A Measure of Inaccuracy in Order Statistics

Richa Thapliyal and H.C. Taneja

Department of Applied Mathematics,
Delhi Technological University,
Bawana Road, Delhi-110042, India

richa31aug@gmail.com, hctaneja@rediffmail.com

Received 20 January 2012

Accepted 5 June 2013

In this article, we consider a measure of inaccuracy between distributions of the i^{th} order statistics and parent random variable. It is shown that the inaccuracy measure characterizes the distribution function of parent random variable uniquely. We also discuss some properties of the proposed measure.

Keywords: Kullback relative information, Kerridge inaccuracy, Order statistics, Survival function.

1. Introduction

In information theory, entropy is a measure of the uncertainty associated with a random variable. This concept was introduced by Shannon [2]. Shannon entropy represents an absolute limit on the best lossless compression of any communication. Shannon entropy of a discrete random variable X with possible values $\{x_1, x_2, \dots, x_n\}$ and probability mass function p is defined as

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i). \tag{1.1}$$

In case of continuous sample, Shannon entropy is given by

$$H(f) = -\int_0^\infty f(x) \log f(x)\, dx. \tag{1.2}$$

Shannon entropy has been used as a major tool in information theory on in almost every branch of science and engineering. Let X and Y be two non-negative random variables with p.d.f. $f(x)$ and $g(x)$, respectively. Let $F(x) = P(X \leqslant x)$ and $G(y) = P(Y \leqslant y)$ be their distribution functions. The Kullback-Leibler [10] measure of discrimination of X about Y and Kerridge [3] measure of

Published by Atlantis Press

inaccuracy are given by

$$H(f \mid g) = \int_0^\infty f(x) \log \frac{f(x)}{g(x)} dx \tag{1.3}$$

$$H(f,g) = -\int_0^\infty f(x) \log g(x) dx \tag{1.4}$$

respectively. Note that

$$H(f \mid g) + H(f) = H(f,g).$$

In this article, we assume X to be a positive continuous random variable.

Suppose that $X_1, X_2, \dots, X_n$ are independent and identically distributed observations from cdf $F(x)$ and p.d.f. $f(x)$. The order statistics of the sample is defined by the arrangement of $X_1, X_2, \dots, X_n$ from the smallest to the largest, denoted as $X_{1:n} \leqslant X_{2:n} \leqslant \cdots \leqslant X_{n:n}$. These statistics have been used in a wide range of problems like detection of outliers, characterizations of probability distributions, quality control and strength of materials; for more details [1,4,6]. In reliability theory, order statistics are used for statistical modeling. The k^{th} order statistics in a sample of size n represents the life lengths of a $(n-k+1)$-out-of-n system.

Several authors have studied the information theoretic properties of an ordered data. Wong and Chen [5] showed that the difference between the average entropy of order statistics and the entropy of parent distribution is a constant. Park [11] obtained some recurrence relations for the entropy of order statistics. Ebrahimi *et al.* [8] explored some properties of the Shannon entropy of order statistics and showed that the Kullback-Leibler information functions involving order statistics are distribution free. We continue this line of research by deriving a measure of inaccuracy in order statistics and exploring some of it's properties.

Shannon's measure of uncertainty associated with i^{th} order statistics $X_{i:n}$ is given by

$$H(X_{i:n}) = -\int_0^\infty f_{i:n}(x) \log f_{i:n}(x) dx, \tag{1.5}$$

where

$$f_{i:n}(x) = \frac{1}{B(i, n-i+1)} (F(x))^{i-1} (1-F(x))^{n-i} f(x) \tag{1.6}$$

is p.d.f. of i^{th} order statistics, for $i = 1, 2, \dots, n$. Here

$$B(a,b) = \int_0^1 x^{a-1} (1-x)^{b-1} dx, \quad a > 0, \ b > 0, \tag{1.7}$$

is beta function with parameters a and b, [1].
Note that for $n = 1$, (1.5) reduces to (1.2). Using probability integral transformation $U = F(X)$, where U follows standard uniform distribution, the entropy of i^{th} order statistics is given by

$$H(X_{i:n}) = H_n(W_i) - E_{g_i}\left[\log(f(F^{-1}(W_i)))\right], \tag{1.8}$$

where

$$H_n(W_i) = \log B(i, n-i+1) - (i-1)[\psi(i) - \psi(n+1)] - (n-i)[\psi(n-i+1) - \psi(n+1)], \tag{1.9}$$

denotes entropy of i^{th} order statistics from standard uniform distribution whose p.d.f. is given by

$$g_i(w) = \frac{1}{B(i, n-i+1)} w^{i-1}(1-w)^{n-i}, \quad 0 < w < 1, \tag{1.10}$$

and $\psi(z) = \frac{d \log \Gamma(z)}{dz}$ is the digamma function (for details [8]).

In this communication, we study a measure of inaccuracy in order statistics. In Section 2, we propose a measure of inaccuracy between distributions of i^{th} order statistics and parent random variable X and study a characterization result based on this measure. In Section 3, we find bounds for inaccuracy measure and calculate the average of inaccuracy measure.

2. A Measure of Inaccuracy

Kullback-Leibler [10] measure of relative information between distribution of i^{th} order statistics and data distribution is given by

$$K_n(f_{i:n}, f_X) = \int_0^\infty f_{i:n}(y) \log\left(\frac{f_{i:n}(y)}{f_X(y)}\right) dy \tag{2.1}$$

Using probability integral transformation $U = F(X)$, this becomes

$$K_n(f_{i:n}, f_X) = K_n(g_i, U) = \int_0^\infty g_i(w) \log g_i(w) dw = -H_n(W_i), \tag{2.2}$$

where $f_X(y)$ is the p.d.f. of parent random variable X, $f_{i:n}$ is p.d.f. of i^{th} order statistics, g_i is the beta distribution (1.10) and U is the uniform distribution (for details [8]).
Adding (1.5) and (2.1), we get

$$\begin{aligned} H(X_{i:n}) + K_n(f_{i:n}, f_X) &= -\int_0^\infty f_{i:n}(y) \log f_{i:n}(y) dy + \int_0^\infty f_{i:n}(y) \log\left(\frac{f_{i:n}(y)}{f_X(y)}\right) dy \\ &= -\int_0^\infty f_{i:n}(y) \log f_X(y) dy. \end{aligned} \tag{2.3}$$

Using probability integral transformation $U = F(X)$, (2.3) reduces to $-E_{g_i}\left[\log(f(F^{-1}(W_i)))\right]$. Further, adding (1.8) and (2.2), we obtain

$$H(X_{i:n}) + K_n(f_{i:n}, f_X) = -E_{g_i}\left[\log(f(F^{-1}(W_i)))\right],$$

which is in confirmation with the result already obtained.

We define the measure

$$I_n(f_{i:n}, f) = -\int_0^\infty f_{i:n}(x) \log f(x) dx = -E_{g_i}\left[\log(f(F^{-1}(W_i)))\right] \tag{2.4}$$

as a measure of inaccuracy associated with distribution of i^{th} order statistics and parent distribution function $f(x)$, analogous to the Kerridge measure of inaccuracy between two density functions f and g given by (1.4).

Next, we show that the inaccuracy measure defined above characterizes the distribution function of parent random variable X uniquely. To prove this characterization result we use the following lemma [12].

Lemma 2.1. *For any increasing sequence of positive integers $\{n_j,\ j \geqslant 1\}$, the sequence of polynomials $\{x^{n_j}\}$ is complete in $L(0,1)$, if and only if $\sum_{j=1}^{\infty} n_j^{-1}$ is infinite.*

Here, $L(0,1)$ is the set of all Lebesgue integrable functions on the interval $(0,1)$.

Theorem 2.1. *Let X and Y be two positive random variables with p.d.f. $f(x)$ and $g(x)$ and absolutely continuous c.d.f. $F(x)$ and $G(x)$, respectively. Then, F and G belong to same family of distributions but for change in location if and only if*

$$I_n(f_{i:n}, f) = I_n(g_{i:n}, g), \quad 1 \leqslant i \leqslant n$$

for $n = n_j, j \geqslant 1$ such that $\sum_{j=1}^{\infty} n_j^{-1}$ is infinite.

Proof. The necessary part is obvious. We only need to prove the sufficiency part. If for all $n = n_j$, $j \geqslant 1$ such that $\sum_{j=1}^{\infty} n_j^{-1}$ is infinite and

$$\begin{aligned} I_n(f_{i:n}, f) = I_n(g_{i:n}, g) - \int_0^{\infty} f_{i:n}(x) \log f(x) dx \\ = -\int_0^{\infty} g_{i:n}(y) \log g(y) dy - \int_0^{\infty} \frac{F(x)^{i-1}(1-F(x))^{n-i} f(x) \log f(x)\, dx}{B(i, n-i+1)} \\ = -\int_0^{\infty} \frac{G(y)^{i-1}(1-G(y))^{n-i} g(y) \log g(y) dy}{B(i, n-i+1)}. \end{aligned}$$

Put $u = 1 - F(x)$ and $u = 1 - G(y)$ and take $n - i = k$, then

$$\int_0^1 (1-u)^{i-1} \left[\log(f(F^{-1}(1-u))) - \log(g(G^{-1}(1-u)))\right] u^k du = 0, \quad \forall k \geqslant 0.$$

Using Lemma 2.1, we have

$$f(F^{-1}(1-u)) = g(G^{-1}(1-u))$$

Take $1 - u = v$, then

$$f(F^{-1}(v)) = g(G^{-1}(v)), \quad \forall v \in (0,1).$$

As,

$$\frac{d(F^{-1}(v))}{dv} = \frac{1}{f(F^{-1}(v))}.$$

Therefore, we have

$$\begin{aligned} F^{-1'}(v) = G^{-1'}(v), \quad \forall v \in (0,1) \\ F^{-1}(v) = G^{-1}(v) + c \end{aligned}$$

where c is a constant and hence concludes the proof. □

3. Properties of Inaccuracy Measure

In this section, we find the bounds of inaccuracy measure (2.3) for order statistics in terms of entropy (1.2). Also, we find the average value of the derived measure.

Theorem 3.1. *For any random variable X with entropy $H(X) < \infty$.*

(i) *If B_i is the i^{th} term of the binomial probability $B(n-1,p_i), p_i = \frac{i-1}{n-1}$, then*

$$nB_i(H(X)+I(A)) \leqslant I_n(f_{i:n},f) \leqslant nB_i[H(X)+I(\bar{A})] \tag{3.1}$$

where $I(A) = \int_A f(x)\log f(x)dx$ and $A = \{x; f(x) \leqslant 1\}$, $\bar{A} = \{x; f(x) > 1\}$.

(ii) *If $M = f(m) < \infty$, where m is the mode of the distribution, then*

$$-\log M \leqslant I_n(f_{i:n},f) \leqslant nB_i[H(X)+\log M] - \log M. \tag{3.2}$$

Proof. The entropy $H(X_{i:n})$ of i^{th} order statistics is bounded as, [8].

$$H_n(W_i) + nB_i(H(X)+I(A)) \leqslant H(X_{i:n}) \leqslant H_n(W_i) + nB_i[H(X)+I(\bar{A})] \tag{3.3}$$

where $H_n(W_i)$ is given by (1.9).
Adding (2.2) and (3.3), we get (3.1).
To prove (ii), we will use result due to Ebrahimi *et al.* (2004) given by

$$H_n(W_i) - \log M \leqslant H(X_{i:n}) \leqslant H_n(W_i) - \log M + nB_i[H(X)+\log M]. \tag{3.4}$$

Adding (2.2) and (3.4), we get (3.2). □

Example 3.1. Let X be a random variable following exponential distribution with p.d.f. $f(x) = \theta e^{-\theta x}$, $x \geqslant 0$, $\theta > 0$. Then, $F(x) = 1 - e^{-\theta x}$.
For $i = 1$, that is the case of sample minima, we have

$$I_n(f_{1:n},f) = -E_{g_1}[\log(f(F^{-1}(W_1)))] = \frac{1}{n} - \log\theta. \tag{3.5}$$

Note that

(i) For a fixed value of n, inaccuracy of sample minimum for exponential distribution decreases with increasing value of θ. Figure 1 shows decrease in inaccuracy for different values of n.
(ii) Similarly, if we keep θ fixed then inaccuracy decreases with increase in sample size. Figure 2 shows decrease in inaccuracy for different values of θ.

For $i = n$, that is the case of sample maxima

$$I_n(f_{n:n},f) = -E_{g_n}[\log(f(F^{-1}(W_n)))] = \gamma + \psi(n) - \log\theta + \frac{1}{n}. \tag{3.6}$$

where $\psi(1) = -\gamma = 0.5772$ is Euler's constant and we use $\psi(n+1) = \psi(n) + \frac{1}{n}$.
Note that

(i) For a fixed value of n, inaccuracy of sample maximum decreases with increasing value of parameter θ.
(ii) $I_n(f_{n:n},f) - I_n(f_{1:n},f) = \gamma + \psi(n) \geqslant 0$, equality holds when $n = 1$. Hence, for exponential distribution we can conclude that inaccuracy about the maximum is always more than the minimum.

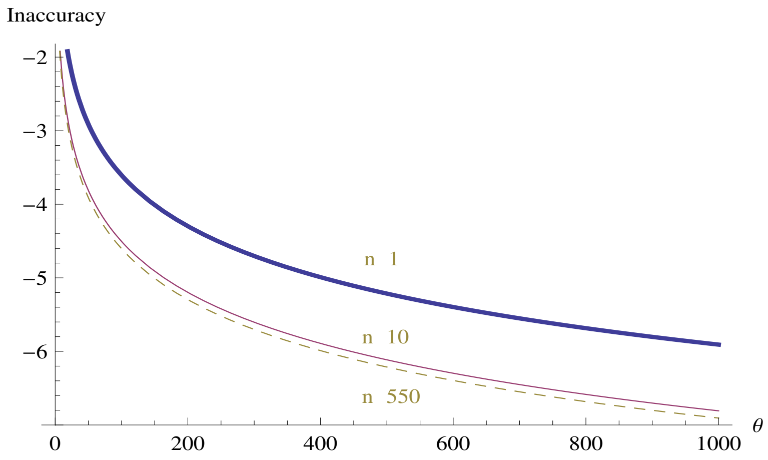

Fig. 1.

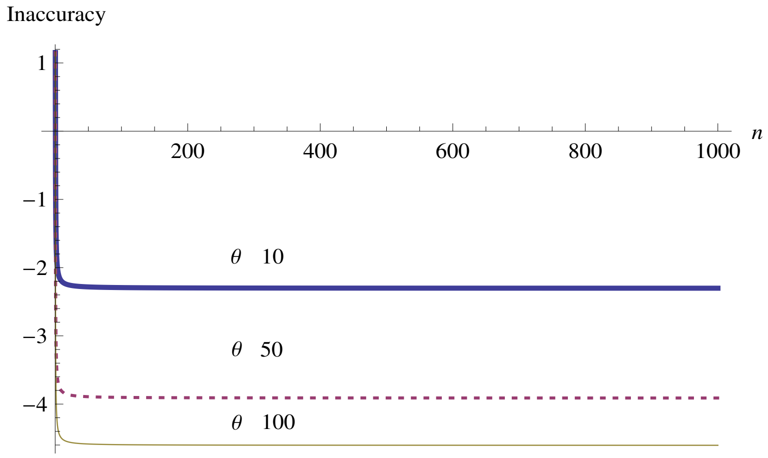

Fig. 2.

Remark 3.1. For exponential distribution with parameter θ we have $M = \theta$ and $H(X) = 1 - \log\theta$. Using (3.2), we have

$$-\log\theta \leqslant I_n(f_{i:n}, f) \leqslant nB_i - \log\theta\,. \tag{3.7}$$

For $i = 1$, (3.7) becomes

$$-\log\theta \leqslant I_n(f_{1:n}, f) \leqslant n - \log\theta\,. \tag{3.8}$$

where as

$$I_n(f_{1:n}, f) = \frac{1}{n} - \log\theta\,. \tag{3.9}$$

The difference between the actual value of $I_n(f_{1:n}, f)$ and the lower bound calculated in (3.8) is $\frac{1}{n}$ which tends to 0 as $n \to \infty$. Therefore, for exponential distribution, lower bound is useful when sample size is large.

Theorem 3.2. *The average value of inaccuracy measure is entropy of the parent random variable X, that is*

$$\frac{1}{n}\sum_{i=1}^{n} I_n(f_{i:n}, f) = H(X). \tag{3.10}$$

Proof. Consider

$$\begin{aligned}
-\sum_{i=1}^{n}\int f_{i:n}(y)\log f(y)dy &= -\sum_{i=1}^{n}\int \frac{1}{B(i,n-i+1)}(F(y))^{i-1}(1-F(y))^{n-i}f(y)\log f(y)dy \\
&= -\sum_{i=1}^{n}\int g_i(F(y))f(y)\log f(y)dy \\
&= -\int \sum_{i=1}^{n} nq_{i-1}f(y)\log f(y)dy \\
&= nH(X),
\end{aligned}$$

where

$$g_i(w) = \frac{1}{B(i,n-i+1)}w^{i-1}(1-w)^{n-i}, \quad 0 \leqslant w \leqslant 1,$$

is the p.d.f. of i^{th} order statistics from standard uniform distribution, and q_{i-1} with $\sum_{i=1}^{n} q_{i-1} = 1$ denotes the $(i-1)^{\text{th}}$ term of $B(n-1,p)$, the Binomial variate with parameters $(n-1)$ and $p = F(x)$. Hence, the desired result (3.10) follows. □

Example 3.2. Let X be a random variable having exponential distribution with p.d.f. $f(x) = \theta\, e^{-\theta x}$, $\theta > 0$, $x \geqslant 0$. Then,

$$f_{i:n}(y) = \frac{1}{B(i,n-i+1)}F(y)^{i-1}(1-F(y))^{n-i}f(y). \tag{3.11}$$

For $i = 1, 2$ and $n = 2$, using (2.4)

$$I_2(f_{1:2}, f) = -\log\theta - \frac{1}{2}$$

and

$$I_2(f_{2:2}, f) = -\log\theta + \frac{3}{2}.$$

Hence,

$$\frac{1}{2}\left(I_2(f_{1:2}, f) + I_2(f_{2:2}, f)\right) = 1 - \log\theta. \tag{3.12}$$

Also, using (1.2) we have

$$H(X) = 1 - \log\theta. \tag{3.13}$$

which is equal to average inaccuracy as calculated in (3.12).

4. Conclusion

The proposed measure of inaccuracy between the i^{th} order statistics and parent random variable characterizes the distribution function of parent random variable uniquely and its average value is the entropy of the parent random variable.

Acknowledgement

The first author is thankful to the Council for Scientific and Industrial Research, India, to provide financial assistance for this work.

References

[1] B.C. Arnold,N. Balakrishnan and H.N. Nagaraja. “A first Course in Order Statistics”, *John Wiley and Sons*, (1992).

[2] C.E. Shannon, “A mathematical theory of communication”, *Bell syst. Tech. J.*, **27**, 379–423 and 623–656, (1948).

[3] D.F. Kerridge, “Inaccuracy and Infrence”, *J. Roy. Statist. Soc. Ser.B.*, **23**, 184–194,(1961).

[4] H.A. David and H.N. Nagaraja, “Order Statistics”, *Wiley, New York*, (2003).

[5] K.M. Wong and S. Chen, “The entropy of ordered sequences and order statistics”, *IEEE Trans. Inform. Theor.*, **36**, 276–284, (1990).

[6] M. Ahsanullah, V.B. Nevzorov, M. Shakil, “An Introduction to Order Statistics”, *Atlantis Press, Paris, France*, (2013).

[7] N. Ebrahimi, “How to measure uncertainty in the residual lifetime distributions”, *Sankhya A*, **58**, 48–57, (1996).

[8] N. Ebrahimi, E.S. Soofi and H. Zahedi, “Information properties of order statistics and spacings”, *IEEE Trans. Inform. Theor.*, **50**, 177–183, (2004).

[9] N. Ebrahimi and S.N.U.A. Kirmani, “A measure of discrimination between two residual lifetime distributions and its applications”, *Ann. Inst. Statist. Math*, **48**, 257–265, (1996).

[10] S. Kullback, “Information theory and Statistics”, *Wiley, New York*, (1959).

[11] S. Park, “The entropy of consecutive order statistics”, *IEEE Trans. Inform. Theor.*, **41**, 2003–2007, (1995).

[12] U. Kamps, “Characterizations of distributions by recurrence relations and identities for moments of order statistics. In: N. Balakrishnan and C.R. Rao, eds.”, *Order Statistics: Theory and Methods. Handbook of Statistics*, **16**, 291–311, (1998).