A BIVARIATE EXTENSION OF FISHER'S Z TEST

BY

D. D. KOSAMBI

(*Fergusson College, Poona*)

A NORMAL distribution in k variates x_1, x_2, ... x_k, each with expectation (population mean) zero is defined by the probability density $c \exp - \phi/2$, where c is always to be understood as a constant so chosen as to make the total probability equal to unity, and ϕ is a positive definite homogeneous quadratic form in the variates, *i.e.*:

$$(1) \qquad p = \frac{1}{\sigma\,(2\pi)^{\frac{k}{2}}} \int_R \cdots \int e^{-\frac{\phi}{2}} dx_1 \cdot \cdot dx_k;$$

$$\phi = \sigma^{ij} x_i x_j.$$

$$\sigma^{ij} = \sigma^{ji}; \; \sigma^{ir} \sigma_{ij} = \delta^i{}_j; \; \sigma^2 = | \sigma_{ij} |.$$

Here, we use the tensor summation convention for repeated indices and the integral is to be taken as extended over that portion of the k-space in which the variates are to lie. The coefficients σ_{ij} are to be formed by taking the normalized co-factors of the corresponding element in $|| \sigma^{ij} ||$, as usual. Alternatively, we can write $\sigma^{ij} = \frac{\partial \log \sigma^2}{\partial \sigma_{ij}}$. The form ϕ being definite, the determinant σ^2 does not vanish, and there is no theoretical difficulty in finding either σ^{ij} or σ_{ij}, the matrix of the other coefficients being given.

Suppose now that a sample of n observations be taken from such a population, the jth sample value of the variate x_i being x_{ij} Then it is known that the best[1] estimates of σ_{ij} are given by

$$(2) \qquad s_{ij} = \frac{1}{n-1} \sum_{r=1}^{n} (x_{ir} - \bar{x}_i)(x_{jr} - \bar{x}_j),$$

$$\text{where} \quad \bar{x}_i = \frac{1}{n} \sum_{j=1}^{n} x_{ij}$$

The best[1] estimate σ^2 is $s^2 = | s_{ij} |$ and of σ^{ij}, the corresponding normalized co-factors, s^{ij}.

It is well known that the quantities s_{ij} are the sample variances when $i = j$, and the sample correlations multiplied by the corresponding standard deviations when $i \neq j$. Again, s^2, the determinant of the sampling coefficients, has a strong claim to be considered as the generalized variance of the multivariate sample. The ratio of two such variances chosen from the same populations would be independent of a linear homogeneous transformation of the co-ordinates, and also of the population parameters. It is natural to ask whether the distribution of this ratio, or rather of its logarithm, has anything in common with Fisher's z, so that the z tables could be used without further ado. The answer is negative in general, but it is the purpose of this note to point out the fact that for a bivariate population ($k = 2$), such an extension is valid.

2. Following the methods given by Uspensky,[2] it is a comparatively simple matter to find the distribution of S, where

$$(3) \qquad S^2 = \det \cdot \left\{ \sum_{r=1}^{n} (x_{ir} - \bar{x}_i)(x_{jr} - \bar{x}_j) \right\}; \quad i, j, = 1, 2.$$

It is to be noted that $s^2 = S^2/(n-1)^2$. By a distribution, we mean the probability that $S^2 < t^2$, the derivative of this with respect to t being then the probability density, which is sometimes called the "distribution" by statistical writers.

For convenience of notation, let the two variates be x and y. The $\phi = ax^2 + 2bxy + cy^2$. But as we mean ultimately to consider the *ratio* of two generalized variances, which is a function independent of linear homogeneous transformations, we might as well consider the transformation to have been performed in advance which brings ϕ to its canonical form: for a positive definite form, $\phi = x^2 + y^2$. The required distribution is then given by

$$(4) \qquad p\,(t) = \frac{1}{(2\pi)^n} \int_R \cdots \int e^{-\frac{1}{2}(x_1^2 + \cdots + x_n^2 + y_1^2 + \cdots + y_n^2)} dx_1 \cdots dx_n dy_1 \cdots dy_n$$

where the region of integration R is defined by the inequality:

$$(5) \qquad S^2 \equiv \Sigma (x_i - \bar{x}) \Sigma (y_i - \bar{y}) - \{\Sigma (x_i - \bar{x})(y_i - \bar{y})\}^2 < t^2;$$

$$\text{with} \quad \bar{x} = \frac{1}{n} \Sigma x_i,$$

$$\bar{y} = \frac{1}{n} \Sigma y_i.$$

The variates x and y have the sampling values $x_1, \ldots x_n$; $y_1, \ldots y_n$, which are independent, being chosen at random by hypothesis, and the formulæ (4 – 5) are then self-evident.

For the reduction of the integral, the

treatment by Uspensky for the distribution of the correlation coefficient is rigorous and can be carried out step by step. Choosing the new variables of integration as the means $\bar{x}$, $\bar{y}$, and $n-1$ each of the differences $x_i - \bar{x}$, $y_i - \bar{y}$, and performing a suitable linear homogeneous transformation, the integral in (4) is reduced to a similar one with $n-1$ in place of n, the usual loss of a degree of freedom for measuring from the sample mean. A second transformation and one integration will reduce the integral further to

$$(6) \qquad p(t) = c \int_{\mathrm{R}} \cdots$$

$$\int e^{-\frac{1}{2}(w_1^2 + \cdots w_{n-1}^2 + \xi_1^2 + \cdots + \xi_{n-2}^2)} dw_1 \cdots dw_{n-1} d\xi_1 \cdot d\xi_{n-2} \quad ;$$

$$\mathrm{R}: (w_1^2 + \cdot\cdot + w_{n-1}^2)(\xi_1^2 + \cdot\cdot + \xi_{n-2}^2) < t^2$$

But we have the two classical formulæ of integration:

$$(7) \quad (a): \int_0^\infty e^{-x^2 - \frac{a^2}{x^2}} dx = \frac{\sqrt{\pi}}{2} e^{-2a}$$

$$(b): \int \cdots \int_{x_1^2 + \cdot\cdot x_r^2 < a} e^{-\frac{1}{2}(x_1^2 + \cdot\cdot + x_r^2)} \mathrm{F}(x_1^2 + \cdots + x_r^2)\, dx_1 \cdot\cdot dx_r = \frac{\pi^{\frac{r}{2}}}{\Gamma\left(\frac{r}{2}\right)} \int_0^a e^{-\frac{u}{2}} u^{\frac{r}{2}-1} \mathrm{F}(u) du$$

These allow us at once to write down dp/dt in the form:

$$(8) \qquad \frac{dp}{dt} = c\, e^{-t}\, t^{n-3} \text{: range } t = 0 - \infty.$$

This is, again, of the form of the integrand for the incomplete gamma function, and so, if we wish to find the distribution of the ratio of two independent sampling observations of S^2, we can proceed as usual. But it is clear that the exponent is not the usual number of degrees of freedom. In fact, the degrees of freedom, as is to be seen by comparing exponents with those in the usual formula, are now $2n-4$. Thus, we must use $(2n-4)^2$ as the divisor for S^2 in place of $(n-1)^2$. Finally, a last correction is necessary for the fact that we have used $\mathrm{S}^2 < t^2$ in place of the usual distribution, which would be the probability $\mathrm{S}^2 < t$. All of this, however, is now quite obvious, and the result can be summed up in a theorem:

If two independent samples of n, n' specimens are taken at random from a bivariate normal population, then the quantity

$$(9) \qquad z = \frac{1}{4} \log \frac{\mathrm{S}^2}{\mathrm{S}'^2} + \tfrac{1}{2} \log \frac{n'-2}{n-2}$$

$$= \log \left\{ \sqrt{\frac{\mathrm{S}}{n-2}} \Big/ \sqrt{\frac{\mathrm{S}'}{n'-2}} \right\}.$$

has the same distribution as Fisher's z for a single variate, with the degrees of freedom $2n-4$, $2n'-4$.

The distribution was known (Wilks,[3] 478) but the adjustment for the proper number of degrees of freedom, and the possibility of using Fisher's tables, have apparently been overlooked. The rule is quite as simple as for a single variate. In the usual notation we calculate the quantity $s^2_x s^2_y (1 - r^2)$, multiply by the correction factor

$$(n-1)^2 / 4(n-2)^2,$$

and take a *quarter* instead of a half of the natural logarithm of the ratio of two such sampling observations. Then, enter Fisher's tables of z as usual, but with the degrees of freedom $2n-4$ instead of $n-1$.

3. The results of the preceding section are *not* extensible to $k \geqslant 3$. The integrals do not reduce so easily, at least by any known formulæ. For example, the case $k = 3$ can be solved completely if an explicit formula for the integral from zero to infinity of $\exp - (x + a^2/x^2)$ is found. But it does not seem possible that this would allow a rigorous use to be made of the z tables.

It would be interesting to see the extended z test for $k = 2$ used for analysis of variance: say for plot experiments with two simultaneous crops sown on each plot. The test is open to the same criticisms levelled against the z test for one variate, in that it does not take the mean values into account, but tests directly on the basis of the observed variances, the hypothesis that both samples might have been drawn from the same normal population. For tests also taking the mean values into account, as in Student's t test, we have the T^2 of Hotelling and its generalizations. But for a bivariate population, the test suggested here is surely more complete than the usual method of testing the variances s^2_x, s^2_y individually, along with the correlation coefficient r.

[1] J. L. Coolidge, *Theory of Probability*, Oxford, 1924, p. 82.

[2] J. V. Uspensky, *Introduction to the Mathematical Theory of Probability*, 1937, p. 332, *et. seq.*

[3] S. S. Wilks, "Certain Generalizations in the Analysis of Variance," *Biometrika*, 1932, **24**, 471–494.