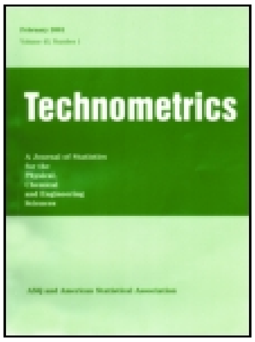

Technometrics

ISSN: 0040-1706 (Print) 1537-2723 (Online) Journal homepage: https://www.tandfonline.com/loi/utch20

Using Extreme Value Theory to Estimate Large Percentiles

Dennis D. Boos

To cite this article: Dennis D. Boos (1984) Using Extreme Value Theory to Estimate Large Percentiles, Technometrics, 26:1, 33-39

To link to this article: https://doi.org/10.1080/00401706.1984.10487919

TECHNOMETRICS ©, VOL. 26, NO. 1, FEBRUARY 1984

Using Extreme Value Theory to Estimate Large Percentiles

Dennis D. Boos

Department of Statistics
North Carolina State University
Raleigh, NC 27650

Weissman (1978) suggested percentile estimators based on the joint limiting distribution of the k largest order statistics. The present work identifies situations where Weissman's estimators are a significant improvement over the usual sample percentile estimators and gives practical advice on how to use these new estimators effectively. In particular, large reductions in mean squared error can be made when the tails of the distributions are approximately exponential and $p \geq .95$.

KEY WORDS: Percentiles; Quantiles; Extreme value theory; Order statistics; Censoring.

1. INTRODUCTION

Often one is interested in estimating the large (or small) percentiles of a distribution based on an ordered sample $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. For example, the sample might arise from Monte Carlo simulations of a new test or pivotal statistic whose distribution is analytically intractable. The standard method for estimating the $100 \times p$th percentile (pth quantile) is to use the empirical percentile given approximately by $X_{(np)}$ or some smoothed version. The SAS routine UNIVARIATE (Chilko 1979) gives four options, and others are mentioned in Section 4. All of these "raw" percentile methods are asymptotically equivalent and basically restricted to $1 \leq np \leq n - 1$. Weissman (1978,1980) suggested a totally different approach based on the joint limiting distribution of the largest k order statistics from random variables in the domain of attraction of $G(x) = \exp(-\exp(-x))$. (The expression "in the domain of attraction of $G(x)$" means that there exist sequences a_n and b_n such that $P((X_{(n)} - b_n)/a_n \leq x) \rightarrow G(x)$ as $n \rightarrow \infty$.) Examples of such random variables include gammas, Weibulls, normals, and lognormals. In practical terms, the approach is based on the fact that the largest k order statistics have a joint distribution that is *approximately* the same as the largest k order statistics from a location and scale exponential distribution. The resulting percentile estimators are simple to compute and can have significantly lower mean squared error (MSE) than $X_{(np)}$ when p lies in $[.95, 1)$. However, when the tails of the distribution are not exactly exponential, the estimators are biased and to control this bias some attention must be given to the number of order statistics k that is used. The purpose of this article is to investigate numerically where the large gains in MSE can be made and to give practical advice on how to use these estimators. Further empirical work and analytic confirmation can be found in Boos (1981) and in Breiman, Stone, and Gins (1979,1981).

The article is organized as follows. Section 2 explains the problem that motivated the research. Section 3 gives the basic theory and percentile estimators, and Section 4 gives the Monte Carlo results and specific recommendations. A numerical example related to Section 2 is presented in Section 5, and Section 6 is a brief summary. Technical details concerning censored situations and confidence intervals are given in the Appendix.

2. A MOTIVATING EXAMPLE

While preparing a paper (Boos 1982) on minimum distance estimation, it became necessary to use Monte Carlo methods to estimate the upper percentiles of a statistic that involved minimization in several dimensions. The minimizations were fairly costly so that only $n = 1{,}000$ replications could be used. A packaged program (see Dickey 1981) processed the Monte Carlo results by placing the output in 800 bins rather than printing out the n exact observations. However, in some situations the range specified in the program was too small, and a few of the largest observations were censored. Likewise, in other cases the largest observations were suspect because certain convergence criteria had not been met. Thus there were situations of Type I censoring and Type II censoring (trimming). The grouping effect of the bins was small compared with the standard error of percentile estimators. Unfortunately, these latter standard errors for the 95th and 99th sample percentile estimators were higher than desired, and no more observations could be taken. The extreme value approach described in the next section turned out to be a useful alternative. We return to this example in Section 5.

3. THE BASIC THEORY AND ESTIMATORS

Weissman's (1978) results can be summarized as follows. Let $X_1, \ldots, X_n$ be a sample from the distribution function F and let $X_{1n} \geq X_{2n} \geq \cdots \geq X_{nn}$ be the order statistics labeled from largest to smallest. Suppose that there are sequences $a_n > 0$ and b_n such that

$$P((X_{1n} - b_n)/a_n \leq x) \rightarrow G(x) \quad \text{as} \quad n \rightarrow \infty \quad (3.1)$$

for all x in the support of G. For convenience, consider only the case $G(x) = \exp(-\exp(-x))$ since results for the other two possible limiting distributions are similar (see Weissman 1978, Secs. 1 and 4). Then for k fixed, Theorem 2 of Weissman yields

$$\left(\frac{X_{1n} - b_n}{a_n}, \ldots, \frac{X_{kn} - b_n}{a_n}\right) \xrightarrow{d} M_k, \quad (3.2)$$

where M_k is the k-dimensional extremal variate with density

$$\psi_k(x_1, \ldots, x_k) = \exp\left(-\exp(-x_k) - \sum_{i=1}^{k} x_i\right)$$

for $x_1 \geq x_2 \geq \cdots \geq x_k$. Thus $(X_{1n}, \ldots, X_{kn})$ has approximately the density

$$\psi_k((x_1 - b_n)/a_n, \ldots, (x_k - b_n)/a_n)/a_n^k,$$

and a_n and b_n can be estimated by maximum likelihood to get $\hat{a}_n = \bar{X}_{kn} - X_{kn}$ and $\hat{b}_n = \hat{a}_n \ln k + X_{kn}$, where $\bar{X}_{kn} = k^{-1}\sum_{i=1}^{k} X_{in}$. Let $\eta_{1-c/n}$ be the upper $100 \times (1 - c/n)$th percentile of F. If (3.1) holds, a further extreme value theory result is $(\eta_{1-c/n} - b_n)/a_n \rightarrow -\ln c$ for each $c > 0$. Thus a natural percentile estimator is $\hat{\eta}_{1-c/n} = \hat{a}_n(-\ln c) + \hat{b}_n = \hat{a}_n \ln(k/c) + X_{kn}$. Modifications for censoring on the right are given in the Appendix. Weissman (1978) also suggested minimum variance unbiased estimators, but Monte Carlo work showed that they are not an improvement over the maximum likelihood estimators, and therefore we will not pursue them further.

In order to see the relationship between the extreme value theory and the exponential distribution, let $X_{1n} \geq \cdots \geq X_{kn}$ be the upper k order statistics from the distribution $F(x) = 1 - \exp(-(x - \mu)/\sigma)$, $x \geq \mu$, $\sigma > 0$. The joint density is then

$$f_k(x_1, \ldots, x_k) = \frac{n!\sigma^{-k}}{(n-k)!}[1 - \exp(-(x_k - \mu)/\sigma)]^{n-k} \times \exp\left(-\sum_{i=1}^{k}(x_i - \mu)/\sigma\right) \quad (3.3)$$

for $x_1 \geq x_2 \geq \cdots \geq x_k$. One can show that the maximum likelihood estimators of σ and η_p based only on these upper k order statistics are exactly the same as those for a_n and η_p. Moreover, one can also show that with proper choice of μ and σ, f_k amd ψ_k are approximately equal. Basically, the standardized upper k order statistics from an exponential distribution converge rapidly to the k-dimensional extremal variate. So we may say that the standardized upper k order statistics from a distribution that satisfies (3.1) have approximately the same distribution as those from an exponential distribution. An advantage of this representation is that one may check assumptions by plotting X_{in} versus $-\ln[i/(n+1)]$ in the usual Q-Q plot for exponential data. For the same purpose Weissman (1978) suggested using the spacings of $X_{1n} \geq \cdots \geq X_{kn}$ because they are approximately distributed as independent exponential random variables. However, considerable information can be lost in transforming to spacings.

4. MONTE CARLO RESULTS

The goal of this section is to indicate empirically some values of n, k, and $p = 1 - c/n$ that can be used in practical situations. Initially, the distributions studied were normal, t distributions with 3 and 8 degrees of freedom (t_3 and t_8), and chi-squared distributions with 1, 4, and 8 degrees of freedom (χ_1^2, χ_4^2, and χ_8^2). These distributions were motivated by the fact that many statistics of interest are approximately normal or chi-squared distributed. A secondary motivation for their use was the availability of the location swindle (see Gross 1973) to reduce Monte Carlo variance. Later it was decided to add two Weibull distributions, $F(x) = 1 - \exp(-x^b)$ with $b = 2$ and $b = 4$, and two lognormals,

$$f(x) = (2\pi\sigma^2 x^2)^{-1/2} \exp(-(\ln x)^2/2\sigma^2)$$

with $\sigma = \frac{1}{2}$ and $\sigma = 1$. Sample sizes studied were 50, 100, 500, 1000, and 5000. Only sample size $n = 500$, and the normal, t_3, t_8, χ_4^2, Weibull ($b = 4$), and the lognormals will be presented here as they are fairly representative of the full study. In these situations the Monte Carlo replication size was $N = 5000$ and all random variables except the Weibull were generated from normal deviates using either the Super-Duper generator of Marsaglia, Anathanarayanan, and Paul (1976) or Marsaglia's modified polar method. The Weibulls were generated from standard exponentials, which were in turn generated using a uniform generator of Schrage (1979).

In order to compare $\hat{\eta}_p$ with "standard" methods, five raw percentile estimators were computed. The first four are listed as options in the SAS routine UNIVARIATE (see Chilko 1979, p. 429) and the fifth estimator (called AV) is motivated by the usual definition of the sample median. If $X_{(1)} \leq \cdots \leq X_{(n)}$ are the order statistics of the sample in ascending order and $[\cdot]$ is the greatest integer function, then this last estimator is

$$\begin{aligned} \mathrm{AV} &= \tfrac{1}{2}(X_{(np)} + X_{(np+1)}) && \text{if } np \text{ is an integer} \\ &= X_{([np]+1)} && \text{otherwise.} \end{aligned}$$

The best of the SAS methods was the default option on UNIVARIATE which we shall call LINT for linear interpolation. LINT is actually obtained by linear inverse interpolation of the empirical distribution function and defined by

$$\mathrm{LINT} = (1 - \varepsilon)X_{([np])} + \varepsilon X_{([np]+1)},$$

where $\varepsilon = np - [np]$. AV performed well in terms of bias. However, when $np \neq [np]$, LINT outperformed AV in terms of MSE. When $np = [np]$, no clear winner emerged. Other "raw" percentile methods such as k-point inverse interpolation or methods based on nonparametric density estimators (see Azzalini 1981) may give small improvements over LINT and AV.

Table 1 lists the ratio of estimates of the MSE of $\hat{\eta}_p$ to that of LINT for the normal and Weibull ($b = 4$). The estimates of MSE individually have relative standard error (s.e./MSE) in the range .01 to .025. The last row of the table is a measure of tail heaviness with respect to the standard exponential distribution introduced by Breiman, Stone, and Gins (1979). It has the form $H(p) = -l''(\eta_p)/[l'(\eta_p)]^2$, where $l(x) = -\ln [1 - F(x)]$. $H(p)$ is scale invariant and is zero at the exponential, negative for lighter than exponential tails, and positive for heavier than exponential tails. This measure shows the normal distribution to have lighter than exponential tails as one would expect since $1 - \Phi(x) \sim \exp(-x^2)/x$ for x large. Likewise, the Weibull distribution ($b = 4$) has even lighter tails since $1 - F(x) = \exp(-x^4)$. It is not as easy to anticipate the values of $H(p)$ for the t distributions and lognormals in Tables 2 and 3.

For the normal distribution we can discern a pattern of good results in Table 1 along diagonals moving from lower left to upper right. The same holds true for the Weibull ($b = 4$). It appears that the optimal strategy in balancing bias and variance in these situations requires $k/c \simeq 4$. This result also holds true at $n = 1000$ and $n = 5000$, but bias is a larger factor there at $p = .95$ and .975. Note that when $k/c \leq 1$, $\hat{\eta}_p$ is uniformly out-performed by LINT. This is to be expected since $k/c \leq 1$ is the case where one extrapolates $\hat{\eta}_p$ from order statistics to the right of it. Three right censored cases at $n = 500$ are included (see the Appendix for the definition of $\hat{\eta}_p$ in these cases). The effects at $p = .95$ and $p = .975$ are relatively small, but the effects are somewhat disturbing at $p = .99$ and $p = .995$.

Table 2 gives results for the χ^2_4, t_8, and lognormal ($\sigma = \frac{1}{2}$) distributions. Here, the tail heaviness measure $H(p)$ shows all three of these distributions to be approximately exponential. Note though that no t distribution satisfies (3.1) (see Galambos 1978, Theorem 2.4.3 (iii)). The χ^2_4 results are similar to those for χ^2_1 amd χ^2_8 and should be generally representative of gamma distributions. Since all three distributions in Table 2 are approximately exponential at the η_p studied, bias plays a smaller role and k can be chosen fairly

Table 1. *Lighter Than Exponential Tails: Ratio of MSE of $\hat{\eta}_p$ to MSE of LINT for Normal and Weibull ($b = 4$) at $n = 500$*

	Normal				Weibull ($b = 4$)			
k \ p	.95	.975	.99	.995	.95	.975	.99	.995
c	25	12.5	5	2.5	25	12.5	5	2.5
10	4.92[a]	1.44[a]	.76	.80[b]	5.00[a]	1.40	.74	.80[b]
20	1.38[a]	.80	.76[b]	.78	1.40[a]	.82	.75[b]	.78
50	.84	.78[b]	.79	1.01	.93	.82[b]	.81	1.23
100	.82[b]	.86	1.89	2.96	.88[b]	.95	2.79	4.91
200	1.63	5.81	5.81	15.5	2.19	9.51	20.7	28.3
				Type II Censoring on the Right				
20−5[c]	1.38[a]	.83	.95	1.11	1.41[a]	.83	.94	1.17
50−5[c]	.80	.81	1.03	1.39	.85	.82	1.13	1.81
100−10[c]	.79	1.33	2.96	4.33	.79	1.67	4.58	7.32
Tail Heaviness $H(p)$	−.20	−.16	−.13	−.11	−.25	−.20	−.16	−.14

[a] $k/c \leq 1$.
[b] $k/c = 4$.
[c] # censored.

Table 2. *Approximately Exponential Tails: Ratio of MSE of $\hat{\eta}_p$ to MSE of LINT for χ^2_4, t_8, and Lognormal ($\sigma = \frac{1}{2}$) at $n = 500$*

	χ^2_4			t_8			Lognormal ($\sigma = \frac{1}{2}$)		
k \ p	.95	.99	.995	.95	.99	.995	.95	.99	.995
c	25	5	2.5	25	5	2.5	25	5	2.5
20	1.38[a]	.75	.76	1.35[a]	.80	.79	1.40[a]	.77	.76
50	.73	.66	.59	.76	.67	.56	.78	.63	.56
100	.72	.56	.49	.75	.59	.47	.77	.48	.41
200	.71	.73	.71	1.14	1.88	1.60	.67	.32	.26
				Type II Censoring on the Right					
20−5[b]	1.41[a]	.92	.99	1.38[a]	.91	.93	1.43[a]	.90	.94
50−5[b]	.76	.73	.66	.78	.70	.59	.80	.68	.61
100−10[b]	.78	.65	.57	.78	.64	.51	.79	.53	.46
Tail Heaviness $H(p)$	−.04	−.02	−.02	−.05	.04	.06	.04	.06	.06

[a] $k/c \leq 1$.
[b] # censored.

large and independent of c. However, after $k = 100$ for χ^2_4 we see that bias begins to have an effect so that $k = 100$ is preferred to $k = 200$. The t_8 distribution behaves similarly but at the lognormal ($\sigma = \frac{1}{2}$), $k = 200$ is still preferred over $k = 100$. An optimal strategy to cover all three distributions in Table 2 suggests that $k/n \simeq .2$ for $n = 500$. For larger sample sizes up to 5000, $k/n \simeq .1$ seems a good compromise. Censoring the five largest observations has considerable effect at $k = 20$ and less so at $k = 50$. It seems clear that the largest observations are important for estimating a_n and η_p. If robustness with regard to possible outliers is desired, then the number trimmed should be quite small or one should stay with LINT.

Table 3 lists results for the t_3 and lognormal ($\sigma = 1$) distributions. These distributions have tails that are considerably heavier than exponential. The results are not encouraging, although there may be some hope around $k/c \simeq 8 - 10$. In several cases the right censored estimators performed better but not uniformly better. I would be generally cautious about using $\hat{\eta}_p$ in such heavy tailed situations.

In practice I suggest a Q-Q plot of X_{in} versus $-\ln(i/(n+1))$ for at least the upper 20% of the order statistics. A plot that is straight suggests approximately exponential tails, whereas a convex upward (downward) plot suggests heavier (lighter) than exponential tails. Then subject to $k > c$, $k \geq 10$, and #

Table 3. *Heavier than Exponential Tails: Ratio of MSE of $\hat{\eta}_p$ to MSE of LINT for t_3 and Lognormal ($\sigma = 1$) at $n = 500$*

	t_3				Lognormal ($\sigma = 1$)			
k \ p	.95	.975	.99	.995	.95	.975	.99	.995
c	25	12.5	5	2.5	25	12.5	5	2.5
20	1.62[a]	1.56	1.58	1.12	1.60[a]	1.26	1.17	.86
50	1.90	1.71	.87	.67	1.69	1.30	.73	.65
100	1.83	1.04	.60	.70	1.59	.79	.81	1.02
200	1.32	.59	.46	.69	.77	.95	1.74	1.98
				Type II Censoring on the Right				
20−5[b]	1.45[a]	1.00	.87	.81	1.52[a]	.96	.89	.83
50−5[b]	1.11	.91	.70	.78	1.10	.85	.77	.87
100−10[b]	.88	.69	.83	1.06	.83	.84	1.35	1.57
Tail Heaviness $H(p)$	.21	.26	.30	.31	.28	.27	.25	.24

[a] $k/c \leq 1$.
[b] # censored.

censored/$k \le .1$, my recommendations are as follows:

1. Lighter than exponential tails; examples studied: normal, Weibull, $b = 2$ and $b = 4$. Use $k/c = 4$ in the range

$$50 \le n \le 500 \quad \text{and} \quad p \ge .95$$

and

$$500 < n \le 5000 \quad \text{and} \quad p \ge .99.$$

2. Approximately exponential tails; examples studied: t_8, χ_1^2, χ_4^2, χ_8^2, lognormal ($\sigma = \frac{1}{2}$). Use

$$k/n = .2 \text{ for } \quad 50 \le n \le 500 \quad \text{and} \quad p \ge .95$$

and

$$k/n = .1 \text{ for } \quad 500 < n \le 5000 \quad \text{and} \quad p \ge .95.$$

3. Heavier than exponential tails; examples studied: t_3, lognormal ($\sigma = 1$). Use a raw percentile estimator such as LINT.

5. NUMERICAL EXAMPLE

The basic situation has been described in Section 2. The specific example considered here is Monte Carlo estimation of the percentiles of $T = \min_\mu d(F_m, F_\mu)$, where F_m is the empirical distribution function of $m = 50$ standard logistic random variables, $F_\mu(x) = [1 + \exp(-(x - \mu))]^{-1}$, and $d(\cdot, \cdot)$ is the Anderson-Darling distance (see Boos 1982 for details). In this particular case the percentiles of T have been tabulated in Stephens (1979, Table 1) and appear as the first row of Table 4. The next four rows of Table 4 are the new estimators $\hat{\eta}_p$ using $k = 50$, 100, 150, and 200 (recall from Section 2 that the replication size is $n = 1{,}000$). Five observations were outside the reporting range [0, 2] and thus were unavailable. However, $\hat{\eta}_p$ based on Type II censoring was used because the gap between the 995th observation and the upper limit 2 was larger than expected, which suggested that the five largest observations would have been trimmed even if they were available. The last row of Table 4 is the raw percentile estimator LINT, which in these four cases is just $X_{(np)}$ since np is an integer. At $p = .95$, all three $\hat{\eta}_p$ with $k/c > 1$ are an improvement over LINT. At $p = .975$ and $p = .995$, all four $\hat{\eta}_p$ are better than LINT. And at $p = .99$, all but $k = 200$ are an improvement over LINT.

The theory in Boos (1982) suggests that T has approximately the same distribution as an infinite sum of chi-squared random variables. Thus it is natural to expect that T has approximately exponential tails and that $k = 100$ is a suitable choice. The Q-Q plot of the upper 20% of the order statistics is roughly straight except for the last four available observations 992–995. If those four are deleted, then at $k = 100$, $\hat{a}_n = .268$ and $\hat{\eta}_p$ becomes 1.036, 1.222, 1.467, and 1.653 respectively for the p given in Table 4. This would still be an improvement over LINT but not as dramatic as for the $k = 100$ row in Table 4. This result demonstrates the importance of the large observations in estimating a_n and thus η_p.

Using results in the Appendix we find that for $k = 100$, $r = 6$, and $c = 10$ we have $EW = -.0440$, var $W = .0705$, $\sqrt{\beta_1} = -.4119$, and $\beta_2 = 3.3249$. From a table of Pearson Curve percentiles (Bouver and Bargmann 1974), the 5th percentile of W is $-.509$ and the 95th percentile is .359. Thus using $\hat{a}_n = .281$, a 90% confidence interval for $\eta_{.99}$ is given by $(1.497 - \hat{a}_n\, .359,\ 1.497 + \hat{a}_n\, .510) = (1.396, 1.640)$. The approximate method (A.4) yields (1.381, 1.613). The usual distribution-free method based on order statistics (see David 1981, Sec. 2.5) yields $(X_{(985)}, X_{(995)}) = (1.349, 1.801)$, which again illustrates the improved precision of $\hat{\eta}_p$.

6. SUMMARY

Weissman's percentile estimators, which are derived from extreme value theory and are based on the k largest order statistics, can have smaller MSE's than conventional "raw" percentile estimators. Thus they have potential application in estimating percentiles of intractable test or pivotal statistics by Monte Carlo methods. Empirical results given in this article provide guidance on when Weissman's estimators are useful and on how to choose k.

APPENDIX

This section adds some minor modifications for censoring on the right and then shows how to con-

Table 4. *Estimates of the Percentiles of T*

p	.95	.975	.99	.995	$\hat{a}_n$	X_{kn}
True Percentiles	1.043	1.238	1.502	1.707		
$k = 50$	1.061	1.249	1.498	1.686	.271	1.061
$k = 100$	1.045	1.240	1.497	1.692	.281	.851
$k = 150$	1.047	1.235	1.484	1.672	.272	.748
$k = 200$	1.044	1.227	1.470	1.654	.265	.676
LINT	1.057	1.206	1.471	1.801		

struct approximate confidence intervals for η_p. Here it is easiest to work with the limiting random variable. Therefore, suppose that $(X_1, \ldots, X_k)$ is an extremal variate with location and scale parameters μ and σ, that is, having density $\psi_k((x_1 - \mu)/, \ldots, (x_k - \mu)/\sigma)/\sigma^k$. To translate back to the practical situation, just let $(X_1, \ldots, X_k)$ be replaced by $(X_{1n}, \ldots, X_{kn})$ and (μ, σ) by (b_n, a_n).

Type I censoring. Suppose that we can only observe those values of $(X_1, \ldots, X_k)$ that are $\leq x_0$. Then, if $X_{r-1} > x_0$ and $X_r \leq x_0$, the likelihood is

$$L = \frac{\exp\left\{-\exp\left[-\left(\frac{X_k - \mu}{\sigma}\right)\right] - \sum_{i=r}^{k}\left(\frac{X_i - \mu}{\sigma}\right) - (r-1)x_0\right\}}{\sigma^{k-r+1}(r-1)!} \tag{A.1}$$

for $X_0 \geq X_r \geq X_k$. Solving the likelihood equations gives

$$\hat{\sigma} = \left(\sum_{i=r}^{k} X_i + (r-1)x_0 - kX_k\right)\Big/(k - r + 1),$$

$$\hat{\mu} = \hat{\sigma} \ln k + X_k. \tag{A.2}$$

Type II Censoring. Here $(X_1, \ldots, X_{r-1})$ are again unavailable but r is not random. The likelihood is just the marginal density of the remaining random variables and is the same as (A.1) except that x_0 is replaced by X_r.

Confidence Intervals. The following results apply only to uncensored and Type II censored data. Weissman (1978, Theorem 3) showed that the spacings $D_i = (X_i - X_{i+1})/\sigma$ are independent exponentials with mean i^{-1} and are also independent of X_k. Writing $X_i/\sigma = \sum_{j=1}^{k} D_j$, one can easily verify that $\hat{\sigma} = \sigma T_{k-r}/(k - r + 1)$, where T_{k-r} is a standard gamma random variable with parameter $k - r$ that is independent of X_k. In this notation Weissman's estimator is $\hat{\eta}_p = \hat{\sigma} \ln (k/c) + X_k$ (recall that $c = n(1 - p)$ when using a sample of size n). An appropriate pivot for $\eta_p \simeq -\sigma \ln c + \mu$ is

$$W = (\hat{\eta}_p - (-\sigma \ln c + \mu))/\hat{\sigma}$$

$$= \ln (k/c) + \left\{\frac{X_k - \mu}{\sigma} + \ln c\right\}(k - r + 1)/T_{k-r}.$$

For the case $r = 1$, $c = 1$, and $2 \leq k \leq 30$, the percentiles of W may be obtained from a table in Weissman (1978). In general, I suggest Pearson curve approximations to the distribution of W (see Solomon and Stephens 1978) since the moments of W are easy to calculate using the independence of X_k and T_{k-r}. The first four moments of $(X_k - \mu)/\sigma$ are given by

$$a_1 = \gamma - S_{1k},$$

$$a_2 = S_{2\infty} - S_{2k} + a_1^2,$$

$$a_3 = 2(S_{3\infty} - S_{3k}) + 3a_1a_2 - 2a_1^3,$$

and

$$a_4 = 6(S_{4\infty} - S_{4k}) + 4a_1a_3 - 12a_1^2a_2 + 3a_2^2 + 6a_1^4,$$

where $\gamma = .5772\cdots$ is Euler's constant and $S_{ik} = \sum_{j=1}^{k-1} j^{-i}$. Since T_{k-r} is a gamma,

$$E(T_{k-r})^{-l} = [(k - r - 1)(k - r - 2)\cdots(k - r - l)]^{-1}.$$

After obtaining the first four moments of W, the approximate percentiles W_α and $W_{1-\alpha}$ are found in a table of Pearson curve percentiles (e.g., Bouver and Bargmann 1974). Then

$$(\hat{\eta}_p - \hat{\sigma}W_{1-\alpha}, \hat{\eta}_p - \hat{\sigma}W_\alpha) \tag{A.3}$$

is an approximate $(1 - 2\alpha) \times 100\%$ confidence interval for η_p. When k is large, normal percentiles can be used since the distribution of W tends to a normal distribution. This can be seen as follows. Under Equation (3.1), $(X_k - \mu)/\sigma$ converges in distribution to $-\ln T_k$ for every fixed k as $n \to \infty$ (see (9.4.3) of David 1981). For large k, T_k and thus $-\ln T_k$ is approaching normality and $(k - r + 1)/T_{k-r}$ tends almost surely to 1, yielding the asymptotic normality of W. Since

Table 5. *95% Confidence Intervals from (A.3) and (A.4) Standardized by $\hat{\eta}_p = 0$ and $\hat{\sigma} = 1$*

k	c	EW	$\sqrt{\beta_1}$	β_2	*From (A.3)*	*From (A.4)*	*From (A.4) With Mean Adjustment*
50	10	−.06	−.56	3.62	(−.37, .56)	(−.44, .44)	(−.38, .50)
100	10	−.04	−.40	3.31	(−.35, .49)	(−.41, .41)	(−.37, .45)
	25	−.02	−.37	3.28	(−.25, .33)	(−.28, .28)	(−.26, .30)
150	10	−.03	−.33	3.20	(−.34, .45)	(−.39, .39)	(−.35, .42)
	25	−.02	−.32	3.19	(−.24, .32)	(−.28, .28)	(−.25, .30)
	50	−.01	−.28	3.16	(−.18, .22)	(−.20, .20)	(−.19, .21)
200	10	−.03	−.28	3.15	(−.33, .42)	(−.37, .37)	(−.34, .40)
	25	−.02	−.28	3.15	(−.24, .30)	(−.27, .27)	(−.25, .29)
	50	−.01	−.26	3.13	(−.18, .22)	(−.20, .20)	(−.19, .21)

NOTE: EW, $\sqrt{\beta_1}$, and β_2 are the mean, skewness, and kurtosis values of W.

$\gamma - S_{1k} \simeq -\ln k$ and $S_{2\infty} - S_{2k} \simeq k^{-1}$, we have for $k \gg r$

$$EW = \ln(k/c) + [\gamma - S_{1k} + \ln c]\left[\frac{k-r+1}{k-r-1}\right] \simeq 0$$

and

$$\begin{aligned}\operatorname{var} W &= \left[S_{2\infty} - S_{2k} + \frac{[\gamma - S_{1k} + \ln c]^2}{k-r-1}\right] \\ &\quad \times \left[\frac{(k-r+1)^2}{(k-r-1)(k-r-2)}\right] \\ &\simeq [1 + (\ln(k/c))^2]/k.\end{aligned}$$

Thus, for large k an approximate $(1 - 2\alpha) \times 100\%$ confidence interval for η_p is given by

$$\hat{\eta}_p \pm \hat{\sigma}\left(\frac{1 + (\ln(k/c))^2}{k}\right)^{1/2} z_{1-2\alpha}, \qquad \text{(A.4)}$$

where z_α is the αth quantile of the standard normal. Table 5 shows that the length of confidence intervals based on (A.3) and (A.4) is about the same. However, even at $k = 200$, (A.3) has an important skewed aspect that (A.4) fails to capture. The last column of Table 5 shows that using the true mean of W rather than $EW \simeq 0$ regains part of that skewness.

[Received January 1982. Revised June 1983.]

REFERENCES

AZZALINI, A. (1981), "A Note on the Estimation of a Distribution Function and Quantiles by a Kernel Method," *Biometrika*, 68, 326-328.

BOOS, D. D. (1981), "On Weissman's Method of Estimating Large Percentiles," Institute of Statistics Mimeo Series 1360, North Carolina State University.

——— (1982), "Minimum Anderson-Darling Estimation," *Communications in Statistics—Theory and Methods*, 11, 2747–2774.

BOUVER, H., and BERGMANN, R. E., (1974), "Tables of the Standardized Percentage Points of the Pearson System of Curves in Terms of β_1 and β_2," Technical Report No. 107, University of Georgia, Dept. of Statistics and Computer Science.

BREIMAN, L., STONE, C. J., and GINS, J. D. (1979), "New Methods for Estimating Tail Probabilities and Extreme Value Distributions," TSC-PD-A226-1, Santa Monica, Calif.: Technology Service Corporation.

——— (1981), "Further Development of New Methods for Estimating Tail Probabilities and Extreme Value Distributions," TSC-PD-A243-1, Santa Monica, Calif.: Technology Service Corporation.

CHILKO, D. M. (1979), "Univariate Procedure," in *SAS User's Guide, 1979 Edition*, eds. J. T. Helwig and K. A. Council, SAS Institute, Inc.

DAVID, H. A. (1981), *Order Statistics*, 2nd ed., New York: John Wiley.

DICKEY, D. A. (1981), "Histograms, Percentiles, and Moments," *The American Statistician*, 35, 164–165.

GALAMBOS, J. (1978), *The Asymptotic Theory of Extreme Order Statistics*, New York: John Wiley.

GROSS, A. M. (1973), "A Monte Carlo Swindle for Estimators of Location," *Applied Statistics*, 22, 347–353.

MARSAGLIA, G., ANATHANARAYANAN, K., and PAUL, N. J. (1976), "Improvements on Fast Methods for Generating Normal Random Variables," *Information Processing Letters*, 5, 27–30.

SCHRAGE, L. (1979), "A More Portable Fortran Random Number Generator," *ACM Transactions on Mathematics Software*, 5, 132–138.

SOLOMON, H., and STEPHENS, M. A. (1978), "Approximations to Density Functions Using Pearson Curves," *Journal of the American Statistical Association*, 73, 153–160.

STEPHENS, M. A. (1979), "Tests of Fit for the Logistic Distribution Based on the Empirical Distribution Function," *Biometrika*, 66, 591–595.

WEISSMAN, I. (1978), "Estimation of Parameters and Large Quantiles Based on the k Largest Observations, *Journal of the American Statistical Association*, 73, 812–815.

——— (1980), "Estimation of Tail Parameters Under Type I Censoring," *Communications in Statistics*, A9, 1165–1175.