TWO-SAMPLE TESTS FOR MULTIVARIATE DISTRIBUTIONS[1]

By Lionel Weiss

Cornell University

1. Introduction and summary. $X(1), X(2), \cdots, X(m), Y(1), Y(2), \cdots, Y(n)$ are independent k-variate random variables. The distribution of $X(i)$ has pdf $f(x)$, say, where x denotes a k-dimensional vector throughout this paper, and the distribution of $Y(j)$ has pdf $g(x)$, say. We assume that $f(x)$ and $g(x)$ are piecewise continuous, and that each has a finite upper bound, which it is not necessary to specify.

Denote by $2R_i$ the distance from $X(i)$ to the nearest of the points $X(1), \cdots, X(i-1), X(i+1), \cdots, X(m)$, and denote by S_i the number of points $Y(1), \cdots, Y(n)$ contained in the open sphere $\{x: \mid x - X(i) \mid < R_i\}$. Clearly, the joint distribution of S_i, S_j is the same as the joint distribution of $S_{i'}$, $S_{j'}$, for any subscripts with $i \neq j$, $i' \neq j'$. Let r be a non-negative integer, and α any fixed positive value. $Q(r)$ denotes the Lebesgue integral

$$\int_{E_k} \frac{2^k \alpha f^2(x)[g(x)]^r}{[g(x) + 2^k \alpha f(x)]^{r+1}}\, dx,$$

where E_k denotes Euclidean k-space. We will show that

$$\lim_{m\to\infty, m/n=\alpha} P_{m,n}[S_1 = s_1, S_2 = s_2] = Q(s_1)Q(s_2),$$

for any non-negative integers s_1, s_2, the approach being uniform in s_1, s_2. Thus, in the limit S_1, S_2 are independently distributed, with

$$\lim_{m\to\infty, m/n=\alpha} P_{m,n}[S_1 = s_1] = Q(s_1).$$

In [1], which discussed the univariate case, S_i was defined as the number of Y's closer to $X(i)$ than to any other X to their right. In the present paper, S_i is defined as the number of Y's in another neighborhood of $X(i)$. Our present definition of S_i does not become for $k = 1$ the same as the definition of S_i in [1]. Rather, in the univariate case, our present definition of S_i is the number of Y's lying within a distance R_i on either side of $X(i)$. However, if

$$\lim_{m\to\infty, m/n=\alpha} P_{m,n}[S_1 = s_1, S_2 = s_2]$$

is computed for the univariate case using the definition of S_i given in [1], the only way in which it differs from $Q(s_1)Q(s_2)$ is that α is replaced by $\alpha/2$. Thus it seems reasonable to treat the S_i as defined here as k-dimensional analogues of the S_i as defined in [1], at least for large samples. An intuitive reason for α being replaced by $\alpha/2$ is that in our present case, $\sum_{i=1}^{m} S_i$ may be less than n, whereas in [1] this sum must always equal n. Thus in our present case, we are

Received September 29, 1958; revised August 17, 1959.

[1] Research sponsored by the Office of Naval Research.

in a sense discarding some of the Y's, which lowers n relative to m and thus raises α by a certain factor (2, as it happens). In our present case, $\sum S_i$ may be less than n because the R_i are chosen to make the spheres around the X's non-overlapping, thus simplifying the analysis. The R_i were chosen to give the largest possible non-overlapping spheres because it would seem intuitively that the larger the spheres, the more rapid the approach of the probabilities to their limiting values.

2. Derivation of the limiting distribution of S_1, S_2. Let $p_m(r_1, r_2 \mid x(1), x(2))$ denote the joint conditional pdf of R_1, R_2 given that $X(1) = x(1)$ and $X(2) = x(2)$. Denote $\int_{|x-a|<b} f(x)\,dx$ by $V(a, b; f)$, where $a \in E_k$ and b is a positive scalar. If $f(x)$ is continuous in an open region containing the open sphere $\{x: |x - a| < b\}$, then $\partial V(a, b; f)/\partial b$ is equal to the surface integral $\int_{S:|x-a|=b} f(x)\,dS$, which we denote by $S(a, b; f)$.

$$P[R_1 \geqq r_1 \text{ and } R_2 \geqq r_2 \mid X(1) = x(1) \text{ and } X(2) = x(2)]$$
$$= [1 - V(x(1), 2r_1; f) - V(x(2), 2r_2; f)]^{m-2}$$

if $r_1 \geqq 0$, $r_2 \geqq 0$, $|x(1) - x(2)| \geqq 2 \max(r_1, r_2)$; and is equal to zero for other values of r_1, r_2, $x(1)$, $x(2)$. If $f(x)$ is continuous in an open region containing the points x with $|x - x(i)| \leqq 2r_i$ for $i = 1, 2$, then

$$p_m(r_1, r_2 \mid x(1), x(2))$$
$$= \frac{\partial^2}{\partial r_1 \partial r_2} P[R_1 \geqq r_1, R_2 \geqq r_2 \mid X(1) = x(1) \quad \text{and} \quad X(2) = x(2)]$$
$$= (m - 2)(m - 3)[1 - V(x(1), 2r_1; f) - V(x(2), 2r_2; f)]^{m-4} \prod_{i=1}^{2} 2S(x(i), 2r_i; f)$$

if $r_1 \geqq 0$, $r_2 \geqq 0$, $|x(1) - x(2)| \geqq 2 \max(r_1, r_2)$; and is equal to zero for other values of r_1, r_2, $x(1)$, $x(2)$. From our continuity assumption on $f(x)$ we have

$$V(x(i), 2r_i; f) = f(x(i))[\pi^{\frac{1}{2}k}(2r_i)^k/\Gamma(\tfrac{1}{2}k + 1)] + r_i^k\,\epsilon(x(i); 2r_i),$$

and

$$S(x(i), 2r_i; f) = f(x(i))k\pi^{\frac{1}{2}k}(2r_i)^{k-1}/\Gamma(\tfrac{1}{2}k + 1) + r_i^{k-1}\epsilon_1(x(i); 2r_i),$$

where $\epsilon(\ )$ and $\epsilon_1(\ )$ approach zero as r_i approaches zero. Furthermore, these quantities approach zero uniformly over any set G of points in $(x(1), x(2))$ space such that $f(x)$ is uniformly continuous over the projection of G on the $x(1)$ hyperplane and on the $x(2)$ hyperplane and $|x(1) - x(2)| > \delta > 0$ over G.

Now introduce the random variables Z_1, Z_2 by the relationship $R_i = (Z_i/m)^{1/k}$, for $i = 1, 2$. Denote by $h_m(z_1, z_2 \mid x(1), x(2))$ the joint conditional pdf of Z_1,

Z_2 given that $X(1) = x(1)$ and $X(2) = x(2)$. By substituting in

$$p_m(r_1, r_2 \mid x(1), x(2))$$

and using the facts developed above, we have

$$\lim_{m\to\infty} h_m(z_1, z_2 \mid x(1), x(2)) = \prod_{i=1}^{2} \frac{(2\sqrt{\pi})^k}{\Gamma(\frac{1}{2}k+1)} f(x(i)) \exp\left\{-\frac{(2\sqrt{\pi})^k}{\Gamma(\frac{1}{2}k+1)} z_i f(x(i))\right\}$$

uniformly over any set G in $(x(1), x(2), z_1, z_2)$ space such that $f(x)$ is uniformly continuous over the projection of G on the $x(1)$ hyperplane and on the $x(2)$ hyperplane, $|x(1) - x(2)| > \delta > 0$ over G, and the projections of G on the z_1 and z_2 axes are bounded from above. We need consider only positive z_1, z_2.

Next, denote by $D_m(s_1, s_2 \mid z_1, z_2, x(1), x(2))$ the conditional probability that $S_1 = s_1$ and $S_2 = s_2$, given that $Z_1 = z_1$, $Z_2 = z_2$, $X(1) = x(1)$, $X(2) = x(2)$. Then

$$D_m(s_1, s_2 \mid z_1, z_2, x(1), x(2)) = \frac{n!}{s_1!\, s_2!(n - s_1 - s_2)!} (1 - V(x(1), r_1; g) - V(x(2), r_2; g))^{n-s_1-s_2} \cdot \prod_{i=1}^{2} V^{s_i}(x(i), r_i; g)$$

if $|x(1) - x(2)| > 2 \max(r_1, r_2)$. It is easily verified that

$$\lim_{\substack{m\to\infty \\ m/n=\alpha}} D_m(s_1, s_2 \mid z_1, z_2, x(1), x(2)) = \frac{1}{s_1!\, s_2!} \prod_{i=1}^{2} \left[\frac{\pi^{\frac{1}{2}k} z_i g(x(i))}{\alpha\Gamma(\frac{1}{2}k+1)}\right]^{s_i} \exp\left\{-\frac{\pi^{\frac{1}{2}k} z_i g(x(i))}{\alpha\Gamma(\frac{1}{2}k+1)}\right\}$$

uniformly over any set G_1 of points in $(x(1), x(2), z_1, z_2)$ space such that $g(x)$ is uniformly continuous over the projection of G_1 on the $x(1)$ hyperplane, and over the projection of G_1 on the $x(2)$ hyperplane, and the projections of G_1 on the z_1 and z_2 axes are bounded from above.

Given any positive ϵ, we can find a subset $K(\epsilon)$ of $(x(1), x(2))$ space such that $P(X(1), X(2) \text{ in } K(\epsilon)) \geqq 1 - \epsilon$, $f(x)$ is uniformly continuous on the projection of $K(\epsilon)$ on the $x(i)$ hyperplane $(i = 1, 2)$, and $|x(1) - x(2)| > \delta(\epsilon) > 0$ at each point of $K(\epsilon)$.

Since

$$P_{m,n}(S_1 = s_1, S_2 = s_2) = \int_{E_k}\int_{E_k}\left[\int_0^\infty\int_0^\infty D_m(s_1, s_2 \mid z_1, z_2, x(1), x(2)) h_m(z_1, z_2 \mid x(1), x(2))\, dz_1\, dz_2\right] \cdot f(x(1)) f(x(2))\, dx(1)\, dx(2),$$

we have

$$\left| P_{m,n}(S_1 = s_1, S_2 = s_2) - \iint_{K(\epsilon)} \left[\int_0^\infty \int_0^\infty D_m(\quad) h_m(\quad)\, dz_1\, dz_2 \right] \cdot f(x(1))f(x(2))\, dx(1)\, dx(2) \right| \leqq \epsilon$$

for all values of m and n. From our discussion above, it can be seen that

$$\lim_{\substack{m,n\to\infty \\ m/n=\alpha}} \int_0^\infty \int_0^\infty D_m(\quad) h_m(\quad)\, dz_1\, dz_2 = \int_0^\infty \int_0^\infty \lim_{\substack{m,n\to\infty \\ m/n=\alpha}} D_m(\quad) h_m(\quad)\, dz_1\, dz_2$$

$$= \prod_{i=1}^{2} \frac{2^k \alpha f(x(i))[g(x(i))]^{s_i}}{[g(x(i)) + 2^k \alpha f(x(i))]^{s_i+1}}$$

uniformly over $K(\epsilon)$. This means

$$\lim_{\substack{m,n\to\infty \\ m/n=\alpha}} \iint_{K(\epsilon)} \left[\int_0^\infty \int_0^\infty D_m(\quad) h_m(\quad)\, dz_1\, dz_2 \right] f(x(1))f(x(2))\, dx(1)\, dx(2)$$

$$= \iint_{K(\epsilon)} \prod_{i=1}^{2} \frac{2^k \alpha f(x(i))[g(x(i))]^{s_i}}{[g(x(i)) + 2^k \alpha f(x(i))]^{s_i+1}} f(x(1))f(x(2))\, dx(1)\, dx(2).$$

This last expression differs from $Q(s_1)Q(s_2)$ by less than ϵ, and this completes the demonstration, since ϵ can be taken arbitrarily close to zero. The uniformity of approach follows from the equalities

$$\sum_{s_1,s_2} \sum P_{m,n}(S_1 = s_1, S_2 = s_2) = \sum_{s_1,s_2} \sum Q(s_1)Q(s_2) = 1,$$

the summations extending over all non-negative integers.

3. Applications. For each non-negative integer r, let $Q_m(r)$ denote the proportion of the values $S_1, \cdots, S_m$ which are equal to r. We will show that $Q_m(r)$ converges stochastically to $Q(r)$ as m, n increase with $m/n = \alpha$. Define the random variable U_i to be equal to one if S_i is equal to r, and to be equal to zero if S_i is not equal to r. Then $Q_m(r) = (U_1 + \cdots + U_m)/m$, and

$$E\{Q_m(r)\} = E\{U_1\} = P(S_1 = r),$$

Variance $\{Q_m(r)\} = (1/m)$ Var $\{U_1\} + ((m-1)/m)$ Cov $\{U_1, U_2\}$. But Variance $\{U_1\}$ is equal to $P(S_1 = r)[1 - P(S_1 = r)]$, and Cov $\{U_1, U_2\}$ is equal to $P(S_1 = r$ and $S_2 = r) - P(S_1 = r)P(S_2 = r)$. Since we showed above that S_1 and S_2 are asymptotically independent, it follows that Variance $\{Q_m(r)\}$ approaches zero. It has also been shown that $P(S_1 = r)$ approaches $Q(r)$, so it follows from Chebyshev's inequality that $Q_m(r)$ converges stochastically to $Q(r)$.

Take the case $r = 0$. In evaluating $Q(0)$, $[g(x)]^0$ is taken to be unity even if $g(x) = 0$. Then

$$\int_{E_k} \left[\frac{\sqrt{2^k \alpha}\, f(x)}{\sqrt{g(x) + 2^k \alpha f(x)}} - \frac{\sqrt{2^k \alpha}\, \sqrt{g(x) + 2^k \alpha f(x)}}{1 + 2^k \alpha} \right]^2 dx \geqq 0,$$

with equality holding if and only if $f(x) = g(x)$ almost everywhere. Expanding the integrand and integrating, we find that $Q(0) \geqq (2^k\alpha)/(1 + 2^k\alpha)$, with equality holding if and only if $f(x) = g(x)$ almost everywhere.

If it were desired to test the hypothesis that $f(x) = g(x)$ almost everywhere, a reasonable test procedure would seem to be to reject if $Q_m(0)$ is "too far" above $(2^k\alpha)/(1 + 2^k\alpha)$. In [1] it was shown that the Wald-Wolfowitz run test [2] is equivalent to rejecting the hypothesis when $Q_m(0)$ is "too large", where $Q_m(0)$ is defined using the S_i of [1]. If we accept the analogy between the S_i as defined in this paper and the S_i as defined in [1], we see that the test which rejects when $Q_m(0)$ is "too large" is a multivariate analogue of the Wald-Wolfowitz run test.

However, there are certain difficulties in the way of using the test based on $Q_m(0)$. Even if $f(x) = g(x)$ almost everywhere, so that the hypothesis is true, the distribution of $Q_m(0)$ depends on the common density function $f(x)$. This is seen from an examination of the expression for the variance of $Q_m(0)$. Thus the test based on $Q_m(0)$ is not similar to the sample space, so the level of significance must be defined as the least upper bound of probabilities of rejecting the hypothesis when it is true. To be more precise, suppose we want the probability of a type I error to be no greater than a preassigned value β $(0 < \beta < 1)$. Fixing m, n, with $m/n = \alpha$, we can write our critical region as

$$Q_m(0) \geqq (2^k\alpha)/(1 + 2^k\alpha) + \delta_m(\beta),$$

where $\delta_m(\beta)$ is chosen so that when $g(x) = f(x)$ almost everywhere,

$$\text{l.u.b.}_{f(x)} P[Q_m(0) \geqq (2^k\alpha)/(1 + 2^k\alpha) + \delta_m(\beta)] \leqq \beta.$$

We can satisfy this inequality trivially by choosing $\delta_m(\beta)$ equal to $1 - (2^k\alpha)/(1 + 2^k\alpha)$, but then our level of significance is zero, and the test is of no interest. What we want, of course, is to have $\delta_m(\beta)$ small so that the power of the test is good. One way of guaranteeing a reasonably small value of $\delta_m(\beta)$ is to limit the class of density functions under consideration in some way. One such way is to assume that all the possible density functions have the following property: given any positive ϵ, there is a positive $\gamma(\epsilon)$ so that the variation of the density function over any sphere of k-dimensional volume $\gamma(\epsilon)$ is no greater than ϵ, and $\lim_{\epsilon \to 0} \gamma(\epsilon)/\epsilon = c > 0$. Then an examination of the argument of Section 2 will show that when $g(x) = f(x)$, and $f(x)$ has the continuity property just described,

$$| E\{Q_m(0)\} - (2^k\alpha)/(1 + 2^k\alpha) | \leqq \Delta_1(c, m), \text{ Variance } \{Q_m(0)\} \leqq \Delta_2(c, m),$$

where $\Delta_1(c, m)$ and $\Delta_2(c, m)$ approach zero as m increases with $m/n = \alpha$, for any fixed positive c. Chebyshev's inequality gives

$$P[Q_m(0) \geqq (2^k\alpha)/(1 + 2^k\alpha) + \Delta_1(c, m) + t] \leqq (1/t)\Delta_2(c, m),$$

and setting $t = (1/\beta)\Delta_2(c, m)$ gives

$$P[Q_m(0) \geqq (2^k\alpha)/(1 + 2^k\alpha) + \Delta_1(c, m) + (1/\beta)\Delta_2(c, m)] \leqq \beta.$$

Thus $\delta_m(\beta) \leqq \Delta_1(c, m) + (1/\beta)\Delta_2(c, m)$, and this upper bound for $\delta_m(\beta)$ approaches zero as m increases with $m/n = \alpha$, so that if the hypothesis is not true, the probability of rejection approaches one as m, n increase. The actual computation of the functions $\Delta_1(c, m)$, $\Delta_2(c, m)$, though possible, would be quite involved, and will not be carried out here.

The test based on $Q_m(0)$ is, as has been noted, not similar to the sample space. To the author's knowledge, no test of the hypothesis under discussion which has reasonable power properties has been shown to be similar to the sample space. The quantities $Q_m(r)$ are invariant under translations and rotations of k-dimensional space, or under linear stretching of each of the k axes by the same factor. Intuitively, then, one would expect tests based on $Q_m(r)$ to be closer to similarity than such tests as the chi-square test or the Kolmogorov-Smirnov test, in the multivariate case.

Professor W. Kruskal has pointed out a lack of symmetry in the test based on $Q_m(0)$ described above, in that interchanging the roles of X and Y gives a test statistic that is not in one-one correspondence with $Q_m(0)$. Most other two-sample tests that have been proposed do not exhibit this lack of symmetry. A test which does not suffer from this lack of symmetry is one based on the average of $Q_m(0)$ and the corresponding quantity given by interchanging the roles of X and Y.

REFERENCES

[1] J. R. Blum and Lionel Weiss, "Consistency of certain two-sample tests," *Ann. Math. Stat.*, Vol. 28 (1957), pp. 242–246.

[2] A. Wald and J. Wolfowitz, "On a test whether two samples are from the same population," *Ann. Math. Stat.*, Vol. 11 (1940), pp. 147–162.