Available online at www.sciencedirect.com
ScienceDirect
Statistics & Probability Letters 78 (2008) 135–143
STATISTICS & PROBABILITY LETTERS
www.elsevier.com/locate/stapro

On the expectation of the maximum of IID geometric random variables

Bennett Eisenberg

Department of Mathematics #14, Lehigh University, Bethlehem, PA 18015, USA

Received 6 December 2006; received in revised form 11 April 2007; accepted 23 May 2007
Available online 8 June 2007

Abstract

A study of the expected value of the maximum of independent, identically distributed (IID) geometric random variables is presented based on the Fourier analysis of the distribution of the fractional part of the maximum of corresponding IID exponential random variables.

Keywords: Geometric random variables; Exponential random variables; Maximum; Asymptotic expectation; Fractional part; Fourier series; Bioinformatics

Introduction

The probability problem studied in this paper is motivated by a statistical problem in bioinformatics. In bioinformatics one often compares long sequences of nucleotides (denoted A, C, G, T) in DNA for similarities or sometimes observes a single string for long runs of a given nucleotide or pattern of nucleotides (see Ewens and Grant, 2005, sec. 5.4 and sec. 6.3.). In the simplest case of comparing two strands of DNA, one might line up the two DNA strands and count the length of matches. Under the null hypothesis of randomness, this might be modeled as a modified geometric random variable (the number of matches until the first non-match) with parameter $p = .75$, the probability of a non-match. One might also try to match triples of nucleotides called codons, which play a major role in molecular biology. Here we might use $p = 1 - .25^3 = .98$. These values for p are approximations since the nucleotides do not have equal probabilities in reality. In testing for statistical significance in these procedures, the test statistic is often the maximum of an extremely large number of independent, identically distributed (IID) modified geometric random variables. In doing such tests approximations are used for the distribution of the test statistic under the null hypothesis. Ewens and Grant state on p. 138, "unless very accurate approximations are used for the mean and variance…, serious errors in P-value approximations can arise…". In this paper we analyze the standard approximation used for the mean of the maximum of IID geometric random variables.

It is well known and easily shown that $E(M_n)$, the expected value of the maximum of n IID exponential random variables with mean $1/\lambda$ is $(1/\lambda)\sum_{k=1}^{n}(1/k)$. This formula is useful for large n since $\sum_{k=1}^{n}(1/k)$ is

E-mail address: BE01@Lehigh.edu

0167-7152/$ - see front matter
doi:10.1016/j.spl.2007.05.011

asymptotic to $(\log n + \gamma)$, where $\gamma = .577\ldots$ is Euler's constant. There is no such simple expression for $E(M_n^*)$, the expected value of the maximum of n IID geometric random variables with mean $1/p$. There is the infinite series expression from the tail probabilities of the maximum of the random variables and even a finite sum expression, but these expressions are not so useful.

It is also easily seen that

$$\frac{1}{\lambda}\sum_{k=1}^{n}\frac{1}{k} \leqslant E(M_n^*) < 1 + \frac{1}{\lambda}\sum_{k=1}^{n}\frac{1}{k},$$

where $q = (1-p) = \mathrm{e}^{-\lambda}$. This is better, but not good enough. The gap of 1 between the upper and lower bounds is significant for moderate values of λ and n. A finer analysis is needed.

Careful asymptotic approximations are given for $E(M_n^*)$ in Szpankowski and Rego (1990). In our notation the Szpankowski and Rego result for the first moment is as follows:

$$E(M_n^*) = \sum_{k=1}^{n}\frac{1}{\lambda k} + \frac{1}{2} - \frac{1}{\lambda}\sum_{m\neq 0}\Gamma\left(\frac{2\pi i m}{\lambda}\right)\mathrm{e}^{-2\pi i m \log n/\lambda} + \mathrm{O}(n^{-1}). \tag{1}$$

Since $|\Gamma(2\pi i m/\lambda)|$ is very small for $\lambda < 2$ and all $m \neq 0$, this result has the interpretation that for $\lambda < 2$ and large n that $E(M_n^* - \sum_{k=1}^{n}\frac{1}{\lambda k})$, is close to $1/2$. This is the interpretation used in Jeske and Blessinger (2004) and applied to bioinformatics by Ewens and Grant (2005).

This interpretation is true, but can be misleading. First, there is no convergence to $1/2$ since the gamma function terms do not go to zero. They are just very small. Also the error term $\mathrm{O}(n^{-1})$ means that here is an unknown constant C such that $\mathrm{O}(n^{-1}) < C/n$. Without knowing the value of C, one does not know what the bound really is. In this case this problem is compounded by the fact that for reasonable values of n and λ, the value $1/n$ is much greater than the value of the gamma function terms. Hence the $\mathrm{O}(n^{-1})$ term can easily dominate the infinite sum in determining how close $E(M_n^* - \sum_{k=1}^{n}\frac{1}{\lambda k})$ is to $1/2$.

In this paper we use simple Fourier analysis to show that $E(M_n^*) - \sum_{k=1}^{n}\frac{1}{\lambda k}$ is very close to $1/2$ not only for moderate values of λ, but also relatively small values of n and that this difference is logarithmically summable to $1/2$ for all values of λ. Moderate λ may be interpreted as $\lambda < 2$. This corresponds to $p < .865$, which is almost always the case. We see, however, that the codon example is an exception to this. A key component of this work is the analysis of the distribution of the fractional part of M_n.

1. A survey of formulas for $E(M_n)$ and $E(M_n^*)$

Let $X_1, X_2, \ldots, X_n$ be IID exponential random variables with $P(X \leqslant x) = 1 - \mathrm{e}^{-\lambda x}$ for $x > 0$ and let $M_n = \max(X_1, \ldots, X_n)$. We then have $P(M_n \leqslant x) = (1 - \mathrm{e}^{-\lambda x})^n$. It follows that

$$\begin{aligned} E(M_n) &= \int_0^\infty P(M_n > x)\,\mathrm{d}x = \int_0^\infty 1 - (1-\mathrm{e}^{-\lambda x})^n\,\mathrm{d}x \\ &= \int_0^1 \frac{1-u^n}{\lambda(1-u)}\,\mathrm{d}u = \int_0^1 \sum_{k=0}^{n-1}\frac{u^k}{\lambda}\,\mathrm{d}u = \sum_{k=1}^{n}\frac{1}{\lambda k}. \end{aligned}$$

This also follows by decomposing M_n as

$$\begin{aligned} M_n &= X_{(1)} + (X_{(2)} - X_{(1)}) + \cdots + (X_{(n)} - X_{(n-1)}) \\ &= Y_1 + Y_2 + \cdots + Y_n, \end{aligned} \tag{2}$$

where $X_{(i)}$ is the ith order statistic of $X_1, \ldots, X_n$. It follows from the lack of memory property of exponential random variables and the fact that the minimum of exponential random variables is exponential that $Y_1, \ldots, Y_n$ are independent exponential random variables with parameters $n\lambda, (n-1)/\lambda, \ldots, 1/\lambda$, respectively. This implies

$$E(M_n) = \sum_{k=1}^{n}\frac{1}{\lambda k} \quad \text{and} \quad \mathrm{Var}(M_n) = \sum_{k=1}^{n}\frac{1}{\lambda^2 k^2}. \tag{3}$$