Sociedad de Estadística e Investigación Operativa
Test (2001) *Vol.* 10, *No.* 1, pp. 121–131

Discrete distributions for which the regression of the first record on the second is linear

Fernando López-Blázquez*
Departamento de Estadística e Investigación Operativa
Universidad de Sevilla, Spain
Jacek Wesołowski
Wydział Matematyki i Nauk Informacyjnych
Politechnika Warszawska, Poland

Abstract

The linearity of regression of the first record on the second is examined for discrete random variables. Both ordinary and weak records are considered. The analysis involves the determination of all possible linear relationships and all possible probability distributions. Several characterizations of geometric distributions are also shown.

Key Words: Discrete distributions, generalized geometric distribution, geometric distribution, linearity of regression, ordinary records, weak records.

AMS subject classification: 60E05, 62E10, 62E15.

1 Introduction

For a sequence $\mathbf{X} = \{X_n\}_{n\geq 1}$ of independent identically distributed (iid) random variables let us define record times as $U(1) = 1$, $U(n) = \inf\{j > U(n-1) : X_j > X_{U(n-1)}\}$, for $n = 2, 3, \ldots$. Then $R_n = X_{U_n}$ is called the n-th record of the sequence $\mathbf{X}$. The linearity of the regression of R_{n+1} given R_n within the class of continuous distributions was studied for the first time in Nagaraja (1977), where a family of three distributions with this property was identified. Nagaraja (1988) also described a class of distributions for which the regression of R_n on R_{n+1} is linear, and observed that the exponential distribution is the only distribution for which both the regressions for the adjacent records are linear. All these results were obtained under the assumption that the common distribution of X_j is continuous.

*Correspondence to: Fernando López-Blázquez. Dpto. de Estadística e I.O., Universidad de Sevilla, Sevilla, Spain. Email: lopez@cica.es

Received: January 2000; Accepted: October 2000

Instead of the regular records defined above, for the discrete distributions Vervaat (1973) proposed to use weak records, which are defined by weak record times $V(1)=1$, $V(n)=\inf\{j>V(n-1):\ X_j\geq X_{V(n-1)}\}$, for $n=2,3,\dots$. Then, $W_n=X_{V(n)}$ is called the n-th weak record. This definition seems to be much more natural in the discrete case, since it gives no priority to the index of the observation, which agrees with the intuition for the iid observations. Observe that in the case of continuous distributions $R_n=W_n$ a.s. Furthermore, in the discrete case weak records are also defined for distributions with bounded support, while for ordinary records this is not possible, without additional assumptions.

We restrict ourselves to supports of X_j's of the form $\{0,1,\dots,N\}$ with N possibly equal to infinity. The joint distributions for weak records can be easily derived

$$P(W_1=k_1,\dots,W_n=k_n)=p_{k_n}\prod_{r=1}^{n-1}\frac{p_{k_r}}{q_{k_r}},\quad 0\leq k_1\leq\cdots\leq k_n\leq N,\quad (1.1)$$

where $p_k=P(X_1=k)$ and $q_k=\sum_{j\geq k}p_j$, $k\geq 0$. Obviously, $k_n<N$ if $N=\infty$ in (1.1).

In the case of ordinary records, as a bounded support is not permitted, $N=\infty$ and the joint distribution is

$$P(R_1=k_1,\dots,R_n=k_n)=p_{k_n}\prod_{r=1}^{n-1}\frac{p_{k_r}}{q_{k_r+1}},\quad 0\leq k_1<\cdots<k_n<\infty.$$

Consequently, $P(W_{n+1}=l|W_n=k)=p_l/q_k$, $0\leq k\leq l$, and $P(R_{n+1}=l|R_n=k)=p_l/q_{k+1}$, $0\leq k<l<\infty$, so both conditional distributions have a simple form. Consequently, the problem of the linearity of the regression of R_{n+1} on R_n was solved in Korwar (1984), where the family of distributions consisting of the geometric tail and negative hypergeometric of the second type tail distributions was characterized - see also comments in the monograph of Arnold, Balakrishnan and Nagaraja (1998). The same problem for the regression of W_{n+1} on W_n was solved in Stepanov (1993) and Wesołowski and Ahsanullah (2000), where the question of the linearity of the regression of W_{n+2} on W_n was also completely resolved. More characterizations through weak records can be seen in Aliev (1998).

Nothing is known about regressions in the opposite direction, i.e. for $E(R_n|R_{n+1})$ or $E(W_n|W_{n+1})$ which seem to be much more complicated.

This is mainly due to the fact that the formulae for the respective conditional distribution, even for the simplest adjacent case, are quite complicated. This paper is devoted to a thorough discussion of the linearity of the regression in the easiest case, i.e. for $n = 1$. Section 2 is devoted to weak records, while ordinary records are considered in Section 3.

2 Linearity of regression for weak records

The joint probability mass function of the first two weak records can be easily determined as

$$P(W_1 = j, W_2 = k) = p_k p_j / q_j, \quad 0 \leq j \leq k. \tag{2.1}$$

From (2.1) it can be immediately deduced that the conditional distribution of W_1 given W_2 is

$$P(W_1 = j \mid W_2 = k) = \frac{c_j}{\sum_{r=0}^{k} c_r}, \quad 0 \leq j \leq k$$

with $c_j = p_j/q_j$, $j \geq 0$. It is obvious that $c_j \in (0, 1]$, for all j in the support of X_1. In particular $c_0 = p_0$ and if N is finite then $c_N = 1$.

Note that given a probability mass function the quantities c_j are calculated as the ratio between p_j and q_j. From the c_j's the probability mass function can be obtained, at least formally, as

$$p_0 = c_0, \quad p_j = c_j \prod_{m=0}^{j-1} (1 - c_m), \ j \geq 1. \tag{2.2}$$

For technical reasons that will be more fully understood in the proof of our main result, we need some conditions that ensure that given positive real numbers c_j's, the sequence of p_j's, defined in (2.2) is a probability mass function with $p_j > 0$, $j \geq 0$.

Lemma 2.1. *Assume that* $\{c_j\}_{j \geq 0}$ *is a sequence of numbers belonging to* $(0, 1]$. *Let* $N = \inf\{j \geq 0 : \ c_j = 1\}$, $(\inf(\emptyset) = \infty)$. *Define a sequence* $\{p_j\}_{j=0}^{N}$ *by* (2.2). *If*

(a) $N < \infty$ *or*

(b) $N = \infty$ *and* $\sum_{j=0}^{\infty} c_j = \infty$

then $\{p_j\}_{j=0}^{N}$ *is a probability mass function and* $p_j > 0$, $j = 0, 1, \ldots, N$.

Proof. If N is finite the proof follows immediately since $q_N = p_N$. Thus we will consider only the case $N = \infty$. It is obvious that $p_j > 0$, for all $j > 0$. Let us consider the partial sums $S_k = \sum_{j=0}^{k} p_j$. By induction, it can be proved that $1 - S_k = \prod_{j=0}^{k}(1 - c_j)$, for all $k \geq 1$. Consequently, we have to show that $\lim_k (1 - S_k) = 0$ or equivalently,

$$\lim_k \sum_{j=0}^{k-1} \log(1 - c_j) = -\infty. \tag{2.3}$$

Condition (2.3) obviously holds, without any additional assumptions, if the sequence $\{c_j\}_{j=0}^{\infty}$ does not converge to zero. If the sequence $\{c_j\}_{j=0}^{\infty}$ has a limit equal to zero the result follows from the inequality, $\log(1 - x) < -x$ for any $x \in (0, 1)$, and the assumption that $\sum_{j=0}^{\infty} c_j = \infty$. □

The distribution defined by a sequence $\{c_j\}$ as in (2.2), under the conditions specified in Lemma 1, is called the generalized geometric distribution, including obviously the ordinary geometric distribution if all c_j's are equal. Such distributions will be identified in this section as the only discrete distributions with the support $\{0, 1, \ldots, N\}$ that have the property of the linearity of the regression

$$E\,(W_1 \mid W_2) = \beta W_2 + \alpha, \tag{2.4}$$

for some real numbers α and β.

Theorem 2.1. *Let* X_j *be a sequence of discrete non-degenerate random variables with the support* $\{0, 1, \ldots, N\}$ *for which the linearity of the regression of* W_1 *given* W_2 *defined by (2.4) holds. Then* $\alpha = 0$, $\beta \in (0, 1)$ *and the probability mass function of* X_j *is of the generalized geometric type defined by (2.2) and with* $\delta = \beta/(1 - \beta)$

(a) if $\beta \in (1/2, 1)$ *then* N *is finite and*

$$c_j = \frac{\Gamma(j + \delta) N!}{\Gamma(N + \delta) j!}, \quad j = 0, 1, \ldots, N;$$

(b) *if* $\beta = 1/2$ *then* $N = \infty$ *and* $c_j = c_0$ *for any* $j \geq 0$, *i.e.* $\{p_j\}_{j \geq 0}$ *is the probability mass function of the geometric distribution:* $p_j = c_0(1-c_0)^j$, $j \geq 0$;

(c) *if* $\beta \in (0, 1/2)$ *then*

$$c_j = \frac{\Gamma(j+\delta)}{\Gamma(\delta) j!} c_0, \quad j \geq 0, \quad c_0 \in (0,1).$$

Proof. The property of the linearity of the regression given in (2.4) implies that

$$\sum_{j=0}^{k} j c_j = (\beta k + \alpha) \sum_{j=0}^{k} c_j, \tag{2.5}$$

for all $k \in \{0, 1, \ldots, N\}$, with $c_j = p_j / q_j \in (0, 1]$. In particular, for $k = 0$, equation (2.5) gives $0 = \alpha c_0 = \alpha p_0$ and, as $p_0 > 0$, we obtain $\alpha = 0$.

It is obvious that the slope β must be a positive number. Moreover, as $W_1 \leq W_2$ a.s. then β must be less than or equal to one. Observe that for $\beta = 1$ we obtain from (2.5) that $c_0 = 1$ and $N = 0$, and consequently the X_j's are concentrated at zero, which is not possible. Finally, we conclude that $\beta \in (0, 1)$.

Subtract from (2.5) evaluated at $k+1$, identity (2.5) evaluated at k to obtain

$$(k(1-\beta)+1) c_{k+1} = \beta \sum_{j=0}^{k+1} c_j, \tag{2.6}$$

for $k \in \{0, 1, \ldots, N-1\}$.

Write expression (2.6) for $k-1$ and subtract this from the original one. It follows that

$$c_{k+1} = \frac{(k-1)(1-\beta)+1}{(k+1)(1-\beta)} c_k, \quad k \in \{1, \ldots, N-1\}. \tag{2.7}$$

Observe that identity (2.5) with $k = 1$ gives $c_1 = \beta c_0 / (1-\beta)$. Therefore, (2.7) is also valid for $k = 0$ and can be rewritten as

$$c_{k+1} = \frac{k+\delta}{k+1} c_k, \quad k \in \{0, \ldots, N-1\}, \tag{2.8}$$

with $\delta = \beta/(1-\beta)$.

By recurrence, it follows from (2.8) that

$$c_k = \frac{\Gamma(k+\delta)}{\Gamma(\delta)k!}c_0, \quad k \in \{0,\dots,N\}. \tag{2.9}$$

Observe further that

$$\log(c_k) = \log(c_0) + \sum_{j=1}^{k} \log(1+(\delta-1)/j), \quad k = 1,2,\dots. \tag{2.10}$$

Since the series $\sum_{j=1}^{\infty} \log(1+(\delta-1)/j)$ for $\delta > 1$ (which is equivalent to $\beta > 1/2$) diverges to ∞ (note that for sufficiently large j $\log(1+(\delta-1)/j) > (\delta-1)/(2j)$), then it follows that $c_k \to \infty$ as $k \to \infty$. Since all c_j's are bounded by 1 it follows that $N < \infty$. Consequently, $c_N = 1$, and by (2.9) it follows that

$$p_0 = c_0 = \frac{\Gamma(\delta)N!}{\Gamma(N+\delta)}$$

and so the first part of the theorem is proved.

On the other hand, it follows immediately from (2.9) that for $\delta = 1$, i.e. $\beta = 1/2$, we have $c_k = c_0$ for any $k = 0,1,\dots,N$. Then $N = \infty$ and $\{p_j\}_{j\geq 0}$ is geometric, i.e. $p_j = p(1-p)^j$, $j = 0,1,\dots$, with $p = c_0$.

For $\delta \in (0,1)$ (which is equivalent to $\beta \in (0,1/2)$) we have $\log(1-(1-\delta)/j) < -(1-\delta)/j$ for any $j = 1,2,\dots$. Consequently the series $\sum_{j=1}^{\infty} \log(1+(\delta-1)/j)$ diverges to $-\infty$ and it follows from (2.10) that $\lim_{k\to\infty} c_k = 0$. Now, by the previous lemma, it suffices to show that, in this case, $\sum_{k=0}^{\infty} c_k = \infty$. To this end we use the Raabe criterion, which says that it suffices to show that $\lim_{k\to\infty} k(1-c_{k+1}/c_k)$ is less than one. But

$$k\left(1-\frac{c_{k+1}}{c_k}\right) = k\left(1-\frac{k+\delta}{k+1}\right) = k\frac{1-\delta}{k+1} \overset{k\to\infty}{\longrightarrow} 1-\delta < 1.$$

□

The discrete distributions for which $E(W_{i+1} \mid W_i)$ is linear for a given fixed $i \in \{1,2,\dots\}$ were studied in Stepanov (1993) and Wesołowski and Ahsanullah (2000). This family consists of the geometric and negative hypergeometric of the first and second type distributions. From these results and Theorem 2.1 we obtain immediately a characterization of geometric distributions which can be considered as the discrete version of the characterization of the exponential law obtained in Nagaraja (1988).

Corollary 2.1. *Assume that X_j has support $\{0, \ldots, N\}$ $(N \leq \infty)$. If both the regressions $E(W_1 \mid W_2)$ and $E(W_2 \mid W_1)$ are linear then the common distribution of the X_j's is geometric.*

Proof. It follows that both the negative hypergeometric probability mass functions are not of the generalized geometric type as specified in Theorem 2.1. □

3 Ordinary records

The joint probability mass function of the first two ordinary records can be easily determined as

$$P(R_1 = j,\, R_2 = k) = p_k p_j / q_{j+1}, \tag{3.1}$$

for any $0 \leq j < k < \infty$. From (3.1) it follows that the conditional distribution of R_1 given R_2 is defined by

$$P(R_1 = j \mid R_2 = k) = \frac{d_j}{\sum_{r=0}^{k-1} d_r}, \quad 0 \leq j < k < \infty$$

with $d_j = p_j / q_{j+1}$, $j \geq 0$.

Note that given a probability mass function the quantities d_j are calculated as the ratio between p_j and q_{j+1}. From the d_j's the probability mass function can be obtained, as

$$p_0 = \frac{d_0}{1 + d_0}, \quad p_j = \frac{d_j}{1 + d_j} \prod_{m=0}^{j-1} \frac{1}{1 + d_m}, \; j \geq 1. \tag{3.2}$$

Again, for technical reasons, we need some conditions that ensure that given positive real numbers d_j, $j \geq 0$, the sequence p_j, $j \geq 0$, is a probability mass function; then, obviously, it is a probability mass function of the generalized geometric type as defined by (3.2) with $c_j = d_j/(1 + d_j)$, $j \geq 0$.

Lemma 3.1. *Let $\{d_j\}_{j \geq 0}$ be a sequence of positive real numbers. If $\sum_{j=0}^{\infty} d_j = +\infty$, then the sequence $\{p_j\}_{j \geq 0}$ defined as in (3.2) is a probability mass function with $p_j > 0$, for all $j \geq 0$.*

Proof. Define $c_j = d_j/(1+d_j)$, $j = 0, 1, \ldots$. Observe that $d_j = c_j/(1-c_j) \leq c_j$ since $c_j \in (0, 1)$, $j \geq 0$. Since the series $\sum_{j=0}^{\infty} d_j$ diverges to infinity it follows that $\sum_{j=0}^{\infty} c_j = \infty$. The result is then an immediate consequence of Lemma 2.1. □

Our aim is to characterize discrete distributions for which

$$E(R_1 \mid R_2) = \beta_1 R_2 + \alpha_1, \tag{3.3}$$

for some real numbers α_1 and β_1. Such a characterization is given in the following theorem.

Theorem 3.1. *Let X_j, $j = 1, 2, \ldots$, be discrete iid random variables with the common support $\{0, 1, \ldots\}$ for which the linearity of the regression of R_1 on R_2 defined by (3.3) holds. Then $\alpha_1 = -\beta_1$, $\beta_1 \in (0, 1)$ and the common probability mass function of X_j's is of the generalized geometric type defined by (3.2) with $d_0 > 0$,*

$$d_k = \frac{\Gamma(k+\delta)}{\Gamma(\delta)k!} d_0, \quad k \geq 1.$$

and $\delta = \beta_1/(1-\beta_1)$.

Proof. The property of the linearity of the regression given in (3.3) implies that

$$\sum_{j=0}^{k-1} j d_j = (\beta_1 k + \alpha_1) \sum_{j=0}^{k-1} d_j, \tag{3.4}$$

for all $k \geq 1$, with $d_j = p_j/q_{j+1} > 0$. In particular, for $k = 1$, expression (3.4) gives $0 = (\beta_1 + \alpha_1)d_0$ and as $d_0 > 0$, we obtain $\alpha_1 = -\beta_1$.

Observe also that the slope β_1 must be a positive number, otherwise for large values of k the right side of (3.4) will be negative, which is impossible, since the left side is always non-negative. Moreover, as $R_1 < R_2$ a.s., $E(R_1 \mid R_2) = \beta_1(R_2 - 1) < R_2$ a.s. or equivalently, $\beta_1 < k/(k-1)$ for all $k > 1$, from which we conclude that $\beta_1 < 1$.

Following similar arguments as in the proof of Theorem 2.1, we get the recurrence formula

$$d_{k+1} = \frac{k+\delta}{k+1} d_k, \quad k \geq 0, \tag{3.5}$$

with $\delta = \beta_1/(1-\beta_1)$.

Performing the recurrence according to (3.5), it can be shown that

$$d_k = \frac{\Gamma(k+\delta)}{\Gamma(\delta)k!} d_0, \quad k \geq 0. \tag{3.6}$$

On comparing (3.6) with (2.9) in the proof of Theorem 2.1, and repeating the argument given there, we get

$$\lim_{k\to\infty} d_k = \begin{cases} \infty, & \text{if } \delta > 1 \\ d_0, & \text{if } \delta = 1 \\ 0, & \text{if } \delta < 1 \end{cases} .$$

Now, according to Lemma 2.1, it suffices to prove that $\sum_{j=0}^{\infty} d_j = \infty$. If $\delta \geq 1$ it is obvious. And for $\delta < 1$ it follows immediately by the Raabe criterion again as in the proof of Theorem 2.1. □

A particular case of Theorem 3.1 occurs when the slope β_1 is 1/2. In that case the probability mass function obtained from (3.2) and (3.6) is geometric. It appears that it is the only distribution for which both the regressions for weak records W_1 onto W_2 and for ordinary records R_1 onto R_2, are both simultaneously linear, as is shown in the following corollary.

Corollary 3.1. *The unique discrete distributions with support on the non-negative integers, for which $E(W_1 \mid W_2)$ and $E(R_1 \mid R_2)$ are both linear, are geometric distributions.*

Proof. Suppose that the X_j's are not geometric with support $\{0, 1, \ldots\}$. Since $E(W_1 \mid W_2)$ and $E(R_1 \mid R_2)$ are both linear then by Theorems 2.1 and 3.1, it follows that

$$E(W_1 \mid W_2) = \beta W_2, \quad E(R_1 \mid R_2) = \beta_1(R_2 - 1),$$

for certain $\beta \in (0, 1/2)$ and $\beta_1 \in (0, 1)$. Let $\delta = \beta/(1-\beta)$ and $\delta_1 = \beta_1/(1-\beta_1)$. As the X_j's are not geometric, δ and δ_1 do not equal 1. From (2.8) and (3.1), we have

$$c_{k+1} = \frac{k+\delta}{k+1} c_k, \quad k \geq 0, \tag{3.7}$$

and

$$d_{k+1} = \frac{k+\delta_1}{k+1} d_k, \quad k \geq 0, \tag{3.8}$$

with $c_k = p_k/q_k$ and $d_k = p_k/q_{k+1}$, $k \geq 0$.. Observe that from the definitions of c_k's and d_k's it follows that

$$c_k = \frac{p_k/q_{k+1}}{(p_k + q_{k+1})/q_{k+1}} = \frac{d_k}{1+d_k}, \quad k \geq 0.$$

Similarly $d_k = c_k/(1-c_k)$, $k \geq 0$. Hence by (3.8) one gets

$$c_{k+1} = \frac{d_{k+1}}{1+d_{k+1}} = \frac{(k+\delta_1)d_k}{(k+1)+(k+\delta_1)d_k} = \frac{(k+\delta_1)c_k}{(k+1)-(1-\delta_1)c_k}, \quad k \geq 0. \tag{3.9}$$

Equating (3.7) and (3.9) and taking into account that $c_k \neq 0$, for all $k \geq 0$, we get

$$c_k = \frac{(k+1)(\delta-\delta_1)}{(k+\delta)(1-\delta_1)}. \tag{3.10}$$

From (3.10), as $c_k \neq 0$, we must have $\delta \neq \delta_1$. Taking limits when k goes to infinity on both sides of (3.10), we have

$$\lim_k c_k = \frac{\delta-\delta_1}{1-\delta_1} \in (0, \infty), \tag{3.11}$$

but for $\delta \neq 1$, see the proof of Theorem 2.1, $\lim_k c_k = 0$ or ∞, which is contradictory to (3.11). $\square$

Similarly, as in the case of weak records as a consequence of Theorem 3.1, and earlier results on characterizations of the distribution of X_j's by linearity of $E(R_2 \mid R_1)$ (geometric tail and negative hypergeometric tail distributions - see Korwar 1984), we derive immediately the following characterization of the geometric distribution (being another discrete version of Nagaraja's (1988) characterization of the exponential distribution).

Corollary 3.2. *Assume that X_j's have the support $\{0, 1, \ldots\}$. If both the regressions $E(R_1 \mid R_2)$ and $E(R_2 \mid R_1)$ are linear then the common distribution of X_j's is geometric.*

References

Aliev, F.A. (1998). Characterizations of distributions through weak records. *Journal of Applied Statistical Science*, **8**, 13-16.

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

Korwar, R.M. (1984). On characterizing distributions for which the second record value has a linear regression on the first. *Sankhya, B*, **46**, 108-109.

Nagaraja, H.N. (1977). On a characterization based on record values. *Australian Journal of Statistics*, **19**, 70-73.

Nagaraja, H.N. (1988). Some characterizations of continuous distributions based on regressions of adjacent order statistics and record values. *Sankhya, A*, **50**, 70-73.

Stepanov, A.V. (1993). A characterization theorem for weak records. *Theory of Probability and Its Application*, **38**, 762-764.

Vervaat, W. (1973). Limit theorems for records from discrete distributions. *Stochastic Processes and Their Applications*, **1**, 317-334.

Wesołowski, J. and M. Ahsanullah (2000). Linearity of regression for non-adjacent weak records. *Statistica Sinica* (to appear).