Statistica Sinica **10**(2000), 1267-1280

WHERE IS THE FISHER INFORMATION IN AN ORDERED SAMPLE?

Gang Zheng and Joseph L. Gastwirth

George Washington University

Abstract: Suppose we have a random sample of size n with multiple censoring. The exact Fisher information in the data is derived and expressed in terms of matrices when each block of censored data contains at least two order statistics. The results are applied to determine how much Fisher information about the location (scale) parameter is contained in the middle (two tails) of an ordered sample. The results show that, for Cauchy, Laplace, logistic, and normal distributions, the middle 40% (extreme half) of the ordered data contains more than 80% of the Fisher information about the location (scale) parameter. These results provide insight into the behavior of two well-known robust linear estimators of the location parameter.

Key words and phrases: Decomposition of Fisher information, limiting Pitman efficiency, location-scale family, matrix expression, multiply censored data.

1. Introduction

Suppose $X_1, \ldots, X_n$ are i.i.d. random variables from c.d.f. $F_\theta(x)$ with continuous density $f_\theta(x)$. Let $X_{1:n}, \ldots, X_{n:n}$ be their order statistics. When only m of the n order statistics are available, denoted by $\mathbf{X} = (X_{k_1:n}, \ldots, X_{k_m:n})$ with joint density $f_{k_1\cdots k_m;n}$, the Fisher information about θ contained in $\mathbf{X}$, under some regularity conditions, is given by

$$I_{k_1\cdots k_m;n}(\theta) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{x_{k_2:n}} (\frac{\partial}{\partial\theta}\log f_{k_1\cdots k_m;n})^2 dF_{k_1\cdots k_m;n}. \qquad (1)$$

How much Fisher information (FI) about θ is contained in $\mathbf{X}$? Tukey (1965) discussed this issue in terms of linear sensitivity of blocks of consecutive order statistics. Nagaraja (1994) studied Tukey's concept of linear sensitivity and related it to asymptotic approximations of FI which were previously used in comparing estimators based on order statistics (Ogawa (1951), Chernoff, Gastwirth and Johns (1967) and David (1981, p.276)). An exact FI expression in the first r order statistics, $I_{1\cdots r;n}(\theta)$, was obtained by Mehrotra, Johnson and Bhattacharyya (1979). Park (1996) also examined the FI in the first r order statistics using a decomposition of FI, and expressed $I_{1\cdots r;n}(\theta)$ as a sum of r single integrals.

All these authors studied exact FI about the location and scale parameters based on consecutive order statistics. In order to examine which parts of the ordered data contain more FI, especially for the scale parameter, we need to consider FI in scattered blocks. The recurrence relations for FI in several order statistics, as provided by Park (1996), are not directly applicable to our problem for scattered blocks. However, using an alternative decomposition, FI contained in multiply censored data can be reduced to FI in right (left) and/or middle censored data. The approach used by Mehrotra, Johnson and Bhattacharyya (1979) to obtain FI in the right (left) censored data is utilized to obtain the result for the middle censored data. For multiply censored data, a matrix formulation enables one to easily obtain exact FI about a parameter. These exact results are compared with their asymptotic approximations for a sample of size 20. In Section 2, we give the exact and matrix expressions of FI. Asymptotic FI about the location and scale parameters in multiply censored data is obtained from the correlation of asymptotically most powerful grouped rank tests in Section 3. Applications are presented in Section 4 where we plot the exact and asymptotic percentages of FI about the location (scale) parameter in the middle portion (two tails) of an ordered sample. For Cauchy, Laplace, logistic, and normal distributions, the asymptotic percentages of FI are close to the exact values for the scale parameter when the sample size equals 20. For the location parameter, however, the proportion of FI in the middle two or four order statistics in a sample of size 20 from the Laplace distribution is not well-approximated. In fact, the sample median, the asymptotically optimum estimator of the location parameter for Laplace distribution, only contains about 77.5% of the FI in a sample of size 15. Our results are used to provide insight into the properties of Tukey's trimean and Gastwirth's estimate of the location parameter (Andrews, Bickel, Hampel, Huber, Rogers and Tukey (1972), Cox and Hinkley (1974, Sec.9.4), Hogg (1974) and Kennedy (1992, p.281)).

2. Exact Results

Let $\mathbf{X} = (X_{i_1:n}, \ldots, X_{i_1+k_1:n}; \cdots; X_{i_p:n}, \ldots, X_{i_p+k_p:n})$ be ordered data of a random sample $X_1, \ldots, X_n$ where $(X_{i_1:n}, \ldots, X_{i_1+k_1:n})$ is the first block of available consecutive order statistics and $(X_{i_p:n}, \ldots, X_{i_p+k_p:n})$ is the last one. Assume these p blocks are disjoint. We will derive expressions for $I_{\mathbf{X}}(\theta)$ from (1).

2.1. Exact Fisher information

Mehrotra, Johnson and Bhattacharyya (1979) defined the following three extended hazard rate functions:

$$K_1(x_{j:n}) = -\frac{F'_\theta(x_{j:n})}{1-F_\theta(x_{j:n})}, K_2(x_{i:n}) = \frac{F'_\theta(x_{i:n})}{F_\theta(x_{i:n})}, K_3(x_{i:n}, x_{j:n}) = \frac{F'_\theta(x_{j:n}) - F'_\theta(x_{i:n})}{F_\theta(x_{j:n}) - F_\theta(x_{i:n})},$$

where $F_\theta' = \partial F_\theta/\partial\theta$. Let $\phi_\theta = \partial \log f_\theta/\partial\theta$ and $\tau(i,j) = E_\theta[\phi_\theta(X_{i:n})\phi_\theta(X_{j:n})]$. The partial derivative of the log likelihood of $\mathbf{X}$ is a linear function of K_i and ϕ_θ, $i = 1,2,3$. They studied and used the moment relations among K_1, K_2, K_3 and ϕ_θ to derive the exact FI about θ contained in the first r order statistics via the $\tau(i,j)$. For a complete sample, it can be shown that $I_{1\,2\cdots n;n}(\theta) = \sum_{i=1}^n \sum_{j=1}^n \tau(i,j) = \sum_{i=1}^n \tau(i,i)$.

The calculation of FI in multiply censored data can be done in stages. This enables one to reduce the problem to methods for obtaining FI in right (left) and/or middle censored data. To describe the method, we consider two blocks of order statistics. The idea is generalized to the multiply censored data in Example 2.1. To calculate $I_{r\cdots u\,v\cdots w;n}(\theta)$, where $1 \le r \le u \le v \le w \le n$, we use three steps: 1) $I_{1\cdots w;n}(\theta) = I_{1\cdots n;n}(\theta) - I_R$; 2) $I_{r\cdots w;n}(\theta) = I_{1\cdots w;n}(\theta) - I_L$; 3) $I_{r\cdots u\,v\cdots w;n}(\theta) = I_{r\cdots w;n}(\theta) - I_M$. Thus, we have $I_{r\cdots u\,v\cdots w;n}(\theta) = I_{1\cdots n;n}(\theta) - I_L - I_M - I_R$ where I_L, I_M, and I_R depend only on where each censored block begins and ends, respectively.

Theorem 2.1. *When a block of order statistics is removed from the left, middle, or right, the change of FI is the same regardless of the previous censoring patterns, i.e.,*

$$I_L = I_{1\cdots n;n}(\theta) - I_{r\cdots n;n}(\theta) = I_{1\cdots r\,j_1\,j_2\cdots j_t;n}(\theta) - I_{r j_1 j_2\cdots j_t;n}(\theta), \quad (2)$$

$$I_M = I_{1\cdots n;n}(\theta) - I_{1\cdots u\,v\cdots n;n}(\theta) = I_{i_1\,i_2\cdots i_s u\cdots v j_1 j_2\cdots j_t;n}(\theta) - I_{i_1\,i_2\cdots i_s u v j_1 j_2\cdots j_t;n}(\theta), \quad (3)$$

$$I_R = I_{1\cdots n;n}(\theta) - I_{1\cdots w;n}(\theta) = I_{i_1 i_2\cdots i_s w\cdots n;n}(\theta) - I_{i_1 i_2\cdots i_s w;n}(\theta). \quad (4)$$

Proof. We prove (3). The other two can be obtained similarly. Let $v \ge u+2$. By the Markov property of order statistics (David (1981, p.20)), for $1 \le i_1 \le i_2 \le \cdots \le i_s \le u < v \le j_1 \le j_2 \le \cdots \le j_t \le n$, $f_{(u+1)\cdots(v-1)|i_1 i_2\cdots i_s\,u\,v\,j_1 j_2\cdots j_t;n} = f_{(u+1)\cdots(v-1)|u\,v;n}$ where $f_{(u+1)\cdots(v-1)|i_1 i_2\cdots i_s\,u\,v\,j_1 j_2\cdots j_t;n}$ is the joint density of $X_{u+1:n}, \ldots, X_{v-1:n}$ given $s+1$ smaller order statistics and $t+1$ larger order statistics. Thus,

$$\frac{\partial}{\partial\theta}\log f_{(u+1)\cdots(v-1)|u\,v;n} = \frac{\partial}{\partial\theta}\log f_{i_1 i_2\cdots i_s\,u\cdots v\,j_1 j_2\cdots j_t;n} - \frac{\partial}{\partial\theta}\log f_{i_1 i_2\cdots i_s\,u\,v\,j_1 j_2\cdots j_t;n}.$$

Hence by the property of conditional expectation (Rao (1973, p.330)), we obtain

$$\begin{aligned} I_{(u+1)\cdots(v-1)|u\,v;n}(\theta) &= I_{i_1 i_2\cdots i_s\,u\cdots v\,j_1 j_2\cdots j_t;n}(\theta) + I_{i_1 i_2\cdots i_s\,u\,v\,j_1 j_2\cdots j_t;n}(\theta) \\ &\quad -2E\Big(\frac{\partial}{\partial\theta}\log f_{i_1 i_2\cdots i_s\,u\cdots v\,j_1 j_2\cdots j_t;n}\frac{\partial}{\partial\theta}\log f_{i_1 i_2\cdots i_s\,u\,v\,j_1 j_2\cdots j_t;n}\Big) \\ &= I_{i_1 i_2\cdots i_s\,u\cdots v\,j_1 j_2\cdots j_t;n}(\theta) - I_{i_1 i_2\cdots i_s\,u\,v\,j_1 j_2\cdots j_t;n}(\theta). \end{aligned} \quad (5)$$

Then (3) follows from (5) since the left hand side of (5) depends only on u and v.

The FI expression in k scattered order statistics of Park (1996) can be obtained by Theorem 2.1. To calculate FI in middle censored data, we have

Theorem 2.2. *If $v > u$, then*

$$I_{1\cdots(u-1)(v+1)\cdots n;n}(\theta) = I_{1\cdots n;n}(\theta) - [\sum_{i=u}^{v} \tau(i,i) - \frac{2}{v-u}\sum_{i=u}^{v-1}\sum_{j=i+1}^{v} \tau(i,j)]. \quad (6)$$

Proof. The log likelihood of $\mathbf{X} = (X_{1:n},\ldots,X_{u-1:n},X_{v+1:n},\ldots,X_{n:n})$ is given by $l \sim \sum_{i=1}^{u-1} \log f_\theta(X_{i:n}) + \sum_{i=v+1}^{n} \log f_\theta(X_{i:n}) + (v-u+1)\log\{F_\theta(X_{v+1:n}) - F_\theta(X_{u-1:n})\}$. Thus $\partial l/\partial\theta = \sum_{i=1}^{n} \phi_\theta(X_{i:n}) - \sum_{i=u}^{v} \phi_\theta(X_{i:n}) + (v-u+1)K_3(X_{u-1:n}, X_{v+1:n})$. Using Lemmas A.3 (iii) and (iv), A.4 (iii), and A.5 (i) of Mehrotra, Johnson and Bhattacharyya (1979) and substituting g by ϕ_θ and h_3 by K_3, we have

$$\begin{aligned} I_{\mathbf{X}}(\theta) = E(\partial l/\partial\theta)^2 &= I_{1\cdots n;n}(\theta) + \sum_{i=u}^{v}\sum_{j=u}^{v} \tau(i,j) + \frac{2(v-u+1)}{v-u}\sum_{i=u}^{v-1}\sum_{j=i+1}^{v} \tau(i,j) \\ &\quad -2\sum_{i=1}^{n}\sum_{j=u}^{v} \tau(i,j) + 2\sum_{i=1}^{u-1}\sum_{j=u}^{v} \tau(i,j) + 2\sum_{i=v+1}^{n}\sum_{j=u}^{v} \tau(i,j) \\ &= I_{1\cdots n;n}(\theta) - \sum_{i=u}^{v} \tau(i,i) + \frac{2}{v-u}\sum_{i=u}^{v-1}\sum_{j=i+1}^{v} \tau(i,j). \end{aligned}$$

Example 2.1. Mehrotra, Johnson and Bhattacharyya (1979) derived

$$I_{1\cdots r;n}(\theta) = I_{1\cdots n;n}(\theta) - [\sum_{i=r+1}^{n} \tau(i,i) - \frac{2}{n-r-1}\sum_{i=r+1}^{n-1}\sum_{j=i+1}^{n} \tau(i,j)]. \quad (7)$$

The FI in the left censored data, $I_{s\cdots n;n}(\theta)$, can be obtained by symmetry. From Theorem 2.1, for $\mathbf{X}$ defined at the beginning of Section 2, we have

$$\begin{aligned} I_{\mathbf{X}}(\theta) = I_{1\cdots n;n}(\theta) - \sum_{j=0}^{p}[&\sum_{u=i_j+k_j+1}^{i_{j+1}-1} \tau(u,u) \\ &-\frac{2}{i_{j+1}-i_j-k_j-2}\sum_{u=i_j+k_j+1}^{i_{j+1}-2}\sum_{v=u+1}^{i_{j+1}-1} \tau(u,v)], \end{aligned} \quad (8)$$

where $i_1 > 2$, $i_p + k_p < n-1$, and $i_{j+1} - i_j - k_j > 2$, $j = 1, \ldots, p-1$, $i_0 = k_0 = 0$, $i_{p+1} = n+1$.

From (8), we can see that FI in multiply censored data is equal to the total FI minus I_L, I_M, and I_R for censored blocks. The advantage of this approach is that once the values $\tau(i,j)$ are tabulated, FI in scattered order statistics can be obtained as easily as that of a block of consecutive order statistics.

2.2. A matrix expression

Let $\mathbf{T}$ be the $n\times n$ symmetric matrix $(\tau(i,j))_{n\times n}$. Define $\mathbf{A}\odot\mathbf{B} = (a_{ij}b_{ij})_{n\times n}$ where $\mathbf{A} = (a_{ij})_{n\times n}$ and $\mathbf{B} = (b_{ij})_{n\times n}$ are two matrices. Suppose a block of consecutive order statistics $\mathbf{P} = (X_{a:n},\ldots, X_{b:n})$ is censored from a full sample. Define a block diagonal weighting matrix $\mathbf{W}_{a\,b}$ for the block $\mathbf{P}$ as $\mathbf{W}_{a\,b} = \text{diag}(\,\mathbf{I}_{a-1}, \mathbf{C}_{b-a},\ \mathbf{I}_{n-b}\,)$, where $\mathbf{I}_{a-1}$ and $\mathbf{I}_{n-b}$ are identity matrices and $\mathbf{C}_{b-a}$ is a $(b-a+1)\times(b-a+1)$ censoring matrix given by $\mathbf{J}/(b-a)$, where all off-diagonals of $\mathbf{J}$ are 1's and all diagonals are 0's. Thus from Theorem 2.1 we have $I_{1\cdots(a-1)(b+1)\cdots n;n}(\theta) = \mathbf{1}'(\mathbf{W}_{ab}\odot\mathbf{T})\mathbf{1}$, where $\mathbf{1}$ is a $1\times n$ vector of 1's. Generally, with $\mathbf{X}$ defined as in the beginning of Section 2, $I_{\mathbf{X}}(\theta) = \mathbf{1}'(\mathbf{W}_{1\,(i_1-1);\cdots;(i_p+k_p+1)\,n}\odot\mathbf{T})\mathbf{1}$, where $\mathbf{W}_{1\,(i_1-1);\cdots;(i_p+k_p+1)\,n} = \text{diag}\ (\,\mathbf{C}_{i_1-2}, \mathbf{I}_{k_1+1}, \mathbf{C}_{i_2-i_1-k_1-2},\ldots,\mathbf{I}_{k_p+1}, \mathbf{C}_{n-i_p-k_p-1})$ depends only on which order statistics are censored. If no censoring occurs, then the weighting matrix is the identity matrix $\mathbf{I_n}$.

Example 2.2. For a random sample from the exponential distribution with the scale parameter θ, $\theta^2 I_{1\,\ldots\,10;10}(\theta) = 10$. To calculate $I_{3\,4\,7;10}(\theta)$, blocks $(1,2)$, $(5,6)$ and $(8,9,10)$ are censored. Therefore, $\mathbf{W}_{1\,2;5\,6;8\,10} = \text{diag}(\mathbf{C}_1,\mathbf{I}_2,\mathbf{C}_1,1,\mathbf{C}_2)$ and $\theta^2 I_{3\,4\,7;10}(\theta) = \mathbf{1}'(\mathbf{W}_{1\,2;5\,6;8\,10}\odot\mathbf{T})\mathbf{1} = 6.8953$. These results complement those of Arnold, Balakrishnan and Nagaraja (1992, p.166) and Nagaraja (1994), giving the FI in consecutive order statistics from an exponential random variable.

3. Asymptotic Results

We derive the asymptotic FI for multiply censored data from $F((x-\theta_1)/\theta_2)$ based on the correlation of asymptotically most powerful grouped rank tests (AMPGRT), see Gastwirth (1965a). The results can also be obtained directly from the exact FI (Zheng (2000)), and from Chernoff, Gastwirth and Johns (1967) and Sen (1967).

Assume only observations in the percentile ranges $[p_i, q_i]$, $i = 1,\ldots,r$, are observed where $0 = q_0 \le p_1 < q_1 < \cdots < p_r < q_r \le p_{r+1} = 1$. Let $E = \bigcup_{i=1}^{r}[p_i, q_i]$. From Gastwirth (1965a), the weight function $K_j(u)$ corresponding to AMPGRT for $H_0 : F(x) = G(x)$ against the alternative H_j where $H_1 : G(x) = F(x-\theta_1)$ and $H_2 : G(x) = F(x/\theta_2)$, when we only observe samples in E, is given by $K_j(u) = J_j(u)$ if $u\in E$, or $K_j(u) = c_{ij}$, if $q_{i-1} \le u < p_i$, where the c_{ij}'s are constants determined by $\int_{q_{i-1}}^{p_i}(J_j(u)-K_j(u))\,du = 0$ for $i = 1,\ldots,r+1$ and $j = 1, 2$, $J_1(u) = -f'(x)/f(x)$, $J_2(u) = -\{1 + xf'(x)/f(x)\}$, and $x = F^{-1}(u)$. Here $J_j(u)$ is the weight function of AMPGRT for θ_j with full samples (Chernoff and Savage (1958), Gastwirth (1965b), and Hájek and Šidák (1967)).

Under a quadratically integrable condition, i.e., $0 < \int_0^1 J_j^2(u)du < \infty$ for $j = 1, 2$, from Hájek (1962) and van Eeden (1963), the limiting Pitman efficiency of tests based on $K_j(u)$ and $J_j(u)$ is given by

$$\rho_j^2 = \frac{\{\int_0^1 J_j(u)\, K_j(u)\, du\}^2}{\int_0^1 J_j^2(u)\, du\, \int_0^1 K_j^2(u)\, du}. \tag{9}$$

This is also a ratio of the asymptotic FI contained in E and in $[0,1]$ since the rank tests based on $K_j(u)$ and $J_j(u)$ are optimal, i.e.,

$$\rho_j^2 = \frac{I_E(\theta_j)}{I_{[0,1]}(\theta_j)} = \frac{I_E(\theta_j)}{I_{1;1}(\theta_j)}. \tag{10}$$

Denote the p_ith percentile of F as λ_{p_i}. Then for the scale parameter,

$$\int_0^1 J_2(u)K_2(u)du = \int_E J_2^2(u)du + \sum_{i=1}^{r+1} \frac{1}{p_i - q_{i-1}} \{\int_{q_{i-1}}^{p_i} J_2(u)du\}^2 = \int_0^1 K_2^2(u)du. \tag{11}$$

If $xf(x) \to 0$ ($f(x) \to 0$ for the location parameter) as $x \to \pm\infty$, then

$$\int_{q_{i-1}}^{p_i} J_2(u)du = -\int_{\lambda_{q_{i-1}}}^{\lambda_{p_i}} f(x)dx - \int_{\lambda_{q_{i-1}}}^{\lambda_{p_i}} xf'(x)dx = -\lambda_{p_i} f(\lambda_{p_i}) + \lambda_{q_{i-1}} f(\lambda_{q_{i-1}}). \tag{12}$$

Note that $\int_0^1 J_2^2(u)du = \theta_2^2 I_{1;1}(\theta_2)$. Therefore, from (9), (10), (11), and (12), we have

$$\theta_2^2\, I_E(\theta_2) = \int_E J_2^2(u)du + \sum_{i=1}^{r+1} \frac{[\lambda_{p_i} f(\lambda_{p_i}) - \lambda_{q_{i-1}} f(\lambda_{q_{i-1}})]^2}{p_i - q_{i-1}}. \tag{13}$$

Similarly, we can obtain the asymptotic FI for the location parameter as

$$\theta_2^2\, I_E(\theta_1) = \int_E J_1^2(u)du + \sum_{i=1}^{r+1} \frac{[f(\lambda_{p_i}) - f(\lambda_{q_{i-1}})]^2}{p_i - q_{i-1}}. \tag{14}$$

Let $E = \bigcup_{i=1}^r [p_i, p_i]$. Then (13) (or (14)) becomes the asymptotic FI about θ_1 (or θ_2) in r percentiles $\lambda_{p_1}, \ldots, \lambda_{p_r}$, denoted as $I_{[p_1,p_1]\cup\cdots\cup[p_r,p_r]}(\theta_i)$, $i = 1, 2$, respectively. The values $\lambda_{p_1}, \ldots, \lambda_{p_r}$ that maximize $I_{[p_1,p_1]\cup\cdots\cup[p_r,p_r]}(\theta_i)$ are the most informative r percentiles for θ_i. These most informative r percentiles are equivalent to r optimum spacings in the sense that the best linear unbiased estimator (BLUE) of θ_i using any r percentiles has maximum asymptotic relative efficiency (R.E.) with respect to the Cramér-Rao lower bound (CRLB) when the optimum spacings are used (Ogawa (1951) and David (1981, p.195)). From (13) and (14), we obtain, for $j = 1, 2$, $\sum_{i=0}^r I_{[p_i,p_{i+1}]}(\theta_j) = I_{[0,1]}(\theta_j) + \sum_{i=1}^r I_{[p_i,p_i]}(\theta_j)$, where $p_0 = 0$ and $p_{r+1} = 1$. Setting $r = 2$ and $p_1 = p_2 = 1/2$, it follows that the sum of the proportions of FI in the first half and the second half of a sample equals 1 plus the proportion of FI in the median. This result formalizes Tukey's (1965) insight that one order statistic "borrows" information from the others.

For a symmetric location-scale family $F((x-\theta_1)/\theta_2)$, we are interested in finding an interval $[p, p+q]$, where $0 \le p < p+q \le 1$ and q is fixed, that contains most of the FI about θ_1. If $p = (1-q)/2$, then $[p, p+q]$ is a symmetric interval of length q with center $1/2$. For a symmetric distribution, we only need to consider $p \in [0, (1-q)/2]$. Let $h(x) = f(x)/\{1-F(x)\}$ and define for the location and scale parameters,

$$g_1(x) = [\frac{d}{dx}\log h(x)]^2, \quad g_2(x) = [1 + x\frac{d}{dx}\log h(x)]^2, \tag{15}$$

respectively, Then, from (13) and (14) for $p \in [0, (1-q)/2]$ and $j = 1, 2$,

$$\frac{d}{dp}\theta_2^2 I_{[p,p+q]}(\theta_j) = g_j(\lambda_{p+q}) - g_j(\lambda_{1-p}). \tag{16}$$

When $p = (1-q)/2$, (16) is zero, which implies that $I_{[p,p+q]}(\theta_j)$ has a local extreme at $p = (1-q)/2$. Then by symmetry, $I_{[p,p+q]}(\theta_1)$ has a global maximum at this point when (16) is non-negative for $p \in [0, (1-q)/2]$. The formal result is given in

Theorem 3.1. *Suppose $F((x-\theta_1)/\theta_2)$ is a symmetric distribution satisfying regularity conditions, and g_j, $j = 1, 2$ are defined as in* (15). *Then for $p \in [0, 1-q]$, $I_{[p,p+q]}(\theta_j)$ is non-decreasing in $p \in [0, (1-q)/2]$ and non-increasing in $p \in [(1-q)/2, 1-q]$ if and only if, for any $p \in [0, (1-q)/2]$ and $j = 1, 2$,*

$$g_j(\lambda_{1-p}) \le g_j(\lambda_{p+q}). \tag{17}$$

Theorem 3.1 can be used to determine whether the middle portion of an ordered sample contains more FI about the location parameter of normal, logistic, Laplace, and Cauchy distributions. For normal and logistic distributions, $g_1(x)$ is strictly decreasing for all x. Thus (17) is satisfied. For the Laplace distribution, $g_1(x)$ is not decreasing for all x but (17) is still satisfied because the left hand side of (17) is identically zero. Thus, for these three distributions the middle portion of the data contains more FI about the location parameter than any other single interval with the same length. For the Cauchy distribution, however, (17) is not satisfied for all t. For $q = 1/2$, Figure 1 plots the asymptotic fraction of FI contained in the percentiles $[p, p+1/2]$, for $p \in [0, 1/4]$. Notice that the middle interval with $p = 1/4$ contains the smallest percentage of FI. This is somewhat surprising, as this interval contains over 90% of the total FI, i.e., $I_{[1/4,3/4]}(\theta_1) = .4526$ and $I_{[0,1]}(\theta_1) = .5$.

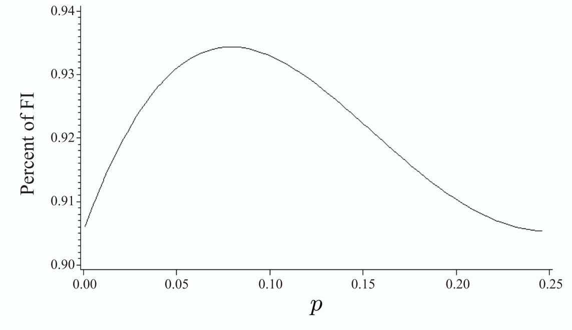

Figure 1. The asymptotic percentage of FI about the location parameter of Cauchy distribution in percentiles p to $p + 1/2$, for p between 0 and 1/4.

For the scale parameter, however, we examine FI in the two tails using (13), where $E = [0, q/2] \cup [1 - q/2, 1]$, since the two tails usually contain more FI than a single block of order statistics. Numerical calculations and plots in Zheng (2000) show that $I_{[0,q/2]\cup[1-q/2,1]}(\theta_2) \geq I_{[p,p+q]}(\theta_2)$ for $0.15 < q < 0.90$ and any $p \in (0, (1 - q)/2]$, for all four distributions, and equality holds for the Cauchy when $p = (1 - q)/2$.

4. Applications

The information in selected order statistics has been useful in several applications. Sometimes very large data sets are collected, but only a few summary measures and values can be stored. Several informative order statistics can then be usefully employed (Eisenberger and Posner (1965)). Sometimes the determination of the status of an observation can be quite costly. In ranked set samples, one uses a cheaper proxy measurement before one selects the sample for a more careful measurement. Öztürk and Wolfe (2000) use the information in the order statistics of the proxy measurements to select those for the second stage. In genetic linkage analysis, Risch and Zhang (1995) found that tests using extreme discordant sib pairs are most powerful. Usually, the cost of measuring the trait, e.g., blood pressure, is much less than the cost of genotyping. Consider the absolute difference between the trait values of the two sibs. The upper quantiles now correspond to extreme discordant sib pairs. The intuition underlying the Risch and Zhang (1995) procedure is supported by an analysis of FI in the upper portion of the data.

4.1. Information about location and scale parameters

For the location-scale family $F((x - \theta_1)/\theta_2)/\theta_2$, we compute the exact percentage of FI of θ_1: $I_{(11-k)\cdots(10+k);20}(\theta_1)/I_{1\cdots 20;20}(\theta_1)$ for $k = 1, \ldots, 8$. For the scale parameter, we calculate $I_{1\cdots k(21-k)\cdots 20;20}(\theta_2)/I_{1\cdots 20;20}(\theta_2)$ for $k = 1, \ldots, 9$.

Similarly, we calculate these percentages for $n = 15$. Table 1 reports some results for the Cauchy, Laplace, logistic, and normal distributions. For the Laplace distribution, $\partial f(x - \theta_1)/\partial \theta_1$ does not exist when $\theta_1 = x$. However, under a quadratically integrable condition, i.e., $0 < \int [f'(x)/f(x)]^2 f(x) dx < \infty$, $\tau(i, j)$ exists and plays the role of Fisher information (Johnson (1974) and Mehrotra, Johnson and Bhattacharyya (1979)).

From Table 1, the middle 40% of the data contains over 80% of the FI about θ_1 for all four distributions. For the scale parameter, the extreme 20% of the order statistics contains nearly 80% of the FI for the Laplace, logistic, and normal distributions. The extreme 50% of the data (25% in each tail) contains at least 80% of the FI for all four distributions. To see what the percentage FI in Table 1 tells us, we calculate the variance of the BLUE of θ_1 (θ_2) based on censored data using David (1981, p.131). For $n = 20$, Table 2 reports the R.E. of the BLUE based on the central statistics (two tails) for θ_1 (θ_2) to the BLUE using the complete sample.

Table 1. Exact percentage of FI contained in ordered data.

	Location parameter (θ_1)				Scale parameter (θ_2)			
	n=15		n=20		n=15		n=20	
	% Data	% FI	% Data	% FI	% Data	% FI	% data	% FI
Cauchy	7%	.7221	10%	.7755	13%	.2588	10%	.2009
	20%	.8160	20%	.8330	27%	.4926	20%	.3932
	33%	.8711	30%	.8689	40%	.6834	30%	.5634
	47%	.9075	40%	.8939	53%	.8254	40%	.7043
	53%	.9382	50%	.9158	67%	.9207	50%	.8140
Laplace	7%	.7747	10%	.8464	13%	.6269	10%	.5707
	20%	.8993	20%	.9271	27%	.8241	20%	.7704
	33%	.9639	30%	.9707	40%	.9127	30%	.8690
	47%	.9901	40%	.9902	53%	.9579	40%	.9241
	53%	.9980	50%	.9974	67%	.9820	50%	.9566
Logistic	7%	.7529	10%	.7857	13%	.6260	10%	.5662
	20%	.8353	20%	.8442	27%	.8350	20%	.7786
	33%	.8971	30%	.8909	40%	.9262	30%	.8828
	47%	.9412	40%	.9273	53%	.9687	40%	.9383
	53%	.9706	50%	.9545	67%	.9885	50%	.9686
Normal	7%	.6556	10%	.6810	13%	.7125	10%	.6601
	20%	.7323	20%	.7368	27%	.8901	20%	.8482
	33%	.7984	30%	.7867	40%	.9553	30%	.9267
	47%	.8552	40%	.8312	53%	.9823	40%	.9640
	53%	.9034	50%	.8708	67%	.9938	50%	.9826

NOTE: % Data is the percentage of the middle portion of the ordered sample for the location parameter and the percentage of sample in two tails for the scale.

In Table 2, the efficiency for the scale parameter based on two tails is not computed for the Cauchy distribution, because the first and last order statistics of the Cauchy distribution have infinite variance. Comparing Table 1 with Table 2 we find, for the scale parameter, the R.E. is close to the exact percentage of FI for all three distributions. For the location parameter, the exact percentage of FI is close to the R.E. for logistic and normal distributions.

Table 2. Efficiency of the BLUE based on censored data relative to the BLUE based on full samples (n=20).

	Cauchy	Laplace		Logistic		Normal	
% data	θ_1	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2
10%	.9011	.9579	.5610	.7947	.5610	.6808	.7051
20%	.9254	.9879	.7847	.8535	.8052	.7365	.8870
30%	.9263	.9982	.8884	.9000	.9127	.7864	.9538
40%	.9311	.9999	.9425	.9354	.9626	.8309	.9814
50%	.9485	1.000	.9721	.9613	.9854	.8706	.9931

NOTE: % Data is defined as in Table 1.

For Cauchy and Laplace distributions, the exact percentage of FI is not close to the R.E. of the BLUE for a small proportion of order statistics in the case of the location parameter, since the CRLB is not attained for finite samples from the Cauchy and Laplace distributions. Asymptotically, however, L-estimators for the location parameters of these distributions are fully efficient. In Zheng (2000), it is shown, through simulation, that the variance of the BLUE based on the entire sample is within 10% of the CRLB for the Cauchy when $n = 50$, and about 11% for the Laplace when $n = 100$.

To assess how well the asymptotic FI approximates exact FI for $n = 20$, for the location parameter, we focus on FI in the central portion of the data and plot $I_{[11.5/21-p,\, 9.5/21+p]}(\theta_1)$ with $I_{(11-k)\cdots(10+k);20}(\theta_1)$, both divided by total FI, where $p = k/21$, $k = 1, \ldots, 8$. For the scale parameter, we concentrate on FI in tail portions of data and plot $I_{[0,p]\cup[1-p,1]}(\theta_2)$ with $I_{1\cdots k(21-k)\cdots 20;20}(\theta_2)$, both divided by total FI, where $p = k/21$, $k = 1, \ldots, 10$. Figures 2 to 5 present the plots.

4.2. On robust linear estimators of the location parameter

Several simple robust estimators based on linear combinations of order statistics for the location parameter were proposed in the 1960's, and examined in the Princeton study (Andrews et al. (1972)). In an ordered sample of size n (n odd), Tukey's trimean (TRI) using the 25th, 50th, and 75th percentiles and Gastwirth's estimator (GAS) (Gastwirth (1966)) using the $33\frac{1}{3}$rd, 50th, and $66\frac{2}{3}$rd percentiles, for the location parameter, are defined as follows:

$$TRI = 0.25X_{[n/4]:n} + 0.5X_{(n+1)/2:n} + 0.25X_{n+1-[n/4]:n},$$
$$GAS = 0.3X_{[n/3]:n} + 0.4X_{(n+1)/2:n} + 0.3X_{n+1-[n/3]:n}.$$

Approximate standard errors for them have been developed by Patel, Mudholkar and Indrasiri Fernando (1988) and Basset and Koenker (1978), who showed how they and other L-estimators can be used in regression analysis. We indicate how FI provides insight into their properties.

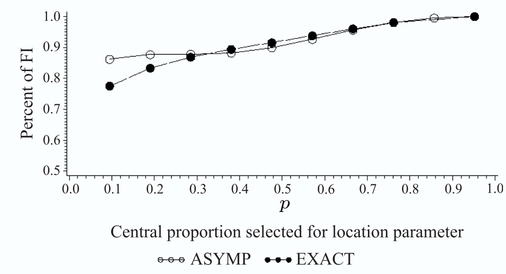

Figure 2a. The percentage of FI for the Cauchy location parameter.

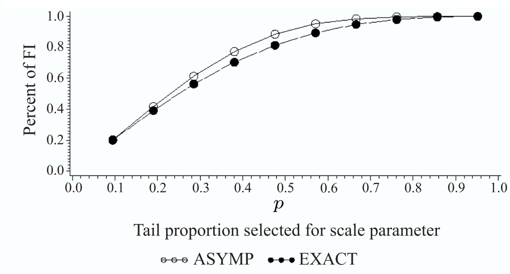

Figure 2b. The percentage of FI for the Cauchy scale parameter.

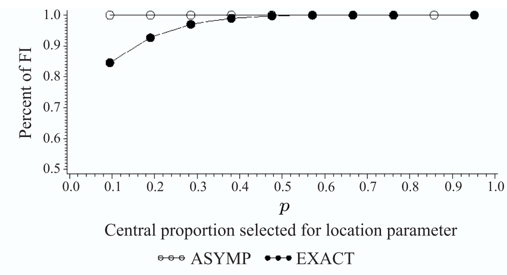

Figure 3a. The percentage of FI for the Laplace location parameter.

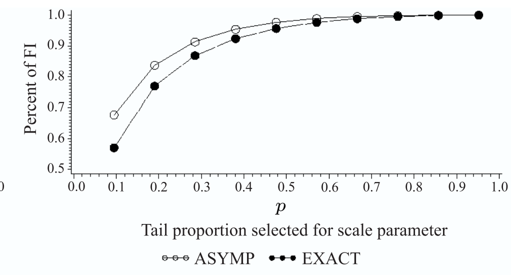

Figure 3b. The percentage of FI for the Laplace scale parameter.

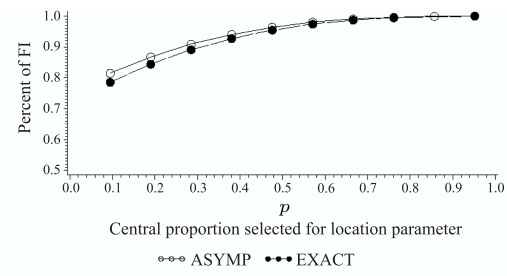

Figure 4a. The percentage of FI for for the logistic scale parameter.

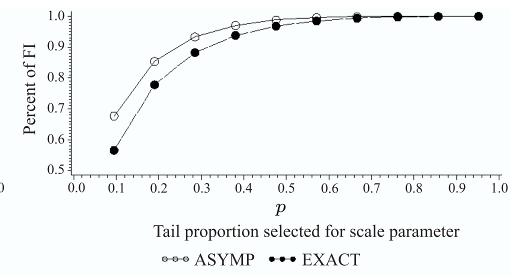

Figure 4b. The percentage of FI the logistic location parameter.

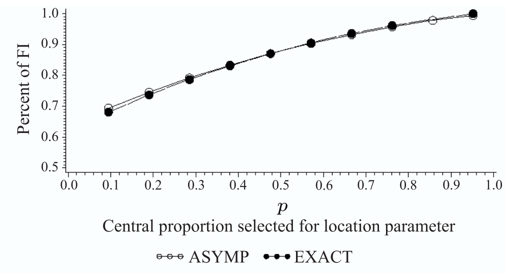

Figure 5a. The percentage of FI for the normal location parameter.

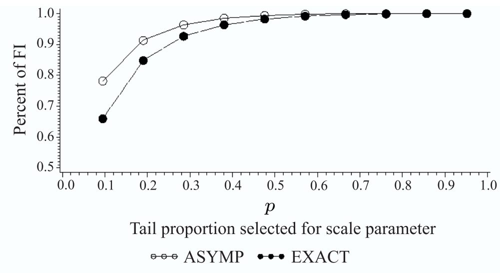

Figure 5b. The percentage of FI for the normal scale parameter.

For $n = 19$, $TRI = 0.25X_{5:19} + 0.5X_{10:19} + 0.25X_{15:19}$ and $GAS = 0.3X_{6:19} + 0.4X_{10:19} + 0.3X_{14:19}$. Andrews et al. (1972) and Huber (1972) indicated that these estimators are efficiency robust when the parent family of distributions underlying the data includes both the normal and a long-tailed distribution, e.g., Cauchy. Table 3 presents the percentage of FI in the middle three scattered order statistics $(X_{10-r:19}, X_{10:19}, X_{10+r:19})$, $r = 3, \ldots, 9$, for four distributions and samples of size 19. From Table 3, three order statistics in TRI (GAS) have minimum FI at the Cauchy (normal) distribution.

The large and small sample performance of TRI, GAS, and other robust estimates were evaluated using their large sample variances and simulation in Andrews et al. (1972). We are using the FI to provide insight into why these estimates have good efficiency properties in small samples, as well as in large samples. From Table 3, two largest minimum percentages of FI are GAS and TRI, about 85%, suggesting that an appropriate combination of these three order statistics might achieve 75% to 80% of the available FI simultaneously for all four models, as GAS and TRI do. The results also indicate that when heavier tailed distributions, e.g., the Cauchy, are not plausible models for the data, the estimator TRI should have slightly higher relative efficiency than GAS. The reverse is true for heavier tailed distributions.

Table 3. Exact percentage of FI about the location parameter in the middle three scattered order statistics for four distributions.

Robust Estimator	Samples	Percent of FI				Minimum
		Cauchy	Laplace	Logistic	Normal	
	$(X_{7:19}, X_{10:19}, X_{13:19})$	.862	.923	.910	.814	.814
GAS	$(X_{6:19}, X_{10:19}, X_{14:19})$	.858	.894	.932	.851	.851
TRI	$(X_{5:19}, X_{10:19}, X_{15:19})$	.845	.863	.940	.879	.845
	$(X_{4:19}, X_{10:19}, X_{16:19})$	.830	.836	.932	.895	.830
	$(X_{3:19}, X_{10:19}, X_{17:19})$	.815	.817	.910	.896	.815
	$(X_{2:19}, X_{10:19}, X_{18:19})$	.800	.804	.872	.875	.800
	$(X_{1:19}, X_{10:19}, X_{19:19})$	.776	.794	.820	.815	.776

Acknowledgements

This research was supported in part by a grant from NSF. The second author completed the research while visiting the Biostatistics Branch of the Division of Cancer Epidemiology and Genetics of the National Cancer Institute. We are grateful to two referees and the editor for their helpful comments regarding presentation.

References

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H. and Tukey, J. W. (1972). *Robust Estimates of Location.* Princeton University Press, Princeton, NJ.

Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (1992). *A First Course in Order Statistics.* John Wiley, New York.

Bassett, G., Jr. and Koenker, R. (1978). Asymptotic theory of least absolute error regression. *J. Amer. Statist. Assoc.* **73**, 618-622.

Chernoff, H., Gastwirth, J. L. and Johns, M. V. (1967). Asymptotic distribution of linear combinations of functions of order statistics with application to estimation. *Ann. Math. Statist.* **38**, 52-72.

Chernoff, H. and Savage, I. R. (1958). Asymptotic normality and efficiency of certain nonparametric statistics. *Ann. Math. Statist.* **29**, 972-994.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics.* Chapman and Hall, London.

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

Eisenberger, I. and Posner, E. C. (1965). Systematic statistics used for data compression in space telemetry. *J. Amer. Statist. Assoc.* **60**, 97-133.

Gastwirth, J. L. (1965a). Asymptotic most powerful rank tests for the two-sample problem with censored data. *Ann. Math. Statist.* **36**, 1243-1247.

Gastwirth, J. L. (1965b). Percentile modifications of two sample rank tests. *J. Amer. Statist. Assoc.* **60**, 1127-1141.

Gastwirth, J. L. (1966). On robust procedures. *J. Amer. Statist. Assoc.* **61**, 929-948.

Hájek, J. (1962). Asymptotically most powerful rank-order tests. *Ann. Math. Statist.* **33**, 1124-1147.

Hájek, J. and Šidák, Z. (1967). *Theory of Rank Tests.* Academic Press, New York.

Hogg, R. V. (1974). Adaptive robust procedures: a partial review and some suggestions for future applications and theory. *J. Amer. Statist. Assoc.* **69**, 909-923.

Huber, P. J. (1972). Robust statistics: a review. *Ann. Math. Statist.* **43**, 1041-1067.

Johnson, R. A. (1974). Asymptotic results for inference procedures based on the r smallest observations. *Ann. Statist.* **2**, 1138-1151.

Kennedy, P. (1992). *A Guide to Econometrics.* 3rd edition. MIT Press, Cambridge, MA.

Mehrotra, K. G., Johnson, R. A. and Bhattacharyya, G. K. (1979). Exact Fisher information for censored samples and the extended hazard rate functions. *Comm. Statist. Theory Methods* **15**, 1493-1510.

Nagaraja, H. N. (1994). Tukey's linear sensitivity and order statistics. *Ann. Inst. Statist. Math.* **46**, 757-768.

Ogawa, J. (1951). Contributions to the theory of systematic statistics, I. *Osaka Math. J.* **3**, 175-213.

Öztürk, Ö. and Wolfe, D. A. (2000). Optimal allocation procedure in ranked set sampling for unimodal and multi-modal distributions. To appear in *Environ. Ecol. Statist.*

Park, S. (1996). Fisher information in order statistics. *J. Amer. Statist. Assoc.* **91**, 385-390.

Patel, K. R., Mudholkar, G. S. and Indrasiri Fernando, J. L. (1988). Student's t approximations for three simple robust estimators. *J. Amer. Statist. Assoc.* **83**, 1203-1210.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. 2nd edition. John Wiley, New York.

Risch, N. and Zhang, H. (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* **268**, 1584-1589.

Sen, P. K. (1967). Asymptoticaly most powerful rank order tests for grouped data. *Ann. Math. Statist.* **38**, 1229-1239.

Tukey, J. W. (1965). Which part of the sample contains the information? *Proc. Nat. Acad. Sci. U.S.A.* **53**, 127-134.

van Eeden, Constance (1963). The relation between Pitman's asymptotic relative efficiency of two-tests and the correlation coefficient between their test statistics. *Ann. Math. Statist.* **34**, 1442-1451.

Zheng, G. (2000). Fisher information in order statistics and ordered randomly censored data. Ph. D. dissertation, George Washington University.

Office of Biostatistics Research, National Heart, Lung and Blood Institute, Rockledge 2, 6701 Rockledge Drive, Bethesda, MD 20892-7938, U.S.A.

E-mail: zhengg@nhlbi.nih.gov

Department of Statistics, George Washington University, Washington, DC 20052, U.S.A.

E-mail: jlgast@research.circ.gwu.edu

(Received June 1999; accepted February 2000)