The number of minima in a discrete sample

Memorandum COSOR 98-02, 1998, Eindhoven University of Technology

The number of minima in a discrete sample

F.W. Steutel *

Abstract

The number of times is considered that the minimum occurs in a sample from a discrete distribution. The special case of the geometric distribution is considered in some detail, and applied to the computation of the expected maximum of a sample from the Cantor distribution.

1. Introduction and summary

Brands et al. (1994) and Baryshnikov et al. (1995) consider the number K_n of values equal to the maximum in a sample from a distribution on the nonnegative integers. For $n \to \infty$, in some cases K_n tends to 1, in many cases it does not converge, and K_n tends to infinity, if the support of the distribution is finite. This last case is rather simple, and has not been given much attention. It is equivalent to the case where the number of *minima* of the sample is considered. In this note we look at this quantity, which we also denote by K_n. In Section 2 the general case is considered briefly. Section

deals in some more detail with the case of the geometric distribution, and in Section 4 these results are applied to a problem, a special case of which appeared in the Problem Section of the American Mathematical Monthly (see Diamond and Reznick (1997); a somewhat similar problem is considered in van Harn (1988)).

2. The general case

Let $M_1, \ldots, M_n$ be independent and distributed as M with

$$P(M = j) = p_j, P(M \leq j) = P_j,$$

for $j = 1, 2, \ldots$. Define

$$N_n = min(M_1, \ldots, M_n)$$

and

$$K_n = \#\{j : M_j = N_n\}.$$

*Department of Mathematics and Computing Science, Eindhoven University of Technology, Eindhoven, The Netherlands

We are interested in the distribution of (N_n, K_n). On account of symmetry one easily verifies the following results.

Proposition 1

$$\begin{aligned} P(N_n = r, K_n = k) &= \binom{n}{k} p_r^k (1 - P_r)^{n-k}, \\ P(N_n = r) &= (1 - P_{r-1})^n - (1 - P_r)^n, \\ P(K_n = k) &= \binom{n}{k} \sum_1^\infty p_r^k (1 - P_r)^{n-k}, \end{aligned}$$

where $P_0 = 0$.

Corollary 1

$$\begin{aligned} P(K_n = k) &= \binom{n}{k} p_1^k (1 - p_1)^{n-k} (1 + e(n)) \\ \mathrm{E} K_n &= n p_1 (1 + e(n)) \end{aligned}$$

as $n \to \infty$*, where* $e(n)$ *denotes an exponentially small contribution.*

Of course, the result in Corollary 1 is not very surprising; for large n the minimum will almost certainly be one, and $K_n$will be the number of ones in the sample, which is binomially distributed with succes probability p_1.

3. The geometric case

Now let M have a geometric distrtibution, i.e., let $p \in (0,1)$ and let

$$p_j = p^{j-1}(1-p), 1 - P_j = p^j, j = 1, 2, \ldots . \tag{1}$$

Whereas in general N_n and K_n are dependent, in this case they are not.

Proposition 2 *If* M *has a geometric distribution as given by (1), then* N_n *and* K_n *are independent, with*

$$P(N_n = r, K_n = k) = (1-p)^{rn} \binom{n}{k} (\frac{p}{1-p})^k .$$

Proof. Follows directly from Proposition 1 with (1), and from the product form of the formula above.

Corollary 2

$$\begin{aligned}
P(K_n = k) &= \binom{n}{k} \frac{(1-p)^k p^{n-k}}{1-p^n}, \\
P(N_n = r) &= p^{n(r-1)}(1-p^n),
\end{aligned}$$

$$\mathrm{E}z^{K_n} = \frac{(p+(1-p)z)^n - p^n}{1-p^n}, \tag{2}$$

$$\mathrm{E}z^{N_n} = \frac{1-p^n}{1-zp^n} z. \tag{3}$$

4. An application

In Problem 10621 (1997) one is requested to compute $1 - m_n$, where

$$m_n = 1 - \int_0^1 F^n(x)dx, \tag{4}$$

whith F a Cantor distribution function (with $p = 1/2$). This means that F is the distribution function of X given by

$$X = \sum_{j=1}^{\infty} (\frac{1}{3})^j Y_j, \tag{5}$$

where $Y_1, Y_2, \ldots$ are independent and distributed as Y with

$$P(Y = 0) = 1 - P(Y = 2) = p,$$

with $p \in (0,1)$. Clearly,

$$m_n = \mathrm{E}max(X_1, \ldots, X_n),$$

with the X's independent and distributed as X above. It follows from (5) that X has the form

$$X = 0.00002\ldots,$$

where the number of zeroes preceding the first 2 equals $M - 1$, with M geometrically distributed as in (1); M indicates the position of the first 2 in the triadic expansion of X. As a result we can apply the formulas derived in Section 3, and so we have

$$max(X_1, \ldots, X_n) =^d 2(\frac{1}{3})^{N_n} + (\frac{1}{3})^{N_n} max(X_1, \ldots, X_{K_n}),$$

with N_n and K_n as in Section 3, and N_n, K_n,and $X_1, X_2, \ldots$ independent. That is $max(X_1, \ldots, X_n)$ equals that X_j for which the first 2 comes first and for which the number represented by the remaining digits is maximal. Taking expectations we obtain

$$m_n = 2\mathrm{E}(\frac{1}{3})^{N_n} + \mathrm{E}(\frac{1}{3})^{N_n}\mathrm{E}m_{K_n},$$

where we used the independence of N_n, K_n and the X's. By (3) of Corollary 2, m_n can be expressed as

$$m_n = \frac{2}{3}\frac{1-p^n}{1-\frac{1}{3}p^n} + \frac{1}{3}\frac{1-p^n}{1-\frac{1}{3}p^n}\mathrm{E}\int_0^1 (1-F^{K_n}(x))dx.$$

Using (2) and (4) we get

$$\begin{aligned} m_n = 2\frac{1-p^n}{3-p^n} + \frac{1-p^n}{3-p^n}\int_0^1 \{1 - \frac{(p+(1-p)F(x))^n - p^n}{1-p^n}\}dx\} = \\ = 2\frac{1-p^n}{3-p^n} + \frac{1}{3-p^n}\int_0^1 \{1-(p+(1-p)F(x)^n\}dx = \\ = 2\frac{1-p^n}{3-p^n} + \frac{1}{3-p^n}\sum_{k=1}^{n}\binom{n}{k}p^{n-k}(1-p)^k m_k. \end{aligned} \tag{6}$$

From (6) we compute the first few values of m_n. For general p we obtain $m_1 = 1-p$, $m_2 = (1+p-3p^2+p^3)/(1+p-p^2)$, $m_3 = (2+5p-2p^2-17p^3+16p^4-4p^5)/(2+5p-2p^2-6p^3+3p^4)$. For $p = 1/2$ we find $m_1 = 1/2$, $m_2 = 7/10$, $m_3 = 4/5$ and $m_4 = 197/230$. The general expression for m_4 involv polynomials of degrees nine and eight in the numerator and denominator, respectively; for $p = 1/2$ the values of m_n can easily be calculated; we give $m_5 = 41/46$, $M_6 = 799/874$, and $m_7 = 8129/8740$. Of course, m_n tends to one as $n \to \infty$.

References

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

Brands, J.A.M., Steutel, F.W. and Wilms, R.J.G., On the number of maxima in a discrete sample, Statist. Prob. Letters 20 (1994), 209-217.

Diamond, H.G. and Reznick, B. (1997), Problem 10621, American Mathematical Monthly 104, Nr. 9, p. 870.

van Harn, K. (1988), Problem 208, solution by D. Gilat, Statistica Neerlandica 42, 70-71.