TWO BY TWO TABLES WITH SMALL NUMBERS: HYPERGEOMETRIC PROBABILITIES

In χ^2 tests of fit for multinomial distributions, if some categories have expected numbers less than 5 for the estimated or hypothesized parameters, then adjoining categories may be combined until each category has expected number larger than 5. In contingency tables, however, combining categories would destroy the structure of the table, unless, for example, two whole rows or whole columns were combined. For a 2×2 table, we would be left with a 1×2 or 2×1 table where testing for independence or homogeneity would no longer make sense. One can deal with 2×2 contingency tables by a method not involving χ^2 tests, as follows.

Suppose we have a 2×2 contingency table and assume for it a π model, with some fixed values of $n_{1\cdot}$ and $n_{2\cdot}$ adding up to some n. We want to test the homogeneity hypothesis H_0 which in this case is $\pi_{11} = \pi_{21} = \pi_1$ for some π_1, from which it follows that also

$$\pi_{12} = 1 - \pi_{11} = 1 - \pi_{21} = \pi_{22} = \pi_2$$

say. To test H_0 we can consider the conditional distribution of n_{11} given the column total $n_{\cdot 1}$. Recall that under H_0, the maximum likelihood estimate of π_1 is $n_{\cdot 1}/n$ and the expected number E_{11} is $n_{1\cdot}n_{\cdot 1}/n$. So if we observe n_{11} much larger or much less than E_{11} we will tend to reject H_0, but if n_{11} is not too different from E_{11} we will not reject H_0. To get quantitative criteria, let's evaluate the conditional probability

$$\Pr(n_{11} = j | n_{\cdot 1} = m) \tag{1}$$

for nonnegative integers j and m. A binomial coefficient $\binom{a}{b}$ is defined for nonnegative integers a and b as $a!/(b!(a-b)!)$ or as 0 if $b < 0$ or $b > a$. Recall that in the π model, n_{ij} for different i are independent. Note that under H_0 n_{11} has a binomial $(n_{1\cdot}, \pi_1)$ distribution, n_{21} has a binomial $(n_{2\cdot}, \pi_1)$ distribution, and $n_{\cdot 1}$ has a binomial (n, π_1) distribution. So for the numerator of (1) we get a product of two binomial probabilities $B_1 B_2$ where

$$B_1 = \Pr(n_{11} = j) = \binom{n_{1\cdot}}{j} \pi_1^j \pi_2^{n_{1\cdot} - j}$$

Date: 18.650, Nov. 6, 2015.

and

$$B_2 = \Pr(n_{21} = m - j) = \binom{n_{2\cdot}}{m-j} \pi_1^{m-j} \pi_2^{n_{2\cdot}+j-m}.$$

The denominator of (1) is another binomial probability B_3,

$$B_3 = \binom{n}{m} \pi_1^m \pi_2^{n-m}.$$

In the conditional probability (1), $B_1 B_2 / B_3$, the exponent of π_1 is $j + (m - j) - m = 0$ and the exponent of $\pi_2 = 1 - \pi_1$ is

$$n_{\cdot 1} - j + n_{2\cdot} + j - m - (n - m) = 0$$

also, so (1) does not depend on π_1 (or π_2) and equals

$$h(j, n_{1\cdot}, m, n) := \binom{n_{1\cdot}}{j} \binom{n_{2\cdot}}{m-j} / \binom{n}{m}. \tag{2}$$

Here is an alternate formulation. Suppose we have a collection of n objects, m of which have a property A. Another subset S of k of the n objects will be said to form a *sample*. Then we have a 2×2 table of nonnegative integers n_{ij}, $i, j = 1, 2$, where n_{11} is number of elements in $A \cap S$, n_{12} the number in $S \setminus A$, n_{21} the number in $A \setminus S$, and n_{22} the number in neither A nor S. Thus the row totals $n_{1\cdot} = k$ and $n_{2\cdot} = n - k$, and the column totals $n_{\cdot 1} = m$, $n_{\cdot 2} = n - m$. These four totals are all fixed, which implies that if we know any n_{ij}, the other three are all determined. This is the π model with fixed $n_{1\cdot} = k$ and $n_{2\cdot} = n - k$ and conditional on the observed first column total $n_{\cdot 1} = m$, as above.

For given n, m and k let $Y = n_{11}$, the number of objects in $S \cap A$, so that $P(Y = j) = h(j, k, m, n)$ for each possible j. The maximum possible value of j is $\min(k, m)$ and the minimum possible value is $\max(0, m + k - n)$. Here $j \geq m + k - n$ because $0 \leq n_{22} = n - k - m + j$.

In R, $P(Y \leq i)$ is found as phyper$(i, m, n - m, k)$. If one wanted $h(j, k, m, n)$, R would find it as dhyper$(j, m, n - m, k)$.

If $Y = n_{11}$ is observed with phyper$(Y, k, n - k, m) \leq \alpha/2$ then we would reject H_0 in a one-sided test. If that does not occur, consider also the probability phyper$(m - Y, n - k, k, m)$ and if that is $\leq \alpha/2$, reject homogeneity H_0. If neither of these things occurs, do not reject H_0.

The test of homogeneity in the given situation is known as "Fisher's exact test" (Rice, §13.2) of homogeneity H_0 in a 2×2 table.