Jour. Ind. Soc. Ag. Statistics
49 (Golden Jub. No.), 1996-97 : 91–98

Some Applications of Concomitants of Order Statistics

H. A. David
Iowa State University, Ames, USA

SUMMARY

Let (X_i, Y_i), $i = 1,......,n$ be independent pairs of variates. If $X_{r:n}$ denotes the r-th ordered X- variate, then the Y- variate paired with $X_{r:n}$ is termed the concomitant of the r-th order statistic and denoted by $Y_{[r:n]}$. After some basic theory is reviewed, the paper gives applications of concomitants of order statistics to (a) selection, (b) estimation of regression and correlation coefficient, (c) the paired t-test under artificial pairing, (d) double sampling and ranked-set sampling, and (e) selection through an associated variable.

Key words : Double sampling, Induced order statistics, Paired t-test, Regression estimator, Selection.

1. *Introduction*

Suppose the top k out of n rams, as judged on a measurement X, are selected for breeding. Here X may represent wool quality or some composite measure of the ram's performance. Then the performance of the ram R_r of rank r ($r = n-k+1, ..., n$) is represented by the r-th order statistic $X_{r:n}$. The properties of $X_{r:n}$ are well understood. Our interest is in a representation of, say, the wool quality of an offspring of R_r. If Y represents wool quality in the offspring generation in the absence of selection, we denote the wool quality of an R_r-offspring by $Y_{[r:n]}$ and call it the *concomitant of the r-th order statistic* (David [4], [5]) or the *induced r-th order statistic* (Bhattacharya [3]).

The general situation can now be stated. Let (X_i, Y_i), $i = 1, ..., n$, be a random sample from a bivariate distribution with cumulative distribution function (cdf) $F(x, y)$. Unless the X_i and Y_i are independent, the ordering of the X's will affect the distribution of the associated Y's. This situation arises whenever we select on a measurement X and are interested in an associated measurement Y. For example, in addition to the above, X may represent a score in a preliminary test and Y in a later test, or X may represent an inexpensive rough measurement and Y the corresponding refined more expensive measurement. In the latter example, we may reduce the number of expensive measurements on the basis of the inexpensive measurements.

After providing some basic theory, we give applications of concomitants of order statistics to (a) selection, (b) estimation of regression and correlation coefficient, (c) the paired t-test under artificial pairing, (d) double sampling and ranked-set sampling, and (e) selection through an associated variable.

For a comprehensive review of the subject, see David and Nagaraja [7]. Apart from additional theory, both finite-sampling and asymptotic, and additional applications, they also deal with multivariate generalizations.

2. *The Model and Basic Theory*

For most of this paper, we will assume that, possibly after transformations, X_i and Y_i $(i = 1,...,n)$ are bivariate normal, with means μ_X, μ_Y, variances σ_X^2, σ_Y^2 and correlation coefficient ρ. As is well known, Y_i may be represented as

$$Y_i = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X_i - \mu_X) + Z_i, \tag{1}$$

where X_i and Z_i are independent variates and $Z_i \sim N(0, \sigma_Y^2(1-\rho^2))$. It is easy to verify that X_i and Y_i have the stated joint distribution.

Ordering on the X_i, we have for $r = 1, ..., n$

$$Y_{[r:n]} = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X_{r:n} - \mu_X) + Z_{[r]}, \tag{2}$$

where $Z_{[r]}$ denotes the particular Z_i associated with $X_{r:n}$. In view of the independence of all the X_i and the Z_i, we see that (a) the $X_{r:n}$ are independent of the $Z_{[r]}$ and (b) ordering the X_i cannot affect the distribution of the Z_i, so that the $Z_{[r]}$, like the Z_i, are independent $N(0, \sigma_Y^2(1-\rho^2))$ variates. It follows from (2) that $Y_{[r:n]}$ is strictly normal only when $\rho = 0$.

The first two moments of the $Y_{[r:n]}$ are easily obtained from (2) in terms of the quantities

$$\alpha_{r:n} = E\left[\frac{X_{r:n} - \mu_X}{\sigma_X}\right] \text{ and } \beta_{rs:n} = \text{cov}\left[\frac{X_{r:n} - \mu_X}{\sigma_X}, \frac{X_{s:n} - \mu_X}{\sigma_X}\right] \tag{3}$$

$r, s = 1, ..., n$. The $\alpha_{r:n}$ are tabulated by, e.g., Harter [8] for $n = 2(1)100(25)250(50)400$ and the $\beta_{rs:n}$ by Tietjen *et al.* [14] for $n \leq 50$. We have

$$E(Y_{[r:n]}) = \mu_Y + \rho\sigma_Y\alpha_{r:n}, \tag{4}$$

$$\text{var}\,(Y_{[r:n]}) = \sigma_Y^2\,(\rho^2\,\beta_{r:n} + 1 - \rho^2), \tag{5}$$

$$\text{cov}\,(Y_{[r:n]}, Y_{[s:n]}) = \rho^2\sigma_Y^2\beta_{rs:n}, \quad r \neq s. \tag{6}$$

Thus, within the limitations of the tables, one can find the exact mean and variance of any linear function of the $Y_{[r:n]}$. Such a function can be shown to be asymptotically normally distributed for $|\rho| < 1$.

For general bivariate distributions, put $m(x) = E(Y \mid X = x)$ and $\sigma^2(x) = \text{var}(Y \mid X = x)$ (Bhattacharya [3]). Then in generalization of (4)-(6), we have (Yang [17])

$$E(Y_{[r:n]}) = E(m(X_{r:n})) \tag{4'}$$

$$\text{var}\,(Y_{[r:n]}) = \text{var}\,[m(X_{r:n})] + E[\sigma^2(X_{r:n})], \tag{5'}$$

$$\text{cov}\,(Y_{[r:n]}, Y_{[s:n]}) = \text{cov}\,[m(X_{r:n}), m(X_{s:n})], \; r \neq s \tag{6'}$$

3. *Applications*

(a) *Selection*

If from n individuals the $k(<n)$ with the largest X-values are chosen, the average improvement, measured in standard deviation units, of the selected individuals is the *selection differential*

$$D_{k:n} = \frac{1}{k}\sum_{i=n-k+1}^{n} \frac{X_{i:n} - \mu_X}{\sigma_X} \tag{7}$$

Similarly, to measure the improvement, due to the same selection, on an associated measurement Y, one can define the *induced selection differential*

$$D_{[k:n]} = \frac{1}{k}\sum_{i=n-k+1}^{n} \frac{Y_{[i:n]} - \mu_Y}{\sigma_Y} \tag{8}$$

The expectation and variance of $D_{[k:n]}$ now follow immediately from (4) to (6) or (4′) to (6′). Thus, when (X,Y) is bivariate normal, we have from (2),

$$D_{[k:n]} = \rho D_{k:n} + \frac{1}{k}\sum_{i=n-k+1}^{n} \frac{Z_{[i]}}{\sigma_Y} \tag{9}$$

and hence

$$E(D_{[k:n]}) = \rho E(D_{k:n}), \tag{10}$$

$$\mathrm{var}(D_{[k:n]}) = \rho^2 \mathrm{var}(D_{k:n}) + \frac{1}{k}(1-\rho^2) \tag{11}$$

We see that the gain due to selection on X is attenuated for Y by the factor ρ. Exact values of the mean and variance of $D_{[k:n]}$ can be obtained with the help of tables of the first two moments of the order statistics. The finite-sample distribution of $D_{k:n}$ has recently been studied by Andrews [2]. For the asymptotic distribution of $D_{[k:n]}$ see Nagaraja [11].

Returning to the original example, if R_r has n_r surviving offspring, then the expected average gain in wool quality of the offspring population is, under bivariate normality,

$$\rho \sum_{i=n-k+1}^{n} n_i \alpha_{i:n} \Big/ \sum n_i$$

(b) Estimation of Regression and Correlation Coefficient

If the regression of Y on the non-stochastic variable x is linear :

$$E(Y \mid x) = \alpha + \beta x, \tag{12}$$

then β may be estimated by the ratio statistic

$$b' = \frac{\bar{Y}'_{[k:n]} - \bar{Y}_{[k:n]}}{\bar{x}'_{k:n} - \bar{x}_{k:n}} \tag{13}$$

where

$$\bar{x}'_{k:n} = \frac{1}{k}\sum_{i=1}^{k} x_{n+1-i:n} \qquad \bar{x}_{k:n} = \frac{1}{k}\sum_{i=1}^{k} x_{i:n}$$

and

$$\bar{Y}'_{[k:n]} = \frac{1}{k}\sum_{i=1}^{k} Y_{[n+1-i:n]} \qquad \bar{Y}_{[k:n]} = \frac{1}{k}\sum_{i=1}^{k} Y_{[i:n]}$$

If X is stochastic, we may interpret (12) as conditional on X = x and have from (13)

$$E(b' \mid x_1, \ldots, x_n) = \beta \tag{14}$$

Since (14) holds whatever the x_i, it also holds uncoditionally; that is,

$$B' = \frac{\overline{Y}'_{[k:n]} - \overline{Y}_{[k:n]}}{\overline{X}'_{k:n} - \overline{X}_{k:n}} \tag{15}$$

is also an unbiased estimator of β. Note that this result does not require either the X's or the Y's to be identically distributed or even to be independent. Barton and Casley [2] show that B′ has an efficiency of the 75-80% when (X_i, Y_i), $i = 1, ..., n$, is a random sample from a bivariate normal, provided k is chosen as about 0.27 n.

Since $\rho = \beta\sigma_X/\sigma_Y$, (15) suggests

$$\begin{aligned}\hat{\rho}' &= B' \frac{(\overline{X}'_{k:n} - \overline{X}_{k:n})/c_{n,x}}{(\overline{Y}'_{k:n} - \overline{Y}_{k:n})/c_{n,x}} \\ &= \frac{(\overline{Y}'_{[k:n]} - \overline{Y}_{[k:n]})/c_{n,x}}{(\overline{Y}'_{k:n} - \overline{Y}_{k:n})/c_{n,y}}\end{aligned}$$

as an estimator of ρ, where $c_{n,x} = E(\overline{X}'_{k:n} - \overline{X}_{k:n})/\sigma_X$, etc. If X and Y have the same marginal distributional form (e.g., both normal), $\hat{\rho}'$ simplifies to

$$\hat{\rho}' = \frac{\overline{Y}'_{[k:n]} - \overline{Y}_{[k:n]}}{\overline{Y}'_{k:n} - \overline{Y}_{k:n}} \tag{16}$$

This estimator has been suggested by Tsukibayashi [15] for k = 1, when the denominator is just the range of the Y_i, and also for a mean range denominator. He points out that (16) can be calculated even if only the ranks of the X's are available. The properties of $\hat{\rho}'$ for k = 1 are currently being investigated in detail (Tsukibayashi [16]).

(c) The Paired t- Test under Artificial Pairing

When a paired t-test cannot be based on natural pairing, such as in before and after or twin experiments, the following procedure is often used : 2n individuals are paired on the basis of closeness or prior related measurements, such as birth weight in experiments on new-born animals. The pairs then correspond to the measurements ($x_{2i:2n}$, $x_{2i-1:2n}$), i = 1, ..., n. The order within pairs is randomized before the two treatments, one of which may be a control, are applied. When a t-test is performed on the signed differences $D_i = \pm (Y_{[2i:2n]} - Y_{[2i-1:2n]})$ of the experimental measurements, the assumptions of the paired t-test no longer hold (David and Gunnink [6]). For example, except when $\rho(X, Y) = 0$, the D_i have unequal variances and are not normally distributed. Nevertheless, simulation indicates that the t distribution with n-1 degrees of freedom continues to hold approximately for all ρ. If (X,Y) is bivariate normal, the reduction in the length of the confidence interval for $E(\overline{D})$, due to the pairing, can be shown to be given by a multiplicative factor $(1-c_n\rho^2)^{\frac{1}{2}}$, where c_n is a constant (< 1) tabulated by David and Gunnink [6]; e.g., for n = 10, $\rho = 0.7$, the reduction is by a factor of 0.733.

(d) Double Sampling and Ranked-Set Sampling

Suppose we wish to estimate μ_Y when Y is expensive to measure. If an inexpensive variable X, correlated with Y, is available, then taking n measurements on X, we can use their ordering to make k (< n) expensive measurements $Y_{[r_j : n]}$, j = 1, ..., k. The average $\overline{Y}_{[r : n]}$ is evidently an unbiased estimator of μ_Y for any symmetric distribution of Y if

$$r_{k+1-j} = n + 1 - r_j \qquad j = 1, \ldots, ½k \text{ (k even)},$$
$$j = 1, \ldots, ½(k+1) \text{ (k odd)}$$

Moreover, if (2) holds, we have

$$\overline{Y}_{[r : n]} = \mu_y + \rho\frac{\sigma_Y}{\sigma_X}(\overline{X}_{r : n} - \mu_X) + \overline{Z}, \tag{17}$$

where $\overline{X}_{r : n}$ is the mean of the k $X_{r_j : n}$ and $\overline{Z}$ of the k $Z_{[r_j]}$. Hence

$$\text{Var}\left[\frac{\overline{Y}_{[r : n]}}{\sigma_Y}\right] = \rho^2 \text{ var}\left[\frac{\overline{X}_{r : n}}{\sigma_X}\right] + \frac{1}{k}(1 - \rho^2) \tag{18}$$

Thus the ranks r_j minimizing var $(\overline{X}_{r : n})$ also minimize var $\overline{Y}_{[r : n]}$, whatever the value of ρ. The optimal choice of r_j for minimizing var $(\overline{X}_{r : n})$ is the integral part of $n\lambda_j + 1$, where $0 < \lambda_1 < \ldots < \lambda_k < 1$, for tabulated values of the λ_j (Mosteller [10]); roughly $\lambda_j = (j - ½)/k$. The first term on the right of (18) is small compared to the second, unless $|\rho|$ is close to 1, and vanishes as $n \to \infty$. Asymptotically, therefore, the variance of $\overline{Y}_{[r : n]}$ is $(1 - \rho^2)$ times the variance of k randomly chosen Y's.

It should be noted that the foregoing double sample procedure, due to O'Connell and David [12], requires only the ranks of the auxiliary X's. An interesting method with a similar aim is *ranked-set sampling*, introduced by McIntyre [9]. Here the sample size $n = k^2$ (or a multiple of k^2) and k subsamples, each of size k, are formed. All X-measurements are made (e.g., visual rankings of the height of k trees) but only one Y is measured (actual height of tree) per subsample, namely $Y_{[j:k]}$ in the j-th subsample ($j = 1, \ldots, k$). McIntyre's estimator of μ_Y is the mean of the $Y_{[j:k]}$. This is easily seen to be unbiased for *any* parent distribution, but is less efficient than $\overline{Y}_{[r:n]}$ above in the normal case (David [5], p. 184). Many extensions of ranked-set sampling are reviewed in a comprehensive paper by Patil *et al.* [13].

(e) Selection Through an Associated Variable

Yeo and David [18] consider the problem of choosing the best k objects out of n when, instead of measurements Y_i of primary interest, only associated measurements X_i $(i = 1, \ldots n)$ are available or feasible. For example, Y_i could represent future performance of an individual, with current score X_i, or Y_i might be an expensive measurement on the i-th object, perhaps destructive, and X_i an inexpensive measurement. It is assumed that the n pairs (X_i, Y_i) are a random sample from a continuous population. The actual values of the X_i are not required, only their ranks. A general expression is developed for the probability π that the s objects with the largest X-values include the k objects $(k \leq s)$ with the largest Y-values. When X and Y are bivariate normal with correlation coefficient ρ, a table of $\pi = {}_{n}\pi_{s:k}(\rho)$, for selected values of the parameters, gives the smallest s for which $\pi \geq P^*$ is preassigned.

Example : From 10 objects it is desired to select a subset of size s that will contain the k best objects (k = 1, 2, 3) with probability at least 0.9. We give a table of s for $\rho = 0.7, 0.8, 0.9$. Thus if we want to be at least 90% certain that the object with the highest Y-value is in the chosen subset for $\rho = 0.8$, we need to select the four objects with the highest X-value. The full table gives the actual inclusion probability $10^{\pi}\, 4:1\ (0.8)$ as 0.9183 and also shows that the object with the highest X-value has probability 0.5176 of having the highest Y-value. Another table tells us that for the object with the highest X-value to have probability ≥ 0.90 of having the highest Y-value, would require $\rho \geq 0.9931$! For a 50 : 50 chance $\rho = 0.783$.

ρ \ k	1	2	3
0.7	5	7	na
0.8	4	6	7
0.9	3	5	6

With the help of a computer program it is also possible to base the selection of the best object on the actual values of the X_i rather than on their ranks (Yeo and David [18]).

REFERENCES

[1] Andrews, D. M., 1995. Moments of the selection differential from exponential and uniform parents. In : *Statistical Theory and Applications* (H. N. Nagaraja, P. K. Sen and D. F. Morrison, eds.), 67-80, Springer - Verlag, New York.

[2] Barton, D. E. and Casley, D. J., 1958. A quick estimate of the regression coefficient, *Biometrika*, **45**, 431-435.

[3] Bhattacharya, P. K., 1974. Convergence of sample paths of normalized sums of induced order statistics. *Ann. Statist.*, **2**, 1034-1039.

[4] David, H. A., 1973. Concomitants of order statistics. *Bulletin of the International Statistical Institute*, **45**, 295-300.

[5] David, H. A., 1981. *Order Statistics*, 2nd ed. John Wiley and Sons, New York.

[6] David, H. A. and Gunnink, J. L., 1995. *The paired t-test under artificial pairing*. Technical Report, Iowa State University, USA.

[7] David, H. A. and Nagaraja, H. N., 1996. Concomitants of order statistics. In: *Handbook of Statistics* (N. Balakrishnan and C. R. Rao, eds.), **15** (in press).

[8] Harter, H. L., 1961. Expected values of normal order statistics. *Biometrika*, **48**, 161-165, Correction **48**, 476.

[9] McIntyre, G. A., 1952. A method of unbiased selective sampling using ranked sets. *Australian J. Agric. Research*, **3**, 385-390.

[10] Mosteller, F., 1946. On some useful "inefficient" statistics. *Ann. Math. Statist.*, **17**, 377-408.

[11] Nagaraja, H. N., 1982. Some asymptotic results for the induced selection differential. *J. Appl. Probab.*, **19**, 253-261.

[12] O'Connell, M. J. and David, H. A., 1976. Order statistics and their concomitants in some double sampling situations. In : *Essays in Probability and Statistics* (S. Ikeda *et al.*, eds.), 451-466, Shinko Tsusho, Tokyo.

[13] Patil, G. P., Sinha, A. K. and Taillie, C., 1994. Ranked set sampling. In : *Handbook of Statistics* (G. P. Patil and C. R. Rao, eds.), **12**, 167-200, Elsevier, Amsterdam.

[14] Tietjen, G. L., Kahaner, D. K. and Beckman, R. J., 1977. Variances and covariances of the normal order statistics for sample sizes 2 to 50. *Selected Tables in Mathematical Statistics*, **5**, 1-73.

[15] Tsukibayashi, S., 1962. Estimation of bivariate parameters based on range. *Reports of Statistical Applied Research*, JUSE, **9**, 10-23.

[16] Tsukibayashi, S., 1996. Private communication.

[17] Yang, S. S., 1977. General distribution theory of the concomitants of order statistics. *Ann. Statist.*, **5**, 996-1002.

[18] Yeo, W. B. and David, H. A., 1984. Selection through an associated characteristic with applications to the random effects model. *J. Amer. Statist. Assoc.*, **79**, 399-405.