Simultaneous Confidence Intervals and Sample Size Determination for Multinomial Proportions

Cristina P. SISON and Joseph GLAZ*

Simultaneous confidence interval procedures for multinomial proportions are used in many areas of science. In this article two new simultaneous confidence interval procedures are introduced. Numerical results are presented to evaluate these procedures and compare their performance with established methods that have been used in statistical literature. From the results presented in this article, it is evident that the new procedures are more accurate than the established ones, where the accuracy of the procedure is measured by the volume of the confidence region corresponding to the nominal coverage probability and the probability of coverage it achieves. In the sample size determination problem, the new procedures provide a sizable amount of savings as compared to the procedures that have been used in many applications. Because both procedures performed equally well, the procedure that requires the least amount of computing time is recommended.

KEY WORDS: Coverage probabilities; Multinomial distribution; Probability approximations; Simultaneous inference.

1. INTRODUCTION

Let $X_1, \ldots, X_k$ be the cell frequencies in a sample of n observations from a multinomial distribution with cell probabilities $p_1, \ldots, p_k$, where $p_i \geq 0$ and $\sum_{i=1}^{k} p_i = 1$. We are interested in deriving simultaneous confidence intervals for $p_1, \ldots, p_k$. These simultaneous confidence intervals play an important role in many areas of statistical applications, including quality control (Goodman 1965; Quesenberry and Hurst 1964), opinion polling (Fitzpatrick and Scott 1987), anthropology (Cochran 1963; Tortora 1978), biology (Thompson 1987), roulette wheel analysis (Ethier 1982), and simulation studies (Angers 1984; Hurtubise 1969). A more thorough discussion about simultaneous confidence intervals, their applications, and additional references were presented by Miller (1981), Aickin (1983), and Hochberg and Tamhane (1988).

In Section 2 two new methods are presented. The first method is based on the approximation for multinomial probabilities using the algorithm of Levin (1981). The second method utilizes the negative dependence structure inherent in the multinomial distribution and related probability inequalities introduced by Glaz and Johnson (1984). The new methods produce approximate simultaneous confidence intervals that are more accurate than the established methods, where the accuracy of a simultaneous confidence region is determined from its volume corresponding to a given probability of coverage and the probability of coverage it achieves.

In Section 3 these simultaneous confidence intervals are used to determine the sample size needed to achieve a specified level of significance and a desired volume for the confidence region. In Section 4 numerical results are presented to evaluate the performance of the simultaneous confidence intervals proposed in Section 2 and the sample size determination algorithm in Section 3.

2. SIMULTANEOUS CONFIDENCE INTERVALS FOR $p_1, \ldots, p_k$

We now present two new methods for deriving simultaneous confidence intervals for $p_1, \ldots, p_k$. The first method is based on approximation (1).

Theorem 2.1. Let $X_1, \ldots, X_k$ be the cell frequencies in a sample of n observations from a multinomial distribution with cell probabilities $p_1, \ldots, p_k$, where $p_i \geq 0$ and $\sum_{i=1}^{k} p_i = 1$. For $i = 1, \ldots, k$, let V_i be independent Poisson random variables with mean np_i and let Y_i be its truncation to $[b_i, a_i]$. Denote the central and factorial moments of Y_i by $\mu_i = EY_i$ and $\sigma_i^2 = \operatorname{var} Y_i$, $\mu_{n,i}$, and $\mu_{(r)}$. Set

$$\gamma_1 = \frac{1}{\sqrt{k}} \frac{(1/k \sum_{i=1}^{k} \mu_{3,i})}{(1/k \sum_{i=1}^{k} \sigma_i^2)^{3/2}}$$

and

$$\gamma_2 = \frac{1}{\sqrt{k}} \frac{(1/k \sum_{i=1}^{k} \mu_{4,i} - 3\sigma_i^4)}{(1/k \sum_{i=1}^{k} \sigma_i^2)^2}.$$

Then

$$P(b_i \leq X_i \leq a_i;\, i = 1, \ldots, k) \approx \frac{n!}{n^n e^{-n}} \left\{ \prod_{i=1}^{k} P(b_i \leq V_i \leq a_i) \right\} \times f_e\left(\frac{n - \sum_{i=1}^{k} \mu_i}{\sqrt{\sum_{i=1}^{k} \sigma_i^2}} \right) \frac{1}{\sqrt{\sum_{i=1}^{k} \sigma_i^2}}, \qquad (1)$$

where

$$f_e(x) = \left(\frac{1}{\sqrt{2\pi}} e^{-x^2/2} \right) \times \left\{ 1 + \frac{\gamma_1}{6}(x^3 - 3x) + \frac{\gamma_2}{24}(x^4 - 6x^2 + 3) + \frac{\gamma_1^2}{72}(x^6 - 15x^4 + 45x^2 - 15) \right\}. \qquad (2)$$

---

* Cristina P. Sison is Ph.D. Candidate and Joseph Glaz is Professor, Department of Statistics, University of Connecticut, Storrs, CT 06269. Research was partially supported by Office of Naval Research Contract No. N00014-94-1-0061 and the University of Connecticut Research Foundation. The authors thank Alan Gelfand, Evarist Gine, the editor, the associate editor, and the referee for helpful comments and suggestions that have improved the presentation of the results in this article.

© 1995 American Statistical Association
Journal of the American Statistical Association
March 1995, Vol. 90, No. 429, Theory and Methods

Table 1. Comparison of Five Simultaneous Confidence Interval Procedures, $1 - \alpha = .95$ for $n = 467$, $X_1 = 56$, $X_2 = 72$, $X_3 = 73$, $X_4 = 59$, $X_5 = 62$, $X_6 = 87$, $X_7 = 58$

i	*Quesenberry–Hurst*	*Goodman*	*Fitzpatrick–Scott*	*(1)*	*(10)*
1	.076 .183	.085 .166	.068 .172	.079 .164	.079 .164
2	.104 .222	.115 .204	.102 .206	.113 .199	.113 .199
3	.106 .225	.116 .207	.104 .209	.116 .201	.116 .201
4	.081 .191	.091 .173	.074 .179	.086 .171	.086 .171
5	.087 .198	.096 .181	.080 .185	.092 .177	.092 .177
6	.131 .258	.143 .239	.134 .239	.146 .231	.146 .231
7	.080 .188	.089 .171	.072 .176	.084 .169	.084 .169
Volume	$.255 \times 10^{-6}$	$.367 \times 10^{-7}$	$.137 \times 10^{-6}$	$.324 \times 10^{-7}$	$.327 \times 10^{-7}$
Coverage	.994	.935	.989	.939	.939

Proof. It follows from the approach of Levin (1981) that

$$P(b_i \le X_i \le a_i; i = 1, \ldots, k)$$
$$\approx \frac{n!}{n^n e^{-n}} \left\{ \prod_{i=2}^{k} P(b_i \le V_i \le a_i) \right\} \{ P(W = n) \}, \quad (3)$$

where for $i = 1, \ldots, k$, V_i are independent observations from a Poisson distribution with mean np_i and $W = \sum_{i=1}^{k} Y_i$, where Y_i are independent observations from a truncated Poisson distribution with range $[b_i, a_i]$, $i = 1, \ldots, k$.

Elaborate but routine calculations show that for $r \ge 1$, the rth factorial moment of the truncated Poisson random variable, Y, with mean λ and range $[b, a]$ is given by

$$\mu_{(r)} = \lambda^r \left(1 + \left[\frac{\sum_{v=b-r}^{b-1} e^{-\lambda}\lambda^v/v! - \sum_{v=a-r+1}^{a} e^{-\lambda}\lambda^v/v!}{\sum_b^a e^{-\lambda}\lambda^v/v!} \right] \right), \quad (4)$$

from which the central moments are easily obtained.

To complete the proof, one approximates $P(W = n)$ using the Edgeworth expansion

$$f_e\left(\frac{n - \sum_{i=1}^{k} \mu_i}{\sqrt{\sum_{i=1}^{k} \sigma_i^2}} \right) \frac{1}{\sqrt{\sum_{i=1}^{k} \sigma_i^2}}, \quad (5)$$

where $f_e(x)$ is given in Equation (2).

For a given value of $1 - \alpha$, consider the region $(\hat{p}_i - c/n, \hat{p}_i + c/n; i = 1, \ldots, k)$, where $\hat{p}_i = X_i/n$ and $c \ge 1$ is an integer determined from the equation

$$P(np_i - c \le X_i \le np_i + c; i = 1, \ldots, k)$$
$$= P\left(\hat{p}_i - \frac{c}{n} \le p_i \le \hat{p}_i + \frac{c}{n}; i = 1, \ldots, k \right)$$
$$= 1 - \alpha. \quad (6)$$

Because $p_1, \ldots, p_k$ are unknown, for given values of $X_1 = x_1, \ldots, X_k = x_k$, let $X_1^*, \ldots, X_k^*$ have a multinomial distribution with parameters n and $\hat{p}_1, \ldots, \hat{p}_k$. It is well known that as $n \to \infty$, the distribution of $(X_1 - np_1)/n^{1/2}, \ldots, (X_k - np_k)/n^{1/2}$ and $(X_1^* - n\hat{p}_1)/n^{1/2}, \ldots, (X_k^* - n\hat{p}_k)/n^{1/2}$ converge to the same multivariate normal distribution (Bishop, Fienberg, and Holland 1975, chap. 14). Therefore, we can approximate the value of c given in Equation (6) by solving

$$v(c) = P(x_i - c \le X_i^* \le x_i + c; i = 1, \ldots, k) = 1 - \alpha. \quad (7)$$

Using Theorem 2.1, we can find an integer c such that $v(c) < 1 - \alpha < v(c + 1)$. Let $\gamma = [(1 - \alpha) - v(c)]/[v(c + 1) - v(c)]$. Because the multinomial distribution is skewed, we recommend the following confidence region for $p_1, \ldots, p_k$:

$$\left(\hat{p}_i - \frac{c}{n} \le p_i \le \hat{p}_i + \frac{(c + 2\gamma)}{n}; i = 1, \ldots, k \right). \quad (8)$$

The performance of this region will be evaluated in Section 4.

The second method is based on the following inequality for the multinomial distribution:

$$P(X_i \le a_i; i = 1, \ldots, k) \le \alpha_{1,m} \prod_{j=m+1}^{k} \frac{\alpha_{j-m+1,j}}{\alpha_{j-m+1,j-1}}, \quad (9)$$

where $\alpha_{i,j} = P(X_i \le a_i, \ldots, X_j \le a_j)$ for $1 \le i < j \le k$ (Glaz and Johnson 1984, thm. 2.8). Because the inequality (9) is tight (Glaz 1990, sec. 4.3), to derive simultaneous confidence intervals for $p_1, \ldots, p_k$, the following approximation will be used:

$$P(X_j \in I_j; j = 1, \ldots, k) \approx \alpha_{1,m}^* \prod_{j=m+1}^{k} \frac{\alpha_{j-m+1,j}^*}{\alpha_{j-m+1,j-1}^*}, \quad (10)$$

where $\alpha_{i,j}^* = P(b_i \le X_i \le a_i, \ldots, b_j \le X_j \le a_j)$ for $1 \le i < j \le k$. Confidence regions given in Equation (8) where the constants c and γ are evaluated via Equation (7) and

Table 2. Comparison of Five Simultaneous Confidence Interval Procedures, $1 - \alpha = .95$ for $n = 250$, $X_i = 5$, $i = 1, \ldots, 50$

i	*Quesenberry–Hurst*	*Goodman*	*Fitzpatrick–Scott*	*(1)*	*(10)*
1, . . . , 50	.001 .240	.005 .075	.000 .091	.000 .053	.000 .053
Volume	$.778 \times 10^{-31}$	$.134 \times 10^{-57}$	$.116 \times 10^{-51}$	$.184 \times 10^{-63}$	$.175 \times 10^{-63}$
Coverage	.724	.113	1.00	.970	.970

Table 3. Comparison of Five Simultaneous Confidence Interval Procedures, $1 - \alpha = .95$ for $n = 250$, $X_1 = 1, \ldots, X_{10} = 1$, $X_{11} = 12, \ldots, X_{20} = 12$, $X_{21} = 5, \ldots, X_{30} = 5$, $X_{31} = 3, \ldots, X_{40} = 3$, $X_{41} = 4, \ldots, X_{50} = 4$

i	*Quesenberry–Hurst*	*Goodman*	*Fitzpatrick–Scott*	*(1)*	*(10)*
1, . . . , 10	.000 .216	.000 .049	.000 .075	.000 .041	.000 .041
11, . . . , 20	.007 .279	.019 .114	.000 .119	.012 .085	.012 .085
21, . . . , 30	.001 .240	.005 .075	.000 .091	.000 .057	.000 .057
31, . . . , 40	.000 .228	.002 .062	.000 .083	.000 .049	.000 .049
41, . . . , 50	.001 .234	.004 .069	.000 .087	.000 .053	.000 .053
Volume	$.533 \times 10^{-31}$	$.949 \times 10^{-59}$	$.626 \times 10^{-52}$	$.401 \times 10^{-63}$	$.379 \times 10^{-63}$
Coverage	.004	.002	1.00	.942	.942

the approximations in (1) and (10) with $m = 3$ will be evaluated in Section 4.

3. SAMPLE SIZE DETERMINATION IN ESTIMATING $p_1, \ldots, p_k$

Assume that estimates $\hat{p}_1, \ldots, \hat{p}_k$ are available from a recent study. We are interested in determining the minimal sample size necessary to achieve a specified coverage probability for a given volume of the confidence region. This problem has been studied by many researchers, including Hurtubise (1969), Angers (1974, 1979, 1984, 1989), Tortora (1978), and Thompson (1987). Most of these procedures are based on the approach of Goodman (1965) or the multivariate normal approximation for the multinomial distribution and the Bonferroni inequality discussed by Angers (1974 and 1989) and Fitzpatrick and Scott (1987).

The new procedures introduced in Section 2 can be easily utilized in an algorithm for sample size determination in multinomial experiments. Let $V = (2w)^k$ be the given volume of the confidence region (8) with $\gamma = 0$. At each iteration of the algorithm, we evaluate

$$\eta(n) = P([n\hat{p}_i - nw + .5] \le X_i \le [n\hat{p}_i + nw]; \quad i = 1, \ldots, k),$$

where $[x]$ is the integer part of x. Let n_0 be a starting value for our algorithm, $\eta(n_0) < 1 - \alpha$. Then the required sample size, n^*, is given by

$$n^* = \min\{n \ge n_0; \eta(n) \ge 1 - \alpha\}.$$

Later, in Table 4, we evaluate the performance of this algorithm for sample size determination.

4. NUMERICAL EXAMPLES

The first example lists personal crimes committed in the city of New Orleans on each of the seven days in a randomly selected week in 1984 (Gelfand, Glaz, Kuo, and Lee 1992). In Table 1 we compare five simultaneous confidence intervals by presenting the intervals for p_i, $1 \le i \le 7$. The simultaneous confidence intervals are targeted to achieve a coverage probability of .95. The achieved coverage probabilities were evaluated using a simulation with 10,000 trials. For each of the five confidence regions, we present the volume of the rectangular region. From Table 1 we see that the new simultaneous confidence intervals are the most accurate ones, because they have the smallest volume and their coverage probability is the closest to .95. The next best procedure is based on work of Goodman (1965). The intervals based on work of Quesenberry and Hurst (1964) and Fitzpatrick and Scott (1987) are conservative, and thus their rectangular regions have larger volumes.

In Tables 2 and 3, two artificial examples with $n = 250$ and $k = 50$ are studied. We present for $1 \le i \le 50$, the confidence intervals for p_i, the volumes of the rectangular regions and the coverage probabilities. The coverage probabilities were evaluated from a simulation with 10,000 trials. It is evident that the proposed new procedures are the most accurate ones, because they have the smallest volume and achieve a coverage closest to .95. The simultaneous confidence intervals based on work of Fitzpatrick and Scott (1987) are conservative, and thus the rectangular region has a larger volume. The intervals based on work of Quesenberry and Hurst (1964) and Goodman (1965) have poor coverage probabilities.

We now evaluate the performance of four procedures for the sample size determination problem. We have chosen the data set discussed in Table 1 and reduced the volumes to the values specified in Table 4. For each procedure we evaluated the sample size needed to achieve a specified coverage probability and a specified volume for the rectangular region. In Table 4 the upper value is the sample size needed to meet these specifications, and the lower value is the simulated coverage based on these sample sizes. It follows from the numerical results in Table 4 that the sample size determination algorithms based on the new procedures are the most accurate ones, because their coverage probability is the closest

Table 4. Comparison of Four Methods in Sample Size Determination, $1 - \alpha = .95$

k	*Specified $1 - \alpha$*	*Volume*	*Goodman*	*Fitzpatrick–Scott*	*(1)*	*(10)*
7	.95	$.672 \times 10^{-8}$	946	1104	750	756
			.982	.991	.941	.947
7	.95	$.250 \times 10^{-9}$	2426	2827	1944	
			.982	.991	.951	

to .95 and the sample size is the smallest. The procedure based on approximation (10) seems to be impractical from the computational viewpoint, and thus only one case was evaluated.

5. SUMMARY AND CONCLUDING REMARKS

Simultaneous confidence intervals for multinomial proportions have been studied. These simultaneous intervals are based on two accurate approximations for the rectangular multinomial probabilities. Both approximations lead to rectangular regions that have a smaller volume and at the same time achieve coverage close to the targeted value. These new simultaneous confidence intervals have been compared to simultaneous intervals that have been used in the statistical literature (Fitzpatrick and Scott 1987; Goodman 1965; Quesenberry and Hurst 1964). The main advantage of the new simultaneous confidence interval procedures presented in this article is that their coverage probability is close to the targeted value. The Quesenberry and Hurst (1964) and Goodman (1965) simultaneous confidence intervals are inconsistent as far as achieving the targeted coverage probability. The Fitzpatrick and Scott (1987) adjusted binomial confidence intervals have a coverage probability that often is too high, leading to a rectangular region with a higher volume.

Achieving the specified coverage probability for simultaneous confidence intervals is of great importance in the sample size determination problem. The new procedures presented in this article lead to a sizable reduction of the sample size needed to achieve a specified volume for the rectangular region and a specified coverage probability. For the examples considered in Section 4, this sample size reduction amounts to about 20%.

In this article we have presented numerical results for coverage probability of .95. Similar results are obtained for other values of coverage probabilities. Moreover, approximations (1) and (10) can be utilized in deriving confidence intervals for $p_{\max} = \max(p_1, \ldots, p_k)$ and $p_{\min} = \min(p_1, \ldots, p_k)$. (For details, see Sison and Glaz 1993.) Because the procedure based on approximation (10) is computationally more time-consuming for large values of n and moderate values of k than the one based on approximation (1) and both procedures performed equally well, we recommend using the procedure based on approximation (1).

[Received February 1993. Revised May 1994.]

REFERENCES

Aickin, M. (1983), *Linear Statistical Analysis of Discrete Data*, New York: John Wiley.

Angers, C. (1974), "A Graphical Method to Evaluate Sample Sizes for the Multinomial Distribution," *Technometrics*, 16, 469–471.

——— (1979), "Sample Size Estimation for Multinomial Populations," *The American Statistician*, 33, 163–164.

——— (1984), "Large Sample Sizes for the Estimation of Multinomial Frequencies From Simulation Studies," *Simulation*, October, 175–178.

——— (1989), "Note on Quick Simultaneous Confidence Intervals for Multinomial Proportions," *The American Statistician*, 43, 91.

Bishop, Y., Fienberg, S., and Holland, P. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: MIT Press.

Cochran, W. G. (1963), *Sampling Techniques* (2nd ed.), New York: John Wiley.

Ethier, S. N. (1982), "Testing for Favorable Numbers on a Roulette Wheel," *Journal of the American Statistical Association*, 77, 660–665.

Fitzpatrick, S., and Scott, A. (1987), "Quick Simultaneous Confidence Intervals for Multinomial Proportions," *Journal of the American Statistical Association*, 82, 875–878.

Gelfand, A. E., Glaz, J., Kuo, L., and Lee, T. M. (1992), "Inference for the Maximum Cell Probability Under Multinomial Sampling," *Naval Research Logistics*, 39, 97–114.

Glaz, J. (1990), "A Comparison of Bonferroni-Type and Product-Type Inequalities in the Presence of Dependence," in *Topics in Statistical Dependence*, eds. H. W. Block, A. R. Sampson, and T. H. Savits, Hayward, CA: Institute of Mathematical Statistics, pp. 223–235.

Glaz, J., and Johnson, B. (1984), "Probability for Multivariate Distributions With Dependence Structures," *Journal of the American Statistical Association*, 79, 436–411.

Goodman, L. A. (1965), "On Simultaneous Confidence Intervals for Multinomial Proportions," *Technometrics*, 7, 247–254.

Hochberg, Y., and Tamhane, A. C. (1988), *Multiple Comparison Procedures*, New York: John Wiley.

Hurtubise, R. (1969), "Sample Sizes and Confidence Intervals Associated With a Monte Carlo Simulation Model Possessing a Multinomial Output," *Simulation*, 12, 71–77.

Levin, B. (1981), "A Representation for Multinomial Cumulative Distribution Functions," *The Annals of Statistics*, 9, 1123–1126.

Miller, R. G. (1981), *Simultaneous Statistical Inference* (2nd ed.), New York: Springer-Verlag.

Quesenberry, C. P., and Hurst, D. C. (1964), "Large-Sample Simultaneous Confidence Intervals for Multinomial Proportions," *Technometrics*, 6, 191–195.

Sison, C. P., and Glaz, J. (1993), "Simultaneous Confidence Intervals and Sample Size Determination for Multinomial Proportions," Technical Report No. 93-05, University of Connecticut, Dept. of Statistics.

Thompson, S. K. (1987), "Sample Size for Estimating Multinomial Proportions," *The American Statistician*, 41, 42–46.

Tortora, R. D. (1978), "A Note on Sample Size Estimation for Multinomial Populations," *The American Statistician*, 32, 100–102.