A Poisson approximation with applications to the number of maxima in a discrete sample

Peter Olofsson

Department of Statistics, Rice University, MS-138, P.O. Box 1892, Houston, TX 77251, USA

Abstract

We present a Poisson approximation with applications to extreme value theory. Let $X_1, X_2, \ldots$ be i.i.d. and let $M_n^{(1)} \geqslant M_n^{(2)} \geqslant \cdots \geqslant M_n^{(j)}$ be the j largest order statistics. Then the asymptotic behavior of the vector $(M_n^{(1)}, \ldots, M_n^{(j)})$ is the same as that of $(M_N^{(1)}, \ldots, M_N^{(j)})$ where N is a random variable which is independent of $X_1, X_2, \ldots$ and has a Poisson distribution with mean n. The distribution of $(M_N^{(1)}, \ldots, M_N^{(j)})$ is easy to obtain since the points $X_1, X_2, \ldots, X_N$ form a Poisson process on the real line. The mean measure is $n \, \mathrm{d}F$ where F is the distribution function of the X_i. We apply this to the problem of multiple maxima in discrete samples, in particular from the geometric distribution where it is known that the number of maxima has no limiting distribution. © 1999 Elsevier Science B.V. All rights reserved

Keywords: Poisson approximation; Maxima in discrete samples; Extreme value theory; Geometric distribution

1. Introduction

Poisson approximation is a technique with many applications in probability theory and statistics. The literature on the subject is vast and we mention only two general references: Aldous (1989) and Barbour et al. (1992). In this paper, we investigate a special case of Poisson approximation with applications to extreme value theory. This was first described in Nordin and Olofsson (1989) where the method was used primarily for extreme values in continuous distributions. In this paper, we will apply the technique to problems of multiple maxima in discrete samples. The idea behind our Poisson approximation, originally suggested by Olle Nerman (Nerman, 1989), is to consider not a fixed number of observations $X_1, \ldots, X_n$ but a random number $X_1, \ldots, X_N$ where N has a Poisson distribution with mean n. Then the maxima of these two sequences will have the same asymptotic behavior as $n \to \infty$ and the second sample is nice because it consitutes a Poisson process. This is definitely part of the "folklore" on Poisson approximation and in the following sections we make it strict.

2. The Poisson process

We first show that a Poisson number of i.i.d. random variables gives rise to a Poisson process. Here and in the sequel, $X_1, X_2, \ldots$ will denote a sequence of i.i.d. random variables distributed as a generic random variable X with distribution function F.

Theorem 2.1. *Let $X_1, X_2, \ldots$ be i.i.d. random variables with the common distribution function F. Let N be independent of $X_1, X_2, \ldots$ and have a Poisson distribution with mean n and define the random counting measure S_N through*

$$S_N(\cdot) = \sum_{k=1}^{N} \delta_{X_k}(\cdot).$$

Then $S_N(\cdot)$ is a nonhomogeneous Poisson process with mean measure $n \, \mathrm{d}F$.

Proof. We will use uniqueness of Laplace functionals for random measures. Let f be a nonnegative bounded function. The Laplace functional of a random measure M on R is $\Psi_M(f) = E[\exp(-\int_R f(x) \, \mathrm{d}M(x))]$. The Laplace functional uniquely determines the random measure and M is a Poisson process with mean measure (μ) if and only if

$$\Psi_M(f) = \exp\left(-\int_R (1 - \mathrm{e}^{-f(x)}) \, \mathrm{d}\mu(x)\right),$$

see, for instance, Resnick (1992). The Laplace functional of S_N is

$$E\left[\exp\left(-\sum_{k=1}^{N} f(X_k)\right)\right] = \sum_{m=0}^{\infty} E\left[\exp\left(-\sum_{k=1}^{m} f(X_k)\right)\right] P(N = m)$$

$$= \sum_{m=0}^{\infty} (E[\mathrm{e}^{-f(X)}])^m \mathrm{e}^{-n} \frac{n^m}{m!}$$

$$= \exp(nE[\mathrm{e}^{-f(X)}] - n) = \exp\left(-\int_R (1 - \mathrm{e}^{-f(x)}) n \, \mathrm{d}F(x)\right)$$

and the result follows. $\quad\square$

3. Joint distributions of the j largest order statistics

The next theorem tells us that the Poisson approximation is indeed useful. Let $M_n^{(1)} \geqslant M_n^{(2)} \ldots \geqslant M_n^{(j)}$ be the j largest order statistics and let the vector $\mathbf{M}_n = (M_n^{(1)}, \ldots, M_n^{(j)})$. Let $\| \cdot \|$ denote total variation, i.e. if μ and v are probability measures, $\| \mu - v \| = \sup_A |\mu(A) - v(A)|$.

Theorem 3.1. *As $n \to \infty$,*

$$\|P(\mathbf{M}_n \in \cdot) - P(M_N \in \cdot)\| \to 0.$$

Proof. First note that $\mathbf{M}_N \neq \mathbf{M}_n \Leftrightarrow M_N^{(i)} \neq M_n^{(i)}$ for some $i \leqslant j$. Hence,

$$\|P(\mathbf{M}_n \in \cdot) - P(\mathbf{M}_N \in \cdot)\| \leqslant \sum_{i=1}^{j} P(M_N^{(i)} \neq M_n^{(i)}).$$

We will show that all the terms go to 0 as $n \to \infty$. Let $(k)_i = k(k-1)\ldots(k-i+1)$. For any fixed $k \leqslant n$, note that $M_k^{(i)} < M_n^{(i)}$ if and only if not all $M_n^{(1)}, \ldots, M_n^{(i)}$ are attained among $X_1, \ldots, X_k$. Hence,

$$P(M_k^{(i)} < M_n^{(i)}) \leqslant 1 - \frac{(k)_i}{(n)_i}$$

with equality if F is continuous (clearly the possibility of ties make distinct maxima less likely). Since $P(N = k - i) = (k)_i / n^i P(N = k)$ we obtain

$$P(M_N^{(i)} < M_n^{(i)}) = \sum_{k=0}^{n} P(M_k^{(i)} < M_n^{(i)}) P(N = k) \leqslant \sum_{k=0}^{n} \left(1 - \frac{(k)_i}{(n)_i}\right) P(N = k)$$

$$\leqslant \sum_{k=0}^{n} P(N = k) - \sum_{k=i}^{n} \frac{(k)_i}{n^i} P(N = k) = \sum_{k=n-i+1}^{n} P(N = k) \leqslant iP(N = n).$$

Finally, note that $E[(N)_i] = n^i$, and hence

$$P(M_N^{(i)} > M_n^{(i)}) = \sum_{k=n+1}^{\infty} P(M_k^{(i)} > M_n^{(i)}) P(N = k)$$

$$\leqslant \sum_{k=n+1}^{\infty} \left(1 - \frac{(n)_i}{(k)_i}\right) P(N = k) \leqslant \sum_{k=n+1}^{\infty} \left(\frac{(k)_i}{(n)_i} - 1\right) P(N = k)$$

$$= E\left[\frac{(N)_i}{(n)_i} - 1\right] + \sum_{k=i}^{n} \left(1 - \frac{(k)_i}{(n)_i}\right) P(N = k) \leqslant \frac{n^i}{(n)_i} - 1 + iP(N = n)$$

by the above calculations. Since $n^i/(n)_i \to 1$ and $P(N = n) \to 0$ as $n \to \infty$, the proof is complete. $\square$

Hence we can use the Poisson approximation for any event involving the j largest values for any fixed j. In the next section we will use this to investigate the problem of multiple maxima in a discrete sample, especially for the geometric distribution.

4. Multiple maxima in a discrete sample

In Råde (1991), the following problem is proposed. Toss n coins, probability p for heads, remove all that fell heads and toss the remaining coins again. Proceed in this way until all coins show heads. What can be said about the asymptotic behavior of the number of coins involved in the last toss as $n \to \infty$?

This problem has been solved and extended by several authors; among these are Baryshnikov et al. (1995), Brands et al. (1993) and Eisenberg et al. (1993). We will deduce some of their results by applying our Poisson approximation which makes some computations shorter and intuitively appealing.

First note that, with X_i denoting the number of tosses needed for the ith coin, the number of coins involved in the last toss will equal the number of X_i that need the maximum number of tosses to show heads.

Generally, consider a discrete sample $X_1, X_2, \ldots$ where the X_i have frequency function f and denote $G(j) = P(X \geqslant j)$. Assume that the X_i are unbounded (otherwise the number of maxima will clearly increase indefinitely). Let K_n denote the number of maxima. The event $\{K_n = k\}$ is then the same as the event $\{M_n^{(1)} = \cdots = M_n^{(k)} > M_n^{(k+1)}\}$ and by Theorem (3.1), the asymptotics of $P(K_n = k)$ will be the same as those of $P(K_N = k)$ where N is Poisson with mean n.

Therefore let N have a Poisson distribution with mean n and fix j. By Theorem 2.1, the number of X_i that equals j has a Poisson distribution with mean $nf(j)$ and these numbers for different j are independent. To

have k maxima attained in j thus has probability

$$P(K_N = k, M_N^{(1)} = j) = \frac{(nf(j))^k}{k!} e^{-nf(j)} e^{-nG(j+1)}$$

$$= \frac{(nf(j))^k}{k!} e^{-nG(j)}$$

since the process must have k points in j and no points from $j+1$ on and $G(j) = G(j+1) + f(j)$. Summing over j yields

$$P(K_N = k) = \sum_{j=1}^{\infty} \frac{(nf(j))^k}{k!} e^{-nG(j)}. \tag{4.1}$$

Of particular interest is the case $k = 1$, a unique maximum. Note that

$$P(K_N = 1) = n \sum_{j=1}^{\infty} f(j) e^{-nG(j)} = nL(n) \tag{4.2}$$

where L is the Laplace–Stieltjes transform of the random variable $G(X)$. This reminds us of the techniques in Baryshnikov et al. (1995) where a Tauberian theorem is used to show that $\lim P(K_n = 1)$ exists (and must equal one) if and only if $f(k)/G(k) \to 0$ as $k \to \infty$.

5. Subsequential limits in the geometric case

For a geometric distribution with success probability $p = 1 - q$ we have $f(j) = pq^j$ and $G(j) = q^j, j = 0, 1, \ldots$ and hence, by (4.1),

$$P(K_N = k) = \sum_{j=1}^{\infty} \frac{(npq^j)^k}{k!} e^{-nq^j}. \tag{5.1}$$

It is known that $P(K_n = k)$ does not converge as $n \to \infty$; there are different subsequential limits along different subsequences. For an intuitively appealing explanation of this slightly surprising fact, see Baryshnikov et al. (1995). The following theorem is from Eisenberg et al. (1993). We prove it by using our Poisson approximation.

Theorem 5.1. *Let K_n denote the number of maxima in a sample of size n from a geometric distribution with success probability $p = 1 - q$. If $n(m)$ is any sequence such that $n(m)q^m \to c$ as $m \to \infty$, then,*

$$P(K_{n(m)} = k) \to \frac{c^k p^k}{k!} \sum_{j=-\infty}^{\infty} q^j e^{-cq^j}$$

as $m \to \infty$.

Proof. By (5.1)

$$P(K_{n(m)} = k) = \sum_{j=1}^{\infty} \frac{(n(m)pq^j)^k}{k!} e^{-n(m)q^j} = \frac{n(m)^k p^k}{k!} \sum_{j=-m+1}^{\infty} q^{k(m+j)} e^{-n(m)q^m q^j}$$

$$= \frac{(n(m)q^m)^k p^k}{k!} \sum_{j=-m+1}^{\infty} q^{kj} e^{-n(m)q^m q^j} \to \frac{c^k p^k}{k!} \sum_{j=-\infty}^{\infty} q^{kj} e^{-cq^j}$$

as $m \to \infty$. $\quad\square$

6. Logarithmic summability of $P(N_n = k)$

We will now prove a result from Baryshnikov et al. (1995). A sequence $\{a_n\}$ is called *logarithmically summable* to a if $(\log m)^{-1} \sum_{n=1}^{m} n^{-1} a_n \to a$ as $m \to \infty$.

Theorem 6.1. *The probabilities $P(K_n = k)$ are logarithmically summable to $p^k/k|\log q|$, that is*

$$\frac{1}{\log m} \sum_{n=1}^{m} \frac{P(K_n = k)}{n} \to \frac{p^k}{k|\log q|}$$

for $k = 1, 2, \ldots$ as $m \to \infty$.

Proof. Plugging in the expression for $P(K_n = k)$, we obtain

$$\sum_{n=1}^{m} \frac{P(K_n = k)}{n} = \sum_{n=1}^{m} \sum_{j=1}^{\infty} \frac{(npq^j)^k}{k!} \frac{1}{n} e^{-nq^j} = \frac{p^k}{k!} \sum_{j=1}^{\infty} \sum_{n=1}^{m} n^{k-1} q^{jk} e^{-nq^j}.$$

The sum $\sum_{n=1}^{m} n^{k-1} q^{jk} e^{-nq^j}$ is decreasing in j and hence the sum over j will differ at most $\sum_{n=1}^{m} n^{k-1} e^{-n}$ from the corresponding integral

$$\int_0^{\infty} \sum_{n=1}^{m} n^{k-1} q^{xk} e^{-nq^x} \, dx. \qquad (6.1)$$

Since $n^{k-1} e^{-n}$ is summable, $(\log m)^{-1} \sum_{n=1}^{m} n^{k-1} e^{-n} \to 0$ as $m \to \infty$ and we can use the integral instead of the sum. The change of variables $y = q^x$ yields

$$\int_0^{\infty} q^{xk} e^{-nq^x} \, dx = \frac{1}{|\log q|} \int_0^1 y^{k-1} e^{-ny} \, dy \sim \frac{1}{|\log q|} \frac{(k-1)!}{n^k}$$

as $n \to \infty$. Hence,

$$\frac{p^k}{k!} \int_0^{\infty} n^{k-1} q^{xk} e^{-nq^x} \, dx \sim \frac{p^k}{k|\log q|} \frac{1}{n}$$

as $n \to \infty$ and hence

$$\frac{1}{\log m} \sum_{n=1}^{m} \frac{P(K_n = k)}{n} \to \frac{p^k}{k|\log q|}$$

as $m \to \infty$. $\square$

References

Aldous, D., 1989. Probability Approximations via the Poisson Clumping Heuristic. Springer, New York.

Barbour, A., Holst, L., Janson, S., 1992. Poisson Approximation. Oxford University Press, Oxford.

Baryshnikov, Y., Eisenberg, B., Stengle, G., 1995. A necessary and sufficient condition for the existence of the limiting probability of a tie for first place. Statist. Probab. Lett. 23, 203-209.

Brands, J.J.A.M., Steutel, F.W., Wilms, R.J.G., 1993. On the number of maxima in a discrete sample. Statist. Probab. Lett. 20, 209-218.

Eisenberg, B., Stengle, G., Strang, G., 1993. The asymptotic probability of a tie for first place. Ann. Appl. Probab. 3, 731-745.

Nerman, O., 1989. Personal communication.

Nordin, P., Olofsson, P., 1989. Convergence of maxima with Poisson approximation (in Swedish). Masters Thesis, Göteborg Universtiy, Göteborg, Sweden.

Resnick, S., 1992. Adventures in Stochastic Processes. Birkhäuser, Boston.

Råde, L., 1991. Problem E 3436. Amer. Math. Monthly. 98, 366.