determining the mathematical expectations of its terms, we get a convergent series, say:

$$\sigma_{r'}'^2 = \frac{t_1'}{N} + \frac{t_2'}{N^2} + \frac{t_3'}{N^3} + \cdots \tag{4}$$

From Slutsky's theorem, mentioned before, it follows that if N increases the ratio of σ_r^2 and $\sigma_{r'}'^2$ will tend to unity. Moreover, if we take N sufficiently large, it will always be possible to fulfill the following inequalities:

$$\left|\frac{t_k'}{t_k}\right| > 1 - \epsilon_k \qquad (k = 1, 2, \cdots n)$$

where ϵ_k $(k = 1, 2, \cdots n)$ and n are arbitrary. Therefore, when n and N are sufficiently large the ratio between the first n terms of the infinite series (3) and the true value of σ_r^2 will differ from 1 by an arbitrary small number. Though the series (3) is divergent for any N, however large, the first n terms of this series will give an approximation of σ_r^2 by taking N sufficiently large.

In this paper we have shown that the procedures which have been followed by the Biometric School and Tschuprow to establish formulas for the standard errors of correlation and regression coefficients and in analogous problems can be made rigorous by the use of conditionally aleatory variables. It was found that their infinite expansions are divergent for some of the values of the random variables involved, however large the number of observations (N) may be. Yet it could be demonstrated, that the first n terms of these series will give an approximation, as close as is wanted, if N is sufficiently large. For practical purposes the case $n = 1$ is the most important.

NETHERLANDS CENTRAL BUREAU OF STATISTICS,
THE HAGUE

A NOTE ON FIDUCIAL INFERENCE

BY R. A. FISHER

In a recent paper [1] Bartlett has written a further justification of his criticism of the test of significance for the difference between means of two samples from normal populations not supposedly of equal or related variance. This test was originally put forward by W. V. Behrens [2], and later [3] found to be very simply derivable by the method of fiducial probability.

It is unfortunate that Bartlett did not restate his own views on this topic without making misleading allusions to mine. Thus, on p. 135 in [1]:

"It is sufficient to note that the distribution certainly provides us with an exact inference of fiducial type, as Fisher himself confirmed [9], p. 375."

I do now know, and Bartlett does not specify, what unguarded statement of mine could be used to justify this assertion. From the time I first introduced

the word, I have used the term *fiducial probability* rather strictly, in accordance with the basic ideas of the theory of estimation. Several other writers have preferred to use it in a wider application, without the restrictions which I think are appropriate. To all, I imagine, it implies at least a valid test of significance expressible in terms of an unknown parameter, and capable of distinguishing, therefore, those values for which the test is significant, from those for which it is not.

Shortly after Bartlett's alternative approach to the problem was put forward [4], I expressed [5] the following opinion. As this occurs prominently in the summary, indeed on the very page to which Bartlett refers in his quotation above, I cannot suppose he has overlooked it, though evidently he must have missed its meaning. I wrote as follows [5] p. 375:

"The criticism of Behrens' test of significance, recently put forward by Bartlett, on the ground that it differs from a possible alternative test, overlooks the inconsistency of assuming for the unknown variances both (a) fiducial distributions in accordance with the samples observed, and (b) values fixed from sample to sample.

The alternative test of significance proposed involves, when the variance ratio of the two populations sampled is unknown, the choice by lot between the value T, used in Behrens' test, and a second value T', which reverses the order of significance of different possible sets of observations. High values of T' are not, therefore, by themselves evidence of inequality of the means."

I submit that the second paragraph quoted above shows, without further argument, that I rejected Bartlett's proposed test of significance, and therefore that I did not confirm his opinion that it provided "an exact inference of fiducial type." Whether my reasons for doing so were strong or weak is, of course, another matter.

What may have led Bartlett to adopt his test of significance is its formal similarity to one appropriate to a different problem. In 1908 "Student" in his now celebrated paper on "The probable error of a mean" [6] applied his solution to what are known as paired observations. Two treatments A and B are applied each to one of a number of pairs of plots, or other experimental units, the members of each pair being chosen to be in other respects closely comparable, although the circumstances of the different pairs are not necessarily closely alike. In order to allow for any, possibly large, variations in the conditions prevailing in the different pairs, attention is confined to the difference, having regard to sign, supplied by each pair.

Thus, if pairs of measurements a_1 , b_1 , a_2 , b_2 , $\cdots$ are obtained, we may write

$$d_k = a_k - b_k ,$$

and test the hypothesis that the differences d are a normal sample having zero mean. This hypothesis will be true if a_k and b_k are distributed, by experimental error, in normal distributions having the same mean, even though this mean is not the same for different pairs. It will be true if the variances of a and b from the hypothetical mean of the pair are unequal, provided these variances are the same

from pair to pair. These are the reasons which make the hypothesis that d is normally distributed about zero appropriate for testing the differential effect of the treatments.

If only two pairs are used, "Student's" test reduces to

$$t = \frac{d_1 + d_2}{d_1 - d_2}.$$

There is one degree of freedom, so that t is distributed in Cauchy's distribution

$$df = \frac{1}{\pi} \frac{dt}{1 + t^2}.$$

If, now, the symbols have a different meaning, so that a_1 and a_2 are a sample of two from a single normal distribution, and b_1 and b_2 a second sample from a different population, having by hypothesis an independent variance, Behrens' problem (limited for comparison with Bartlett to samples of 2) is to test whether the two populations can be regarded as having the same mean, or whether there is reason to regard the means also as being different. Note that the pairs 1 and 2 are not supposed to differ in treatment or situation. The difference $a_1 - a_2$ is not to be ascribed partly to differences between the hypothetical means of these pairs, but wholly to the error variance of the observations a, about which it is the only source of information; the like is true of the difference $b_1 - b_2$. The sign of these two differences is arbitrary, only their positive values concern our problem. There is no real correspondence between the suffices assigned to the two pairs of letters. They could be interchanged for a, and not for b, without affecting the problem.

Behrens' test reduces for this case to taking

$$T = \frac{a_1 + a_2 - b_1 - b_2}{|a_1 - a_2| + |b_1 - b_2|},$$

using for the probability function, "Student's" distribution for one degree of freedom. Bartlett's test involves choosing at random between T and T', where

$$T' = \frac{a_1 + a_2 - b_1 - b_2}{||a_1 - a_2| - |b_1 - b_2||}.$$

It will be noticed that, if $|b_1 - b_2| < |a_1 - a_2|$, and if, keeping $b_1 + b_2$ constant, $|b_1 - b_2|$ is *increased*, a change which must make us suspect larger errors, and therefore a lower significance, the value of the difference $|a_1 - a_2| - |b_1 - b_2|$ is continuously diminished, and that of T' continuously increased, without limit. In fact, the probability of exceeding any limit of significance, however high, may be made to exceed 50% by this process. The order of significance of such a series of possible observations is thus reversed. The fact that choosing at random T and T' will give us a quantity which, on the null hypothesis, is distributed in "Student's" distribution is, thus, insufficient to justify its use as a test of significance.

It is also irrelevant, and this may be at the present time the most important point to make, that the sampling distribution of T above is not given by "Student's" distribution, if the populations to which statements of probability refer is supposed to consist of samples taken repeatedly from populations having a fixed variance ratio. Such a supposition, as I noted in the passage quoted above, is inconsistent with the fiducial distributions derived from the samples. Bartlett comes near to discussing this point on p. 136 in [1]. He says:

"While Fisher suggests that this in no way invalidates his fiducial argument, in my view if an inference is to be independent of an unknown parameter, it should in particular be independent of it if we imagine that we are being supplied with pairs of samples, for all of which the ratio has the same value."

In its natural meaning this statement seems to be true. The problem concerns what inferences are legitimate from a unique pair of samples, which supply the data, in the light of the suppositions we entertain about their origin; the legitimacy of such inferences cannot be affected by any supposition as to the origin of other samples which do not appear in the data. Such a population of samples is really extraneous to the discussion. Nor has Bartlett shown that Behrens' inference from a unique pair of samples is so affected. What he seems to rely on is that an aggregate of samples fulfilling the null hypothesis, but drawn from pairs of populations having a fixed variance ratio, will show differences between their means exceeding the limits fixed by the test for significance, with a frequency other than that indicated by the test. This, however, is a circumstance common to all the well known tests of significance, and has been obvious from their very origin.

In "Student's" test for significance, for example, if a sample of n' observations are taken from a population normally distributed about zero, we calculate

$$\bar{x} = \frac{1}{n'} S(x), \qquad n = n' - 1, \qquad s^2 = \frac{1}{n} S(x - \bar{x})^2$$

and count $\bar{x}$ as significant, if

$$\bar{x} > s t_n / \sqrt{n'}$$

where t_n is "Student's" test for n degrees of freedom, corresponding to the level of significance chosen.

However, in repeated samples of n' from a population having a given variance σ^2, it is highly improbable that $\bar{x}$ will exceed the limit assigned with the frequency chosen. The limit it will exceed with this frequency is

$$\sigma t_\infty / \sqrt{n'}$$

which will usually differ from that assigned from the sample. This, however, has not hitherto been considered an adequate reason for calling the test inaccurate, or biased. It is merely a recognition of the fact that, if we did know σ, we could make a better test. Just as, in Behrens' problem, if we knew the relative weights x of the observations in the two samples, we could make a

weighted "Student's" test, and should be wise to do so—if the information were available.

Naturally, it may be said that although the limit of significance assigned to $\bar{x}$ will not be verified in repeated sampling from populations having the same variance, the distribution of t will be so verified. In this respect the distribution of t in "Student's" case is analogous to the simultaneous distribution of t_1 and t_2 in Behrens' case, where

$$t_1 = \frac{\bar{x}_1 - \mu}{s_1}, \qquad t_2 = \frac{\bar{x}_1 - \mu}{s_2},$$

μ is the hypothetical common mean of the two populations, and s_1^2, and s_2^2 are the estimated variances of the means of the two samples. The quantity d which Sukhatmé [7] has conveniently tabulated, in such a way that

$$d\sqrt{s_1^2 + s_2^2}$$

supplies a significance limit for $\bar{x}_1 - \bar{x}_2$, naturally does not possess the property that

$$\bar{x}_1 - \bar{x}_2 > d\sqrt{s_1^2 + s_2^2}$$

with the probability assigned, in a population consisting of pairs of samples from populations having the same variance ratio.

If the populations were fixed, the corresponding limit would be

$$t_\infty\sqrt{\sigma_1^2 + \sigma_2^2},$$

and if the variance ratio were fixed so that w is the weight of $\bar{x}_2$ relative to that of $\bar{x}_1$, it would be

$$t_{n_1+n_2}\sqrt{\frac{(n_1 s_1^2 + w n_2 s_2^2)\left(1 + \frac{1}{w}\right)}{n_1 + n_2}}$$

provided always, if we wish to express ourselves in terms of repeated sampling, that the absolute values of σ_1, or σ_1 were fiducially distributed. Behrens' problem refers to the case in which neither the variances nor their ratio is known, so that the unknown variances, independently, must be given their fiducial distributions.

In this note I have not touched on the logical background of Behrens' test, or the practical conditions on which it is appropriate, since I have recently discussed these more fully [8]. Recently also [9] Yates has given a careful explanation of the basis of the test.

Summary

The statement of Bartlett that the author (Fisher) has confirmed that Bartlett's approach to Behrens' problem provides an exact inference of fiducial type is incorrect. The only exact test appropriate to his problem seems to be that given by Behrens.

REFERENCES

[1] M. S. BARTLETT, "Complete simultaneous fiducial distributions," *Annals of Math. Stat.*, Vol. X(1939), pp. 129–138.

[2] W.-V. BEHRENS, "Ein Beitrag zur Fehlerberchnung bei wenigen Beobachtungen," *Landw. Jb.*, Vol. LXVIII(1929), pp. 807–37.

[3] R. A. FISHER, "The fiducial argument in statistical inference," *Ann. Eugen.*, Vol. VI(1935), pp. 91–98.

[4] M. S. BARTLETT, The information available in small samples. *Proc. Camb. Phil. Soc.*, Vol. XXXII(1936), pp. 560–66.

[5] R. A. FISHER, "On a point raised by M. S. Bartlett on fiducial probability," *Ann. Eugen.*, Vol. VIII(1937), pp. 370–75.

[6] "Student," "The probable error of a mean," *Biometrika*, Vol. VI(1908), pp. 1–25.

[7] P. V. SUKHATME, "On Fisher and Behrens' test of significance for the difference in means of two normal samples," *Sankhyā*, Vol. IV(1938), pp. 39–48.

[8] R. A. FISHER, "Samples with possibly unequal variances," *Ann. Eugen.*, Vol. IX(1939), pp. 174–180.

[9] F. YATES, "An apparent inconsistency arising from tests of significance based on fiducial distributions of unknown parameters," *Proc. Camb. Phil. Soc.* (in press).

UNIVERSITY COLLEGE,
LONDON.

A NOTE ON NEYMAN'S THEORY OF STATISTICAL ESTIMATION[1]

BY SOLOMON KULLBACK

In this note we shall examine a section of a recent paper by Neyman[1] dealing with statistical estimation. Consider the following quotation from that section[2] which deals with the statement of the problem:

"Consider the variables $[x_1, x_2, \cdots, x_n]$ and assume that the form of the probability law $[p(x_1, \cdots, x_n \mid \theta_1, \theta_2, \cdots, \theta_l)]$ is known, that it involves the parameters $\theta_1, \theta_2, \cdots, \theta_l$ which are constant (not random variables), and that the numerical values of these parameters are unknown. It is desired to estimate one of these parameters, say θ_1. By this I shall mean that it is desired to define two functions $\bar{\theta}(E)$ and $\underline{\theta}(E) \leq \bar{\theta}(E)$, determined and single valued at any point E of the sample space, such that if E' is the sample point determined by observation, we can (1) calculate the corresponding values of $\underline{\theta}(E')$ and $\bar{\theta}(E')$ and (2) state that the true value of θ_1, say θ_1^0, is contained within the limits

$$\underline{\theta}(E') \leq \theta_1^0 \leq \bar{\theta}(E') \tag{18}$$

this statement having some intelligible justification on the ground of the theory of probability.

[1] Specifically we refer to J. Neyman "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability," *Phil. Trans. Roy. Soc.*, vol. A236 (1937), pp. 333–380.

[2] J. Neyman, loc. cit., p. 347. The material in brackets are slight alterations of the original text in order that the quotation do not refer to previous matter in the original paper.