Approximations of variances and covariances for order statistics from the standard extreme value distribution

Mir Nabi Pirouzi Fard

and

Björn Holmquist

2006:1

DEPARTMENT OF STATISTICS

S-220 07 LUND
SWEDEN

APPROXIMATIONS OF VARIANCES AND COVARIANCES FOR ORDER STATISTICS FROM THE STANDARD EXTREME VALUE DISTRIBUTION

Mir Nabi Pirouzi Fard and Björn Holmquist
Department of Statistics
Lund University
Box 743
S-220 07 Lund, Sweden
bjorn.holmquist@stat.lu.se

Key Words: order statistics; extreme value distribution; variances and covariances for order statistics.

ABSTRACT

We consider simple approximations of variances and covariances for order statistics from the standard extreme value distribution. Exact values and simulation results of the variances and covariances for certain sample sizes are used to determine the validity of the suggested approximations.

1. INTRODUCTION

Let $X_i,\ i = 1, \ldots, n$ be a random sample of size n from the standard extreme value distribution with probability density function (pdf)

$$f(x) = \exp(x - e^x), \qquad -\infty < x < \infty$$

and cumulative distribution function (cdf)

$$F(x) = 1 - \exp(-e^x), \qquad -\infty < x < \infty$$

The mean and variance are $-\gamma$ and $\pi^2/6$ respectively where $\gamma \approx 0.57721$ is Euler's constant.

Let the order statistics $X_{1:n} < X_{2:n} < \ldots < X_{n:n}$ be the sample ordered from the smallest to the largest. The order statistics has been discussed by many authors such as Blom (1958), Sarhan and Greenberg (1962), David (1970) and David and Nagaraja (2003).

The pdf of $X_{i:n}$, $1 \le i \le n$, is given by

$$f_i(x) = n\binom{n-1}{i-1} F(x)^{i-1}(1-F(x))^{n-i} f(x), \quad -\infty < x < \infty$$

and the joint density of $X_{i:n}$ and $X_{j:n}$, $1 \le i < j \le n$

$$f_{i,j}(x,y) = n(j-i)\binom{n-1}{j-1}\binom{j-1}{i-1} F(x)^{i-1}(F(y)-F(x))^{j-i-1}(1-F(y))^{n-j} f(x)f(y)$$

for $-\infty < x \le y < \infty$. These densities can be used to derive expressions for variances and covariances for the order statistics which gives

$$\mathrm{E}(X_{(i)}) = -n\binom{n-1}{i-1}\sum_{k=0}^{i-1}\binom{i-1}{k}(-1)^k \frac{\gamma + \ln(n-i+k+1)}{n-i+k+1}.$$

and

$$\mathrm{E}(X_{(i)}^2) = n\binom{n-1}{i-1}\sum_{k=0}^{i-1}\binom{i-1}{k}(-1)^k \frac{\pi^2/6 + (\gamma + \ln(n-i+k+1))^2}{n-i+k+1}.$$

$$\begin{aligned}\mathrm{E}(X_{(i)}X_{(j)}) \;=\; & \binom{n}{j}\binom{j}{i-1}\sum_{r=0}^{j-i-1}\sum_{s=0}^{n-j}(-1)^{r+s}\binom{j-i-1}{r}\binom{n-i}{s} \\ & \xi(i+r, j-i-r+s)\end{aligned}$$

where the function ξ is the double integral

$$\xi(t,u) = \int_{-\infty}^{\infty}\int_{-\infty}^{y} xy\, e^{(x-te^x)}\, e^{(y-ue^y)}\, dx\, dy \qquad\qquad t,u > 0$$

An explicit expression for ξ has been given by Leiblein (1953). The calculation of variances and covariances using these expressions causes problems due to the binomial coefficients and rounding errors in the evaluation even for moderate sample sizes n. For this reason tabulated values of the variances of the order statistics (White 1967, 1969) and of variances and covariances (Leiblein and Zelen 1956; Balakrishnan and Chan 1992a, 1992b) have been presented for selected sample sizes.

2. APPROXIMATIONS OF VARIANCES AND COVARIANCES

The variance of the ith order statistic can be approximated by using Taylor's series expansion.

By using a Taylor's series expansion the variance of a function $h(U)$ of a random variable U may be approximated by

$$\mathrm{V}(h(U)) \approx (h'(\mathrm{E}(U)))^2\mathrm{V}(U)$$

Hence, since $X_{i:n} \stackrel{d}{=} F^{-1}(U_{i:n})$ where $U_{i:n}$ is the ith order statistic from a sample of size n from a uniform distribution on the interval $(0,1)$, the variance of $F^{-1}(U_{i:n})$ can be approximated by

$$\mathrm{V}(F^{-1}(U_{i:n})) \approx (\frac{1}{f(F^{-1}(\mathrm{E}(U_{i:n})))})^2\mathrm{V}(U_{i:n})$$

since $\mathrm{E}(U_{i:n}) = i/(n+1)$ and $\mathrm{V}(U_{i:n}) = i(n+1-i)/((n+1)^2(n+2))$ and further $F^{-1}(u) = \ln(-\ln(1-u))$ so that $f(F^{-1}(u)) = -(1-u)\ln(1-u)$ we have that

$$\mathrm{V}(F^{-1}(U_{i:n})) \approx \frac{i}{(\ln(\frac{n+1-i}{n+1}))^2(n+1-i)(n+2)} \qquad (1)$$

The covariance between functions $h(U_i)$ and $h(U_j)$ of random variables U_i and U_j can by using Taylor's series expansion be approximated by

$$\mathrm{C}(h(U_i), h(U_j)) \approx h'(\mathrm{E}(U_i))h'(\mathrm{E}(U_j))\mathrm{C}(U_i, U_j)$$

By noting that $X_{i:n} \stackrel{d}{=} F^{-1}(U_{i:n})$ and $X_{j:n} \stackrel{d}{=} F^{-1}(U_{j:n})$ where $U_{i:n}$ and $U_{j:n}$ are the ith and jth order statistics from a sample of size n from a uniform distribution on $(0,1)$, we then have the following approximation

$$\mathrm{C}(F^{-1}(U_{i:n}), F^{-1}(U_{j:n})) \approx \frac{1}{f(F^{-1}(\mathrm{E}(U_{i:n})))f(F^{-1}(\mathrm{E}(U_{j:n})))}\mathrm{C}(U_{i:n}, U_{j:n})$$

Since $\mathrm{E}(U_{i:n}) = i/(n+1)$, $\mathrm{E}(U_{j:n}) = j/(n+1)$ and $\mathrm{C}(U_{i:n}, U_{j:n}) = i(n+1-j)/((n+1)^2(n+2))$ we thus have

$$\mathrm{C}(F^{-1}(U_{i:n}), F^{-1}(U_{j:n})) \approx \frac{i}{\ln(\frac{n+1-i}{n+1})\ln(\frac{n+1-j}{n+1})(n+1-i)(n+2)} \qquad (2)$$

The approximation for the variance in equation (1) is a special case of equation (2) when $i = j$.

By introducing correction terms, we will study the following approximation of the covariances and variances:

$$\mathrm{C}(X_{i:n}, X_{j:n}) \approx \frac{i+\varphi_1}{\ln(\frac{n+1-i-\varphi_3}{n+1-\varphi_4})\ln(\frac{n+1-j-\varphi_5}{n+1-\varphi_4})(n+1-i-\varphi_3)(n+2-\varphi_2)} \qquad (3)$$

$(1 \le i \le j \le n)$

A table of variances and covariances for the order statistics were given by Balakrishnan and Chan (1992a) for $n = 2(1)15(5)30$. Values from this table have been used to obtain parameters $\varphi = (\varphi_1, \varphi_2, \varphi_3, \varphi_4, \varphi_5)$ that minimizes

$$Q_1(\varphi) = \sum_{i \le j}(C_{ij} - g_{ij}(\varphi))^2$$

where g_{ij} is given by the RHS of (3) and C_{ij} are the values from Balakrishnan and Chan's table. The variance of the first order statistic is $\pi^2/6$ independent of sample size, so for $i = j = 1$, we define $C_{11} = g_{11} = \pi^2/6$.

Table 1 presents the optimum values of φ for a range of different sample sizes n. For $n > 30$ the covariances C_{ij} have been replaced by the empirical covariances between order statistics based on 1500000 Monte Carlo simulations.

The obtained values of φ indicate that the dependence of the optimal values of φ_k on the sample size is fairly weak. We use the average value of the optimal values from sample sizes 10(5)30 given i Table 2 as a suggested approximation for the covariances and variances for the order statistics.

The suggested approximation for the covariance between $X_{i:n}$ and $X_{j:n}$ is thus

$$K_{ij} = \begin{cases} \pi^2/6 & \text{if } i = j = 1 \\ \dfrac{i-0.469}{\ln(\frac{n+0.831-i}{n+0.356})\ln(\frac{n+0.779-j}{n+0.356})(n+0.831-i)(n+0.073)} & \text{otherwise} \end{cases}$$

for $1 \le i \le j \le n$.

3. EVALUATION OF THE APPROXIMATION

We define the deviation between our proposed model, K_{ij}, and, either the exact values (for $n \le 30$) or the results based on simulations (for $n > 30$), C_{ij}, as

$$D_{ij} = K_{ij} - C_{ij}$$

n	φ_1	φ_2	φ_3	φ_4	φ_5
10	−0.452	1.815	0.127	0.589	0.180
15	−0.464	1.885	0.157	0.627	0.208
20	−0.471	1.934	0.175	0.651	0.226
25	−0.476	1.974	0.189	0.668	0.239
30	−0.483	2.029	0.199	0.685	0.254
75	−0.465	2.066	0.235	0.697	0.265
100	−0.457	2.008	0.241	0.696	0.262
150	−0.448	1.952	0.249	0.697	0.259
200	−0.440	2.214	0.245	0.687	0.249

Table 1: Optimization results for $\varphi_k, k = 1, 2 \ldots, 5$.

$\hat{\varphi}_1$	$\hat{\varphi}_2$	$\hat{\varphi}_3$	$\hat{\varphi}_4$	$\hat{\varphi}_5$
−0.469	1.927	0.169	0.644	0.221

Table 2: The average value of $\varphi_k, k = 1, 2 \ldots, 5$.

for $1 \leq i \leq j \leq n$. These differences are shown in Figure 1 for some selected sample sizes and shows that the errors in the approximations are less than two decimal places.

For a given sample size the maximum absolute deviations

$$Z_1 = \max_{i \leq j} |D_{ij}|$$

as well as the percentual absolute deviation

$$\Delta_{ij} = \frac{100|D_{ij}|}{C_{ij}}$$

were calculated together with the mean and maximum

$$T = \sum_{i \leq j} \Delta_{ij}/(n(n+1)/2), \qquad Z_2 = \max_{1 \leq i \leq j \leq n} \Delta_{ij}$$

These measures are presented in Table 3 for some selected sample sizes.

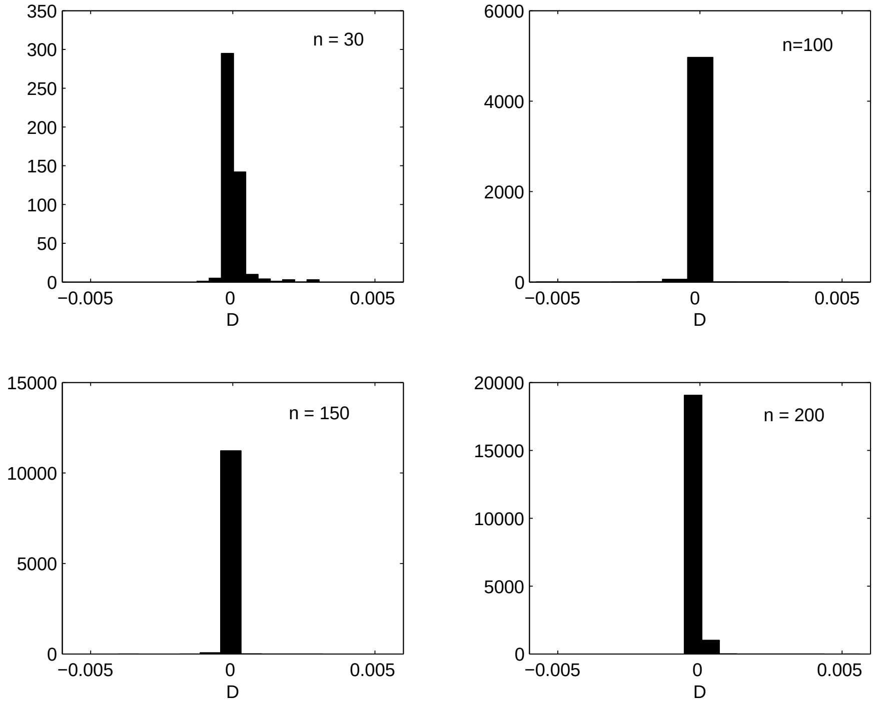

Figure 1: Differences D_{ij} between K_{ij} and C_{ij} for $1 \leq i \leq j \leq n$.

n	10	20	30	75	100	150	200
$Z_1 \times 10^2$	0.8	0.2	0.3	0.6	0.6	0.4	0.6
T	0.6	0.3	0.2	0.7	0.6	0.4	0.7
Z_2	1.9	1.4	2.2	12.8	14.0	24.0	37.9

Table 3: Maximum absolute deviation and maximum and mean percentual absolute deviation between K_{ij} and C_{ij}.

The maximum absolute deviation Z_1 is less than two decimal places for all sample sizes. The mean percentual absolute deviation is less than 1 % for all sample sizes. However the maximum percentual absolute deviation may be as high as 38 % for large sample sizes. The percentual large deviations occur for covariances close to zero, where the absolute deviation between the approximation and the true covariance still is small.

4. APPROXIMATIONS OF VARIANCES

The approximation studied in previous sections were applied to covariances as well as variances for the order statistic. Occasionally only variances are of interest. Tables of variances for order statistics from the extreme value distribution were presented by White (1967) for sample sizes 1(1)50(5)100.

Based on the appproximation in Eq. (1) we use White's table of variances of order statistics, $W_i, i = 1, 2, \ldots, n$ to obtain parameters $\psi = (\psi_1, \psi_2, \psi_3)$ that minimizes

$$Q_2(\psi) = \sum_{i=1}^{n} (W_i - g_i(\psi))^2$$

where

$$g_i(\psi) = \frac{i + \psi_1}{(\ln(\frac{n+1-i-\psi_3}{n+1-\psi_2}))^2 (n + 1 - i - \psi_3)(n + 2 - \psi_2)}$$

We note that in Q_2, $W_1 = g_1 = \pi^2/6$, independent of n. The principle for introducing correction terms has in this case been than increasing factors in i has one term, decreasing factors in i has another term and factors independent of i has a third term.

The optimum values of ψ are presented in Table 4 for selected sample sizes n. For sample sizes for which variances are not tabulated, the true variances were replaced by empirical variances of order statistics based om 1500000 Monte Carlo simulations for $n = 200$.

n	10	15	20	35	50	70	85	100	200
ψ_1	1.643	1.427	1.245	0.906	0.730	0.598	0.536	0.491	0.325
ψ_2	-0.112	-0.043	0.016	0.130	0.193	0.244	0.268	0.287	0.352
ψ_3	0.169	0.197	0.210	0.225	0.232	0.237	0.240	0.243	0.252

Table 4: Optimization results for ψ_1, ψ_2, ψ_3.

As seen from Table 4, the optimum parameters are quite dependent on sample size n, and as such the approximation is not useful if not the optimum values for a given sample size are known, or can be obtained by regression on the sample size by using the optimum values from Table 4.

The main reason for this strong dependence of sample size n, is that all factors independent of i (i.e. in the denominator of the logarithm as well as in the denominator) have been assigned the same correction term.

5. CONCLUSION

While the explicite expressions for variances and covariances for order statistics from the standard extreme value distribution are complicated and troublesome to handle, the approximation presented for variances and covariances for order statistics of arbitrary sample size is easy to use and gives good approximate values.

BIBLIOGRAPHY

Blom, G. (1958). *Statistical Estimates and Transformed Beta-Variables*, Almqvist and Wiksell, Uppsala.

Balakrishnan, N. and Chan, P.S. (1992a). Order statistics from extreme value distribution I: Tables of means, variances and covariances. *Communications in Statistics, Simulation and Computation*, **21**, 1199–1217.

Balakrishnan, N. and Chan, P.S. (1992b). Extended tables of means, variances, and covariances of order statistics from the extreme value distribution for sample sizes up to 30, Report, Department of Mathematics and Statistics, McMaster University, Hamilton, Canada.

David, H.A. (1970). *Order Statistics*, John Wiley, New York.

David, H.A. and Nagaraja, H.N. (2003). *Order Statistics*, 3rd Ed., John Wiley, New York.

Lieblein, J. (1953). On the exact evaluation of the variances and covariances of order statistics in samples from the extreme-value distribution, *Ann. Math. Statist.*, **24**, 282–287.

Lieblein, J. and Zelen, M. (1956). Statistical investigation of the fatigue life of deep groove ball bearing. *J. Res. Nat. Bur. Stand.*, **57**, 273–316.

Sarhan, A.E. and Greenberg, B. G. (eds.) (1962). *Contributions to Order Statistics*, John

Wiley, New York.

White, J.S. (1967). The moments of log-Weibull order statistics, *Research Publication GMR-717, Research Laboratories, General Motors Corp.*, Warren, Mich.

White, J.S. (1969). Moments of Log-Weibull Order Statistics, *Technometrics*, **11**, 373–386.