EXACT CUMULATIVE PROBABILITIES FOR THE MULTINOMIAL DISTRIBUTION

KENNETH J. BERRY AND PAUL W. MIELKE, JR.

Colorado State University

Kenneth J. Berry
Department of Sociology
Colorado State University
Fort Collins, CO 80523
Office: (303) 491-5070
Department: (303) 491-6044
FAX: (303) 491-2191
Internet: berry@lamar.colostate.edu

Paul W. Mielke, Jr.
Department of Statistics
Colorado State University
Fort Collins, CO 80523
Office: (303) 491-6465
Department: (303) 491-7277
FAX: (303) 491-7895
Internet: mielke@lamar.colostate.edu

EXACT CUMULATIVE PROBABILITIES FOR THE MULTINOMIAL DISTRIBUTION

[Abstract]

This article describes a FORTRAN program which computes exact cumulative probabilities for the multinomial distribution. Two recursive routines eliminate the need for factorial expressions and provide for efficient calculation.

EXACT CUMULATIVE PROBABILITIES FOR THE MULTINOMIAL DISTRIBUTION

It is often of interest to evaluate the outcome of a multinomial experiment. Consider k mutually exclusive exhaustive categories with associated probabilities $p_1, \ldots, p_k$, where

$$\sum_{j=1}^{k} p_j = 1 . \tag{1}$$

If N independent observations are made at random, then the probability that exactly n_j observations are in category j $(j = 1, \ldots, k)$ and

$$\sum_{j=1}^{k} n_j = N \tag{2}$$

is given by

$$P(n_1, \ldots, n_k) = \frac{N!}{\prod_{j=1}^{k} n_j!} \prod_{j=1}^{k} p_j^{n_j} . \tag{3}$$

The purpose of this paper is to describe subroutine MULT which calculates the exact multinomial point and cumulative probabilities for a realized arrangement of $n_1, \ldots, n_k$.

There are three distinct calculation problems inherent in the computation of the exact multinomial probability distribution. First, all possible arrangements of the N observations to the k categories must be generated. Second, for each arrangement, the number of permutations of observations to categories must be computed. Third, as each permutation is equally probable within an arrangement, the probability of a permutation must be calculated for each arrangement.

When $k \geq 2$, the number of possible frequency arrangements of the N observations in the k categories $(n_1, \ldots n_k)$ is given by

$$M = \binom{N+k-1}{k-1}. \tag{4}$$

Subroutine MULT utilizes a routine developed by Dunlap, Myers, and Silver (1984) to generate the M possible arrangements of the N observations to the k categories.

If $(n_1, \ldots, n_k)$ implicitly denotes the ith of M distinct arrangements, then there are

$$W_i = \frac{N!}{\prod_{j=1}^{k} n_j!} \tag{5}$$

possible permutations of the N observations to the k categories ($i = 1,\ldots,M$). The usual method of computing the required permutations, and the associated probabilties, is to compute log factorials with Stirling's approximation. Subroutine MULT utilizes a recursive routine which completely eliminiates all factorials, log factorials, and Stirling's approximation. Given

$$W_i = \frac{N!}{n_1!\,n_2!\cdots n_{k-1}!\,n_k!}, \tag{6}$$

then the implicitly-defined subsequent value given by

$$W_{i+1} = \frac{N!}{n_1!\,n_2!\cdots(n_{k-1}+1)!\,(n_k-1)!} \tag{7}$$

can be represented as

$$W_{i+1} = W_i\frac{n_k}{n_{k-1}+1}. \tag{8}$$

As the initial arrangement is always of the form $n_1 = n_2 = \cdots = n_{k-1} = 0$ and $n_k = N$, W_1 is always equal to 1. As the recursive routine accumulates products, there is a possibility that a W_i value might exceed the upper range of the computer. For this reason it is prudent to start the recursion with a small initial value (e.g., 10^{-50}) yielding M

scaled W_i values. The unscaled W_i values can then be obtained by dividing by the initial starting value.

For the ith of the M distinct arrangements of $n_1, \ldots, n_k$, it is necessary to compute the product of probabilities,

$$\prod_{j=1}^{k} p_j^{n_j} . \tag{9}$$

Given

$$Q_i = p_1^{n_1} p_2^{n_2} \cdots p_{k-1}^{n_{k-1}} p_k^{n_k} , \tag{10}$$

then the implicitly-defined subsequent value given by

$$Q_{i+1} = p_1^{n_1} p_2^{n_2} \cdots p_{k-1}^{n_{k-1}+1} p_k^{n_k-1} \tag{11}$$

can be represented as

$$Q_{i+1} = Q_i \frac{p_{k-1}}{p_k} . \tag{12}$$

Again, a small initial value (e.g., 10^{-50}) of Q_1 yields M scaled Q_i values.

Subroutine MULT consists of two distinct steps. The first step involves the number of recursions required to obtain the value associated with the observed arrangement of $n_1, \ldots, n_k$. The second step involves M recursions to obtain the conditional sum, S, of the recursively-defined values which are less than or equal to the value of the observed arrangement and, the unconditional sum, T, of all M recursively-defined values. The exact cumulative probability value, P-value$(n_1, \ldots, n_k)$, of the observed arrangement of $n_1, \ldots, n_k$ is given by P-value$(n_1, \ldots, n_k) = S/T$.

Program Language

Subroutine MULT is written in FORTRAN-77 in double precision. Comment lines provided for input/output specification and documentation. Input consists of the number of categories (k), the observed frequency for each category ($n_1, \ldots, n_k$), and the associated probabilities for each category ($p_1, \ldots, p_k$). Output consists of a summary table based on the input, and the point and cumulative probability values of the observed arrangement of observations to categories.

Application

Consider an example where $N = 200$ adult subjects are classified into $k = 4$ categories of marital status (i.e., single, married, widowed, and divorced) with $n_1 = 30$, $n_2 = 80$, $n_3 = 40$, and $n_4 = 50$, respectively. Census records yield the associated proportions of $p_1 = 0.20$, $p_2 = 0.35$, $p_3 = 0.15$, and $p_4 = 0.30$. The exact point probability is $P(30,80,40,50) = 4.78\times10^{-6}$ and the exact cumulative probability is P-value(30,80,40,50) = 0.030837. For this example, $k = 4$, $N = 200$, $M = 1,373,701$, and computation time on an IBM RISC-6000 work station was 2.47 seconds.

Availability

A listing of subroutine MULT and an appropriate driver program is available from Kenneth J. Berry, Department of Sociology, Colorado State University, Fort Collins, CO 80523, or by e-mail from Internet address: berry@lamar.colostate.edu.

Reference

Dunlap, W. P., Myers, L., & Silver, N. C. (1984). Exact multinomial probabilities for one-way contingency tables. *Behavior Research Methods, Instruments, & Computers, 16*, 54-56.