DIFFERENCE OF PROPORTIONS OF DEVIATES IN SELECTION

Ying Y. Wang

BU-234-M January, 1967

ABSTRACT

In problems involving genetic selection, one may wish to evaluate the proportion of genetic deviates in some specified region. This paper shows that such a problem may be reduced to the more familiar form of the distribution of a weighted sum of independent chi-squares, and that available tables and computational formulae can be used readily in the actual computations.

Department of Plant Breeding and Biometry, Cornell University, Ithaca, New York.

DIFFERENCE OF PROPORTIONS OF DEVIATES IN SELECTION

Ying Y. Wang

BU-234-M January, 1967

In experiments involving genetic selection, one may wish to estimate the difference in the proportion of deviates in some specified region for a population with genetic and environmental variation and for a population with only environmental variation. We shall, for want of a better term, call this difference in the positive direction the proportion of plus genetic deviates. If the selection is for the smaller values we shall call this the proportion of negative genetic deviates. Because of symmetry in the normal distribution, we need only consider the plus genetic deviates. Thus, the greater the proportion of plus genetic deviates in a prescribed region, the greater the chances of success in selection if the genetic variation exists.

Thompson and Federer have shown that under normality assumption, a simple procedure exists for the univariate problem. Let the variable X be normally distributed with density function denoted by $\phi_1 = \phi(x;\ \mu,\ \sigma_g^2 + \sigma_e^2)$, where σ_g^2 is the genetic variance, and σ_e^2 the phenotypic variance. If there were no genetic variation so that X has density $\phi_2 = \phi(x;\ \mu,\ \sigma_e^2)$, then the proportion of plus genetic deviates is the difference between the upper tail areas of the two distributions

$$D(x_0) = H_1(x_0) - H_2(x_0)$$

where x_0 is defined by

$$\phi_1(x_0) = \phi_2(x_0) \ .$$

Department of Plant Breeding and Biometry, Cornell University, Ithaca, New York.

Thus

$$D(x_0) = H\left(\frac{x_0 - \mu}{\sigma_e}\right) - H\left(\frac{x_0 - \mu}{\sqrt{\sigma_g^2 + \sigma_e^2}}\right)$$

where $H(x)$ is the right tail area of the standard normal.

The following is a multivariate generalization of the problem under the assumption of normality. Let $\underline{X}$ be a p-dimensional normal vector with joint density $\phi_1(\underline{x}) = \phi(\underline{x};\ \underline{\mu},\ \Sigma_g + \Sigma_e)$, if $\Sigma_g = 0$, then $\phi_2(\underline{x}) = \phi(\underline{x};\ \underline{\mu},\ \Sigma_e)$, and the proportion of plus genetic deviates is

$$D(R) = H_1(R) - H_2(R)$$

where H_i, $i = 1, 2$, is the probability content of the normal density ϕ_i in the region R defined by

$$R = \left\{\underline{x} \mid \phi_1(\underline{x}) \geq \phi_2(\underline{x}),\ \underline{x} \geq \underline{\mu}\right\} .$$

Since

$$\frac{\phi_1(\underline{x})}{\phi_2(\underline{x})} = \left(\frac{|\Sigma_e|}{|\Sigma_e + \Sigma_g|}\right)^{p/2} e^{-\frac{1}{2}(\underline{x} - \underline{\mu})'\left((\Sigma_e + \Sigma_g)^{-1} - \Sigma_e^{-1}\right)(\underline{x} - \underline{\mu})}$$

$$R = \left\{\underline{x} \mid (\underline{x} - \underline{\mu})'\left(\Sigma_e^{-1} - (\Sigma_e + \Sigma_g)^{-1}\right)(\underline{x} - \underline{\mu}) \geq p\ \ell g(|\Sigma_e + \Sigma_g|/|\Sigma_e|),\ \underline{x} \geq \underline{\mu}\right\}.$$

If Σ_e and Σ_g are non-singular, there exists a non-singular matrix C such that

$$C\ \Sigma_e\ C' = I$$

$$C(\Sigma_e + \Sigma_g)C' = \Lambda ,$$

and the diagonal matrix Λ has elements λ_i which are roots of the equation

$$|\Sigma_g - \lambda\Sigma_e| = 0 .$$

Let $p\ \ell g(|\Sigma_e + \Sigma_g|/|\Sigma_e|) = p\ \ell g|I + \Lambda| = p \sum_{i=1}^{k} \ell g(1 + \lambda_i) = q$.

Transforming $\underline{x}$ to standard spherical normal $\underline{y} = C(\underline{x} - \underline{\mu})$

$$H_2(R) = \int_R \phi_1(\underline{x})d\underline{x} = K \int_{R_2} e^{-\underline{y}'\underline{y}/2}\, d\underline{y}$$

$$= H(R_2)$$

where

$$R_2 = \left\{\underline{y} \geq \underline{0}\ ,\quad \underline{y}'\left(\Lambda (I + \Lambda)^{-1}\right)\underline{y} \geq q\right\} ,$$

$$K = (2\pi)^{-p/2} ,$$

and similarly

$$H_1(R) = H(R_1)$$

$$R_1 = \left\{\underline{y} \geq \underline{0}\ ,\quad \underline{y}'\Lambda\,\underline{y} \geq q\right\} .$$

Due to symmetry,

$$H(R_1) = 2^{-p}\ \Pr\left\{\sum_1^p \lambda_i y_i^2 \geq q\right\}$$

$$H(R_2) = 2^{-p}\ \Pr\left\{\sum_1^p \lambda_i y_i^2/(1 + \lambda_i) \geq q\right\} ,$$

therefore they are just the upper tail probability contents of weighted sum of independent chi-squares, or definite quadratic forms. This reduces the problem to the familiar form discussed by Ruben and others.

In carrying out the computations for D(R), we note that it is only necessary to find the roots λ_i from the equation

$$|\Sigma_g - \lambda\Sigma_e| = 0,$$

to obtain q and the relevant quadratic forms. Grad and Solomon have provided tables for the distribution of definite quadratic forms for p = 2, 3, for various sets of λ_i. DiDonato and Jarnagin gave formulae for numerical evaluation of the integrals. For the special case when an even number of the λ_i's are alike, the distribution is a linear combination of gamma functions.

For the bivariate case Ruben cited a useful relationship, due to Kleinecke, between the weighted sum of two chi-squares each with one degree of freedom and the non-central chi-square with two degrees of freedom,

$$\Pr\left\{ay_1^2 + by_2^2 \geq c\right\} = G_v(u^2) - G_u(v^2)$$

where

$$u = \tfrac{1}{2}\left((c/a)^{\frac{1}{2}} + (c/b)^{\frac{1}{2}}\right)$$

$$v = \tfrac{1}{2}\left|(c/a)^{\frac{1}{2}} - (c/b)^{\frac{1}{2}}\right|$$

and G_v is the cumulative distribution function of a non-central chi-square with two degrees of freedom and non-central parameter v. One may thus use the tables of non-central chi-squares, as given by Fix, Burington and May, to obtain D(R).

A numerical example follows:

$$\Sigma_g = \begin{pmatrix} .105 & -.099 \\ -.099 & .210 \end{pmatrix}$$

$$\Sigma_e = \begin{pmatrix} .806 & -.344 \\ -.344 & 7.840 \end{pmatrix},$$

so

$$(\lambda_1, \lambda_2) = (.01470, \quad .13436)$$

and

$$q = .2814 .$$

Using Grad and Solomon's table, the probabilities are approximately

$$Q_1 = \Pr\left\{.01470\ y_1^2 + .13436\ y_2^2 \geq .2814\right\} \approx .16$$

$$Q_2 = \Pr\left\{.01449\ y_1^2 + .11844\ y_2^2 \geq .2814\right\} \approx .14$$

$$D(R) = \tfrac{1}{4}(Q_1 - Q_2) \approx .005$$

while individually for each variable

$$D_1(R_1) = H(.96) - H(1.01) = .010$$

$$D_2(R_2) = H(.86) - H(\ .87) = .003$$

References

Burington, R. S. and May, D. C. (1953). Probability and Statistics. Handbook Publishers, Ohio. 102-105.

DiDonato, A. R. and Jarnagin, M. P. (1962). A method for computing the circular coverage function. Math. Comp. 16, 347-355.

Fix, E. (1949). Table of Non-Central X^2. University of California.

Grad, A. and Solomon, H. (1955). Distribution of quadratic forms and some applications. Ann. Math. Statist. 26, 464-477.

Ruben, H. (1961). Probability content of regions under spherical normal distributions. III. The bivariate normal integral. Ann. Math. Statist. 32, 171-186.

Thompson, K. H. and Federer, W. T. (1962). Estimation of the proportion of genetic deviates under the assumptions of normality. Biometrics Unit, Cornell University, Mimeo BU-143-M.