Computational Statistics and Data Analysis 78 (2014) 21–32

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Statistical inference for the geometric distribution based on δ-records

Raúl Gouet [a], F. Javier López [c], Lina P. Maldonado [b], Gerardo Sanz [c,*]

[a] Dpto. Ingeniería Matemática y Centro de Modelamiento Matemático (UMI 2807, CNRS), Universidad de Chile, Chile

[b] Dpto. Métodos Estadísticos, Facultad de Ciencias, Universidad de Zaragoza, Zaragoza, Spain

[c] Dpto. Métodos Estadísticos and BIFI, Facultad de Ciencias, Universidad de Zaragoza, Zaragoza, Spain

ARTICLE INFO

Article history:
Received 1 August 2013
Received in revised form 2 April 2014
Accepted 4 April 2014
Available online 12 April 2014

Keywords:
δ-records
Geometric distribution
Maximum likelihood estimation
Bayes estimation
Empirical Bayes
Maximum likelihood prediction
Bayes prediction

ABSTRACT

New inferential procedures for the geometric distribution, based on δ-records, are developed. Maximum likelihood and Bayesian approaches for parameter estimation and prediction of future records are considered. The performance of the estimators is compared with those based solely on record-breaking data by means of Monte Carlo simulations, concluding that the use of δ-records is clearly advantageous. An example using real data is also discussed.

1. Introduction

Records are quite ubiquitous and interesting objects per se. They are naturally encountered in sports but also in climatology, seismology, finance, insurance, etc. Their mathematical theory has been under development for several decades and reached maturity, as can be seen in Ahsanullah (1995); Arnold et al. (1998) and Nevzorov (2001). In parallel but somewhat later, the theory of statistical inference based on record-breaking data began to develop and rapidly attained a good level of sophistication; see Gulati and Padgett (2003). A good reason for exploring record-based inference procedures is that record-breaking data are readily available in many situations. Consider, for example, the standard experimental setup of destructive stress-testing, where the data consists of (lower) records only; see Glick (1978) for an account of this issue. This strategy provides valuable information for estimating population quantiles, say, at a fraction of the measurement costs of classical sampling.

The concept of δ-record was introduced in Gouet et al. (2007) as a natural generalization of classical records, simple enough to allow for rigorous analysis of mathematical properties, such as the asymptotic behaviour of their counting process; see also Gouet et al. (2012). Loosely speaking, a δ-record is an observation which is either a record or falls short of being one (see below for definitions) and, as such, it is not surprising that we consider them as candidates for upgrading the so-called inference from record-breaking data. In fact, the destructive stress-testing setup mentioned above needs only a

* Correspondence to: Dpto. Métodos Estadísticos, Facultad de Ciencias, Universidad de Zaragoza, C/ Pedro Cerbuna, 12, 50009 Zaragoza, Spain. Tel.: +34 976762860; fax: +34 976762860.

E-mail addresses: rgouet@dim.uchile.cl (R. Gouet), javier.lopez@unizar.es (F.J. López), lmguaje@unizar.es (L.P. Maldonado), gerardo.sanz@unizar.es (G. Sanz).

http://dx.doi.org/10.1016/j.csda.2014.04.002

minor adjustment to allow for the measurement of δ-record data. Of course, the total cost will increase as more items are destroyed but the type of additional data obtained, close to records, is likely to enhance the performance of procedures originally designed to accept only records as input.

A first example showing the usefulness of δ-records can be found in Gouet et al. (2012). Preliminary results are encouraging but much remains to be done, both in terms of theory and applications. This paper is meant to contribute in that direction, in the context of an important discrete model such as the geometric distribution.

Let $(X_n)_{n\geq 1}$ be a random sequence, $M_n = \max\{X_1, \ldots, X_n\}$, for $n \geq 1$, and $M_0 = -\infty$. Observation X_n is a record if it is greater than all previous observations, that is, if $X_n > M_{n-1}$. On the other hand, X_n is a δ-record if $X_n > M_{n-1}+\delta$, where δ is a fixed real parameter. If $\delta > 0$ every δ-record is a record but obviously not all records are δ-records. So, in this case, δ-records are even scarcer than ordinary records and clearly not adequate to replace records in inference procedures. The opposite situation is observed if $\delta < 0$ because every record is a δ-record and additionally, non-record observations within distance $-\delta$ of the current record are also δ-records. In this case δ-records correspond to records together with near-records, as defined in Balakrishnan et al. (2005). For discrete distributions, δ-records with $\delta = -1$ correspond to weak records, which have been largely analysed in the literature; see Castaño Martínez et al. (2013), Gouet et al. (2008) and Hashorva and Stepanov (2012) for recent developments in the study of weak records. Since we observe more δ-records than records we can expect better performance of δ-record-based inference than inference based only on records. Accordingly, Gouet et al. (2012) consider maximum likelihood estimation for continuous distributions, in particular exponential and Weibull distributions, showing that δ-records-based estimators outperform those based only on records. Also, López-Blázquez and Salamanca-Miño (2013) analyse, in Section 7, properties of maximum likelihood (ML) estimation based on δ-records in the exponential distribution.

The literature on record-based statistical inference for discrete models is rather scarce, when compared to that for continuous ones. Interesting references are Stepanov et al. (2003), dealing with the Fisher information contained in records, and Doostparast and Ahmadi (2006), on the statistical analysis of the geometric distribution, both from Bayesian and non-Bayesian viewpoints.

The aim of this paper is to assess the usefulness of δ-records by following a path similar to that of Doostparast and Ahmadi (2006), namely developing new (δ-record-based) point and interval estimators of the parameter and predictors of future records, for the geometric distribution, in both frequentist and Bayesian inferential frameworks. Monte Carlo simulations clearly show superior performance of procedures using δ-records, in all instances considered.

The paper is organized as follows. In Section 2 we introduce the probabilistic framework, define δ-records and related random variables and then focus on the estimation of parameter p. We consider maximum likelihood (in Section 2.2), Bayes and empirical Bayes estimation (in Section 2.3). The prediction of future records is addressed in Section 3, with maximum likelihood framework in Section 3.1 and Bayesian in Section 3.2. Conclusions and some ideas for future work are presented in Section 4.

2. Parameter estimation

2.1. Probabilistic framework and definitions

We consider a sequence $(X_n)_{n\geq 1}$ of independent, geometrically distributed random variables, with parameter $p \in (0, 1)$. That is, $P(k) := P[X_1 = k] = pq^{k-1}$, for $k = 1, 2, \ldots$, with $q := 1 - p$. Let $(M_n)_{n\geq 1}$ denote the sequence of partial maxima, with $M_n = \max\{X_1, \ldots, X_n\}$, for $n \geq 1$, $M_0 = -\infty$ and let $(R_n)_{n\geq 1}$ be the sequence of record values, obtained from partial maxima by discarding repetitions. Record times L_n, $n \geq 1$, are defined as $L_1 = 1$ and $L_n = \min\{m > L_{n-1}; X_m > X_{L_{n-1}}\}$, $n \geq 2$. Then, clearly $R_n = X_{L_n}$, for $n \geq 1$.

For δ-records we consider a fixed, negative-integer-valued parameter δ, because the random variables X_n are integer valued. However, it is convenient to include 0 as possible value of δ in order to see records as particular case of δ-records. Observe that in this context weak records, satisfying $X_n \geq M_{n-1}$, are δ-records with $\delta = -1$.

Recall that X_1 is a δ-record by convention and that X_n is a δ-record if $X_n > M_{n-1} + \delta$, $n \geq 2$. We say that X_n is a δ-record associated to the mth record R_m if $X_n > M_{n-1} + \delta = R_m + \delta$ and $L_m \leq n < L_{m+1}$. The number of δ-records associated to R_m, not counting R_m itself, is denoted by S_m and the vector of δ-records associated to R_m (excluding R_m) by $(Y_m^1, \ldots, Y_m^{S_m})$. Observe that $0 \leq S_m < L_{m+1} - L_m$ and also that $R_m + \delta < Y_m^j \leq R_m$, for $j = 1, \ldots, S_m$.

Proposition 2.1. *Let $n \geq 1$ and $\delta \leq -1$, then,*

(i) *conditional on $R_i = r_i$, S_i has geometric distribution (starting at 0) with success parameter $p_i := \bar{F}(r_i)/\bar{F}(r_i + \delta)$, for $i = 1, \ldots, n$, where $\bar{F}(x) := 1 - F(x)$ is the survival function of X_1.*

(ii) *Conditional on $R_i = r_i$, $S_i = s_i$, the random variables $Y_i^1, \ldots, Y_i^{S_i}$ are independent, with common (conditional) probability mass function*

$$P_i(k) := \frac{P(k)}{\bar{F}(r_i + \delta) - \bar{F}(r_i)}, \quad k = r_i + \delta + 1, \ldots, r_i.$$