First moment approximations for order statistics from the extreme value distribution

Mir Nabi Pirouzi Fard

and

Björn Holmquist

2005:1

LUND
UNIVERSITY

DEPARTMENT OF STATISTICS

S-220 07 LUND
SWEDEN

APPROXIMATIONS OF THE FIRST MOMENT OF ORDER STATISTICS FROM THE STANDARD EXTREME VALUE DISTRIBUTION

Mir Nabi Pirouzi Fard and Björn Holmquist

Department of Statistics
Lund University
Box 743, SE-220 07 Lund, Sweden

ABSTRACT. Approximate expressions of the first moment of the order statistics of standard extreme value distributions are proposed. We compare different previously given approximations with the exact values. The results show that the here given approximation fits the exact values better than previously given models.

Keywords: Order statistics; extreme value distribution; expected value; approximation.

October 11, 2005

1. INTRODUCTION

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the standard extreme value distribution of type I with probability density function

$$f(x) = \exp(x - e^x), \quad -\infty < x < \infty \tag{1}$$

and cumulative distribution function (cdf)

$$F(x) = 1 - \exp(-e^x), \qquad -\infty < x < \infty \tag{2}$$

The extreme value distribution of type I sometimes refers to as the Gumbel distribution.

The mean and variance of a r.v. X having the cdf given in eq. (2) are

$$\mathrm{E}(X) = -\gamma \quad \text{and} \quad \mathrm{V}(X) = \pi^2/6$$

where $\gamma \approx 0.5772156649\ldots$ is Euler's constant.

Let $X_{1:n} \le X_{2:n} \le \cdots \le X_{n:n}$, be the order statistics obtained by arranging the n variables $X_i, i = 1, \ldots, n$ in ascending order. The order statistics are discussed by Blom (1958), Sarhan and Greenberg (1962), Gringorten (1963), David (1970), Filliben (1975), Hosking (1990), Yu and Huang (2001) and David and Nagaraja (2003) among others.

The extreme value distribution probability plot, like a number of other tests of fit for the extreme value distribution such as those discussed by Mann et al. (1973) and Tiku and Singh (1981), require the expected values of the order statistics from the standard extreme value distribution. Explicite expressions for the order statistics means have been given by Lieblein (1953), although these are not particularly useful, since the accuracy in the numerical computation is severely reduced for large sample sizes.

The expected value of the ith order statistic of a random sample from the cdf (2) is

$$\mathrm{E}(X_{i:n}) = \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n+1-i)} \int_{-\infty}^{\infty} xF(x)^{i-1}(1-F(x))^{n-i} f(x)\, dx$$

which by binomial expansion can be written

$$\mathrm{E}(X_{i:n}) = \frac{n!}{(n-i)!(i-1)!} \sum_{k=0}^{i-1} \binom{i-1}{k} (-1)^k \int_{-\infty}^{\infty} x \exp(-(n-i+k+1)e^x + x)\, dx$$

i.e

$$\mathrm{E}(X_{i:n}) = \frac{n!}{(n-i)!(i-1)!} \sum_{k=0}^{i-1} \binom{i-1}{k} (-1)^k h(n-i+k+1)$$

where

$$h(c) = \int_{-\infty}^{\infty} x \exp(-ce^x + x)\, dx = \int_0^{\infty} \ln u \exp(-cu)\, du$$

From the derivative of the gamma function $\Gamma(z) = \int_0^{\infty} t^{z-1}e^{-t}\, dt$ we get

$$h(c) = (\Gamma'(1) - \ln c)/c = (-\gamma - \ln c)/c$$

so that

$$\mathrm{E}(X_{i:n}) = -n\binom{n-1}{i-1} \sum_{k=0}^{i-1} \binom{i-1}{k} (-1)^k \frac{\gamma + \ln(n-i+k+1)}{n-i+k+1}. \tag{3}$$

For example $\mathrm{E}(X_{1:n})$ is given by

$$\mathrm{E}(X_{1:n}) = nh(n) = -\gamma - \ln n \tag{4}$$

and

$$\begin{aligned} \mathrm{E}(X_{2:n}) &= n(n-1)(h(n-1) - h(n)) \\ &= -\gamma - n(n-1)\left(\frac{\ln(n-1)}{n-1} - \frac{\ln n}{n}\right) \end{aligned}$$

while for general i,

$$\mathrm{E}(X_{i:n}) = -\gamma + n\binom{n-1}{i-1} (-1)^i \sum_{\ell=0}^{i-1} (-1)^{\ell} \binom{i-1}{\ell} \frac{\ln(n-\ell)}{n-\ell}. \tag{5}$$

The explicit expressions for the means of the order statistics given in (3) and (5) encounter numerical problems due to rounding errors in the evaluation for higher orders even for moderate n.

White (1967, 1969), by using a set of 100 decimal place logarithms of integers, presents a table for the means of the order statistics of the standard extreme value distribution for $n \leq 100$. Balakrishnan and Chan (1992a) presented tables of means (as well as variances and covariances) of all order statistics for sample sizes $n = 1(1)15(5)30$. Tables for all sample sizes up to 30 have also been presented by Balakrishnan and Chan (1992b).

In this paper we propose approximations for the expected values of the order statistics from the standard extreme value distribution. In section 2 we present an approximation for the first moment of the order statistics of the cdf in (2). A corresponding approximation for the mean of the order statistics of the standard extreme value distribution for maxima is discussed in section 3.

2. Approximations of $\mathrm{E}(X_{i:n})$

2.1. **An approximation based on the inverse cdf.** A well-known approximation for $\mathrm{E}(X_{i:n})$ for sufficiently large n is provided by

$$\mathrm{E}(X_{i:n}) \approx F^{-1}(\frac{i}{n+1})$$

where F^{-1} is the inverse of the cumulative distribution function of X, i is the rank of the ordered sample and n is the size of the sample.

Blom (1958) suggested α, β corrections and writes

$$\mathrm{E}(X_{i:n}) \approx F^{-1}(\frac{i-\alpha}{n-\alpha-\beta+1}) \tag{6}$$

where $\alpha, \beta \leq 1$.

To select values of the parameters α and β we use the method of least squares to minimize the squared difference between the expected values of the order statistics and the approximation given by (6). The expression for the mean of the first order statistic in (4) is fairly simple and will be adopted. We thus minimize

$$Q(\alpha, \beta) = \sum_{i=2}^{n} (W_i - g_i(\alpha, \beta))^2 \tag{7}$$

where W_i represents the expected values, also given in White's table, and

$$g_i(\alpha, \beta) = F^{-1}(\frac{i-\alpha}{n-\alpha-\beta+1}) = \ln(-\ln(1 - \frac{i-\alpha}{n-\alpha-\beta+1})) \tag{8}$$

For the minimization of Q in (7) we used the numerical algorithm fminsearch in MATLAB to obtain solutions for a set of different sample sizes n. The optimal values of α and β for some specific sample sizes are given in Table 1. As seen from

n	5	10	20	35	50	70	85	100
α	0.4854	0.4884	0.4886	0.4878	0.4872	0.4865	0.4862	0.4859
β	0.2854	0.3001	0.3140	0.3252	0.3322	0.3387	0.3424	0.3454

TABLE 1. The results of optimization.

this table, the optimal values depend on the sample size, but the dependence is less pronounced for larger sample sizes.

Using the averages of the parameter values in Table 1, giving $\alpha = 0.4870$ and $\beta = 0.3229$, we present the following approximation of the uniform order statistics

$$m_{i1} = \begin{cases} 1 - \exp(-e^{-\gamma}/n) & i = 1 \\ (i - 0.4870)/(n + 0.1901) & i = 2, 3, \ldots, n \end{cases} \tag{9}$$

from which we have the approximations of the means of the order statistics from the cdf in (2):

$$\mathrm{E}(X_{i:n}) \approx F^{-1}(m_{i1})$$

2.2. **Other approximations.** We use some other previously proposed models to compare with our model. Filliben (1975) proposed uniform order statistics medians as

$$m_{i2} = \begin{cases} 1 - 0.5^{1/n} & i = 1 \\ (i - 0.3175)/(n + 0.365) & i = 2, 3, \ldots, n - 1 \\ 0.5^{1/n} & i = n \end{cases}$$

Filliben's model was used to approximate the median of the ith order statistic from the normal distribution for use in normal probability plots. Filliben (1975) and Vogel (1986) used this model to develop the probability plot correlation coefficient (PPCC) test for normality. Evans et al. (1989) however also used Filliben's model to develop a goodness-of-fit test for the Weibull distribution.

Gringorten (1963) proposed the approximation

$$m_{i3*} = \frac{i - 0.44}{n + 0.12}$$

for the uniform order statistics. This model was applied to the means of the order statistics from the cdf

$$G(x) = \exp(-e^{-x}), \qquad -\infty < x < \infty \tag{10}$$

with inverse

$$G^{-1}(u) = -\ln(-\ln u).$$

If a r.v. Y have the cdf in (10) then the distribution of $-X$ and Y are the same. The ith order statistic of X has therefore the same distribution as the ith order statistic of $-Y$ which is the $(n - i + 1)$th order statistic of Y with reversed sign.

Thus if $m_{i3*} = (i - \alpha_3)/(n + \psi_3)$, with $\alpha_3 = 0.44$ and $\psi_3 = 0.12$ is the uniform order statistics of the cdf in (10) then the uniform order statistic of the cdf in (2) is given by

$$m_{i3} = \frac{n - i + 1 - \alpha_3}{n + \psi_3} = 1 - \frac{i + \psi_3 - 1 + \alpha_3}{n + \psi_3} = 1 - \frac{i - \alpha_{3*}}{n + \psi_{3*}} \tag{11}$$

where $\alpha_{3*} = -(\psi_3 - 1 + \alpha_3)$ and $\psi_{3*} = \psi_3$.

Hence the parameters in m_{i3} are

$$\alpha_{3*} = -(0.12 - 1 + 0.44) = 0.44 \quad \text{and} \quad \psi_{3*} = 0.12$$

and Gringorten's model have thus the same parameter values also for the uniform order statistic of the cdf in (2) as for the cdf in (10).

Blom (1958) introduced

$$m_{i4} = \frac{i - 0.375}{n + 0.25}$$

to approximate means of order statistics from the normal distribution. Although this is not linked to the extreme value distribution we will also use this model in our comparison. To approximate the means of order statistics in the extreme value distribution Blom (1958) suggested

$$m_{i5*} = \frac{i - 0.25}{n + 0.25}$$

for the cdf in (10) which, using the transformation (11), corresponds to

$$m_{i5} = \frac{i - 0.50}{n + 0.25}$$

for the cdf in (2).

2.3. **Comparisons.** We compare the exact means from White's table with those obtained using the inverse cdf-transformation on Filliben's model, Blom's two models, simulation results and our proposed model in (9). The differences between White's table values (W_i) and the other models are expressed by

$$\delta_{ij} = W_i - p_{ij}, \quad i = 1, \ldots, n, j = 1, \ldots, 6$$

where $p_{ij} = \ln(-\ln(1 - m_{ij}))$ for $j = 1, \ldots, 5$ and where p_{i6} represents the mean of 15 000 replications of the order statistics from the cdf (2).

Figure 1 shows the differences for the different models for a selected number of sample sizes. It shows that the deviations for our proposed model are reasonbly small.

The plot also verify that the magnitude of the differences between the exact values (W_i) and the simulation results (p_{i6}) are of the size to be expected according to the size of the Monte-Carlo simulation.

According to the results found here the differences based on p_{i2} and p_{i4} have large variability for lower order statistics. This is also the case for differences based on

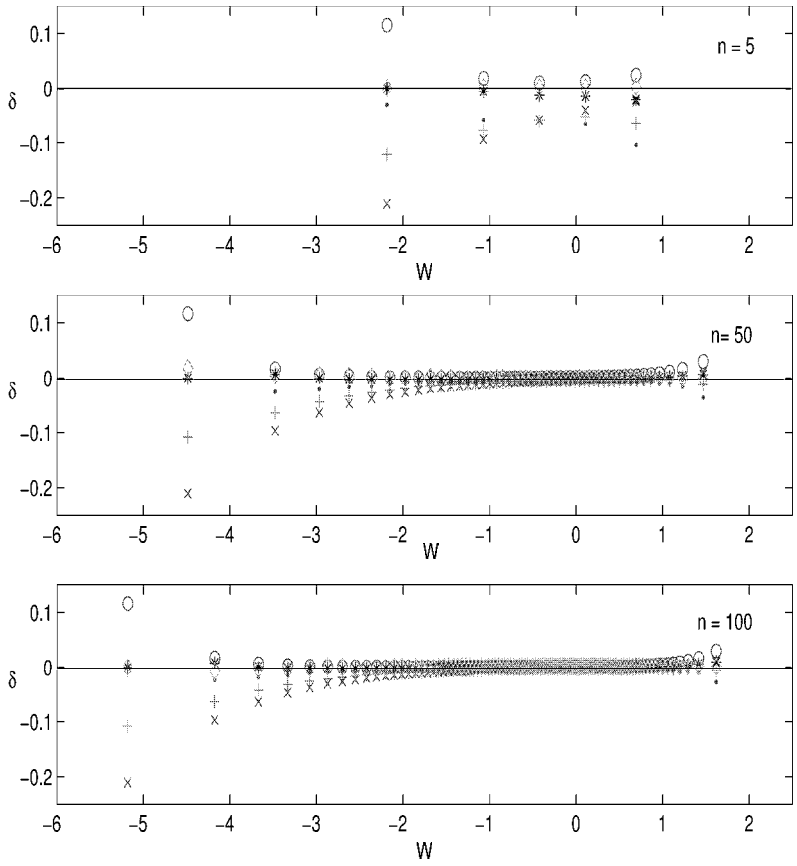

FIGURE 1. Differences between White's tables values W_i and p_{ij}'s, $j = 1, 2, \ldots, 6$. Here $*$ describe $\delta_{i1} = W_i - p_{i1}$; $\times$ symbolize $\delta_{i2} = W_i - p_{i2}$; $\bullet$ symbolize $\delta_{i3} = W_i - p_{i3}$; $+$ symbolize $\delta_{i4} = W_i - p_{i4}$; $\circ$ symbolize $\delta_{i5} = W_i - p_{i5}$; $\diamond$ symbolize $\delta_{i6} = W_i - p_{i6}$.

p_{i5}. Also by study of δ_{ij} for $j = 1, \ldots, 6$, we observed that while the differences between our proposed model and the exact values are up to one decimal place for $n \leq 8$ and less than two decimal places for larger sample sizes ($n > 8$), differences in the other models continue to stay in one decimal place.

For $n > 100$, for which we have no numerical values of the mean order statistics, we apply simulation results to make the comparisons between p_{ij} and p_{i6} for $j = 1, 2, \ldots, 5$ with differences defined by

$$\phi_{ij} = p_{i6} - p_{ij}, \quad i = 1, \ldots, n, j = 1, \ldots, 5$$

The plot of differences in Figure 2 indicates that our model has less variability over different order and fits better than the other models. It shows that Gringorten's model is not suitable for small sample sizes but gives better results for larger sample

size. The results also show that Blom's and Filliben's model have good precision over the right tail only.

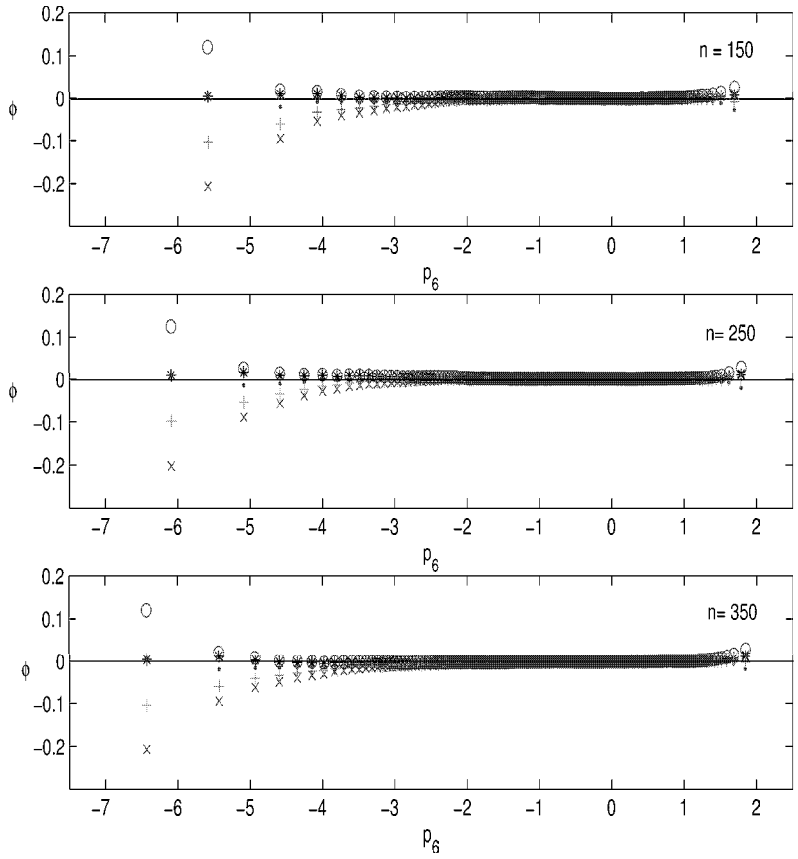

FIGURE 2. Differences between the simulation results p_{i6} and p_{ij}'s, $j = 1, \ldots, 5$. Here $*$ symbolize $\phi_{i1} = p_{i6} - p_{i1}$; $\times$ symbolize $\phi_{i2} = p_{i6} - p_{i2}$; $\bullet$ symbolize $\phi_{i3} = p_{i6} - p_{i3}$; $+$ symbolize $\phi_{i4} = p_{i6} - p_{i4}$; $\circ$ symbolize $\phi_{i5} = p_{i6} - p_{i5}$.

3. Approximations of the mean of order statistics of the cdf in (10)

As discussed in section 2, the parameter values in Gringorten's model of the uniform order statistics in cdf (10) are the same as in the cdf (2). By the same approach we transform our proposed model for the means of the order statistics of the cdf in (2) to obtain the uniform order statistics of the cdf in (10) as

$$m_{i1*} = \begin{cases} (i - 0.3229)/(n + 0.1901) & i = 1, 2, \ldots, n-1 \\ \exp(-e^{-\gamma}/n) & i = n \end{cases}$$

We compare this model, Blom's model and Gringorten's model with the exact values. It is obvious that White's tables can be used in reversed order and sign to give the means of the order statistics of the cdf in (10).

We define $q_{ij}, i = 1, 2, \ldots, n$ and $j = 1, 3, 5, 6$ as

$$q_{i1} = -\ln(-\ln m_{i1*}), \quad q_{i3} = -\ln(-\ln m_{i3*}), \quad q_{i5} = -\ln(-\ln m_{i5*})$$

and q_{i6} represent then mean of 15 000 replications of the ith order statistic from the cdf in (10).

We also define differences between the exact values (W_i^*) and q_{ij} as

$$\varepsilon_{ij} = W_i^* - q_{ij}$$

for $i = 1, \ldots, n$ and $j = 1, 3, 5, 6$.

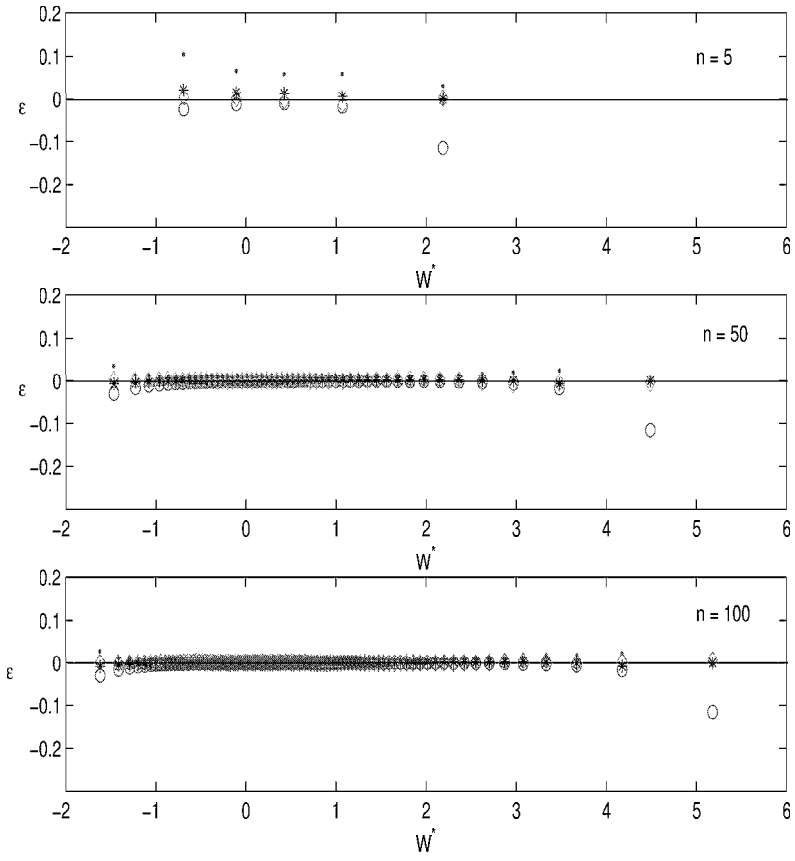

FIGURE 3. Differences between the transformed White's table values, W_i^* and q_{ij}'s, $j = 1, 3, 5, 6$. Here $*$ symbolize $\varepsilon_{i1} = W_i^* - q_{i1}$; $\bullet$ symbolize $\varepsilon_{i3} = W_i^* - q_{i3}$; $\circ$ symbolize $\varepsilon_{i5} = W_i^* - q_{i5}$; $\diamond$ symbolize $\varepsilon_{i6} = W_i^* - q_{i6}$.

Figure 3 shows the results of ε_{ij} for $i = 1, 2, \ldots, n$ and $j = 1, 3, 5, 6$. The plot reveals that the differences ε_{i1} and ε_{i6} are small over all orders i. It also shows that ε_{i3} has more variability over different orders in the left tail for small sample sizes and ε_{i5}, if we ignore the last order observation, gives good results than ε_{i3}.

4. Conclusions

The expression for the first moment of the order statistics from the standard extreme value distribution is difficult to use in practical applications especially for large sample sizes. The proposed methods for approximating the standard extreme value distribution gives approximated values with good precision. The differences between our model and the exact values are less than two decimal places and are of the same magnitude as the differences between the Monte-Carlo estimates of the mean order statistics and the true values. They provide better fit than previously given models.

References

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

[2] 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.

[3] 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 ans Statistics, McMaster University, Hamilton, Canada.

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

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

[6] Evans, J.W., Johnson, R.A. and Green, D.W. (1989). Two- and Three- Parameter Weibull Goodness-of-Fit Tests, *Research Paper FPL-RP-493*, United States Department of Agriculture, Forest Products Laboratory

[7] Filliben, J.J. (1975). The probability plot correlation coefficient test for normality, *Technometrics* **17**, 111-117.

[8] Gringorten, I.I. (1963). A plotting rule for extreme probability paper, *J. Geophys. Res.* **68**, 813-814.

[9] Hosking, J.R.M. (1990). L-moments: Analysis and Estimation of Distributions using Linear Combinations of Order Statistics, *J. R. Statist. Soc. B* **52**, 105-124.

[10] 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.

[11] Mann, N.R., Scheuer, E.M. and Fertig, K.W. (1973). A new goodness of fit test for the two-parameter Weibull or extreme value distribution, *Communications in Statistics* **2**, 383-400.

[12] Sarhan, A.E. and Greenberg, B. G. (eds.) (1962). *Contributions to Order Statistics*, John Wiley, New York.

[13] Tiku, M.L. and Singh, M. (1981). Testing for the two parameter Weibull or extreme value distribution, *Communications in Statistics, Theory and Methods* **10**, 907-918.

[14] Vogel, R.M. (1986). The Probability Plot Correlation Coefficient Test for the Normal, Lognormal, and Gumbel Distributional Hypotheses, *Water Resources Research* **22**, 587-590

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

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

[17] Yu G.H. and Huang C.C. (2001). A distribution free plotting position, *Stochastic Environmental Research and Risk Assessment* **15**, 462-476.