Sample Size for Estimating Multinomial Proportions

STEVEN K. THOMPSON*

~~J. J. HASBROUCK~~

[Th]is article presents a procedure and a table for selecting sample size for simultaneously estimating the parameters of a multinomial distribution. The results are obtained by examining the "worst" possible value of a multinomial parameter, vector, analogous to the case in which a binomial parameter equals one-half.

KEY WORDS: Multinomial distribution; Simultaneous inference.

1. INTRODUCTION

Sample size problems rarely have satisfyingly simple answers, because the anticipated precision of estimates may depend on one or more unknown parameters. An exception in which a definite answer can be given in the absence of prior information is the binomial distribution, for which we can compute the sample size sufficient to achieve a specified degree of precision for the "worst case" parameter value of one-half. For the multinomial distribution, with more than two categories, it is not realistic to consider a "worst case" in which all parameters equal one-half, since the sum of the parameters must equal one. In this article I answer the question, what is the "worst" possible value of a multinomial parameter vector, and give a table of sample sizes [for] specified precision levels.

The problem of choosing a sample size for the simultaneous estimation of multinomal proportions is closely akin to the problem of simultaneous confidence intervals for multinomial proportions, with the difference that in the sample size problem the acceptable widths of the intervals are specified in advance and sample size is chosen to control the probability that the intervals will cover the true proportions.

Queensbury and Hurst (1964) presented a method for constructing simultaneous confidence intervals for multinomial proportions based on the approximate chi-squared distribution of the sum of the observed minus expected frequencies squared divided by the expected frequencies. Goodman (1965) improved on this result by constructing less conservative (shorter) intervals satisfying the stated level of significance. Goodman's method was based on the normal approximation for a binomial proportion and used Bonferroni's inequality to put a bound on the probability that all of the intervals would be simultaneously correct. Goodman did not address the problem of sample size selection, but his approach has been the basis for subsequent work on the subject.

Angers (1974) applied Goodman's method to the problem of sample size for multinomial proportions and presented a graphical method for selecting a sample size based on prior knowledge of the parameter values. Tortora (1978) based sample size on the "worst case" individual parameter, which is the parameter closest to .5 when the precision criteria are the same for all parameters. Angers (1979) pointed out that Tortora's method was more conservative than necessary in some cases. For selecting sample size with limited prior information about the parameters in the form of inequalities, Angers proposed using the value closest to .5 for each parameter. Angers (1979, 1984) described the general procedure for selecting sample size using prior estimates of parameter values. The method is computationally tedious to carry out but results in substantial reductions in sample size over previous methods. For equal interval widths, Angers (1984) gave an empirical result on the "worst case" parameter vectors for small α levels, based on Monte Carlo selection of large numbers of parameter vectors.

In this article I establish the form of the "worst case" multinomial parameter vector when equal interval widths are specified for each parameter component. A formula for sample size under this worst case is given, and a table is provided that makes sample size easy to determine for selected significance levels and any interval width. Angers's empirical result is proved and extended to all α levels. The form of the worst case parameter vector is also established for the case in which the possible parameter values are constrained by prior knowledge to satisfy inequalities. It is noted that when different acceptable confidence interval widths are specified for different parameters, determination of sample size remains a computational problem.

Examples from the literature are reworked, illustrating the possible reductions in sample size or simplification of the sample size determination procedure using the methods of this article.

2. METHOD

The objective is to select the smallest sample size n for a random sample from a multinomial population such that the probability will be at least $1 - \alpha$ that all of the estimated proportions will simultaneously be within specified distances of the true population proportions, that is,

$$\Pr\{\bigcap_{i=1}^{k} |p_i - \pi_i| \le d_i\} \ge 1 - \alpha,$$

relates to $P(p-d<\hat{p}<p+d)\ge 0.95$ for 2 categories)

where π_i is the proportion in the ith category in the population, p_i is the observed proportion, and k is the number of categories.

In this article, it is assumed that the population is large enough for finite population correction factors to be ignored if sampling is done without replacement and that sample sizes are large enough for the normal approximation to be used.

For an individual parameter π_i, the probability α_i that the estimate p_i lies outside the specified interval is, by the nor-

*Steven K. Thompson is Assistant Professor, Department of Mathematical Sciences, University of Alaska, Fairbanks, AK 99775-1110. The work for this article was initiated while the author was Biometrician, Alaska Department of Fish and Game, Kodiak, AK. The author would like to thank the referees for valuable suggestions.

"case" value of m used to compute (1): For $0 < \alpha < .0344$, the worst case is $m = 2$; for $.0344 \le \alpha < .3466$, $m = 3$; for $.3466 \le \alpha < .6311$, $m = 4$; for $.6311 \le \alpha < .8934$, $m = 5$; and for $.8934 \le \alpha < 1$, $m = 6$. Angers (1984) reported the empirical discovery of the first cutoff point, $\alpha = .0344$.

If it is assumed a priori that each parameter π_i is within a specified interval $a_i \le \pi_i \le b_i$, then a smaller sample size may be determined to be sufficient. In this case, with $d_i = d$ for all categories, the worst case is shown in Appendix A to be of the form having m of the π_i equal and the other $k - m$ parameters on the boundaries of their respective intervals (i.e., $\pi_j = a_j$ or b_j) for some integer m ($0 \le m \le k$).

If, instead of absolute precision d_i, a relative precision r_i is desired for each parameter estimate, we require

$$\Pr\{\cap |p_i - \pi_i| \le r_i \pi_i\} \ge 1 - \alpha.$$

For this criterion, no sample size will be sufficient for all possible parameter values, since the necessary size increases without bound as any of the parameters approach zero. If every parameter is constrained a priori to be no closer than a given distance from the boundary, a sufficient sample size can be determined by considering the sum of the error probabilities (substituting $r_i \pi_i$ for d_i in the computations) for all possible parameter values in the constrained parameter space. Simplifying results helpful in the actual computation of sample size are not available for this case, however.

In many applications, the complete sample must be collected before measurements can be taken and proportions estimated. Without prior information about the parameters it is necessary to choose a sample size adequate for any possible parameter value, as described in this article. In other applications it may be possible to estimate sequentially population proportions and base sample size on a multistage procedure (see, e.g., Angers 1984; Cochran 1977, p. 79). In such situations, a further reduction in sample size may be possible through the sequential procedure.

3. COMPARISON WITH OTHER METHODS

Tortora (1978) proposed the following method for selecting sample size n for simultaneously estimating k proportions of a multinomial population: Compute $n_i = z^2 \pi_i(1 - \pi_i)/d_i^2$, where z is the upper $(\alpha/2k) \times 100$th percentile of the standard normal distribution for each of the k categories in the population, and choose the largest n_i for sample size. If there is no prior knowledge about the value of the π_i, then calculate n_i based on the "worst case" value $\pi_i = .5$ as $n_i = z^2(.5^2)/d_i^2$. The use of the $(\alpha/2k) \times 100$th percentile restricts each α_i to be less than or equal to α/k, which assures that $\Sigma\alpha_i$ will be no greater than α, but may result in a sample size larger than necessary if this restriction is not required.

Tortora applied this method to an example in which $\alpha = .05$, $d = .05$, and prior knowledge indicates that $\pi_1 = .27$, $\pi_2 = .43$, $\pi_3 = .19$, and $\pi_4 = .11$. Computing n for the category that gives the largest n_i (category 2) gives $n = 613$, whereas the individually worst possible case, using $\pi_i = .5$, gives 624. Since each parameter value is exactly specified by prior estimates, we can obtain a lower sample size by using Angers's method, which removes the constraint that each α_i be less than or equal to α/k. The (somewhat tedious) procedure, which involves computing

$$\alpha_i = 2(1 - \Phi(d_i \sqrt{n}/\sqrt{\pi_i(1 - \pi_i)}))$$

for selected values of n, yields $n = 469$ as the smallest value of n for which $\Sigma\alpha_i \le .05$. Note that if we do not wish to assume any prior knowledge of the parameters, we could use Table 1 to determine very easily that a sample size of 510 would be sufficient to meet the criteria.

When the parameters are not completely specified but the prior information is given in the form of inequalities, Angers (1979) proposed using for π_i the value closest to .5 in the interval to which π_i is restricted, for $i = 1, \ldots, k$. Angers applied this method to an example with $\pi_1 \ge .6$, $\pi_2 \le .3$, $\pi_3 \le .3$, and $d_1 = .05$, $d_2 = d_3 = .025$, and $\alpha = .05$. Using $\pi_i = .6$, $\pi_2 = .3$, and $\pi_3 = .3$, he thus obtained $n = 1{,}689$ as sufficient to meet the criteria. Note, however, that the parameter values used sum to more than one and hence are not among the possible cases we need to consider. Since different interval widths are specified for different parameters, we cannot simplify the computational procedure for determining sample size, but we can find a smaller sample size that meets the criteria. A numerical search of the restricted parameter space reveals that the worst possible value of the parameter vector is $\boldsymbol{\pi}_0 = (.6, .3, .1)$ or, equivalently, $\boldsymbol{\pi}_0 = (.6, .1, .3)$. We find that a sample size of $n = 1{,}322$ is sufficient to satisfy the criteria. The actual procedure for obtaining this value involves computing sufficient sample size for a large number of parameter values, selected either randomly or systematically from the restricted parameter space.

If in the last example equal interval widths had been specified, for example, $d = .05$ for all three parameters, the solution would be greatly simplified using Theorem 2 of Appendix A. Theorem 2 states that the "worst" case has the form that all parameters not on the boundary—that is, not equal to 0, .6, or .3—must be equal. The only candidates within the restricted parameter space are (.6, .2, .2), (.6, .3, .1), and (.7, .3, 0). By computing the sufficient sample size for just these three values, we find that (.6, .3, .1) is still the "worst" parameter vector and that a sample size of 455 would be sufficient.

APPENDIX A: THE WORST POSSIBLE PARAMETER VECTOR OF A MULTINOMIAL DISTRIBUTION

Theorem 1. Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$ $(0 \le \pi_i \le 1, i = 1, \ldots, k)$. Let $\alpha_i = 2(1 - \Phi(z_i))$, where Φ is the normal cdf, and

$$z_i = d\sqrt{n}/\sqrt{\pi_i(1 - \pi_i)}, \quad d > 0.$$

The value of $\boldsymbol{\pi}$ that maximizes $\Sigma\alpha_i$, subject to $\Sigma\pi_i = 1$, is the form $\pi_i = 1/m$ for m of the parameters and $\pi_j = 0$ for the other $k - m$ parameters, where m is a nonnegative integer ($m \le k$).

Proof. The proof has four parts. In (1), I show that a value of $\boldsymbol{\pi}$ with all parameters equal satisfies the *necessary*

α_i is fixed. Denote by b the sum of these $k - m$ α_i. The problem of maximizing $\Sigma_{i=1}^{k} \alpha_i$ becomes that of maximizing $\Sigma_{i=1}^{m} \alpha_i$ for the m remaining parameters, subject to $\Sigma_{i=1}^{m} \pi_i = 1 - b$.

Set $G(\pi, \lambda) = \Sigma \alpha_i - \lambda(\Sigma_{i=1}^{m} \pi_i - b)$. Except for the straint $\Sigma \pi_i = b$, the derivative equations of G are exactly same as in the proof of Theorem 1, and the only interior extremum point is provided by $\boldsymbol{\pi}^m$, in which the π_i are equal, that is, $\pi_i = (1 - b)/m$ $(i = 1, \ldots, m)$. If $\boldsymbol{\pi}^m$ is not in the interior of A, there is no extremum point in the boundary region of A under consideration. If $\boldsymbol{\pi}^m$ is in the interior of A, an extremum is provided by the parameter vector with $\pi_i = (1 - b)/m$ for m parameters and the remaining π_i on the boundary of A.

APPENDIX B: A RESULT USEFUL IN SAMPLE SIZE COMPUTATION

The function $f(m) = [\Phi^{-1}(1 - \alpha/2m)]^2(1/m)(1 - 1/m)$ decreases for all $m > 2$ such that

$$[\Phi^{-1}(1 - \alpha/2m)]^2 \geq 2(1 - 1/m)/(1 - 2/m).$$

Proof. Let $g(u) = [\Phi^{-1}(1 - au)]^2 u(1 - u)$, where $u = 1/m$ and $a = \alpha/2$.

$$g'(u) = -2au(1 - u)\Phi^{-1} \times (1 - au)[\phi(\Phi^{-1}(1 - au))]^{-1} + [\Phi^{-1}(1 - au)]^2(1 - 2u).$$

An inequality involving Mills ratio (Patel and Read 1982, p. 64) gives $x\phi(x) \leq 1/[1 - \Phi(x)]$ for $x > 0$, so with $x = \Phi^{-1}(1 - au)$,

$$g'(u) \geq -2(1 - u) + [\Phi^{-1}(1 - au)]^2(1 - 2u).$$

Thus g is increasing and f is decreasing whenever the right side is positive, that is, for all $m > 2$ such that

$$[\Phi^{-1}(1 - \alpha/2m)]^2 \geq 2(1 - 1/m)/(1 - 2/m).$$

[*Received July 1984. Revised December 1985.*]

REFERENCES

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

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

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

Cochran, William G. (1977), *Sampling Techniques* (3rd ed.), New York: John Wiley.

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

Hancock, Harris (1960), *Theory of Maxima and Minima*, New York: Dover Publications.

Miller, Rupert G., Jr. (1980), *Simultaneous Statistical Inference* (2nd ed.), New York: Springer-Verlag.

Patel, J. K., and Read, C. B. (1982), *Handbook of the Normal Distribution*, New York: Marcel Dekker.

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

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

Optimal Allocation in Multivariate, Two-Stage Sampling Designs

JAMES R. WATERS and ALEXANDER J. CHESTER*

Procedures for determining optimal two-stage sampling designs are well documented when a single variable is measured. Many surveys, however, measure several variables simultaneously, in which case the overall optimal sampling plan is traditionally chosen from among the optima for each variable considered individually. We use a graphical approach to illustrate that the traditional method is not always appropriate and that intersections of variance constraints should be considered as candidates for the overall optimum.

KEY WORDS: Subsampling; Optimal survey design; Sample size; Graphical solution.

1. INTRODUCTION

Analyses of two-stage designs are well documented when a single variable is measured, and methods to calculate the optimal allocation of sampling effort among sampling stages are readily available (Cochran 1977, chap. 10; Snedecor and Cochran 1980, chap. 21; Sokal and Rohlf 1981, chap. 10). Many surveys, however, measure several variables simultaneously, and procedures for determining optimal sampling effort are not well defined. The traditional approach has been to estimate optimal sample size for each variable individually and then choose the final sampling design from among the individual solutions. The shortcoming of this approach is that it does not necessarily identify all possible candidates for the overall solution.

Our purpose is to determine the optimal allocation of sampling effort in a two-stage design when several parameters are estimated simultaneously. We begin with a brief discussion of optimal survey design when estimating a single

*James R. Waters and Alexander J. Chester are with the National Marine Fisheries Service, Southeast Fisheries Center, Beaufort Laboratory, Beaufort, NC 28516. The authors thank David Colby, Douglas Vaughan, and two anonymous reviewers for their helpful comments on the manuscript.