TECHNOMETRICS VOL. 6, NO. 2 MAY, 1964

Large Sample Simultaneous Confidence Intervals for Multinomial Proportions

C. P. QUESENBERRY* AND D. C. HURST

Montana State College and
Virginia Polytechnic Institute

In this paper a method is presented for obtaining a set of simultaneous confidence intervals for the probabilities of a multinomial distribution. Large sample size is assumed and a lower bound given for the confidence coefficient. Use of the method to obtain a ranking of the proportions is illustrated. A computational scheme is illustrated by an example.

1. INTRODUCTION

Problems involving sampling from a multinomial distribution are of frequent occurrence in science and engineering. The classification of sample units according to a number of specified quality levels, studies of the proportions of failures of a machine due to different modes of failure, etc., provide examples where multinomial models may be appropriate. In sampling from a multinomial distribution the object is often to estimate the cell proportions with confidence intervals. One approach to this estimation problem is to consider each cell versus the remaining ones as a binomial distribution, and to make a set of binomial confidence interval estimates for the individual cell proportions (cf. Cochran, 1963). This approach does not allow one to assess the value of the confidence coefficient for the entire set of intervals, however, or to make statements concerning the relative values of the proportions.

Let $n_1, n_2, \cdots, n_k$ denote the observed cell frequencies in a sample of size N from the multinomial distribution, i.e.

$$P(n_1, n_2, \cdots, n_k) = N!\, \pi_1^{n_1}\pi_2^{n_2} \cdots \pi_k^{n_k}/n_1!\, n_2! \cdots n_k!, \tag{1.1}$$

where

$$\sum_{i=1}^{k} \pi_i = 1 \quad \text{and} \quad \sum_{i=1}^{k} n_i = N.$$

We wish to obtain a set of intervals $\{S_i\}$ $i = 1, 2, \cdots, k$; such that

$$P\left\{\bigcap_{i=1}^{k} (\pi_i \epsilon S_i)\right\} \geq 1 - \alpha, \tag{1.2}$$

for a specified value of α. In words, we require that the probability that every interval S_i contains its corresponding probability parameter be at least $1 - \alpha$.

* An earlier version of this paper was prepared while both authors were in residence at V. P. I. The earlier paper was sponsored by the Army Ballistic Missile Agency, Contract No. DA-36-034-ORD-2531 RD.

2. Derivation of Intervals

Pearson (1900) has shown that if $E(n_i) = N\pi_i$ is sufficiently large for all $i = 1, 2, \cdots, k$; the statistic

$$X^2 = \sum_{i=1}^{k} (n_i - N\pi_i)^2/N\pi_i \tag{2.1}$$

is distributed approximately as a chi-square variate with $(k - 1)$ d.f. We shall assume that N is large enough for this approximation to be adequate. If N is large enough to make the *smallest* expected cell frequency $N\pi_i$ at least 5, the approximation should be good. This recommendation is probably conservative. See Cochran (1954) for a discussion of this approximation. In order to use this recommendation in practice, it will be necessary to have some knowledge as to the smallest cell probability. Assuming the approximation is good we have

$$P\left\{\sum_{i=1}^{k} (Np_i^2/\pi_i) - N \leq \chi^2_{\alpha,k-1}\right\} \doteq 1 - \alpha, \qquad p_i = n_i/N \tag{2.2}$$

where $\chi^2_{\alpha,k-1}$ is the upper α percentage point of the chi-square distribution. For fixed α, k and N the inequality

$$\sum_{i=1}^{k} p_i^2/\pi_i \leq (\chi^2_{\alpha,k-1}/N) + 1 \equiv C \tag{2.3}$$

defines a region in the parameter space, and if the equals sign is taken the equation gives the bounding surface of the region. The intersection of this surface with the hyperplane represented by

$$\sum_{i=1}^{k} \pi_i = 1 \tag{2.4}$$

is a k-dimensional hypercurve. Maximum and minimum values will now be found for each π_i $(i = 1, 2, \cdots, k)$ on this curve.

From equation (2.3) we have

$$\pi_i = p_i^2\Big/\left(C - \sum_{\substack{j=1\\ j\neq i}}^{k} p_j^2/\pi_j\right), \qquad (i = 1, 2, \cdots, k) \tag{2.5}$$

So the function to be maximized and minimized is

$$Q = (1 - \lambda)p_i^2\Big/\left(C - \sum_{\substack{j=1\\ j\neq i}}^{k} p_j^2/\pi_j\right) - \lambda\left(\sum_{\substack{j=1\\ j\neq i}}^{k} \pi_j - 1\right), \tag{2.6}$$

where λ is the Lagrange multiplier. Differentiating with respect to π_j and π_m and setting these expressions equal to zero gives

$$\left.\begin{aligned} \lambda p_i^2/(\lambda - 1)\pi_i^2 &= p_j^2/\pi_j^2 \\ \lambda p_i^2/(\lambda - 1)\pi_i^2 &= p_m^2/\pi_m^2 \end{aligned}\right\} \qquad \begin{aligned} &i,\ j,\ m = 1, 2, \cdots, k \\ &j \neq i, \quad m \neq i, \quad m \neq j \end{aligned} \tag{2.7}$$

Therefore

$$\pi_m = p_m\pi_j/p_j \tag{2.8}$$

Solving these $(k - 2)$ equations with equations (2.3) and (2.4) gives

$$\pi_i = \frac{C + 2p_i - 1 \pm [(C + 2p_i - 1)^2 - 4Cp_i^2]^{\frac{1}{2}}}{2C} \tag{2.9}$$

Or, rewriting in a more convenient computing form

$$\pi_i = \frac{\chi^2 + 2n_i \pm \{\chi^2[\chi^2 + 4n_i(N - n_i)/N]\}^{\frac{1}{2}}}{2(N + \chi^2)}, \tag{2.9$'$}$$

where $\chi^2 = \chi^2_{\alpha,k-1}$.

3. Conservativeness of the Intervals

If we denote the interval obtained from (2.9) by S_i then the set $\{S_i\}$ $i = 1, 2, \cdots, k$ satisfies (1.2). We denote the Cartesian product of the sets S_1, $S_2, \cdots, S_k$ by S, i.e.

$$S = S_1 \times S_2 \times \cdots S_k. \tag{3.1}$$

Also, let Ω denote the set of points satisfying (2.3) and (2.4), i.e.

$$\Omega = \left\{(\pi_1, \pi_2, \cdots, \pi_k) \,\middle|\, \sum_{i=1}^{k} p_i^2/\pi_i \le C \quad \text{and} \quad \sum_{i=1}^{k} \pi_i = 1\right\}. \tag{3.2}$$

From the method of derivation of the S_i's it is clear that

$$\Omega \subset S. \tag{3.3}$$

Since Ω is a confidence region with coefficient $1 - \alpha$, then S is a confidence region with coefficient not less than $1 - \alpha$. We shall call this the conservative property of the region S. In general, the conservativeness of the region S will depend upon the true values of the proportions $\pi_1, \pi_2, \cdots, \pi_k$ and upon their number k. Intuitively, one would expect S to become more conservative with increasing k, and for this reason the method to be most useful for small values of k, i.e. for $k = 3$ or 4.

We know of no way to assess the conservativeness of S in general, however, some empirical studies have been made for selected values of the parameters. These studies were performed by drawing samples from multinomial distributions with known parameter values, computing a set of intervals for each sample, and counting the number of sets of intervals which actually covered all of the parameters. The whole procedure was performed on an I.B.M. 650 Computer.

One thousand samples of size one hundred each were drawn from a trinomial distribution with parameter values (.3333, .3333, .3334). Sets of confidence intervals were computed with $1 - \alpha = .95$. The actual proportion of sets which contained all parameter values was .981. Smaller samples were drawn from the trinomial for other values of the parameters, and appeared to give about the same level of conservativeness. Some computing was also performed for larger values of k and the procedure appeared to become more conservative with increasing k; however the sample size here was small and the results can not be considered conclusive.

4. Example

A manufacturer knows that a machine which he produces commonly fails by one of 9 distinct modes of failure. In order to institute a product improvement program he would like to assess the relative importance of the different modes of failure.

Records are available for 870 machines which have failed and the data are given in Table 1. Cell 1 consists of all failures which could not be classified into the other 9. (It may be noted that the cells are numbered here according to the value of the observed cell frequency. This is merely for computational convenience.)

TABLE 1

Data for Example

Mode of Failure (Cell)	1	2	3	4	5	6	7	8	9	10
Frequency	5	11	19	30	58	67	92	118	173	297

A set of simultaneous confidence intervals for the cell probabilities is computed in Table 2. Here $\alpha = .1$ was chosen and from tables of the chi-square

TABLE 2

Computations for Example

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10) $\frac{(8)}{2(N+\chi^2)}$	(11) $\frac{(9)}{2(N+\chi^2)}$
n_i	$n-n_i$	$4n_i(N-n_i)$	$(3)/N$	$\chi^2[\chi^2+(4)]$	$\sqrt{(5)}$	χ^2+2n_i	$(7)-(6)$	$(7)+(6)$		
5	865	17,300	19.885	507.60	22.53	24.68	2.15	47.21	.0012	.0267
11	859	37,796	43.444	853.53	29.21	36.68	7.47	65.89	.0042	.0372
19	851	64,676	74.340	1307.20	36.15	52.68	16.53	88.83	.0093	.0502
30	840	100,800	115.862	1916.89	43.78	74.68	30.90	118.46	.0175	.0670
58	812	188,384	216.533	3395.12	58.27	130.68	72.40	188.96	.0409	.1068
67	803	215,204	247.361	3847.79	62.03	148.68	86.65	210.71	.0490	.1191
92	778	286,304	329.085	5047.80	71.05	198.68	127.63	269.73	.0721	.1525
118	752	354,944	407.982	6206.30	78.78	250.68	171.90	329.46	.0972	.1862
173	697	482,324	554.395	8356.18	91.41	360.68	269.27	452.09	.1522	.2556
297	573	680,724	782.441	11704.74	108.19	608.68	500.49	716.87	.2829	.4052
Checks:										
$N = 870$	$9N = 7830$	2428456	2791.328	43143.14	601.40*	1886.80	1285.40	2488.20	.7265	1.4065

$\chi^2_{.1,9} = 14.6837,\ 2(N+\chi^2_{.1,9}) = 1769.3674,$

* Obtained by summing column.

distribution we obtain $\chi^2_{.1,9} = 14.6837$. Table 2 presents a systematic procedure for computing the intervals. Columns (10) and (11) give the lower and upper points of the intervals, respectively.

Since interest here is primarily in comparing the cell probabilities, we present these results for quick comparisons in the following form:

$\pi_1 \quad \pi_2 \quad \pi_3 \quad \pi_4 \quad \pi_5 \quad \pi_6 \quad \pi_7 \quad \pi_8 \quad \pi_9 \quad \pi_{10}$

Here we have underscored the cell probabilities in subgroups according to whether or not the intervals obtained have points in common. For example, the interval (.0012, .0267) for π_1 has points in common with the intervals

(.0042, .0372), (.0093, .0502), and (.0175, .0670) for π_2, π_3 and π_4, respectively; but not with the interval (.0409, .1068) for π_5.

In view of the suspected increasing conservativeness of the method with increasing k, with $k = 10$ as in this example, we should have a quite conservative procedure. Even so, the set of intervals yields useful information.

References

Cochran, W. G., 1954. Some methods for strengthening the common χ^2 tests, *Biometrics, 10,* 417.

Cochran, W. G., 1963. *Sampling Techniques*, 2nd edition, John Wiley and Sons, Inc., New York.

Pearson, K., 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling, *Philos. Mag., Series 5, 50,* 157.