〈Research Paper〉

Journal of the Korean
Statistical Society
Vol. 12, No. 2, 1983

Selection Conditional on Associated Measurements

Woon Bang Yeo*

ABSTRACT

In this paper, a random subset selection procedure for the choice of the k best objects out of n primary measurements Y_i is considered when only the associated measurements X_i are available. In contrast to Yeo and David (1982), where only the ranks of the $X's$ are needed, the present procedure uses the observed X-values. The approach is illustrated numerically when X and Y are bivariate normal and the standard deviation of X is known.

1. Introduction

We consider the selection of the k best objects out of n when, instead of measurements y_i $(i=1, \cdots, n)$ of primary interest, only associated measurements x_i are available. For definiteness we assume that high Y-values (i.e., large y_i) are desirable and that X and Y are positively associated. The best object then corresponds to the largest Y-value. As in Yeo and David (1982), henceforth referred to as *YD*, it is assumed that the n pairs (x_i, y_i) are realizations of (X_i, Y_i), which are n independent random couples with c.d.f. $F(x, y)$ and p.d.f. $f(x, y)$. We wish to choose the smallest subset which includes the k best objects, with a probability at least equal to a pre-assigned value $P^*(0<P^*<1)$. In *YD* this problem is treated by a procedure requiring only the ranks of the x_i. Here we show how the actual x_i-values can be used provided the standard deviation of the X_i may be taken as known. Madsen (1982) has recently considered an approach of this kind in a different context. See also Portnoy (1982).

For given n, $F(x, y)$, and P^*, the size s of the chosen subset will vary with the

*Korea Development Institute

observed x_i-values. There is an interesting connection with YD where s is fixed. Some illustrative numerical results are provided in the bivariate normal case.

2. Probability of Correct Selection

To construct the desired subset of size s containing the largest k objects with probability $\prod_{n\ s:k}(\underline{x}) \geqq P^*$ (specified), we require the conditional probability

$$\prod_{n\ s:k}(\underline{x}) = \Pr[\{Y_{[n]}, \cdots, Y_{[n-s+1]}\} \supset \{Y_{(n)}, \cdots, Y_{(n-k+1)}\}] \tag{1}$$

where $Y_{[j]}$ is defined to be $Y_j | X_j = x_{(j)}$, $Y_{(j)}$ is the j-th order statistic with $Y_{(1)} \leqq Y_{(2)} \leqq \cdots \leqq Y_{(n)}$, and $x_{(j)}$ is the realization of the j-th order statistic $X_{(j)}$. Y-variate associated with $X_{(j)}$ is called the concomitant of j-th order statistic (See e.g. David, 1981) but in our case $Y_{[j]}$ means the concomitant of $X_{(j)}$ conditional on $X_{(j)} = x_{(j)}$ and $Y_{[j]}$ $(j=1, \cdots, n)$ are independent with the distribution $F_{Y_{[j]}}(y) \equiv F_j(y_j)$. The event in (1) is explained as follows.

(i) $Y_{(n-k+1)}$ is one of $\{Y_{[n-s+1]}, Y_{[n-s+2]}, \cdots, Y_{[n]}\}$

(ii) $Y_{[1]}, \cdots, Y_{[n-s]} < Y_{(n-k+1)}$

(iii) Among $Y_{[n-s+1]}, \cdots, Y_{[n]}$, exactly $(k-1)$ $Y_{[i]}$ are greater than $Y_{(n-k+1)}$ and exactly $(s-k)$ $Y_{[i]}$ are less than $Y_{(n-k+1)}$.

Hence noting (i)—(iii) and conditioning on $Y_{(n-k+1)}$, the probability in (1) can be expressed by

$$\prod_{n\ s:k}(\underline{x}) = \int_{-\infty}^{\infty} \prod_{i=1}^{n-s} F_i(y) \sum \prod_{i=1}^{k-1} [1 - F_{j_i}(y)] \prod_{i=k}^{s-1} F_{j_i}(y) dF_{j_s}(y), \tag{2}$$

where the summation extends over all permutation $(j_1, j_2, \cdots, j_s)$ of $(n-s+1, n-s+2, \cdots, n-1, n)$ for which $j_1 < \cdots < j_{k-1}$ and $j_k < \cdots < j_{s-1}$. Note that

$$\prod_{n\ 1:1}(\underline{x}) = \int_{-\infty}^{\infty} \prod_{i=1}^{n-1} F_i(y)\ dF_n(y),$$

$$\prod_{n\ s:1}(\underline{x}) = \sum_{j=n-s+1}^{n} \int_{-\infty}^{\infty} \prod_{\substack{i=1 \\ i \neq j}}^{n} F_i(y) dF_j(y)\ \ (1 < s < n). \tag{3}$$

It will be convenient to call the present approach conditional and that in YD unconditional.

We consider now $\prod_{n\ s:1}(\underline{x})$ when (X, Y) is bivariate normal $BvN(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$, with $0 < \rho < 1$. Substituting in (3) gives

$$\prod_{n\ s:1}(\underline{x})=\sum_{j=n-s+1}^{n}\int_{-\infty}^{\infty}\prod_{\substack{i=1\\ i\neq j}}^{n}\Phi\left(z+\frac{\rho}{(1-\rho^2)^{1/2}}\ \frac{X_j-X_i}{\sigma_x}\right)\phi(z)dz, \tag{4}$$

where $\Phi(z)$ and $\phi(z)$ are the standard normal c.d.f. and p.d.f. Note that $\prod_{n\ s:1}(\underline{x})$ depends on σ_x and ρ, but not on the other parameters. In practice, the standard deviation of the auxiliary variate X can often be taken as known. From previous experience one is also likely to have an idea of the range of ρ.

Example 1. Arranged in increasing order of magnitude the first 10 random normal deviates given in Beyer (1968) are reproduced at the top of Table 1, the body of which gives values of $\prod_{10\ s:k}(\underline{x})$ for selected values of ρ. See the Appendix for the method of computation. The entries for $\rho=0$ or $s=10$ are obvious and agree with (4).

Suppose that $k=1$ and $\rho=0.7$. Then table 1 shows that for $P^*=0.9$ the required subset size is $s=6$, with actual conditional probability of correct selection 0.9386. When $k=2$ with the same values of ρ and P^*, the subset size must be $s=7$ with conditional probability 0.9388.

3. Relation to the Unconditional Approach

Averaging (2) over repeated samples $\underline{x}$ gives the unconditional probability

$$\prod_{n\ s:k}=\Pr\{s \text{ objects with the largest } X_i \text{ include the } k \text{ largest } Y_i\},$$

for which expressions are developed in Yeo (1982). With the help of these expressions *YD* have prepared tables in the bivariate normal case which immediately provide the desired subset in the unconditional case (i.e., when only the ranks of the x_i are used).

Example 1 (continued). For $n=10$, $\rho=0.7$ we read off the following values from Table 1 of *YD*.

Probability $\prod_{10\ s:k}$ that s objects out of 10 with largest x_i-values include the objects with the k largest y_i-values.

k/s	1	2	3	4	5	6	7	8	9	10
1	0.4261	0.6396	0.7706	0.8563	0.9134	0.9511	0.9752	0.9896	0.9971	1.0
2	—	0.2332	0.4470	0.6201	0.7532	0.8513	0.9199	0.9644	0.9896	1.0

Table 1. Values of $\prod_{10\,s:k}(x)$ in (2), for the $X_{(i)}$ shown when (X, Y) is $BvN(0,0,1,1,\rho)$

$X_{(i)}$		−1.501	−0.690	0.060	0.464	0.906	1.022	1.179	1.372	1.394	1.486
s	k/ρ	0.0		0.7		0.8		0.9		0.95	0.99
1	1	0.1000		0.2201		0.2503		0.3042		0.3648	0.5408
2	1	0.2000		0.4131		0.4607		0.5389		0.6172	0.7839
	2	0.0222		0.0960		0.1216		0.1721		0.2347	0.4163
3	1	0.3000		0.6000		0.6622		0.7590		0.8472	0.9800
	2	0.0667		0.2711		0.3370		0.4593		0.5979	0.9060
	3	0.0083		0.0726		0.1044		0.1778		0.2858	0.6918
4	1	0.4000		0.7398		0.7984		0.8792		0.9406	0.9986
	2	0.1333		0.4649		0.5517		0.6905		0.8172	0.9895
	3	0.0333		0.2297		0.3065		0.4522		0.6145	0.9462
	4	0.0048		0.0698		0.1080		0.1973		0.3224	0.7104
5	1	0.5000		0.8487		0.8931		0.9487		0.9801	0.9999
	2	0.2222		0.6553		0.7417		0.8548		0.9322	0.9989
	3	0.0833		0.4461		0.5523		0.7122		0.8401	0.9917
	4	0.0238		0.2470		0.3428		0.5134		0.6758	0.9354
	5	0.0040		0.0876		0.1414		0.2599		0.3996	0.6835
6	1	0.6000		0.9386		0.9691		0.9933		0.9994	1.0000
	2	0.3333		0.8393		0.9096		0.9753		0.9967	1.0000
	3	0.1667		0.7050		0.8162		0.9377		0.9885	1.0000
	4	0.0714		0.5389		0.6812		0.8657		0.9651	1.0000
	5	0.0238		0.3487		0.4966		0.7318		0.8998	0.9993
	6	0.0048		0.1533		0.2589		0.4804		0.7078	0.9836
7	1	0.7000		0.9791		0.9927		0.9994		1.0000	1.0000
	2	0.4667		0.9388		0.9755		0.9972		0.9999	1.0000
	3	0.2917		0.8753		0.9435		0.9912		0.9996	1.0000
	4	0.1667		0.7827		0.8888		0.9767		0.9983	1.0000
	5	0.0833		0.6530		0.7981		0.9423		0.9925	1.0000
	6	0.0333		0.4770		0.6473		0.8577		0.9642	1.0000
	7	0.0083		0.2490		0.3931		0.6206		0.7838	0.9775
8	1	0.8000		0.9968		0.9996		1.0000		1.0000	1.0000
	2	0.6222		0.9887		0.9981		1.0000		1.0000	1.0000
	3	0.4667		0.9729		0.9945		0.9999		1.0000	1.0000
	4	0.3333		0.9448		0.9864		0.9997		1.0000	1.0000
	5	0.2222		0.8969		0.9689		0.9986		1.0000	1.0000
	6	0.1333		0.8159		0.9302		0.9939		0.9999	1.0000
	7	0.0667		0.6765		0.8398		0.9687		0.9971	1.0000
	8	0.0222		0.4300		0.6100		0.8279		0.9425	0.9999
9	1	0.9000		0.9997		1.0000		1.0000		1.0000	1.0000
	2	0.8000		0.9988		0.9999		1.0000		1.0000	1.0000
	3	0.7000		0.9967		0.9998		1.0000		1.0000	1.0000
	4	0.6000		0.9920		0.9994		1.0000		1.0000	1.0000
	5	0.5000		0.9826		0.9980		1.0000		1.0000	1.0000
	6	0.4000		0.9635		0.9941		1.0000		1.0000	1.0000
	7	0.3000		0.9235		0.9811		0.9995		1.0000	1.0000
	8	0.2000		0.8343		0.9332		0.9924		0.9998	1.0000
	9	0.1000		0.6113		0.7347		0.8761		0.9593	1.0000
10	k	1.0000		1.0000		1.0000		1.0000		1.0000	1.0000

For $k=1$ in Example 1, use of the actual values of the x_i led to $s=6$. If in repeated samples we were always to use $s=6$, then the resulting average probability of correct selection is $\prod_{10\ 6:1}=0.9511$. Repeated use of $s=5$ gives $\prod_{10\ 5:1}=0.9134>P^*$. In other words, as inspection confirms intuitively, the x_i have happened to come out rather unfavorably for our aim of choosing a small subset. But when $k=2$, the required subset size is $s=7$ in both cases to satisfy the pre-assigned probability $P^*=0.9$.

APPENDIX

The entries of Table 1 were calculated from (2) (with $\sigma_x=1$) by use of the 64-point Gauss-Hermite quadrature of

$$\int_{-\infty}^{\infty} e^{-t^2} G(t)dt,$$

where in the present case

$$G(t)=\frac{1}{\pi^{1/2}}\prod_{i=1}^{n-s}\Phi[\sqrt{2}t+A(x_{j_s}-x_i)]$$

$$\sum\prod_{i=1}^{k-1}[1-\Phi\{\sqrt{2}t+A(x_{j_s}-x_{j_i})\}]\prod_{i=k}^{s-1}\Phi[\sqrt{2}t+A(x_{j_s}-x_{j_i})],$$

$$A=\rho/(1-\rho^2)^{1/2}.$$

The acuracy of the numerical evaluation is up to four decimal places. As a check we can see the values corresponding to $\rho=0.0$ or $s=n$.

REFERENCES

(1) Beyer, W.H. (Ed.) (1968). *Handbook of Probability and Statistics*, 2nd ed., Cleveland: The Chemical Rubber Company.

(2) David, H.A. (1981). *Order Statistics*. 2nd ed., John Wiley & Sons, New York, N.Y., 360.

(3) Madsen, R.W. (1982). A Selection Procedure Using a Screening Variate, *Technometrics*, Vol. 24, 301—306.

(4) Portnoy, S. (1982). Maximizing the Probability of Correctly Ordering Random Variables Using Predictors, *Journal of Multivariate Analysis*, Vol. 12, 256—269.

(5) Yeo, W.B. (1982). Selection Through an Associated Characteristic, Unpublished Ph.D. thesis, Iowa State University.

(6) Yeo, W.B., and H.A. David (1982). Selection through an Associated Characteristic, with Applications to the Random Effects Model, submitted for publication to the *Journal of the American Statistical Association.*