J. Tukey to Fisher: 27 *August* 1954

It now begins to look as if I may be in England briefly this fall. If so, I look forward very much to seeing you again. I hope to find you in Cambridge.

Perhaps you will remember a little document on the 'Purposes of Fiducial Inference', a copy of which I sent to you some time back. Both you and Owen were kind enough to send back some comments, to which I have not made organized reply. I am afraid, if what I hear by devious routes is correct, that you and Owen misinterpreted my intent in giving a number of examples. All but the last type of example were intended to illustrate the need for the restrictions currently imposed on pivotal functions (e.g. in Owen's paper[1] in *Sankhyā*. The *last* type of example was intended to illustrate, and still illustrates to my satisfaction, how these restrictions *do not* imply uniqueness, *even* in the Behrens–Fisher case.

I look forward hopefully to discussing this matter with both Owen and yourself.

[1] Owen, A.R.G. (1948). Ancillary statistics and fiducial distributions. *Sankhyā* 9, 1–18.

Fisher to J. Tukey: 1 *September* 1954

. . .It would not be surprising if there were a theoretical ambiguity in Behrens' problem, since there is no sufficient set of estimates appropriate to the three parameters, i.e. the two variances, and the distance between the means. The amount of information dropped, however, must be quite minute, and any ambiguity of a serious magnitude, e.g. such as the difference between my solution and those of Bartlett and Welch, could not be explained in this way.

J. Tukey to Fisher: 22 *April* 1955

You may remember our previous correspondence about the nonuniqueness of fiducial probability. I suspect that you will class me as 'still unregenerate', but I still adhere to my belief that examples 6 and 7 are relevant. I have finally revised the material with a view toward publication, separating the more factual material on nonuniqueness from the more 'opinionated' material on the place and support of fiducial inference.

I enclose a copy of a draft of the material on nonuniqueness. I should not like to misrepresent your views on what restrictions should today be attached to pivotal quantities, so I send this in the hope you will care to look it through for misrepresentations. If you and/or Owen would care to look at examples 6 and 7 any further, that too would be fine. . . .

Fisher to J. Tukey: 27 *April* 1955

I am interested to get your long screed about probability statements concerning parameters derived from observational data. I think I can see what you are trying to do, but I do not think you are getting very warm, or are trying the right way.

If you must write about someone else's work it is, I feel sure, worth taking even more than a little trouble to avoid misrepresenting him. One safeguard is to use actual quotations from his writings; better still a series of comparative quotations, for it may be that, like yourself, he is still capable of revising his forms of words.

On page 7, you ascribe to me, without reference, a belief that I have never held. I have rather often emphasized my belief that statements of fiducial probability are in fact statements of probability, and can be verified by observational frequencies in exactly the same way as the statements of probability of the eighteenth century writers, Montmort, de Moivre, or Bayes. Clearly you have no support in my own writings for what you say I 'admit and, in fact, advertise'. Could you let me know from whom this mischievous statement was obtained?

[Start of page 2 of letter.]

It occurs to me that you may be referring to my objection to identifying the level of significance with certain frequency criteria. This is because such criteria have been arbitrarily or unsuitably chosen, and in some cases because no such criteria can be found for all cases of a composite hypothesis. This of course throws no doubt on the verification of any probability statement by observed frequency, if the appropriate frequency is observable, as you would see if you could ever get your bull-headed mind to stop and think. A level of significance is a probability derived from a *hypothesis*, not one asserted in the real world. A paper of mine in the *J.R.S.S.* [*CP* 261] may help you here. I have only seen the proof so far.

I expect you know really, whatever you may pretend, what the process of reasoning was for which I coined the term 'The fiducial argument'. You know too that I have, since 1930, laid stress on it as a form of reasoning of a particular kind. Probably you could reconstruct its steps. The probability integral of the exact frequency distribution, in finite samples, of an exhaustive statistic is used to form a continuum of probability-statements, of the form

$$Pr\{T < T_P(\theta)\} = P,$$

and using the monotonic property of the functions T_P for all P, this is transformed to the equivalent

$$Pr\{\theta_{1-P}(T) < \theta\} = P,$$

a complete set of probability statements about θ, in terms known for a given sample value T.

[Start of page 3 of letter.]

If you are sensitive to inductive processes of reasoning you will not think that the argument stops there. It requires also the observation that though the particular pair of values (T, θ) relevant to the experimenter has been shown to belong to a reference set in which the frequencies of all such inequality statements as the above are exactly known, yet the possibility of recognizing a sub-set having a different probability has to be excluded, and that this can be done if T is exhaustive, and in the absence of knowledge *a priori*, or from sources other than the sample under discussion.

In speaking of fiducial probability in connection with your numerous examples, I am sure you would have done better to examine in each case whether the material does or does not exist for the development of an argument of this type. This would at least spare you from arriving at contradictions, for the argument can be used quite rigorously. In examples 6 and 7, I should not expect to be able to complete such an argument, but it may be you will be more successful, if you actually try to see whether a fiducial statement can be inferred or not, instead of asserting that it can.

You may like to verify in a couple of lines that it can be inferred in the only case in which I have given a bivariate frequency distribution, namely that for the two parameters of the Normal distribution, so sternly rejected by Bartlett!

[P.S.] p.15. If you have a frequency distribution, of course you can derive a set of Tests of Significance. I suppose this is what you mean by a family of confidence sets. It is the converse argument which fails.
p.13. I have never spoken of the 'desirability' of uniqueness. Uniqueness is guaranteed by the fiducial argument when conducted with sufficient care to be rigorous. It is not guaranteed by short-cuts which by-pass the argument. Lack of uniqueness is a good check on them.

J. Tukey to Fisher: 10 *May* 1955

Thank you very much for your letter of 27 April. I will endeavor to do what I can to get my 'bull-headed mind' to stop and think. I would suggest, however, that what I was engaged in when I sent this document to you was 'taking even more than a little trouble to avoid misrepresenting you'. I shall give the material in your letter a very careful thinking, and I am sure that this will enforce various changes in the document. I hope you will not mind if I send the next version on too so that you may kick it around again if you think it needs it. . . .

J. Tukey to Fisher: 12 *July* 1955

I have been slow in getting back to your letter of 27th April . . .

I appreciated your letter very much, and have been trying to get things through my 'bull-headed mind'. There are some points which appear to me to require careful consideration and on which I cannot feel sure of your position from your letter. So I am venturing to write again.

(1) The 'mischievous statement' about 'admit and in fact, advertise' was not obtained from anyone else and was solely my own responsibility.

(2) You say that you have emphasized your belief (and presumably continue to believe) that 'statements of fiducial probability are in fact statements of probability, and can be verified by observational frequencies'. Under what circumstances can I verify the fiducial probabilities resulting from the numerical results of the Behrens-Fisher procedure as observational frequencies? (And are these circumstances generally reasonable in connection with actual experiments?)

(3) If I understand the top paragraph of your second page correctly, I agree. Therefore let us postpone any discussion here.

(4) The lower paragraph of your page 2 certainly must be accepted by anyone.

The emphasis on statements of inequality and the 'monotonic property of the function T_P' causes me to ask the following questions.

Define ψ as a $1-1$ function of θ by

$$\psi = \begin{cases} \theta, \text{ when } \theta = \{\theta\}, \text{ that is when } \theta \text{ is an integer,} \\ \{\theta\} + (1 - (\theta - \{\theta\})), \text{ otherwise,} \end{cases}$$

and then let, for example, x be a sample of 1 from a normal distribution with average value θ and variance 1. Can we then make fiducial probability statements about ψ? I cannot predict your answer.

(5) I come now to the top paragraph of your page 3, where you say: 'yet the possibility of recognizing a sub-set having a different probability has to be excluded . . .'. The words are here, but I am not sure that I have caught all the thought.

Do you mean that the formal inferential procedure is unsatisfactory for your purposes, unless it has been so restricted as to produce a unique result? And that if it is unsatisfactory, it should not be called fiducial probability? If your answer is 'yes', then I would assume that each proposed application of fiducial probability would be incomplete until a uniqueness theorem had been proved. Would this be your position?

It is of course true that monotonicity provides additional uniqueness in one-dimensional cases, and in some two-dimensional cases. I am not sure whether you rely on monotonicity in cases involving more than one parameter. Do you? And if you do, what definition of monotonicity do you use?

(6) Later on in the same paragraph, you say 'in the absence of knowledge *a priori*, or from sources other than the sample under discussion'. Before reading these clauses, I would have felt that you believed that it was always possible, and often wise, to summarize the inferences from any single body of

experience. Therefore I would have assumed that you would have been willing to begin such a statement 'in the actual, or assumed, absence. . .' and thus encompass both senses. But your detailed words now leave me in doubt. What is your position here? If there exists a further, entirely separate, body of experience, may we properly make statements of fiducial probability (i) about either body of experience separately, *and* (ii) about both bodies of experience jointly?

(7) Until the top paragraph of your page 3 is clearer to me, I don't think that I can usefully discuss the next paragraph.

(8) You then suggest that I verify something for the case of the two parameters of the normal distribution. What may I verify? That some specific formal procedure gives a unique answer? Which formal procedure? I am truly at a considerable loss here!

(9) You say in that paragraph 'in the only case in which I have given a bivariate frequency distribution'. Am I in any sense to interpret these words as disclaiming the application of fiducial probability to any other example with two or more parameters? What about the bivariate normal with known second moments? What about the between variance component in a simple analysis of variance? What about the Behrens - Fisher problem? You have me at another loss!

(10) In your discussion of the one-parameter case you laid great stress, properly in my view, on the monotonicity of distribution functions. As we are both well aware, a distribution of one statistic is monotone in a handy way, while distribution functions of two or more statistics maintain a positive increase in terms of higher differences which are not nearly so handy to deal with. What use would you plan to make of monotonicity in situations involving more than one statistic? . . .

Fisher to J. Tukey: 18 *July* 1955

I believe you are trying to understand what I write, though this would be easier if you were more familiar with what I have written in the past. If I could be sure of that, I could assume that you were using words in senses conformable with my own usage, or, in case you thought my usage incorrect, that you would draw my attention to a better way of speaking.

I see from your letter that the word 'probability' does not mean to you even approximately what it means to me, and I am not unfamiliar with the mathematical and scientific literature involving this word. In the first place you seem to express surprise at my belief that 'statements of probability can be verified by observational frequencies'. It would have been more explicit had I said 'statements of probability in the Natural Sciences' or 'statements of probability referring to the real world', for it is possible as an abstract mathematical exercise to make statements which refer to nothing on earth. Your view apparently is that no statements asserting a mathematical probability are capable of such experimental verification, which seems to imply, in the Natural Sciences, that all such statements are meaningless.

In the second place you do not seem to grasp the central characteristic of the concept of mathematical probability, namely that it enables a statement of uncertainty to be made with rigorous exactitude. This requires a specification not only of what is known, or can be validly asserted on the data, but also of what is unknown, in order that the probability statement should be distinguishable from a statement of certainty. To specify what is unknown requires, naturally, an exhaustive scrutiny of the data; this, of course, is not necessary when 'certain' inferences are deduced, since for them only a portion of the axiomatic material available may be required, and other axioms may be ignored.

I do not see how anyone even slightly acquainted with what I have written on statistical topics could not have seized my meaning of these points.

Having accepted page 2, if I understand you rightly, you ought to read the first sentence of page 3, when perhaps the rest of the paragraph will become clear to you. There is not a word out of place, and scarcely one that is not needed for an exactly rigorous statement. If you cannot understand the first paragraph of page 3, then I am afraid you cannot know what the phrase 'fiducial probability' means, and as you want to write about this, you do owe the effort necessary to grasp its meaning.

I do not know what your trouble is about monotonicity. Is there more than one definition for me to choose from, for you ask me what definition I prefer? The function T of θ and ϕ may, of course, be monotonic for θ uniformly for all ϕ. Is this the kind of definition you are asking for? The point of monotonicity was, as you would have seen from my letter, merely to establish the validity of the transformation of an inequality statement. Do you claim it is impossible validly to transform simultaneous statements of inequality? You have exhibited, of course, that it is rather easy to make bogus, or invalid, transformations of this kind. If the probability that $x>a$, and simultaneously that $y>b$, is given by a numerical value P over any well-defined field of possibilities, or reference set, then I believe that the same probability may be applied to the simultaneous inequality $a<x$ and $b<y$. I suggested that you should verify, what is not difficult, that the simultaneous distribution of the parameters of a normal population demonstrably had the simultaneous distribution I gave for them long ago, and that the method outlined on pages 2 and 3 supplied a rigorous form for such a demonstration. The case you mention of the five parameters defining a bivariate sample also offers no particular difficulty. If you could discipline your mind to carry out the argument step-by-step and with exactitude, even in a single case, you would, of course, put yourself in a position to recognize the many cases in which a parallel argument cannot be developed.

I was wrong in saying that the bivariate frequency distribution of the normal parameters was the only case I had published; I have also published

more extensive problems. I do not suppose that you are such an idiot as 'to interpret these words as disclaiming the application of fiducial probability to any other example with two or more parameters'. The process of addition in arithmetic is widely applicable, even though neither you nor I have published every possible example of its use. In Behrens' problem I confirmed that Behrens had given a valid and precise test of significance. There is no Sufficient simultaneous estimate of the three parameters, so that a three-fold fiducial distribution is not obtainable.

But surely you can see these things for yourself!

J. Tukey to Fisher: 25 *July* 1955

1. Your immediate, though partial, reply to my letter about fiducial [probability] is a great help to me, but I still am in certain difficulties, which I shall endeavor to explain below.

2. I certainly believe I am trying hard to understand what you write. I agree that a maximum of familiarity with what you have written in the past would be a good thing, but even the entropy of the universe only tends toward a maximum. In this connection, however, may I remind you of your letter of 27 April, in which you suggested that, in writing about someone else's work, it is well to use 'a series of comparative quotations, for it may be that, like yourself, he is still capable of revising his forms of words'. I hope that I am engaged in following this advice.

3. I will plead guilty to the charge that probability means many things to me, but will assert that one of them is what you had in mind. Moreover, I did *not* intend to express surprise that 'statements of probability can be verified by observational frequencies', for surely such a 'probability' is a probability in the most useful and esteemed sense. It was my intention to accept this as the meaning of the word 'probability' for this part of the discussion, and then to inquire 'Under what circumstances can I verify the (fiducial) probabilities resulting from the numerical results of the Behrens–Fisher procedure as observational frequencies?' (I have now added parentheses around fiducial in view of your comments of 18 July.) I should like to repeat this question in what I hope is now a far more clearly defined context. Under what circumstances?

4. You go on to say that I do not seem to grasp 'the central characteristic of the concept of mathematical probability, namely that it enables a statement of uncertainty to be made with rigorous exactitude'. This is, I suspect, a place where even carefully chosen words may have two meanings, and where I have been likely to have missed the intended point in the past. There are two ways in which, as it seems to me, the words 'rigorous exactitude' may be construed:

(1) as meaning that the asserted probabilities will be exactly confirmed by observational frequencies, if and when we observe under the specified circumstances,

(2) as meaning not only (1) but also that among the statements with property (1) the statement being made is, to a suitable extent, rigorously unique.

(I will maintain that either construction is logical and grammatical). I inquire whether I am now correct in believing that (2) is what you really mean.

5. We come now to the first sentence of page 3 (your letter of 27 April). Does the preceding paragraph suggest that this and its paragraph are now clear(er) to me? (There remain some more formal questions, which I propose to take up below).

6. I am still unclear as to your position on where the line may be drawn and still have the result fiducial probability. May I quote you from two places for comparison:

'. . . In following his example it is not necessary to deny the existence of knowledge based on previous experience, which might modify his result. It is sufficient that we shall deliberately choose to examine the evidence of the sample on its own merits only. This has not only the advantages of giving simplicity and definition to the problem, it has the profoundly important effect that modern tests of significance, treating each body of data as unique, can thereby derive from them *independent* evidence which may be compared, knowing it to be independent, with evidence from other sources. In applying this principle, there is, of course, nothing to prevent us from combining the evidence of several different samples. We can do so and at the same time treat the whole body of available material as a unique body of data. Without methods of treating unique samples, we should have no real guidance in these more complex cases.' (*Annals of Eugenics*, Volume 9, page 175) [*CP* 162].

'. . . and that this can be done if T is exhaustive, and in the absence of knowledge *a priori*, or from sources other than the sample under discussion.' (letter of 27 April, page 3).

Am I to interpret the phrase in the letter as meaning that we *must* always include *a priori* information and other samples, or should I follow the longer quotation and hold that we are entitled to a choice as to whether we include it or not?

7. As I interpret your statements in the letter of 18 July: 'I do not know what your trouble is about monotonicity'. 'The point of monotonicity was, as you would have seen from my letter, merely to establish the validity of the transformation of an inequality statement'. You do *not* feel that a monotonic connection between the parameters and the sufficient statistics is essential. Is this correct?

8. You state 'If the probability that $x>a$, and simultaneously that $y>b$, is given by a numerical value P over any well-defined field of possibilities, or reference set, then I believe that the same probability may be applied to the simultaneous inequality $a<x$ and $b<y$.' Let me ask your opinion about a specific application of this sentence. Suppose that

(*) u is normally distributed with mean μ and unit variance,
(**) v is normally distributed with mean υ and unit variance,
(***) u and v are independently distributed,
(****) α is a fixed real number,

$$(*****)\quad x = \sqrt{(u-\mu)^2 + (v-\upsilon)^2} \quad \cos\left\{\tan^{-1}\frac{v-\upsilon}{u-\mu} + \frac{\alpha\mu}{1+(u-\mu)^2+(v-\upsilon)^2}\right\}$$

$$(******)\quad y = \sqrt{(u-\mu)^2 + (v-\upsilon)^2} \quad \sin\left\{\tan^{-1}\frac{v-\upsilon}{u-\mu} + \frac{\alpha\mu}{1+(u-\mu)^2+(v-\upsilon)^2}\right\}$$

Then we have

$$P = \frac{1}{2\pi}\int_a^\infty e^{-\frac{1}{2}x^2}\,dx \int_b^\infty e^{-\frac{1}{2}y^2}\,dy$$

and, if $|\alpha| \leqq 1$, I have shown that the simultaneous inequality $x \leqq a$, $y \leqq b$ can be inverted into a statement concerning μ and υ, the inversion being by the solution of (*****) and (******) for μ and υ in terms of x and y, for u and v fixed, with this solution being unique, continuous and differentiable. *Does your statement apply here*? If not, why not? If it does apply, is any one value of α to be distinguished from another? If no value is distinguished, then, since the result depends on α and is thus not unique, must not all results fail to give the fiducial distribution of μ and υ given u and v (even the one with $\alpha = 0$)? (In this case, would not the means of two independent unit normals fail to have a fiducial distribution?) If some value of α is distinguished, by what general principle is it distinguished?

9. You suggest that I 'should verify, what is not difficult, that the simultaneous distribution of the parameters of a normal population demonstrably had the simultaneous distribution I gave for them long ago, and that the method outlined on pages 2 and 3 supplied a rigorous form for such a demonstration'. At this point, I am still baffled as to just what is to be shown. Consider the following statements:

(A) There exists a pair of inequalities whose simultaneous probability is independent of the parameters.

(B) There exists a unique inversion of the functions appearing in these inequalities.

(C) These inverse functions are continuous.

(D) These inverse functions are differentiable.

(E) There exists no essentially different pair of inequalities with property (A).

(F) There exists no essentially different pair of inequalities combining properties (A) and (B).

(G) There exists no essentially different pair of inequalities combining properties (A), (B), (C).

(H) There exists no essentially different pair of inequalities combining properties (A), (B), (C) and (D).

Just what combination (or one of which combinations) would I have to establish in order to be sure that I had a fiducial probability distribution for the two parameters of a general univariate normal?

10. In your last paragraph, you say that there is no three-fold fiducial distribution in Behrens' problem because 'There is no sufficient simultaneous estimate of the three parameters . . .'. I believe I may possibly understand the facts of the situation, but let me recapitulate them so that we may see if we agree:

(A) $\bar{x}_1 - \bar{x}_2$ is a sufficient estimate of $\mu_1 - \mu_2$.

(B) s_1^2 is *not* a sufficient estimate of σ_1^2, *nor* is s_2^2 a sufficient estimate of σ_2^2, *nor* are s_1^2 and s_2^2 jointly sufficient estimates of σ_1^2 and σ_2^2, *because* $(\bar{x}_1 - \bar{x}_2)^2$ provides information about $\sigma_1^2 + \sigma_2^2$ in the special case where $\mu_1 = \mu_2$ by assumption.

(C) However, $\bar{x}_1 - \bar{x}_2$, s_1^2 and s_2^2 *are* jointly sufficient *statistics* for $\mu_1 - \mu_2$, σ_1^2 and σ_2^2.

Apparently, the subtle distinction between 'sufficient statistics' and 'sufficient estimates' is very important here. Is this your view??

11. The following simplified version of the Behrens' problem should prove illuminating:

(A) $\{x_{1i}\}$ is a sample from a normal population of (unknown) mean μ_1 and (known) variance 1.

(B) $\{x_{2j}\}$ is a sample from a normal population of (unknown) mean μ_2 and (unknown) variance σ^2.

(C) We desire either
  (i) to make a valid test of significance concerning $\mu_1 = \mu_2$,
  (ii) to make a valid test of significance concerning $\mu_1 = \mu_2 + \theta$ for some preassigned θ,
  (iii) to establish the fiducial probability distribution of $\mu_1 = \mu_2$ (if this is legitimate).

As I understand the situation under (i), evidence concerning σ^2 is provided both by s_2^2 and by $(\bar{x}_1 - \bar{x}_2)^2$ and, if we are to make exhaustive use of what the observations bring us, we should combine these in some way. Is this your view?

Fisher to J. Tukey: 29 *July* 1955

Your long letter shows that you have been thinking, but still you do not easily take notice of what I have written. Instead you keep reiterating any irrelevant puzzle which in the past you have been unable to get clear about, as though

you were unable to think clearly about our interchange of ideas until everything had been restated to you to your own satisfaction. If you would concentrate on one question at a time I am sure you would make better progress.

I am glad that you do not now challenge my statement that probabilities (as the word is used in the natural sciences) can be verified by observed frequencies. That is one point gained. The next step might be to consider the second point in which you do not (or apparently did not) understand the word in the same way as myself. However, you break in with a confused remark about Behrens' test, knowing, I believe, exactly what my answer is, but anxious to avoid the serious consideration of what you are overlooking in the semantics of the word probability.

On the (irrelevant) Behrens question, you know how to verify 'Student's' t distribution experimentally. Do this with two t distributions simultaneously, with n_1 and n_2 degrees of freedom; for any assigned area marked off on a (t_1, t_2) diagram compare the number of points observed with the number calculated by integration.

You seem to dispute my statement, in my last letter, that there does not exist a simultaneously sufficient set of estimates for the common mean and true variance of a pair of independent samples. So you are still confusing this case with those in which sufficient estimation is possible. This is a fair test case: Do you only want to argue, or to make some progress in understanding?

I hope now I have set aside your obstacles, doubtless erected under a subconscious urge, to your consideration of the 'central characteristic'. 'Rigorous exactitude' is not an unfamiliar idea. The phrase would be quite intelligible to you if you wanted it to be. Neither of your 'two ways' is near the mark. For (i) please note that I am not harking back and reiterating what was said about observational frequencies; I hoped that that point had been done with. I was making a *second* point with a new paragraph and page. About (2) I suppose all rigorously exact statements are unique. Correct me if you disagree about this, but it was not what I was saying, for, as I tried to suggest, I was not using unnecessary words.

The words following 'rigorous exactitude' now deserve your attention. The passage is less than four lines long. Probably you are unwilling to agree with it, and will resolutely shut you eyes to its meaning. The paradox you attempt to build from it is phoney, as you perfectly well know. There is no contradiction between the two quotations, though written 20 years apart or so. At the risk of Dame-School simplicity, may I suggest that it would be reasonable to say 'On the data as it existed before 1950 the probability that x exceeded a was over 90%', or 'evidence that has accrued in recent years has led to very different views of these probabilities'. If you can understand statements in this form, you will be able to distinguish, as men of science do, between statements referring to the real world, and statements referring only to those hypothetical worlds created by hypotheses. When speaking about the

real world we must, if we are making a statement involving uncertainty with the care such a statement requires, always include all the data available for the inference. In a hypothetical world also we must use all that, within that hypothesis, is available.

You will have a better chance of getting this if I stop here without discussing your very clever paradoxes.

Fisher to J. Tukey: 9 *August* 1955

Thanks for your note[1]. At the risk of being tedious, I am sure I ought to suggest again that you try to carry out a strict fiducial argument based on inequalities, instead of relying merely on the pivotal transformation. I feel sure you might find in this way where your difficulty lies.

[1] Tukey had written thanking Fisher for a pleasant time during their recent discussions in Cambridge.

J. Tukey to Fisher: 18 *October* 1955

Your note of 9th August has just reached me . . . In it you suggest that I 'try to carry out a strict fiducial argument based on inequalities, instead of relying merely on the pivotal transformation'. I am afraid that I have my old trouble with this remark — wondering which of the several meanings that seem plausible to me you meant — and fearing that you meant still another.

(1) Do you mean that following through a chain of inequalities is likely to bring to the surface subtle misbehaviours which are easier to overlook in the pivotal analysis?

(2) Do you suggest that cases where inversion is in closed form are safer? Or better in principle?

(3) Would you feel that a correct pivotal argument which showed that the mapping from the pivotal plane to the parameter plane was everywhere smooth and one-to-one was not enough?

(4) Or have I missed the point again?

P.S. The day with you was a great enlightener, and the effect has steadily increased.

Fisher to J. Tukey: 22 *October* 1955

No, I do not mean any of the three things you say, in so far as I can understand your numbers 1, 2 and 3. I only meant what I said, namely that you should 'try to carry out a strict fiducial argument based on inequalities, instead of relying merely on the pivotal transformation'. In fact I am sure that the pivotal transformation as used by you, and without safe-guards, which if you were a little forward-looking you might be able to think up, is thoroughly misleading.

The direct use of a strict fiducial argument, which I recommend, I have illustrated rather elaborately and specifically for your benefit on page two of my letter of April 27th, so it is really disappointing that you should not know what I meant by my suggestion.

J. Tukey to Fisher: 10 *January* 1956

There appears to be a certain chance that I could be in England for a while this summer. I found the day I spent with you last summer very stimulating and helpful. Consequently, a major factor in my decision about the coming summer, should a trip to England become possible, would be the possibility of more extended contact.

Do you plan to be in Cambridge this summer? Would you desire to spend a substantial amount of time in discussion, educating me further? (Please do not hesitate to say 'no' if you feel so inclined. I should not like to impose on you!)

Fisher to J. Tukey: 16 *January* 1956

I expect to be in Cambridge during the summer, at least until about mid August and perhaps later, though there will be a short interval for a congress on human genetics in Copenhagen.

Of course I should like to resume talks with you, and these could be fruitful if you can get over the sort of caginess or inhibition which seems to be preventing you from hearing or seeing anything that could possibly remove any misapprehension.

At least I expect I have got you to understand that I do not use, or recommend, the process of mechanical transformation by the Jacobians, in which you and Savage seem to have been exercising yourselves following the earlier work of Segal and Diananda. I know that an unguarded phrase of mine has been read to this effect, but this misinterpretation has not only led to you barking eagerly up the wrong tree, but to your missing the point I was trying to make, namely the immense analytic abbreviation made possible by recognition of the fiducial argument. One of the best examples of that is its application to *Behrens' problem*.

J. Tukey to Fisher: 11 *February* 1958

. . . I am writing to you to ask your advice and sanction on a matter of terminology. I have been asked to give the Wald lectures at the Institute of Mathematical Statistics meeting next fall and have accepted. The purpose of this set of lectures is to give a connected account of some branch of statistics which is not easily available as such in the literature. It is my present intention to say what I can about those matters connected with fiducial inference which I think I understand. (And perhaps go on and make suggestions rather than statements about other matters.) I am not at all sure, as always, that what I would say would be exactly what you would like to see said. This perhaps offers a good reason for using a name slightly modified from 'fiducial' such as, for example, 'fiduciary'. On the other hand, the material which is to be discussed is built on a foundation of your own ideas. From this point of view, it would be desirable to retain the word 'fiducial'.

Do you have any wishes, desires or advice in this matter? I should be glad to try to be guided by any preference you might have as to which choice of terminology I should adopt. . . .

Fisher to J. Tukey: 17 *February* 1958

I have your letter of February 11th and I suppose its purpose is to know whether I like the word 'fiduciary' as a variant of 'fiducial'. The answer is no, not a bit. Of course I knew of the use of the word 'fiduciary' in financial circles and was therefore anxious to avoid its use in another connection. This use fiducial is as an adjective qualifying the word 'argument' in order to contrast it with a Bayesian argument, though in particular examples they may be brought indistinguishably close together, as perhaps you have seen in Section 6 of Chapter V (*Statistical Methods and Scientific Inference*).

It is not an adjective qualifying the word 'probability' save in the sense of a probability derived by the fiducial argument. I use the word 'probability', unlike Neyman, Wald, and most of your friends, in the sense in which it was used by the old masters, Fermat, Pascal, Leibnitz, Bernouilli, Montmort, de Moivre, and Bayes. These all thought that you could make probability statements about parameters characteristic of the real world. They did not think that such probabilities were necessarily zero or unity. So far your publications on this subject have not clarified it.

Fisher to S.S. Wilks: 15 *August* 1939

I was hoping you would visit us again before leaving London. However, I expect you found that you had a great deal to do in your last days here.

I am sending a short note,[1] submitted for the *Annals of Mathematical Statistics*, written chiefly to correct the impression given by Bartlett's recent paper[2] that I have confirmed the validity of his own approach.

As I think it was just this so-called 'exact inference of fiducial type' which led him into all his other criticisms, it may be worth making clear that I have written nothing which could be taken as confirming its validity. The relation of tests of significance to the various possible hypothetical succession of repeated samples may, I hope, be made a little clearer by the latter part of my note.

[1] Subsequently published as 'A note on fiducial inference'. *Ann. Math. Stat.* 10, 383–8 (1939) (*CP* 164). Wilks was editor of the journal.

[2] Bartlett, M.S. (1939). Complete fiducial distributions. *Ann. Math. Stat.* **10**, 129–38.

S.S. Wilks to Fisher: 27 *October* 1939

In reading your 'Note on fiducial inference' several points have occurred to me which I should like to pass on to you. Perhaps you can straighten me out on them. I am glad that you have indicated on the first page that when you introduced the concept of fiducial probability, you had in mind its use in connection with the theory of estimation, and hence in making significance tests. It seems to me that this point cannot be emphasized too strongly and that it might well be brought in at several points, for the phrase 'inference of fiducial type' has apparently reached a stage of wider interpretation (perhaps unfortunately) than you originally had in mind. To take a rather trivial and extreme case, if one has a sample of n objects from a normal population and considers the quantity

$$t = (x_1 - a)/s \qquad (1)$$

where x_1 is one of the items taken at random from the sample, then t has the 'Student' distribution with $n - 1$ degrees of freedom. One can make 'inferences' of 'fiducial type' about a in this wider sense from t given by (1). The inference about a is of course far from being satisfactory from an estimation point of view as compared with the inference which can be made by using the mean $\bar{x}$. In practice, one would obviously fix up a function (a t function in this case) that would enable him to make inferences about a as close as possible. Now in this wider sense, I think everyone will agree that Bartlett's T and T' (the choice determined at random) furnishes a method of making an 'inference of fiducial type' but, as you have pointed out, the inferences by this method from the estimation point of view are unsatisfactory on account of the peculiarity of the functional form of T'. In practice, one would not want to use T' for the purpose of making inferences about the difference between the population means. Now my own feeling, based on what I know of the controversy between yourself and Bartlett together with what a number of others have said about it, is that readers will misinterpret the first sentence in your summary by regarding your statement as being equivalent to the statement that Bartlett's test does not provide an 'inference of fiducial type'. This will also in my opinion leave open the way for another rebuttal from Bartlett. In other words, is not the following statement true: Bartlett's test provides a basis for making an 'inference of fiducial type' about the difference $(a_1 - a_2)$ between the population means in the wider sense, but the inference is not satisfactory from an estimation point of view, as you have shown?

On page 5 . . ., the first three lines of the last paragraph are not clear to me. I am probably misinterpreting your statement. Apparently $\bar{x}$ will exceed $st_n/\sqrt{n'}$ with the chosen frequency regardless of whether σ^2 is known or not. Of course, if the value of σ^2 is actually used, we can say that $\bar{x}$ will exceed $\sigma t_\infty/\sqrt{n'}$ with the assigned frequency. I believe you will confirm me in saying that this latter statement is not inconsistent with the statement that $\bar{x}$ will exceed $st_n/\sqrt{n'}$ with the chosen frequency. . . .

Fisher to S.S. Wilks: 6 *November* 1939

Much of my note on fiducial inference is devoted to the question 'Did Fisher assert that Bartlett's distribution provided an exact inference of fiducial type?' Much of your own letter, on the other hand, is concerned with the different question, whether Bartlett's T and T' furnish a method of making an inference of a fiducial type. The second question may be the more important, and would be, if there were agreement as to the use of the word 'fiducial'; but it was the first, which is at least capable of being settled definitely, that I was concerned to settle in my note.

The case which you take on your first page is not nearly so extreme as the case of T and T'. Your case is of the familiar type, where the fault of the method is equivalent simply to ignoring part of the data. I have always protested against speaking of such inferences explicitly or implicitly as inferences from the data, though I can have no objection to speaking of them as fiducial inferences from a selected portion of the data. In Bartlett's case, on the other hand, I do not think that the distribution leads to a test of significance at all.

In your allusion to my statement on p.5, that it is highly improbable that $\bar{x}$ will exceed the limit assigned with the frequency chosen, I think you have missed the point of the distinction I was drawing. In any fixed population the limit will be $\sigma t_\infty/\sqrt{n'}$ and this will only be equal to $st_n/\sqrt{n'}$ for a particular value of s. It is highly improbable that this particular value will be the value of s observed. As I say later, on p. 6, the distribution of t will be so verified. This implies that $\bar{x}$ exceeds $st_n/\sqrt{n'}$ with the correct frequency when each observed $\bar{x}$ is associated with its own observed s. . . .

I want you to remember that the occasion of my note is that Bartlett has ascribed to me a view which I do not hold. It is always rash in controversy to express the views of an opponent in any other way than by specific quotation. Misrepresentation, I always think, is particularly to be avoided, since, in political and forensic controversy, no artifice is commoner than that of misrepresenting an opponent's views in order to create prejudice. To do so in scientific controversy seems to me a very serious offence against good taste, since we are each doing our best to express the truth clearly. . . .

S.S. Wilks to Fisher: 21 *November* 1939

I am enclosing proof and *ms* of your note. You will note that I took the liberty of making a few minor changes which, however, I think are consistent with the understanding contained in our recent exchange of letters. I still feel that the first sentence in the second paragraph may be a bit too strong and not sufficiently impersonal but to leave out the word misleading may leave the statement too empty.

On p. 2 of your letter (first paragraph),[1] I think I understand the distinction you are drawing, but there is some question in my own mind as to whether the statement might be improved in wording. I simply mean this: If for each

sample we compute $\bar{x}$ and s, we can make the statement that $\bar{x} > st_n/\sqrt{n'}$ and be correct with the frequency assigned, say 0.025, understanding of course that $\bar{x}$ and s are inserted for each sample. If s for a particular sample is used over and over, then obviously we cannot quite say that $\bar{x} > st_n/\sqrt{n'}$ and be correct with the chosen frequency (the only thing changing from sample to sample being $\bar{x}$). On the other hand if σ is known we can say that $\bar{x} > \sigma t_\infty/\sqrt{n'}$ and be correct with the chosen frequency, where the same value of σ is used repeatedly. The whole question hinges on whether one has in mind using the value of s for a particular sample repeatedly for various samples. The statement in my other letter on this point presupposed that in considering $\bar{x} > st_n/\sqrt{n'}$, $\bar{x}$ and s changed from sample to sample according to the sampling law of s and $\bar{x}$. Presumably, you do not mean to change s from sample to sample. . . .

[1] i.e. third paragraph of Fisher's letter.

Fisher to S.S. Wilks: 8 *December* 1939

I have sent back the corrected proof by air-mail, the main alterations being in the punctuation, especially where formulae are concerned, and one important error, which I hope you will see is not messed up, namely where S^2 is used for s^2.

Of course, in the case of 'Student's' problem, the practice of obtaining a fiducial limit from a single sample, e.g.

$$st_n/\sqrt{n'}$$

and then complaining that a population of samples drawn from a fixed population will not be divided by this limit in the ratio specified by the level of significance chosen is a ridiculous one, in that it shows a complete misunderstanding of 'Student's' test. It is, however, actually this argument which has been used, or implied, as though it were an effective criticism of Behrens' test, the critics being, I suppose, genuinely misled by the fact that, in 'Student's' test, we can use actual tables of a quantity t whose distribution is known, whereas, in Sukhatme's case, the simultaneous distribution of t_1 and t_2 cannot be used directly to show the integral over any relevant area, so that it is necessary, for purely practical purposes, to have a secondary table of d. I imagine that Bartlett's puzzlement arose from him thinking that d was the analogue of t, not only in being used for tabular reference, but in its logical position in the fiducial argument. Actually, of course, a number of different functions, such as d' defined by

$$d' = d\sqrt{s_1^2 + s_2^2} \ / f(s_1, s_2)$$

might be used for tabular purposes in place of d, giving an exactly equivalent test of significance where f is any homogeneous function of the first degree.

Fisher to E.B. Wilson: 20 *May* 1935

I have now looked through W.E. Deming's paper[1] . . . What first Treloar and Wilder,[2] and later, it seems, Deming failed to see is that, in setting our level of significance at any value such as 1%, we are choosing voluntarily to make the mistake of rejecting the hypothesis when it is true in this proportion of cases: and as, on the hypothesis discussed, we are always equally wrong in such rejection, it is a matter of complete indifference, provided the proportion is kept right, in which particular samples this mistake is made.

Obviously, of course, when the population variance is known, the test which utilises this knowledge is preferable to one using only an estimate, and the reason for this preference is not that we should prefer to be deceived by one sort of sample rather than by another.

It is curious that the simple history of the problem should not have kept the author straight. From the beginning of the Theory of Errors, it must have been clear that, knowing the population variance, one could use the normal curve to test the significance of a mean. The test can easily be extended to a weighted mean, or to the regression coefficient, as soon as this is recognised as merely a weighted mean.

In the majority of practical problems, it was easily perceived that the population variance was not provided in the data, but, if the sample were large, could be estimated with some confidence, so that the problem was discussed by Gauss, what estimate it was best to take. Gauss satisfied himself that the mean square error, with allowance for degrees of freedom, was most proper, though, under the influence of Peters, astronomers were largely persuaded to use the mean error instead. As soon, however, as any estimate is used rather than a known value, the question of its errors of estimation arises; and throughout the nineteenth century it was usual to say that we needed some large number, e.g. 50 observations, before relying on the test. 'Student's' work showed that, using the mean square error, the frequency of rejection of the hypothesis, when true, can be calculated exactly, and for small samples was larger than that derived by substituting the estimate in the formula for the normal curve. 'Student's' tables, in fact, show how to make the classical test exact. . . .

[1] See also Fisher's letter of 19 September 1935 to Deming (p.82).
[2] See Fisher's letter of 13 May 1935 to Immer (p.266).

Fisher to E.B. Wilson: 8 *March* 1955

. . . A good deal of modern work, in fact, prompts in me the question that it is often very difficult for a mathematician to recognize the difference between what we do know, and what we do not know, a distinction which is not difficult to the worker in the natural sciences, but being irrelevant to a good deal of mathematical work, mathematicians are often surprised to find that it matters.

I find that the word 'fixed' is often used by statistical writers to claim for a quantity supposedly unknown, the properties of one actually within our cognizance.

Fisher to E.B. Wilson: 26 *September* 1955

. . . I do not think the existence of innate or spontaneous emotional reactions to logical situations should prevent there being rational analysis, for we can calculate and communicate the extent of the evidence or data having such effect, and the clear recognition of this extent is at least a rational as well as a communicable effect of the evidence.

I think the typical example is a simple test of significance in which we normally experience a reluctance or resistance to a belief involving a very long chance. The emotional state accompanying this resistance may be quite sub-rational, but we can invite any other rational being to re-calculate the length of the odds, and verify the nature of the evidence available to create such a rational reluctance. In most cases, indeed, there is no means of going further, e.g. of deriving a probability statement concerning the probability of any hypothesis subject to which such long odds have been calculated. I am inclined, therefore, to speak of the hypothetical improbability as affording a rational basis for disbelief in a degree measured by the smallness of the improbability calculated.

Fisher to E.B. Wilson: 8 *August* 1956

You will be glad to hear that my new book is on the point of emerging from the binders under the title of *Statistical Methods and Scientific Inference*, and I hope you will be amused by my various hair-splittings in the attempt to get clarity on a subject which has been so gravely confused in the literature.

A great deal of the thought is new, even to me, and I have only just realized that Keynes's excellent phrase, as I have always thought it, 'the degree of rational belief', is not really well applied to the concept of mathematical probability, in its classical sense implying the possibility of verification by observations of frequency, but can be seen to be more appropriate to the more primitive type of inference represented by those tests of significance from which no statement of probability can be derived.

I have discussed and, I think, made this clear, in essence though not in terminology, in the opening section of Chapter III, but I am sure that I have often commended this qualitative definition of Keynes's as appropriate to probability *senso strictu* and I now realize that it is not really the concept of mathematical probability to which it is appropriate.

I hope I may look forward to your comments some day when you have glanced at this little book. . . .

Fisher to E.B. Wilson: 28 *March* 1957

It was nice to see your letter of March 16th.

There were a sprinkling of new things in my book on scientific inference on which I should certainly like your opinion.[1] Section 4 of Chapter II, on 'The meaning of probability', is, I think, of some permanent importance as stressing an aspect of the meaning of probability which seems to have been overlooked by the leading twentieth century writers, such as Kolmogoroff and Feller. The statement of the fiducial argument in Section 3 of Chapter III is certainly more complete than anything I have done before, and I believe clears up the slight confusions left by Kolmogoroff's and Jeffreys' discussions of the subject. As a mathematical *beispiel*, the asymptotic application of the fiducial argument to discontinous data has, I believe, not been before attempted, and leads to some interesting comparisons. I hope to get the misprints in this analysis corrected for the second edition, which has already been asked for. However, from you I would particularly want to get the reader's reaction to the forewords that I have inserted before some of the chapters, such as that of Chapter V, pages 106–110, on axiomatics and Gödel's theorem, which may well amuse you and perhaps help others. The example on page 169 gives the correct fiducial inference for a case which has been mauled over a good deal, I think by one of the Princeton men, for there has been quite a cult, led I fancy by Savage of Chicago and Tukey from Princeton, for exhibiting garbled versions of the fiducial argument without any reference to the strict application of that argument itself.

On your point about the 2×2 table,[2] when making an exact calculation I always use the single tail, and if I want to compare significance with cases where both tails are used, I simply double the value obtained, without regard to the question of how lumpy the other tail may be. Usually, indeed, I think that the single tail is appropriate, though of course not always. For example if, being new to the subject, I test 500 students with phenyl-thiocarbamide and find there are more non-tasters among the men than among the women, I should only be sure there was a sex difference if on doubling the probability the result was still small enough to satisfy me, but having heard that X, Y, and Z had found fewer non-tasters among women than among men, I should be satisfied that my population was telling the same story as theirs if the probability for the single tail was satisfactorily small; but in the first case I would not feel at all concerned with the discontinuities at the tail other than that which had been observed.

No, I do not discuss this in the new book, where in Chapter IV I am principally concerned to get across the important point that it is frequencies within the least recognizable subset of hypothetical possibilities that are relevant to scientific inference. The 2×2 table gives a fairly good preliminary example of this, though in such a discussion I suppose I ought to emphasize that if, without my meaning to perform an experiment, three rabbits die

unexpectedly, and on enquiry I find that a mistake has probably been made leading to their food being poisoned, I believe any rational man would think it probable that the cause of these unexpected deaths had been revealed, even if the particular poison suspected had not previously been shown to be lethal to rabbits. As a recommender of scientific procedure, I should say that at that point it would be reasonable to perform a definitive and well-controlled experiment capable of verifying this inference, that in such an experiment there would be an adequate number of untreated animals, and verification would only be claimed if there was a significant *contrast* between the reactions of the treated and the untreated, and that contrast in the expected direction. In fact I would distinguish for this purpose between being sure, and being in a position to claim an experimental demonstration.

[1] Wilson had written to Fisher that he had read *SMSI* with great pleasure but had not found much new material in it.
[2] Wilson had asked about the use of one or two tails when calculating exact probabilities with 2 × 2 tables.

Fisher to F. Yates: 27 *July* 1955

I am sending herewith the last chapter,[1] including something about estimation and the use of ancillary values. It will be mostly old stuff to you, apart perhaps from section 12, which I put in because some Americans, Savage at Chicago and John Tukey, have been rather working themselves up about the uniqueness of simultaneous fiducial distributions, and have been trying to use Jacobians in which some of the elements, or minors, are not uniform in sign. I fancy it is one more example of the difficulty felt by many mathematicians at the idea that the kind of conclusion which can be drawn depends on the analytic form of the data. Anyway, it occurred to me that in my early work the logical notions used in this book had been developed chiefly in connection with the theory of estimation, and there were one or two things, such as the definition of 'consistency', in which I had not been very consistent myself, and in which, inevitably, people like Kendall had chosen the less satisfactory definition, without a hint that the idea had been expressed in any other way!

[1] A draft of chapter 6 of *Statistical methods and scientific inference.*

Fisher to F. Yates: 13 *October* 1955

Quite a long while ago I sent George Barnard a note on Pearson and Hartley's table 11, and he has taken some time to mull over it, and perhaps has not yet sent you the copy I sent him. Anyway, enclosed is a more complete copy, with various corrections and additions in the form in which I should, at present, be inclined to print it. Together with this I am sending a letter from George Barnard, written yesterday, and a copy of my reply.[1] This interchange may be useful as a means of getting the point of the problem

clearly realized. I think George shows some confusion on his second page, for in the case of Behrens' values as against Welch's, so far as I know both are aiming to test the significance of the same hypothesis, and both are using to that end identically the same statistic. They differ only in respect of the sampling distribution of that statistic which is regarded as relevant for testing the hypothesis.

However, confusion is to be expected, for in the course of time every sort of lunacy has been written about this problem.

Could you send me back the copy of my note that I am herewith enclosing, including the diagram? I have other copies of George's letter and my reply.

[1] See p.29.

F. Yates to Fisher: 1 *November* 1955

I am now returning your paper.[1] By and large I am in agreement with it, but the fact that people such as Barnard find difficulties suggests that some further discussion may be worth while.

The crux of the matter is, I suppose, what is meant by 'Once in ten (or $1/P$) trials' (your p.8). Welch can claim for his test that this property holds in repeated sampling from the populations with the same σ_1/σ_2, but only by ignoring the actually observed ratio s_1/s_2 (or should I rather say, using it to enter the table with, but ignoring the other information it contains).

Logically I think every right minded person would agree that all relevant information should as far as possible be taken into account when making a test of significance. The snag is in the phrase 'as far as possible'. We are, I believe, fully agreed that vague and indefinite information (such as that of the *a priori* distribution type) is best excluded from formal tests and taken into account in the subsequent overall assessment (or should I say, considered judgement) which rational beings normally make after doing their figuring. But once it is admitted that certain information can be excluded the question arises where to draw the line? In the analysis of the results of an experiment we agree to ignore the information provided by the actual pattern of treatments in that particular experiment and merely to utilize the information that the design was selected at random from a set of designs having certain properties. This is justified by just the argument which you condemn in your reply to George Barnard, and the purist, using the same argument, might well say the fact that he gets too small a chance of significance when he was lucky enough to select a Knut Vik 5 × 5 Latin square[2] is not compensated for by the fact that he gets too high a chance in the diagonal square. And that, if that is the best the statistician can do, he proposes to use Knut Vik squares in future, which have been shown statistically (as common sense would indicate) to be the more accurate. (I know there are other arguments against this course, but I don't think they are relevant to the present issues.)

I bring up this point because it does show that the statistician considers

himself free to ignore inconvenient information, and also emphasises, and rightly I think, that some[3] value is attached to getting $1/P$ significance in repeated sampling from the same population. The fact that in some common cases, such as the ordinary t test, this occurs while at the same time taking into account all relevant information, has rather blinded people to the fact that it does not *necessarily* occur. But it might be worth while giving examples in this paper — I have particularly in mind the linear regression case, and the 2×2 contingency table. The second paragraph of your letter of 13th is also very relevant.

On the paper as it is at present I have little detailed comment, except that I am not very happy about your comments on the quotations. James, I think, might well be dropped. As I see it the reason for his footnote was that he had himself cast doubts in the text on the mathematical accuracy of Welch's solution. I don't think it is *a propos de rien* — indeed, I might well have added just such a footnote myself in similar circumstances. The other people have taken the stand they do because they regard $1/P$ significance in repeated sampling *the* desirable feature. I doubt if Bartlett will now regard Welch's solution as in the least intolerable. Perhaps he ought to, but that is a different matter.

The paragraph which you added contains the phrase 'the error of these calculations'. The error is surely that of refusal to take into account relevant information, and not of calculation. . . .

[1] Subsequently published as *CP* 264.
[2] Fisher has written in the margin, 'Is there a test valid for Knut Vik?'
[3] Fisher has underlined 'some' and put '!' in the margin.

Fisher to F. Yates: 2 *November* 1955

I thought you must have got into some very deep entanglement from the time you have taken in replying, and I see that the analogy which has misled you is that of randomization.

Over the period in which I was putting forward this recommendation to experimenters, I naturally gave a great deal of thought to the effects of this procedure and the purposes that it could usefully fulfil; and, so far as statistical methods are concerned, what I was constantly pointing out was that its object was to guarantee the validity of the test of significance; that is to say that with all the great advantages of the Knut Vik square, if the results of using it were reduced by an analysis of variance, or one of the cruder techniques that preceded it, the probability statements obtained in the z test would be erroneous, whereas if proper randomization were applied, as I think you and Eden once demonstrated experimentally, the z test was made to be reliable.

Of course if a method were available to give a reliable test of significance for the use of the Knut Vik square, there would be no advantage in wider

randomization in this respect. In the analogous case of eliminating blocks in a randomized block arrangement, or rows and columns in a Latin square, we do, and I think you will agree, properly and inevitably consider an experiment laid out in randomized blocks as one of a population arranged subject to this restriction, and not as one of the larger population, to which it also belongs, in which there is indiscriminate randomization regardless of blocks. If you can be clear in your own head as to why an experimenter who knows not only that the plots did in fact fulfil the conditions of a Latin square, but also that they were laid out by choosing one such arrangement out of all possible, by such a random process as you, yourself, have explained, would it not be simply erroneous, the error being due perhaps to prejudice or ignorance, if he insisted on drawing conclusions as though the plots had been distributed at random over the whole area?

In my view it would be simply erroneous in exactly the same way as a rain maker who claimed significant success by comparing the frequency of rain following his experiments with that of the annual frequency observable in his neighbourhood, although it is within his knowledge that the frequency of rain is greater than the annual frequency in that part of the year during which his experiments were carried out. Of course we do not know a probability unless we know it, and it is only when it is within our knowledge, that it is erroneous to substitute for it a less appropriate probability. It is when we lack this knowledge, that randomization provides the safeguard.

I do not quite know what you mean about Barnard, but perhaps you have had some correspondence with him. What he said in his letter of October 17th was 'Thank you for your letter of the 14th, and that of the 13th, which make all clear, on Welch's test'. He refers to having quoted Yates' 1939 paper indicating why one was concerned with fixed s_1/s_2, but I have not looked this up so I do not know if, and if so, why you thought at that time that s_1/s_2 needed to be fixed. From my point of view this 'fixing' is not a voluntary act to be done or abstained from, but a fact to be recognized, as Behrens clearly perceived, in obtaining the appropriate test of significance.

The tests put forward by Behrens and by Welch respectively are (a) attempts to solve the same problem, and (b) attempts using exactly the same statistic, the appropriate notation for which was, I fancy, first fixed in Sukhatme's paper. The two tests differ only in the frequency distribution ascribed to this statistic. Such a difference can surely be resolved without reference to the intangible elements of judgement, but if you do employ horse-sense on this problem, do you not feel any difficulty in Welch's value being actually less than the value of t for the number of degrees of freedom supplied by the two samples jointly? What hit me in the face when I first realized this was the idiocy of supposing that a pair of samples showing no significant difference between the means when tested by 'Student's' test, on the assumption that the sums of squares could properly be pooled, should provide significant evidence that the means are unequal to a man who

professes complete ignorance, apart from the evidence supplied by the samples, as to the relative precision of the two empirical means. 'Thank God I am not absolutely sure', he says, 'that these two varieties have the same variability, else I could not claim that one gives significantly a larger yield than the other'. . . .

Click here for next section