JIRSS (2014)
Vol. 13, No. 1, pp 31-42

Characterizations Using Entropies of Records in a Geometric Random Record Model

M. Fashandi, A. Khosravi, Jafar Ahmadi

Department of Statistics, Ordered and Spatial Data Center of Excellence, Ferdowsi University of Mashhad, Iran.

Abstract. Suppose that a geometrically distributed number of observations are available from an absolutely continuous distribution function F, within this set of observations denote the random number of records by M. This is called geometric random record model. In this paper, characterizations of F are provided in terms of the subsequences entropies of records conditional on events $\{M \geq n\}$ or $\{M = n\}$ in a geometric random record model. Characterization results for symmetric distributions are also presented based on entropies of upper and lower records in a random record model.

Keywords. Geometric distribution, Müntz-Szász Theorem, records, Shannon information, symmetric distribution.

MSC: 62E10, 62G30.

1 Introduction

Let $\{X_i,\ i \geq 1\}$ be a sequence of independent and identically distributed (iid) random variables with an absolutely continuous cumulative distribution function (cdf) $F(x)$ and probability density function (pdf) $f(x)$. An observation X_j is called an *upper record value* if it exceeds all previous observations, i.e., X_j is an *upper record* if $X_j > X_i$ for every $i < j$.

M. Fashandi(fashandi@um.ac.ir), A. Khosravi(khosravi.a_66@yahoo.com), Jafar Ahmadi(✉)(ahmadi-j@um.ac.ir)
Received: April 2013; Accepted: November 2013

An analogous definition can be given for *lower record values*. For notational convenience, we shall denote the i-th upper record and the j-th lower record values by U_i and L_j, respectively. The zero-th upper and lower records are set as $U_0 = L_0 \equiv X_1$, which is referred to as the reference value or the trivial record. Interested readers may refer to the book by Arnold *et al.* (1998) and the references contained therein for an elaborate treatment on theory, methods and applications of record values. In some situation, instead of assuming the availability of an infinite sequence of observations, we have to consider a sequence $X_1, X_2, ..., X_N$ where N is a positive integer-valued random variable independent of the X_i-sequence. Such a situation arises, for instance, when the observations arrive at time points determined by an independent point process observed over a finite time, see Arnold *et al.* (1998, Chapter 7) and also see job search models in labor economics considered by Nagaraja and Barlevy (2003). Then, record values are observed from the sequence with random length $X_1, X_2, ..., X_N$, in this case the model is called *random record model*. We refer the reader to Arnold et al. (1998, Chapter 7) for more pertinent details on random record models. Let X be a random variable having an absolutely continuous cdf F_X with pdf f_X, and M denote the number of non-trivial records. If N has a geometric probability mass function (pmf) with parameter p, i.e.

$$P(N = k) = q^{k-1}p, \quad k = 1, 2, ..., \ 0 < p < 1, \ q = 1 - p, \tag{1}$$

then the joint likelihood of $U_0, U_1, ..., U_n$ and the event $\{M \geq n\}$ is given by

$$f(u_0, u_1, ..., u_n; M \geq n) = f_X(u_n) \prod_{i=0}^{n-1} \frac{q f_X(u_i)}{1 - q F_X(u_i)}, \quad u_0 < u_1 < ... < u_n, \tag{2}$$

where $q = 1 - p$, see Arnold et al. (1998, p. 229). It may be noted that when N is infinite with probability one, then the classical record model is deduced.

The problem of characterizing a probability distribution based on record values in the context of classical record model have been studied by several researchers. For example, see Chapter 4 of Arnold et al. (1998) and references therein which contain most of the results on characterizations based on record values in classical model up to 1998. For recent works, see for example, Balakrishnan and Stepanov (2004), Gupta and Ahsanullah (2004) and Su et al. (2008). In the context of random record model, Nagaraja and Barlevy (2003) proved that appropriately chosen

subsequences of $E(U_n|M \geq n)$ or $E(U_n|M = n)$ characterize F, for N with a geometric random variable with pmf (1). In this paper, we intend to develop their results in terms of Shannon *entropy* of record statistics, it was originally introduced by Shannon (1948). Let X be a random variable having an absolutely continuous cdf F_X with pdf f_X, then the basic uncertainty measure for distribution F_X is defined as

$$H(X) = -\int f_X(x) \log f_X(x)\ dx, \tag{3}$$

provided the integral exists. In the literature, $H(X)$ is often referred to as the entropy of X or Shannon information about F_X. We refer the reader to Cover and Thomas (1991) for more details and references therein. Entropy properties of record values have been studied by several authors. See, for example, Baratpour et al. (2007a) and Razmkhah et al. (2012). Recently, some works have been done in the subject of characterization based on Shannon's entropy of records for the classical record model, see, Baratpour et al. (2007b), Raqab and Awad (2000, 2001), Ahmadi and Fashandi (2009) and Ahmadi (2009). No previous work has been done on entropy properties of records in random record model.

The rest of this paper is organized as follows. Section 2 contains some preliminaries, the basic definitions and calculations which will be used in the next sections. In Section 3, we prove that appropriately chosen subsequences of $H(U_n|M \geq n)$ or $H(U_n|M = n)$ characterize F up to a location shift. In Section 4, we obtain characterization results for symmetric distributions.

2 Basic Tools

In this section, we present some preliminaries and basic tools to establish the new characterization results. From (2) and also equations (6) and (9) in Nagaraja and Barlevy (2003), it follows that

$$f(u_n; M \geq n) = \frac{(-\log(1 - qF_X(u_n)))^n}{n!} f_X(u_n), \tag{4}$$

and

$$f(u_n; M = n) = \frac{p}{1 - qF_X(u_n)} \frac{(-\log(1 - qF_X(u_n)))^n}{n!} f_X(u_n). \tag{5}$$

Using (4), for a geometric random record model, we derive a relation between the entropy of records from an arbitrary absolutely continuous cdf F and the entropy of records from standard uniform distribution, which are stated in the next lemma.

Lemma 2.1. *Let U_n and L_n be the n-th upper and lower records, respectively, in a geometric random record model. Then we have:*

$$H(U_n|M \geq n) = H(U_n^*|M \geq n) - E\left[\log(f_X(F_X^{-1}(U_n^*)))|M \geq n\right] \quad (6)$$

and

$$\begin{aligned} & H(L_n|M \geq n) \\ &= H(L_n^*|M \geq n) - E\left[\log(f_X(F_X^{-1}(L_n^*)))|M \geq n\right] \\ &= H(U_n^*|M \geq n) - E\left[\log(f_X(F_X^{-1}(1-U_n^*)))|M \geq n\right], \end{aligned} \quad (7)$$

where U_n^* and L_n^* stand for the n-th upper and the n-th lower records from uniform distribution, respectively.

Proof. From (4) and (3), we have

$$\begin{aligned} H(U_n|M \geq n) &= -\int f_n(u|M \geq n) \log f_n(u|M \geq n)\, du \\ &= -\int \frac{(-\log(1-qF_X(x)))^n}{n!P(M \geq n)} f_X(x) \\ &\quad \times \log\left(\frac{(-\log(1-qF_X(x)))^n}{n!P(M \geq n)} f_X(x)\right) dx \\ &= -\int_0^1 \frac{(-\log(1-qu))^n}{n!P(M \geq n)} \log\left(\frac{(-\log(1-qu))^n}{n!P(M \geq n)}\right) du \\ &\quad -\int_0^1 \frac{(-\log(1-qu))^n}{n!P(M \geq n)} \log\left(f_X(F_X^{-1}(u))\right) du. \end{aligned}$$

The above expression is equal to the right hand side of (6). Similarly, for lower record values, we have

$$\begin{aligned} H(L_n|M \geq n) &= -\int \frac{(-\log(1-q\bar{F}_X(x)))^n}{n!P(M \geq n)} f_X(x) \\ &\quad \times \log\left(\frac{(-\log(1-q\bar{F}_X(x)))^n}{n!P(M \geq n)} f_X(x)\right) dx \end{aligned}$$

$$
\begin{aligned}
&= -\int_0^1 \frac{(-\log(1-q(1-u)))^n}{n!P(M \geq n)} \\
&\quad \times \log\left(\frac{(-\log(1-q(1-u)))^n}{n!P(M \geq n)}\right) du \\
&\quad -\int_0^1 \frac{(-\log(1-q(1-u)))^n}{n!P(M \geq n)} \log\left(f_X(F_X^{-1}(u))\right) du \\
&= H(L_n^*|M \geq n) - E\left[\log(f_X(F_X^{-1}(L_n^*)))|M \geq n\right].
\end{aligned}
$$

The second equality in (7) follows by the use of the fact that U_n^* is identical in distribution with $1 - L_n^*$ and the location shift does't change the entropy. □

We will use the completeness property to obtain the results, so let us recall the basic definition of a complete sequence.

Definition 2.1. A sequence $\{\phi_n\}_{n\geq 1}$ in a Hilbert space $\mathcal{H}$ is called *complete* if the only element of $\mathcal{H}$ which is orthogonal to every ϕ_n is the null element, that is

$$< f, \phi_n >= 0, (n \geq 1) \Rightarrow f = o,$$

here o stands for the zero element of $\mathcal{H}$.

Now, we recall the following theorem, which is well-known as Müntz-Szász Theorem and is used in the proofs of the results in this paper.

Theorem 2.1. *(Higgins, 2004, pp. 95-96) The set* $\{x^{\lambda_1}, x^{\lambda_2}, ... : 1 \leq \lambda_1 < \lambda_2 < ...\}$ *forms a complete sequence in* $L^2(0,1)$ *if and only if*

$$\sum_{j=1}^{+\infty} \lambda_j^{-1} = +\infty, \text{ where } 1 \leq \lambda_1 < \lambda_2 < \cdots . \tag{8}$$

See for example, Higgins (2004) and Michel (2013) for more details about complete sequences in the Hilbert space. In what follow, we provide some characterization results.

3 Characterization results

Nagaraja and Barlevy (2003) proved that in the geometric random record model the sequence $E(U_{n_j}|M \geq n_j)$, where $\sum_j n_j^{-1} = \infty$, characterizes F in the family of continuous distributions. Here, we show

that similar characterizations hold based on entropies of records in a geometric random record model.

Theorem 3.1. *Let $X_1, X_2, ..., X_N$ be a sequence of iid random variables from continuous cdf $F_X(x)$ and pdf $f_X(x)$, and N be a geometric random variable independent of X_i-sequence with pmf (1). Moreover, denote the number of upper observed non-trivial records in $X_1, X_2, ..., X_N$ by M. Then, the sequence $H(U_{n_j}|M \geq n_j)$, where $\sum_j n_j^{-1} = \infty$, with $n_1 < n_2 < ...$, characterizes F_X in the family of continuous distributions but for a location shift.*

Proof. Let $Y_1, Y_2, ..., Y_N$ be a sequence of iid random variables from continuous cdf G_Y and pdf g_Y. Suppose that for two sequences X_1, X_2, $..., X_N$ and $Y_1, Y_2, ..., Y_N$, we have

$$H(U_n^X|M \geq n) = H(U_n^Y|M \geq n), \tag{9}$$

where U_n^X and U_n^Y stand for the n-th upper record from X- and Y-sequences, respectively. Then, using Lemma 2.1 and by (6) and (9) it is deduced that

$$E\left[\log(f(F^{-1}(U_n^*)))|M \geq n\right] = E\left[\log(g(G^{-1}(U_n^*)))|M \geq n\right]. \tag{10}$$

Using (4), the identity (10) is equivalent to saying that

$$\begin{aligned}\int_0^1 \frac{(-\log(1-qu))^n}{n!P(M \geq n)} \log\left(f(F^{-1}(u))\right) du \\ = \int_0^1 \frac{(-\log(1-qu))^n}{n!P(M \geq n)} \log\left(g(G^{-1}(u))\right) du.\end{aligned} \tag{11}$$

Then, from (11), we have

$$\int_0^1 (-\log(1-qu))^n \left[\log\left(f(F^{-1}(u))\right) - \log\left(g(G^{-1}(u))\right)\right] du = 0. \tag{12}$$

By taking $z = \frac{\log(1-qu)}{\log p}$, the equation (12) can be rewritten as

$$\int_0^1 \left[\log\left(f\left(F^{-1}\left(\frac{1-p^z}{q}\right)\right)\right) - \log\left(g\left(G^{-1}\left(\frac{1-p^z}{q}\right)\right)\right)\right] p^z z^n \, dz = 0. \tag{13}$$

If (13) holds for any increasing sequence $\{n = n_j, \, j \geq 1\}$, such that $\sum_j n_j^{-1} = \infty$, then by appealing the Müntz-Szász Theorem, see Theorem 2.1, it follows that

$$f\left(F^{-1}\left(\frac{1-p^z}{q}\right)\right) - g\left(G^{-1}\left(\frac{1-p^z}{q}\right)\right) = 0, \tag{14}$$

for almost all $z \in (0,1)$. As in Nagaraja and Barlevy (2003), the identity (14) is equivalent to

$$f(F^{-1}(u)) - g(G^{-1}(u)) = 0, \quad \text{for almost every} \quad u \in (0,1). \tag{15}$$

From (15), it follows that

$$F^{-1}(u) = G^{-1}(u) + c,$$

where c is a constant. This means F and G belong to the same family of distributions, but for a location shift. □

For the classical record model, it may be mentioned that Baratpour et al. (2007) proved that X and Y have the same distribution with common lower boundary, if and only if $H(U_n^X) = H(U_n^Y)$, for all $n \geq 1$. Nagaraja and Barlevy (2003) also proved that in the geometric random record model the sequence $E(U_{n_j}|M = n_j)$, where $\sum_j n_j^{-1} = \infty$, characterizes F. Here, we show that similar characterizations hold based on entropy.

Theorem 3.2. *By the assumptions of Theorem 3.1, the sequence* $H(U_{n_j}^X|M = n_j)$, *where* $\{n_j, j \geq 1\}$ *satisfied in* (8), *characterizes* F_X *in the family of continuous distributions but for a location shift.*

Proof. To prove the Theorem, first from (5) we find

$$\begin{aligned} H(U_n|M=n) &= -\int_0^1 \frac{p(-\log(1-qu))^n}{n!P(M=n)(1-qu)} \\ &\quad \times \log\left(\frac{p(-\log(1-qu))^n}{n!P(M=n)(1-qu)}\right) du \\ &\quad -\int_0^1 \frac{p(-\log(1-qu))^n}{n!P(M=n)(1-qu)} \log\left(f_X(F_X^{-1}(u))\right) du \\ &= H(U_n^*|M=n) \\ &\quad -E\left[\log(f_X(F_X^{-1}(U_n^*)))|M=n\right]. \end{aligned} \tag{16}$$

Next, suppose for two cdfs F and G

$$H(U_n^X|M=n) = H(U_n^Y|M=n). \tag{17}$$

Then, from (16) and (17), we have

$$\int_0^1 \frac{(-\log(1-qu))^n}{1-qu} \left\{\log\left(f(F^{-1}(u))\right) - \log\left(g(G^{-1}(u))\right)\right\} du = 0. \tag{18}$$

Similar to the proof of Theorem 3.1, let $z = \frac{\log(1-qu)}{\log p}$, then from equation (18), we have

$$\int_0^1 \left[\log\left(f\left(F^{-1}\left(\frac{1-p^z}{q}\right)\right)\right) - \log\left(g\left(G^{-1}\left(\frac{1-p^z}{q}\right)\right)\right)\right] z^n \, du = 0.$$

By proceeding as in the proof of Theorem 3.1, the required result follows. □

Results similar to those in Theorems 3.1 and 3.2 hold in terms of lower records, which are stated in the next result.

Theorem 3.3. *Under the assumptions of Theorem 3.1, either of the sequences*
(i) $H(L_{n_j}^X | M \geq n_j)$ *and*
(ii) $H(L_{n_j}^X | M = n_j)$,
where $\sum_j n_j^{-1} = \infty$, *with* $n_1 < n_2 < ...$, *characterizes* F_X *in the family of continuous distributions but for a location shift.*

Proof. (i) The result follows from equation (7) and proceeding similarly to the proof of Theorem 3.1. To prove part (ii), first note that, we have

$$\begin{aligned} H(L_n|M=n) &= H(L_n^*|M=n) - E\left[\log(f_X(F_X^{-1}(L_n^*)))|M=n\right] \\ &= H(U_n^*|M=n) \\ &\quad -E\left[\log(f_X(F_X^{-1}(1-U_n^*)))|M=n\right]. \end{aligned} \tag{19}$$

The rest of proof is similar to the proof of Theorem 3.2. □

4 Characterization of Symmetric Distributions

For the classical record model, Fashandi and Ahmadi (2012) have proved that the equality of the entropy of upper and lower records is a characteristic property of symmetric distributions. In this section, we show that the same result holds for the the geometric random record model. First, we recall the following lemma.

Lemma 4.1. (Fashandi and Ahmadi, 2012) *Let* X *be a continuous random variable with cdf* F_X *and pdf* f_X *with support* S_X. *Then, the identity*

$$f_X(F_X^{-1}(u)) = f_X(F_X^{-1}(1-u)), \quad \textit{for almost all } u \in (0,1), \tag{20}$$

implies that there exists a constant c such that $F_X(c-x) = 1-F_X(c+x)$ *for all* $x \in S_X$.

The identity (20) is equivalent to the fact that F_X is symmetric about c.

Theorem 4.1. *Suppose the assumptions of Theorem 3.1 hold, in addition assume that* M_1 *and* M_2 *are the number of upper and lower observed nontrivial records in* $X_1, X_2, ..., X_N$. *Then, the following two statements are equivalent:*
(i) X *has a symmetric distribution;*
(ii) $H(U_{n_j}|M_1 \geq n_j) = H(L_{n_j}|M_2 \geq n_j)$, *where the sequence* $\{n_j, j \geq 1\}$ *satisfied in* (8).

Proof. When X has a symmetric distribution about zero (without loss of generality), then it is obvious that U_n is identical in distribution with $-L_n$. Thus, (i) implies (ii). We shall prove that also (ii) $\Rightarrow$ (i) holds. From (6) and (7)

$$\begin{aligned} & H(U_n|M \geq n) - H(L_n|M \geq n) \\ &= H(U_n^*|M \geq n) - E\left[\log(f_X(F_X^{-1}(U_n^*)))|M \geq n\right] \\ &\quad -H(L_n^*|M \geq n) + E\left[\log(f_X(F_X^{-1}(L_n^*)))|M \geq n\right] \\ &= E\left[\log(f_X(F_X^{-1}(1-U_n^*)))|M \geq n\right] \\ &\quad -E\left[\log(f_X(F_X^{-1}(U_n^*)))|M \geq n\right]. \end{aligned} \tag{21}$$

So by (21), Part (ii) is equivalent to

$$E\left[\log(f_X(F_X^{-1}(U_n^*)))|M \geq n\right] - E\left[\log(f_X(F_X^{-1}(1-U_n^*)))|M \geq n\right] = 0.$$

Then from (4), we find

$$\int_0^1 (-\log(1-qu))^n \left[\log\left(f(F^{-1}(u))\right) - \log\left(f(F^{-1}(1-u))\right)\right] du = 0. \tag{22}$$

As in the proof of Theorem 3.1, by taking $z = \frac{\log(1-qu)}{\log p}$, the identity (22) can be written as

$$\int_0^1 \left[\log\left(f\left(F^{-1}\left(\frac{1-p^z}{q}\right)\right)\right) - \log\left(f\left(F^{-1}\left(1-\frac{1-p^z}{q}\right)\right)\right)\right] p^z z^n \, dz = 0. \tag{23}$$

By appealing the Müntz-Szász Theorem, if (23) holds for any increasing sequence $\{n_j, j \geq 1\}$, such that $\sum_j n_j^{-1} = \infty$, then,

$$f\left(F^{-1}\left(\frac{1-p^z}{q}\right)\right) - f\left(F^{-1}\left(1-\frac{1-p^z}{q}\right)\right) = 0,$$

for almost every $z \in (0,1)$ or

$$f\left(F^{-1}(u)\right) - f\left(F^{-1}(1-u)\right) = 0, \quad \text{for almost every } u \in (0,1). \quad (24)$$

Thus, by Lemma 4.1 the proof is completed. □

The equality $H(U_{n_j}|M_1 = n_j) = H(L_{n_j}|M_2 = n_j)$ is also a characteristic property of symmetric distributions which is stated in the following theorem. The proof is similar to that of Theorem 4.1 and is omitted for the sake of brevity.

Theorem 4.2. *Suppose the conditions of Theorem 4.1 hold, then, the following two statements are equivalent:*
(i) X has a symmetric distribution;
(ii) $H(U_{n_j}|M_1 = n_j) = H(L_{n_j}|M_2 = n_j)$, where $\{n_j, j \geq 1\}$ satisfied in (8).

5 Summary

It is well-known that characterization problems in mathematical statistics are statements in which the description of possible distributions of random variables follows from properties of some functions in these variables. In this work, we have tackled the characterization problems of distribution F in terms of the subsequences entropies of records conditional on two different events in a geometric random record model. We proved that the equality of entropy of the sequence of upper and lower record values in the setup of random record model implies the symmetry of the parent distribution. The obtained results are useful in testing goodness-of-fit and symmetry. Because, as pointed out in Chapter 4 of Arnold et al. (1998), a characterization can be of use in the construction of goodness-of-fit tests and in the examination of the consequences of modeling assumptions made by an applied scientist.

Acknowledgements

The authors would like to thank the reviewer for his valuable comments and suggestions to improve the presentation of the paper.

References

Ahmadi, J. (2009), Entropy properties of certain record statistics and some characterization results. Journal of the Iranian Statistical Society (JIRSS), **8**, 1–13.

Ahmadi, J. and Fashandi, M. (2009), Some characterization and ordering results based on entropies of current records. Statistics & Probability Letters, **79**, 2053–2059.

Arnold, B. C., Balakrishnan, N., and Nagaraja, H. N. (1998), Records. New York: John Wiley & Sons.

Balakrishnan, N. and Stepanov, A. (2004), Two characterizations based on order statistics and records. Journal of Statistical Planning and Inference, **124**, 273–287.

Baratpour, S., Ahmadi, J., and Arghami, N. R. (2007a), Entropy properties of record statistics. Statistical Papers, **48**, 197–213.

Baratpour, S., Ahmadi, J., and Arghami, N. R. (2007b), Some characterizations based on entropy of order statistics and record values. Communications in Statistics-Theory and Methods, **36**, 47–57.

Cover, T. M. and Thomas, J. A. (1991), Elements of Information Theory. A Wiley-Interscience Publication, New York: Wiley.

Gupta, R. C. and Ahsanullah, M. (2004), Some characterization results based on the conditional expectation of a function of non-adjacent order statistic (record value). Annals of the Institute of Statistical Mathematics, **56**, 721–732.

Higgins, J. R. (2004), Completeness and Basis Properties of Sets of Special Functions. New York: Cambridge University Press.

Michel, V. (2013), Lectures on Constructive Approximation, Fourier, Spline, and Wavelet Methods on the Real Line, the Sphere, and the Ball. New York: Birkhäuser.

Nagaraja, H. N. and Barlevy, G. (2003), Characterizations using record moments in a random record model and applications. Journal of Applied Probability, **40**, 826–833.

Raqab, M. Z. and Awad, A. M. (2000), Characterizations of the Pareto and related distributions. Metrika, **52**, 63–67.

Raqab, M. Z. and Awad, A. M. (2001), A note on characterization based on Shannon entropy of record statistics. Statistics, **35**, 411–413.

Razmkhah, M., Morabbi, H. and Ahmadi, J. (2012), Comparing two sampling schemes based on entropy of record statistics. Statistical Papers, **53**, 95–106.

Shannon, C. E. (1948), A mathematical theory of communication. Bell System Technical Journal, **27**, 379–423.

Su, J.-C., Su, N.-C., and Huang, W.-J. (2008), Characterizations based on record values and order statistics. Journal of Statistical Planning and Inference, **138**, 1358–1367.