EXACT SMALL SAMPLE TESTS FOR CONTINGENCY TABLES.

A Query by Katri Veldi, Department of Zoology, University of Toronto.

by

BU-365-M D. S. Robson April, 1971

Abstract

Calculation of the exact sampling distribution of the contingency chi-square statistic is illustrated in detail for a 2×3 table. This calculation entails the enumeration of all possible 2×3 tables with the same marginal totals, calculation of the conditional probability of each outcome (as a product of hypergeometric probabilities) and calculation of the chi-square statistic of each outcome. A simplification results when the test statistic is calculated as the sum of single degree of freedom chi-squares corresponding to each probability factor in the product of hypergeometric probabilities.

Paper No. BU-365-M in the Biometrics Unit Mimeograph Series, Department of Plant Breeding and Biometry, Cornell University, Ithaca, New York 14850.

EXACT SMALL SAMPLE TESTS FOR CONTINGENCY TABLES.

A Query by Katri Veldi, Department of Zoology, University of Toronto.

by

D. S. Robson

BU-365-M April, 1971

Calculation of the exact sampling distribution of the contingency chi-square statistic of a 2 × 3 table with fixed marginal totals

			a
			n-a
b	c'	n-b-c'	n

involves an enumeration of all possible 2 × 3 tables with these fixed marginal totals, calculation of the probability of each such table (under the null hypothesis) and calculation of the chi-square statistic for each such table. Any one such table is of the form

x	y'	a-x-y'	a
b-x	c'-y'	n-a-b-c'+x+y'	n-a
b	c'	n-b-c'	n

and the probability of this outcome is

$$P(x,y') = \frac{a!b!c'!(n-a)!(n-b-c')!}{n!x!y'!(b-x)!(c'-y')!(a-x-y')!(n-a-b-c'+x+y')!}$$

and the corresponding value of the X^2 statistic is

$$X^2(x,y') = \frac{1}{a(n-a)}\left[\frac{(nx-ab)^2}{b} + \frac{(ny-ac')^2}{c'} + \frac{\{n(x+y')-a(b+c')\}^2}{n-b-c'}\right].$$

Paper No. BU-365-M in the Biometrics Unit Mimeograph Series, Department of Plant Breeding and Biometry, Cornell University, Ithaca, New York 14850.

In order to utilize existing probability tables and also facilitate the enumeration of all possible outcomes, we may relabel the table entries as

x	y-x	a-y	a
b-x	c-b-y+x	n-c-a+y	n-a
b	c-b	n-c	n

and write

$$P(x,y) = P(y)P(x|y)$$

$$= \frac{\binom{a}{y}\binom{n-a}{c-y}}{\binom{n}{c}} \cdot \frac{\binom{y}{x}\binom{c-y}{b-x}}{\binom{c}{b}}$$

while now

$$\chi^2(x,y) = \frac{1}{a(n-a)}\left[\frac{(nx-ab)^2}{b} + \frac{\{n(y-x)-a(c-b)\}^2}{c-b} + \frac{(ny-ac)^2}{n-c}\right].$$

The two probability distributions

$$P(y) = \frac{\binom{a}{y}\binom{n-a}{c-b-y}}{\binom{n}{c-b}} \quad \text{and} \quad P(x|y) = \frac{\binom{y}{x}\binom{c-y}{b-x}}{\binom{c}{b}}$$

are hypergeometric functions and are tabulated by Lieberman and Owens. The enumeration proceeds by letting y vary through the range (of integers)

$$\max(0,a+c-n) \leq y \leq \min(a,c)$$

and, for each such value of y, letting x vary through the (integer) range

$$\max(0,b+y-c) \leq x \leq \min(b,y).$$

Numerical Example

For the fixed marginal totals:

			4 = a
			5 = n-a
4 = b	2 = c-b	3 = n-c	9 = n

we have

$$P(y) = \frac{\binom{4}{y}\binom{5}{6-y}}{\binom{9}{6}} \quad \text{and} \quad P(x|y) = \frac{\binom{y}{x}\binom{6-y}{4-x}}{\binom{6}{4}}$$

with

$$\max(0,1) = 1 \le y \le 4 = \min(4,6)$$

and for

$$y = 1, \quad \max(0,-1) = 0 \le x \le 1 = \min(4,1)$$

$$y = 2, \quad \max(0,0) = 0 \le x \le 2 = \min(4,2)$$

$$y = 3, \quad \max(0,1) = 1 \le x \le \min(4,3) = 3$$

$$y = 4, \quad \max(0,2) = 2 \le x \le 4 = \min(4,4)$$

so there are altogether 2 + 3 + 3 + 3 = 11 possible outcomes to enumerate:

$$\underline{y = 1} \qquad P(y = 1) = \frac{\binom{4}{1}\binom{5}{5}}{\binom{9}{6}} = \frac{2}{42}$$

$$\underline{x = 0} \qquad P(x = 0|y = 1) = \frac{\binom{1}{0}\binom{6-1}{4-0}}{\binom{6}{4}} = \frac{5}{15}$$

$$X^2(0,1) = \frac{1}{4(5)}\left[\frac{(0-16)^2}{4} + \frac{(9-8)^2}{2} + \frac{(9-24)^2}{3}\right]$$

$$= 3.2 + .025 + 3.75 = \boxed{6.975}$$

$$P(0,1) = \frac{2}{42} \cdot \frac{5}{15} = \boxed{\frac{10}{630}}$$

$\underline{x = 1}$ $\quad P(1|1) = \frac{10}{15}$

$$\chi^2(1,1) = \frac{1}{20}\left[\frac{5^2}{4} + \frac{8^2}{2} + \frac{15^2}{3}\right] = \boxed{5.6625}$$

$$P(1,1) = \boxed{\frac{20}{630}}$$

$\underline{y = 2}$ $\quad P(y = 2) = \frac{15}{42}$

$\underline{x = 0}$ $\quad P(0|2) = \frac{1}{15}$ $\quad \chi^2(0,2) = \frac{1}{20}\left[\frac{16^2}{4} + \frac{10^2}{2} + \frac{6^2}{3}\right] = \boxed{6.3}$

$$P(0,2) = \boxed{\frac{15}{630}}$$

$\underline{x = 1}$ $\quad P(1|2) = \frac{8}{15}$ $\quad \chi^2(1,2) = \frac{1}{20}\left[\frac{5^2}{4} + \frac{1^2}{2} + \frac{6^2}{3}\right] = \boxed{.9375}$

$$P(1,2) = \boxed{\frac{120}{630}}$$

$\underline{x = 2}$ $\quad P(2|2) = \frac{6}{15}$ $\quad \chi^2(2,2) = \frac{1}{20}\left[\frac{2^2}{4} + \frac{8^2}{2} + \frac{6^2}{3}\right] = \boxed{2.25}$

$$P(2,2) = \boxed{\frac{90}{630}}$$

$\underline{y = 3}$ $\quad P(y = 3) = \frac{20}{42}$

$\underline{x = 1}$ $\quad P(1|3) = \frac{3}{15}$ $\quad \chi^2(1,3) = \frac{1}{20}\left[\frac{5^2}{4} + \frac{10^2}{2} + \frac{3^2}{3}\right] = \boxed{2.9625}$

$$P(1,3) = \boxed{\frac{60}{630}}$$

$\underline{x = 2}$ $\quad P(2|3) = \frac{9}{15}$ $\quad \chi^2(2,3) = \frac{1}{20}\left[\frac{2^2}{4} + \frac{1^2}{2} + \frac{3^2}{3}\right] = \boxed{.225}$

$$P(2,3) = \boxed{\frac{180}{630}}$$

$\underline{x = 3}$ $\quad P(3|3) = \frac{3}{15}$ $\quad \chi^2(3,3) = \frac{1}{20}\left[\frac{11^2}{4} + \frac{8^2}{2} + \frac{3^2}{3}\right] = \boxed{3.2625}$

$$P(3,3) = \boxed{\frac{60}{630}}$$

$\underline{y = 4}$ $\quad P(y = 4) = \frac{5}{42}$

$\underline{x = 2}$ $\quad P(2|4) = \frac{6}{15}$ $\quad \chi^2(2,4) = \frac{1}{20}\left[\frac{2^2}{4} + \frac{10^2}{2} + \frac{12^2}{3}\right] = \boxed{4.95}$

$$P(2,4) = \boxed{\frac{30}{630}}$$

$\underline{x = 3}$ $\quad P(3|4) = \frac{8}{15}$ $\quad \chi^2(3,4) = \frac{1}{20}\left[\frac{11^2}{4} + \frac{1^2}{2} + \frac{12^2}{3}\right] = \boxed{3.9375}$

$$P(3,4) = \boxed{\frac{40}{630}}$$

$\underline{x = 4}$ $\quad P(4|4) = \frac{1}{15}$ $\quad \chi^2(4,4) = \frac{1}{20}\left[\frac{20^2}{4} + \frac{8^2}{2} + \frac{12^2}{3}\right] = \boxed{9.0}$

$$P(4,4) = \boxed{\frac{5}{630}}$$

Summary

chi-square	probability	tail probability
0.225	.2857	.
0.9375	.1905	.
2.25	.1429	.
2.9625	.0952	.
3.2625	.0952	.
3.9375	.0649	.
4.95	.0476	.1270
5.6625	.0317	.0794
6.3	.0238	.0476
6.975	.0159	.0238
9.0	.0079	.0079

Thus a chi-square value of 6.3 or greater would be significant at the 5% level (actually 4.76% level), and a value of 9.0 would be significant at the 1% level (actually 0.79% level). These results agree very closely with those obtained from the chi-square table which, for 2 degrees of freedom, gives $\chi^2_{.05} = 5.99$ and $\chi^2_{.01} = 9.21$.

Partitioned Chi-Square

An alternative approach to calculating essentially this same table involves partitioning $\chi^2(x,y)$ into

$$\chi^{*2}(x,y) = \chi^2(y) + \chi^2(x|y)$$

where $\chi^2(y)$ is calculated from the 2 × 2 table:

y	a-y	a
c-y	n-a-c+y	n-a
c	n-c	n

$$\chi^2(y) = \frac{n(ny-ac)^2}{ac(n-a)(n-c)}$$

and $\chi^2(x|y)$ from the 2 × 2 table:

x	y-x	y
b-x	c-b-y+x	c-y
b	c-b	c

$$\chi^2(x|y) = \frac{c(cx-by)^2}{by(c-b)(c-y)}$$

so that

$$\chi^{*2}(x,y) = \frac{n(ny-ac)^2}{ac(n-a)(n-c)} + \frac{c(cx-by)^2}{by(c-b)(c-y)} .$$

Following through exactly the same calculations for $P(x,y) = P(y)\ P(x|y)$ but calculating the somewhat simpler $\chi^{*2}(x,y)$ instead of $\chi^2(x,y)$ for each (x,y)-pair gives the following sampling distribution of $\chi^{*2}(x,y)$:

(chi-square)	chi-square*	probability	tail probability
(0.225)	0.225	.2857	.
(0.9375)	1.275	.1905	.
(2.25)	2.4	.1429	.
(2.9625)	3.225	.0952	.
(3.2625)	3.225	.0952	.
(3.9375)	3.975	.0649	.
(4.95)	5.1	.0476	.1270
(5.6625)	6.225	.0317	.0794
(6.3)	6.9	.0238	.0476
(6.975)	8.025	.0159	.0238
(9.0)	9.6	.0079	.0079

As an illustration, let

$$\underline{y = 3} \qquad P(y=3) = \frac{20}{42} \quad \chi^2(y=3) = \frac{9(27-24)^2}{4(6)(5)(3)} = \frac{9}{40}$$

$$\underline{x = 1} \qquad P(1|3) = \frac{3}{15} \quad \chi^2(1|3) = \frac{6(6-12)^2}{4(3)(2)(3)} = 3$$

$$\underline{x = 2} \qquad P(2|3) = \frac{9}{15} \quad \chi^2(2|3) = \frac{6(12-12)^2}{4(3)(2)(3)} = 0$$

$$\underline{x = 3} \qquad P(3|3) = \frac{3}{15} \quad \chi^2(3|3) = \frac{6(18-12)^2}{4(3)(2)(3)} = 3$$

so that

$$P(1,3) = \frac{60}{630} \quad \chi^{*2}(1,3) = \frac{9}{40} + 3 = 3.225$$

$$P(2,3) = \frac{180}{630} \quad \chi^{*2}(2,3) = \frac{9}{40} + 0 = .225$$

$$P(3,3) = \frac{60}{630} \quad \chi^{*2}(3,3) = \frac{9}{40} + 3 = 3.225$$

Note that the significance of $\chi^2(y)$ and $\chi^2(x|y)$ may be tested separately, and were it not for the discrete nature of these probability distributions the two tests would be statistically independent. If for each value of y the test of $\chi^2(x|y)$ were of constant size α then the tests would be independent even though the probability distribution of $\chi^2(x|y)$ depends on y; i.e., even though $\chi^2(y)$ and $\chi^2(x|y)$ are not themselves statistically independent. Introducing a randomization step in the test of $\chi^2(x|y)$ would enable us to achieve constant α, and hence independence, but this device lacks intuitive appeal. Asymptotically, of course, the two statistics $\chi^2(y)$ and $\chi^2(x|y)$ are independent and identically distributed as chi-square variables on one degree of freedom, and asymptotically $\chi^2(x,y) = \chi^{*2}(x,y)$ in probability.

The separate testing of $\chi^2(y)$ and $\chi^2(x|y)$ can in practice be achieved without actually calculating the numerical values of these two statistics. Tables are already available giving the critical values of y in the hypergeometric distribution $P(y)$ and critical values of x in the hypergeometric distribution $P(x|y)$. In order to combine these two separate tests into one test, however, it is necessary to calculate the exact sampling distribution of $\chi^{*2}(x,y)$ for small sample sizes. Fortunately, the chi-square approximation improves very rapidly with increasing sample size and usually, as in the preceding numerical example, one will find that he wasted his time in calculating exact results.

The partitioning approach outlined here may be extended to an $r \times c$ contingency table. Combining any c-1 columns (say the first c-1) produces an $r \times 2$ table which may then be partitioned in the above manner into r-1 2×2 tables. Deleting the i^{th} column leaves an $r \times c-1$ table upon which this process may be repeated. Asymptotically, all of the resulting 2×2 tests produced by this method are statistically independent.