The Korean Communications
in Statistics Vol. 5, No. 2, 1998
pp. 307-313

Graphical Study on the Entropy of Order Statistics[1]

Sangun Park[2]

Abstract

The entropy measure is considered to denote the uncertainty of order statistics filters and choose the length of consecutive order statistic filters. However, it needs much calculations to get the amount of the entropy of all possible sets of consecutive order statistics, and the results of those calculations return many numerical values. Thus we provide an efficient graphical presentation of those numerical values, which make it easy to understand the distribution of the entropy among order statistics.

1. Introduction

Order statistic filters have been considered for various purposes. Their wide applications in engineering have been noted in Wong and Chen (1990). For example, the median filter and the linear combination of order statistics have received high attentions, since they are robust to some possible outliers in the signal. In quality control, only some consecutive order statistics are available in the case of Type II censored samples. In these kinds of situations, we are primarily concerned with the loss of information or concentration due to employing only a part of samples, which may lead us to a wise selection of the length of consecutive order statistics.

Wong and Chen (1990) considered the entropy measure, which is a measure of uncertainty, and studied the calculation and presentation of the entropy in each individual order statistic. The study on the entropy of order statistics provides us the amount of uncertainty reduction or a guide to the choice of order statistic. Park (1995) extended the study to the simple calculation of the entropy in the consecutive order statistics case. However, the calculations for all possible sets of consecultive order statistics return too many numerical results, since the number of all possible sets of consecutive order statistics is $n(n+1)/2$ for a sample of size n, while the number of single order statistics alone is just n. Thus it is difficult to figure out the changes of the entropy directly from those numerical results.

In this paper, we provide what we call the entropy plot, a graphical presentation of the entropy of all possible sets of consecutive order statistics. The entropy plot reflects the entropy of all possible sets of consecutive order statistics in one plot. Thus we can read from

1) This work was supported by Yonsei University Research Fund of 1997
2) Assitant Professor, Dept. of Applied Statistics, Yonsei University, Seoul, 120-749, Korea

2. Preliminaries

Suppose that a random variable X has a distribution function F(x), with a density function f(x), a continuous positive function. According to Shannon (1948), the entropy of the random variable is defiend to be, $-\int_{-\infty}^{\infty} f(x)\log f(x)dx$. Let $X_{(r;n)}$ be the rth order statistic based on independently and identically distributed (i.i.d.) samples of size n from F(x). The joint density of $X_{(r;n)},\cdots,X_{(s;n)}$ can be written as

$$f_{r\cdots s;n}=\frac{n!}{(r-1)!\,(n-s)!}F(x_{(r;n)})^{r-1}f(x_{(r;n)})\cdots f(x_{(s;n)})(1-F(x_{(s;n)}))^{n-s}.$$

Then the entropy of a set of consecutive order statistics, $X_{(r;n)},\cdots,X_{(s;n)}$, is defined, in view of (1), to be

$$H_{r\cdots s;n}=-\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{x_{(r+1;n)}}\log f_{r\cdots s;n}dF_{r\cdots s;n}$$

where $F_{r\cdots s;n}$ is the joint distribution of $X_{(r;n)},\cdots,X_{(s;n)}$.

It has been shown in (2) that $H_{r\cdots s;n}$ can be obtained as $H_{1\cdots s;n}+H_{r\cdots n;n}-H_{1\cdots n;n}$, where $H_{1\cdots r;n}$and $H_{r\cdots n;n}$ can be calculated as a linear function of $H_{1;i}$ and $H_{i;i}$, $i=1,\cdots,n$ as follow.

$$H_{1\cdots r;n} = \sum_{i=n-r+1}^{n} C_{i-2,n-r-1}C_{n,i}(-1)^{i-n+r-1}(H_{1;i}+\log i)+\log(n-r)!-\log n!$$

$$H_{r\cdots n;n} = \sum_{i=r}^{n} C_{i-2,r-2}C_{n,i}(-1)^{i-r}(H_{i;i}+\log i)+\log(r-1)!-\log n!$$

Let $H_{r+1\cdots n|r;n}(x_{(r;n)})$ be the conditional entropy of $X_{(r+1;n)},\cdots,X_{(n;n)}$ given $X_{(r;n)}=x_{(r;n)}$, which can be written as

$$H_{r+1\cdots n|r;n}(x_{(r;n)})=E(-\log f_{r+1\cdots n|r;n}|X_{(r;n)}=x_{(r;n)})$$

where $f_{r+1\cdots n|r;n}$ is the joint conditional density of $X_{(r+1;n)},\cdots,X_{(n;n)}$ given $X_{(r;n)}=x_{(r;n)}$. We define $H_{r+1\cdots n|r;n}$ to be the expectation of $H_{r+1\cdots n|r;n}(x_{(r;n)})$ about $X_{(r;n)}$. Then we can extend the decomposition of the entropy of order statistics in [2, Lemma 2.2], which provides a basic idea in the entropy plot described in the next.

Lemma 2.1

$$H_{1\cdots n;n}=H_{1\cdots r-1|r;n}+H_{r\cdots s;n}+H_{s+1\cdots n|r;n}$$

3. Entropy plot

We will denote A(s) to be $H_{1\cdots s;n}$, which is $H_{1\cdots r-1|r;n}+H_{r\cdots s;n}$, and B(r) to be $H_{1\cdots r-1|r;n}$.

Then $H_{r\cdots s;n}$ can be represented to be A(s)-B(r) in view of Lemma 2.1. Thus if we calculate A(r)'s and B(r)'s for $r=1,\cdots,n$, we can automatically calculate $H_{r\cdots s;n}$'s for $r, s=1,\cdots,n$. In calculating A(r) and B(r), it is enough to calculate $H_{1\cdots r;n}$ and $H_{r\cdots n;n}$, since

$$\begin{aligned} A(r) &= H_{1\cdots r;n} \\ B(r) &= H_{1\cdots n;n} - H_{r\cdots n;n} \end{aligned}$$

In Figure 1, we provide two entropy plots of the normal distributions, based on the above calculations, whose scales are 0.1, 10 for the samples of size 7. The amount of the entropy can be read from the plot as follow;

$$H_{r\cdots s;n} = A(s) - B(r)$$

For example, to read $H_{2\cdots 6;7}$ from (a) in figure 1, we need to read A(6) - B(2). In figure 1, A(6) is about -13, and B(2) is about -2. Thus $H_{2\cdots 6;7}$ can be approxmately read to be -11, while the exact value is -10.9778. Like that, we can read the entropy of any set of consecutive order statistics from the entropy plot. We can instantly see that the amount of the entropy decreases by adding an additional order statistic to any set of consecutive order statistics in the case (a) in figure 1, while we can see that the amount of the entropy increases by adding an additional order statistic to any set of consecutive order in the case (b) in figure 1. However, the entropy of the sample median is smallest for both cases.

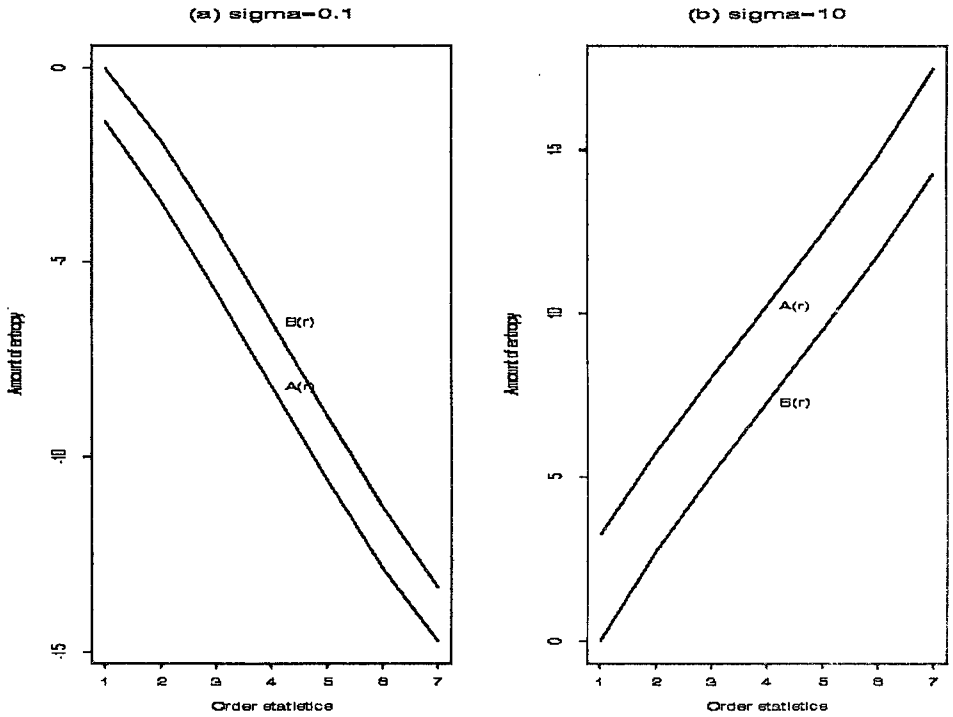

Figure 1. Entropy plots of the normal distribution (n = 7: sigma = 0.1, 10)

4. Examples

We here study the entropy of order statistics with the entropy plot described in the previous section for some well-known parametric distribution functions. The entropy of order statistics depends on the scale parameter and the sample size, but is invariant about the location parameter. Generally, it is expected that the small value of the scale parameter reduces the uncertainty, and the large number of the sample size also reduces the uncertainty.

Figure 2 shows the change of the entropy of order statistics of the normal distribution according to the sample sizes. The length of consecutive order statistics which has the smallest amount of entropy gets larger, as the sample size gets larger. However, the entropy of all order statistics is not smallest.

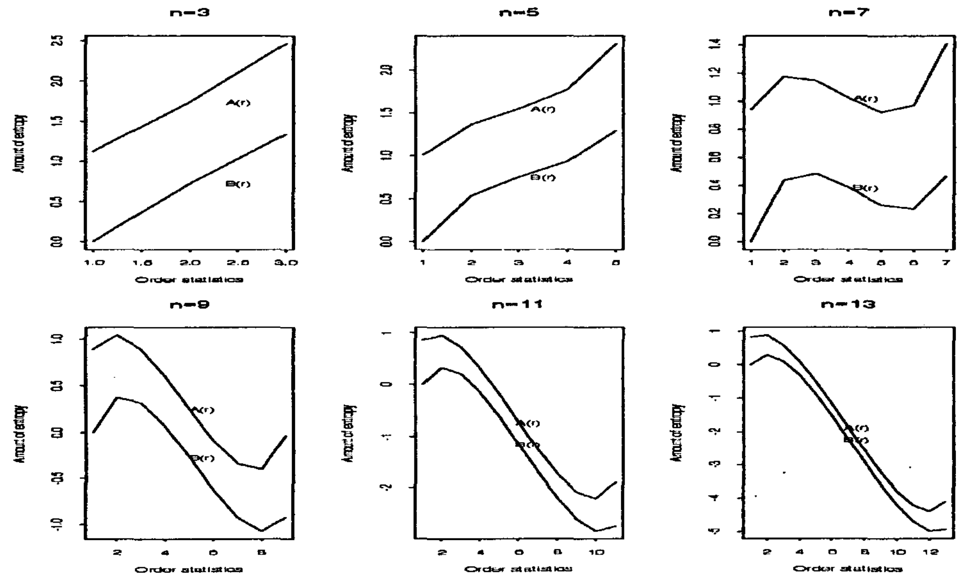

Figure 2. Change of the entropy in normal distribution according to sample sizes

Figure 3 shows the change of the entropy of order statistics of the normal distirbutions according to the values of the scale parameter. The length of consecutive order statistics which has the smallest amount of entropy gets larger, as the scale parameter gets smaller.

We also provide the entropy plots for the exponential and uniform distributions. For the exponential case, the entropy plot is not symmetric. It is noted that the consecutive order statistics on the left has the smaller amount of entropy than those on the right. For the unform case, the order statistics on both ends have smaller amount of entropy than those in the center.

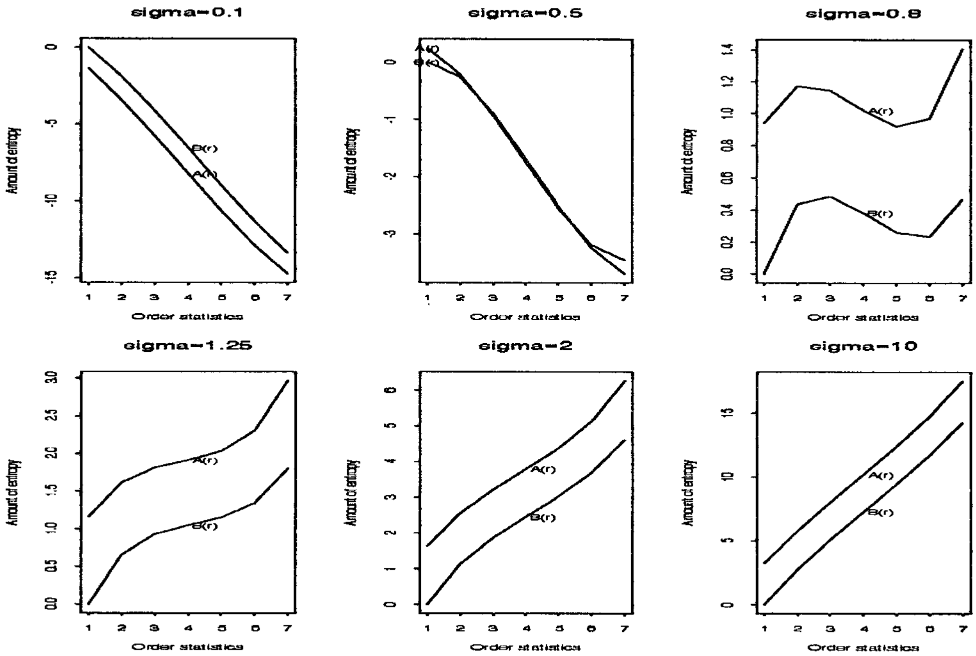

Figure 3. Change of the entropy in normal distribution according to the scale

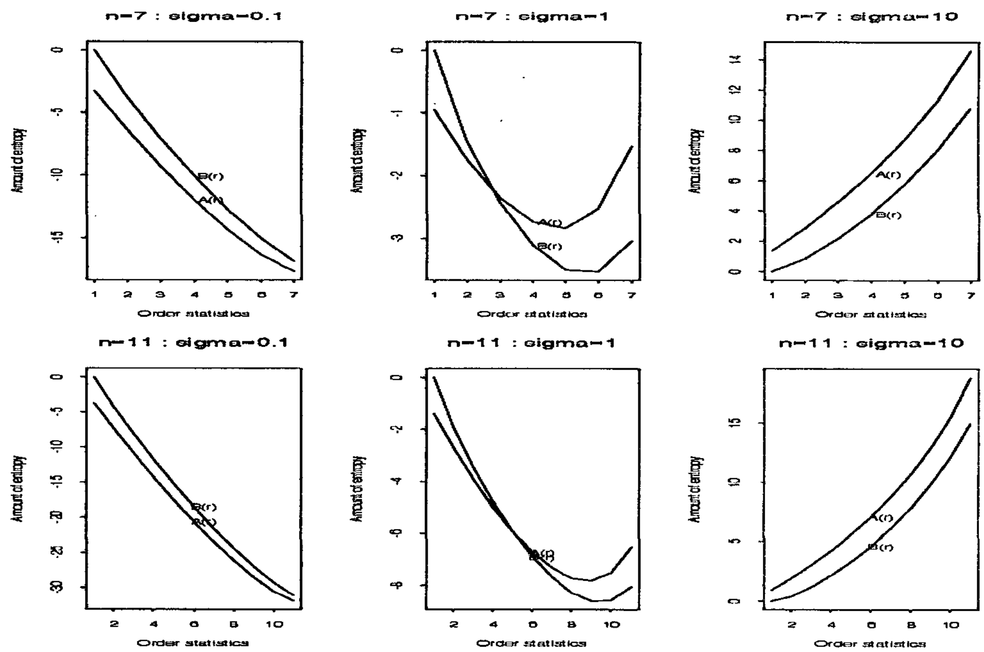

Figure 4. Entropy plot : Exponential distribution $(\frac{1}{\sigma}e^{-\frac{x}{\sigma}}, x\geq 0)$

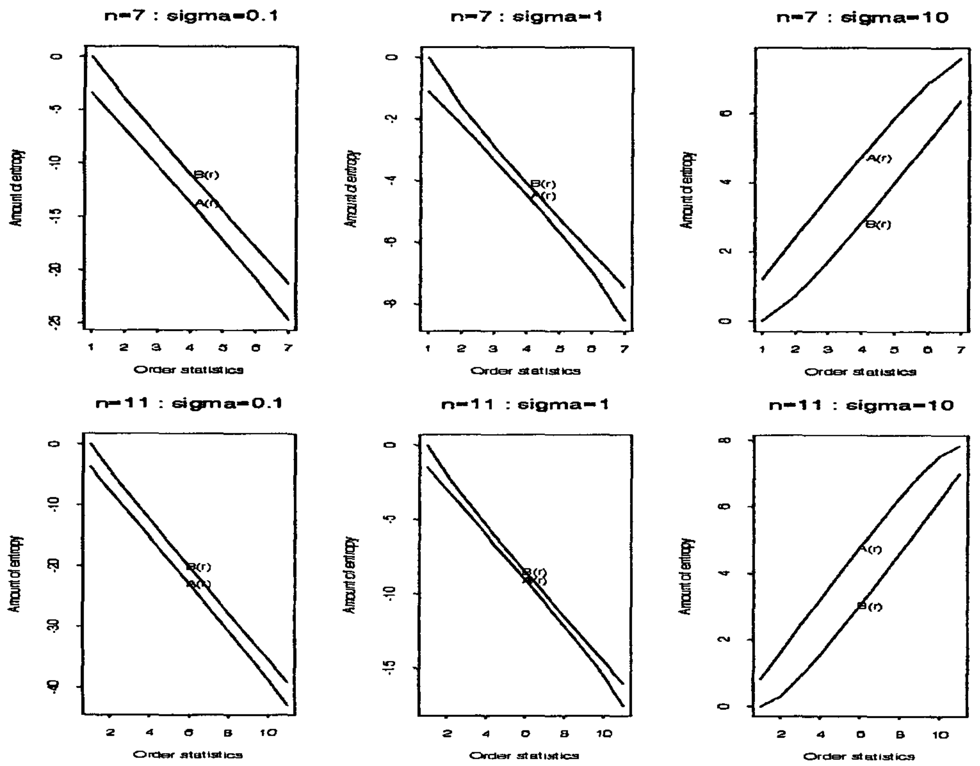

Figure 5. Entropy plot : Uniform distribution $(\frac{1}{\sigma}, 0 \leq x \leq \sigma)$

5. Conclusions

We have observed that the entropy plot provides a simple presentation of the entropies of consecutive order statistics, which enables us to read the entropy of any set of consecutive order statistics. Thus the entropy plot is an efficient way of representing the entropy of order statistics. Since the entropy is a measure of uncertainty, we prefer a set of order statistics or a sample size which has a minimum amount of entropy. Thus the entropy plot can be basically used, in the study of the selection of consecutive order statistics or the sample size, for investigating how the sample size or the parameter value affects the distribution of the entropy among order statistics for an assumed parametric probability model. The application of the entropy of order statistics has been done in Wong and Chen (1990) to study the selection of order statistics and in Brooks (1982), Turrero (1989), and Ebrahimi and Soofi (1990) to study the loss of information due to censoring.

Reference

[1] R. J. Brooks (1982). On the loss of information through censoring, *Biometrika*, Vol. 69, pp 137-144.

[2] H. A. David (1981). *Order statistics*, 2nd ed. New York : Wiley.

[3] N. Ebrahimi and E. S. Soofi (1990). Relative information loss under Type II censored exponential data, *Biometrika*, Vol. 77, pp 429-435.

[4] S. Park (1995). The entropy of consecutive order statistics, *IEEE Trans. Inform. Theory,* Vol. 41, pp. 2003-2007, Nov.

[5] C. E. Shannon (1948). A mathematical theory of communication, *Bell Syst. Tech. J.*, Vol. 27, 379-423, 623-656.

[6] Turrero, A. (1989). On the relative efficiency of grouped and censored survival data, *Biometrika*, Vol. 76, pp 125-131.

[7] K. M. Wong and S. Chan (1990). The entropy of ordered sequences and order statistics, *IEEE Trans. Inform. Theory,* Vol. 36, 276-284, Mar.