THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN TWO MEANS WHEN THE POPULATION VARIANCES ARE UNEQUAL

BY B. L. WELCH, Ph.D.

1. Introduction. Suppose that we have samples of sizes $n_1$ and $n_2$ from populations $\pi_1$ and $\pi_2$ respectively. Let the populations be normal in form, $\pi_1$ having mean and standard deviation $\sigma_1$ and $\pi_2$ having mean and standard deviation $\sigma_2$. Let it be required to test whether $x_1 = x_2$. Two cases may be distinguished: (i) $\sigma_1$ and $\sigma_2$ may be equal or (ii) they may be unequal. In the first case the most appropriate test for the equality of the $x$'s is made by referring the criterion

$$ u = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sum_1 + \sum_2}{(n_1 + n_2 - 2)} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} $$

......(1)

to the $t$ distribution with $f = (n_1 + n_2 - 2)$.* In the second case, if the ratio of the two $\sigma$'s is known, a similar criterion can be used: if, however, this ratio is unknown, no criterion quite so simple is available. A solution of the problem of testing the hypothesis in this instance has been proposed by R. A. Fisher,† using the concept of fiducial distributions. Fisher notes the equivalence of his test to that given previously by W. V. Behrens‡ in 1929. The validity of this test has, however, been questioned by M. S. Bartlett.§ An alternative criterion which has been often employed is

$$ v = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sum_1}{n_1(n_1 - 1)} + \frac{\sum_2}{n_2(n_2 - 1)}}} $$

......(2)

This may be referred to the normal probability table if the samples are large enough, but for small samples it does not yield an exact test and it is not clear how it may best be made to furnish approximations.

It has been pointed out by Fisher that in many practical situations where $u$ is used, the fact that the $\sigma$'s must be equal for the criterion to be distributed as $t$ does not necessarily mean that an assumption of equality is involved. It may

* $\Sigma_1$ denotes the sum of squares of the observations in the first sample from their mean. $\Sigma_2$ similarly for the second sample.
|| It should be noted that, if $n_1 = n_2$, the criteria $u$ and $v$ are identical.
mean that the equality of the $\sigma$'s is being regarded as part of the hypothesis under test. In such situations it may be argued that there is no point in testing whether $z_1 = z_2$ unless we have also $\sigma_1 = \sigma_2$. However, even if the question posed is one of testing whether two normal populations are identical, $u$ will not necessarily be the best criterion to use. $u$ will afford a valid* test, in the sense that it will control satisfactorily the chance of rejecting the hypothesis when it is actually true, but it is only one of many such. The choice of criterion must depend on what sort of departure from the hypothesis under test we are most interested in detecting. $u$ is demonstrably the best criterion when we wish to detect differences in means without attendant differences in standard deviations. It is conceivable, however, that the test based on $u$ may sometimes operate in such a fashion that differences in the standard deviations $\sigma_1$ and $\sigma_2$ may mask differences in the means $z_1$ and $z_2$, with the result that judgments of non-significance may be too frequently made. The investigations in this paper throw some light on this point, although explicitly they are concerned with cases where it is reasonable to test whether $z_1 = z_2$, whatever the ratio of $\sigma_1$ to $\sigma_2$.

In the first place I shall consider the problem—how far is the criterion $u$ valid even when $\sigma_1 \neq \sigma_2$? (That the test is liable to be biased in this instance is generally realized, but the extent of the bias has not hitherto received any detailed discussion.) In the second place I shall consider the validity of testing the hypothesis by referring $v$ to the $t$ distribution with $f = (n_1 + n_2 - 2)$. Finally, I wish to make some observations about the test of Fisher and Behrens, mentioned above.

It is easily seen that $u$ in general is not distributed as $t$. For whereas the square of the standard error of $(\bar{x}_1 - \bar{x}_2)$ is $(\sigma_1^2/n_1 + \sigma_2^2/n_2)$, the quantity under the root in (1) is an unbiased estimate of

$$\frac{(n_1 - 1) \sigma_1^2 + (n_2 - 1) \sigma_2^2}{(n_1 + n_2 - 2)} \left( \frac{1}{n_1} + \frac{1}{n_2} \right).$$

This is equal to $(\sigma_1^2/n_1 + \sigma_2^2/n_2)$ only if $\sigma_1 = \sigma_2$ or $n_1 = n_2$. The criterion $v$ does not suffer from this objection, but its distribution still depends to a certain extent on $\sigma_1/\sigma_2$. The first problem will be to obtain the distributions of $u$ and $v$. The exact distributions will not be derived here, but only certain approximations adequate, I believe, for the purpose in hand.

2. The distributions of $u$ and $v$. When $z_1 = z_2$ we may write

$$(\bar{x}_1 - \bar{x}_2) = \chi' \sqrt{n_1 \sigma_1^2/n_1 + n_2 \sigma_2^2/n_2}; \quad \Sigma_1 = \chi_1^2 \sigma_1^2; \quad \Sigma_2 = \chi_2^2 \sigma_2^2, \quad \ldots \ldots (3)$$

where $\chi^2$, $\chi_1^2$ and $\chi_2^2$ are independently distributed as $\chi^2$ with degrees of freedom 1.

* The term "validity" applied to a test is used throughout this paper in the sense here indicated. The term "unbiased" is also used with the same meaning, which should not be confused with the meaning which J. Neyman and E. S. Pearson have attached to it in recent papers on testing statistical hypotheses.
(n_1 - 1) and (n_2 - 1) respectively. It is therefore possible to write both \( u \) and \( v \) in the form
\[
y = \frac{\chi'}{\sqrt{a\chi_1^2 + b\chi_2^2}} = \frac{\chi'}{\sqrt{w}} \quad \text{(say),} \quad ......(4)
\]
where \( a \) and \( b \) are constants depending on the \( n \)'s and \( \sigma \)'s. \( w \) is always distributed independently of \( \chi' \) and, when \( a = b \) or when either \( a \) or \( b \) is zero, \( w \) is distributed as \( \chi^2 \) multiplied by some constant. In these cases \( y \) will be distributed as \( t \) multiplied by some constant. For other values of \( a \) and \( b \) the distribution is not so simple, but a useful approximation may be obtained. Following the lines adopted in a previous paper* let us first approximate to the distribution of \( w \) by the Pearsonian Type III Curve
\[
p(w) = \frac{1}{(2g)^{1/2} \Gamma (1/2)} \cdot \frac{w^{|w|}}{w^{2a}}, \quad .....(5)
\]
where \( f \) and \( g \) are so chosen that the first two moments of the curve agree with the true moments of \( w \). For the curve we have
\[
\text{mean} = af; \quad \mu_2 = 2g^2f
\]
and for the true moments of \( w \)
\[
\text{mean} = (af_1 + bf_2); \quad \mu_2 = 2(a^2f_1 + b^2f_2),
\]
\( f_1 \) and \( f_2 \) now being written instead of \( (n_1 - 1) \) and \( (n_2 - 1) \). Hence
\[
g = \frac{a^2f_1 + b^2f_2}{af_1 + bf_2}; \quad f = \frac{(af_1 + bf_2)^2}{a^2f_1 + b^2f_2}. \quad .....(6)
\]
With these values of \( f \) and \( g \) we see from (5) that \( w/g \) is distributed approximately as \( \chi^2 \) with \( f \) degrees of freedom. Hence \( \chi' \) divided by \( \sqrt{w/g} \) is distributed approximately as \( t \). Therefore from (4) we have \( y = ct_f \), where
\[
c = \frac{1}{\sqrt{fg}} = \frac{1}{\sqrt{af_1 + bf_2}}, \quad .....(7)
\]
and \( t_f \) is distributed approximately as \( t \) with degrees of freedom \( f \)† given by (6). This approximation is sufficiently close for the purpose of the comparisons made in this paper‡ and it will be used throughout. The term "approximation" will be omitted.

From (1) and (3) it is seen that \( u \) is of the form (4), where
\[
a = \frac{\sigma^2_1\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{(n_1 + n_2 - 2)(\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2})}; \quad b = \frac{\sigma^2_2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{(n_1 + n_2 - 2)(\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2})}.
\]

† \( f \) is, of course, not now necessarily an integral number of degrees of freedom. It is simply a term in a mathematical approximation. This approximation is of the same form as a true \( t \) distribution and hence we may regard \( f \) as effectively a number of degrees of freedom.
‡ For some further discussion of the adequacy of this approximation, see a Note by Miss E. Tanburn at the end of this paper.
Hence from (6) and (7) we have \( u = ct_f \), where

\[
f = \left( \frac{n_1 - 1}{n_1} \sigma_1^2 + \frac{n_2 - 1}{n_2} \sigma_2^2 \right) \cdot \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \left( \frac{n_1 + n_2 - 2}{n_1 + n_2 - 1} \right) \left( \frac{n_1}{n_1} + \frac{n_2}{n_2} \right),
\]

\[
c = \sqrt{\left( \frac{n_1 + n_2 - 2}{n_1 + n_2 - 1} \right) \left( \frac{n_1}{n_1} + \frac{n_2}{n_2} \right)}. \quad \ldots (8)
\]

Similarly for \( v \) we have

\[
a = \frac{n_1 (n_1 - 1)}{\left( \frac{\sigma_1^2 + \sigma_2^2}{n_1} \right) n_2}, \quad b = \frac{n_2 (n_2 - 1)}{\left( \frac{\sigma_1^2 + \sigma_2^2}{n_2} \right) n_1},
\]

and we can write \( v = ct_f \), where

\[
f = \frac{\left( \frac{\sigma_1^2 + \sigma_2^2}{n_1} \right)^2}{\frac{\sigma_1^4}{n_1 (n_1 - 1)} + \frac{\sigma_2^4}{n_2 (n_2 - 1)}}, \quad c = 1. \quad \ldots \ldots (9)
\]

3. The validity of the criterion \( u \). Suppose that \( u \) is being used to test the hypothesis that \( \alpha_1 = \alpha_2 \) and that the risk of rejecting the hypothesis when true is to be fixed at some prescribed level \( \epsilon \). If it can be assumed that \( \sigma_1 = \sigma_2 \), then from \( t \) tables with \( (n_1 + n_2 - 2) \) degrees of freedom it is possible to choose \( u_0 \) such that the chance \( P(|u| > u_0) = \epsilon \). If \( u_0 \) is so chosen, but it happens that \( \sigma_1 + \sigma_2 \), then the test, which consists in rejecting the hypothesis when \( |u| > u_0 \), will be biased. We shall have

\[
P(|u| > u_0) = P(|ct_f| > u_0) = P\left( |t_1| > \frac{u_0}{c} \right), \quad \ldots \ldots (10)
\]

where \( c \) and \( f \) are given by (8). Owing to the connection between the \( t \) distribution and the Beta-function it can be shown that

\[
P(|t_1| > t_0) = I_{t_0} \left( \frac{1}{2}, \frac{1}{2} \right),
\]

where

\[
I_{t_0}(p, q) = \frac{1}{B(p, q)} \int_0^{z_0} z^{p-1} (1 - z)^{q-1} dz.
\]

Hence from (10)

\[
P(|u| > u_0) = I_{z_0} \left( \frac{1}{2}, \frac{1}{2} \right), \quad \ldots \ldots (11)
\]

where

\[
z_0 = \left( \frac{u_0^2}{f + \frac{u_0^2}{c^2}} \right). \quad \ldots \ldots (12)
\]

Since for given sample sizes, \( c \) and \( f \) depend only on the ratio \( \theta = \sigma_1^2 / \sigma_2^2 \), it is possible from (11) and (12) to obtain for any \( \theta \) the chance of rejection of the hypothesis when it is true. This dependence on \( \theta \) is best illustrated by taking particular examples.
Difference Between Two Means

Example I. Let \( n_1 = n_2 = 10 \). Suppose the chance of rejection \( \epsilon \) is to be fixed at 0-05. The value of \( u_0 \) appropriate when \( \theta = 1 \) is found to be 2-101. In this case, as always when the \( n \)'s are equal, \( c \) is unity. \( f \) is \( 9 (\theta + 1)^2/(\theta^2 + 1) \). The values of \( P (|u| > 2-101) \) for different \( \theta \) were obtained from (11), using the Incomplete Beta-function Tables,* and are plotted in Fig. 1 as curve (a). For convenience \( \theta \) is on a logarithmic scale. It is seen that \( P \) always lies between 0-05 and 0-065, the latter value being attained when the variation in one of the populations is zero. The test is therefore never very seriously biased.

\[ \theta = \sigma_1^2/\sigma_2^2 \] (logarithmic scale)

Fig. 1. Probability of rejection of hypothesis \( \alpha_1 = \alpha_2 \) when true plotted against \( \theta \). (a) \( n_1 = n_2 = 10 \). \( P (|u| > 2-101) \); (b) \( n_1 = 5, n_2 = 15, P (|u| > 2-101) \); (c) \( n_1 = 5, n_2 = 15, P (|v| > 2-101) \).

Example II. Let \( n_1 = 5, n_2 = 15, \epsilon = 0-05 \). In this case \((n_1 + n_2 - 2)\) is 18 as before and \( u_0 = 2-101 \). (8) gives

\[ f = \frac{(4\theta + 14)^2}{4(4\theta + 14)}; \quad c^2 = \frac{18(3\theta + 1)}{4(4\theta + 14)}. \]

\( P (|u| > 2-101) \) is plotted against \( \theta \) in Fig. 1 (b). It will now be seen that \( P \) varies from 0-0024 when \( \theta = 0 \) to 0-05 when \( \theta = 1 \) and then to 0-313 when \( \theta = \infty \). There is therefore the possibility of a considerable bias in the test. The significance of the difference between the two means will tend to be under-estimated when \( \sigma_1 < \sigma_2 \).

and overestimated when \( \sigma_1 > \sigma_2 \). The reason for this is not so much that \( f \) may differ from 18, but that \( c \) can differ considerably from unity. In general the greater the disparity between the \( n \)'s the more likely is this \( c \) factor to bias the test. For equal-sized samples, except perhaps when they are as small as two, the test is never very much biased, whatever \( \theta \).

4. The validity of the criterion \( v \). The validity of the procedure of testing the hypothesis by referring the criterion \( v \) to the \( t \) distribution with \((n_1 + n_2 - 2)\) degrees of freedom may be investigated in the same manner. When the \( n \)'s are equal there is no need for separate discussion as \( u \) and \( v \) are then identical. Let us consider the case \( n_1 = 5, n_2 = 15 \). We find

\[
f = \frac{28(3\theta + 1)^2}{(63\theta^2 + 2)}; \quad c = 1.
\]

\( P(|v| > 2.101) \), obtained from an equation similar to (11), is plotted against \( \theta \) in Fig. 1 (c). As \( \theta \) increases from 0 to 2/21 \( P \) decreases from 0.054 to 0.050 and then increases again to 0.104 at \( \theta = \infty \). It is seen that the test formulated in this way is only unbiased when \( \theta = 2/21 \). There is not, however, the possibility of so large a bias as occurs for some values of \( \theta \) when using the \( u \) criterion. The reason of course is that \( c \) is always unity, this being guaranteed by the fact that the expectation of the square of the denominator of \( v \) is \((\sigma_1^2/n_1 + \sigma_2^2/n_2)\). Bias is due solely to \( f \) being in general less than 18. When \( \theta = 0 \), \( f \) is only 14. \( f \) increases to 18 at \( \theta = 2/21 \) and then decreases to 4 at \( \theta = \infty \). When the smaller sample comes from the more variable population the effective number of degrees of freedom \( f \) is liable to be much smaller than \((n_1 + n_2 - 2)\). Even when \( \theta = 1 \) the effective degrees of freedom in the present case are 6.89, as against 18 for \( u \). If it is known that \( \sigma_1 = \sigma_2 \), then there can be no doubt that \( u \) is a better, more sensitive, \( \ast \) criterion than \( v \). If, however, there exists the possibility that \( \sigma_1 \) and \( \sigma_2 \) differ, then \( u \) may give very misleading results and it will be safer to use \( v \).\( \dagger \)

5. The comparison of regressions. It has been found that a criterion based on an estimate of an assumed common variance may lead to a biased test if \( n_1 \) and \( n_2 \) are different. In the majority of cases it is possible to arrange that \( n_1 \) and \( n_2 \) are equal or almost so, and hence, practically, serious errors of the kind discussed in the previous sections will not often occur. But there is a more general class of problem where the assumption of equal variances may lead to trouble. An instance is afforded by the test for the equality of linear regression coefficients. Here the usual criterion is

\[
u = \frac{(b_1 - b_2)}{\sqrt{\left(\frac{\sum (2) + \sum (2)}{n_1 + n_2 - 4}\right) \left(\frac{1}{\sum (x - x_1)^2} + \frac{1}{\sum (x - x_2)^2}\right) }}
\]

\( \ast \) For further discussion of what is meant here by “sensitivity”, see § 6.
\( \dagger \) Fig. 1 clearly shows that \( \theta \) need not differ much from unity before \( v \) becomes less biased than \( u \). (This refers of course to the particular sample sizes \( n_1 = 5, n_2 = 15 \).)
where $b_1$ and $b_2$ are sample regressions and $\Sigma_1$ and $\Sigma_2$ are now sums of squares from the fitted regression straight lines. Considerations similar to the above show that even if the sample sizes are equal, unless $\Sigma (x - \bar{x}_1)^2$ and $\Sigma (x - \bar{x}_2)^2$ are also equal, the test is biased when the residual variances about the two population regression lines are not the same.

More generally suppose that we have a situation where the samples yield independent statistics $T_1$, $\Sigma_1$, $T_2$, $\Sigma_2$. Let $T_1$ be normally distributed about $\alpha_1$ with standard deviation $\sqrt{\Sigma_1} \sigma_1$ and $\Sigma_1$ be distributed as $\chi^2_{\alpha_1} \sigma_1^2$ with $f_1$ degrees of freedom. Let $T_2$ be normally distributed about $\alpha_2$ with standard deviation $\sqrt{\Sigma_2} \sigma_2$ and $\Sigma_2$ be distributed as $\chi^2_{\alpha_2} \sigma_2^2$ with $f_2$ degrees of freedom. To test whether $\alpha_1 = \alpha_2$ we may use the criterion

$$u = \frac{(T_1 - T_2)}{\sqrt{\Sigma_1 + \Sigma_2}} \frac{1}{V_1 + V_2}$$

This is appropriate if $\theta = 1$. Otherwise $u$ is approximately distributed as $ct_f$, where

$$f = \frac{(f_1 \theta + f_2)^2}{(f_1 \theta^2 + f_2)}; \quad c^2 = \frac{(f_1 + f_2)(V_1 \theta + V_2)}{(f_1 \theta + f_2)(V_1 + V_2)}.$$ 

The effective number of degrees of freedom is $(f_1 + f_2)$ only when $\theta = 1$. The chief cause of bias, however, is likely to arise from the $c^2$ factor. When $\theta = 0$, $c^2$ is $V_2(f_1 + f_2)/f_2(V_1 + V_2)$ and when $\theta = \infty$, $c^2$ is $V_1(f_1 + f_2)/f_1(V_1 + V_2)$. $c^2$ increases or decreases steadily between these limits according as $V_1/f_1$ is greater or less than $V_2/f_2$. $c^2$ is only uniformly equal to unity if $V_1$ and $V_2$ are inversely proportional to $f_1$ and $f_2$. In the simple regression case $f_1$ and $f_2$ are $(n_1 - 2)$ and $(n_2 - 2)$. $V_1$ and $V_2$ are $1/\Sigma (x - \bar{x}_1)^2$ and $1/\Sigma (x - \bar{x}_2)^2$. When $n_1 = n_2$, $c^2$ is uniformly unity for all $\theta$ only if $\Sigma (x - \bar{x}_1)^2 = \Sigma (x - \bar{x}_2)^2$.

The alternative criterion which makes use of separate estimates of $\sigma_1^2$ and $\sigma_2^2$ is

$$v = \frac{(T_1 - T_2)}{\sqrt{V_1 \Sigma_1 + V_2 \Sigma_2}}.$$ 

This leads to

$$f = \frac{(V_1 \theta + V_2)^2}{(V_1 \theta^2 + V_2)}; \quad c = 1.$$ 

Any bias is now due only to the effective number of degrees of freedom which is never less than the smaller of $f_1$ and $f_2$. It is clear that in certain situations where a criterion of the $u$ type is customarily used, the condition $\theta = 1$ needs to be satisfied very stringently. A criterion of the $v$ type will be much safer.

6. **Choice of effective number of degrees of freedom for $v$.** The question remains—if $v$ is used, what is the best value to take for $f$? In the above discussion the
consequences of referring $v$ to $t$ tables with $(f_1 + f_2)$ degrees have been considered. It was seen that this was absolutely valid only if $\theta$ had a particular value. (In the example of Fig. 1 this was $\theta = 2/21$. In general it is $\theta = V_2 f_1 / V_1 f_2$.) When there is strong a priori reason for believing that $\theta$ is in the neighbourhood of a certain value, then it will be better to take $f$ obtained by substituting this value in (13). For instance suppose that we have reason to believe that $\theta \approx 1$. Then, for the $v$ test, it will be preferable to take

$$f = \frac{(V_1 + V_2)^2}{\left(\frac{V_1^2}{f_1} + \frac{V_2^2}{f_2}\right)}.$$  

......(14)

In the example of section 4 this gives $f = 6.89$ and the corresponding critical value $\nu_0$ is 2.374. In Fig. 2 a comparison is made for different $\theta$ between the two rules:

![Figure 2](image)

Fig. 2. Probability of rejection of hypothesis $\alpha_1 = \alpha_2$ when true plotted against $\theta$. (a) $n_1 = 5, n_2 = 15$, $P(|u| > 2.101)$; (b) $n_1 = 5, n_2 = 15$, $P(|v| > 2.374)$; (c) $n_1 = 5, n_2 = 15$, $P(|z| > 1.861)$.

(a) reject the hypothesis that $\alpha_1 = \alpha_2$ if $|u| > 2.101$ and (b) reject if $|v| > 2.374$. As arranged, now, both these rules have the property that for $\theta = 1$, the chance of rejection of the hypothesis when true is 0.05.

It may be objected that it is illogical to use the $v$ criterion and at the same time regard it as having effective degrees of freedom $f$ given by substituting $\theta = 1$ in (13). For, if $\theta = 1$, it is known that the $u$ test is better from the point of view of sensitivity, i.e. any real difference ($\alpha_1 - \alpha_2$) will then be detected more frequently
by \( u \) than by \( v \). However in taking the value of \( f \) given by (14) we are not assuming that \( \theta \) is exactly unity, but simply making use of our reasons for believing that it is near unity. The \( v \) test based on this value of \( f \) is biased when \( \theta = 1 \) but the bias is seen to be less than that of the \( u \) test. The small* advantage which \( u \) enjoys with respect to sensitivity when \( \theta = 1 \) is soon offset by this gain of \( v \) in controlling the chance of rejecting the hypothesis that \( \alpha_1 = \alpha_2 \) when it is actually true.

When there is no very precise a priori information about \( \theta \) available it might seem permissible to use the ratio \( \Sigma_1/\Sigma_2 \) from the samples to estimate \( \theta \) and hence \( f \). Complications arise, however, owing to the fact that the distribution of \( v \) is not independent of that of \( \Sigma_1/\Sigma_2 \). A discussion of this point is beyond the scope of this paper.

7. The fiducial test of R. A. Fisher. I have so far considered only two of many criteria which may be proposed to test the hypothesis that \( \alpha_1 = \alpha_2 \). I have moreover been concerned with one aspect only of the tests based upon these two criteria, viz. whether they control satisfactorily the risk of rejecting the hypothesis when it is actually true. A test which does control this risk is termed in the present paper either a "valid" or an "unbiased" test. The special sense in which these terms have been used should be noted, for a valid test is not necessarily a good test nor vice versa. In the present section I propose to discuss the fiducial test suggested by R. A. Fisher from the single point of view only of how far it is valid, in the sense defined above, for all values of the ratio \( \theta = \sigma_1^2/\sigma_2^2 \).

The manner of developing this fiducial test is as follows†: let

\[
\begin{align*}
    s_1 &= \sqrt{\frac{\Sigma_1}{n_1(n_1-1)}}, \\
    s_2 &= \sqrt{\frac{\Sigma_2}{n_2(n_2-1)}}, \\
    d &= (\bar{x}_1 - \bar{x}_2), \\
    \delta &= (\alpha_1 - \alpha_2), \\
    \epsilon &= (\delta - d); \\
    t_1 &= \frac{(\bar{x}_1 - \alpha_1)}{s_1}, \\
    t_2 &= \frac{(\bar{x}_2 - \alpha_2)}{s_2}.
\end{align*}
\]

From (15) we obtain

\[
\epsilon = (\delta - d) = s_2 t_2 - s_1 t_1.
\]

The fiducial distribution of \( \delta \) is taken by Fisher to be the distribution obtained from (16) by treating \( d, s_1 \) and \( s_2 \) formally as fixed and allowing \( t_1 \) and \( t_2 \) to be distributed independently as \( t \) with degrees of freedom \((n_1-1)\) and \((n_2-1)\) respectively.

Now if \( A_1 \) and \( A_2 \) are any constants the distribution of \( (A_2 t_2 - A_1 t_1)/\sqrt{A_1^2 + A_2^3} \) clearly depends only on \( A_1/A_2, n_1 \) and \( n_2 \). Hence we can theoretically determine

* See J. Neyman's paper, "Statistical Problems in Agricultural Experimentation", J. Roy. Statist. Soc. Supplement, II, No. 2 (1935), pp. 130–6. The sensitivity of \( t \) criteria to real population differences was seen to depend on \( f \) in a pronounced fashion only when \( f \) was very small (say < 6). In the present case the increase in sensitivity of \( u \) (with \( f = 18 \)) over \( v \) (with \( f = 6.89 \)) will not be large.

a function $F(A_1/A_2, n_1, n_2, \gamma)$ such that the probability is $\gamma$ that the inequality
\[ |A_2 t_2 - A_1 t_1| > \sqrt{A_2^2 + A_2^2} F(A_1/A_2, n_1, n_2, \gamma) \] ......(17)
is satisfied. The corresponding statement in terms of fiducial probability is that the inequality
\[ |\delta - d| > \sqrt{s_1^2 + s_2^2} F(s_1/s_2, n_1, n_2, \gamma) \] ......(18)
is satisfied with fiducial probability $\gamma$. The corresponding fiducial test of the hypothesis $\delta = 0$, which Fisher has suggested, consists in rejecting the hypothesis if
\[ \left| \frac{d}{\sqrt{s_1^2 + s_2^2}} \right| > F(s_1/s_2, n_1, n_2, \gamma), \] ......(19)
where $\gamma$ is now the level of significance. However, in repeated sampling from fixed populations $\pi_1$ and $\pi_2$ with $\alpha_1 = \alpha_2$, the repeated application of this test will not, in general, lead to rejection in the prescribed proportion, $\gamma$, of cases. Although the inequality (17), which can be written
\[ \left| \frac{A_2 (\bar{x}_2 - \alpha_2) - A_1 (\bar{x}_1 - \alpha_1)}{s_2} \right| > \sqrt{A_2^2 + A_2^2} F(A_1/A_2, n_1, n_2, \gamma) \] ......(20)
is satisfied with probability $\gamma$ whatever constant values are assigned to $A_1$ and $A_2$, this does not imply that the inequality (18) which can be written
\[ \left| \frac{s_2 (\bar{x}_2 - \alpha_2) - s_1 (\bar{x}_1 - \alpha_1)}{s_2} \right| > \sqrt{s_1^2 + s_2^2} F(s_1/s_2, n_1, n_2, \gamma) \] ......(21)
will also be satisfied with the same probability $\gamma$.

It may be noted that in the case $n_1 = n_2 = 2$, Fisher has himself drawn attention to the difference between (20) and (21). As he has shown for that case*, the fiducial test involves the rejection of the hypothesis $\delta = \alpha_1 - \alpha_2 = 0$ at a 5% level of significance when a certain criterion $T$, which is a function of the sample observations only, numerically exceeds 12.7062. He shows also, however, that if we are sampling from fixed normal populations $\pi_1$ and $\pi_2$ for which $\delta = 0$, the probability of $|T|$ exceeding 12.7062 is only equal to 0.05 if $\theta = \sigma_1^2/\sigma_2^2 = 0$ or $\infty$; in general the probability will be less than 0.05.

That the statement (18) may be associated with a specially defined measure of fiducial probability which could be used by the experimenter as a guide in deciding whether to reject the hypothesis $\delta = 0$ is quite possible. But it seems to me important to make clear that the rule of the test involved in (19), if applied to repeated samples taken from fixed normal populations $\pi_1$ and $\pi_2$, would not lead in the long run on a proportion $\gamma$ of occasions to the rejection of the hypothesis $\delta = 0$, when it is true, whatever be the value $\theta$.

8. **Exact tests.** Dr Bartlett has pointed out to me that the exact test which he
gives for the case \( n_1 = n_2 = 2 \), is capable of easy generalization. For instance, if
\( n_1 = n_2 = n > 2 \), let \( l_{11}, l_{12}, \ldots, l_{1,n-1} \) be \( n - 1 \) linear functions of the observations in
the first sample; let the \( l \)'s be orthogonal to each other and to \( \bar{x}_1 \): further, let them
all have expectation zero and standard deviation \( \sigma_1 \). Linear functions satisfying
these conditions can always be defined. Similarly define \( l_{21}, l_{22}, \ldots, l_{2,n-1} \) for the
second sample, these having standard deviation \( \sigma_2 \). Then \( \sqrt{n (\bar{x}_1 - \bar{x}_2)} \) divided by
\[
\sqrt{\frac{\sum_{i=1}^{n-1} (l_{1i} + l_{2i})^2}{(n-1)}}
\]
will be a criterion distributed as \( t \) with \( n - 1 \) degrees
of freedom, whatever the ratio \( \sigma_1^2/\sigma_2^2 \). Clearly a like test can be evolved when
\( n_1 \neq n_2 \), the degrees of freedom of the corresponding \( t \), then being one less than the
smaller of \( n_1 \) and \( n_2 \). Bartlett would not advocate the use of this test in practice
for the reason that it is not more efficient than using an inexact test based on \( v \), and
expressing the significance level of the sample as lying between two limits. The
number of degrees of freedom for the criterion defined above is \( n_1 - 1 \) (if \( n_1 < n_2 \)),
whereas the effective number of degrees of freedom for \( v \) is never less than \( n_1 - 1 \)
and may be as much as \( n_1 + n_2 - 2 \).

While on the subject of exact tests, it is of interest to note that other criteria
of the form \( (\bar{x}_1 - \bar{x}_2)/\sqrt{d \Sigma_1 + e \Sigma_2} \) may be less dependent on \( \sigma_1^2/\sigma_2^2 \) even than \( v \). For
instance, if \( n_1 \) and \( n_2 \) are both \( > 3 \) we might expect
\[
z = (\bar{x}_1 - \bar{x}_2)/\sqrt{\frac{\Sigma_1}{n_1 (n_1 - 3)} + \frac{\Sigma_2}{n_2 (n_2 - 3)}}
\]
to be such a criterion. The reason for taking these particular values of \( d \) and \( e \) is
that they give to \( \sigma_2^2 \) the same value both when \( \theta = \sigma_1^2/\sigma_2^2 = 0 \) and when \( \theta = \infty \). The
curve (c) in Fig. 2 shows the dependence of \( z \) on \( \theta \). Arranging the test so that
the probability of rejection of the hypothesis \( x_1 = x_2 \) is 0.05 when \( \theta = 1 \), it is seen that
no matter what \( \theta \), the probability of rejection departs from 0.05 less for \( z \) than
for either of the criteria \( u \) and \( v \). It is not proposed in the present paper to discuss
whether tests such as these are of practical value.

9. **Summary.** Three tests of the hypothesis that the means of two normal
populations are equal have been considered in some detail. The object has been to
study how closely each of these controls the risk of rejecting the hypothesis when
it is actually true. None of the tests was exact in the sense that it would control
this risk precisely, whatever the unknown ratio \( \theta \) of the variances of the two
populations.

The first criterion \( u \), which is the best criterion when it is known that \( \theta \) is
unity, can under certain circumstances be seriously biased when \( \theta \neq 1 \).

* By an exact test is meant one depending on a known probability distribution; that is, independent of irrelevant unknown parameters (e.g. in the present case independent of \( \theta = \sigma_1^2/\sigma_2^2 \)). See, for instance, M. S. Bartlett, *Proc. Roy. Soc. A*, CLX (1937), p. 271.
The second criterion \( v \), which employs separate estimates of the unknown variances of the two populations, was seen to be very much less liable to bias. Unless therefore it is definitely known that \( \theta = 1 \), the general use of \( v \) rather than \( u \) is worth serious consideration.

I agree with M. S. Bartlett’s criticism of the third test, which has been put forward from considerations of fiducial probability. The bias of this test depends on \( \theta \), but I have not considered the relationship in any detail.

**NOTE ON AN APPROXIMATION USED BY B. L. WELCH**

**BY ELIZABETH TANBURN, B.A.**

In the preceding paper on the “Significance of the Difference between Two Means”, B. L. Welch has considered two criteria, viz.

\[
\begin{align*}
    u &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\Sigma_1 + \Sigma_2}{(n_1 + n_2 - 2) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}}, \\
    v &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\Sigma_1}{n_1 (n_1 - 1)} + \frac{\Sigma_2}{n_2 (n_2 - 1)}}}.
\end{align*}
\]

He discusses the distribution of these in the case where the means \( \mu_1 \) and \( \mu_2 \) of the normal populations sampled are equal, but where the standard deviations \( \sigma_1 \) and \( \sigma_2 \) are not necessarily equal. He shows that \( u \) and \( v \) will be distributed approximately as \( ct_f \), where \( c \) is a constant and \( t_f \) is distributed as “Student’s” \( t \) having \( f \) degrees of freedom; \( c \) and \( f \) which are functions of \( n_1, n_2 \) and \( \theta = \sigma_1^2/\sigma_2^2 \) are given by equations (8) and (9) of p. 353 above.

The present writer has been studying the same problem both theoretically and also by means of practical sampling experiments and the results of this investigation will shortly be published. It may be of interest, however, to note here some points which have bearing on the approximation Welch has used. The approximation was made by fitting the quantities under the square roots in the criteria by Pearson Type III curves. The fitting was performed by making the Type III curves have the correct first two moments. This was a convenient method, but, of course, not the only one. We might for instance write \( u = ct_f \) and choose \( c \) and \( f \) so that the \( \mu_2 \) and \( \beta_2 \) of \( ct_f \) are the same as the true \( \mu_2 \) and \( \beta_2 \) of \( u \), in other words represent \( u \) (and \( v \)) by a Pearson Type VII curve having the correct 2nd and 4th moments. In general, the objection to this is the more complicated form of the moments of \( u \). (Also for very small samples these moments become infinite.) I have, however, obtained \( \mu_2 \) and \( \beta_2 \) for \( u \) and \( v \) in one particular instance, viz. \( n_1 = 5, n_2 = 15, \theta = \sigma_1^2/\sigma_2^2 = 0.25 \). They are given in the first line of Table I. Let us now compare them with the moments of Welch’s approximation obtained by taking \( u \) (or \( v \)) = \( ct_f \). For this we have

\[
\mu_2 = \frac{cf^2}{(f-2)}, \quad \beta_2 = \frac{3(f-2)}{(f-4)}.
\]
Difference Between Two Means

TABLE I

<table>
<thead>
<tr>
<th>True moments</th>
<th>$\mu_2(u)$</th>
<th>$\beta_2(u)$</th>
<th>$\mu_2(v)$</th>
<th>$\beta_2(v)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moments of Welch’s approximation</td>
<td>0.60022</td>
<td>3.493</td>
<td>1.512</td>
<td>3.503</td>
</tr>
<tr>
<td></td>
<td>0.6011</td>
<td>3.509</td>
<td>1.608</td>
<td>3.575</td>
</tr>
</tbody>
</table>

For $u$ equation (8) gives $f = 15.79$, $c^2 = 0.5250$. For $v$ equation (9) gives $f = 14.44$, $c^2 = 1$. Substituting these we obtain the values given in the second line of Table I. The agreement with the true values is close enough to indicate that no great difference in the values of $c$ and $f$ would have occurred if Welch had used the moment method to represent $u$ (or $v$) by a Type VII (“Student”) curve rather than to represent the square of the denominator of $u$ (or $v$) by a Type III ($\chi^2$) curve.

A further comparison, which is worth making, is that between the approximate theoretical distribution of $u$ and the actual distribution of 500 values of $u$ which were obtained in a sampling experiment using Tippett’s Random Numbers ($n_1 = 5$, $n_2 = 15$, $\theta = 0.25$ as before). In the second line of Table II are shown the numbers of these sample $u$'s whose absolute values lay between the limits given

TABLE II

| $|u|$ | 0.000–0.362 | 0.362–0.500 | 0.500–0.627 | 0.627–0.777 | 0.777–0.969 | 0.969–1.266 | 1.266–1.538 | 1.538–1.874 | >1.874 |
|-----|-------------|-------------|-------------|------------|------------|------------|------------|------------|--------|
|      | Frequency of samples | 252 | 50 | 57 | 44 | 53 | 26 | 12 | 1 | 5 |
|      | Approx. theoretical chances | 0.50 | 0.10 | 0.10 | 0.10 | 0.10 | 0.05 | 0.03 | 0.01 | 0.01 |
|      | Approx. expectations | 250 | 50 | 50 | 50 | 25 | 15 | 5 | 5 |

in the first line. These limits are so chosen that, if the representation, $u = cf$, were exact ($f$ being 15.79 and $c^2$ being 0.5250), then the true theoretical chances of $u$ falling in the ranges would be those given in the third line of the table. The corresponding expectations for 500 samples are given in the last line. The sampling results are seen to agree very well with the theoretical distribution as it has been approximated.