2
Statistical Theory and Method

C.I. Bliss to Fisher: 13 *September* 1938

. . . I enclose an extract from my correspondence with Snedecor, upon which I would appreciate your opinion.

Letter from Snedecor dated Sept. 6, 1938:

'I found that in applying the inverse sine transformation to Moore's data (the numerical example in my Russian paper containing the table of angles),[1] the relative size of the means was reversed in two of the stages. That is, stage A had a larger mean percentage than stage B, but in the transformed values mean A was smaller than mean B. In that case what are you testing, the significance of the difference between percentages or between inverse sines? This seems to me to raise a question as to the validity of the inference made from the transformed values.'

My reply:

'Your question concerning the inverse sine transformation raises an interesting point. I take it that we are testing the significance of the difference between inverse sines on the assumption that these are the better index to the biological differences in which we are really interested. From the analysis of this experiment we know that the fumigations differed markedly, so that the successive estimates for any one stage form a heterogeneous series rather than samples from a single population. In averaging a series of this type, one would probably weight each contribution to the total by the information it contains. This the transformation does automatically, so that in testing the significance of a difference between inverse sines we are in effect testing the difference between corresponding weighted mean percentages. It is to be expected that some differences may be reversed by the transformation, especially when they are well within the sampling error as in the case you cite. In other cases they may change in magnitude from non-significant to significant or vice versa as a result of the transformation. If this did not occur, one would question the utility of using the transformation in the first place. At least this is the way it appears to me although I may have missed the point of your criticism.'

Is my interpretation correct?

[1] The passage in parentheses is presumably an insertion by Bliss.

Fisher to C.I. Bliss: 6 *October* 1938

I was interested in your correspondence with Snedecor, and enclose my own reaction to it . . .

'Shorter catechism' on transformations

A. Are you testing the significance of the *mean percentage*, or of the mean of the corresponding *angle*?
B. Strictly speaking I am testing the hypothesis that mortality is unaffected by a certain specified change of treatment.
A. But that implies that both the mean percentage and the mean angle are unaffected.
B. Yes.
A. Then how can it matter which you use?
B. I believe the mean angles may be the more sensitive.
A. What do you mean by more sensitive?
B. That the experiment tested with the mean angles will more probably show a significant deviation, if in fact the hypothesis is untrue.
A. Do you mean that every possible sort of deviation from the hypothesis will be more obviously apparent using the mean angle?
B. No, that would not be true of every conceivable sort of deviation. From my experience of variation in the experimental material, and of experimental conditions, I judge that the change in treatment under examination will, if it has any effect, cause more nearly a constant increase in the angle than a constant increase in the percentage.
A. Then your test of significance is chosen by judgement based on experience. Have not Neyman and Pearson developed a general mathematical theory for deciding what tests of significance to apply?
B. I believe that is the aim of much they have written. The absence of reference to experience seems to me a serious flaw in their work. Their method only leads to definite results when mathematical postulates are introduced, which could only be justifiably believed as a result of extensive experience. They do not, unfortunately, discuss the nature of the evidence for these postulates. If they did so, they would find that in practice it could amount to no more than the experience that one test of significance gave more frequently significant results than another. In general we may come to the conclusion that this is so, partly by direct trials on analogous material, partly from our general concepts as to how the observable effects arise. The introduction of hidden postulates only disguises the tentative nature of the process by which real knowledge is built up.

C.I. Bliss to Fisher: 2 *October* 1941

. . . In regard to table XX [in *Statistical tables*], one possible application has occurred to me which probably is based upon a misconception but would be extraordinarily useful if it were legitimate. It often happens that in varietal trials involving many varieties one wants to use the pooled variance for error in testing the significance between varieties arranged *a posteriori* in order of yield. In this case, of course, the just significant difference as usually computed has serious limitations. It has occurred to me, however, that it might be possible on the basis of the intervals between successive scores in Table XX to form an adjusted series of mean yields to which the just significant difference might be applied. On this scale, varieties near the ends of the distribution would be moved toward the center and those in the middle

of the distribution would be spaced more widely as determined by the intervals in Table XX. Hotelling gave me no encouragement when I suggested this possibility to him last July, in part because no paper is yet available describing just how Table XX was computed and the conditions under which it may be used. I have tried it myself with scores in a collaborative study on techniques for analyzing soils where 31 soils were ranked by 20 different observers. The transformation seemed to work nicely and its logic seemed perfectly obvious and convincing, but whether it could be extended to these other uses where tests of significance are particularly unsatisfactory at the present time is a problem for which you may have the answer. . . .

Fisher to C.I. Bliss: 24 *November* 1941

. . . In respect to the type of varietal trial you have in mind, I think the varieties giving intermediate yields, although in fact they will be crowded together, have, properly speaking, the same precision as those with the highest and lowest yields. The function of Table XX is to make the best use of observations in which, as in a race, order may be recorded more easily than any numerical measure can be obtained. I had thought that the formula on p. 14 and the explanation starting on p. 13 gave the theory of the thing sufficiently. . . .

C.I. Bliss to Fisher: 19 *July* 1956

Whenever I get really stuck on a problem and no one hereabouts seems to have the answer, I remember the ease with which you have found solutions for me in the past. Here is a problem that I believe has potentialities but is not treated effectively in any place that I have looked, although I suspect that it has been solved more than once in the past.

In a recent paper in *Science* (June 1, 1956), Peter D. King lists the number of births in each hour of the day in five hospitals. At my request, he has sent me the complete record for each of the five hospitals. The three largest, comprising some 86 percent of the data, differed less drastically in the number of cases. I have computed the analysis of variance on the attached sheet, the unit being the number of births per hour per hospital.

In fitting the data, I have followed in part Aitken's *Statistical Mathematics*, where he describes Fourier analysis under the heading 'Periodic Regression'. The first two terms seem to fit quite well as you can see from the data and diagram. Since the mean square for the interaction of series by scatter ($s^2 = 360.10$) is even less than the mean ($\bar{y} = 398.99$), I have assumed that the variation is essentially Poisson. Although the average scatter about the fitted curve shows a high significance, the distribution of the plotted points does not suggest any advantage in adding additional terms to the Fourier analysis. The observed small number at 7.00 a.m. coincides with the hour at which most

hospitals change nursing shifts, and the new and old shifts meet to discuss current nursing problems.

The problem on which I need your help is to determine from the equation of the fitted curve two biologically meaningful statistics and their standard errors or fiducial limits. (1) At which hour is the number of births a maximum or a minimum? (2) What is the amplitude of the cycle in number of births (or percentage)? These questions may need restatement but they represent the statistics which should be most interesting, I believe, to the biologist.

Perhaps you can refer me to a paper where all this is set out explicitly in terms that I can follow. If by any chance the problem should intrigue you sufficiently, I would be delighted to have you take over the data and such calculations as I have made and write a paper for *Biometrics* describing the analysis. . . .

Hour of birth in man. Data from Sparrow Hospital (A), W.C.A. Hospital (B), and Jamestown General Hospital (C) over 3 to 9 year periods. P.D. King, *Science*, *123*: 985–986, 1956.

Hour starting	Births per hour, y A	B	C	Total T_h	cos 0 a_1	sin 0 b_1	cos 2θ a_2	sin 2θ b_2	Mean $\bar{y}$	Predicted Y
12 Mt	370	447	421	1238	1	0	1	0	412.7	401.31
1 am	349	447	415	1211	0.966	0.259	0.866	0.5	403.7	416.81
2	357	475	434	1266	0.866	0.5	0.5	0.866	422.0	431.08
3	411	508	447	1366	0.707	0.707	0	1	455.3	443.16
4	422	469	440	1331	0.5	0.866	−0.5	0.866	443.7	452.24
5	457	498	474	1429	0.259	0.966	−0.866	0.5	476.3	457.70
6	415	505	427	1347	0	1	−1	0	449.0	459.14
7	398	436	448	1282	−0.259	0.966	−0.866	−0.5	427.3	456.50
8	425	490	450	1365	−0.5	0.866	−0.5	−0.866	455.0	449.92
9	435	478	470	1383	−0.707	0.707	0	−1	461.0	439.88
10	406	501	415	1322	−0.866	0.5	0.5	−0.866	440.7	427.06
11	382	429	420	1231	−0.966	0.259	0.866	−0.5	410.3	412.33
12 M	381	422	388	1191	−1	0	1	0	397.0	396.67
1 pm	357	427	337	1121	−0.966	−0.259	0.866	0.5	373.7	381.17
2	339	413	356	1108	−0.866	−0.5	0.5	0.866	369.3	366.90
3	308	375	300	983	−0.707	−0.707	0	1	327.7	354.82
4	355	401	353	1109	−0.5	−0.866	−0.5	0.866	369.7	345.74
5	295	355	344	994	−0.259	−0.966	−0.866	0.5	331.3	340.28
6	296	416	331	1043	0	−1	−1	0	347.7	338.84
7	292	340	302	934	0.259	−0.966	−0.866	−0.5	311.3	341.48
8	331	434	339	1104	0.5	−0.866	−0.5	−0.866	368.0	348.06
9	339	441	365	1145	0.707	−0.707	0	−1	381.7	358.10
10	309	383	380	1072	0.866	−0.5	0.5	−0.866	357.3	370.92
11	331	456	365	1152	0.966	−0.259	0.866	−0.5	384.0	385.65
Total	8760	10546	9421	28727	0	0	0	0	398.99 =	$\bar{y}$

		i	$\Sigma(a_1 y_i)$	$\Sigma(b_1 y_i)$
$\Sigma(a_1 T_h) =$	83.502	A	−130.455	726.644
$\Sigma(b_1 T_h) =$	2165.474	B	107.951	635.635
$\Sigma(a_2 T_h) =$	34.316	C	106.006	803.195
$\Sigma(b_2 T_h) =$	−143.434	T_h	83.502	2165.474
$\Sigma a_1^2 = \Sigma b_1^2 =$	12.000168			
$\Sigma a_2^2 = \Sigma b_2^2 =$	11.999648			

Predicted $Y = 398.99 + 2.3195\, a_1 + 60.1538\, b_1$

Source	DF	Sum squares	Mean square	$\chi^2 = SS/\bar{y}$	P
Between hospitals	2	67 949.20	33 974.60	170.30	<0.001
Hours: 1st term	2	130 449.57	65 224.78	326.95	<0.001
" 2nd "	2	604.21	302.10	1.51	0.5
" Scatter	19	17 837.88	938.84	44.71	<0.001
Hospitals x hours, 1st term	4	4 304.77	1 076.19	10.79	<0.05
" " 2nd term	4	453.44	113.36	1.14	0.89
" " Scatter	38	13 683.92	$s^2 = 360.10$	34.30	0.64

Fisher to C.I. Bliss: 26 July 1956

The problem you sent me looks to me very like the topic I discussed in *The Design of Experiments* under the heading of Section 64 ('Wider Tests based on the Analysis of Variance'). That is to say one should not, I fancy, be satisfied with a neat statement of fiducial limits for any one particular parameter, without regard to the other.

If one were to apply the method of that section, it would go somewhat like this:

The sum of squares for 2 degrees of freedom, which you call the first term, is, I fancy, $\{(83.502)^2 + (2165.474)^2\}/18$. Had you been considering not the hypothesis that these two components should be zero, but that your measures of them should have true values α and β, this sum of squares would be replaced by $\{(\alpha-83.502)^2 + (\beta-2165.474)^2\}/18$. To decide how large this expression could be without over-turning the hypothesis, consider your error for 19 degrees of freedom, which has a mean square 938.84, a natural logarithm 6.83393, and one-half the natural logarithm 3.41696, to which can be added the tabular z for 5% significance with 2 and 19 degrees of freedom, namely 0.6295, bringing the total to 4.0465, which is the natural logarithm of 57.2 approximately. The variance for 2 degrees of freedom would then be $(57.2)^2 = 3271.84$ and the sum of squares for the 2 degrees of freedom would be 6543.7, so multiplying by 18 one would have $(\alpha-83.502)^2 + (\beta-2165.474)^2 < 117786$, if the hypothesis is not to be contradicted at the 5% level.

Graphically this puts the hypothetical first harmonic pair of terms within a circle, or rather a series of concentric circles, for different levels of significance. I think I should be inclined to express the apparent precision of the observations in some such graphical form, rather than to particularize as to the fiducial limits within which the phase angle and the amplitude separately lie. Of course, for the phase angle separately I should use the method of the ratio of two means, as in Section 62.1, while I suppose the amplitude treated in isolation from the phase angle would be dealt with exactly by Mahalanobis's D^2 (of which, perhaps, tables have been worked by R.C. Bose), but as

you cannot help having a multiplicity of hypotheses in view, it seems to me more rational to treat them all simultaneously. . . .

C.I. Bliss to Fisher: 3 *February* 1958

Thanks for the offprint of 'The underworld of probability' [*CP* 267] . . .

Now that I have worked on periodic regression, I am discovering additional series where it would seem to apply. Also I am a little frightened by the many papers which discuss serial correlation. It seems to be the main hazard in applying the analysis of variance to periodic data. I hope I am not courting disaster in omitting this aspect which still seems to me a blind alley. . . .

Fisher to C.I. Bliss: 17 *February* 1958

Thanks for your note . . .. I do not believe you need worry about serial correlation. It leads, I think, to stochastic processes and general frustration, but perhaps there are better uses for it which I do not know about. . . .

Fisher to C.I. Bliss: 21 *August* 1958

Thank you for your letter. . . .

The trouble with variance components[1] seems only to have arisen through Tukey thinking that the matter involved simultaneous decision functions which in many cases would be exceedingly complex, as Duncan particularly has shown. If, without having heard of John Tukey, your own researches had given you measures of soporific power of seven different drugs showing significant differences, though of course every chosen pair would be significant at a different level, do you think that you would want to carry out a lot of further calculations on the seven values before you? Of course, if the second on the list was very much cheaper than the first, you might well want supplementary information, perhaps using these two only, from a further experiment.

[1] In his letter Bliss said he was 'struggling with a chapter on variance components' for a book that he was writing.

G.E.P. Box to Fisher: 19 *May* 1955

I have at some time heard or read that you said that whereas for a properly designed experiment we can perform a valid analysis, for an experiment which is not properly designed we can often only carry out a post mortem to decide what the experiment died of.

The above are not the exact words, and I would be very grateful if you would be kind enough to tell me if you did say something like this and if possible give me the correct version because, with your permission, I would very much like to quote it.

Fisher to G.E.P. Box: 21 *May* 1955

Yes, that was the kind of thing, and very much the *ipsissima verba* of what I used to say a goodish while ago at Rothamsted, say 1930–33 or so.[1] It served to bring home the fact that colleagues might take an interest in an experiment not only for its findings, but for its technique and even for its pathology.

[1] See *CP* 159, p.17.

W.G. Cochran to Fisher: 6 *September* 1938

I am interested to see your transformation for dealing with ordinal or rank data,[1] and should like to trouble you with a few questions about it. I should like to have seen an example of its practical use contained in the introduction. A good deal of labour has been and still is being devoted to the construction of tests of significance for ranked data, which your transformation, when its usefulness has been realised, will obviate. Can you refer me to any such practical example?

For amusement, I applied the transformation to the example *X X O X O O X O O O* which you discussed in your lectures. Comparing the difference between the mean score of *X* and *O* with the standard deviation (8 degrees of freedom) computed from differences within *X* and within *O*, I find *t* equal to 2.024, so that the probability of a better result is 0.039. According to your combinatorial solution, six cases were better, one hundred and ninety worse and fourteen neither better nor worse. If these are classed as better or worse according to the transformation, one finds that three of them are equal, one better and ten worse. This gives a significance level (for as good or better) of 10/210 or 0.0476, which shows a reasonable agreement with the *t* test.

Did you find any simple way to compute the values in the table? This at first sight seems to me rather a tedious business. Further, assuming that the data were originally normally distributed, what is the loss of information produced by replacing each value by its mean position?

[1] See Table XX in Fisher and Yates's *Statistical tables*, first published in 1938.

Fisher to W.G. Cochran: 7 *September* 1938

Thanks for your letter. I had made a few trials to satisfy myself that the *t* test would work well, and I am glad to see that you have done the same. The exact test of significance is, of course, that based on permutations, and from this point of view the scores are only conventional.

I think I arrived at them first by considering the problem: given the ordinal series for two variates drawn from uncorrelated normal distributions, what weighting would give an efficient estimate of a suspected correlation. The correlation of course is just the sum of products of tabulated values divided by the sum of squares given in the second table. This also, except for coarse

grouping in small samples, naturally agrees well with the correlation from normal deviates.

The computation of the table is considerably simplified by expanding the series of values for each value of n in odd orthogonal polynomials. The series terminates, e.g. for $n = 2$ or 3 the linear term is exact; for $n = 4$ or 5 the cubic term is exact, and so on; but the diminution of the numerical values is also very helpful, e.g. I suppose that the 13th degree is very nearly right, apart from the last value, for numbers up to 20 or 25. The coefficients of these polynomials are a good deal simpler than the original integrals, being free from the variable r.

Years ago — I write from memory — I found that the loss of information due to replacing variates by ordinal values is not very large, if it is remedied by replacing ordinal values by mean variates. My impression is that the sum of squares of the true mean deviates, which differs a little from what we have tabulated, is $(n - 1)$ times the efficiency. If this is so the percentage loss of information falls off proportionately to $(\log n)/n$ and is moderately small when n exceeds 10.

If you are interested you might look in some time and I will try to hunt up what I have in the way of algebra, as I have a strong impression that a good many pretty, though intricate, results could be obtained by proper algebraical treatment.

E.A. Cornish to Fisher: 31 *August* 1938

I should be very grateful if you would be so good as to give me some advice on the following problem.

I have three mutually correlated normal variables which I may designate x, y and z, and I wish to perform a separate analysis of variance on the observations of each variate derived from a rather comprehensive randomized block experiment. The various values of x are independent of each other, similarly with y and z.

Now we know that apart from variation due to blocks, treatments, etc., a certain amount of the variation in x can be explained in terms of y and z. We can, therefore, obtain from the analysis of variance and covariance of x, y and z the correction to adjust the residual term in the analysis of variance of x. May we now proceed in that same way using y as the dependent variate and x and z as the independent variates, and then with z as dependent and x and y as independent? Some doubts as to the validity of the procedure are in my mind owing to the assumptions about absence of error in the independent variates.

Fisher to E.A. Cornish: 12 *September* 1938

. . . With respect to your 3-variate problem, what you can logically do is to choose one of them, say x, and perform an analysis of variance without regard to y and z, then to do the same for y, making this time allowance for variation in x, and finally, to analyse z making allowance for x and y. These three analyses will give independent information, e.g. if the order of the variates is so chosen, you might find that the treatments had a significant effect on x, but that after this is allowed for they had no significant effect on y or z, jointly and severally, or you might find that there was also an independent effect on y, but no further independent effect on z. Since one wants the simplest interpretation, it is desirable to discover if it is true that one or more variates are unaffected by treatment when allowance is made for the variates on which the treatments had some real effect.

To allow for the other two variates in all three cases gives non-independent results, which might, indeed, all be non-significant even though each variate singly was largely affected.

H.E. Daniels to Fisher: 24 *September* 1937

. . . I have been interested recently in the problem of how the z test is affected by a slight correlation between the members of the sample. The question arose out of an attempt to detect periodic variations in twist from a series of consecutive twist tests on a length of yarn, trial periods p being taken in the usual way, the totals T_r of r, $r + p$, $r + 2p$, . . . found, and the variation between the T_r tested against the variance within the groups. The consecutive twist tests can be shown to be slightly positively correlated, and this must affect the significance of the results.

It had always seemed to me in an intuitive way that the effect of the correlation would be in a sense to reduce the 'degrees of freedom' by a fractional amount, but a tentative investigation, outlined below, seemed to show that neglecting squares of small quantities, the effect of the correlation is to retain the z distribution with the same number of degrees of freedom, but to introduce a bias into z. . . .

Fisher to H.E. Daniels: 27 *September* 1937

Thanks for your letter. A slightly different, but analogous, specification may be relevant to your problem, e.g. if, instead of considering a correlation between successive values, i.e. a specification in which $x_{p+1} - \rho x_p$ constitute a series of independent normal deviates, you take the not very different specification, that the normal and independent deviates are deviations of each observation from the expected mean ζ from a run of p consecutive values to which it belongs, these means themselves showing some variability, the effect of correlation reduces itself to precisely a bias in z, i.e. then $z - \zeta$ satisfies the z distribution where ζ depends on the variability of the successive expected means.

The main point of this is that you can sometimes clear your head and

simplify the algebra by choosing a specification involving the features you want and differing only in non-essentials from the one you are concerned with.

Fisher to C. Edelstam: 9 *December* 1948

Thank you for your letter of December 5th on the limits of significance for binomial and Poisson distributions. I rather think, however, that the tabulation you suggest would appear to be a step backward rather than forward.[1]

In the introduction to the book of tables you refer to, Example 1 shows how to get single tail values of 1% and 5% for the Poisson series, and Example 2 for the binomial, using the z distribution as tabulated.

In most problems of this kind very decent accuracy is obtained by interpolating for log P, if, and this is usually very exceptional, the experimenter wants to be particular as to what level of significance he chooses for making his test.

This is all part of a rather bigger question lying in the past between cases in which utility is best served by a compact table, which requires some expert knowledge to use, such as Stevens has aimed at in Table VIII.1, and large comprehensive tables, which are justifiable when in particular lines of work a very large number of separate entries have to be made, in which case it is convenient if the tables can be made almost foolproof. A great deal of scientific money has in the past been sunk in such large comprehensive tables for various purposes, and the corpus of such tabular publications which a comprehensive mathematical library can now acquire is indeed a very large one.

The situation has been in recent years greatly changed first by the introduction of convenient computing machines of desk size, which, for example, often make it easier to make a special tabulation for one's own use *ab initio*, rather than to seek by mathematical transformations to reduce the problem to expressions in terms of known functions which have already to some extent been tabulated, and then to use existing tabulations of the latter for one's own special purpose. A still larger change in the same direction appears to be impending in the construction in many parts of the world of high power electronic digital computers, capable of making rapidly *ad hoc* tabulations appropriate to particular problems.

I feel, therefore, that at the present time one should be very sure that a real need will be met before sinking labour and money in the production of largely expanded tables of functions which many may feel are already sufficiently available.

[1] Edelstam had suggested the preparation of an expanded version of Table VIII.1 in *Statistical tables*, incorporating more values for a, N, and P.

Fisher to D.J. Finney: 27 *September* 1954[1]

. . . I suppose it must be to assuage some uneasiness of conscience under this head that one finds in statistics, unlike let us say chemistry, an urge to warn students against genuinely accurate methods using various rather rhetorical grounds, instead of pointing out that some care is usually needed in choosing the amount of technical labour it is appropriate to give to any given job. I have been impressed lately by A.E. Mourant in the admirable book that he has written on the distribution of blood groups, putting in his chapter on computational methods quite a lot of counter-propaganda against the use of the method of maximum likelihood. It does not seem to occur to him that it is a purely factual question whether, at any stage of enquiry, doing more sums or collecting more blood is the more economical way of adding to our information. Rather it seems to be settled on emotional grounds by passionate appreciation of the supposedly insuperable difficulties which the more exact methods involve. Is the difference simply that we do not usually expect to do our chemical analysis for ourselves but that many people who have not the equipment, the time, or the technical training appropriate for the task, think that they ought to do their own statistical analysis, or does the trouble run more deeply than this?

[1] For the first part of this letter, see p.95.

D.J. Finney to Fisher: 4 *October* 1954

. . . Of course your comments on the economic aspects of information wasting are important, and I have often emphasized this point. I have not yet seen Mourant's book, but am sorry to learn of his timidity in respect of maximum likelihood. I would have said that the effort of collecting human genetic records was so great that, even if M.L. estimation took twice as long as some cruder method, it was worth while for as little as an extra 5%–10% of information. It is strange how the biologist who will cheerfully spend weeks on field work is scared by the thought of a few hours of computing.

I imagine that the simplicity with which electronic machines can cope with elaborate iterative procedures will shift the economic balance still further in the direction of extracting all the information from data. . . .

Your last sentence provokes the question: should we attempt to train biologists and others to do any statistical analyses for themselves, or should we seek to reserve their time and abilities for their own field of work and pass all statistical work through bigger and better Statistical Departments? I do not think the matter is quite dealt with on the analogy of chemical analysis, because I suspect the investigator can learn truths about his material from his own statistical analyses that the statistician might not observe or might fail to transmit to him, and this is perhaps less true of chemical analyses. Unless and until the statistical analyses of a much greater range of problems become

purely mechanical tasks, isn't the answer that the best procedure depends upon the particular problem and upon the particular people involved?

Fisher to D.J. Finney: 7 *October* 1954

. . . You raise a very big question in speaking of the influence of high power computing machines. What needs, I think, to be safe-guarded is the practice of the scientist or, at least, a statistical consultant giving personal consideration to what is to be learnt from his data as to the hypothesis (or model) in terms of which they are to be interpreted. Short of that it seems to me that the geneticist, or ecologist, or whoever manages to gather the data, has virtually put himself into the position of asking the machine's masters what he has to observe for them. I suppose the answer is that no one can stop the human race from selling themselves into slavery, still one would like somehow to try.

I had a Ph.D. thesis to examine recently in which the young man, impressed by the large numbers of observations needed to distinguish ratios different to a practically important extent, mobilised the aid of a giant computer, but still passed through to it for programming the approximate methods which he had been accustomed to use when labour saving was really important to him. I think the point of this story is that he thought it was the job of the machine only to do the work for him, instead of realizing that it was a competent method of finding out whether his approximative processes were sufficiently accurate for the work required.

So I do think we should try to train biologists to do their own work with at least enough competence to know when further consultation is needed. I rather think the chemists have succeeded in this respect.

D.J. Finney to Fisher: 3 *April* 1956

I should be very interested sometime to have your opinion on what constitutes an analysis of variance. I think that authors of papers and reports frequently include details of these analyses unnecessarily, but obviously when questions of statistical technique are under discussion inclusion is essential. I find an increasing tendency to present, under the title of 'analysis of variance', a tabulation of degrees of freedom and mean squares *only* for the various components, the column for sum of squares being omitted. Since it is the total sum of squares of deviations that is analyzed, omission of this and its components seems to me very unfortunate. I regard the preparation of a full table of the analysis of variance as an excellent discipline to impose on computations, and the author of an expository paper who suggests that parts of it can be omitted may be doing a dis-service to those who follow him.

This is particularly relevant to papers submitted for a journal such as *Biometrics*. Do you think that an author should be asked to show the full table, or are you content to see merely the mean squares? I have heard it said that you dislike the practice of adding a further column showing ratios of

treatment mean squares to error. I would never include these myself as I think they encourage excessively rigid interpretation of significance tests but I am not sure how strongly one ought to discourage them. A paper in front of me for refereeing this moment has a column headed *P* showing the probability levels for the ratios of mean squares to error mean square. What are your views on this?

Scarcely any statistical technique is more widely useful than the analysis of variance, but the novice today is confused by many styles of presentation. Some put the total at the top and some put it at the bottom. Some put degrees of freedom before the sum of squares and some reverse these. I am conservative enough to prefer the order that you have always used, but I have difficulty in deciding how far editorial pressure should be exercised in persuading other people to follow a standard pattern.

Fisher to D.J. Finney: 16 *April* 1956

. . . About the analysis of variance, I agree with you that such analysis must include a column for sums of squares, or its equivalent, and of course to the early workers, like myself, one of its main attractions lay in the automatic verification that everything had been accounted for by the check, or checks, provided by the totals in this column. Now that the procedure is widely used, I believe academic exposition should insist, far more than I have done in more or less broaching the subject in *Statistical Methods*, on tearing the sums of squares to pieces, in as many ways as possible, to see that every aspect of the data is behaving approximately as it should; you do not need to identify the microbe before using an antiseptic! I think editors could very usefully make this point, subject of course to the author's right, either to present an analysis of variance *senso strictu*, or a summary of an analysis, stated as such, or an elaboration with additional columns, such as you mention, with variance ratios, *P* values, or what you will. . . .

Fisher to D.J. Finney: 7 *November* 1956

. . . I notice in a review, in *J.R.S.S., A,* the last number, of a recent book by Federer, the reviewer reproves the author for quoting, apparently without disputing, the allowance I calculated for the loss of information about a mean owing to using an empirical estimate of error instead of the true, but unknown, variance.[1] As I think you were in these discussions, perhaps you could enlighten me as to what the reviewer can refer to as 'alternative methods'. Of course, if additional data are supplied, e.g. the value of t expected, the problem is reduced to 'Student's' test itself. My calculation is integrated over the whole t distribution, and seems to me appropriate in the realistic case in which the value of t, or probably the whole crop of values of t, for which comparisons are as yet unknown, and the planner of the experiment is concerned to judge how much he should sacrifice in order to have say 20

degrees of freedom, in place of 10, for the estimation of error. I believe confusion may have been caused by trying to bring in the sampling distribution of s for given σ whereas in reality the experimental planner knows that only s will be available, and that σ, if thought of at all, will have a rigorous fiducial distribution appropriate to that s.

[1] See also Fisher's letters to J.A. Nelder (p.280).

D.J. Finney to Fisher: 10 *November* 1956

I too thought that the reviewer of Federer's book had expressed himself rather badly over the point you mention, though I feel in considerable sympathy with him over the irritating manner in which the book presents its methods.

I imagine that the reviewer has in mind the kind of thing that Cochran and Cox discuss in their Chapter 2, and particularly on pages 26–29. They give a brief but sensible general summary and state clearly that the answer depends upon the use to which the results are put; I imagine that you would agree with this sentiment, though it might perhaps be better expressed by saying that the results depend upon the question that is asked. I don't know whether anyone has suggested alternative answers to your question, but clearly it is possible to frame alternative questions bearing on the same subject, as you point out in your letter. . . .

Fisher to D.J. Finney: 12 *November* 1956

Thank you for your reference to Cochran and Cox pages 26–29. I believe if you will re-read the final paragraphs of this section (2.31) you will find that, after considering what Neyman and Welch had written, these authors conclude, 'For general purposes it is suggested that this table' (the one embodying my form for the correction) 'be used to take account of the differences in degrees of freedom for error in two designs that are being compared.' They then give an example and an explanation of the cases in which such an expression might be useful.

The table on page 27, which I suppose is Neyman's or Welch's, clearly treats the question which conceivably could arise if the level of significance were known in advance, although in my opinion this problem had been completely solved by 'Student' himself.

Anyone reading Nelder's review could see, and doubtless you can see, that the reviewer was criticizing Federer for recommending what Nelder regards as an unsound method, for in addition to suggesting, as he doubtless believed, that alternative solutions had been given the same problem, he complains that my demonstration of it is peculiar, and that I have not used the method on other occasions; whereas the formula I used for assessing the amount of information about a parameter distributed in a known distribution was in my 1921 paper 'On the Mathematical Foundations' [*CP* 18], and had been

illustrated in the same book in almost step by step parallelism in the previous section.

The question I was trying to answer was what allowance should be made in anticipating the precision of an experimental design for the number of degrees of freedom on which estimates of error are to be based. Apart from the question of alternative methods of solution of this problem, which, so far as I know have not been put forward, you say that 'clearly it is possible to frame alternative questions bearing on the same subject' and imply that this is in some way a justification for scolding Federer for advocating, however stupidly, the only method so far suggested, for what is in fact a question which does arise in practice.

I am not concerned with the merits of Federer's book, but with the use made of a review for an oblique attack on a method which the reviewer has not perhaps considered very deeply, but has been led to think, in the way that prevails in centres of statistical teaching, is unsound, not for any tangible reason but because someone else disagrees with it.

D.J. Finney to Fisher: 20 *November* 1956

. . . When I returned here, I reread the letter from you that arrived in my absence. I am in complete agreement with you about the misleading nature of the remarks in the review of Federer's book. The further comments that I made in my previous letter were not intended in any way to modify this opinion which relates to the factual content of the book and of the review. Since I had myself reviewed the book, however, I was disposed to criticize it on quite different grounds of style of presentation, not in any unique instance but in the manner in which the author presents references to a great number of statistical writers.

M. Fréchet to Fisher: 8 *February* 1936

. . . Je sollicite donc de votre obligeance l'envoi (à mon adresse: 12, Square Desnouettes, Paris, 15e) d'une Note indiquant — en quelques lignes, et sous forme de règles concises — en quelles circonstances et sous quelles réserves le coefficient r peut être utilisé *pour repérer* la rigueur d'une dépendance fonctionnelle. (Pour éviter tout risque de malentendu, je vous serais très obligé de bien vouloir laisser *en dehors* de votre note les applications bien connues de ce coefficient r à la représentation mathématique des lois statistiques, à l'ajustement, à diverses questions de physique, etc. . . .). A cette courte Note pourrait être adjointe, s'il était jugé nécessaire, des explications plus développées mais traitées séparément. . . .

Fisher to M. Fréchet: 19 *February* 1936

I must begin my reply to your circular letter of February 8th by a purely

verbal difficulty, namely as to the exact meaning which should be attached to the word 'repérer', underlined in the middle of your second page. Should it be taken to mean simply to discover? In my ignorance I had recourse to my copy of Petit Larouse where I find 'Marquer des repères', while a repère is 'Marque faite à différentes pièces d'assemblage pour les reconnaître et les ajuster plus facilement'. This literal definition cannot be what is wanted, but I have missed the metaphor which should make it intelligible.

In the sense of simple discovery I have seen the correlation coefficient used with success, e.g. a soil physicist studying the plastic qualities of agricultural soils might find in the samples examined no association with the total percentage of calcium carbonate in the soil, but be delighted to find that a significant association appears when particles of calcium carbonate, too large to pass through holes 1 mm in diameter, are excluded from the analysis. The search for such a significant association among the many variables, simple and compound, which may be examined, is carried out not inconveniently in terms of the correlation coefficient, though certainly also not inconveniently in other ways. It is used, I think, in preference to other measures chiefly because the method of calculation is widely known and easily remembered, and because relatively precise tests of significance are easy to apply, owing to the availability of published tables.

The test of significance, i.e. the test of the hypothesis that there is no real association in the material from which the existing observations are regarded as a random sample, is necessarily equivalent to the test whether the coefficient b in the equation

$$y = a + bx$$

fitted to the data by least squares, as a means of predicting the value of y from that of x, differs significantly from zero. This is often the more appropriate and fruitful way of viewing the matter, but since the tests of significance are equivalent, and since, in the business of *discovery*, we are only concerned with significance and not necessarily with estimation, or even with knowing what we want to estimate, the correlation coefficient may be legitimately used in this way, even when its numerical value is without scientific meaning beyond that given it by its mathematical definition as a calculable function of observable quantities.

To say that this procedure is legitimate is, of course, not to say that it is the best that can be recommended. In particular cases other tests may be much more sensitive, that is to say they may be capable of demonstrating a significant association on the basis of fewer or less careful observations. This is an important advantage in statistical theory, but it is of little interest to the experimenter who, having found a significant correlation, is led, perhaps, to abandon a false theory by which he has hitherto guided himself, or to undertake a new line of research with different aims and different instruments.

M. Fréchet to Fisher: 22 *February* 1936

I thank you for your interesting answer to my circular letter.

I. I want to explain that the word 'repérer' was intended as a substitute for 'to measure'. I wanted to avoid the idea that a 'measure' could be found for the degree of strictness of dependency — as there is no meaning for the sum of two degrees of such strictness. I should think that the verb 'to graduate' is the proper equivalent of 'repérer'; for instance we can 'repérer' (graduate) but not 'mesurer' (measure) the physical feeling of heat.

II. Your answer, being based on the approximate translation of 'repérer la rigueur d'une dépendance' by 'discover the existence of a dependence' relates to a more restricted problem than the one I meant. Still, this restricted problem is important.

However, I do not see well your exact position and, if I am not too troublesome, I should be glad if you could precise [*sic*] the following point. For, your first page (first example) should evoke the impression that your soil physicist operated this way. As it is intended as an example of the uses of the coeff. of correlation, I understand that in his first experience he found this coefficient small and that it is this way that he concluded that there was no association between . . . . Then in the second experience, he found the coeff. near 1 and it is that way that he found a significant association in the second experience. These conclusions were obtained, as you say, not inconveniently in terms of the correl. coeff. as they might have been got other ways.

In your second page and end of third page, I gather the impression that a test of significance should be limited to the case when r is near 1, and that no conclusion from the value of r should be derived when r is near 0. The first page should correspond to what is very often done, the second one to a more safeguarding method.

But I have perhaps been mistaken in both cases, and a few words of comment would be much appreciated.

P.S. In my recent travelling in Russia, Prof. Kolmogoroff — who is one of the half dozen best 'probabilists' in the world — observed to me that you have recently come to a differ. equation[1] without probably knowing that it has been met before in a physical problem by Fokker and Planck. It may be interesting to you to hear it is called Fokker–Planck's equation by the physicists and that it may be written:

$$\frac{\partial p}{\partial t} = -\frac{\partial Ap}{\partial x} + \tfrac{1}{2}\frac{\partial^2(B^2p)}{\partial x^2}$$

[1] See *CP 86*.

Fisher to M. Fréchet: 5 *March* 1936

You ask me to explain more clearly the kind of case I had in mind as an

example of the detection or discovery of an association by means of the correlation coefficient. This is easy, for in both cases to which I referred I had the same criterion in mind, namely whether the correlation coefficient does or does not exceed the least value which should be judged significant in regard to the number of observations available. Thus, if the soil physicist had 22 samples of different soils available, the properties of which he had examined and measured, he would probably choose the value 0.4227 as the least observed correlation which should be regarded as significant, the reason for this choice being simply that from uncorrelated and normally distributed populations this is the value that would be exceeded by chance just once in twenty trials. In order to judge a correlation to be insignificant, therefore, it is not necessary that the observed value should be very small or near to zero. He would ignore it with confidence coming from so small a sample if it were between plus and minus 0.35. On the other hand, a value as high as 0.55 would only occur by chance once in several hundred trials in samples of this size in a population from which correlation was absent, so that the value of the correlation need not be near to unity in order to satisfy the experimenter that the value he has determined does demonstrate the existence of a real connection. It is the reality of this association which is vital to the experimenter, for on his confidence in its existence he is willing to spend, and perhaps to waste, the work he is capable of doing for perhaps several years.

What is important in this is that he should be provided with appropriate and sufficiently exact tests of significance. It is a matter of no consequence that in making these tests he should use the particular function of the observations (statistic) which we denote by r. In fact, we arrive at an exactly equivalent test if, ignoring the correlation coefficient, we calculate the regression of either one variate on the other by the familiar formulae of least squares and compare this with its standard error,using the exact procedure introduced by 'Student' for allowing for the limited number of degrees of freedom upon which the estimate of error has been based. It is, in fact, only the test of significance in its entirety which is of value to the experimenter, and not the particular statistic in which he happens to find it convenient to carry out the test.

M. Fréchet to Fisher: 13 *March* 1936

I thank you very much for your answer. It throws light upon some points which your letter *published in my enquiry*, left obscure for me. In that former letter you treated simultaneously the two very different points: is there a dependence; what best can be done to *describe* the simultaneous variability of 2 quantities. Now in your last two letters the first point is treated separately, and then I can see that:

your method is based on the fundamental assumption that we have to deal with quantities satisfying to the so called normal law of errors (with 2 variables).

On this assumption your method — very different from the usual practice — looks very satisfactory.

And I think I will be able to recall it in my report.

Now, I am afraid to retain much of your time, but there remains the cases — very frequent ones—: 1st when the observer does not know whether the variation is normal or even suspects it to be slightly different; 2nd when the observer knows that the variation is far from being normal.

Is your position such that: I. we *must not* use the coeff. of correlation r in such cases *as far* as concerns the *existence* of a dependence. II. or that you possess in such cases a method to use it for *this same purpose*, a method necessarily different from the one you mentioned?

Fisher to M. Fréchet: 18 *March* 1936

The test of significance mentioned in my letter is somewhat less restricted in its application than you suggest in your reply of March 13. I mentioned that the formal conventional test of significance of the correlation coefficient was, in fact, equivalent to the test whether the linear regression of one variate y on the other, x, differed significantly from zero. The test is, therefore, exact not only for the case of a normal bivariate distribution, but also for one in which one variate is distributed in any manner whatever, while for given values of x the second variate y is distributed normally about a mean which is a linear function of x. The marginal distributions of both x and y may, therefore, both be far from normal, and for the test of significance it is immaterial which of the variables is regarded as the independent, and which as the dependent variate.

M. Fréchet to Fisher: 22 *March* 1936

I understand that the test you mentioned in your second letter is applicable in a wider range of cases than the cases in which there is a normal bivariate distribution, but that it still requires that one of the 2 lines of means be linear (and even it requires a little more).

Now my question is:

as far as concerns the cases in which none of the 2 lines of the means is a straight line, do you think that the use of r (for judging whether there is a significant dependency) is incorrect?

or do you know, for these cases, of a method — necessarily different from the one you mentioned — enabling to use correctly the coefficient of correlation for the same precise purpose?

Your previous letters were very useful to me; I hope that you can answer the above questions and so help me to define your position in the previous enquiry.

Fisher to M. Fréchet: 24 *March* 1936

I will try to answer your further questions of your letter of March [22].

The investigator who has obtained a single correlation and who merely infers that the variables are not independent is quite immune from criticism on the ground that the regression may be non-linear; for this [proposition] if true, would only [decrease] the sensitiveness of his test, and not increase the frequency with which, in the absence of association, high values of r would be observed. The investigator who finds no significant correlation should always be cautious not to say that no correlation exists, for it may be, as in the case of the physicists of whom I first spoke, that he has failed to detect an important relationship, through using a physical quantity less suitable than he might have used, e.g. total calcium carbonate instead of the calcium carbonate in small particles only. He may have failed equally for the reason you have in mind, namely, through seeking for correlation with a linear function of his observed value, instead of using a function of more complicated kind which would have revealed what it is in his interest to detect. In fact, I should say that the choice of one form rather than another for the regression equation to be examined is in the same sense a matter of the individual judgement, or intuition, of the investigator as is the choice of the physical attributes to which he is to devote his attention.

It may be worth noting that an investigator who feels more sure of what physical quantities to use than he does of what functions of them would be the most appropriate, can satisfy himself that the data before him deserve no more elaborate tests than he has applied by making a test of linearity of regression (Blakeman's criterion is, of course, quite incorrect). For, if there is no indication of departure from linearity in a test specially designed for detecting such departure, he has good ground for confidence that any curvature which exists will not have appreciably disturbed his judgement based on a linear regression function.

Fisher to L. Goossens: 28 *April* 1950

Thank you for your kind letter which I have just read. . . .

I think one can best get an idea of the argument[1] in paragraph 21.1 [*SMRW*] by putting the problem to oneself quite independently of the χ^2 test. Suppose we are told that a hypothesis is suspect, though difficult to test directly, and that efforts to do so by summarising groups of data which should be to some extent sensitive to the truth of the hypothesis have led on different occasions, using entirely independent data, to reliable values

$$p_1, p_2, p_3, \ldots$$

for the probability of obtaining a more extreme deviation calculated on the

supposition that the hypothesis is, in fact, true.

On this hypothesis, therefore, these values p must be distributed between zero and unity in a rectangular distribution.

If the hypothesis is false we expect an excessive number of small values and a deficiency of large values compared with this standard rectangle. Arbitrarily, therefore, we might choose to multiply the values of p together to determine the sampling distribution of the product and to base a test of significance on whether this product appears to be significantly smaller than one would expect it to be by chance.

One may then bring in the idea of χ^2 for two degrees of freedom as supplying an easy way of solving the problem of the sampling distribution of the product of such a group of values p. Of course, such an approach is only valid if the distribution of p is truly rectangular and continuous, as it is effectively in many problems.

I hope this approach will remove your difficulties.

[1] i.e. on the combination of probabilities for tests of significance.

Fisher to H.W. Heckstall-Smith: 25 *July* 1957

Thanks for your letter of 24th.[1] The remark in *Statistical Methods* was, of course, paradoxical and intended to make people think, as it obviously has in your case. Of course the phrase 'In these cases the hypothesis considered (etc.)' means the hypothesis 'that the data under discussion were derived from a particular hypothetical system of causation'. In fact, logically the data are a part of the hypothesis. In most cases the hypothesis is more complex than this because its full specification involves assertions about independence, which are essential to the mathematical specification of the expectations, but often pass unnoticed by the mathematician discussing scientific questions. In the case of your coins one might imagine that the performer counted the result of the first throw, say a head, and then the next time tails turned up, and then the next heads, and so on, omitting all repetitions. As you say, the data should be rejected on the grounds that they were not obtained by independent random trials, and, as I have put it, the hypothesis that the data were obtained by independent random trials with a true chance of fifty-fifty (or any other for that matter) is to be rejected.

[1] Heckstall-Smith had written querying the statement in Section 20 (p.81) of *SMRW* that in tests of goodness of fit where P is over 0.999, 'the hypothesis considered is as definitely disproved as if P had been 0.001'. He suggested instead that 'either the evidence must be rejected or else the hypothesis must be considered as disproved'. He asked in particular about a hypothesis that a coin is unbiased and supposed that on tossing the coin, he found 500 heads and 500 tails; should he then reject the hypothesis that the coin is unbiased?

Click here for next section