162

THE COMPARISON OF SAMPLES WITH POSSIBLY UNEQUAL VARIANCES

Author's Note (CMS 35.173a) *

* Reproduced from "Contributions to Mathematical Statistics" (1950) by permission of John Wiley & Sons, Inc.

In republishing this paper I should like to invite the reader's attention to the first section, in which the logic of the test is discussed. The principles brought to light seem to the author essential to the theory of tests of significance in general, and to have been most unwarrantably ignored in at least one pretentious work on "Testing statistical hypotheses."[1] Practical experimenters have not been seriously influenced by this work, but in mathematical departments, at a time when these were beginning to appreciate the part they might play as guides in the theoretical aspects of experimentation, its influence has been somewhat retrograde.

With respect to the particular problem, first discussed by Behrens,[2] who arrived, I believe, essentially at the right solution, the origin of the controversy may be distinctly recognised. Pearson and Neyman have laid it down axiomatically that the level of significance of a test must be equated to the frequency of a wrong decision "in repeated samples from the same population." This idea was foreign to the development of tests of significance given by the author in 1925,[3] for the experimenter's experience does not consist in repeated samples from the same population, although in simple cases the numerical values are often the same; and it was, I believe, this coincidence of values in simple cases which misled Pearson and Neyman, who were not very familiar with the ideas of "Student" and the author. It was obvious from the first, and particularly emphasised by the present writer, that Behrens' test rejects a smaller proportion of such repeated samples than the proportion specified by the level of significance, for the sufficient reason that the variance ratio of the populations sampled was unknown.

This point was early emphasised (Fisher, 1937)[4] by giving in a simple case the exact formula for the proportion rejected; it is irrelevant to the purpose of the test, for the experimenter is not con-

Annals of Eugenics, 9: 174-180, (1939).

cerned with repeated samples from the same population. The population of cases which concern him is specified by the properties of his sample, not by the functions of an entirely hypothetical population. The objection raised against Behrens' test thus seemed merely irrelevant. Later S. Wilks[5] has stated that he has proved that no test can exist in this problem, satisfying the conditions laid down by Neyman and Pearson. This, one might have thought, would have settled the matter. It is obviously not an objection to a test of significance that it does not satisfy conditions which cannot possibly be satisfied! However, as the point seems still to be disputed on these grounds, ignoring Wilks' note, the opening section of this paper may still serve a useful purpose. Perhaps, also, Professor Wilks may be induced to publish the proof of his statement, and so clarify the nature of the "requirement" which Neyman and Pearson have, apparently inadvertently, introduced.

REFERENCES

[1] J. Neyman and E. S. Pearson, 1932. "Testing statistical hypotheses in relation to probabilities *a priori*." *Proc. Camb. Phil. Soc.*, vol. 29, pp. 492–516.

[2] W. V. Behrens, 1929. "Ein Beitrag zur Fehlerberechnung bei weinige Beobachtungen." *Landw. Jb.*, vol. 68, pp. 807–837.

[3] R. A. Fisher, 1925–46. *Statistical Methods for Research Workers*, Edinburgh, Oliver & Boyd.

[4] R. A. Fisher, 1937. "On a point raised by M. S. Bartlett on fiducial probability." *Ann. Eugen.*, vol. 7, pp. 370–375.

[5] S. S. Wilks, 1940. "On the problem of two samples from normal populations with unequal variances." *Ann. Math. Stat.*, vol. 11, pp. 475–476 (abstract).

THE COMPARISON OF SAMPLES WITH POSSIBLY UNEQUAL VARIANCES

By R. A. FISHER

1. The nature of the problem

For many years, prior to the introduction of exact tests of significance, it was customary, when a number of mean values had been obtained, as in a replicated experiment, each based on two or more independent observations, to calculate independently a standard error for each mean, and thence to obtain a different standard error for each possible comparison to be made.

This procedure, besides being laborious, is open to the objection, in many cases, that the observed estimates of standard errors, ascribed to different treatments, or varieties, do not differ more than would be expected merely from errors of random sampling. When this is the case, it is reasonable to conclude that the greater part of the observed differences is in fact due to random sampling, and that a more precise, as well as a simpler, analysis would be possible by pooling the sums of squares of deviations obtained from different varieties, and using the pooled estimate for all the tests of significance required.

This change (Fisher, 1925–38), which made it possible to make exact tests of significance, had the advantage of giving precision to the null hypothesis, which the tests were required to substantiate, or to discredit. For the null hypothesis is now simply that all treatments or varieties, or those of them chosen for comparison, are equivalent in the circumstances of the test, and in respect of the measurements used. Consequently, the pooling of the estimates of error is now habitual in all experimental trials.

Critics concerned to uphold the older biometrical tradition, misunderstanding the nature of the hypothesis to be tested, argued that such tests were invalid on the ground that the variances of the different varieties were *assumed* to be equal. The equality of the variances is, however, a characteristic of the null hypothesis chosen. This hypothesis is never assumed to be true, and the whole point of the procedure is to give the facts an opportunity of demonstrating its falsity if, in fact, it is not true. It is an hypothesis particularly appropriate to experimental trials, in that, if a treatment has any effect on the variances of the observed values, it must in some circumstances increase them, and in others diminish them; so that any hypothesis involving a difference in variances is only of interest when it is already admitted that the treatment has any relevant effect at all.

The advances of statistical science have consisted largely in the provision of exact tests of significance appropriate to an increasing variety of useful hypotheses, and occasionally, though not characteristically in experimental work, some interest attaches to hypotheses implying that the means of two populations are equal, while their variances are unequal.

At least a theoretical problem of this sort can be framed. The solution has been known for nearly ten years (Behrens, 1929), though it has been obscured by some controversy (Bartlett, 1936), arising, I believe, from a misunderstanding of the nature of the problem. Useful tables of the solution have recently been published (Sukhatme, 1938), and it is the purpose of the present note to clarify the hypothesis of which they furnish the exact test.

In putting forward his test of significance "Student" (1908) specifies that the problem with which he is concerned is that of a *unique* sample. His clear intention in this is to exclude from his discussion all possible suppositions as to the "true" distribution of the variances of the populations which might have been sampled. If such a distribution were supposed known, "Student's" method would be open to criticism and to correction. 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.

This principle is important for our problem, because it might be thought, that in testing the significance of the difference between two means of normal samples, when the hypothetical equality of the variances of the populations from which they are drawn is deleted, it is to be replaced by some supposition, based on previous experience, as to the true ratio of the variances, or as to the distribution of this true ratio. On the contrary, when any such previous experience, sufficiently valid to demand inclusion, exists, I suggest that it should be treated in exactly the same way as the evidence supplied by a unique pair of samples. In this way it will, of course, add to our information and, in consequence, allow of the rejection of the hypothesis that the means are equal, in cases in which such a rejection would otherwise be inadmissible, but its possible existence does not supply any reason for neglecting the problem of a pair of samples regarded as unique.

For the case in which the variances are by hypothesis equal, any difference between the estimated variances is evidence only of sampling error. The element of our hypothesis by which the equality of the variances is replaced, is that the observed ratio between the variances is no evidence that this ratio is in error in one direction, rather than the other. We suppose, indeed, that it will be affected by sampling error, but the increment or decrement in the logarithm of the estimate due to errors of random sampling will be supposed, in the material to which the test is applied, to be distributed exactly as such errors are known to be distributed in general, for estimates based on the same numbers of degrees of freedom, i.e. in the z distribution.

The implication of this supposition is that whereas, supposing the variances σ_1^2 and σ_2^2 equal, the estimate s_1^2 derived from one sample is equally relevant to the estimation of σ_2^2 as to that of σ_1^2, now, when the variances are no longer supposed equal, we specify for exactitude that the value s_1^2 is of no relevance for the estimation of σ_2^2, nor is s_2^2 of relevance for the estimate of σ_1^2. In this precise sense the unknown variances σ_1^2 and σ_2^2 may be spoken of as "independent by hypothesis". Such variances may, of course, be near to equality, or may differ to any possible extent.

In contrasting this hypothesis with that of equality, it is worth noting that, just as the latter is appropriate when the variances of the populations sampled are not exactly equal, but differ by an amount small compared with the errors of sampling, so that hypothesis of independence implies that real differences are to be expected which are large compared with the sampling errors. Evidently, in the same material, we may be more interested to test the hypothesis of equal variances when the samples are small, and the hypothesis of independent variances when the samples are large. Equally, the investigator will be free, without incurring the charge of inconsistency, to test the same body of data from these two contrasted standpoints.

2. Analytic properties of the solution

Let

$$n_1(n_1+1)\,s_1^2 = \overset{n_1+1}{\underset{1}{S}}\,(x-\bar{x})^2,$$

$$n_2(n_2+1)\,s_2^2 = \overset{n_2+1}{\underset{1}{S}}\,(x'-\bar{x}')^2.$$

A statistician who also knew the true variance ratio of the populations would know the true relative weights of the means $\bar{x}$, $\bar{x}'$; let these be as $1:w$.

Then

$$(n_1+1)\,(n_1s_1^2+wn_2s_2^2)$$

would be the sum of n_1+n_2 homogeneous squares, from which, by dividing by n_1+n_2 and n_1+1, the sampling variance of $\bar{x}$ can be estimated. Hence the sampling variance of the difference $\bar{x}-\bar{x}'$ is

$$\frac{1}{n_1+n_2}\left(1+\frac{1}{w}\right)(n_1s_1^2+wn_2s_2^2).$$

If any limit,

$$\bar{x}-\bar{x}' = d\sqrt{(s_1^2+s_2^2)},$$

where d depends on n_1 and n_2, and also on the ratio $s_1:s_2$, were proposed, such a statistician could calculate

$$t^2 = \frac{d^2(n_1+n_2)\,(s_1^2+s_2^2)}{\left(1+\frac{1}{w}\right)(n_1s_1^2+wn_2s_2^2)},$$

and from this value, and the number of degrees of freedom, n_1+n_2, could read the probability that a pair of samples, from populations having the same mean, should give a difference between the observed means greater than the limit proposed.

The inclusion of the sum of squares, $s_1^2+s_2^2$, in the formula above is not arbitrary, but merely conventional, since d is supposed to vary when the ratio $s_1:s_2$ is changed. Any limit of the kind proposed could therefore be put into the form chosen.

The probability obtained by this process clearly involves w, and cannot be ascertained so long as w is unknown. We may, however, suppose that in the material to which the test is to be applied w takes different values in accordance with a known law. The average value of the probability will then be the probability, on repeated trials with varying values of w, that a statistician, knowing for each trial the true relative weight but ignorant of the absolute variability, would find the limit proposed to be exceeded by chance, by the means of samples from populations having in fact the same mean.

If v_1 and v_2 are the population variances, then

$$w=\frac{(n_2+1)\,v_1}{(n_1+1)\,v_2};$$

and, whatever the values of v_1 and v_2 may be, if

$$z=\tfrac{1}{2}\log\frac{s_1^2}{s_2^2 w},$$

then z will be distributed in its familiar distribution with n_1 and n_2 degrees of freedom. Hence fiducially w may be taken to be distributed as is

$$s_1^2/s_2^2e^{2z}.$$

For example, consider the case $n_1=n_2=6$, $s_1=s_2$, for which according to Sukhatme's table, based on Behrens' formula, $d=2{\cdot}435$.

As typical of the variation of w we may take the medians of sixteen ranges of equal frequency, for which P is an odd number over 32; as the case is symmetrical, only 8 values, from 1/32 to 15/32, need be tabulated (Table I). The second column then gives the fraction of the total weight contributed by the less weighty sample, either

$$\frac{w}{w+1}=\frac{1}{e^{2z}+1},\quad\text{or}\quad\frac{1}{w+1}=\frac{e^{2z}}{e^{2z}+1}. \tag{1}$$

Table I. *Frequencies with which the tabulated values of d are exceeded for various possible values of the true relative weight*

P	$\frac{w}{w+1}$	$\frac{t}{d}$	t	%
1/32	0·15906	0·73146	1·7811	5·02
3/32	0·24050	0·85477	2·0814	2·98
5/32	0·29479	0·91190	2·2205	2·33
7/32	0·33927	0·94692	2·3058	1·99
9/32	0·37865	0·97010	2·3622	1·80
11/32	0·41505	0·98546	2·3996	1·68
13/32	0·44966	0·99492	2·4226	1·61
15/32	0·48332	0·99944	2·4336	1·59
				19·00

Knowing w, we can calculate $$\frac{t}{d}=\frac{2\sqrt{w}}{w+1},$$

the values of t are obtained by using Sukhatme's value of d, and those for the percentage falling outside the fiducial limits, from "Student's" (1925) table in *Metron*. The average of the eight values is 2·38 %. Seeing that a finer graduation would doubtless have increased the contribution of the tails, the agreement with 2·50 % is entirely satisfactory.

3. Equivalence with previous solution.

The analytic equivalence of the two approaches is most easily perceived by means of the analysis of variance.

We have to consider the independent variation of t for n_1+n_2 degrees of freedom, and of z for n_1 and n_2 degrees of freedom, and in this double distribution to calculate the total probability that

$$t^2>\frac{d^2(n_1+n_2)(s_1^2+s_2^2)}{(s_1^2+s_2^2e^{2z})(n_1+n_2e^{-2z})}.$$

Now if A, B and C are the sums of squares respectively of 1, n_1 and n_2 homogeneous degrees of freedom,

$$\frac{(n_1+n_2)A}{B+C}=t^2$$

for n_1+n_2 degrees of freedom, while

$$\frac{n_2B}{n_1C}=e^{2z}$$

for n_1 and n_2 degrees of freedom. Consequently, the inequality defining the region of significance may be written in terms of A, B and C as

$$A>\frac{d^2(s_1^2+s_2^2)BC}{(n_1s_1^2C+n_2s_2^2B)},$$

or, more simply, as $$\frac{d^2(s_1^2+s_2^2)}{A}<\frac{n_1s_1^2}{B}+\frac{n_2s_2^2}{C}. \qquad (2)$$

Using trilinear coordinates A, B, C the boundary is a conic through the vertices of the triangle of reference.

We obtain the same analysis of variance, and therefore the same simultaneous distribution of A, B and C by putting

$$B=n_1(n_1+1)s_1^2/v_1,$$
$$C=n_2(n_2+1)s_2^2/v_2,$$
$$A=(s_1t_1-s_2t_2)^2\Big/\left(\frac{v_1}{n_1+1}+\frac{v_2}{n_2+1}\right),$$

where t_1 and t_2 have respectively n_1 and n_2 degrees of freedom, and are distributed independently.

Substituting these values for A, B and C in equation (2), we find

$$(s_1 t_1 - s_2 t_2)^2 > d^2(s_1^2 + s_2^2),$$

which is the inequality given by Fisher (1935) and used by Sukhatme.

The values calculated in Table I show that the reason why d is as high as it is, is that it makes allowance for the possibility that the relative weight of the two means compared differs materially from what the samples indicate. In that example, however, the apparent weights are equal, and to obtain a clear understanding of the test it is worth while to consider a case in which we can distinguish between the effects of two different possibilities, which the experimenter will certainly wish to consider: (*a*) that the true variances are unequal, and (*b*) that their ratio differs from that of the estimates derived from the samples.

If, in the general formula,

$$t^2 = \frac{d^2(n_1+n_2)(s_1^2+s_2^2)}{(1+1/w)(n_1 s_1^2 + w n_2 s_2^2)},$$

we consider first the supposition that the variances are equal, we shall put

$$w = \frac{n_2+1}{n_1+1},$$

so that

$$d^2 = \frac{n_1+n_2+2}{n_1+n_2} \cdot \frac{\{n_1(n_1+1)s_1^2 + n_2(n_2+1)s_2^2\}}{(n_1+1)(n_2+1)(s_1^2+s_2^2)} t^2.$$

Note that d^2 is not in general equal to t^2 in this case, though it is so when $n_1 = n_2$. Thus, if $n_1 = 6$, $n_2 = 8$, the 5 % value of t for 14 degrees of freedom is 2·145. Taking s_1^2/s_2^2 equal successively to 3, 1 and 1/3, we find

$$d = 2{\cdot}033,\ 2{\cdot}109,\ 2{\cdot}320.$$

If, in the second case, it were supposed that the real relative precision of the two means were exactly equal to the apparent relative precision, we should have

$$w = s_1^2/s_2^2,$$

whence

$$t^2 = d^2.$$

If, on the contrary, we allow for the possibility that the apparent relative precision differs from the true precision by sampling errors given by the z distribution, the appropriate values of d are those given in Sukhatme's table, with which we may compare the values obtained above.

It will be seen from this table that the possibility that the true ratio of weights differs materially from what it appears to be, is the major factor in requiring a larger value of d

Table II. *Values of d appropriate to different suppositions*, $n_1 = 6$, $n_2 = 8$

s_1^2/s_2^2	3	1	1/3
Equal variances	2·033	2·109	2·320
Estimated variances	2·145	2·145	2·145
Independent variances	2·398	2·364	2·332

when the samples are small. To take into consideration only the possibility that the true variance ratio is equal to that observed is quite insufficient. When the smaller apparent variance is associated with the smaller number of degrees of freedom, this test may actually diminish the value of d below that obtained for equal variances. The problem of a test of significance for samples with possibly unequal variances has, however, often been conceived as though in this case the only danger to be considered was that the true variances should differ from equality as much as appeared from the estimates. The danger that, owing to random sampling, the estimated ratio should be in error, has not apparently been appreciated.

This may explain why Bartlett (1936) should have thought it could be inferred (for the case $n_1 = n_2$) that the probability of exceeding d must always be greater when $s_1 = s_2$, than when s_1 or $s_2 = 0$. He says: (p. 565)

> An examination of Behrens' complete table ($n_1 = n_2$) might be sufficient to make us suspect its validity, for in all cases the fiducial probability given is less for $s_1/s_2 = 1$ than $s_1/s_2 = 0$ or ∞, whereas given T, we should expect to be more sure that the observed difference is significant if $s_1/s_2 = 1$, since in that case there is evidence that $\sigma_1^2 + \sigma_2^2$ is more efficiently estimated.

Sukhatme's work has now shown that at the 5 % level the facts are the reverse of Bartlett's statement when $n > 5$. It is probable, however, in any case that Bartlett would not now be inclined to press an argument of this sort, for the errors of $s_1^2 + s_2^2$ regarded as an estimate of $\sigma_1^2 + \sigma_2^2$ fail to specify the errors of s_1^2/s_2^2. It would be impossible, without entering exactly into the analysis, to make inferences as to the relative values of d appropriate to different observed ratios. There seems to be no justification for Bartlett's procedure of taking the value of d when s_1, or s_2, is zero as an upper limit for other cases. At the time of writing, however, Bartlett was evidently under the impression that an analytic error of some kind underlay Behrens' formula, and this perhaps made him expect to find some unreasonable feature in the table.

There can be now no doubt that the supposed error was non-existent. Behrens proposed and gave the correct solution of a perfectly definite problem. Opinions may differ as to the occasions in practical research to which this problem is appropriate, but a discussion of this topic cannot be furthered by the suggestion that the numerical results to which his solution leads are inaccurate. It is probable that, at the time he wrote, Bartlett imagined that he had found a better approach to the same problem, but, as has already been shown (Fisher, 1937), the test of significance on which he relied is irrelevant to the work he was discussing.

REFERENCES

W.-V. Behrens (1929). "Ein Beitrag zur Fehlerberechnung bei weinige Beobachtungen." *Landw. Jb.* **68**, 807–37.

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

P. V. Sukhatme (1938). "On Fisher and Behrens' test of significance for the difference in means of two normal samples." *Sankhyā*, **4**, 39–48.

"Student" (1908). "The probable error of a mean." *Biometrika*, **6**, 1–25.

—— (1925). "New tables for testing the significance of observations." *Metron*, **5**, 18–21.

R. A. Fisher (1925–38). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

—— (1935). "The fiducial argument in statistical inference." *Ann. Eugen., Lond.*, **6**, 391–8.

—— (1937). "On a point raised by M. S. Bartlett on fiducial probability." *Ann. Eugen., Lond.*, **7**, 370–5.