THE FREQUENCY DISTRIBUTION OF DEVIATES FROM MEANS AND REGRESSION LINES IN SAMPLES FROM A MULTIVARIATE NORMAL POPULATION

By D. J. Finney

Oxford University, England

1. Summary. The joint frequency distribution has been found for any set of the $(n - k)$ deviates from their sample mean of each of the t variates in a sample from a multivariate normal population. Expressions for the variance of any single deviate in this distribution, the correlation coefficient between any pair of deviates, and certain partial correlation coefficients between any pair have also been obtained.

These results have been generalized so as to include the corresponding properties of deviates from a set of t multiple linear regression equations estimated from the sample, the m independent variates being the same for each of the t dependent.

2. Introduction. Some years ago, Irwin published results relating to the frequency distribution of the deviations of individual observations from the mean of a sample drawn from a normal population (see [1]). He derived an expression for the joint distribution of any number of these deviates, which distribution is always of the normal multivariate form, and thence obtained the total and partial correlation coefficients between any pair of the deviates.

The purpose of this paper is to discuss the generalization of Irwin's problem, firstly to the properties of the deviates of individual observations from the mean in a sample from a multivariate normal population and secondly to the properties of deviates from a regression equation instead of from a mean. So far as is known to the writer Irwin's results are of little practical importance, and these generalizations are probably of no practical value whatsoever. Nevertheless, they have some interest as additions to the knowledge of the mathematical properties of the normal frequency function, and for that reason alone they are put on record here.

3. Deviations from the sample mean. Irwin based his discussion on a normal population with mean m and variance σ^2, but the algebra is simplified a little, without any real loss of generality in the final results, if, by means of a preliminary transformation, these parameters of position and scale are made zero and unity respectively. The multivariate normal distribution in the t variates $_iy$, $(i = 1, 2, \cdots t)$, each with mean zero and variance unity, has the frequency function

$$(1) \qquad \frac{1}{(2\pi)^{\frac{1}{2}t} R^{\frac{1}{2}}} \exp\left\{-\frac{1}{2R}\rho^{ij}\,{}_iy\,{}_jy\right\},$$

where $i, j = 1, 2, \cdots t$; ρ^{ij} is the cofactor of the element ρ_{ij} in the determinant of population correlation coefficients

$$R = \begin{vmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1t} \\ \rho_{12} & 1 & \rho_{23} & \cdots & \rho_{2t} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{1t} & \rho_{2t} & \rho_{3t} & \cdots & 1 \end{vmatrix} \tag{2}$$

A summation convention for the affixes i, j is understood throughout this paper, except when the contrary is explicitly stated.

Let $({}_iy_p)$ represent a sample of n independent sets of values of the t variates randomly selected from the population, $(p = 1, 2, \cdots, n)$. Then the element of probability for the sample is

$$\frac{1}{(2\pi)^{\frac{1}{2}nt} R^{\frac{1}{2}n}} \exp\left\{-\frac{1}{2R}\rho^{ij}\sum_p {}_iy_p\,{}_jy_p\right\} \prod_{\substack{i=1,\cdots,t \\ p=1,\cdots,n}} \{d({}_iy_p)\}. \tag{3}$$

If ${}_i\bar{y}$ is the mean of the n sample values of ${}_iy$, the deviates from the mean are $({}_iY_p)$, where

$${}_iY_p = {}_iy_p - {}_i\bar{y} = \sum_q \left(\delta_{pq} - \frac{1}{n}\right) {}_iy_q, \tag{4}$$

the summation being taken over $q = 1, 2, \cdots n$ with

$$\delta_{pq} = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q. \end{cases}$$

Now the ${}_iY$ are linear combinations of normally distributed variates, and are therefore themselves normally distributed. Clearly

$$E({}_iY_p) = 0 \tag{5}$$

and, from an expansion by means of equation (4) using

$$\begin{aligned} E({}_iy_p\,{}_jy_q) &= \delta_{pq}\rho_{ij}, \\ E({}_iY_p\,{}_jY_q) &= \left(\delta_{pq} - \frac{1}{n}\right)\rho_{ij}, \end{aligned} \tag{6}$$

where $\rho_{ii} = 1$ (not summed). Consequently the variance of any one deviate is

$$\sigma^2({}_iY_p) = \frac{n-1}{n}, \tag{7}$$

and the correlation coefficient between any pair is

$$\rho({}_iY_p, {}_jY_q) = \frac{n\delta_{pq} - 1}{n-1}\rho_{ij}. \tag{8}$$

Equation (7) and equation (8) for the particular case of $i = j$ agree with the well-known results that Irwin has already given as equations (10) of his paper.

For any i, only $(n-1)$ of the deviates ${}_iY_p$ are functionally independent. The joint distribution of these for $p = 1, 2, \cdots, (n-k)$ may be obtained from an inversion of the matrix of correlation coefficients. If Δ is the determinant of this matrix and $\Delta({}_iY_p, {}_jY_q)$ the cofactor corresponding to the two elements specified, this inversion shows that

$$\frac{\Delta({}_iY_p, {}_jY_q)}{\Delta} = \left(\delta_{pq} + \frac{1}{k}\right)\frac{\rho_{ij}}{R} \cdot \frac{n-1}{n}. \tag{9}$$

The joint distribution is therefore

$$\text{const.} \times \exp\left\{-\frac{1}{2R}\rho^{ij} \sum_{p \le q \le n-k}\left(\delta_{pq} + \frac{1}{k}\right){}_iY_p\,{}_jY_q\right\}\prod(dY). \tag{10}$$

Now Δ may be evaluated as

$$\Delta = \left(\frac{n}{n-1}\right)^{t(n-k-1)}\left(\frac{k}{n-1}\right)^{t} R^{n-k},$$

and the constant multiplier in equation (10) is therefore

$$\frac{\left(\frac{n}{k}\right)^{\frac{1}{2}t}}{\{(2\pi)^t R\}^{\frac{1}{2}(n-k)}}. \tag{11}$$

From equation (9), the partial correlation coefficient between any two of the variates in the distribution (10), the remaining $t(n-k)-2$ being held constant, is written down as

$$\text{partial correlation coefficient between } {}_iY_p, {}_jY_q = -\frac{k\delta_{pq}+1}{k+1} \cdot \frac{\rho^{ij}}{(\rho^{ii}\rho^{jj})^{\frac{1}{2}}}; \tag{12}$$

the summation convention is suspended for this equation.

4. Deviations from regression equations. The results obtained in section three may be generalized so as to relate to the frequency distribution of deviates from linear or polynomial regression equations instead of to deviates from means. Suppose that there are m independent variates x^α, $(\alpha = 1, 2, \cdots, m)$, which take values x_p^α corresponding to the sample observations ${}_iy_p$; polynomial regressions may be included by taking powers of an x as separate variates. If a conventional variate x^0, whose value is always unity, be introduced, the regression equation of ${}_iy$ on x^α, $(\alpha = 0, 1, 2 \cdots, m)$, may be written

$${}_i\eta = {}_ib^\alpha x^\alpha, \tag{13}$$

where a summation convention is understood for $\alpha = 0, 1, \cdots, m$ and the regression coefficients are the solutions of the normal equations.

$${}_ib^\alpha \sum_p x_p^\alpha x_p^\beta = \sum_p {}_iy_p x_p^\beta. \tag{14}$$

Write

$$B^{\alpha\beta} = \sum_p x_p^\alpha x_p^\beta \tag{15}$$

and let $(B_{\alpha\beta})$ be the inverse matrix of $(B^{\alpha\beta})$.

Then the solutions of equations (14) are

$$ {}_ib^\alpha = B_{\alpha\beta} \sum_p {}_iy_p x_p^\beta . \tag{16}$$

If the deviation of ${}_iy_p$ from the regression equation (13) is ${}_iZ_p$, then

$$\begin{aligned} {}_iZ_p &= {}_iy_p - {}_i\eta_p \\ &= \sum_q (\delta_{pq} - B_{\alpha\beta} x_p^\alpha x_q^\beta)\, {}_iy_q , \end{aligned} \tag{17}$$

the summation for q being over $q = 1, 2, \cdots, n$. As for equation (5),

$$E({}_iZ_p) = 0. \tag{18}$$

Also

$$E({}_iZ_p {}_jZ_q) = (\delta_{pq} - B_{\alpha\beta} x_p^\alpha x_q^\beta)\rho_{ij}. \tag{19}$$

since by definition

$$B^{\alpha\beta} B_{\alpha\gamma} = \delta_{\beta\gamma}.$$

Write now θ for the square matrix of $(m + 1)$ rows and columns whose elements are the $B^{\alpha\beta}$, and X_p for the single column matrix of values x corresponding to the pth observation; i.e.

$$\theta = (B^{\alpha\beta}) \tag{20}$$

and

$$X_p = \begin{pmatrix} x_p^0 \\ x_p^1 \\ x_p^2 \\ \cdot \\ \cdot \\ \cdot \\ x_p^m \end{pmatrix} \tag{21}$$

Write also

$$\theta_{p,q,r,\ldots} = \theta - X_p X_p' - X_q X_q' - X_r X_r' - \cdots \tag{22}$$

Then

$$|\, \theta_p \,| = |\, \theta \,| \cdot (1 - B_{\alpha\beta} x_p^\alpha x_p^\beta),$$

and

$$|\, \theta_{pq} \,| - |\, \theta_{pq} + X_p X_q' \,| = - |\, \theta \,| \cdot B_{\alpha\beta} x_p^\alpha x_q^\beta .$$

Hence, from equation (19), the variance of a deviate may be written

$$\sigma^2({}_iZ_p) = \frac{|\theta_p|}{|\theta|}, \tag{23}$$

and the correlation coefficient between any pair of deviates is

$$\rho({}_iZ_p, {}_jZ_q) = \begin{cases} \rho_{ij} & (p = q) \\ \rho_{ij} \dfrac{|\theta_{pq}| - |\theta_{pq} + X_p X_q'|}{\{|\theta_p| \cdot |\theta_q|\}^{\frac{1}{2}}} & (p \neq q) \end{cases} \tag{24}$$

For any i, only $(n - m - 1)$ of the deviates ${}_iZ_p$ are functionally independent. The joint distribution of these for $p = 1, 2, \cdots, (n - k)$ and any $k \geq m + 1$ may be found by inversion of the matrix of correlation coefficients obtained from equation (24). The multiplier of the exponential in this distribution of $t(n - k)$ variables is

$$\frac{|\theta|^{\frac{1}{2}t(n-k)}}{(2\pi)^{\frac{1}{2}t(n-k)} R^{\frac{1}{2}(n-k)} D^{\frac{1}{2}t}},$$

where

$$D = \begin{vmatrix} |\theta_1| & |\theta_{12}| - |\theta_{12} + X_1X_2'| & \cdots & |\theta_{1,n-k}| - |\theta_{1,n-k} + X_1X_{n-k}'| \\ |\theta_{12}| - |\theta_{12} + X_1X_2'| & |\theta_2| & \cdots & |\theta_{2,n-k}| - |\theta_{2,n-k} + X_2X_{n-k}'| \\ \cdots & \cdots & \cdots & \cdots \\ |\theta_{1,n-k}| - |\theta_{1,n-k} + X_1X_{n-k}'| & |\theta_{2,n-k}| - |\theta_{2,n-k} + X_2X_{n-k}'| & \cdots & |\theta_{n-k}| \end{vmatrix}$$

Since θ is positive definite, there exists a non-singular matrix K such that

$$K\theta K' = I.$$

Then the X_p may be transformed to new column matrices W_p by

$$KX_p = W_p = \begin{pmatrix} w_p^0 \\ w_p^1 \\ w_p^2 \\ \cdot \\ \cdot \\ \cdot \\ w_p^m \end{pmatrix}$$

and consequently

$$X_p = K^{-1}W_p.$$

It follows that

$$|\theta_p| = |\theta| \cdot |I - W_pW_p'|,$$

which may be reduced to the form

$$|\theta_p| = |\theta| \cdot (1 - w_p^\alpha w_p^\alpha).$$

Similarly

$$|\theta_{pq}| - |\theta_{pq} + X_p X'_q| = -|\theta| w^\alpha_p w^\alpha_q.$$

Hence

$$D = |\theta|^{n-k} \begin{vmatrix} 1 - w^\alpha_1 w^\alpha_1 & -w^\alpha_1 w^\alpha_2 & \cdots & -w^\alpha_1 w^\alpha_{n-k} \\ -w^\alpha_1 w^\alpha_2 & 1 - w^\alpha_2 w^\alpha_2 & \cdots & -w^\alpha_2 w^\alpha_{n-k} \\ \cdots & \cdots & \cdots & \cdots \\ -w^\alpha_1 w^\alpha_{n-k} & -w^\alpha_2 w^\alpha_{n-k} & \cdots & 1 - w^\alpha_{n-k} w^\alpha_{n-k} \end{vmatrix}$$

This may be transformed into

$$D = |\theta|^{n-k} \begin{vmatrix} & & & W'_1 \\ & & & W'_2 \\ & & & \cdot \\ & I_{n-k} & & \cdot \\ & & & \cdot \\ & & & W'_{n-k} \\ W_1 & W_2 \cdots & W_{n-k} & I_{m+1} \end{vmatrix}$$

$$= |\theta|^{n-k} \cdot |I - W_1 W'_1 - W_2 W'_2 - \cdots - W_{n-k} W'_{n-k}|$$

$$= |\theta|^{n-k-1} \cdot |\theta_{1,2,\cdots,(n-k)}|.$$

Thus, finally, the constant in the distribution is found to be

$$\frac{1}{\{(2\pi)^t R\}^{\frac{1}{2}(n-k)}} \left\{\frac{|\theta|}{|\Omega_k|}\right\}^{\frac{1}{2}t}, \tag{25}$$

in which Ω_k has been written for $\theta_{1,2,\cdots,(n-k)}$, a matrix of the same form as θ but calculated from the last k sets of observations only.

The cofactors of the matrix of correlation coefficients, required for the coefficients of the quadratic form in the distribution, can be derived in a similar manner. The distribution may be written

$$\frac{1}{\{(2\pi)^t R\}^{\frac{1}{2}(n-k)}} \left\{\frac{|\theta|}{|\Omega_k|}\right\}^{\frac{1}{2}t} \cdot \exp.\left\{-\frac{1}{2R}\rho^{ij} \sum_{p \le q \le n-k} \left(\delta_{pq} - 1 + \frac{|\Omega_k + X_p X'_q|}{|\Omega_k|}\right) {}_iZ_p \, {}_jZ_q\right\} \Pi\,(dZ), \tag{26}$$

of which the distribution (10) is easily seen to be the particular case for $m = 0$.

From (26), the partial correlation coefficient between any pair of deviates, ${}_iZ_p$ and ${}_jZ_q$, may be written down as

$$-\frac{|\Omega_k + X_p X'_q| + (\delta_{pq} - 1)|\Omega_k|}{\{|\Omega_k + X_p X'_p| \cdot |\Omega_k + X_q X'_q|\}^{\frac{1}{2}}} \frac{\rho^{ij}}{(\rho^{ii}\rho^{jj})^{\frac{1}{2}}}; \tag{27}$$

in this expression the summation convention is again suspended.

REFERENCE

[1] J. O. Irwin, "On the frequency distribution of any number of deviates from the mean of a sample from a normal population and the partial correlations between them," *Roy. Stat. Soc. Jour.*, Vol. 92 (1929), pp. 580–584.