Estimation of Parameters and Larger Quantiles Based on the k Largest Observations

Author(s): Ishay Weissman

Source: *Journal of the American Statistical Association*, Dec., 1978, Vol. 73, No. 364 (Dec., 1978), pp. 812-815

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/2286285

Estimation of Parameters and Large Quantiles Based on the k Largest Observations

ISHAY WEISSMAN*

We consider an estimation problem when only the k largest observations of a sample of size n are available. It is assumed that the underlying distribution function F belongs to the domain of attraction of a known extreme-value distribution and that k remains fixed as $n \to \infty$. We present estimators for the location and scale parameters and for p-quantiles of F, where p is of the form $1 - c/n$ (c fixed). These estimators are either asymptotically maximum likelihood or minimum variance.

KEY WORDS: Censored samples; k-dimensional extremal distribution; Quantile estimation; Sample extremes.

1. INTRODUCTION

Consider the following two situations taken from the sporting and insurance fields. The Athletics Council of a certain country tested all 15-year-old male students in shot putting and formed a team of the ten best shot-putters (i.e., those with the ten longest throws). For purposes of future planning, the estimates of some extreme quantiles of the throw-distance distribution of that age-group were requested. However, only the records of the selected students were kept.

An insurance company can protect itself against high losses by reinsurance. One type of reinsurance is ECOMOR (Thepaut 1950), where the reinsurer covers the k largest claims of the year. The value of k, which is agreed upon at the beginning of the year, is usually small (less than ten), while the value of n, the number of policy holders, is in the thousands. The information available to the reinsurer, accumulated in one year, consists of the k largest claims.

In this article we derive estimators for large quantiles based on the available data, as illustrated by the two examples. Specifically, let $X_1, \ldots, X_n$ be a sample from a distribution function (df) F and let $X_{1n} \geq X_{2n} \geq \ldots \geq X_{kn} \geq \ldots \geq X_{nn}$ be the order statistics. The available data are $X_{1n}, \ldots, X_{kn}$ for some fixed k. Suppose that F is in the domain of attraction of a known df G; that is, there exist sequences $a_n > 0$ and b_n such that

$$\Pr\{(X_{1n} - b_n)/a_n \leq x\} = F^n(a_n x + b_n) \to G(x) \quad \text{as} \quad n \to \infty \tag{1.1}$$

for all x in the support of G. We shall be concerned with the estimation of a_n, b_n, and η_p, the p-quantile of F, for p near 1; that is, for p of the form $1 - c/n$ ($c > 0$ fixed).

The available data form a *censored sample* of the type II which has attracted considerable attention in the literature (Sarhan and Greenberg 1962; Johnson and Kotz 1970). Recent related discussions of inference based on extreme order statistics are Johnson (1974), Pickands (1975), and Hill (1975).

As is well-known (Gnedenko 1943, or David 1970, p. 205), the df G which appears in (1.1) must be one of the following df's (up to location and scale parameters):

$$\begin{aligned} \Lambda(x) &= \exp\{-e^{-x}\}, && -\infty < x < \infty \\ \Phi_\alpha(x) &= \exp\{-x^{-\alpha}\}, && x > 0, \alpha > 0 \\ \Psi_\alpha(x) &= \exp\{-(-x)^{\alpha}\}, && x < 0, \alpha > 0 . \end{aligned} \tag{1.2}$$

Note that apart from (1.1), we stipulate nothing concerning F. Our results are asymptotic in nature and can also be considered nonparametric within the wide class of df's satisfying (1.1). For example, the gamma, exponential, Weibull, normal, lognormal, and logistic distributions all satisfy (1.1) with the same $G = \Lambda$. Thus the estimators suggested here are useful for tail inference (based on the sample tail) even when all the observations are available, but an exact parametric family is not assumed.

The limiting distribution of $(X_{1n}, \ldots, X_{kn})$ is discussed in Section 2. In Section 3 we derive estimators for location and scale parameters of the limiting process, and in Section 4 we apply them to the sample extremes. In Section 5 we compare our estimators to those obtained when F in (1.1) is exponential.

2. THE k-DIMENSIONAL EXTREMAL DISTRIBUTION

Let $m_{in} = (X_{in} - b_n)/a_n$ and let $I_n(A)$ be the number of m_{in}, $i = 1, 2, \ldots, n$, in the Borel set A. Clearly, $I_n((x, \infty]) \equiv I_n(x)$ is binomial with probability of success $1 - F(a_n x + b_n)$. Since (1.1) is equivalent to $n[1 - F(a_n x + b_n)] \to -\ln G(x) \equiv \lambda(x)$, $I_n(x)$ converges in distribution (denoted by $\xrightarrow{D}$) to a Poisson random variable $I(x)$, say, whose mean is $\lambda(x)$. This convergence holds in a stronger sense.

Theorem 1 (Weissman 1975): Under (1.1), there exists on R a Poisson random measure I, for which $EI((x, \infty])$

* Ishay Weissman is Senior Lecturer in Statistics, Faculty of Industrial and Management Engineering, Technion, Haifa, Israel. The author wishes to thank Benjamin Epstein for helpful discussions. The percentage points of U_k were computed by Paul D. Feigin, to whom the author expresses his appreciation.

© Journal of the American Statistical Association
December 1978, Volume 73, Number 364
Theory and Methods Section

$= \lambda(x)$, such that

$$(I_n(A_1), \ldots, I_n(A_q)) \xrightarrow{D} (I(A_1), \ldots, I(A_q)) \quad (2.1)$$

for every finite collection of Borel sets A_i.

Note that a Poisson random measure I is a counting measure for which $I(A)$ is Poisson and $I(A_1), I(A_2), \ldots$ are independent whenever $A_1, A_2, \ldots$ are disjoint. Now denote $I(x) = I((x, \infty])$ and define $m_i = \inf\{x: I(x) \leq i - 1\}$ and $M_k = (m_1, \ldots, m_k)$; the latter is known as the k-dimensional extremal variate (Dwass 1966).

Theorem 2: Under (1.1), for each fixed k,

$$(m_{1n}, \ldots, m_{kn}) \xrightarrow{D} M_k \quad \text{as} \quad n \to \infty \ . \quad (2.2)$$

(See Dwass 1966; Weissman 1975 for the proof.) It follows from the definition of m_i that

$$\Pr\{m_i \leq x\} = \Pr\{I(x) \leq i - 1\}$$
$$= G(x) \sum_{j=0}^{i-1} \lambda^j(x)/j! = \frac{1}{(i-1)!} \int_{\lambda(x)}^{\infty} e^{-t} t^{i-1} dt \ . \quad (2.3)$$

Hence, $\lambda(m_i)$ is a gamma variate (parameters 1 and i).

We turn now to the particular case $G = \Lambda$. Here $\lambda(x) = e^{-x}$, the density of m_i is

$$\phi_i(x) = [1/(i-1)!] \exp\{-e^{-x} - ix\} \ , \quad -\infty < x < \infty \ , \quad (2.4)$$

and the density of M_k is

$$\psi_k(x_1, \ldots, x_k)$$
$$= \exp\{-e^{-x_k} - \sum_{i=1}^{k} x_i\} \quad (x_1 \geq x_2 \geq \ldots \geq x_k) \ . \quad (2.5)$$

The joint density of the spacings $D_i = m_i - m_{i+1}$ and m_k is

$$\tilde{\psi}_k(d_1, \ldots, d_{k-1}, m_k) = \phi_k(x_k) \prod_{i=1}^{k-1} i \exp\{-id_i\}$$
$$(d_i \geq 0; i = 1, \ldots, k-1) \ . \quad (2.6)$$

Thus, the following result has been obtained:

Theorem 3: If $\lambda(x) = e^{-x}$, then the process $M = \{m_i: i = 1, 2, \ldots\}$ has the property that for each k, $\{m_k, D_1, \ldots, D_{k-1}\}$ are independent and each D_i is exponential with mean i^{-1}.

This result is not surprising since the exponential distribution is itself in the domain of attraction of Λ. It is well-known that for each n, the spacings $D_{in} = X_{in} - X_{(i+1)n}$, $1 \leq i \leq n-1$, are independent exponential variates with means proportional to i^{-1}. The converse is also true.

Theorem 4: If m_k and $m_l - m_k$ are independent for some $l < k$, then $\lambda(x) = \exp\{-\alpha(x - \beta)\}$, $\alpha > 0$.

Proof: By the independence of increments of I,

$$\Pr\{m_l - m_k \leq z \mid m_k = x\} = \Pr\{I(x+z) \leq l-1 \mid I(x-\epsilon) - I(x) > 0 \ \forall \epsilon > 0, I(x) = k-1\}$$
$$= \Pr\{I(x+z) \leq l-1 \mid I(x) = k-1\} \ . \quad (2.7)$$

But $I(x) = [I(x) - I(x+z)] + I(x+z)$ is a sum of two independent Poisson random variables, thus (2.7) is just $\sum_{i=0}^{l-1} \binom{k-1}{i} \pi^i (1-\pi)^{k-1-i}$ with $\pi = \lambda(x+z)/\lambda(x)$. Hence (2.7) does not depend on x if and only if π does not. By a routine argument, $\lambda(x)$ must be of the form $\exp\{-\alpha(x - \beta)\}$.

For $\lambda(x) = e^{-x}$, straightforward calculations yield

$$\mu_k = Em_k = \gamma - \sum_{j=1}^{k-1} j^{-1} \equiv \gamma - S_k \ , \quad (S_1 = 0) \ , \quad (2.8)$$

where $\gamma = .5772\ldots.$ is Euler's constant and

$$\sigma^2_k = \operatorname{var} m_k = \sum_{j=k}^{\infty} j^{-2} = \frac{\pi^2}{6} - \sum_{j=1}^{k-1} j^{-2}$$
$$(\sigma^2_1 = \pi^2/6) \ . \quad (2.9)$$

Remark: The asymptotic exponentiality of the spacings in the case $\lambda(x) = e^{-x}$ is known (Darwin 1957; Gumbel 1958, p. 198), but we have not found any reference for the asymptotic independence.

3. ESTIMATION BASED ON M_k

We say that M is extremal with parameters θ and δ ($\delta > 0$) (and write $\operatorname{Ext}(\theta, \delta)$) if $\lambda(x) = \exp\{-(x-\theta)/\delta\}$. Suppose M is $\operatorname{Ext}(\theta, \delta)$ and that we have observed M_k. The likelihood function is, by (2.5),

$$L(\theta, \delta) = \delta^{-k} \exp\{-e^{-(m_k - \theta)/\delta} - \sum_{i=1}^{k} (m_i - \theta)/\delta\} \ .$$

Hence the pair $(m_k, \sum_{i=1}^{k} m_i)$ is a sufficient statistic for (θ, δ).

Theorem 5: The maximum likelihood estimators (MLEs) of δ and θ based on M_k are

$$\hat{\delta} = \bar{m}_k - m_k \quad \text{and} \quad \hat{\theta} = \hat{\delta} \ln k + m_k \ , \quad (3.1)$$

where $\bar{m}_j = j^{-1} \sum_{i=1}^{j} m_i$. The minimum variance unbiased estimators (MVUEs) are

$$\delta^* = \bar{m}_{k-1} - m_k = \hat{\delta} k (k-1)^{-1} \ ; \quad \theta^* = \delta^*(S_k - \gamma) + m_k \ , \quad (3.2)$$

with variances

$$\operatorname{var}\delta^* = \delta^2/(k-1) \ ; \quad \operatorname{var}\theta^* = \delta^2\{(S_k - \gamma)^2/(k-1) + \sigma^2_k\} \ . \quad (3.3)$$

Proof: The MLEs can be verified directly. The unbiasedness of the estimators (3.2) can be verified using (2.8) and Theorem 3. Since δ^* and θ^* are functions of a sufficient statistic, one need only show completeness of the sufficient statistic. Note that the pair $(m_k, T = \sum_{i=1}^{k-1} iD_i)$ is also sufficient, m_k and T are independent, $(m_k - \theta)/\delta$ has density (2.4) (with $i = k$), and T/δ is a gamma variate. Suppose now that $h(m_k, T)$ is a statistic, and $Eh = 0$ for all $\delta > 0$ and all θ; that is,

$$\int_0^{\infty} \int_{-\infty}^{\infty} h(x, t) \exp\{-e^{-(x-\theta)/\delta} - k(x - \theta)/\delta\} dx \cdot \exp\{-t/\delta\} t^{k-2} dt = 0 \ ,$$

Percentage Points $U_k(p)$

	p								
k	0.010	0.025	0.050	0.100	0.500	0.900	0.950	0.975	0.990
2	−58.3	−23.0	−11.2	−5.32	−0.543	1.06	2.71	6.00	15.9
3	−14.7	−8.90	−5.99	−3.92	−1.04	−0.093	0.225	0.649	1.48
4	−9.69	−6.77	−5.07	−3.70	−1.37	−0.470	−0.275	−0.0825	0.216
5	−8.00	−5.99	−4.72	−3.66	−1.61	−0.724	−0.552	−0.404	−0.221
6	−7.15	−5.60	−4.59	−3.67	−1.81	−0.925	−0.755	−0.620	−0.467
8	−6.38	−5.25	−4.47	−3.75	−2.10	−1.24	−1.06	−0.931	−0.791
10	−6.02	−5.10	−4.46	−3.83	−2.33	−1.48	−1.31	−1.170	−1.03
12	−5.83	−5.04	−4.47	−3.91	−2.51	−1.67	−1.50	−1.36	−1.22
14	−5.71	−5.02	−4.50	−3.98	−2.67	−1.85	−1.66	−1.53	−1.39
16	−5.64	−5.01	−4.53	−4.05	−2.79	−1.99	−1.82	−1.68	−1.53
18	−5.59	−5.02	−4.57	−4.11	−2.91	−2.13	−1.94	−1.80	−1.66
20	−5.57	−5.02	−4.60	−4.18	−3.02	−2.24	−2.05	−1.92	−1.77
25	−5.54	−5.06	−4.70	−4.32	−3.24	−2.47	−2.31	−2.17	−2.02
30	−5.54	−5.10	−4.79	−4.44	−3.42	−2.69	−2.52	−2.38	−2.23

or equivalently,

$$\int_0^\infty \int_{-\infty}^\infty [h(\theta + u/k, t) \exp\{-e^{-u/k\delta}\} t^{k-2}] \cdot \exp\{-(u+t)/\delta\} du dt = 0$$

for all $\delta > 0$ and θ. By the uniqueness of the Laplace transform, the expression in the square brackets is 0, a.e., which is possible only if $h(x, t) = 0$ a.e.

Note that $2(k-1)\delta^*/\delta = 2T/\delta$ is a χ^2 variate with $2(k-1)$ degrees of freedom; thus confidence intervals and tests of significance for δ can be constructed. Similarly, the distribution of $U_k = (m_k - \theta)/\delta^*$ is parameter-free and its percentage points enable us to construct confidence intervals for θ. Based on the df of U_k,

$$\Pr\{U_k \le (k-1)z\} = \int_0^\infty [\sum_{j=0}^{k-1} \exp\{-e^{-yz}\} e^{-yzj}/j!] e^{-y} y^{k-2} dy ,$$

my colleague Paul Feigin computed some percentage points $U_k(p)$ which appear in the table.

4. APPLICATIONS TO THE SAMPLE EXTREMES

According to the type of G in (1.1), three cases arise:

(i) In case $G = \Lambda$, by (2.2), $(X_{1n}, \ldots, X_{kn})$ is approximately (for large n) Ext(b_n, a_n), and thus by substituting X_{in} in (3.1) and (3.2), we get estimators for a_n and b_n:

$$\hat{a}_n = \bar{X}_{kn} - X_{kn} \; ; \qquad \hat{b}_n = \hat{a}_n \ln k + X_{kn} \; ; \tag{4.1}$$

$$a_n^* = \bar{X}_{(k-1)n} - X_{kn} \; ; \quad b_n^* = a_n^*(S_k - \gamma) + X_{kn} \; . \tag{4.2}$$

It is well-known (Gnedenko 1943) that $b_n = \eta_{1-(1/n)}$ and $a_n = \eta_{1-e^{-1}/n} - b_n$. By a theorem of Mejzler (1949) (which can be found also in De Hann 1971, p. 76), (1.1) holds with $G = \Lambda$ if and only if $(\eta_{1-c/n} - b_n)/a_n \to -\ln c$ as $n \to \infty$ for every $c > 0$. Hence, $\eta_{1-c/n}$ is approximated by $a_n(-\ln c) + b_n$ and estimated either by the MLE

$$\hat{\eta}_{1-c/n} = \hat{a}_n(-\ln c) + \hat{b}_n = \hat{a}_n \ln (k/c) + X_{kn} \; , \tag{4.3}$$

or by the MVUE

$$\eta_{1-c/n}^* = a_n^*(-\ln c) + b_n^* = a_n^*(S_k - \gamma - \ln c) + X_{kn} \; . \tag{4.4}$$

(ii) In case $G = \Phi_\alpha$, the relation $\Phi_\alpha(x) = \Lambda(\alpha \ln x)$ $(x > 0)$ implies that the limiting distribution of the $Y_{in} = \ln X_{in}$ is the same as in case (i); thus for large n,

$$\hat{\alpha}^{-1} = \bar{Y}_{kn} - Y_{kn} \quad \text{and} \quad \hat{\eta}_{1-c/n} = (k/c)^{\hat{\alpha}^{-1}} X_{kn} \; .$$

(iii) In case $G = \Psi_\alpha$, $x_* = \inf\{x: F(x) = 1\}$ must be finite and $\Psi_\alpha(x) = \Lambda(-\alpha \ln(-x))$, $x < 0$. In this case the limiting distribution of the $Z_{in} = -\ln(x_* - X_{in})$ is the same as in case (i). Thus, when x_* is known,

$$\hat{\alpha}^{-1} = \bar{Z}_{kn} - Z_{kn} \quad \text{and} \quad \hat{\eta}_{1-c/n} = x_* - (c/k)^{\hat{\alpha}^{-1}}(x_* - X_{kn}) \; .$$

When x_* is unknown, we have a three-parameter problem, and a different approach is needed.

5. EXPONENTIAL EXAMPLE

Let $F(x) = 1 - \exp\{-(x-\theta)/\delta\}$ $(x > \theta, \; \delta > 0)$. Then the likelihood function based on the k largest observations is

$$\frac{n!}{(n-k)!} \delta^{-k}[1 - \exp\{-(x_k - \theta)/\delta\}]^{n-k} \cdot \exp\{-\sum_{j=1}^{k} (x_i - \theta)/\delta\} \; . \tag{5.1}$$

By straightforward calculations, we get the MLEs and MVUEs. It turns out that $\hat{\delta}$, δ^*, and $\hat{\eta}_{1-c/n}$ are the same as in (4.1), (4.2), and (4.3), respectively. However, $\eta_{1-c/n}^* = \delta^*[S_k - S_{n+1} - \ln(c/n)] + X_{kn}$ differs from (4.4) by a term of order $\delta^*/2n$ (for large n).

Numerous works in the literature deal with estimation of parameters under censoring (under right-censoring in particular). A chief contributor to the area is Epstein (1962). A detailed survey of the subject is given by Johnson and Kotz (1970, Ch. 18), including a long list of references.

6. CONCLUDING REMARKS

To estimate $\eta_{1-c/n}$, one has to assume knowledge of G in (1.1). It is possible to test whether $G = \Lambda$ by testing whether the set $D_{1,n}, 2D_{2,n}, \ldots, (k-1)D_{k-1,n}$ is a set of iid (one-parameter) exponential random variables.

Similarly, the hypothesis that $G = \Phi_\alpha$ or the hypothesis that $G = \Psi_\alpha$ can be tested by applying the above test to the spacings of $\log X_{i,n}$ or of $-\log (x_* - X_{i,n})$. There are a number of procedures for testing exponentiality; e.g., Pyke (1965); Epstein (1960).

Finally, by an obvious transformation, our results are applicable for lower-tail and k smallest observations.

[*Received December 1976. Revised January 1978.*]

REFERENCES

Darwin, J.H. (1957), "The Difference between Consecutive Members of a Series of Random Variables Arranged in Order of Size," *Biometrika*, 44, 211–218.

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

De Hann, Laurens (1971), *On Regular Variation and Its Application to Weak Convergence of Sample Extremes*, Mathematical Center Tract 32, Amsterdam: Mathematisch Centrum.

Dwass, Meyer (1966), "Extremal Processes, II," *Illinois Journal of Mathematics*, 10, 381–391.

Epstein, Benjamin (1960), "Tests for the Validity of the Assumption that the Underlying Distribution of Life is Exponential. Part I," *Technometrics*, 2, 83–101.

——— (1962), "Simple Estimates of the Parameters of Exponential Distributions," in *Contributions to Order Statistics*, eds. Ahmad E. Sarhan and Bernard B. Greenberg, New York: John Wiley & Sons, 361–371.

Gnedenko, Par B. (1943), "Sur la Distribution Limite du Terme Maximum d'une Série Aléatoire," *Annals of Mathematics*, 44, 423–453.

Gumbel, E.J. (1958), *Statistics of Extremes*, New York: Columbia University Press.

Hill, Bruce M. (1975), "A Simple General Approach to Inference about the Tail of a Distribution," *Annals of Statistics*, 3, 1163–1174.

Johnson, Norman L., and Kotz, Samuel (1970), *Continuous Univariate Distributions–1*, Boston: Houghton–Mifflin Co.

Johnson, Richard A. (1974), "Asymptotic Results for Inference Procedures Based on the r Smallest Observations," *Annals of Statistics*, 2, 1138–1151.

Mejzler, David G. (1949), "On a Theorem of Gnedenko," *Sb. Trudov Inst. Mat. Akad. Nauk. Ukrain. R.S.R.*, 12, 31–35 (Russian).

Pickands, James, III (1975), "Statistical Inference Using Extreme Order Statistics," *Annals of Statistics*, 3, 119–131.

Pyke, R. (1965), "Spacings," *Journal of the Royal Statistical Society*, Ser. B, 27, 395–436.

Sarhan, Ahmad E., and Greenberg, Bernard G. (eds.) (1962), *Contributions to Order Statistics*, New York: John Wiley & Sons.

Thepaut, A. (1950), "Le Traite d'Excedent du Cout Moyen Relatif (Ecomor)," *Bulletin Trimestriel de l'Institut des Actuaires Français*, 192.

Weissman, Ishay (1975), "Multivariate Extremal Processes Generated by Independent Non-Identically Distributed Random Variables," *Journal of Applied Probability*, 12, 477–487.