

EXACT GOODNESS-OF-FIT PROBABILITY TESTS FOR
ANALYZING CATEGORICAL DATA

PAUL W. MIELKE, JR. AND KENNETH J. BERRY
Colorado State University, Fort Collins

EXACT GOODNESS-OF-FIT PROBABILITY TESTS FOR
ANALYZING CATEGORICAL DATA

[Abstract]

Algorithms and associated FORTRAN subroutines for exact goodness-of-fit probability tests are presented. Classifications from two to six categories are considered. The use of recursion and an arbitrary initial value ensures computational efficiency.

EXACT GOODNESS-OF-FIT PROBABILITY TESTS FOR ANALYZING CATEGORICAL DATA

The assessment of goodness of fit for unordered categories is common in educational and psychological research. This paper presents high-speed recursion algorithms and associated subroutines for exact goodness-of-fit probability tests for up to six mutually-exclusive, unordered, exhaustive categories. Specifically, five separate FORTRAN-77 subroutines are provided for the analysis of 2, 3, 4, 5, and 6 categories, where the subroutines are goodness-of-fit analogs to the exact probability tests for $r \times c$ cross-classification tables presented by Mielke and Berry (1992).

The exact subroutines are conditional and have the advantage in practice of accounting for category probabilities under the null hypothesis when the probabilities may be either (1) fully specified or (2) fitted from the raw frequency data under a hypothesized model. Compared with chi-square goodness-of-fit tests, exact tests are free from any asymptotic assumptions, such as large expected category frequencies.

The same general algorithm is used for each of the five subroutines. Beginning with an arbitrarily-defined initial

value, a recursion procedure generates relative frequency values for all possible configurations of the N objects in the k categories, given that N and k are fixed. The required probability value is obtained by summing the relative frequency values equal to or less than the observed relative frequency value, and dividing by the unrestricted relative frequency total.

Subroutines

The k -fold multinomial probability distribution is given by

$$P(m_1, \dots, m_k | p_1, \dots, p_k, N) = N! \prod_{i=1}^k \frac{p_i^{m_i}}{m_i!}$$

where k is the number of categories, N is the number of objects, m_i is the observed frequency of category i ($i = 1, \dots, k$), $p_i > 0$ is the expected proportion for

category i ($i = 1, \dots, k$), $\sum_{i=1}^k p_i = 1$, $\sum_{i=1}^k m_i = N$, and m_1, \dots, m_k

are k non-negative integers. Exact goodness-of-fit probability tests for $2 \leq k \leq 6$ are obtained. Since $k - 1$ nested loops are associated with any k categories, the restriction of $k \leq 6$ is only because computation becomes

prohibitive for $k > 6$. The nested loops must account for all

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

distinct configurations of the category frequencies.

Because each of the five subroutines is a variation of the others, the following subroutine description is confined to $k = 4$.

Let $m_1 = x$, $m_2 = y$, $m_3 = z$, and $m_4 = N - x - y - z$. (Since N and $k = 4$ are fixed, the cell frequencies are totally described by x , y , and z .) Also let x_0 , y_0 , and z_0 denote the observed values of x , y , and z . If the order specification of x , y , and z implies that z depends on x and y , and y depends on x , then the following bounds hold for x , y , and z :

$$0 \leq x \leq N,$$

$$0 \leq y \leq N-x,$$

and

$$0 \leq z \leq N-x-y.$$

In accordance with the previously-defined order specification of x , y , and z , the conditional recursively-defined probability adjustments from (x_1, y_1, z_1) on step 1 of the recursion to (x_2, y_2, z_2) on step 2 of the recursion are given by

$$\frac{P(x_2, y_2, z_2 | p_1, p_2, p_3, N)}{P(x_1, y_1, z_1 | p_1, p_2, p_3, N)}$$

where

$$P(x, y, z | p_1, p_2, p_3, N) = \frac{N! p_1^x p_2^y p_3^z (1-p_1-p_2-p_3)^{N-x-y-z}}{x! y! z! (N-x-y-z)!}$$

Starting with an arbitrarily-defined machine-dependent initial value, e.g., 10^{-200} , the subroutine consists of two distinct steps. The first step involves $N - m_k$ recursions to obtain the value associated with the observed frequency configuration, and the second step involves M recursions to obtain the probability value. For $k = 4$, the first step obtains the value

$$U(x_o, y_o, z_o) = D \times P(x_o, y_o, z_o | p_1, p_2, p_3, N)$$

where D is the initial value, and the second step determines

(1) the conditional sum, S , of the recursively-defined values of

$$U(x, y, z) = D \times P(x, y, z | p_1, p_2, p_3, N)$$

satisfying $U(x, y, z) \leq U(x_0, y_0, z_0)$, and (2) the unconditional sum, T , of all M values of $U(x, y, z)$. The exact probability value, P , associated with the observed frequencies x_0 , y_0 , and z_0 is given by $P = S/T$.

Example

Consider an example where $N = 10$ learning-disabled elementary school children are classified into $k = 4$ categories of learning disability with $m_1 = 4$, $m_2 = 3$, $m_3 = 2$, and $m_4 = 1$. Previous research indicates the expected proportions to be $p_1 = 0.18$, $p_2 = 0.12$, $p_3 = 0.40$, and $p_4 = 0.30$. Under the null hypothesis of no difference between the observed and expected frequencies, the exact goodness-of-fit probability is $P = 0.0361$.

Program Language

The five subroutines (i.e., EXGOF2, EXGOF3, EXGOF4, EXGOF5, and EXGOF6) are written in ANSI-standard FORTRAN-77 in double precision. Comment lines provide input/output

specification and documentation. Input for each of the subroutines consists of the number of categories (k), the observed cell frequencies (m_1, \dots, m_k), and the expected cell proportions (p_1, \dots, p_k). Each of the five subroutines returns a summary of the input and the exact goodness-of-fit probability value.

Availability

Listings of the five subroutines and an appropriate driver program are 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

Mielke, P. W. and Berry, K. J. (1992). Fisher's exact probability test for cross-classification tables. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 52, 97-101.