters of the population, but in section 74, it is derived not from the likelihood function based on $(x-\mu)/\sigma$ but from the t-distribution $(x-\mu)/s$. This latter may be regarded as derived from the likelihood function by integrating out σ according to its fiducial distribution. There are thus two processes, the elimination of the unwanted parameter σ and the formation of the information function. If, however, we carry out these processes in the reverse order, the information function gives us $1/\sigma^2$, and its integration over the fiducial distribution of σ gives $1/s^2$. The point on which I am not clear is the logical basis for preferring one order of operations to the other. I should be most grateful if you could elucidate this point for me.

Fisher to J.A. Nelder: 3 *December* 1956

. . . I suppose your misapprehension as to what Cochran and Cox said about my formula $(n+1)/(n+3)[s^2]$ was due either to your misunderstanding one of the teachers here,[1] or to his own misapprehension on this point. . . .

In the superfluous little table provided by Cochran and Cox on page 28,

they say 'equal amounts of information', and perhaps you have read into this the more elaborate concept of equivalence. In the table on the previous page, the author, perhaps Welch, has fallen between two stools by calculating the *average* limits of error at various levels of significance, whereas 'Student' had given the actual limits at each level, and this is what the experimenter needs.

If I understand your letter aright, Federer is to be scolded for quoting my formula not because the latter had been shown to be inexact, and not because any effective alternative has been offered, but because you hope at some future time to offer such an alternative, not being yourself satisfied with the matter as I have left it.

With respect to the calculation of the amount of information, this has been the same and examples of it have been available every year, at least of the last 35; and again it must be the peculiarities of your teaching at Cambridge which led you to think that some other method is more authentic. Common forms are

$$\sum \frac{1}{m}\left(\frac{\partial m}{\partial \theta}\right)^2, \quad \int \frac{1}{y}\left(\frac{\partial y}{\partial \theta}\right)^2 dx.$$

Probably, however, you were not taught to regard the fiducial distribution of μ as a frequency distribution at all.

[1] University of Cambridge.

J.A. Prescott to Fisher: 4 *October* 1930

For some time I have been intending to write to you about one or two problems of sampling and experimental method. The first deals with the matter of systematic sampling over a large area of, say, several thousand, or even million, acres. Two come to my mind particularly, one the selection of a number of sites in Egypt for the daily counting of the flowers of cotton plants, and the second relating to the determination of the probable distribution of salt in a large suspected area where it would be almost hopeless to map accurately the actual distribution within a reasonable time. My recollection of the former case is that a grid was recommended to be laid down over a map of the Nile delta and that localities for observation were to be selected say at points of intersections. I am no longer particularly interested in this matter, but the second one is very vital at the moment in Western Australia. . . . In this case an examination of soil samples from 650 sites on the so-called forest soils in an area of about 2000 square miles indicated that about one-half of the samples contained dangerous quantities of salt.

A similar instance is the systematic examination of soil profiles from an area of about 500 000 acres in Victoria commanded by the waters of the Murray impounded by the Hume dam. In this latter case the soils have been sampled on a grid basis, while in the former the sampling was at regular

intervals along and on each side of prospectors' tracks through the virgin 'forest'.

The problem is in effect what system of systematic or random sampling should be adopted in order to obtain a reasonably accurate estimate of the probable distribution of certain soil types or soil conditions without actually mapping the area.

. . . It is interesting to note that, having selected an area as suitable for survey, we frequently meet all the major soil types in the course of a few days' work and that thereafter for several weeks no new types are met.

Another problem of systematic as against random sampling has been referred to me. . . . In this case the examination of records from areas showing definite yield gradient across the field has given . . . some difficulty. This appears to apply particularly to pasture work. . . .

Am I not right in assuming that your system of randomized blocks more or less takes care of the above problem?

There is one final point which has appealed to me personally and which may be already covered by your general theory of sampling and that is that sometimes one gets an impression that the older statistical method underestimates the reliability of the final result. . . . In a latin square of four treatments, there is actually one-fourth of the area under any given treatment, and with the four plots distributed at random the estimate of the yield from the whole area should be very near the truth. I can quite see that it is possible to treat this problem rather differently from the older method of taking say 1000 small plots and grouping them into different combinations until the reliability is sufficiently increased. Would it not be better to group these results into as many possible areas of say one-twentieth of an acre, scattered over the whole area? . . . The size of plot is not only a matter for determination by statistical analysis, but there is also the important question of field technique, the implements and labour available, weighing appliances, and so on, which all have a decided bearing on the final size of the plot. In Egypt we found this to be 1/40th acre for wheat, 1/20th acre for maize, and 1/10th acre for cotton.

I should very much appreciate any remarks you may have to make on the above-mentioned problems.

Fisher to J.A. Prescott: 10 *November* 1930

The problem you mention of the large scale survey is one of those which are dealt with very satisfactorily by multiple regression. The standard example of this treatment of an analogous problem is in Section 29 of my book *Statistical Methods for Research Workers* where I treat of the relationship of rainfall to Latitude, Longitude and Altitude, from the point of view of one wishing to estimate the probable rainfall at a given point, or within a given area, in view of the values observed at a number of stations.

When there is an appreciable gradient in any quantity (say) from North to South, one must distinguish between the probable error of an estimate at a known Latitude from that at a point of undetermined Latitude. The first will evidently be the smaller. In field experiments, as you say, the method of randomised blocks eliminates the greater part of any such gradient effect; it should be noticed that it also eliminates much of the heterogeneity which is not simple enough to be represented by a gradient, and what is most important, it provides a valid estimate of the errors not eliminated. In this it differs from all systematic arrangements.

I quite agree with you that higher accuracy can always be attained by increasing an experiment from one to several Latin squares; usually it can also be done without increasing the area by sub-dividing the experimental area available. However, parallel Latin squares at different stations have the great advantage of broadening the inductive basis of the result.

Personally, I always advise that size and shape of the plot should be determined by practical agricultural convenience, while the structure and extent of the experiment must be governed by the accuracy aimed at. The advantage of using the easiest and most expeditious field methods is thus applied to make further replication possible.

C.R. Rao to Fisher: 18 *September* 1952

I am writing this to refer to you some difficulty which I had in reading Fairfield Smith's paper on discriminant function for plant selection.[1] In the problem of predicting the genetic component $\psi = a_1\psi_1 + a_2\psi_2 + \ldots$ by a linear combination of the observations $l_1\bar{x}_1 + l_2\bar{x}_2 + \ldots$ I find the constants $l_1, l_2, \ldots$ depend on the sample size on which $\bar{x}_1, \bar{x}_2, \ldots$ are based so that no single set is ideally suited to predict genotypic values of various lines providing different numbers of observations. So the least square estimates weighted by n_r

$$\sum_r (\psi - l_1\bar{x}_1 + \ldots)^2 \, n_r$$

do not provide the best functions for individual cases except when $n_r = n$ for all r as in Fairfield Smith's problem. But in Panse's problem of poultry n_r are all different. But of course one can ask for a common linear function with which the genotypic values can be ordered. In this case the least square solution weighted by n_r is a good one. Having estimated the covariance matrix of the genotypic values $\psi_1, \psi_2, \ldots$ and also of within lines from covariance analysis it is possible to calculate the appropriate constants for each sample size within the lines and obtain the genotypic value for the ith line as

$$l_{1i}\bar{x}_1 + \ldots + l_{pi}\bar{x}_p$$

and then arrange the lines in decreasing order for selection. Am I correct? I

am also developing tests for the adequacy of an assigned discriminant function. If you think I am on right lines I propose to send the manuscript for publication in *Annals of Eugenics* or *Heredity*. I will send you a copy in advance.

[1] Smith H.F. (1936). A discriminant function for plant selection. *Ann. Eugen.* 7, 240–50.

Fisher to C.R. Rao: 23 *September* 1952

Thanks for your letter letting me know that you propose to send a manuscript for publication in the *Annals of Eugenics* or *Heredity*. It is only of the latter that I am now editor. I think it was intrinsic of Fairfield Smith's approach that he had a single multivariate sample to work on so that the number of cases was the same for all variates. I do not think I have heard of Panse's problem with poultry. I only know of his work with cotton. Another thing to remember about Fairfield Smith's paper, which perhaps does not concern you, is that he was developing a method supposedly based on the covariation of yield factors in one year with measurements available for selection in the previous year. Unfortunately the only material he had at hand to illustrate this method consisted in wheat data for a single year and this has led Mather and others somewhat to misunderstand the nature and purpose of the discriminant function.

Fisher to P.R. Rider: 19 *January* 1932

I was much interested in your long letter on B. Peirce's method for the rejection of observations;[1] the final step is certainly Maximum Likelihood in which Peirce was presumably following Gauss. I am not clear, however, whether or not Peirce is trying to do more than is possible.

Any set of residuals from a mean (and the same applies to residuals from more complicated regressions) may be represented by a point on a many dimensional sphere, and any criterion for rejection must consist in choosing certain portions of this sphere which shall be deemed inadmissible. If the criterion also determines how many observations are to be rejected, it must divide the sphere into portions

(i) in which no observation is rejected,

(ii) in which one is rejected,

(iii) in which two are rejected and so on.

Necessarily, the volumes of these portions must be fixed *a priori*; i.e. the probability of rejecting 1, 2, 3, . . . observations out of n, when no gross errors have really occurred. We might for example decide to reject one or more observations in one case out of 20, two or more in one case out of $(20)^2$ and so on; and we shall obtain such a series if we decide to reject observations

one at a time by the repeated application of the same criterion. The question is what particular values does Peirce reject, and how do the frequencies run?

This puzzles me a good deal, as it is only when the rejected area can be mapped out that one can judge of the plausibility of the method.

[1] Rider had asked if the final step in the paper by B. Peirce (1852), Criterion for the rejection of doubtful observations, *Astron J.* 2, 161–3, was not 'just the principle of maximum likelihood'.

Fisher to H.L. Rietz: 25 *February* 1930[1]

The approach to sampling problems by way of the geometry of generalised Euclidean space is an exceedingly fruitful one, but one which at present has to rely to some extent on the mathematical intuition of the reader. At least I know of no source to which one can be referred for the elementary general properties of such space; I was looking only recently at Baker's big work to see if he treated of the matter, and though there is much about cubic and biquadratic loci, the simple properties of the generalised Euclidean space are not explicitly developed.

The solution of the 1915 paper [*CP* 4], while obtained geometrically, may be easily verified analytically as in 'Sur la solution de l'équation intégrale de M.V. Romanovsky' (1925), (*Comptes Rendus de l'Académie des Sciences*, vol. 181, pp. 88–89) [*CP* 46]. I have also given in 'Applications of 'Student's' distribution', (*Metron*, vol. V, pp. 90–104, 1925) [*CP* 43], a general analytic argument equivalent to the very widespread use of hypergeometry in considering degrees of freedom. I fancy the only result of mine which now rests wholly upon the geometrical argument is that for the 'Distribution of the partial correlation coefficient' (*Metron*, vol. III, pp. 329–332), (1924) [*CP* 35], where it is beautifully direct; however, in this case also, the thing can be reduced to the analytic dissection of a sum of squares.

[1] Rietz had written seeking references to the properties of n-dimensional space used by Fisher in studying the distribution of correlation coefficients.

Fisher to H.L. Rietz: 4 *March* 1937

My attention has been called to a statement ascribed to you by one, S. Kullback,[1] which is either misleading, or, if not so in its original connection, appears to have misled Kullback.

This is to the effect that the analysis of variance is obtained largely by inferences based on the number of degrees of freedom of the variates, rather than upon formal mathematical proofs. Mr. Kullback seems to take this to mean, *inter alia*, that the distribution of the multiple correlation coefficient has not, in the course of my work, ever been demonstrated.

I need scarcely remind you that this and a great many other similar

problems were solved independently prior to 1925, before the particular arithmetical arrangement known as 'the analysis of variance' had been developed. A reference for the correlation ratio and the multiple correlation in the null case is *J.R.S.S. 85*: 595–612 [*CP* 20]. It was about 1923 that I recognised the unity of these different solutions, giving the general form of the z distribution at the Toronto Conference in 1924 [*CP* 36], and an analytical demonstration, of which all analyses of variance are more or less direct applications, in *Metron 5*, part 3, pp. 90–104 [*CP* 43].

The analysis of variance merely presents the properties of the quadratic forms involved in, I think, a very simple manner, and it is too much like deliberately ignoring the body of work on which it was based to suggest to students that no adequate proofs have been given. Whether you have overlooked this previous work, which was, of course, deliberately ignored by my distinguished predecessor, Professor Karl Pearson, or whether you think that the proofs therein given are not complete, I do not, of course, know. I have, however, never been inaccessible to correspondents who wish for a more explicit statement of the steps of my proof.

[1] Kullback, S. (1935). A note on the analysis of variance. *Ann. Math. Stat.* **6**, 76–7.

H.L. Rietz to Fisher: 22 *March* 1937

In reply to your letter of March fourth, I regret the matter very much if I have misled Kullback or any one else by my remark about arguments based on degrees of freedom without formal proof. Those arguments have been difficult for me to follow and I feel sure the same difficulty has been experienced by many of your interested readers.

Perhaps I should say first that Kullback has never been a student of mine. I met him in Washington on the occasion of his examination for the doctorate. That is the only personal contact I have had with him.

The introduction to Kullback's note is unfortunate. So far as I see, he was merely attempting to present, by the use of characteristic functions, an alternate demonstration of the distribution of the multiple correlation coefficient from uncorrelated normally distributed material.

With respect to the reader's difficulties in following arguments based on degrees of freedom, I venture to say it would be very useful to your interested readers to be shown in considerable detail the meaning of degrees of freedom in relation to properties of the quadratic forms under consideration.

I shall be glad to avail myself of the kind suggestion in the last sentence of your letter.

It will be much appreciated if you will include my name on the list to whom you send your current reprints.

Fisher to H.L. Rietz: 1 *April* 1937

Thanks for you courteous letter. I think the most general approach to the analysis of quadratic forms is that given, in 1925 or 6, in Applications of 'Student's' distribution, *Metron* V: 90–104 [*CP* 43]. In Section 6 on the significance of regression coefficients, I see that the printer has put ζ for ξ, but I think it is sufficiently correct in other ways, and you will readily recognise the generality of the demonstration. I am afraid I have now no spare copies of this paper. I have added your name to my mailing list.

Fisher to V. Satakopan: 30 *January* 1939

Thanks for your letter and kind congratulations.

For my own part I should never have pitted orthogonal polynomials and harmonic components against each other as rival methods,[1] since the cases to which they are appropriate seem to me quite distinct. However, Whipple has evidently for years harboured a feeling that polynomials have no place in meteorology, and in this I think he is wrong, as soon as meteorologists interest themselves, not merely in the average of a number of years, but in the differences between these years as individuals.

The difference arises in the fact that, whereas the average of any element at different times of the year through a number of years is necessarily a periodic function, yet the same element for any one individual year may have quite a different value at the end from what it had at the beginning. This is true of ordinary meteorological elements, though it was brought to my mind most forcibly when the element I was studying was the average effect of a given amount of rainfall on a crop harvested at the end of August, when it was obviously more than ridiculous to suppose that the average effect on August 31st in the year in which the crop was harvested approximated to the average effect on September 1st, 6 weeks before the crop was sown.

Otherwise the parallelism is close, apart from the fact that users of harmonic analysis have not always confined themselves to the truly orthogonal components, namely, those with period n/s where n is the number of observations in the record and s is a whole number from 1 to $(n-1)/2$ in case n is odd, and $(n-2)/2$ when n is even. Periodogram values for periods other than this series are complicated compounds, involving the rejection of varying amounts of the data, and I do not think anything useful can be done with them, but the series of coefficients of sines and cosines having these periods are strictly comparable to the coefficients of the orthogonal polynomials, save that they come in pairs instead of in a single series. . . .

[1] Satakopan had said in his letter that he had been asked to initiate a discussion on 'harmonic functions vs. orthogonal polynomials in meteorology and geophysics'.

Fisher to W.H. Sayers: 20 *August* 1941

Thanks for your long letter, which, if I have followed it correctly, turns

largely on the meaning of the word *experiment*, and perhaps partly on that of the word *design*.[1] Perhaps I can make my meaning clear by a discussion of this point.

The experiments with which I was concerned are observations, or sets of observations, intended and adapted to the intention — that is to say *designed* — to increase knowledge, or, in other words, to remove doubt. Whatever degree of confidence we may feel in various assertions which can be made about the subject matter, there are, in these cases, at least some which we are willing to revise, or reject, in consequence of our observations. They may, of course, be very easy to reject, i.e. very crude, or few, or inaccurate observations may be sufficient to disprove them, or again they may be difficult to disprove in that they are compatible with common or crude observations, while a doubt still remains as to their truth. Further, it may be noted that only such assertions as influence our future expectations, beyond observations already made, have scientific relevance; e.g. the statement that Napoleon died at some exact age, though it may be a verifiable fact, is a contribution rather to history than to science. The statements which experiment is designed to test are to this extent all in the nature of generalisations, i.e. statements concerning some aggregates of possible future observations. For example, the statement that, on killing a particular sheep, we shall find in the intestine some particular species of parasitic worm can be proved true or false; but, except for its implication that this parasite may be found in other sheep which we have not killed, its verification has no scientific value. The statement that between x and y per cent of the sheep of the country contain the worm, or that some particular drug diminishes its incidence, are, on the other hand, both within the field of scientific experimental enquiry. Logically I suggest that the contribution to knowledge of any experiment lies in its capacity to disprove more or less conclusively some statement or generalisation respecting future observations which previously was tenable. If this is clearly borne in mind, a certain amount of supposed experimentation may be set aside as non-functional or irrelevant.

I do not clearly follow on what grounds you feel that doubt as to the precision of units plays any determining part in the nature of experimentation, nor really do I think that the amount of statistical work required in interpretation plays any central part. It is not very long since statistical methods have been brought into any relation at all with experimental design, and clearly throughout they are only one aspect of the subject, though co-extensive with it. Obviously the best designed experiments are those of which the interpretation is most plain, and these often require so little statistical investigation that we are scarcely aware there is any at all.

[1] Sayers, a physicist, had questioned Fisher's choice of title for his book *The design of experiments*, claiming that not all experiments come within its scope and that the role of statistics in the field of experiment is 'merely that of providing a special type of metrology'.

W.H. Sayers to Fisher: 24 *August* 1941

Many thanks for your letter of August 20th. I agree that our difference ultimately is in respect of the definition of 'experiment'. I think your interpretation is a result of long association with experimental work of a special type and that it is unduly narrow.

One of your fellow biologists (T.H. Morgan) says that experiment is 'The method by which Science separates the wheat from the chaff — namely the use of working hypotheses controlled by quantitative measurement'.

I do not regard this as a completely satisfactory definition but it does stress the essential characteristics of experiment — there must be a hypothesis, which is to some extent in doubt, and there must be an experimental result which can be expressed in the form of a measure of the validity of that hypothesis. It may be accepted that the term 'experiment' should be restricted to cases where the hypothesis to be tested is not trivial.

I imagine that you will agree so far.

Now on any such basis it is not permissible to dismiss, as you appear to do, the test of a single sheep for infestation by a particular parasite as 'not an experiment'.

I understand exactly why you have so dismissed this particular test — it is because you have tacitly made certain assumptions both as to the nature of the hypothesis which might be tested in this way, and as to the 'significance' of the result so obtainable on hypotheses of this kind. These assumptions would as a matter of fact usually be valid in your everyday practice. But suppose that the parasite in question is one which infests sheep in say Australia, but has never been known to occur in Britain. Suppose further the life history of the parasite to be unknown, but that you have formed the hypothesis that it enters the sheep with its food in a larval form, this larval form being known, but hitherto assumed to be a separate species. The suspect larval form is, like the parasite, unknown in Britain. You obtain a culture of the suspect organism from Australia, you introduce this into the food of a British sheep, and after a suitable interval, examine the sheep and find the adult parasite. Would this not be an 'experiment' and a highly significant one at that?

This example is a very interesting one from my point of view. Your hypothesis is that the suspect larva, α say, develops into the adult parasite β. From this hypothesis you deduce a secondary hypothesis from the first, that if α is introduced into a British sheep, β will subsequently appear in that sheep. If the test confirms the secondary hypothesis it confirms that the primary hypothesis is probably valid, the probability under the particular circumstances being obviously very high indeed.

But a more direct test of the primary hypothesis is conceivable. You may be able to find a suitable culture medium in which you can observe that α does in fact develop into β. If you can do this there is no question of confirming the primary hypothesis only to a high order of probability — you have confirmed

the initial hypothesis absolutely.

I do not suppose you will deny that both methods of testing the initial hypothesis are deserving of the title of experiment. But there is a fundamental difference between the character of the two tests. One confirms — or otherwise — by indirect inference what is to some degree uncertain — the significance of the result obtained is measurable only in terms of the probability that the indirect inference is valid. In the other there is no indirect inference with its attendant uncertainty. And for this reason I still maintain that there is a field of experiment which does not call for the use of statistical concepts in the analysis of results.

This field is fairly easily recognisable. You pose your hypothesis in the form of a question — for example, 'Does α develop into β?' and you answer it by saying, 'I believe (estimate) the answer is Yes'.

Now if your experiment, or test, gives you an answer to the hypothetical question in precisely the same terms as those of the question — if the answer, that is, is 'Yes, α does develop into β', no indirect or uncertain inference is involved and the experiment lies outside the field in which statistical methods are required in its interpretation. But if the answer given by your experiment is not in this form, if it is merely that certain results have been observed which are consistent with the hypothesis that α develops into β, you have then to examine what are the probabilities that this observed consistency is not the consequence of the validity of your assumption. And this probability is — under such conditions — never zero. In such cases, statistics — the metrology of probability — must play a part in interpreting the result.

I am quite sure you will not attempt to obscure the argument by pointing out that in fact there is always some measure of indirect inference in every real experiment, since your measurements are of uncertain precision. If the uncertainty of the measurements are within limits too small significantly to affect the final result, they do not introduce any practical element of uncertain inference into the interpretation of the result.

Fisher to W.H. Sayers: 1 *September* 1941

Thanks for your long letter. I see that there is a type of work which may involve experiment, the object of which is merely identification. In the framework of ideas every adult has a unique larval form, and *vice versa*, and if we exclude such real possibilities as that those of polymorphism in adults or larvae, or the possibility that critical differences in larvae may have been overlooked, then rearing a single larva, e.g. from marine plankton in an experimental tank up to a recognisable adult, would, in ordinary language, be an experiment of this logically simple type. It is analogous to the position of a palaeo-botanist who has recognised certain leaf impressions and fruiting bodies separately at a given horizon, that is, searching for a plant fragment that should prove the two to belong to the same species. Confirmation or refutation of such identification is usually supplied by totally independent workers, but is certainly felt to be desirable.

Personally I do not much mind if the word 'experiment' is commonly applied to such rather ideal special cases. For the worker with any future applications in view, frequency, e.g. of successive infestation, will always be of essential importance.

W.H. Sayers to Fisher: 8 *September* 1941

Thanks for your letter of September 1st. I am not quite clear that I have yet made my point quite clear to you. If an observed experimental result is expressible in precisely the same terms as the hypothesis which that result is to be used to check, then you may classify the experiment as pure 'observation'. It is this factor of the commensurate relation between experimental answer and hypothetical question which determines whether statistical analysis has any part to play in interpreting the result.

In a very large number of experiments in the realm of physics, you can so design your experiment that your result is in such commensurate terms — apart from the question of the precision of the instruments you are using.

In general the determination of this precision is a matter of calibrating instruments — and the calibration is an experiment on the instruments separate from the experiment for which you use the instruments. Whether practically speaking you need to apply statistical methods to the analysis of the calibration experiment depends on circumstances, but ultimately calibration must, to be of any real value, have a statistical basis.

Your type of biological experiment is rather like trying to carry out chemical analyses using reagents of unknown and variable purity and concentration. You can 'calibrate' chemical reagents, but you can't 'calibrate' a plot of soil for fertility, or a spell of weather for its influence on plant growth, etc. Therefore you have so to design your experiments that each experiment provides in itself the data from which you can derive some equivalent to a calibration of the reagents you are using in that particular experiment.

It is I suggest this experience with forms of experiment in which 'calibration' of the experimental equipment and measurement of the experimental result are necessarily part of one and the same experiment that has led you to make somewhat too extensive claims for the rôle of statistical analyses in experiment generally.

Fisher to W.H. Sayers: 29 *September* 1941

Of course I agree with the generality of the distinction you make between physical experiments in which the necessary calibration of the instruments is carried on on a different occasion, often at a different time, and in a different

institution from their experimental use, owing to the legitimacy in much physical material of the assumption that the calibration remains valid, whereas for biological material this assumption is often so palpably untenable that a good biological experiment has to contain within itself the controls needed to give confidence in the conclusions to be drawn. It is, I imagine, this common simplification of physical research which has led to the problem of experimentation being much more thoroughly elaborated in biological and psychological than in physical research. A good biological experiment, for example in agriculture, is in fact from the logical standpoint a far more perfect whole than a physical experiment need ordinarily be.

Click here for next section