

A test for identity of multivariate means*

Leigh M. Van Valen
 Department of Biology
 University of Chicago
 915 East 57th Street
 Chicago, Illinois 60637

Received February 2, 1980; July 30, 1980

ABSTRACT: An alternative generalization of the $\underline{t}$ test reduces the multivariate case to the univariate case by a geometrical approach. The test appears to be more robust than Hotelling's $\underline{T}^2$ and can be modified for a perhaps maximally robust discriminant analysis.

*

*

*

The standard statistic for testing the identity of two mean vectors (multivariate means, centroids) is Hotelling's $\underline{T}^2$, which is one possible generalization of the univariate $\underline{t}$. Because $\underline{T}^2$ is relatively sensitive to violations of some of its assumptions, and is awkward to interpret geometrically, I give an alternative generalization of $\underline{t}$.

The approach is geometrical and is simple in principle. In the bivariate case, diagrammed in Figure 1, each sample is represented by an equiprobability ellipse at one standard deviation in every direction, and a line joins the mean vectors. Each semi-axis is the square root of the corresponding eigenvalue and is therefore one sample standard deviation along a principal axis.

Because the distance from one mean vector to its equiprobability envelope is one standard deviation in the direction toward the other mean vector, just as in every other direction, we can restrict our attention to this single dimension. This is not a matter of projecting the distribution onto the line between the mean vectors, which procedure would ordinarily give a greater spread, but of looking at the distances along the line itself. If two distributions are distinct they will be at least as distinct along this line as will the optimal projections of the distributions onto any line. We have a difference between two points, and the standard error of the location of each point along the line connecting them, irrespective of what the distributions may do elsewhere in hyperspace. We can therefore apply univariate analysis to this geometrically univariate situation.

The distance within each ellipse traversed by the line joining the mean vectors will then, by assumption, be one standard deviation of the sample in the desired direction. If $\underline{s}$ is this distance, $\underline{a}$ is the major semiaxis, $\underline{b}$ is the minor semiaxis, and θ is the angle the major axis makes with the line joining the mean vectors,

$$s^2 = \frac{a^2 b^2}{a^2 \sin^2 \theta + b^2 \cos^2 \theta} . \quad (1)$$

In the p -dimensional case, if all angles θ_i are measured from the same reference axis (say the first principal axis) and λ_i are the eigenvalues, λ_1 being that for the reference axis,

$$s^2 = \frac{1 + \sum_{i=2}^p \tan^2 \theta_i}{\lambda_1^{-1} + \sum_{i=2}^p \lambda_i^{-1} \tan^2 \theta_i} . \quad (2)$$

* * * * *

*Contribution 17, Lothlorien Laboratory of Evolutionary Biology
 Evolutionary Theory 5: 135-138 (December, 1980)

The editors thank two referees for help in evaluating this paper.

© 1980, the author

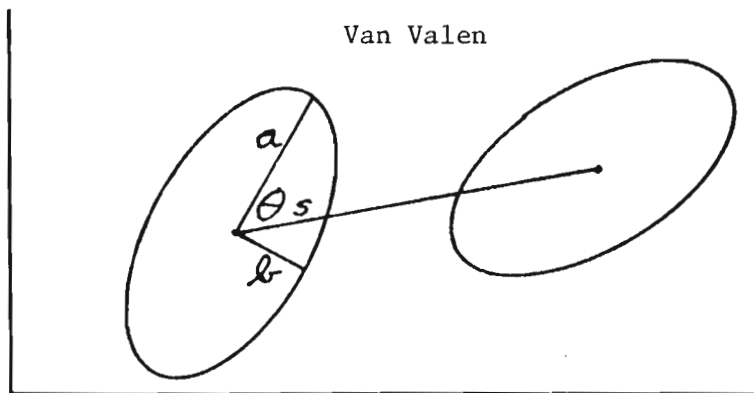


Figure 1. Diagram for the calculation of the variance (s^2) of the left sample in the direction of the line joining the mean vectors. Equiprobability ellipses represent the samples.

* * * * *

Given these variances, one for each sample, the problem of testing the identity of the mean vectors has been reduced to the unidimensional case. The sample means (mean vectors) are asymptotically normally distributed in any single direction by the univariate Central Limit Theorem applied to the marginal distributions in that direction. We have, from (2), the sample variances in the appropriate direction. The only complexity comes in the calculation of these variances, but conceptually the multivariate situation collapses into the univariate one. An ordinary two-tailed t test can therefore be applied to the absolute Euclidean distance $(\Sigma d_i^2)^{1/2}$ between the mean vectors.

In one of several variant explicit forms, then, the test is as follows:

$$t_f = \left[\frac{\sum_{i=1}^p (\bar{x}_{1i} - \bar{x}_{2i})^2}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]^{1/2}, \quad (3)$$

where s^2 is found by Equation (2), $\bar{x}_i$ is the sample mean for variable i , and n is sample size. There are

$$f = \frac{(n_1-1)(n_2-1)(n_2s_1^2 + n_1s_2^2)^2(s_1^4 + s_2^4)}{(s_1^2 + s_2^2)^2[n_2^2(n_2-1)s_1^4 + n_1^2(n_1-1)s_2^4]} \quad (4)$$

degrees of freedom (cf. Kendall and Stuart, 1973, p.154).

Because θ_i is not a component of an eigenvector, it is useful to relate θ_i to the results of a principal-components analysis so that it may be calculated easily. The data we have for each sample are (1) the cosines of the angles between the first principal axis and each original coordinate axis, and (2) the mean vector. We can take one original coordinate axis as standard. Then the difference between the mean vectors determines, for the plane of the i th original coordinate axis, the tangent of angle β_i in Figure 2. The cosine of angle ϕ_i will be negative in the case figured. $\cos\phi_i$ is the i th term in the first eigenvector. γ_i is the projection of θ_i on the plane determined by original coordinate axes 1 and i . In general,

$$\gamma_i = \frac{\pi}{2} + \phi_i - \beta_i, \quad (5)$$

where the signs of ϕ_i and β_i are those of their cosine and tangent respectively, and γ_i is measured clockwise from the first principal axis. Therefore

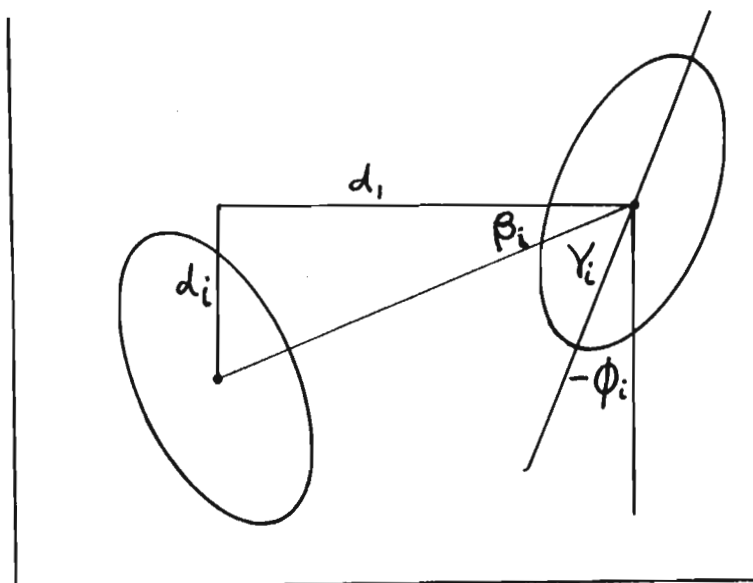


Figure 2. Diagram showing the variables needed to relate θ_i to the first eigenvector, for the sample on the right.

* * * * *

$$\tan \gamma_i = \frac{1 + \frac{d_i \sqrt{1 - \cos^2 \phi_i}}{d_1 \cos \phi_i}}{\frac{d_i}{d_1} - \frac{\sqrt{1 - \cos^2 \phi_i}}{\cos \phi_i}} \quad (6)$$

d_i is the signed difference between the means of the populations for the i th character (original coordinate axis), and d_1 is that difference for the one original coordinate axis used in each comparison.

To combine all planes,

$$\tan^2 \theta_j = \sum_{i=2}^p \tan^2 \gamma_i \quad (7)$$

My proof of Equations 5-7 is exceptionally long and exceptionally ugly. If the results prove useful someone will probably find a more elegant derivation.

The method assumes additivity and linear relationships among all variables within each population, although not necessarily the same relationships in the two populations. Any serious violation of these assumptions, if detected, can usually be corrected by transforming one or more variables before the principal-components analysis. Because the procedure is simply a test rather than an estimate, there can be no biological objection to any transformation, however artificial, that corrects such violations. The only assumptions of normality and homogeneity of covariance matrices come by the univariate t test, which is rather robust to their relaxation. T^2 is sensitive to the assumption of homogeneity of covariance matrices (cf. the results of Holloway and Dunn [1967]). Extension to more than two populations, as an analog of manova, is geometrically obvious but algebraically cumbersome.

A third test would be to extend the line joining the mean vectors and use the projections of the points in each sample onto this line as univariate distributions to which the t test could perhaps be applied, although the line here becomes a random variable itself. This test is probably a little simpler than the one described above but also uses less information and would often not give high discrimination because the projections would be larger than the actual standard deviations

along the line between the mean vectors.

Any method of discriminant analysis must depend on knowledge of the shapes of the distributions being discriminated. However, assuming multivariate normality but nothing more, the present method can clearly be applied to the problems of discrimination. I conjecture that the best (hyper)plane will pass through the point of equal multiples of $\underline{s}_1$ and $\underline{s}_2$ on the line joining the mean vectors and will bisect the angle formed by the (hyper)planes tangent to the equiprobability ellipsoids at that point. Analogous treatments are possible for other known or assumed shapes of the distributions.

Canonical analysis is also relatively nonrobust, and moreover it measures distance in units of standard deviation (assumed identical for all populations) rather than the original unit of measurement. It is useful in those rare cases where units of measurement are incommensurable among different variables, but otherwise variables can be made of equal importance (if this is desired) by direct multiplicative transformation. A principal-components analysis of the mean vectors is ordinarily to be preferred to canonical analysis.

The present method is of course also applicable to testing the position of the mean vector against a null hypothesis.

It is sometimes said that the distance between mean vectors, and even a unique value for the vectors themselves, are undefinable. When each variable is measured in the same units, however, there is no real problem except when, as is often done, the distance between mean vectors is combined with the fundamentally distinct concept of discrimination between distributions to become a function of the proportion of overlap of the distributions. I have elsewhere (Van Valen [1974]) criticized and proposed a substitute for Mahalanobis's D^2 , using the same geometrical approach as in the present note.

I thank Dr. D. Brillinger for bibliographic help.

LITERATURE CITED

- Holloway, L. N., and O. J. Dunn. 1967. The robustness of Hotelling's T^2 .
Journal of the American Statistical Association 62: 124-136.
- Kendall, M. G., and A. Stuart. 1973. The Advanced Theory of Statistics, vol.2,
ed.3. New York: Hafner. 723pp.
- Van Valen, L. 1974. Multivariate structural statistics in natural history.
Journal of Theoretical Biology 45: 235-247; 48: 501.