Expectation Value of the Lowest of a Set of Randomly Selected Integers

Expectation Value of the Lowest of a Set of Randomly Selected Integers

Adolph Baker
Ellen Baker

TR-11-98
November 1998

Expectation Value of the Lowest of a Set of Randomly Selected Integers

Adolph Baker
Department of Physics and Applied Physics
University of Massachusetts, Lowell
1 University Avenue
Lowell, MA 01854
abaker@erols.com

Ellen Baker
Division of Engineering and Applied Sciences
Harvard University
29 Oxford Street
Cambridge, MA 02138
ellie@eecs.harvard.edu

Abstract

Consider the set of positive integers 0, 1, 2, ..., D. If we pick N of them at random, where N < (D+1), what is the expectation (or average value) of the lowest-valued of the N picks? We briefly describe the image database search question that gave rise to this problem, and present a proof that the answer is (D-N+1)/(N+1).

1 Introduction

This paper addresses a problem that arises in the context of an image database search problem. Suppose that the criteria for determining similarity (or distance) between images is well-defined, and that a database of (D+1) images can thus be sorted automatically in order of similarity to any one of the images within it. If a user wishes to search the database for a particular image, this sorting mechanism can be used to reduce (on average, as compared to sequential search) the number of image inspections required to find the desired "target" image. One strategy is to pick N images at random from the database, and sort the database by distance from the one of these that is most similar in appearance to the "target". The question is, how far (on average) should one expect to search in this sorted list of images before finding the target? This problem can be reformulated mathematically as follows [EB98]:

Consider the set of positive integers 0, 1, 2, ..., D. If we pick N of them at random, where N < (D+1), what is the expectation (or average value) of the lowest-valued of the N picks?

The answer is (D-N+1)/(N+1).

2 Proof

$E(D, N)$, the expectation (or average) of the lowest-valued of the N picks, is:

$$\sum_{i=0}^{D-N+1} i \cdot (\text{The probability of having i as the lowest pick})$$

The probability of having i as the lowest pick is the number of combinations of $(D-i)$ of the integers taken $(N-1)$ at a time, divided by the total number of combinations of all the $(D+1)$ integers taken N at a time. Thus, noting that the first term in the summation vanishes, we have:

$$E(D,N) = \sum_{i=1}^{(D-N+1)} i \cdot \frac{\binom{D-i}{N-1}}{\binom{D+1}{N}} \qquad \text{Eq. 1}$$

Using the binomial coefficient identity (or recursion formula):

$$\binom{n}{m} = \binom{n+1}{m+1} - \binom{n}{m+1} \qquad \text{Eq. 2}$$

Equation 2 gives:

$$\binom{D-i}{N-1} = \binom{D-i+1}{N} - \binom{D-i}{N} \qquad \text{Eq. 3}$$

Substituting Equation 3 in Equation 1,

$$\binom{D+1}{N} \cdot E(D,N) = \sum_{i=1}^{(D-N+1)} \left[i \cdot \binom{D-i+1}{N} - i \cdot \binom{D-i}{N} \right] \qquad \text{Eq. 4}$$

In (only) the first term on the right of Equation 4, let $i = i' + 1$, $i' = i - 1$. This term then becomes:

$$\sum_{i=1}^{(D-N+1)} i \cdot \binom{D-i+1}{N} = \sum_{i'=0}^{(D-N)} (i'+1) \cdot \binom{D-i'}{N} \qquad \text{Eq. 5}$$

Since i' is just a dummy variable in the summation (you can call it anything; it disappears when the summation is carried out), we can now drop the prime and substitute Equation 5 in Equation 4 without the prime. The last term on the right of Equation 4, on the other hand, can just as easily be summed from $i = 0$ instead of $i = 1$, and breaking off the $(D-N+1)$ term of this sum, Equation 4 becomes:

$$\binom{D+1}{N} \cdot E(D,N) = \sum_{i=0}^{(D-N)} \left[i \cdot \binom{D-i}{N} + \binom{D-i}{N} - i \cdot \binom{D-i}{N} - (D-N+1)\binom{N-1}{N} \right]$$

Eq. 6

Cancelling the first and the third terms on the right of Equation 6, and since the last term vanishes because $\binom{n}{r} = 0$ for $n < r$,

$$\binom{D+1}{N} \cdot E(D,N) = \sum_{i=0}^{(D-N)} \binom{D-i}{N} \qquad \text{Eq. 7}$$

The sum of the form on the right in Equation 7 has been evaluated in the literature [PBM81]. It is:[1]

$$\sum_{k=0}^{n} \binom{a-k}{b} = \binom{a+1}{b+1} - \binom{a-n}{b+1} \qquad \text{Eq. 8}$$

Using Equation 8 to solve Equation 7:

$$\binom{D+1}{N} \cdot E(D,N) = \binom{D+1}{N+1} - \binom{N}{N+1} \qquad \text{Eq. 9}$$

Since the last term on the right of Equation 9 vanishes, we have

$$E(D,N) = \frac{(D+1)!/[(N+1)! \cdot (D-N)!]}{(D+1)!/[N! \cdot (D-N+1)!]}$$

Simplifying,

$$E(D,N) = \frac{D-N+1}{N+1}$$

3 References

[EB98] E. Baker, The Mug-Shot Search Problem: A Study of the Eigenface Metric, Search Strategies, and Interfaces in a System for Searching Facial Image Data. Doctoral thesis, Harvard University, Division of Engineering and Applied Sciences, Appendix B, November 1998.

[PBM81] A.P. Prudnikov, Y.A. Brichkov, and O.I. Marichev, Integrals and Series: Elementary Functions. Moscow “Science,” page 608, Eqn. 26, 1981 (in Russian).

1. This sum can be proved by the method of induction. It is seen to be satisfied for n=0 by means of identity 2. Then if we assume it to be true for a particular value of n, this can be used to show that is is likewise true for (n+1).