butes no information about λ in the likelihood sense, it does enable us to convert the likelihood statement into a probability statement. I hope I make myself clear and that I have not misunderstood the situation. I would like to discuss this and the other question when I come to Cambridge.

[1] See *CP* 289.

G.A. Barnard to Fisher: 13 *July* 1961

David Sprott, whom I mentioned to you as having written two papers concerning the connection between the likelihood function and the necessary form for the fiducial distribution, has come over here and we have been discussing a number of points in connection with the fiducial argument. We have reached a stage where we feel very much in need of your help and I wonder if we could come to see you on Monday of next week (the 17th) to discuss it with you.

If I may briefly indicate the nature of the difficulty it is this. At the end of *Statistical Methods and Scientific Inference* you derive the fiducial distribution for the normal bi-variate [parameters] ρ, σ_1, σ_2. At the same time you point out that one could use another set of pivotal quantities, one of which is the variance about the regression line, which might conceivably be misused to derive another 'fiducial distribution', for the same set of parameters. The difference between David Sprott (and incidentally Quenouille) on the one hand, and myself on the other, is that they seem to think that the second set of pivotal quantities are appropriate for a fiducial distribution of the parameters corresponding to them; but that this fiducial distribution is not transformable to a distribution of ρ, σ_1 and σ_2. For my part it seems to me that this second 'fiducial distribution' is just wrong, since when it is transformed in the normal way to its expression in terms of ρ, σ_1 and σ_2 it contradicts the validity principle.

This is just to indicate the kind of issue involved and I hope we may explain it more fully when we see you. . . .

Fisher to G.A. Barnard: March (?) 1962[1]

Your letter Feb. 15 was forwarded from India, and I have just read it. I think we have arrived at nearly indistinguishable conclusions. Let me now formulate my own position.

A pivotal quantity is a function of parameters and statistics, the distribution of which is independent of all parameters. To be of any use in deducing probability statements about parameters, let me add

(a) it involves only one parameter,
(b) the statistics involved are jointly exhaustive for that parameter,
(c) it varies monotonically with the parameter.

As you have observed, and as the last example in *Statistical Inference* was intended to make clear, parameters and their corresponding exhaustive statistics may arrange themselves in strata.

For the normal sample s alone is exhaustive for σ, consequently the marginal distribution of σ is expressible in terms of s, using the pivotal s/σ. Both s and $\bar{x}$ are needed for exhaustive estimation of μ using $(\mu - \bar{x})/s$. In this case it is noticeable and probably essential that the two pivotals together each involves one parameter. *Jointly* they involve a set of statistics exhaustive for both. Each is monotonic in its parameter *uniformly* for variations of the other. Also, I should have stipulated that their simultaneous distribution is independent of all parameters.

That I think is enough. In the bivariate case s_1/σ_1 and s_2/σ_2 have a joint distribution independent of σ_1, σ_2 but not of ρ. However, for arbitrarily assigned ρ, they suffice to give the simultaneous distribution of σ_1 and σ_2 in an array (with ρ constant) and this suffices for the trivariate distribution.

For sets of pivotals then I add

(d) the joint distribution is independent of parameters (of as high or higher stratum)
(e) all are monotonic, uniformly for variations of parameters of as high or higher stratum.

See if you can formulate a justification for ignoring parameters of lower strata once their simultaneous marginal distribution is determined.

Fraser wrote hopefully about the trivariate case, but he may not see all the difficulties, e.g.

$$\frac{r_{12}}{\sqrt{1-r_{12}^2}}\chi_{N-2} - \frac{\rho_{12}}{\sqrt{1-\rho_{12}^2}}\chi_{N-1}$$

is distributed in a standardized normal distribution, and is uniformly monotonic in ρ_{12} but is unlikely to be distributed independently of

$$\frac{r_{13}}{\sqrt{1-r_{13}^2}}\chi_{N-2} - \frac{\rho_{13}}{\sqrt{1-\rho_{13}^2}}\chi_{N-1}\,.$$

In fact the distribution of r_{13}, r_{23} conditional on r_{12} is not at all alluring.

I expect I told you I had run out the simultaneous distribution of r_{ij} given the system ρ_{ij} for any number of variables.

[1] This letter was published in Barnard, G.A. (1963). Fisher's contributions to mathematical statistics. *J. R. Statist. Soc.* A **126**, 165–6.

M.S. Bartlett to Fisher: 25 *September* 1933

Though no doubt you have seen my paper in the last Royal Society *Proceedings*,[1] may I take the liberty of sending you a reprint? Since this paper was accepted, Jeffreys' latest paper in the *Proceedings* was published;[2] and perhaps my intrusion into a 'controversy' which is primarily between Jeffreys and yourself can very easily become somewhat superfluous. I do not, however, remember that Jeffreys' paper had in it anything to alter the point of view I have tried to take up in my own paper, though I feel conscious that this view point I may have rather inadequately expressed.

[1] Bartlett, M.S. (1933). Probability and chance in the theory of statistics. *Proc. R. Soc.* **A 141**, 518–34.
[2] Jeffreys, H. (1933). Probability, statistics, and the theory of errors. *Proc. R. Soc.* **A 140**, 523–35.

Fisher to M.S. Bartlett: 26 *September* 1933

I had not before seen your paper which I am sure will be interesting. It is as you suggest a pity to have published before Jeffreys wrote, as effectively his reply is an attempt to defend a simple blunder, or howler as it is called with boys, by a confusion of definition. Of course it is easy to find ample precedent for such confusion in the controversial history of the subject.

M.S. Bartlett to Fisher: 30 *December* 1935

In your recent book *Design of Experiments*, I was interested to see your concluding section on the information supplied when the variance σ^2 is unknown, particularly because I had looked at this problem some time ago and had not been satisfied with the solution I reached. It was for this reason that in a paper[1] I wrote at the time (not yet out) in which I was attempting to look at the general problem of small samples with more than one unknown, I classed the problem you discuss in a different category from other problems in which the reduction in the amount of information available when another parameter is estimated is more straightforward.

Looking at the t-distribution problem again in the light of your solution, I feel, though I would agree that we can, if we want to, regard the intrinsic accuracy of the t-distribution we are using to estimate the mean μ as

$$\frac{n+1}{(n+3)s^2}, \tag{1}$$

that the interpretation of this result rather raises the whole question of the interpretation of information for small samples. Thus it is clearly not permissible to imagine any actual set of samples from which we are going to estimate μ, since s^2 would vary for such a set. We might, however, consider such an actual set, s^2 varying as well as x, and our estimate would then, according to my reckoning, contain the information per sample,

$$\frac{n-1}{(n+1)\sigma^2}. \tag{2}$$

This vanishes for $n = 1$, but this simply means that our method of estimation (which is equivalent to supposing we know nothing about σ^2 *for each sample*) fails when s^2 is based on only 1 d.f.

Without going into the question here of the interpretation of the information in small samples generally, I rather favour the consistency of (2) if we want some measure of the information we are using, though it is certainly not quite what we wanted. (1) has the disadvantage that it cannot correspond to any actual procedure; it obviously cannot, for example, be averaged for s^2.

[1] Bartlett, M.S. (1936). Statistical information and properties of sufficiency. *Proc. R. Soc.* **A 154**, 124–37.

Fisher to M.S. Bartlett: 31 *December* 1935

Thanks for your letter of December 30th.

I am afraid I don't yet see where your formula

$$\frac{n-1}{(n+1)\sigma^2}$$

comes from and therefore what it is put forward as meaning. The meaning of (1) is that different experiments for which the formula has the same value supply estimates of equal intrinsic accuracy, though the error curves are different in form.

Are you not perhaps bringing in the different conception of forming a weighted average from samples which are not known to have the same true variance?

I shall be glad if you are able to give more time to examining the meaning of information for small samples, for it is just in this case that the concept seems to me valuable.

M.S. Bartlett to Fisher: 2 *January* 1936

Many thanks for your reply to my letter. I am sorry that I did not make myself altogether clear. The second formula

$$I = \frac{n-1}{(n+1)\sigma^2}$$

I gave is the information obtained from a sample when we are estimating the mean from several samples for which we do not wish to assume the constancy of σ^2. This is admittedly a different conception from the intrinsic accuracy of an estimate obtained from any experiment as given by your formula. My point was that before this type of problem was considered, there had been no need to make such a distinction, intrinsic accuracy in the case of one parameter being defined in terms of the accuracy obtained from a large set of estimates. If we go on to take the case of a regression coefficient (σ^2 assumed known) the information in any particular sample will depend on our value of $\Sigma(x - \bar{x})^2$ but when we consider the corresponding problem of the information in a large number of samples when the x's vary, the information per sample is simply the average value of the information for any one sample. The same sort of problem can occur in some cases where instead of $\Sigma(x - \bar{x})^2$, we have an estimate of another unknown. It does not apply in the problem we are considering, (the reason being that s^2 is not the theoretically complete statistic for σ^2).

I agree that the second formula involves a different conception from the first, and is hardly what we wanted; I referred to it because, given its proper interpretation, it seemed to be immediately consistent with preceding definitions of information. I was somewhat puzzled in connection with the first formula to know how you were regarding the definition of intrinsic accuracy, in view of the fact that the idea of several values from the same t distribution of $x - \mu$ would not appear to have any direct practical procedure to correspond to it.

Fisher to M.S. Bartlett: 8 *January* 1936

I am sorry, but I have still not got at the origin of your formula

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

I suppose n here stands for the degrees of freedom, as it does in the other case.

M.S. Bartlett to Fisher: 9 *January* 1936

If we confine ourselves to the problem of estimating m from a large set of samples with the same true variance (possibly out of a larger set with differing variances), instead of the equation of estimation

$$\Sigma\,\frac{x-m}{\sigma^2} = 0,$$

which we should have if σ^2 were known, we might anticipate the equation

$$\Sigma\,\frac{x-m}{ns^2 + (x-m)^2} = 0,$$

since the denominator is a (theoretical) sufficient statistic for σ^2 (n = no. of degrees of freedom). If we maximise the likelihood from the t distribution, or from the equivalent distribution of

$$r = (x-m)/\{ns^2 + (x-m)^2\}^{1/2},$$

then the equation

$$\Sigma\,\frac{\partial L}{\partial m} = 0 \quad (L = \log \text{likelihood})$$

is in fact equivalent to the equation above. The only remaining problem is to evaluate the information in the estimate provided by this equation when x and s are allowed to vary at random. There is a snag here in that $\Sigma(\partial L/\partial m)$ does not go to normality. It does go to normality, however, for fixed $ns^2 + (x - m)^2$, so that using the r distribution, for which

$$\Sigma\,\frac{\partial L}{\partial m} \equiv \Sigma\,\frac{(n-1)(x-m)}{ns^2 + (x-m)^2},$$

we find:

$$E_r\left(-\frac{\partial^2 L}{\partial m^2}\right) = E_r\left(\frac{\partial L}{\partial m}\right)^2 = \frac{(n-1)^2}{(n+1)\{ns^2 + (x-m)^2\}}$$

of which the average value when we allow $ns^2 + (x - m)^2$ to vary is

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

Fisher to M.S. Bartlett: 14 *January* 1936

Thanks for your letter. I still do not see how you get your formula, unless there has been some confusion of degrees of freedom with numbers in sample. I have put down on the attached sheet[1] a discussion of your problem of a number of different samples supposed to be drawn from populations having the same mean, but, possibly, different variances. The samples are supposed to be of different numbers. I get an equation of estimation rather like yours, with a factor $n + 1$ in each term, but the amount of information in the estimate derived from it seems to be given by my formula.

[1] Not in Fisher's file.

M.S. Bartlett to Fisher: 15 *January* 1936

The two points where we differ are as follows:—

(1) I had the equation

$$\Sigma\,\frac{(n-1)(x-m)}{ns^2 + (x-m)^2} = 0$$

whereas you had

$$\Sigma \frac{(n+1)(x-m)}{ns^2+(x-m)^2} = 0.$$

You will agree, however, that if we confine ourselves to samples with the same number of degrees of freedom, the estimate from either equation should contain the same amount of information.

(2) When considering the average value of $(\partial L/\partial m)^2$, I averaged first for t (or $r \propto t/\sqrt{\{1 + t^2/n\}}$) and then for $ns^2 + (x - m)^2$, *which is independent of t*. Your averaging proceeds on the lines indicated in your book, and it is this that I have been doubtful about,—namely, the averaging for t, ignoring s^2 (which is not independent of t), and leaving it in the formula. I do not follow the interpretation of this. Thus it surely cannot represent the information available for any actual problem, such as the one I proposed, since if we considered samples with one degree of freedom, our average information, if the true variance were in fact constant, would appear infinite.

Fisher to M.S. Bartlett: 17 *January* 1936

I am sorry you do not discuss the origin of your equation of estimation, as this might throw light on what you are aiming at. The whole point of my procedure, as I think must be clear in my book, is to retain s^2 as the sole available fact about the precision of the average. I should say it must certainly represent the information available for the actual problem I have in view, namely, one in which s^2 supplies the only available knowledge about the variance. From this point of view it is not appropriate to average the value of $1/s^2$ for different samples from the same population, but to average the value of $1/\sigma^2$ for the different populations leading to the observed values.

M.S. Bartlett to Fisher: 20 *January* 1936

Many thanks for your letter. I'm afraid, however, that I am still not clear on the interpretation of your procedure. It is true that if the fiducial distribution of σ^2, when s^2 is given, were taken, the average value of $1/\sigma^2$ would be $1/s^2$ but I do not see that the fiducial distribution can be used in this way.

It is also true that I considered for simplicity, in order to get a definite answer, a constant value of σ^2. But before averaging by means of any distribution depending on σ^2, I had the formula

$$I = \frac{(n-1)^2}{(n+1)\{ns^2+(x-m)^2\}}$$

and if σ^2 were not constant, our average information would still be the average value of this. But I do not see why we should not be able to find out what information we should have used if σ^2 were constant after all.

I hope to be up on Thursday for your lecture. Perhaps if you care to make any further comment, you would let me know then. I must apologise for this correspondence becoming so lengthy.

M.S. Bartlett to Fisher: 16 *April* 1937

Cochran has mentioned to me that you do not agree with the discussion I gave at the top of p. 565 of my last *Proc. Camb. Phil. Soc.* paper (Part 4, 1936)[1] in connection with testing differences in means. I refer there to the special (and practically trivial) case where there is only one degree of freedom each for the estimates of the unknown and unequal variances of the two means.

I have not detected any error in my remarks, but I should be glad to examine them again in the light of any comments you care to make.

[1] Bartlett, M.S. (1936). The information available in small samples. *Proc. Camb. Phil. Soc.* **32**, 560–6.

Fisher to M.S. Bartlett: 19 *April* 1937

I have written a note on the discrepancy between your results and those obtained by the fiducial argument, which will appear in a forthcoming number of *Annals* [*CP* 151]. I think the cause of the discrepancy lies in the introduction of fixed values for σ_1 and σ_2 into an argument in which the distribution of these values has already been determined.

M.S. Bartlett to Fisher: 14 *July* 1937

As I was in town yesterday afternoon, I called in at the Galton Lab. on the chance of being able to see you for a few minutes, but found you away. I thought perhaps we might have been able to resolve our apparent difference of opinion on the Behrens' test business, on which you comment in a note in the last number of *Annals of Eugenics*.

With regard to the particular point at issue there (the one degree of freedom case), your representation of my view did not appear to me to be altogether fair. You refer, for example, to the distribution of 'my ratios', whereas the composite statistic ($T\ T'$) (your notation) essentially is not to be split up into T and T'. Granting this, you appear to recognize in your concluding paragraphs that the test would do what I claimed for it, but thought that no experimenter would be likely to use it. With this last remark I entirely agree, but I had no intention of recommending any test based on one degree of freedom—it would be a rotten test anyway!

In the practical example I considered at the end of my own paper (*Proc. Camb. Phil. Soc.*), I used an inequality statement as being a convenient test in the circumstances. Practically, the curious feature of the Behrens' test for small numbers of degrees of freedom is the tendency to get a less significant result for s/s' finite than where it is 0 or ∞. Here Behrens' original table seems a little misleading, for one would infer from it that this was always the case, whereas on examination the significance level for small numbers greater than one appears to change over in comparison with the ordinary t test, (assuming $\phi = \sigma_1^2/\sigma_2^2 = 0$), as the probability value of this level ranges from 1/2 to 0. Whether or not any exact probability interpretation of the Behrens' test can be made, these results appear rather anomalous. . . .

Fisher to M.S. Bartlett: 15 *July* 1937

I am sorry you thought my reference to your paper unfair, for I doubt if you will ever receive fairer treatment from those who differ from you on mathematical points. You say that I appear to recognise, in my concluding paragraph, that the test has done what you claimed for it. Since for paired data this is 'Student's' test of 1908, I imagine there could be no doubt on the subject. The difficulty is that, for data not paired, different possible samples are placed in the wrong order in respect of significance, and therefore, though the distribution is correct, as a test of sigificance it is not available except in the case to which 'Student' applies it. This is quite a different criticism from saying that it is based on one degree of freedom only.

The point has, I think, only this importance, that you adduce your test as disagreeing with Behrens' and as showing, in consequence, that Behrens' test must be wrong.

I do not follow the statement in your last paragraph, e.g. that Behrens' test has a tendency to get a less significant result for s/s' finite than where it is zero or infinity. There must be some meaning behind this, but I do not know what it is. Like any other test, Behrens' accepts certain bodies of data as significant, and distinguishes them from other bodies which, in this test, are non-significant. Do you mean that for a given difference in the means significant results are got less frequently when the ratio s/s' is near to unity than when it is larger or smaller? If so, I cannot see its bearing as a criticism of the test. . . .

M.S. Bartlett to Fisher: 21 *July* 1937

. . . By the statement I made in my letter (3rd paragraph) about the Behrens' test, I meant the following. The statistic $z = (x_1 - x_2)/(\sqrt{s_1^2 + s_2^2})$ (equal numbers of degrees of freedom assumed for simplicity) is used either in the ordinary t-test or in the Behrens' test. In the latter it is considered in conjunction with the value of s_1^2/s_2^2. If the latter value is ignored, the distribution of the statistic z will depend on the unknown ϕ ($\equiv \sigma_1^2/\sigma_2^2$), but we know that, whatever ϕ is, we shall err *on the side of caution* if we consider it zero and consider z as a t with the same number of degrees of freedom as s_1^2 (or s_2^2). On the other hand, if we use Behrens' test and brought in s_1^2/s_2^2, we might (for small numbers of degrees of freedom) consider z as even *less* significant than we thought it when we knew we were already erring on the cautious side. I think I should prefer the 'inequality test'. For the case of $n_1 = n_2 = 1$, the significance level given by Behrens' test is P_1 where

$$P_1(s_1^2/s_2^2) < P(\phi = 0), \quad \text{whatever } s_1^2/s_2^2,$$

($P(\phi = 0)$ being the ordinary t test level for $n = 1$); and the significance level given by the 'inequality test',

$$P_2(\phi) \geqq P(\phi = 0), \quad \text{whatever } \phi.$$

I don't think I am altogether clear yet about your objections to my 'randomisation test', but the above comments appear to me also to require some consideration.

Fisher to M.S. Bartlett: 22 *July* 1937

If we are to get on any further I am afraid you will have to tell me why we know that 'whatever ϕ is, we shall err on the side of caution if we consider it zero and consider z as a t with the same number of degrees of freedom as s_1^2 or s_2^2'. You must be suggesting some objective procedure, but I do not know what it is. We cannot, without ignoring some of the information supplied by the sample, consider σ_1 to be zero if s_1 is not zero.

In the same paragraph I do not know what *the inequality test* means.

M.S. Bartlett to Fisher: 23 *July* 1937

I apologise for being obscure, but since I was referring to a test whose application I had illustrated in my *Proc. Camb. Phil. Soc.* paper (final paragraph) I may have neglected to make myself clear about it in my letter. Perhaps if I enclose a formal mathematical statement of what I meant by the test (which I called for convenience the 'inequality test'), it will make my arguments clearer. The proof[1] I give in this enclosure of the 'inequality' is of course merely a verification that the test is mathematically correct, and is perhaps hardly necessary.

I agree with you that we are ignoring the information supplied by s_1^2/s_2^2, but that does not seem to me to imply that the possible value or relevance of the test is nil.

[1] This proof, not reproduced here, was subsequently included in §7 of Bartlett's 1939 paper, Complete simultaneous fiducial distributions. *Ann. Math. Stat.* **10**, 129–38.

W.U. Behrens to Fisher: 13 *August* 1929

Ich interessiere mich für die Anwendung der Wahrscheinlichkeitslehre auf landwirtschaftliche Probleme. Auf der Tagung der Internationalen Bodenkundlichen Gesellschaft hatte ich Gelegenheit, mit ausländischen Herren hierüber zu diskutieren und ich wurde immer wieder auf Ihr Werk: *Statistical Methods for Research Workers* aufmerksam gemacht. Ich habe es mir daher sofort gekauft und lese es jetzt mit grossem Interesse. Ich bedaure ausserordentlich, dass ich es nicht früher gekannt habe, ich hätte mir sonst viel Arbeit ersparen können. Aus meiner Arbeit:[1] 'Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen', von der ich Ihnen gleichzeitig ein Separatum schicke, werden Sie ersehen, dass ich zum Teil zu Ergebnissen gekommen bin, wie sie schon längst den englischen Statistikern bekannt sind. Ich hatte die Arbeit vor der Drucklegung an einen der bedeutensten deutschen Wahrscheinlichkeitstheoretiker geschickt und angefragt, ob seines Wissens bereits Arbeiten mit ähnlicher Problemstellung vorliegen. Er antwortete mir damals, dass ihm keine derartigen Arbeiten bekannt seien. Ich werde selbstverständlich bei der nächsten sich bietenden Gelegenheit darauf aufmerksam machen, dass ich *nicht* die Priorität habe.

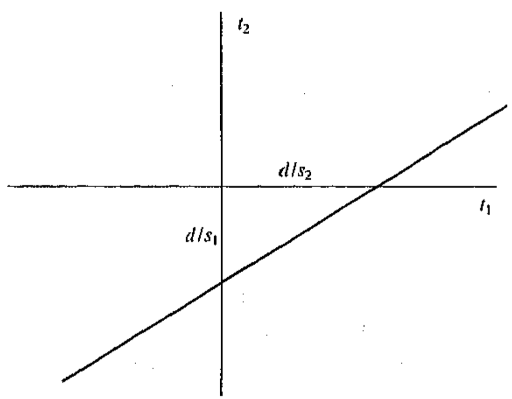
Auf Seite 820–822 meiner Arbeit komme ich zu anderen Ergebnissen als die englischen Statistiker resp. Sie. Ich wäre Ihnen sehr dankbar, wenn Sie hierzu Stellung nehmen und mir Ihre Ansicht mitteilen könnten.

Grüssen Sie bitte Herrn Dr. Crowther von mir.

[1] *Landw. Jbr.* **68**, 807–37 (1929).

Fisher to W. U. Behrens: 5 *September* 1929

I received your letter with very great interest during my holiday, and since my return have studied the reprint of your paper. It is a great pleasure to see that you have arrived independently at 'Student's' distribution, the importance of which I have now for many years been endeavouring to make clear to the English speaking statisticians. Your attack upon the more complex problem of comparing the means of samples having different variances seems to me wholly original.



The integration of the simultaneous distributions of two values t_1 and t_2, dependent from [? upon] two different samples of sizes n_1 and n_2, on one side of the line, making intercepts d/s_1 and d/s_2 respectively, would be of very great interest. Unfortunately this probability depends from [? upon] four parameters, n_1 and n_2, in addition to d/s_1 and d/s_2 or as you have it $d/\sqrt{s_1^2+s_2^2}$ and s_1/s_2. The numerical values you give are confined, I believe, to the case $n_1 = n_2$; the more general case is, however, of considerable interest for experimentalists. The tabulation would be simplified and the results expressed in a form convenient for tests of significance if one could tabulate the 5 per cent and 1 per cent values of $d/\sqrt{s_1^2+s_2^2}$ for different values of n_1, n_2 and s_1/s_2, replacing the last modulus by its arc tan for purposes of interpolation. Probably 10 values at intervals of 10° would then be sufficient, while for n_1 and n_2 I should utilise the harmonic series 6, 8, 12, 24, ∞, which works excellently for 'Student's' distribution.

I am glad that you are calling attention in Germany to recent advances in the theory of errors, which besides their intrinsic importance are having an increasing influence upon the design of biological and agricultural experi ments.

I enclose two recent papers which I hope may be of interest to you.

W. U. Behrens to Fisher: 10 *September* 1929

Verbindlichsten Dank für Ihren werten Brief vom 5.9. und Ihre interessanten Separata. Aus Ihrem Brief ersehe ich, dass Sie es für nützlich halten, grössere Tabellen für die 'tests of significance' von Beobachtungsdifferenzen aus zurechnen. Ihr Plan findet meinen Beifall, ich möchte vorschlagen, auch die 2 per cent values von $d/\sqrt{s_1^2+s_2^2}$ zu berechnen. Sollten Sie auf meine Mitarbeit Wert legen, so bin ich gern bereit, mich an einer gemeinsamen Arbeit zu beteiligen, und ich erwarte dann von Ihnen nähere Vorschläge. Allerdings bin ich jetzt durch andere Arbeiten in Anspruch genommen, aber von Dezember dieses Jahres bis Marz 1930 hoffe ich mehr Zeit zur Verfügung zu haben. Die Ergebnisse könnten wir dann gemeinsam publizieren, Sie könnten vielleicht einen Aufsatz für eine englische Zeitschrift schreiben, ich würde den Text für eine deutsche Publikation in den Landwirtschaftlichen Jahrbüchern schreiben.

Fisher to W. U. Behrens: 9 *February* 1957

I am sending you herewith a recent offprint [*CP* 265] on the test of significance which bears your name, at least in English-speaking countries.

I do not know if it has ever come to your knowledge that a number of attacks have been made on the validity of this test by J. Neyman, and others influenced by him, such as Pearson and Bartlett; indeed, tables of a rival test have been published—Table 11 in the *Biometrika Tables* of Pearson and Hartley. I have recently criticized these tables, which are, in my opinion, grossly inaccurate, and make the observed difference significant on much too easy terms.

The fact is that the problem which you first discussed, I think about 1929, is a good test case for the validity of the rather academic exposition of tests of significance put forward by Neyman and Pearson more than twenty years ago, of which the key thought is the frequency with which a statement would be found to be correct in 'repeated samples from the same population', which is in fact very far from being a measure of the strength of the evidence provided by the data against some specific and well defined hypothesis or group of hypotheses.

I hope, at least, you find these two offprints [*CP* 264, 265] interesting.

I have been wondering if we shall see you at the Stockholm meeting of the International Statistical Institute. I feel sure an invitation could be obtained if this would enable you to get any necessary financial assistance.

W.U. Behrens to Fisher: 22 *September* 1957

I beg your pardon that I have not answered you before today. I had planned to come to Stockholm, but I had not the time. The money was not missing, but the time. My occupation is the agricultural chemistry in the fertilizer industry, the biomathematics are my hobby. So I have not the time for study of the international literature as it is necessary, and I thank you very much for your information about the so-called Behrens–Fisher test. . . .

Fisher to C.I. Bliss: 4 *October* 1956

. . . I shall be much interested to learn how you, and others among your countrymen, bear with my efforts [in *SMSI*] to draw the necessary logical distinctions, and to use words accurately, in a subject which has been so deeply entangled and knotted together as the theory of probability had become early in the century, and the theory of testing hypotheses since about 1930.

I believe, now, I should have stressed early and loudly that Keynes was mistaken in defining probability as 'the measure of rational belief', in that whereas the phrase 'measure of rational belief' was a penetrating one and needed in the subject, yet it is not the classical and mathematically defined concept of probability that fills this bill.[1] This does appear, indeed, gradually and by stages in the course of Chapters II and III, but perhaps you could give me an opinion as to whether the reader might like to have fair warning of so

large a semantic change. The effort to find probability statements appropriate to every case in which belief, or disbelief, in some degree can be supported on good and communicable reasons, is, of course, the main cause of the numerically erroneous values arrived at on the basis of Pearson and Neyman's theory of testing hypotheses, of which I give an example in the current *J.R.S.S.*, Series B [*CP* 264], which I hope you may see before I have an offprint to send you. . . .

[1] A review by Fisher of J.M. Keynes's book *A treatise on probability* (1921) was published in *Eugen. Rev.* **14**, 46–50, (1923).

O.K. Buros to Fisher: 16 *February* 1959

In your *Statistical Methods for Research Workers*, you describe biased intraclass correlation coefficients, methods for correcting for the bias, and unbiased intraclass coefficients obtained by means of analysis of variance. Have you written more extensively on these topics elsewhere?

An article, 'Unbiased estimation of certain correlation coefficients' in the March 1958 issue of *Annals of Mathematical Statistics* seems (I am unable to follow the proof) to indicate that your unbiased estimates of the intraclass correlation coefficient are biased. Would you be willing to comment on this? . . .

O.K. Buros to Fisher: 17 *February* 1959

Since writing you yesterday, I have received the following statement from a friend:

'I raised the question (regarding what has been described as Fisher's unbiased estimate of the population intraclass correlation coefficient) with Professor S.S. Wilks. He has written down a proof that no ratio of linear functions of 'among' and 'between' sums of squares can be an unbiased estimate of the population intraclass correlation. Of course the numerator and denominator taken separately are unbiased estimates of important population parameters, but this does not make the ratio itself an unbiased estimate.'

I have always talked about biased and unbiased estimates of intraclass correlations, thinking that I was interpreting you correctly. Now that Professor Wilks states that such a ratio cannot be an unbiased estimate, I should like to learn what the facts are from you. . . .

Fisher to O.K. Buros: 23 *February* 1959

I have your letters of February 16th and 17th on a point which has needed clarification in statistical teaching for a long while, but as a good many points can be introduced it would be too long to discuss in a single letter. In my recent book *Statistical Methods and Scientific Inference* I say something about the use of 'unbiased' as a criterion in estimation (p. 140)[1] because as used by Neyman, and perhaps other teachers in Berkeley, it seems to have given a great deal of complexity to otherwise simple problems.

Perhaps the simplest interpretation of Professor Wilks's recent demonstration is that he must be using the word 'biased' in a sense which is not usefully applicable by practical statisticians, e.g. the sense in which the estimated standard deviation is biased, although found by taking the square root of an unbiased estimate of the variance.

In Section 2 of the chapter to which I have referred, there is developed with small sample exactitude the notion of Consistent estimation, which fulfils the common sense requirement of unbiasedness by stipulating that the estimate shall be a function of the observed frequencies of such a kind that if expected frequencies are substituted for those observed, the estimate shall be exactly equal to the true parametric value. This criterion, which I put forward I think about as early as 1922 (Mathematical foundations of theoretical statistics: *Phil. Trans.* A, vol. 222) [*CP* 18], is invariant for transformations of the parameter, such as squaring, and so avoids the difficulty which formal bias, as defined by Wilks and others, falls into.

I think the only practical importance of what has become by now a rather intricate subject lies in the avoidance of bias in the methods of sample survey, where it is often desired to total or average numerous small sample estimates. Of course you know Dr. Frank Yates's book *Sampling Methods for Censuses and Surveys* (1949) in which he gives to this point all the care it deserves. If you do refer to my work, however, it might be better, in regard to the current use of language, to speak of the estimates I recommend rather as 'Consistent' than as 'unbiased'. This, of course, is just as strict a criterion but one more suitable for the situation of statistical estimation.

[1] *SMSI*, p. 146.

N. Campbell to Fisher: 18 *July* 1922

On my return from a holiday I have received the reprint of your paper [*CP* 18], for which I am greatly obliged. Just before leaving I had read your paper with the very greatest interest, for it appeared to me a most welcome reversion to realities after the wholly abstract and impracticable dissertations to which we had become accustomed from mathematicians.

As you say, I think our ideas, insofar as they cover the same ground, are essentially similar. But there are differences. The difference in nomenclature is nothing. I prefer to get away from the connotations of the term probability, but even if I used that term *my* probability ('chance') would not be the same as that which you define or describe on p. 312. My chance is something essentially experimental and 'an infinite number of throws' means nothing to an experimenter—you might as well talk of a round square. Moreover randomness is essential to my chance, whereas, so far as I can see, your die would have a 'probability' if it turned up six regularly at the 1st, 7th 13th, . . . throw. But as a matter of fact, the *use* you make of your term seems to me to make it approximate much more nearly to mine than the definition suggests. If a more 'technical' word than chance is thought desirable for my conception, I should suggest 'statistical frequency'—but most European languages have a word exactly corresponding to our 'chance'.

For the rest the problem of statistics seems to me that of identifying a system by means of a measurement of its chances. And that is a problem which (as I believe you recognise) mathematics alone can never solve. Whatever degree of agreement is found between the measured chances and those calculated for some system with which it is proposed to identify that examined experimentally, there is always room and need for the characteristically scientific decision whether, taking everything into account, there is sufficient evidence to permit the identification to be made. That kind of judgement is distinctive of all physical work and by no means confined to statistical study. Still there is room for the mathematician (and it is here that I find your conception of a statistic of maximum efficiency so helpful) in selecting for measurement those chances which provide the best basis for comparison. I make no pretence to have mastered your methods completely and indeed it is probably impossible for one so little practised in statistics to do so, but I am thoroughly convinced that you are on the right lines. The only other method which seems to offer a plausible alternative is some kind of systematic search for regularity in 'residuals'. The less regularity such search discloses, the more weighty the evidence that the system under investigation really has a chance of the kind and of the magnitude which is proposed.

But whether you consider my remarks on such matters sensible or no, pray accept my best thanks and warmest congratulations.

N. Campbell to Fisher: 21 *July* 1922

I hope you will not think that because you were good enough to write to me I am going to drag you into an endless controversy.[1] But I do want to make my position clear; for I am rather alarmed that so sympathetic a critic should seem to miss the essential feature of my contention.

You say that I nominally reduce chance to a synonym for frequency. I protest violently. I maintain on the contrary that the bare knowledge that an event has occurred 15 times out of 100 trials tells me nothing whatsoever about its chance. For that knowledge does not tell me that the event has a chance at all. Even if I know beforehand that the event has a chance, it gives me but the vaguest knowledge what that chance is. It tells me perhaps that the chance is *not* 1 in 1000 and is not 999 in 1000, but it tells me little more. In order to estimate chance I must have a whole series of frequencies, and the chance which I estimate will be determined by the whole of the series and not by a single member of it.

I think it is very useful to have before one's eyes a concrete example. Accordingly on the enclosed diagram[2] I have shown how I should determine the chance that the last figure of a seven-figure logarithm is either 3, 6, or 9. The ordinates give the number of such last figures in the first n logarithms, where n is the abscissa. After I have plotted 250 points I ask myself whether I am justified in drawing a straight line through the points. If I knew nothing about the events, I should be doubtful in this instance, for the earlier points tend to lie above the mean line, the later below it. I should have to take many more points (or plot the diagram again from a different page of the book) to be sure whether this regularity was real. But if I assume that there is a chance, it is pretty closely defined. I have drawn the 'theoretical' line; and it is clear that any line I could reasonably draw would not differ as much as 10% in slope from it. But that slope, I want you to observe, is determined by all the frequencies taken together; it is not determined by any one of them; in fact at only four points (marked by arrows) do the frequencies coincide with the line that determines the chance.

I suggest that what you mean by the 'true' frequency, or the frequency of a hypothetical infinite population, is simply the frequency corresponding to points on the line. And of course this true frequency can only agree with the actual frequency (in this instance) when the number of trials is a multiple of 10. What you are essentially doing in interpreting statistics is finding a theoretically determined line (such as I have drawn on the diagram) which will give an adequate representation of the scattered actual points. But I won't go further into that. I only want you to be clear what *I* mean, not to discuss what *you* mean.

[1] There is no copy in Fisher's files of his reply to Campbell's letter of 18 July 1922. Fisher's letter of 22 July 1922 to Campbell is the earliest of all his letters included in this volume; the original copy, in Fisher's hand, is reproduced here.
[2] Not shown here.

Fisher to N. Campbell: 22 July 1922

Many thanks for your letter. I do not think there is any danger of an endless controversy, especially as I imagine we are both concerned to make sense of a

Telegrams: LABORATORY, HARPENDEN.
Telephone: 21 HARPENDEN.

Railway Station:
HARPENDEN (MIDLAND RAILWAY).

Lawes Agricultural Trust.

Director:
E. J. RUSSELL, D.Sc., F.R.S.

R. A. FISHER, M.A., *Chief Statistician.*
WINIFRED A. MACKENZIE, B.Sc. (ECON).

Rothamsted Experimental Station,
Harpenden.

July 22 1922

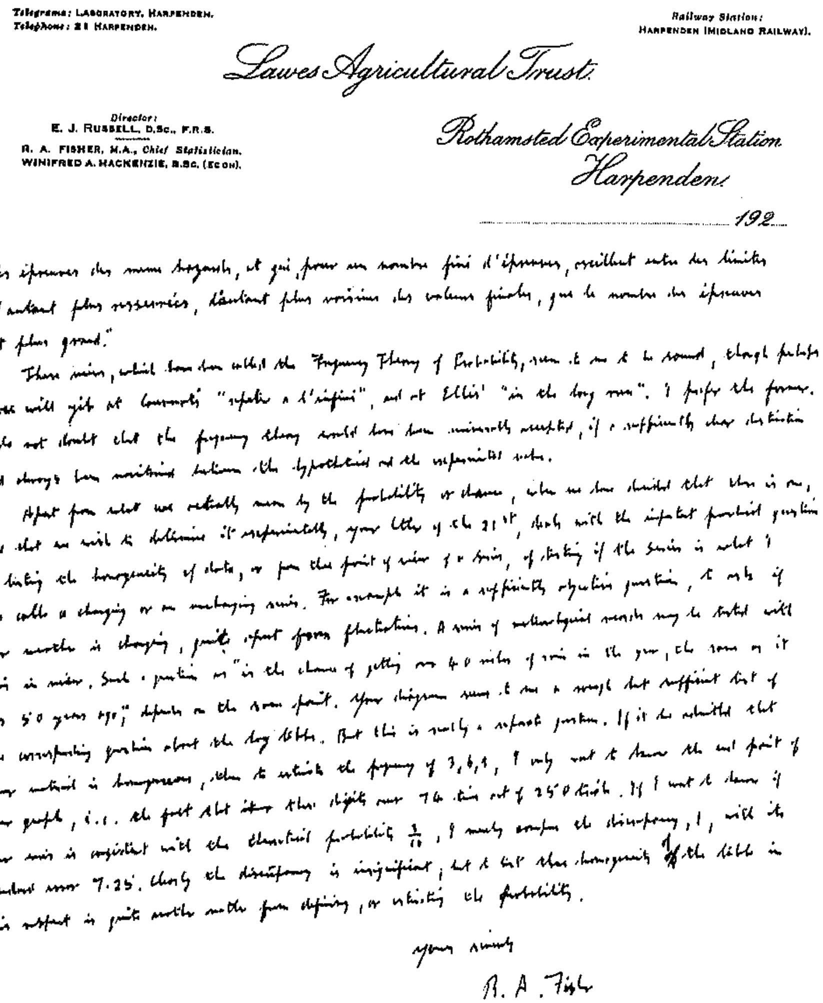
Dear Dr Campbell,

Many thanks for your letter. I do not think there is any danger of an endless controversy, especially as I imagine we are both concerned to make sense of a rather intricate subject, about which a good many false assumptions have been widely accepted. Your suggestion, in your letter of the 18th, that apart from the difference in nomenclature, your "chance" and my "probability" were intended to designate somewhat different ideas, interested me; the main point, as you say, was for me to get a clear idea of what you meant. I had judged from your remarks in the Phil. Mag., that as the chance of an event occurring would mean a physical property of the material system concerned, that we did mean exactly the same thing.

In your letter of the 18th the distinction you draw (apart from the point about randomness in which I fully agree with you) was that "My chance is something essentially experimental and 'an infinite number of throws' means nothing to an experimenter", and again "If a more 'technical' word than chance is thought desirable for my conception, I should suggest 'statistical frequency'". This, of course, would make your chance quite distinct from any notion of probability, but I did not think, after what you had written in the Phil. Mag. that you really wished to identify chance with frequency, and I am glad that you "protest violently" against the suggestion.

A certain number of writers on Probability have tried to define the notion on an objective basis. Thus Leslie Ellis (1843). "If the probability of ~~an event~~ a given event be correctly determined, the event will on a long run of trials tend to recur with frequency proportional to their probability. This is generally proved mathematically. It seems to me to be true *a priori*", and again "I have been unable to sever the judgment that one event is more likely to happen than another, from the belief that in the long run it will occur more frequently."

About the same time Cournot wrote "La théorie des probabilités a pour objet certains rapports numériques qui prendraient des valeurs fixes et complètement déterminées, [illegible] si l'on pouvait répéter à l'infini



Telegrams: Laboratory, Harpenden.
Telephone: 21 Harpenden.

Railway Station:
Harpenden (Midland Railway).

Lawes Agricultural Trust.

Director:
E. J. Russell, D.Sc., F.R.S.

R. A. Fisher, M.A., Chief Statistician.
Winifred A. Mackenzie, B.Sc. (Econ).

Rothamsted Experimental Station
Harpenden.

192

les épreuves des mêmes hasards, et qui, pour un nombre fini d'épreuves, oscillent entre des limites d'autant plus resserrées, d'autant plus voisines des valeurs finales, que le nombre des épreuves est plus grand."

These views, which have been called the Frequency Theory of Probability, seem to me to be sound, though perhaps you will jib at Cournot's "répéter à l'infini", and at Ellis' "in the long run". I prefer the former. I do not doubt that the frequency theory would have been universally accepted, if a sufficiently clear distinction had always been maintained between the hypothetical and the experimental value.

Apart from what we actually mean by the probability or chance, when we have decided that there is one, and that we wish to determine it experimentally, your letter of the 21st, deals with the important practical question of testing the homogeneity of data, or from the point of view of a series, of testing if the series is what I call a changing or an unchanging series. For example it is a sufficiently objective question, to ask if our weather is changing, quite apart from fluctuations. A series of meteorological records may be tested with this in view. Such a question as "is the chance of getting over 40 inches of rain in the year, the same as it was 50 years ago?" depends on the same point. Your diagram seems to me a simple but sufficient test of the corresponding question about the [illegible] tables. But this is really a separate question. [illegible]

Yours sincerely

R. A. Fisher

rather intricate subject, about which a good many false assumptions have been widely accepted. Your suggestion, in your letter of the 18th, that apart from the difference in nomenclature, your 'chance' and my 'probability' were intended to designate somewhat different ideas, interested me; the main point, as you say, was for me to get a clear idea of what you meant. I had judged from your remark in the *Phil. Mag.*,[1] that the chance of an event occurring was a physical property of the material system concerned, that we did mean exactly the same thing.

In your letter of the 18th the distinction you draw (apart from the point about randomness in which I fully agree with you) was that 'My chance is something essentially experimental and "an infinite number of throws" means nothing to an experimenter', and again, 'If a more "technical" word than chance is thought desirable for my conception, I should suggest "statistical frequency".' This, of course, would make your chance quite distinct from my notion of probability, but I did not think, after what you had written in the *Phil. Mag.*, that you really wished to identify chance with frequency, and I am glad that you 'protest violently' against the suggestion.

A certain number of writers on Probability have tried to define the notion on an objective basis. Thus Leslie Ellis (1843);[2] 'If the probability of a given event be correctly determined, the event will on a long run of trials tend to recur with frequency proportional to their probability. This is generally proved mathematically. It seems to me to be true *a priori*', and again, 'I have been unable to sever the judgement that one event is more likely to happen than another from the belief that in the long run it will occur more frequently.'

About the same time Cournot wrote[3] 'La théorie des probabilités a pour objet certains rapports numériques qui prendraient des valeurs fixes et complétement déterminées, si l'on pouvait répéter à l'infini les épreuves des mêmes hasards, et qui, pour un nombre fini d'épreuves, oscillent entre des limites d'autant plus resserrées, d'autant plus voisines des valeurs *finales*, que le nombre des épreuves est plus grand.'

These views, which have been called the Frequency Theory of Probability, seem to me to be sound, though perhaps you will jib at Cournot's 'répéter à l'infini' and at Ellis' 'in the long run'. I prefer the former. I do not doubt that the frequency theory would have been universally accepted if a sufficiently clear distinction had always been maintained between the hypothetical and the experimental value.

Apart from what we actually mean by the probability or chance, when we have decided that there is one, and that we wish to determine it experimentally, your letter of the 21st deals with the important practical question of testing the homogeneity of data, or from the point of view of a series, of testing if the series is what I call a changing or an unchanging series. For example, it is a sufficiently objective question to ask if our weather is changing, quite apart from fluctuations. A series of meteorological records may be tested with this in view. Such a question as 'Is the chance of getting over 40 inches of rain in the year the same as it was 50 years ago?' depends on the same point. Your diagram seems to me a rough but sufficient test of the corresponding question about the log tables. But this is really a separate question. If it be admitted that your material is homogeneous, then to estimate the frequency of 3,6,9, I only want to know the end point of your graph, i.e. the fact that these digits occur 74 times out of 250 trials. If I want to know if your series is consistent with the theoretical probability 3/10, I merely compare the discrepancy, 1, with its standard error 7.25. Clearly the discrepancy is insignificant, but to test the homogeneity of the table in this respect is quite another matter from defining, or estimating, the probability.

[1] Campbell, N. (1922). The measurement of chance. *Phil. Mag.* **44**, 67–79.
[2] Ellis, R.L. (1843). On the foundations of the theory of probability. *Proc. Camb. Phil. Soc.* **8**, 1–6.
[3] Cournot, A. (1843). *Exposition de la théorie des chances et des probabilités*. Paris.

[The following paragraph seems worthy of inclusion here. It is the final part of a 'Note on Dr. Norman Campbell's alternative to the method of least squares', written by Fisher probably about 1925. See Campbell, N. (1924). The adjustment of observations. *Phil. Mag.* **47**, 816–26.]

Finally perhaps a word may be said with an eye to the underlying motive of Dr. Campbell's innovation. He is evidently oppressed by the arithmetical labour involved in the recognised processes of reduction. These processes, like commercial book-keeping, do require some arithmetical labour; and would constitute an intolerable burden on a man whose mind was full of other and more important business. The commercial man gets over this difficulty by hiring such computational assistance as he needs; the alternative of inventing for himself new book-keeping methods, so quick and simple that he could do it all himself during train journeys, would not, I suggest, make a strong appeal. Need the organisation of scientific research be so much behind that of an ordinary small business establishment? In my experience, much of the troublesome arithmetic with which research workers, physicists and biologists, harass themselves, could be done *better* by a boy of 16 of average intelligence, working in a properly organised statistical laboratory. The constant crop of inefficient methods, of which Dr. Campbell's is by no means the worst, which are continually being produced by laboratory workers, who have some but too little experience of statistical problems, is an index of the amount of highly skilled labour which is being wasted in these ways, through lack, it would seem, of organised computational assistance.

H.E. Daniels to Fisher: 29 *January* 1938

Some time ago I wrote to you in regard to the question of intercorrelated observations and the z test, and you were kind enough to send me a most

illuminating reply.[1] The following point is now troubling me, and I should be grateful for your advice.

Consider e.g. two samples of size n_1, n_2 respectively and we wish to test whether the means differ significantly. Suppose that the population variances are not identical, but are σ_1^2, σ_2^2. The expected value of

$$\{\Sigma_r(x_{1r}-\bar{x}_1)^2+\Sigma_r(x_{2r}-\bar{x}_2)^2\}/(n_1+n_2-2)$$

is

$$\{(n_1-1)\sigma_1^2+(n_2-1)\sigma_2^2\}/(n_1+n_2-2)=\nu_1$$

and that of

$$n_1n_2(\bar{x}_1-\bar{x}_2)^2/(n_1+n_2) \text{ is } (n_2\sigma_1^2+n_1\sigma_2^2)/(n_1+n_2)=\nu_2$$

so that

$$\nu_2-\nu_1=(\sigma_1^2-\sigma_2^2)\,(n_1+n_2-1)\,(n_2-n_1)/\{(n_1+n_2-2)\,(n_1+n_2)\}.$$

When

$$\sigma_2>\sigma_1 \text{ [and] } n_2>n_1, \text{ [then] } \nu_2 \text{ is less than } \nu_1.$$

In case[2]

$$\begin{aligned}\sigma_1^2 &= 4,\ \sigma_2^2=24,\ n_1=5,\ n_2=25,\\ \nu_1 &= 21,\ \nu_2\cong 7\end{aligned}$$

and while the expected value of t^2 will not quite equal ν_2/ν_1 it seems very unlikely that it should exceed unity. If correct, this would appear to contradict your remark[3] (*Statistical Methods* p. 121) that a real difference in variance would always enhance t. One might argue that only those cases which do enhance t are important since when significance is not obtained, no positive conclusion is drawn. But the test would seem to suffer as a test of the null hypothesis that the samples arose from the same normal population.

[1] See p. 253.
[2] Daniels had written $n_1 = 25$, $n_2 = 5$, $\nu_1 \simeq 82$ and $\nu_2 = 21$, but the values shown here are evidently those intended.
[3] '. . . a difference in variance between the populations from which the samples are drawn will tend somewhat to enhance the value of t obtained.' (*SMRW*, 5th edition, 1934, Section 24.1, p. 121). However, with the 6th edition in 1936, Fisher replaced 'somewhat' in the above sentence by 'sometimes'. See *SMRW*, p. 124.

Fisher to H.E. Daniels: 18 *February* 1938

Your letter of January 29th was given me on my return, and I agree with it, except that I do not think I ever said that a real difference in variances would always enhance t. Perhaps you will let me know whether it was I or some one else who gave you this impression. I think Stevens mentioned that the possibility that the populations sampled differed in their variances does not affect the exactitude of the test of the null hypothesis that they are drawn from the same population. Such a case as you mention, supposing t to be non-significant on 'Student's' test, would seem to be nicely treated by the supplementary z which I suggest for the differences between the variances, for this would show that the variances were very significantly different, although on the data it could not be said that the means were different.

Now in practice this would advance the investigation as much as in the standard case; for, supposing the test were made between two varieties of the plant, the fact of a real difference in the variances shows that in some circumstances one variety is the better and in other circumstances that it is the worse. The situation is thus proved to be more complicated than perhaps the experimenter originally thought. He will see now that it is useless to compare the means unless he has some specification of the circumstances, or range of circumstances, in which the test is to be made. His preliminary enquiry 'Are the samples from the same population?' is answered definitely in the negative and, this being so, it will depend entirely on the circumstances of the case whether any comparison of the means is desired at all.

The point has, I think, received the rather large amount of theoretical attention that it has chiefly through lack of contact with the practical experimental situation. Some years ago Behrens published a test of significance appropriate to the rather academic question 'Might these samples have been drawn from different normal populations having the same mean?', and more recently Sukhatme has been preparing tables needed for Behrens' test. I have, however, always doubted whether the test has any real importance.

Click here for next section