

Prediction of Future Order Statistics Coming from Pareto Model

S. M. T. K. Mirmostafae and J. Ahmadi

Abstract In this paper, we study the problem of predicting the j -th order statistic and also the geometric mean of a future sample of size m from two parameter Pareto distribution on the basis of the observed censored records coming from the same distribution. Numerical computations are given for illustration of the proposed procedures.

Key words: Gamma distribution, Pivotal quantity, Statistical computing, Two-sided prediction interval.

1 Introduction

Let $X_1, \dots, X_m$ be a random sample, if these random variables are arranged in ascending order of magnitude and then written as $X_{1:m} \leq X_{2:m} \leq \dots \leq X_{m:m}$, then $X_{i:m}$ is said to be the i -th order statistic in a sample of size m . Order statistics and functions of these statistics play an important role in a wide range of theoretical and practical problems such as characterization of probability distributions and goodness-of-fit tests, entropy estimation, analysis of censored samples, reliability analysis, quality control and strength of materials; see David and Nagaraja (2003), Arnold et al. (1992) and the references therein for more details.

Next, let $X_1, X_2, \dots$ be an infinite sequence of random variables. Then, an observation X_j is said an upper record value if it exceeds all its previous observations, i.e., if $X_j > X_i$ for every $i < j$. Record data arise in a wide variety of practical situations

This paper is supported in part by the Iranian National Foundation of Elites.

S. M. T. K. Mirmostafae

Ferdowsi University of Mashhad, Mashhad, Iran, 91775-1159., e-mail: ta_mi182@stu-mail.um.ac.ir ;

J. Ahmadi

Ferdowsi University of Mashhad, Mashhad, Iran, 91775-1159. e-mail: ahmadi-j@um.ac.ir

including industrial stress testing, meteorological analysis, hydrology, seismology, sporting and athletic events, and oil and mining surveys. Interested readers may refer to the book by Arnold et al. (1998) and the references therein.

Several authors have studied the problem of predicting future order statistics based on observed order statistics in parametric or nonparametric models and also prediction future records on the basis of observed records from the same distribution. See for example, Lawless (1977), Dunsmore (1983), Raqab and Nagaraja (1995), Ahmadi et al. (2005) and the references therein. Recently, Ahmadi and Balakrishnan (2009) have obtained several nonparametric prediction intervals for future order statistics(record values) from an independent Y-sequence based on order statistics(records) from X-sequence. Also, Ahmadi and Mirmostafae (2009) have discussed the problem of predicting future order statistics based on observed usual records from an exponential distribution.

Here, we study the same problem for a Pareto model but with the assumption that the records are censored. Let X have a two parameter Pareto distribution, denoted by $P(\alpha, \beta)$, with cdf

$$F(x; \alpha, \beta) = 1 - \left(\frac{\beta}{x}\right)^\alpha, \quad x \geq \beta > 0, \alpha > 0. \quad (1)$$

This distribution is called Pareto type I distribution. It has been used often to model naturally occurring phenomena in which the distributions of random variables of interest has long tails. See e.g. Arnold (1983).

The rest of this paper is organized as follows: First we study the problem of prediction of the j -th order statistic in a future sample from $P(\alpha, \beta)$ on the basis of observed censored records from the same underlying distribution in section 2. In section 3, we obtain prediction intervals for the geometric mean of a future sample of size m based on record values. Finally, we use a simulated data set in section 4 in order to illustrate the proposed procedures.

2 Prediction Intervals for Future Order Statistics

Suppose $R_1, R_2, \dots, R_n$ are the upper record values from $P(\alpha, \beta)$ but we have failed to observe the first $(s-1)$ records i.e. we observed $(r_s, r_{s+1}, \dots, r_n)$ for $1 \leq s < n$. Let $Y_{j:m}$ be the j -th smallest observation in a future sample of size m from the same model. In this section we intend to construct prediction intervals for $Y_{j:m}$ on the basis of these observed upper record values. First, consider the following lemma which will be helpful in finding the results.

Lemma 1. *Let $\{X_i, i \geq 1\}$ be a sequence of iid continuous variables with cumulative distribution function $F(x)$ and $H(w) = -\log(1 - F(w))$, then*

(i) $H(R_k) \stackrel{d}{=} E_{R_k}$ where E_{R_k} is the k -th upper record value arising from a sequence

of iid standard exponential random variables and we have $2E_{R_k} \sim \chi^2(2k)$.

(ii) $T = E_{R_n} - E_{R_s} \sim \Gamma(n-s, 1)$ and also T and E_{R_s} are independent.

Proof. If X is a continuous random variable, then $H(X)$ has a standard exponential distribution. Part (i) follows by notifying that $H(x)$ is strictly increasing with respect to x . For part (ii), see Arnold et al. (1998).

Now we are ready to introduce a suitable pivotal quantity in order to fulfill our intentions. Consider the quantity

$$W = \frac{\log Y_{j:m} - \log R_s}{\log R_n - \log R_s} \quad (2)$$

which does not depend on α and β . Let $S_{j:m} = H(Y_{j:m}) = -\alpha \log \frac{\beta}{Y_{j:m}}$, $T_s = H(R_s) = -\alpha \log \frac{\beta}{R_s}$ and $T_n = H(R_n) - H(R_s) = \alpha(\log R_n - \log R_s)$. By Lemma 1, T_s has a Gamma distribution with the shape parameter s and the scale parameter 1, $2T_n$ has a chi-square distribution with $2(n-s)$ degrees of freedom and T_s and T_n are independent. Since $H(x)$ is strictly increasing with respect to x , it is easy to confirm that $S_{j:m} \stackrel{d}{=} E_{j:m}$ where $E_{j:m}$ is the j -th order statistic of a sample of size m from standard exponential distribution. Therefore the pdf of $S_{j:m}$ can be given by (see e.g. Arnold et al., 1992)

$$f_{S_{j:m}}(s) = j \binom{m}{j} (1 - e^{-s})^{j-1} e^{-s(m-j+1)}, \quad s > 0.$$

So W can be rewritten as $W = \frac{S_{j:m} - T_s}{T_n}$. We obtained the cdf of W in the next result, which is free of any parameter.

Lemma 2. Let $(R_s, R_{s+1}, \dots, R_n)$, $1 \leq s < n$, be the upper record values from a Pareto model with cdf defined in (1) and $Y_{j:m}$ be the j -th smallest observation in a future sample of size m from the same model. Then the cdf of W , as defined in (2), is

$$Pr(W \leq w) = \begin{cases} j \binom{m}{j} \sum_{i=0}^{j-1} \sum_{k=0}^{s-1} \frac{(-1)^{i+k} \binom{j-1}{i} \binom{n-s+k-1}{k} [1 - (m-j+i+2)^{k-s}] w^k}{(m-j+i+1)(1-w)^{n-s+k}} & \text{for } w \leq 0, \\ 1 - j \binom{m}{j} \sum_{i=0}^{j-1} \binom{j-1}{i} (-1)^i \frac{[1+w(m-j+i+1)]^{s-n}}{(m-j+i+1)(m-j+i+2)^s} & \text{for } w > 0. \end{cases}$$

Proof. First we find the distribution of $V = S_{j:m} - T_s$. As $S_{j:m}$ and T_s are independent, we get

$$\begin{aligned} Pr(V \leq v) &= \int_0^\infty Pr(S_{j:m} - T_s \leq v | S_{j:m} = t) dF_{S_{j:m}}(t) \\ &= \int_0^\infty Pr(T_s \geq t - v) dF_{S_{j:m}}(t). \end{aligned}$$

By assumptions $T_s \sim \Gamma(s, 1)$, so

$$Pr(T_s \geq t - v) = \begin{cases} 1 & \text{if } t - v \leq 0, \\ \frac{1}{(s-1)!} \int_{t-v}^{\infty} x^{s-1} e^{-x} dx & \text{if } t - v > 0. \end{cases}$$

Thus we have

$$Pr(V \leq v) = Pr(S_{j:m} \leq v) + \int_v^{\infty} \int_{t-v}^{\infty} \frac{x^{s-1} e^{-x}}{(s-1)!} dx f_{S_{j:m}}(t) dt. \quad (3)$$

Using the following equality

$$\int_t^{\infty} \frac{\sigma^n x^{n-1}}{(n-1)!} e^{-\sigma x} dx = \sum_{j=0}^{n-1} e^{-\sigma t} \frac{(\sigma t)^j}{j!}, \quad (4)$$

and by interchanging the order of integrals, for $v < 0$, (3) can be rewritten as

$$\begin{aligned} Pr(V \leq v) &= \int_0^{\infty} \int_{t-v}^{\infty} \frac{x^{s-1} e^{-x}}{(s-1)!} dx f_{S_{j:m}}(t) dt \\ &= \int_{-v}^{\infty} \int_0^{x+v} \frac{x^{s-1} e^{-x}}{(s-1)!} j \binom{m}{j} (1 - e^{-t})^{j-1} e^{-t(m-j+1)} dt dx \\ &= j \binom{m}{j} e^v \sum_{i=0}^{j-1} \sum_{k=0}^{s-1} \frac{(-1)^{i+k} \binom{j-1}{i} [1 - (m-j+i+2)^{k-s}] v^k}{(m-j+i+1)k!}. \end{aligned} \quad (5)$$

For $v \geq 0$, from (3) and (4) and by interchanging the order of integrals, we have

$$\begin{aligned} Pr(V \leq v) &= 1 - \int_v^{\infty} f_{S_{j:m}}(t) dt + \int_v^{\infty} \int_{t-v}^{\infty} \frac{x^{s-1} e^{-x}}{(s-1)!} dx f_{S_{j:m}}(t) dt \\ &= 1 - \int_v^{\infty} f_{S_{j:m}}(t) dt + \int_0^{x+v} \frac{x^{s-1} e^{-x}}{(s-1)!} j \binom{m}{j} (1 - e^{-t})^{j-1} e^{-t(m-j+1)} dt dx \\ &= 1 - \int_0^{\infty} \int_{x+v}^{\infty} \frac{x^{s-1} e^{-x}}{(s-1)!} j \binom{m}{j} (1 - e^{-t})^{j-1} e^{-t(m-j+1)} dt dx \\ &= 1 - j \binom{m}{j} \sum_{i=0}^{j-1} \frac{(-1)^i \binom{j-1}{i} e^{-v(m-j+i+1)}}{(m-j+i+1)(m-j+i+2)^s}. \end{aligned} \quad (6)$$

Let $g_{T_n}(u)$ denote the pdf of T_n , using the conditional argument we get

$$\begin{aligned} Pr(W \leq w) &= \int_0^{\infty} Pr(W \leq w | T_n = u) g_{T_n}(u) du \\ &= \int_0^{\infty} Pr(V \leq wu) \frac{u^{n-s-1}}{(n-s-1)!} e^{-u} du, \end{aligned} \quad (7)$$

where the second equality is obtained by independency of V and T_n . By substituting (5) and (6) in (7), we have the result. $\square$

The next theorem is the main result of this section which is an immediate consequence of Lemma 2.

Theorem 1. *Suppose that the conditions of Lemma 2 hold, then a two-sided prediction interval for $Y_{j:m}$ is*

$$\left(R_s \left[\frac{R_n}{R_s} \right]^{w_\delta(n,m,s,j)}, R_s \left[\frac{R_n}{R_s} \right]^{w_\eta(n,m,s,j)} \right)$$

whose coefficient is free of α and β and is given by $\eta - \delta$, where $w_\gamma(n,m,s,j)$ is the γ th quantile of W , i.e., $\Pr(W \leq w_\gamma(n,m,s,j)) = \gamma$.

One can compute $w_\gamma(n,m,s,j)$ numerically for desired values of γ , n , m , s and j . We give an example in section 4.

3 Prediction Intervals for Future Geometric Mean

Geometric mean is of importance in economical situations. So in this section we want to obtain interval predictors for the geometric mean of a future sample of size m based on censored record values from $\text{Pareto}(\alpha, \beta)$. Let $R_s, R_{s+1}, \dots, R_n$ be the $n - s + 1$ record values from $P(\alpha, \beta)$ where $1 \leq s < n$. Let $G_m = (\prod_{i=1}^m Y_i)^{\frac{1}{m}}$ be the geometric mean of a future sample of size m from the same distribution. We wish, on the basis of observed $r_s, r_{s+1}, \dots, r_n$ to construct prediction intervals for G_m . To do this, consider the quantity

$$U = \frac{\log G_m - \log R_s}{\log R_n - \log R_s}. \quad (8)$$

Let $T_u = \frac{1}{m} \sum_{i=1}^m H(Y_i) = -\alpha \log \frac{\beta}{G_m}$, $T_2 = -\alpha \log \frac{\beta}{R_s}$ and $T_3 = \alpha(\log R_n - \log R_s)$. As mentioned before, T_2 has a Gamma distribution with the shape parameter s and the scale parameter 1, $2T_s$ has a chi-square distribution with $2(n-s)$ degrees of freedom and T_2 and T_3 are independent. Also mT_u has a gamma distribution with shape parameter m and scale parameter 1 as $H(Y_i)$ has a standard exponential distribution. So we can express U in terms of T_u, T_2 and T_3 as $U = \frac{T_u - T_2}{T_3}$, whose distribution is obtained in the next result.

Lemma 3. *Let $(R_s, R_{s+1}, \dots, R_n), 1 \leq s < n$, be the record values from $P(\alpha, \beta)$ and G_m be the geometric mean of a future sample of size m from the same distribution. Then the cdf of U , as defined in (8), is*

$$\Pr(U \leq u) = \begin{cases} \sum_{k=0}^{s-1} \sum_{i=0}^k \binom{k-i+m-1}{m-1} \binom{n-s+i-1}{i} q_1^i p_1^{n-s} q_3^m p_3^{k-i} & \text{for } u \leq 0, \\ 1 - \sum_{k=0}^{m-1} \sum_{i=0}^k \binom{k-i+s-1}{s-1} \binom{n-s+i-1}{i} q_2^i p_2^{n-s} q_3^s p_3^{k-i} & \text{for } u > 0, \end{cases}$$

where $p_1 = \frac{1}{1-u}$, $p_2 = \frac{1}{mu+1}$, $p_3 = \frac{1}{m+1}$ and $q_j = 1 - p_j$, $j = 1, 2, 3$.

Proof. Proof is similar to that of Lemma 2 and therefore is omitted. $\square$

According to Lemma 3, the distribution of $U = \frac{T_u - T_2}{T_3}$ does not depend on the parameters and depends only on n, s and m . Therefore, we can construct prediction intervals for G_m needless of knowing the values of α and β . We have the following result.

Theorem 2. Under the assumptions of Lemma 3,

$$\left(R_s \left[\frac{R_n}{R_s} \right]^{u_{\delta}(n,s,m)}, R_s \left[\frac{R_n}{R_s} \right]^{u_{\eta}(n,s,m)} \right)$$

is a prediction interval for G_m whose confidence coefficient is free of α and β and is given by $\eta - \delta$, where $u_{\gamma}(n,s,m)$ is the γ th quantile of U , i.e., $\Pr(U \leq u_{\gamma}(n,s,m)) = \gamma$.

One can compute $u_{\gamma}(n,s,m)$ numerically for desired values of γ , n , m and s .

4 Illustrative Example

In order to illustrate the procedures developed in this paper, we use the data set simulated by Raqab et al. (2007). The first 6 upper record values were simulated from a two parameter Pareto distribution with $\alpha = 3.7676$ and $\beta = 1.5096$. They are as follows:

$$1.9144, 4.1615, 8.2738, 17.8995, 20.1482, 20.3658.$$

Here we can see that $s = 1$ and $n = 6$. So based on Theorems 1 and 2, we can construct prediction intervals for order statistics and the geometric mean of a future sample from the same distribution. Table 1 includes 90% and 95% two-sided equal-tailed prediction intervals for $Y_{17:20}$, $Y_{20:20}$ and the geometric mean in a future sample of size $m = 20$.

Table 1 90% and 95% two-sided prediction intervals (PIs) for $Y_{17:20}$, $Y_{20:20}$ and the geometric mean.

	$Y_{17:20}$	$Y_{20:20}$	Geometric Mean
90% PI	(0.9270, 9.6383)	(1.9588, 19.4251)	(5.8148, 10.0478)
95% PI	(0.5847, 14.5575)	(1.3831, 25.7981)	(5.7937, 12.0116)

Acknowledgements The authors are thankful to an anonymous referee for the valuable comments.

References

1. Ahmadi, J., Balakrishnan, N.: Prediction of order statistics and record values from two independent sequences. *Statistics* (2009) DOI: 10.1080/02331880903189133
2. Ahmadi, J., Doostparast, M., Parsian, A.: Bayesian estimation and prediction for some life distributions based on record values. *Comm. Statist. Theory Methods* **34**, no.4, 795–805 (2005)
3. Ahmadi, J., MirMostafaei, S. M. T. K.: Prediction intervals for future records and order statistics coming from two parameter exponential distribution. *Statist. Probab. Lett.* **79**, 977–983 (2009)
4. Arnold, B. C., Balakrishnan, N., Nagaraja, H. N.: *Pareto Distributions*. Maryland: International Co-operative Publishing House, Fairland (1983)
5. Arnold, B. C., Balakrishnan, N., Nagaraja, H. N.: *A First Course in Order Statistics*. John Wiley & Sons, New York (1992)
6. Arnold, B. C., Balakrishnan, N., Nagaraja, H. N.: *Records*. John Wiley & Sons, New York (1998)
7. David, H. A., Nagaraja, H. N.: *Order Statistics*. 3rd edition, John Wiley & Sons, New York (2003)
8. Dunsmore, I. R.: The future occurrence of records. *Ann. Inst. Statist. Math.* **35**, 276–277 (1983)
9. Lawless, J. F.: Prediction intervals for the two parameter exponential distribution. *Technometrics* **19**, 469–472 (1977)
10. Raqab, M. Z., Ahmadi, J., Doostparast, M.: Statistical inference based on record data from Pareto model. *Statistics* **41**, 105–118 (2007)
11. Raqab, M. Z., Nagaraja, H. N.: On some predictors of future order statistic. *Metron* **53**, 185–204 (1995)