*Biometrika* (1982), **69**, 3, *pp.* 635–40
*Printed in Great Britain*



# A new distribution-free quantile estimator

BY FRANK E. HARRELL
*Clinical Biostatistics, Duke University Medical Center, Durham, North Carolina, U.S.A.*

AND C. E. DAVIS
*Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina, U.S.A.*

## SUMMARY

A new distribution-free estimator $Q_p$ of the $p$th population quantile is formulated, where $Q_p$ is a linear combination of order statistics admitting a jackknife variance estimator having excellent properties. The small sample efficiency of $Q_p$ is studied under a variety of light and heavy-tailed symmetric and asymmetric distributions. For the distributions and values of $p$ studied, $Q_p$ is generally substantially more efficient than the traditional estimator based on one or two order statistics.

*Some key words*: Distribution-free estimator; Nonparametric estimator; Order statistic; Percentile; Quantile.

## 1. INTRODUCTION

The estimation of population quantiles or percentiles is of great interest, particularly when the statistician is unwilling to assume a parametric form for the distribution or even to assume the distribution to be symmetric. Sample quantiles have many desirable properties. However, they also have drawbacks. They are not particularly efficient estimators of location for distributions such as the normal, good estimators of the variance of sample quantiles do not exist for general distributions, sample quantiles may not be jackknifed, and the sample median differs in form and in efficiency depending on the sample size being even or odd. Maritz & Jarrett (1978) have developed an estimator of the sample median that performs well for some distributions.

We propose to estimate the $p$th quantile by a linear combination of the order statistics. For most distributions the new estimator offers a significant gain in efficiency over the traditional one and admits a jackknife variance estimator that performs well. The properties of the estimator and its variance estimator are studied over a wide variety of light- and heavy-tailed symmetric and asymmetric distributions, with emphasis on small sample results.

## 2. ESTIMATORS

Let $X_1, \ldots, X_n$ denote a random sample of size $n$ from a continuous distribution having distribution function $F(.)$. Let $X_{(1)} \leqslant \ldots \leqslant X_{(n)}$ denote the order statistics of the sample and $X = (X_{(1)}, \ldots, X_{(n)})$. One traditional estimator of the $p$th population quantile $F^{-1}(p)$ is

$$T_p = (1-g)\, X_{(j)} + g X_{(j+1)}, \tag{1}$$

where $(n+1)p = j+g$ and $j$ is the integral part of $(n+1)p$. When $p = \frac{1}{2}$, $T_p$ is the usual sample median.

The expected value of the $k$th order statistic is given by

$$E(X_{(k)}) = \frac{1}{\beta(k, n-k+1)} \int_{-\infty}^{\infty} xF(x)^{k-1}\{1-F(x)\}^{n-k}\, dF(x)$$

$$= \frac{1}{\beta(k, n-k+1)} \int_0^1 F^{-1}(y)\, y^{k-1}(1-y)^{n-k}\, dy.$$

Since $E(X_{\{(n+1)p\}})$ converges to $F^{-1}(p)$ for $p \in (0,1)$, we take as our estimator of $F^{-1}(p)$ something which estimates $E(X_{\{(n+1)p\}})$ whether or not $(n+1))p$ is an integer, namely

$$Q_p = \frac{1}{\beta\{(n+1)p, (n+1)(1-p)\}} \int_0^1 F_n^{-1}(y)\, y^{(n+1)p-1}(1-y)^{(n+1)(1-p)-1}\, dy,$$

where $F_n(X)$ is the sample distribution function, $F_n(x) = n^{-1}\Sigma I(X_i \leqslant x)$, $I(A)$ being the indicator function of the set $A$. The estimator can be reexpressed as

$$Q_p = \sum_{i=1}^{n} W_{n,i} X_{(i)}, \tag{2}$$

where

$$W_{n,i} = \frac{1}{\beta\{(n+1)p, (n+1)(1-p)\}} \int_{(i-1)/n}^{i/n} y^{(n+1)p-1}(1-y)^{(n+1)(1-p)-1}\, dy$$

$$= I_{i/n}\{p(n+1), (1-p)(n+1)\} - I_{(i-1)/n}\{p(n+1), (1-p)(n+1)\} \tag{3}$$

and $I_x(a,b)$ denotes the incomplete beta function. Maritz & Jarrett (1978) used a similar idea to estimate the second moment of the sample median.

For $p = \frac{1}{2}$ or $n \geqslant 100$ with $p \neq \frac{1}{2}$, the weights $W_{n,i}$ in (3) can be adequately calculated with numerical integration using Simpson's rule with 2 intervals between $(i-1)/n$ and $i/n$. For other cases, the incomplete beta function should be calculated exactly. The algorithm of Majumdar & Bhattacharjee (1973) is very efficient for this calculation.

Following David (1981, p. 273), $Q_p$ is asymptotically normally distributed under mild assumptions on $F(.)$. Monte Carlo studies using the Kolmogorov–Smirnov statistic have shown that for the uniform and normal distributions, the normal approximation is adequate for samples as small as 20 for $p = \frac{1}{2}$ or 30–50 for $p = 0{\cdot}95$. For asymmetric distributions such as the exponential, sample sizes as large as 80–100 may be required for $p = 0{\cdot}9$ or above. For the calculation of tail probabilities using the variance estimator given below and a normal approximation, sample sizes necessary for accurate confidence intervals may be smaller. However, more research is needed in this area.

In order to calculate the jackknife variance estimate of $Q_p$ (Miller, 1974), consider removing order statistic $j$ from the sample. The resulting estimate is

$$S_j = s_j^{\mathrm{T}} X, \quad (s_j)_i = I(i \neq j)\, W_{n-1, i-I(i>j)}.$$

The jackknife variance estimator is

$$V_p = \frac{n-1}{n} \sum_{j=1}^{n} (S_j - \bar{S})^2 = (n-1)(n^{-1}\Sigma S_j^2 - \bar{S}^2), \tag{4}$$

where

$$\bar{S} = n^{-1} \sum_{j=1}^{n} S_j = n^{-1} u^{\mathrm{T}} X, \quad (u)_i = (i-1)\, W_{n-1,i-1} + (n-i)\, W_{n-1,i},$$

$$\Sigma S_j^2 = X^{\mathrm{T}} \Lambda X, \quad \Lambda_{ll} = (l-1)\, W_{n-1,l-1}^2 + (n-l)\, W_{n-1,l}^2,$$

$$\Lambda_{lm} = (l-1)\, W_{n-1,l-1}\, W_{n-1,m-1} + (m-l-1)\, W_{n-1,l}\, W_{n-1,m-1} + (n-m)\, W_{n-1,l}\, W_{n-1,m} \quad (m > l),$$

$$\Lambda_{ml} = \Lambda_{lm}, \quad W_{n-1,i} \equiv 0 \quad (i < 1 \text{ or } i > n-1).$$

Simulations show that the jackknife version of $Q_p$, while having lower bias than $Q_p$, has larger variance, resulting in an estimator with similar efficiency to $T_p$. The extreme order statistic weights for the jackknife estimator for small $n$ and $p = \frac{1}{2}$ are sometimes negative, resulting in nearly unbiased estimators of extreme quantiles although having large variance. The jackknifed quantile estimator will not be discussed further, only $V_p$, the associated variance estimator.

Kaigh & Lachenbruch (1982) have proposed a quantile estimator which is the average of subsample $T_p$-like estimators. Their estimator has properties similar to $Q_p$ and does not require numerical integration. It may require larger sample sizes for estimating extreme quantiles, and a variance estimator has not been studied.

## 3. Efficiency of $Q_p$ relative to $T_p$ for various distributions

To investigate the performance of $Q_p$ with respect to $T_p$ for a wide variety of distributions, the generalized lambda distribution (Ramberg *et al.*, 1979) was considered. The distribution is defined by

$$F^{-1}(p) = \mu + \sigma\{p^a - (1-p)^b\},$$

where $\mu$ and $\sigma$ are respectively location and scale parameters, set to 0 and 1 for this investigation, and $a$ and $b$ are shape parameters. Table 1 shows the distributions used with the standardized skewness, $\alpha_3$, and kurtosis, $\alpha_4$, values.

Table 1. *Generalized lambda distributions*

| $a$ | $b$ | $\alpha_3$ | $\alpha_4$ | Description |
|---|---|---|---|---|
| 1 | 1 | 0 | 1·8 | Light-tailed symmetric |
| 0·1349 | 0·1349 | 0 | 3 | Normal-like |
| −0·1359 | −0·1359 | 0 | 9 | Very heavy-tailed symmetric, like $t$ distribution with 5 degrees of freedom |
| −1 | −1 | $\infty$ | $\infty$ | Cauchy-like |
| 0·0251 | 0·0953 | 0·9 | 4·2 | Medium-tailed asymmetric |
| 0 | 0·0004 | 2 | 9 | Exponential-like |

The mean squared error of an estimator was used to measure its efficiency, for example

$$\mathrm{MSE}(Q_p) = E\{Q_p - F^{-1}(p)\}^2, \quad \mathrm{MSE}(T_p) = E\{T_p - F^{-1}(p)\}^2.$$

These mean squared errors were estimated by generating 1000 random samples for each distribution and each $n$ and averaging the squared errors. Random order statistics were generated using the method of Lurie & Hartley (1972) incorporating a Tausworthe uniform random number generator (Whittlesey, 1968).

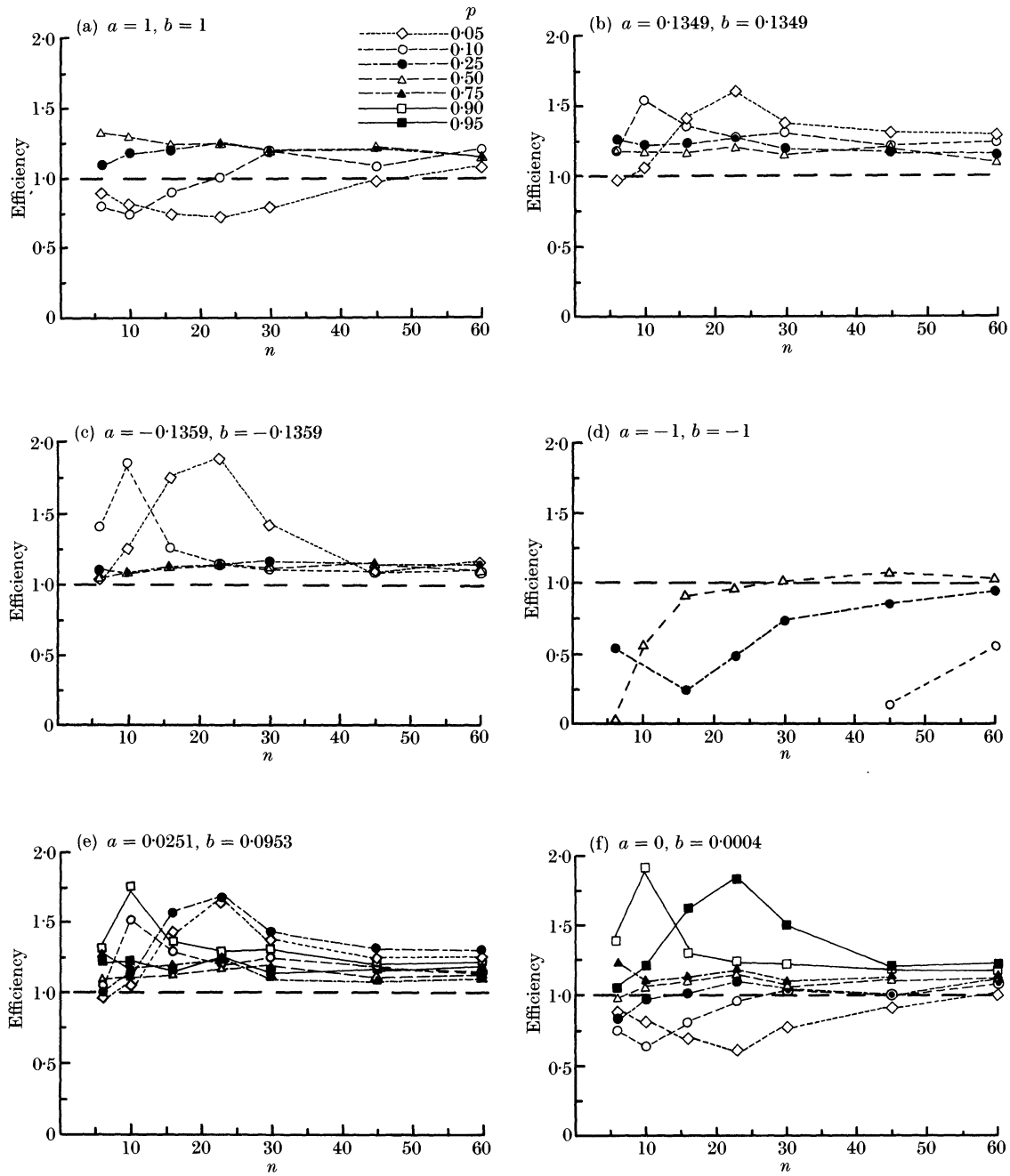


Fig. 1. Simulated mean squared error efficiency of $Q_p$ against $T_p$ for generalized lambda distributions with different parameters of the distribution.

The efficiency of $Q_p$ relative to $T_p$ is $\text{MSE}(T_p)/\text{MSE}(Q_p)$. This was estimated by taking the ratio of estimated mean squared errors. Monte Carlo experiments were performed for $n = 6, 10, 16, 23, 45, 60$ and $p = 0{\cdot}05, 0{\cdot}10, 0{\cdot}25, 0{\cdot}50, 0{\cdot}75, 0{\cdot}90, 0{\cdot}95$. Results for $p > \frac{1}{2}$ are not displayed for symmetric distributions. The results are shown in Fig. 1. One additional experiment was performed for the normal distribution for $n = 250$ and $p = \frac{1}{2}$ resulting in an estimated relative effiency of $Q_p$ of 1·07. From Fig. 1 we see that, except for the Cauchy-like distribution, $Q_p$ is generally more than 1·1 times as efficient as $T_p$.

## 4. Exact efficiency of $Q_p$ and $T_p$ relative to parametric estimators for the normal distribution

For the normal distribution, the uniform minimum variance unbiased estimator of $F^{-1}(p)$ is $\hat{X}_p = \bar{X} + \Phi^{-1}(p)\, s/E(s)$, where $\bar{X}$ and $s$ are the sample mean and standard deviation respectively, $\Phi(\,.\,)$ is the normal distribution function, and

$$E(s) = \{2/(n-1)\}^{1/2}\, \Gamma(\tfrac{1}{2}n)/\Gamma\{\tfrac{1}{2}(n-1)\}.$$

For $2 \leqslant n \leqslant 20$, exact efficiencies of $Q_p$ and $T_p$ can be calculated using the moments of normal order statistics tabled by Sarhan & Greenberg (1962, p. 193). For $p = 0{\cdot}1$ and $p = 0{\cdot}5$, the efficiencies are shown in Fig. 2.

We do not recommend the use of $Q_p$ for small $n$ and extreme $p$; however, the relative performance of $\hat{X}_p$ is actually worse in this situation than for larger $n$. From Fig. 2 we see that the new estimator has much to offer over $T_p$ especially for extreme quantiles.

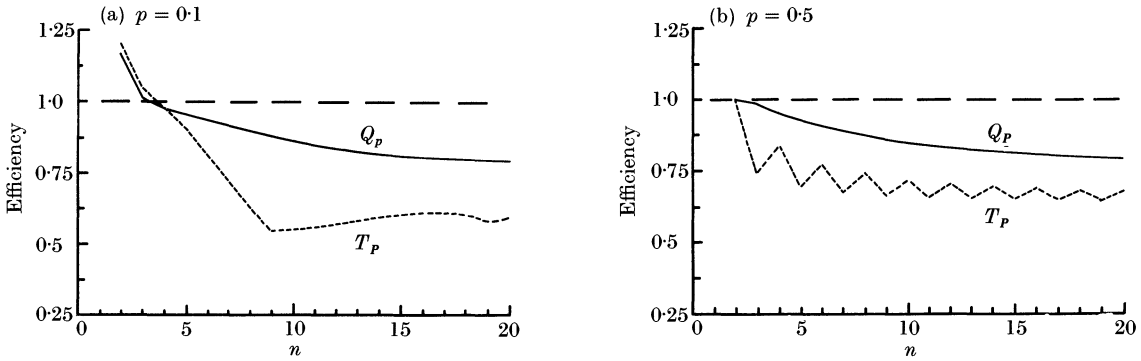


Fig. 2. Exact mean squared error efficiency of new estimator, $Q_p$, and sample quantile $T_p$, with respect to $\hat{X}_p$ for the normal distribution with quantiles 0·1 and 0·5.

## 5. Performance of the variance estimator $V_p$

For the 1000 repeated samples in the simulations for each distribution and sample size $n$, the sample variance of the simulated $Q_p$ estimator was calculated. This was compared to the sample mean of simulated $V_p$ estimates. The ratio of the mean $V_p$ to the simulated $V(Q_p)$ was used to estimate $E(V_p)/V(Q_p)$ to measure the bias of $V_p$. The results indicated that $E(V_p)$ was seldom different from $V(Q_p)$ by more than a factor of 1·15, even for the Cauchy-like distribution with $n > 16$. Thus confidence intervals for $Q_p$ can be readily constructed using the asymptotic normality of $Q_p$.

The authors are grateful to the referee for constructive comments that improved the clarity of the paper.

## References

David, H. A. (1981). *Order Statistics*, 2nd edition. New York: Wiley.

Kaigh, W. D. & Lachenbruch, P. A. (1982). A generalized quantile estimator. *Comm. Statist.* A **11**. To appear.

Lurie, D. & Hartley, H. O. (1972). Machine-generation of order statistics for Monte Carlo computations. *Am. Statistician* **26**, 26–7. Errata (1972) **26**, 56—57.

Majumdar, K. L. & Bhattacharjee, G. P. (1973). The incomplete beta integral (Algorithm as 63). *Appl. Statist.* **22**, 409–11.

MARITZ, J. S. & JARRETT, R. G. (1978). A note on estimating the variance of the sample median. *J. Am. Statist. Assoc.* **73**, 194–6.

MILLER, R. (1974). The jackknife—a review. *Biometrika* **61**, 1–15.

RAMBERG, J. S., DUDEWICZ, E. J., TADIKAMALLA, P. R. & MYKYTKA, E. F. (1979). A probability distribution and its uses in fitting data. *Technometrics* **21**, 201–14.

SARHAN, A. E. & GREENBERG, B. G. (Eds). (1962). *Contributions to Order Statistics*. New York: Wiley.

WHITTLESEY, J. R. B. (1968). A comparison of the correlational behavior of random number generators for the IBM 360. *Comm. Assoc. Comp. Mach.* **11**, 641–4.

[*Received October* 1980. *Revised February* 1982]