Journal of Statistical Planning and Inference 157–158 (2015) 77–89

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Asymptotic results for runs and empirical cumulative entropies

Rita Giuliano [a], Claudio Macci [b,*], Barbara Pacchiarotti [b]

[a] Dipartimento di Matematica, Università di Pisa, Largo Bruno Pontecorvo 5, I-56127 Pisa, Italy

[b] Dipartimento di Matematica, Università di Roma Tor Vergata, Via della Ricerca Scientifica, I-00133 Roma, Italy

ARTICLE INFO

Article history:
Received 3 April 2014
Received in revised form 7 August 2014
Accepted 1 October 2014
Available online 14 October 2014

MSC:
60F10
62G20
62G30

Keywords:
Large deviations
Moderate deviations
Order statistics
Spacings

ABSTRACT

We prove large and moderate deviations for two sequences of estimators based on the order statistics and, more precisely, on spacings. In the first case we deal with runs, and we have sums of independent Bernoulli distributed random variables. In the second case we deal with empirical cumulative entropies, and we have linear combinations of independent exponentially distributed random variables.

1. Introduction

In this paper we prove results for sequences of runs and of empirical cumulative entropies. Runs are commonly used in different applications including reliability, quality control and molecular biology; see e.g. Balakrishnan and Koutras (2002) and Fu and Lou (2003) for a review on theory and applications, and the recent paper (Balakrishnan and Stepanov, 2013) for a complete bibliography. Cumulative entropy is one of the generalizations or modifications in the literature of the classical differential entropy (or Shannon information measure) defined by

$$\mathcal{H}(X) := -\int_0^\infty f(x)\log f(x)dx,$$

where X is a nonnegative and absolutely continuous random variable, with probability density function f. This concept, which plays a crucial role in the field of information theory, was introduced in Shannon (1948).

There are several definitions of runs in the literature and here we have in mind the ones based on differences and ratios of consecutive order statistics as in Eryilmaz and Stepanov (2008) (see also Stepanov (2011a)) and in Stepanov (2011b), respectively. In both cases we have sums of independent Bernoulli distributed random variables; so we present the results by referring to this scheme and later we illustrate the connection with the random variables studied in Eryilmaz and

* Corresponding author.
E-mail addresses: giuliano@dm.unipi.it (R. Giuliano), macci@mat.uniroma2.it, claudio.macci@gmail.com (C. Macci), pacchiar@mat.uniroma2.it (B. Pacchiarotti).

http://dx.doi.org/10.1016/j.jspi.2014.10.001

Stepanov (2008), Stepanov (2011a) and Stepanov (2011b). For the empirical cumulative entropies we refer to the sequence of estimators in Di Crescenzo and Longobardi (2009a), and we have linear combinations of independent exponentially distributed random variables.

Large deviations give an asymptotic computation of small probabilities on an exponential scale (see e.g. Dembo and Zeitouni (1998) as a reference on this topic). The asymptotic behavior of sequences of estimators is often formulated in terms of consistency and asymptotic normality only, as it happens for instance for the estimators in this paper; we propose for them an approach based on large deviations, which is advisable in order to have a more complete analysis.

In this paper we prove large deviation principles for sequences which converge almost surely to a constant, i.e. Proposition 3.1 for the convergence in (7) concerning runs, and Proposition 4.1 for the convergence in (21) concerning the empirical cumulative entropies. Moreover we prove the corresponding results on *moderate deviations*, i.e. Propositions 3.3 and 4.3. The term moderate deviations is used when, for all sequences positive numbers $\{a_n : n \geq 1\}$ such that

$$a_n \to 0 \quad \text{and} \quad na_n \to \infty \quad (\text{as } n \to \infty),$$

we have a suitable large deviation principle governed by the same quadratic rate function (for each choice of $\{a_n : n \geq 1\}$) which uniquely vanishes at the origin. Typically moderate deviations fill the gap between the following two kinds of results:

1. almost sure convergence to zero of centered random variables (see (6) and (20)), which in some sense corresponds to the case $a_n = \frac{1}{n}$ (for all $n \geq 1$);
2. asymptotic normality results (see (11) and (24)), which in some sense corresponds to the case $a_n = 1$ (for all $n \geq 1$).

Large deviation rate functions are often expressed in terms of suitable relative entropies; see the discussion in Varadhan (2003). Here we briefly recall its definition (see e.g. Section 2.3 in Cover and Thomas (1991)): let Q_1 and Q_2 be two probability measures on the same measurable space $(\Omega, \mathcal{F})$ and we write $Q_1 \ll Q_2$ to mean that Q_1 is absolutely continuous with respect to Q_2; then the relative entropy of Q_1 with respect to Q_2 is

$$H(Q_1|Q_2) = \begin{cases} \int_\Omega \log\left(\frac{dQ_1}{dQ_2}(\omega)\right) dQ_1(\omega) & \text{if } Q_1 \ll Q_2 \\ \infty & \text{otherwise.} \end{cases}$$

Moreover we recall that $H(Q_1|Q_2) \geq 0$ and we have $H(Q_1|Q_2) = 0$ if and only if $Q_1 = Q_2$. Then Propositions 3.2 and 4.2 provide upper bounds for the rate functions in Propositions 3.1 and 4.1, respectively, in terms of suitable integrals of relative entropies; more precisely we mean relative entropies between Bernoulli distributions in Proposition 3.2 and relative entropies between exponential distributions in Proposition 4.2.

Spacings (i.e. the differences of consecutive ascending order statistics) of i.i.d. random variables play a crucial role in this paper for both runs and empirical cumulative entropies. In the statistical literature spacings come up for instance in some goodness of fit tests; therefore it is not surprising that they appear explicitly in the definition of the empirical cumulative entropies in Section 4 where the distribution function is estimated by the empirical distribution function. On the other hand spacings also come up when we illustrate the connection of the sums of Bernoulli distributed random variables in Section 3 with the random variables in Eryilmaz and Stepanov (2008); Stepanov (2011a) and Stepanov (2011b). In particular in both cases studied in Sections 3 and 4 it is important to deal with i.i.d. exponentially distributed random variables; indeed it is one of the few cases in which the exact distribution of spacings can be easily handled (see e.g. Section 2 in Pyke (1965) for the exact distributions of spacings).

Now we introduce some symbols used throughout the paper. We start with some well-known distributions: $\mathcal{N}(\mu, \sigma^2)$ is the Normal distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$; for $p \in [0, 1]$, $\mathcal{B}(p)$ is the Bernoulli distribution, i.e. the distribution of a random variable Z such that

$$P(Z = 1) = p = 1 - P(Z = 0);$$

for $\lambda > 0$, $\mathcal{E}(\lambda)$ is the exponential distribution, i.e. the distribution of a random variable Z such that

$$P(Z \leq t) = \begin{cases} 1 - e^{-\lambda t} & \text{if } t \geq 0 \\ 0 & \text{if } t < 0. \end{cases}$$

Moreover we write $\Rightarrow$ to mean weak convergence, and $a_n \sim b_n$ to mean that $\frac{a_n}{b_n} \to 1$ as $n \to \infty$. The (ascending) order statistics of some random variables $X_1, \ldots, X_n$ will be denoted by $X_{1:n}, \ldots, X_{n:n}$ (therefore $X_{1:n} \leq \cdots \leq X_{n:n}$); for both runs and empirical cumulative entropies we deal with the spacings $\{X_{j+1:n} - X_{j:n} : j \in \{1, \ldots, n-1\}\}$.

The outline of the paper is as follows. We start with some preliminaries in Section 2. The results for runs and empirical cumulative entropies are presented in Sections 3 and 4, respectively.

2. Preliminaries

Here we briefly recall some basic preliminaries on large deviations (see e.g. Dembo and Zeitouni (1998), pages 4–5). Let $\mathcal{X}$ be a topological space equipped with its completed Borel σ-field. A sequence of $\mathcal{X}$-valued random variables $\{Z_n : n \geq 1\}$