2nd November, 1955.

Dear Frank,

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 ~~employed by~~ 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 ~~a~~ 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."

I do not think James threw any doubt at all on Welch's tables when he said "if they are correct", or words to that effect. He was not developing a reductio ad absurdum, and I think it was merely a piece of propaganda in support of Pearson's view that all the world now agrees with Neyman and Pearson in the interpretation of tests of significance.

Sincerely yours,