A HIGH DIMENSIONAL TWO SAMPLE SIGNIFICANCE TEST[1]

By A. P. Dempster

Bell Telephone Laboratories[2], Murray Hill, New Jersey

0. Summary. The classical multivariate 2 sample significance test based on Hotelling's T^2 is undefined when the number k of variables exceeds the number of within sample degrees of freedom available for estimation of variances and covariances. Addition of an a priori Euclidean metric to the affine k-space assumed by the classical method leads to an alternative approach to the same problem. A test statistic F which is the ratio of 2 mean square distances is proposed and 3 methods of attaching a significance level to F are described. The third method is considered in detail and leads to a "non-exact" significance test where the null hypothesis distribution of F depends, in approximation, on a single unknown parameter r for which an estimate must be substituted. Approximate distribution theory leads to 2 independent estimates of r based on nearly sufficient statistics and these may be combined to yield a single estimate. A test of F nominally at the 5% level but based on an estimate of r rather than r itself has a true significance level which is a function of r. This function is investigated and shown to be quite near 5%. The sensitivity of the test to a parameter measuring statistical distance between population means is discussed and it is shown that arbitrarily small differences in each individual variable can result in a detectable overall difference provided the number of variables (or, more precisely, r) can be made sufficiently large. This sensitivity discussion has stated implications for the a priori choice of metric in k-space. Finally a geometrical description of the case of large r is presented.

1. Introduction. The statistical problem here treated is that of significance testing for the difference of the means of 2 k-variate populations which may be assumed to have the same structure of variances and covariances, the test being based on a sample from each population with sample sizes denoted by n_1 and n_2. It is intended to provide a method applicable to data where the number k of characteristics measured on each individual is large but where the number of individuals measured may be quite small. The usual method of classical multivariate statistics encounters a mathematical barrier and becomes inapplicable when $k > n_1 + n_2 - 2$, but certainly the need has arisen in applied statistical work for techniques handling small samples of highly described individuals.

The classical method has 2 equivalent formulations in terms of the T^2 statistic of Hotelling [2] or the best linear discriminator of Fisher [3]. For this method the

Received July 8, 1957; revised June 27, 1958.

[1] Most of the material presented here is also contained in the author's PhD thesis [1] at Princeton University. His work in Princeton was supported principally by the National Research Council of Canada.

[2] Now at Harvard University.

space of the k characteristics is thought of as k-dimensional affine space and needs no further structure: the method is invariant over the choice of any k linear combinations of full rank of the given k variables to be used in place of the given variables. The 2 populations are assumed to be probability distributions over affine k-space and the samples constitute $n_1 + n_2$ points of this space. In the formulation of [3] the sample points are projected along a family of parallel $(k - 1)$-dimensional hyperplanes onto any line, the family being chosen so that the one-dimensional Student's t for the 2 samples is maximized. This $t_{\max}$ is then used to test the significance of the difference in population means. However, if $k > n_1 + n_2 - 2$, a family of $(k - 1)$-dimensional hyperplanes can be chosen which projects the points into 2 points, one for each sample. Then $t_{\max} = \infty$ regardless of the populations and so is useless as a test-statistic. In the formulation of [2] the samples are used to define a Euclidean metric in the affine k-space and the test-statistic is the distance between the 2 sample means in this metric. This metric is based on the variation of the samples about their means, and if the samples are shifted to have a common mean point and $k > n_1 + n_2 - 2$ the variation spans only a subspace of $n_1 + n_2 - 2$ dimensions. Thus it is not surprising that in this case the method of defining the metric breaks down. Furthermore it is heuristically evident that no metric for a whole affine space can be well-defined from variation taking place in a flat subspace. For these reasons we are forced to give up the classical approach with its elegant mathematical property of affine invariance.

The approach of this paper is based on the observation that, whatever metric is chosen for k-space, the distance between sample means is a statistic which may yield evidence of separation of the populations, and, rather than be preoccupied with a choice of optimum metric from the data, we should try to use a metric determined apart from the data and analyze the information yielded through this metric.

For much of the theory the population distributions will be assumed to be (multivariate) normal.

2. The general method. It is assumed that a Euclidean metric has been assigned to the affine k-space of the k characteristics; that is, k independent linear combinations of the given variables have been chosen which define distance along k mutually orthogonal axes of Euclidean k-space. The metric may be thought of as chosen from a priori knowledge (precise or imprecise) of the joint distributions of the k characteristics, in the hope of roughly sphericalizing these distributions. More detailed remarks on the choice of a metric are to be found in section 5.

Suppose that the 2 population distributions have means denoted by $k \times 1$ vectors ν_1 and ν_2 and common $k \times k$ matrix of variances and covariances denoted by Λ. We are seeking evidence that $\nu_0 = \nu_1 - \nu_2$ is different from zero and are naturally led to consider V_0 the vector joining the sample means. V_0 is an unbiased estimate of ν_0. Having a metric at hand we will try to direct a significance test at the detection of a non-zero length of ν_0 and will use the

length of V_0 in estimating this length. Rejection of the null hypothesis $\nu_0 = 0$ will result from evidence that the length of V_0 is significantly greater than zero.

So far this use of V_0 has been justified mostly on heuristic grounds. It makes sense geometrically. If however we assume that the populations are multivariate normal $N(\nu_1, \Lambda)$ and $N(\nu_2, \Lambda)$ a more mathematical reason may be given. Suppose the $n_1 + n_2$ individuals are regarded as defining a set of orthogonal axes in a Euclidean space of $n_1 + n_2$ dimensions. The space may be regarded as "degree of freedom" (d.f.) space and any set of orthogonal axes defines a set of orthogonal d.f. Such a new set of d.f. may be defined as follows: first choose the d.f. measuring the grand mean of the $n_1 + n_2$ individuals, second choose the d.f. measuring the difference between the means of the 2 samples, and third choose any set of $n_1 + n_2 - 2$ d.f. which together with the first 2 form an orthogonal set. This last set represents "within sample" d.f. Their number $n_1 + n_2 - 2$ will henceforth for convenience be denoted by m. The data, which consists of k points in this $(n_1 + n_2)$-space, can be described by a set of $n_1 + n_2$ $k \times 1$ vectors corresponding to the new d.f. Let U_0 be the vector corresponding to the mean difference d.f. and $U_1, U_2, \cdots, U_m$ be the vectors corresponding to the within sample d.f. It can be easily checked that

$$V_0 = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{\frac{1}{2}} U_0,$$

that $U_1, U_2, \cdots, U_m$ have mean 0, and that $U_0, U_1, \cdots, U_m$ are uncorrelated and each have Λ for matrix of variances and covariances. Finally, assuming normality and defining

$$\xi = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-\frac{1}{2}} \nu_0,$$

it is seen that $U_0, U_1, \cdots, U_m$ are independent, the first being distributed as $N(\xi, \Lambda)$ and the remainder as $N(0, \Lambda)$. With the normality assumption it is clear that $U_0, U_1, \cdots, U_m$ are sufficient for the parameters ν_0 and Λ, for apart from an irrelevant overall translation of both samples the original data can be reconstructed. But since U_0 is the only one of these vectors involving the parameter ν_0 it is natural to choose a property of U_0 alone in testing significance.

Three methods of testing whether or not U_0 is significantly long will be described, but only the third of these will be pursued. The first is the non-parametric randomization test based on the method of Pitman and Welch [4, 5]. For each of the $\binom{n_1 + n_2}{n_1}$ divisions of the $n_1 + n_2$ individuals into 2 groups of n_1 and n_2 there is a corresponding d.f. for group difference and corresponding vector U. Under the null hypothesis that the $n_1 + n_2$ individuals are a sample from one distribution the lengths of all these vectors U have a joint distribution symmetric under permutation of the vectors. Accordingly U_0 is significantly long at level α if the length of U_0 is beyond the $(1 - \alpha)$ point of the sample cumulative distribution of the set of $\binom{n_1 + n_2}{n_1}$ lengths of vectors U. The second method is

the continuous analogue of the first method which comes into play when normal distributions are assumed. Suppose a set of p d.f. are chosen independently at random uniformly with regard to direction in that part of $(n_1 + n_2)$-space orthogonal to the d.f. for the grand mean. The set of p corresponding vectors together with U_0 have, under the null hypothesis of identical normally distributed populations, joint distributions which are again symmetric under permutations so that a significance test may be defined as in the first method. The limiting test as $p \to \infty$ is uniquely defined and may be regarded as the continuous analogue of the Pitman and Welch procedure. For $k = 1$ this amounts to the usual t test, but for general k the distribution associated with the limiting test appears difficult to handle analytically. However the test could be approximated using a suitable p and experimental sampling.

The third method, which is the concern of most of the subsequent discussion, is also based on normal distribution theory. The idea here is to compare the length of U_0 directly against the lengths of $U_1, U_2, \cdots, U_m$, since under the null hypothesis they form a sample of size $m + 1$ from a certain distribution. Define Q_i = squared length of $U_i (i = 0, 1, \cdots, m)$ and

$$F = Q_0 \Big/ \frac{1}{m} \sum_{i=1}^{m} Q_i$$

Then U_0 will be declared significantly long if F is significantly large. If the null hypothesis distribution of F involved no unknown parameters then an exact test could be based on F; since this is not the case a type of "non-exact significance test" will be introduced.

3. Distribution theory. The distributions involved in the non-exact significance test are those of properties of the vectors $U_0, U_1, U_2, \cdots, U_m$, in particular their lengths and angles between pairs of them. We suppose in this section normal distributions and so may deal with a typical vector U distributed as $N(0, \Lambda)$ or a typical sample of such vectors. Under these assumptions Q, the squared length of U, has the distribution of a quadratic form in k normal variables. Since this distribution in precise form involves k parameters, all unknown, we will rely on the well-known [6] approximation which treats Q as distributed as $\mu\chi_r^2$ depending only on 2 unknown parameters μ and r. The parameters μ and r are generally fitted by equating the first 2 moments, and this results in the inequality $r \leqq k$.

This approximation, at least for integral r, corresponds to approximating the distribution of vector U by a spherical normal distribution lying in a flat subspace of dimension r in k-space. Stated more precisely this says that in the metric chosen for k-space there is an orthogonal transformation to coordinates $(y_1, y_2, \cdots, y_k)$ such that the distribution of U is defined by

(i) density $\dfrac{1}{(2\pi)^{r/2}} \exp\left(-\dfrac{1}{2\mu}\sum_{i=1}^{r} y_i^2\right)$ for $y_1, y_2, \cdots y_r$, and

(ii) $y_{r+1}, y_{r+2}, \cdots, y_k$ are zero with probability one.

Having this approximate underlying distribution for U it is possible to define from it approximate distributions for other statistics based on U.

As a first example consider the angle θ between a pair of vectors U and U' independently distributed according to (i) and (ii) above. Due to the spherical symmetry of the distribution in r-space the conditional distribution of θ given U' does not depend on the particular U' so that the distribution of θ is the distribution of the angle between U and any fixed direction e.g.,

$$y_1 = 1, y_2 = y_3 = \cdots = y_k = 0.$$

Thus $\cos^2 \theta$ is distributed as

$$y_1^2/(y_1^2 + y_2^2 + \cdots + y_r^2)$$

i.e. $\cos^2 \theta$ has the β distribution $\beta_{1/2,(r-1)/2}$ defined by density

$$\frac{\Gamma\left(\frac{r}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{r-1}{2}\right)} x^{\frac{1}{2}-1}(1-x)^{(r-1)/2-1}.$$

This will be used as an approximation to the distribution of $\cos^2 \theta$ under the circumstances where $\mu\chi_r^2$ is used as an approximation to the distribution of Q.

Accepting these approximations it is natural to attempt to estimate μ and r. In particular, estimation of r plays a significant role in our non-exact test. The distribution theory leading to estimates of r will now be discussed. The vectors $V_1, V_2, \cdots, V_m$ may be described by the set of their lengths and the set of angles between pairs of them, and under the sphericalizing approximation these 2 sets of random variables are independent of one another. From each of these sets a statistic is defined which contains nearly all the information about r in the set and whose distribution may be approximated by a fast-converging limiting form as $r \to \infty$, namely $[(1/r) + (c/r^2)]\chi_h^2$ where c and h must be determined for each set. This leads to 2 independent estimates of r which may be combined into a single estimate.

Taking the set of lengths define Q_i as the squared length of U_i and consider $Q_i, \cdots, Q_m$ as m independent observations from $\mu\chi_r^2$ with μ and r unknown. The results of this paragraph are found in [7]. The joint density function of $Q_1, Q_2, \cdots, Q_m$ is

$$\left(2^{r/2}\Gamma\left(\frac{r}{2}\right)\mu^{r/2}\right)^{-m}\left(\prod_{i=1}^m Q_i\right)^{(r/2)-1}\exp\left(-\frac{1}{2\mu}\sum_{i=1}^m Q_i\right)$$

so that $\prod_{i=1}^m Q_i$ and $\sum_{i=1}^m Q_i$ are a pair of sufficient statistics for μ and r. It is now natural to look at

$$v = \frac{\prod_{i=1}^m Q_i}{\left(\sum_{i=1}^m Q_i\right)^m}$$

as a statistic not involving μ for the purpose of estimating r. From joint characteristic functions v and $\sum_{i=1}^{m} Q_i$ are seen to be independent. Thus

$$v \cdot \left(\sum_{i=1}^{m} Q_i\right)^m = \prod_{i=1}^{m} Q_i$$

where the 2 factors on the left are independent as are the m factors on the right. Since the distributions of $\sum_{i=1}^{m} Q_i$ and Q_i are known this equation makes it possible to immediately write down the moments of v about 0 or the cumulants and characteristic function of $\log v$. In this way we approach the limiting χ^2 distribution of $\log v$ as $r \to \infty$ and show that the power series expansion in terms of $(1/r)$ of the cumulants of the actual and asymptotic distributions agree up to the terms in $(1/r)^2$. This asymptotic distribution is stated in [7] to be remarkably good with agreement of the first 4 cumulants to within 5% when r is as small as 5.

Asymptotic expansions for the cumulants may be derived as follows. Define $t = -\log(m^m v)$, and K_s as meaning sth cumulant. Then for any s

$$K_s(\log v) + m^s K_s\left(\log \sum_{i=1}^{m} Q_i\right) = mK_s(\log Q_i),$$

or

$$K_s(\log v) + m^s K_s(\log \chi^2_{mr}) = mK_s(\log \chi^2_r).$$

From [8] asymptotic formulas for the cumulants of $\log \chi^2_n$ are given by

$$K_1(\log \chi^2_n) = \log n - \frac{1}{n} - 2\sum_{j=1}^{\infty} \frac{(-4)^{j-1}B_j}{jn^{2j}}$$

$$= \log n - \frac{1}{n} - \frac{1}{3n^2} + \frac{1}{15n^4} - \frac{16}{63n^6} + \cdots,$$

and

$$K_s(\log \chi^2_n) = (-)^s 2^s \left[\frac{(s-2)!}{2n^{s-1}} + \frac{(s-1)!}{2n^s} + \frac{2}{n^s}\sum_{j=1}^{\infty} \frac{(-4)^{j-1}B_j(2j+s-1)!}{(2j)!\,n^{2j-1}}\right]$$

$$= (-)^s 2^s \left[\frac{(s-2)!}{2n^{s-1}} + \frac{(s-1)!}{2n^s} + \frac{s!}{6n^{s+1}} + \frac{0}{n^{s+2}} + \cdots\right] \quad \text{for} \quad s \geqq 2,$$

where B_j are Bernoulli numbers. Thence

$$K_1(t) = -m\log m - K_1(\log v)$$

$$= -m\log m - mK_1(\log \chi^2_r) + mK_1(\log \chi^2_{mr})$$

$$= (m-1)\left[\frac{1}{r} + \frac{1+\frac{1}{m}}{3r^2} - \frac{2\left(1+\frac{1}{m}+\frac{1}{m^2}+\frac{1}{m^3}\right)}{15r^4} + \cdots\right],$$

and for $s \geqq 2$

$$K_s(t) = (-)^s K_s(\log v) = (-)^s[mK_s(\log \chi^2_r) - m^s K_s(\log \chi^2_{mr})]$$

$$= 2^{s-1}(s-1)!(m-1)\left[\frac{1}{r^s} + \frac{s\left(1+\frac{1}{m}\right)}{3r^{s+1}} + \frac{0}{r^{s+2}} + \cdots\right]$$

Since χ^2_{m-1} has cumulants $K_s = 2^{s-1}(s-1)!\,(m-1)$ it is seen that $t \sim (1/r)\chi^2_{m-1}$ with agreement in first terms of the expansions, and

$$t \sim \left(\frac{1}{r} + \frac{1 + \frac{1}{m}}{3r^2}\right)\chi^2_{m-1}$$

with agreement in the first 2 terms, for all cumulants.

Thus r may be estimated by $\hat{r}$ defined by

$$t = \left(\frac{1}{\hat{r}} + \frac{1 + \frac{1}{m}}{3\hat{r}^2}\right)(m-1)$$

and for r moderately large the distribution χ^2_{m-1} can be used to put confidence limits on r.

Consider next the set of $\frac{1}{2}m(m-1)$ angles among $U_1, U_2, \cdots, U_m$. Set $n = \frac{1}{2}m(m-1)$ and denote by S_i $(i = 1, 2, \cdots, n)$ the squared sines of these angles. Under the approximate model any S_i is considered distributed as $\beta_{(r-1)/2,1/2}$, but as a further consequence of spherical symmetry in r-space it may be noted that any set of angles containing no closed subset is a mutually independently distributed set, and in particular the angles are pairwise independent. Extending this approximation to complete independence the joint density of the S_i becomes

$$\left(\frac{\Gamma\left(\frac{r}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{r-1}{2}\right)}\right)^n \prod_{i=1}^n (S_i)^{(r-3)/2} \prod_{i=1}^n (1 - S_i)^{-\frac{1}{2}}$$

so that $\prod_{i=1}^n S_i$ or $\sum_{i=1}^n \log S_i$ appear as equivalent sufficient statistics for r, so contain approximately all the information about r in the directional properties.

This leads to a consideration of $-\log \beta_{(r-1)/2,1/2}$. The density of $1 - \beta_{(r-1)/2,1/2}$ is easily seen to be asymptotically as $r \to \infty$ the density of $1/r\ \chi^2_1$ and since $\beta_{(r-1)/2,1/2} \to 1$ in probability as $r \to \infty$ it follows that

$$\frac{-\log \beta_{(r-1)/2,1/2}}{1 - \beta_{(r-1)/2,1/2}} \to 1$$

in probability as $r \to \infty$ so that $-\log \beta_{(r-1)/2,1/2}$ is also asymptotically distributed as $1/r\ \chi^2_1$.

Direct asymptotic expansions for the cumulants of $-\log \beta_{(r-1)/2,1/2}$ show that, as with statistic t, this last asymptotic distribution can be modified to have agreement in the first 2 terms. For, since $\chi^2_{r-1} = \beta_{(r-1)/2,1/2} \cdot \chi^2_r$ with independence on the right (as may be seen by computing the characteristic functions of the logs of these random variables),

$$\begin{aligned} K_s\,(-\log \beta_{(r-1)/2,1/2}) &= (-)^s K_s\,(\log \beta_{(r-1)/2,1/2}) \\ &= (-)^s[K_s\,(\log \chi^2_{r-1}) - K_s\,(\log \chi^2_r)] \end{aligned}$$

for all s. Thence

$$K_1(-\log \beta_{(r-1)/2,1/2}) = \log r + \left[-\frac{1}{r} - \frac{1}{3r^2} + \frac{2}{15r^4} + \cdots\right]$$
$$- \log (r-1) - \left[-\frac{1}{r-1} - \frac{1}{3(r-1)^2} + \frac{2}{15(r-1)^4} + \cdots\right]$$
$$= \frac{1}{r} + \frac{3}{2r^2} + \frac{2}{r^3} + \frac{9}{4r^4} + \cdots,$$

and for $s \geqq 2$

$$K_s(-\log \beta_{(r-1)/2,1/2}) = 2^s\left[\frac{(s-2)!}{2(r-1)^{s-1}} + \frac{(s-1)!}{2(r-1)^s} + \frac{s!}{6(r-1)^s} + 0 + \cdots\right]$$
$$- 2\left[\frac{(s-2)!}{2r^{s-1}} + \frac{(s-1)!}{2r^s} + \frac{s!}{6r^{s-1}} + 0 + \cdots\right]$$
$$= 2^{s-1}(s-1)!\left[\frac{1}{r^s} + \frac{3s}{2r^{s+1}} + \frac{s(s+1)}{r^{s+2}} + \cdots\right]$$

so that

$$-\log \beta_{(r-1)/2,1/2} \sim \left(\frac{1}{r} + \frac{3}{2r^2}\right)\chi_1^2$$

with agreement to second terms in the expansions and therefore usable accuracy for quite small r.

Now we may regard

$$-\sum_{i=1}^{n} \log S_i \sim \left(\frac{1}{r} + \frac{3}{2r^2}\right)\chi_n^2$$

and obtain a new estimate of r. Since in approximation the angles were more than pairwise independent the first 2 moments of this last are asymptotically faithful to the approximate model. The remaining moments however will be distorted slightly on account of non-independence in a way which is difficult to investigate.

Finally an estimate of r can be obtained from $t - \sum_{i=1}^{n} \log S_i$ regarded as asymptotically $(1/r)\ \chi^2_{m-1+n}$ or an appropriate refinement for small r.

4. The non-exact significance test. The question is discussed here of what can be had in the way of a significance test based on $F = Q_0/1/m\sum_{i=1}^{m} Q_i$ considered as $F_{r,mr}$ under the null hypothesis where r is unknown but estimated from a statistic w considered distributed as $f(r)\chi_n^2$ independent of F with $f(r)$ equal to $1/r$ or an asymptotically equivalent refinement of $1/r$. The point estimate of r found from the equation $w = f(r)\cdot n$ will be denoted by $\hat{r}$ and the term $100p\%$ confidence point of r will indicate the value of r satisfying $w = f(r)\chi^2_{n(p)}$ where $\chi^2_{n(p)}$ denotes the $100p\%$ point of χ_n^2. Similar notation will be used for percentage points of other distributions.

A statistical test may be termed exact if the distribution of the test statistic under the null hypothesis does not depend on any unknown parameters. If r

were known the statistic F would have this property and the natural test would be to regard F as significant if $F > F_{r,mr(.95)}$. (Assume for this discussion a standard 5% nominal significance level.) Since r is unknown any test based on F must be non-exact and the natural non-exact test appears to be to regard F as significant if $F > F_{\hat{r},m\hat{r}(.95)}$. This test can also be formulated in terms of quantities α and $\hat{\alpha}$. Define α as the significance level of the observed F as a function of the true parameter r, i.e. α satisfies $F = F_{r,mr(1-\alpha)}$. Similarly $\hat{\alpha}$ as a function of the observed statistics F and w can be determined from $F = F_{\hat{r},m\hat{a}(1-\hat{r})}$. The unattainable exact test is that F is significant if $\alpha < .05$; the non-exact test defined is that F is significant if $\hat{\alpha} < .05$.

The non-exact test still has a significance level (or size or probability of type I error) but this is now a function of r. Denoting this function by $\gamma(r)$ we have

$$\begin{aligned}\gamma(r) &= \Pr(\hat{\alpha} < .05) \\ &= \Pr(F > F_{\hat{r},m\hat{r}(.95)}) \\ &= \operatorname*{ave}_{w}\{\Pr(F > F_{\hat{r}(w),m\hat{r}(w)(.95)} \mid w)\}\end{aligned}$$

where F is distributed as $F_{r,mr}$. The last version of $\gamma(r)$ indicates how $\gamma(r)$ can be calculated for given r i.e. by averaging a set of fairly well tabled probabilities over a χ^2 distribution. The major interest of this section is to determine the relation between $\gamma(r)$ and the nominal significance level .05.

The distributions of α and $\hat{\alpha}$ can be compared by fixing α and looking at the variability of the corresponding $\hat{\alpha}$. This amounts to conditioning the various random variables by fixing F to produce the desired α, but leaving w unconditioned. For any fixed α, if r is known, percentage points of w can be translated into percentage points of $\hat{r}$ and thence to percentage points of $\hat{\alpha}$. These are denoted $(\hat{\alpha} \mid \alpha)_{(p)}$. Alternatively, for fixed α, r unknown, but w observed, confidence points for r can be translated into confidence points for α and these will also indicate how much $\hat{\alpha}$ varies about α.

Short of actually calculating $\gamma(r)$ for various values of r, m, n and .05, two arguments will be advanced to show that it is near .05. The first argument is to use a table to back up the belief that the disturbance caused by going from α to $\hat{\alpha}$ is not very great relative to the (0, 1) range of α and is well balanced with regard to direction, so that the unconditional distribution of $\hat{\alpha}$ is not much different from the uniform (0, 1) distribution of α. Table 1 shows quartiles of $(\hat{\alpha} \mid \alpha)$ for $m = 10, n = 64, \alpha = .05$ and .10, and $r = 6$ and ∞. This table indicates that the

TABLE 1

r	α	$(\hat{\alpha}\mid\alpha)_{(.25)}$	$(\hat{\alpha}\mid\alpha)_{(.75)}$
6	.050	.043	.057
	.100	.091	.107
∞	.050	.030	.070
	.100	.072	.125

disturbance in α caused by using $\hat{\alpha}$ is well-balanced near the 5% level and is a slight shift towards 0 near 10%. The indication is that $\gamma(r)$ is very near .05.

The second argument involves computing the non-trivial limit $\gamma(\infty) = \lim_{r\to\infty} \gamma(r)$. Define

$$(F_{r,mr} - 1)^+ = 0 \quad \text{if} \quad F_{r,mr} - 1 \leqq 0$$
$$= F_{r,mr} - 1 \quad \text{otherwise.}$$

As

$$r \to \infty \; F_{r,mr} \sim N\left(1, \frac{2}{r}\left\{1 + \frac{1}{m}\right\}\right)$$

so that

$$(F_{r,mr} - 1)^+ \sim \left(N\left(0, \frac{2}{r}\left\{1 + \frac{1}{m}\right\}\right)\right)^+$$

or

$$[(F_{r,mr} - 1)^+]^2 \sim 0 \quad \text{or} \quad \frac{2}{r}\left(1 + \frac{1}{m}\right)\chi_1^2$$

each with probability $\frac{1}{2}$. Similarly if $(1/\hat{r}) = (1/rn)\chi_n^2$ is put in for $1/r$,

$$[(F_{\hat{r},m\hat{r}} - 1)^+]^2 \sim 0 \quad \text{or} \quad \frac{2}{r}\left(1 + \frac{1}{m}\right)\chi_1^2 = \frac{2}{r}\left(1 + \frac{1}{m}\right)\frac{1}{n}\chi_n^2 \cdot \chi_1^2$$

each with probability $\frac{1}{2}$ where χ_n^2 and χ_1^2 are independent. From this

$$\gamma(\infty) = \lim_{r\to\infty} \Pr\,(F_{r,mr} > F_{\hat{r},m\hat{r}(.95)})$$
$$= \lim_{r\to\infty} \Pr\,([(F_{r,mr} - 1)^+]^2 > [(F_{\hat{r},m\hat{r}} - 1)^+]^2_{(.95)})$$
$$= \tfrac{1}{2} \Pr\,(\chi_1^2 > \frac{1}{n}[\chi_n^2 \cdot \chi_1^2]_{(.90)})$$

Now

$$\text{ave}\,\{\chi_1^2\} = 1 \quad \text{and} \quad \text{var}\,\{\chi_1^2\} = 2,$$

and

$$\text{ave}\left\{\frac{1}{n}\chi_n^2 \cdot \chi_1^2\right\} = 1 \quad \text{and} \quad \text{var}\left\{\frac{1}{n}\chi_n^2 \cdot \chi_1^2\right\} = 2 + \frac{6}{n},$$

which indicates strongly that

$$\left[\frac{1}{n}\chi_n^2 \cdot \chi_1^2\right]_{(.90)} > \left[\chi_1^2\right]_{(.90)}$$

i.e. $\gamma(\infty) < .05$ so that asymptotically the test is conservative as r gets large. Since the foregoing table indicates smaller spread in $(\hat{\alpha} \mid \alpha)$ for finite r than for $r = \infty$ we might hope that $\gamma(r)$ is as near .05 as $\gamma(\infty)$ is.

In any particular case the spread in $(\hat{\alpha} \mid \alpha)$ can be examined by computing confidence points of α from confidence points of r.

One feature of this test which might be regarded as a practical drawback is the non-uniqueness of the vectors $U_1, U_2, \cdots, U_m$. These vectors resulted from a choice of an orthogonal set of m d.f. chosen arbitrarily in a space of $m = n_1 + n_2 - 2$ dimensions. The symmetry of the normal distribution over any choice of an orthogonal set of d.f. assures that the distribution theory of the test holds for any such set, but there is no assurance that the observed statistics are unchanged by different choices. In fact it can be easily seen that $\sum_{i=1}^{m} Q_i$ is invariant under all choices so that $F = Q_0/(1/m) \sum_{i=1}^{m} Q_i$ is also. Thus it is only in the estimation of r that variations occur, and, since we have heuristic evidence of having used almost all the information about r in our estimates $\hat{r}$ and since the $\hat{r}$ plays only a secondary role, the non-uniqueness of the significance test should be of minor importance.

5. The sensitivity of the test. A natural parameter measuring separation of the 2 populations is distance between their means in a metric defined as follows from their second order moments. Suppose that the metric inserted into affine k-space from a priori information and used heretofore is denoted by G_1, and suppose that the ellipsoid in affine space which appears as the unit sphere in G_1 is denoted by E_1. If, in an affine coordinate system for k-space a sample point is represented by $k \times 1$ vector u and the corresponding $k \times k$ matrix of variances and covariances is Λ, then an ellipsoid E_2 can be defined as $u'\Lambda^{-1}u = 1$. It is easily seen that the same ellipsoid is defined by the same prescription in any affine coordinate system so that given the distribution over affine space E_2 is uniquely defined. E_2 may now be used to define a Euclidean metric G_2 in affine space as that metric in which E_2 appears to be the unit sphere (so that the distribution is sphericalized). Suppose the G_2-distance between population means is τ i.e. ν_0 has G_2-length τ. Then τ, which is also the ratio of mean difference to standard deviation for that linear combination of original variables which maximizes this ratio, may be taken as the parameter measuring difference between population means, and we would like to know if our test is sensitive to large τ.

Now ave$\{V_0\} = \nu_0$ with G_2-length τ so ave$\{U_0\} = \xi = (1/n_1 + 1/n_2)^{-1/2}\nu_0$ with G_2-length $(1/n_1 + 1/n_2)^{-1/2}\tau = \tau_1$ say. Denote by $Q_0(\xi)$ the squared G_1-length of U_0, by $P_0(\xi)$ the squared G_2-length of U_0 and by $R(U_0)$ the squared ratio of the radius of E_1 to the radius of E_2 both radii in the direction of U_0. (This ratio of lengths in one direction is independent of the particular metric.) Then

$$Q_0(\xi) = R(U_0)P_0(\xi). \tag{5.1}$$

Assuming normality the distributions appear in G_2 as spherical unit normals so that $P_0(\xi)$ has the non-central χ^2 distribution $\chi_k^2(\tau_1)$ defined as the distribution of $(v_1 + \tau_1)^2 + v_2^2 + \cdots + v_k^2$ where $v_1, v_2, \cdots, v_k$ are NID(0, 1). Unfortunately $R(U_0)$ has a distribution depending on the direction of ξ as well as its length.

This may be contrasted with the statistic T^2 usable if $k \leqq m$, which is non-centrally distributed as

$$T^2 = \frac{m\chi_k^2(\tau_1)}{\chi_{m-k+1}^2}$$

with the numerator independent of the denominator [9] which distribution depends only on τ and on no other properties of ξ. The present situation has the undesirable feature that the G_1-metric may have been selected in such a way that the G_1-length of ξ is too small to cause a significant disturbance in Q_0 whereas τ_1 is large enough to cause a significant disturbance in $\chi_k^2(\tau_1)$. An extreme example of this would occur if the populations were not of full rank but lay in separate parallel hyperplanes but still very close in G_1. Here $\tau = \infty$ but $Q_0(\xi)$ could very well be little disturbed. As long as ξ is regarded as having non-random direction and G_1 cannot be chosen to coincide with G_2 there is danger of insensitivity to a large τ arising from this source. On the other hand if it is permissible to assume randomness for ξ then this danger can be controlled on the average, and further discussion proceeds along these lines.

The high-dimensional case is likely to arise in practice when little or nothing is known about the separating power of individual variables. If nothing is supposed known it may be reasonable to think of ξ as random with all directions intuitively equally likely. The only affine choice consistent with this intuitive notion is to make ξ uniformly distributed with respect to G_2-direction, so the first case considered will be where ξ has constant G_2-length τ_1 and is uniformly distributed with respect to G_2-direction independently of the within sample variation.

Under this assumption and normality $U_0(\xi) = \xi + U_0(0)$ where the 2 vectors on the right are independent each with directions distributed uniformly in G_2. Also, due to the G_2-spherical symmetry of the distribution of $U_0(0)$, the G_2-length of $U_0(0)$ is distributed independently of its G_2-direction. It follows that $U_0(\xi)$ has independently distributed G_2-length and G_2-direction, so that in the equation (5.1) the 2 terms on the right are independent. Also the distribution of $R(U_0)$ is independent of ξ. In our standard approximation of $Q_0(0)$ by $\mu\chi_r^2$ by fitting first 2 moments we have $\text{ave}\{Q_0(0)\} = \mu r$ and $\text{ave}\{Q_0^2(0)\} = \mu^2 r(r+2)$. Also $\text{ave}\{P_0(0)\} = \text{ave}\{\chi_k^2\} = k$ and $\text{ave}\{P_0^2(0)\} = k(k+2)$. Thus

$$\text{ave}\{R(U_0)\} = \frac{\text{ave}\{Q_0(0)\}}{\text{ave}\{P_0(0)\}} = \frac{\mu r}{k}$$

and

$$\text{ave}\{R^2(U_0)\} = \frac{\text{ave}\{Q_0^2(0)\}}{\text{ave}\{P_0^2(0)\}} = \frac{\mu^2 r(r+2)}{k(k+2)}$$

so that

$$\text{ave}\{Q_0(\xi)\} = \text{ave}\{R(U_0)\} \cdot \text{ave}\{\chi_k^2(\tau_1)\} = \mu r\left(1 + \frac{\tau_1^2}{k}\right),$$

$$\text{ave } \{Q_0^2(\xi)\} = \text{ave } \{R^2(U_0)\} \text{ ave } \{\chi_k^4(\tau_1)\}$$
$$= \mu^2 r(r+2)\left(1 + 2\frac{\tau_1^2}{k} + \frac{\tau_1^4}{k(k+2)}\right),$$

and

$$\text{var } \{Q_0(\xi)\} = 2\mu^2 r\left(1 + 2\frac{\tau_1^2}{k} + \frac{k-r}{k+2}\frac{\tau_1^4}{k^2}\right).$$

The distribution of $Q_0(\xi)$ is clearly not non-central χ^2 since the variance of the latter does not involve a term in τ_1^4. For practical purposes it would be reasonable to fit a χ^2 shape to this distribution by fitting first 2 moments, i.e. $\lambda\chi_q^2$ where

$$\lambda = \mu \frac{1 + 2\frac{\tau_1^2}{k} + \frac{k-r}{k+2}\frac{\tau_1^4}{k^2}}{1 + \frac{\tau_1^2}{k}}$$

$$q = r\left(1 + \frac{\frac{r+2}{k+2}\frac{\tau_1^4}{k^2}}{1 + 2\frac{\tau_1^2}{k} + \frac{k-r}{k+2}\frac{\tau_1^4}{k^2}}\right)$$

Now it is possible to compute approximate "power functions" and "confidence limits" for τ by assuming F for $\tau > 0$ approximately distributed as $\lambda/\mu\, F_{q,mr}$ and by adopting the procedure used with significance testing of replacing r by $\hat{r}$. These "power functions" and "confidence limits" are actually estimates of the true power functions and confidence limits associated with the non-exact test just as $\hat{\alpha}$ was an estimate of α. The deviation of the estimated power from the true power may again be expected to be near zero and balanced about zero. Using confidence points of r confidence points for any particular value of the power function may be found and these will indicate the order of the disturbance caused by replacing r by $\hat{r}$.

For convenience a criterion different from the power function will be used to measure the sensitivity of the test, namely τ_c the value of τ which will produce on the average a barely significant test statistic. Regarding $(1/m)\sum_{i=1}^m Q_i$, the denominator of F, as $(\mu/m)\,\chi_{mr}^2$

$$\text{ave } \{F\} = \text{ave } \{Q_0(\xi)\} \cdot \text{ave}\left\{\frac{m}{\mu} \cdot \chi_{mr}^{-2}\right\}$$
$$= \mu r\left(1 + \frac{\tau_1^2}{k}\right) \cdot \frac{1}{\mu r}\frac{mr}{mr-2}$$
$$= \left(1 - \frac{2}{mr}\right)^{-1}\left(1 + \frac{\tau_1^2}{k}\right)$$

so that τ_c satisfies

$$1 + \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1}\frac{\tau_c^2}{k} = \left(1 - \frac{2}{mr}\right) F_{r,mr(.95)}$$

or, since, for large r, $F_{r,mr} \sim N(1, (2/r)[1 + (1/m)])$, τ_c asymptotically satisfies

$$1 + \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} \frac{\tau_c^2}{k} = 1 + 1.65 \left(\frac{2}{r}\right)^{\frac{1}{2}} \left(1 + \frac{1}{m}\right)^{\frac{1}{2}}$$

or

$$\tau_c^2 = N \left(\frac{1}{n_1} + \frac{1}{n_2}\right) k r^{-\frac{1}{2}}$$

where $N = 1.65(2)^{\frac{1}{2}}(1 + (1/m))^{\frac{1}{2}} \doteq 2.3$. Note that for a given experiment $r^{-\frac{1}{2}}$ is the only factor in τ_c^2 which depends on G_1. This result is encouraging, for suppose we have a set of variables with equal but possibly small individual separation parameters ρ. If the within sample variation is independent from variable to variable then $\tau^2 = k\rho^2$. Thus if G_1 can be chosen such that

$$r \geqq \frac{N^2}{\rho^4} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^2$$

then separation would show on the average. This implies that regardless of how small ρ is we need only go on adding variables of separation ρ until r has been built up to correct size. Whether it is possible to continue indefinitely adding variables with small separations in a practical case is uncertain, but the example does show how small individual separations can produce something that will show.

If there is some feeling that ξ is not uniformly distributed relative to G_2-direction an alternative would be to suppose it uniform relative to a different metric G_3 with ellipsoid E_3, i.e. when E_3 appears as the unit sphere ξ appears of length σ and uniform with regard to direction independent of $Q_0(0)$. Then a priori knowledge of the separating powers of the variables could be supposed to consist of some information about E_3. Suppose the mean square G_1-length of ξ is $A^2\sigma^2$ where A^2 depends only on E_3, and suppose the mean square G_1-length of the centrally distributed U_0 is $B^2 = \mu r = \text{ave}\{Q_0(0)\}$. Then

$$\text{ave}\{Q_0(\xi)\} = B^2 + A^2\sigma = \mu r \left(1 + \frac{A^2}{B^2}\sigma^2\right)$$

so that σ_c^2 producing significance on the average is given, in the asymptotic case by

$$\sigma_c^2 = N \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \frac{A^2}{B^2} r^{-\frac{1}{2}}$$

where now the choice of G_1 can influence both B/A and r.

We are now in a position to discuss theoretical issues concerning the original choice of metric G_1. These suggest that for most purposes the aim should be to make G_2 and G_1 coincide as nearly as possible except for a scale factor. The practical question of how well this can be accomplished is not discussed, nor is it crucial for the use of the method. There are 2 issues in the choice of G_1 : sensitivity of the test and safety of the assumptions.

If G_1 is related to G_2 by a scale factor, then the statistic Q is distributed as $\mu\chi_k^2$ i.e., $r = k$ and under normality the approximation to the distribution of Q by $\mu\chi_r^2$ is exact. This is the way in which choice of G_1 can be made to improve the assumptions. It is heuristically evident that a larger value of r results in less likelihood that approximations of this kind will go wrong.

Regarding sensitivity it can be seen that only when E_2 is G_1-spherical is there equal sensitivity to a separation of τ in all directions and so no danger of insensitivity to large τ. Also it has been seen that when the direction of ξ is assumed G_2-uniformly distributed τ_c^2 depends on $r^{-\frac{1}{2}}$ so again there is evidence that maximizing r to k gives greatest sensitivity. However, under the alternative randomness assumption of ξ uniform over ellipsoid E_3 the situation appears more complicated, for the factor in σ_c^2 to be minimized by choice of G_1 is $A^2/(B^2)\, r^{-\frac{1}{2}}$. This suggests that if something is known about the shape of E_3 as well as E_2 then E_1 should be chosen to give more weight to those directions in which E_3 is long relative to E_2 provided this does not too greatly depress r. It is felt that this last suggestion may be occasionally useful but the general rule will be to try to make $r = k$.

6. Asymptotic behavior. In the foregoing are many results asymptotically true as $r \to \infty$ with m fixed. Certainly these are a mathematical convenience. The question of whether indefinitely large r can be practically obtained remains open. Certainly if k can be made arbitrarily large and each of the k variables contains a part independent of the rest then in theory r can be made arbitrarily large because a metric can be chosen such that $r = k$. What is much more in doubt is whether or not variables could be chosen which would give r indefinitely increasing and τ also increasing at a rate such that the sensitivity of our method would continue to improve.

Whether it is practically attainable or just mathematically useful the following geometrical picture of the asymptotic case is illuminating. Consider throughout the approximate model of section 3 and its asymptotic behavior. As $r \to \infty$ the coefficient of variation of Q (i.e. $\mu\chi_r^2$) tends to 0, so that if we back away from the picture at the correct rate as r increases the vectors $U_1, \cdots, U_m$ will appear to all approach in probability the same constant length. Also since $1 - S_i \sim 1/(r)\, \chi_1^2$ each angle between vectors tends in probability to $\pi/2$ so they tend to an orthogonal set of m equal length vectors. Vector U_0 also becomes perpendicular to $U_1, \cdots, U_m$ but its length depends on τ. However if its length should differ from the common limiting lengths of the rest by a factor as great as $(1 + Nr^{-\frac{1}{2}})^{\frac{1}{2}}$ this is roughly what would be called significant, so that asymptotically a significant U_0 could be indistinguishably different from the rest.

An implication of this asymptotic picture is as follows. For small r it would be natural to compare $Q_0(\xi)$ from U_0 more closely with those Q_i from U_i making the smallest angles with U_0, because if U_0 and U_i are close then $R(U_0)$ and $R(U_i)$ are likely to be more nearly the same. T^2 accomplishes this in a neat manner which disappears when $k > m$, but the present method makes no attempt to do

it. The asymptotic picture says that in the limit there is no hope of making such a correction, for if U_0 is nearly at right angles with every U_i then the radii of E_1 and E_2 in the direction of U_0 bear no relation to the radii in the direction of the U_i, i.e. there are too many directions for U_0 to take to hope that it will be near enough to any U_i to make any difference.

7. Acknowledgments. Heartiest thanks are due to Professor John W. Tukey of Princeton University for his generous guidance in this research.

REFERENCES

[1] A. P. DEMPSTER, "The multivariate two sample problem in the degenerate case," unpublished PhD. thesis, Princeton University, 1956.

[2] H. HOTELLING, "The generalisation of 'Student's' ratio," *Ann. Math. Stat.*, Vol. 2 (1931), p. 360.

[3] R. A. FISHER, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*," Vol. 7 (1936), p. 179.

[4] E. J. G. PITMAN, "Significance tests which may be applied to samples from any population," J. Roy. Stat. Soc., Supplement, Vol. 4 (1937), p. 119–130, 225–232.

[5] B. L. WELCH, "On the z-test in randomized blocks and latin squares," *Biometrika*, Vol. 29 (1937), p. 21.

[6] G. E. P. BOX, "Some theorems on quadratic forms applied in the study of analysis of variance I," *Ann. Math. Stat.*, Vol. 25 (1954), p. 290.

[7] M. S. BARTLETT AND M. G. KENDALL, "The statistical analysis of variance heterogeneity and the logarithmic transformation," J. Roy. Stat. Soc., Supplement, Vol. 8 (1946), p. 128.

[8] J. WISHART, "The cumulants of the Z and of the logarithmic χ^2 and t distributions," *Biometrika*, Vol. 34 (1947), p. 170.

[9] R. C. BOSE AND S. N. ROY, "The distribution of Studentized D^2 statistic," *Sankhya*, Vol. 4 (1938), p. 337.