ELSEVIER

Journal of Statistical Planning and Inference 82 (1999) 251–262

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Simultaneous confidence intervals for multinomial proportions

Joseph Glaz[a,*], Cristina P. Sison[b]

[a] *Department of Statistics, University of Connecticut, 196 Auditorium Road, U-120 MSB 428, Storrs, CT 06269-3120, USA*

[b] *Division of Biostatistics, North Shore University Hospital, Manhasset, NY 11030, USA*

Received 1 January 1997; accepted 1 February 1998

Abstract

In this article approximate parametric bootstrap confidence intervals for functions of multinomial proportions are discussed. The interesting feature of these confidence intervals is that they are obtained via an Edgeworth expansion approximation for the rectangular multinomial probabilities rather than the resampling approach. In the first part of the article simultaneous confidence intervals for multinomial proportions are considered. The parametric bootstrap confidence interval appears to be the most accurate procedure. The use of this parametric bootstrap confidence region in the sample size determination problem is also discussed. In the second part of the article approximate parametric bootstrap equal-tailed confidence intervals for the minimum and maximum multinomial cell probabilities are derived. Numerical results based on a simulation study are presented to evaluate the performance of these confidence intervals. We also indicate several problems for possible future research in this area. © 1999 Elsevier Science B.V. All rights reserved.

MSC: primary 62F25; secondary 62H12

Keywords: Approximations; Confidence regions; Edgeworth expansion; Multinomial distribution; Order statistics; Parametric bootstrap; Simultaneous inference

1. Introduction

Let $\boldsymbol{X} = (X_1, \ldots, X_k)$ be the vector of cell frequencies in a sample of n observations from a multinomial distribution with cell probabilities $\theta_0 = (p_1, \ldots, p_k)$, where $p_i > 0$ and $\sum_{i=1}^{k} p_i = 1$. In this article we are interested in approximate parametric bootstrap confidence intervals for functions of θ_0 using the methods of Hall (1992). The interesting feature of these bootstrap confidence intervals is that they are implemented via an Edgeworth expansion approximation rather than resampling. The bootstrap methods

* Corresponding author. Fax: 860-486-4113.

E-mail address: glaz@uconnvm.uconn.edu (J. Glaz)

0378-3758/99/$ - see front matter © 1999 Elsevier Science B.V. All rights reserved.
PII: S0378-3758(99)00047-6

without resampling for different problems have been discussed in Efron (1987), DiCiccio and Efron (1992), Hall (1992), and Laird and Louis (1987), among others.

In Section 2 we briefly review the parametric bootstrap approach presented in Hall (1992, Chapter 1). We show that the confidence region derived in Sison and Glaz (1995a) is an approximate parametric bootstrap confidence region for the multinomial proportions. The implementation of this parametric bootstrap confidence region is based on the Edgeworth expansion approximation derived in Sison and Glaz (1995a). These simultaneous confidence intervals play an important role in many areas of statistical applications, to list just a few: anthropology (Cochran, 1963; Tortora, 1978), biology (Thompson, 1987), opinion polling (Fitzpatrick and Scott, 1987), quality control (Quesenberry and Hurst, 1964; Goodman, 1965), roulette wheel analysis (Ethier, 1982), and simulation studies (Hurtubise, 1969; Angers, 1984). The importance of simultaneous confidence intervals and their applications is also discussed in Hochberg and Tamhane (1988). Based on the numerical results in Sison and Glaz (1995a) we discuss the performance of this confidence region and its use in the sample size determination problem.

In Section 3 we derive approximate parametric bootstrap confidence intervals for the minimum and the maximum multinomial cell probability. The inference for the extreme multinomial probabilities has applications in the area of allocation of resources (Gelfand et al., 1992) and opinion polling where one is interested in estimating the preference for the most popular or the least popular candidate or product. In Section 3 we focus on two methods for deriving approximate parametric equal-tail confidence intervals for the minimum and the maximum multinomial cell probability. The implementation of these confidence intervals is based on the Edgeworth expansion approximation for the multinomial rectangular probabilities studied in Sison and Glaz (1995a). Numerical results are presented to evaluate the performance of these parametric bootstrap confidence intervals.

In Section 4 of this article we present concluding remarks and outline some future work in this area.

2. A parametric bootstrap confidence region for multinomial proportions

Many statistical inference procedures are based on the relationship between the observed sample and the population from which it is obtained. The following parametric bootstrap formulation is presented in Hall (1992, Chapter 1). Let F_0 be the distribution function of the population under study and let F_1 be the distribution function obtained from F_0 by estimating its parameters from the observed sample. In general, given a specified class of functionals $f_t, t \in T$, corresponding to the particular inference procedure under study, the value of t that solves the following *population equation* has to be determined:

$$E\{f_t(F_0, F_1) \mid F_0\} = 0. \tag{2.1}$$

In the above equation the conditioning on F_0 is used to emphasize that the expectation is evaluated with respect to F_0. Since F_0 is unknown we cannot solve the population equation. One hopes that the relationship between the distribution functions F_0 and F_1 can be captured in the relationship between the distribution functions F_1 and F_2, where the distribution function F_2 is obtained from F_1 by replacing its parameters by the updated sample estimates based on the observed sample from the distribution F_1. Therefore, instead of solving the above population equation one solves the following *sample equation*:

$$E\{f_t(F_1,F_2)\,|\,F_1\}=0. \tag{2.2}$$

The above equation is called the sample equation since conditional on F_1 all the parameters in it are specified. Hall (1992, Chapter 1) refers to this as the *bootstrap principle*. The distribution functions F_1 and F_2 are called the *parametric bootstrap* estimates of F_0 and F_1, respectively, since they are obtained by estimating the unknown parameters from the respective observed samples.

For the problem at hand, denote the multinomial distribution function of the population under study by F_0. The vector of parameters θ_0 can be estimated by its maximum likelihood estimate $\hat{\theta}=\hat{p}_1=x_1/n,\ldots,\hat{p}_k=x_k/n)$. Let F_1 be the distribution function obtained from F_0 by replacing the cell probabilities with their maximum likelihood estimates. Let $\boldsymbol{X}^*=(X_1^*,\ldots,X_k^*)$ be the random vector of cell frequencies in a sample of n observations from the distribution F_1. Denote by F_2 the distribution function obtained from F_1 by replacing $\hat{\theta}$ with $\hat{\theta}^*=(\hat{p}_1^*=x_1^*/n,\ldots,\hat{p}_k^*=x_k^*/n)$. For $1\leqslant i\leqslant k$, let $\theta_i(F_0)=p_i$, $\theta_i(F_1)=X_i/n$, and $\theta_i(F_2)=X_i^*/n$. We will assume that conditional on F_1, $\theta_i(F_1)=x_i/n$. To derive a $(1-\alpha)100\%$ parametric bootstrap confidence region for the multinomial proportions the following functional is considered.

$$f_t=I\{\theta_i(F_1)-t\leqslant\theta_i(F_0)\leqslant\theta_i(F_1)+t;1\leqslant i\leqslant k\,|\,F_0\}-(1-\alpha), \tag{2.3}$$

where $I\{A\}$ denotes the indicator function of the event A. It follows from the definition of the functionals f_t, $\theta_i(F_0)$ and $\theta_i(F_1)$ that the population equation (2.1) can be expressed by

$$P\left\{\frac{X_i}{n}-t\leqslant p_i\leqslant\frac{X_i}{n}+t;1\leqslant i\leqslant k\,\middle|\,F_0\right\}=1-\alpha,$$

which is equivalent to

$$P\{np_i-c\leqslant X_i\leqslant np_i+c;1\leqslant i\leqslant k\,|\,F_0\}=1-\alpha, \tag{2.4}$$

where $c=nt$ is an integer. The sample equation (2.2) for the functional f_t is given by

$$P\left\{\frac{X_i^*}{n}-t\leqslant\frac{x_i}{n}\leqslant\frac{X_t^*}{n}+t;1\leqslant i\leqslant k\,\middle|\,F_1\right\}=1-\alpha,$$

which is equivalent to

$$P\{x_i-c\leqslant X_i^*\leqslant x_i+c;1\leqslant i\leqslant k\,|\,F_1\}=1-\alpha, \tag{2.5}$$

where $c=nt$ is an integer. The solution to the sample equation (2.5) is called a parametric bootstrap estimate of the solution to the population equation (2.4) and the

resulting approximate confidence region is a parametric bootstrap confidence region for the multinomial proportions.

The following result (Bishop et al., 1975, Chapter 14) supports the use of the parametric bootstrap estimate of the solution to the population equation for constructing a confidence region for the multinomial proportions. Let $X_1,\ldots,X_k$ have a multinomial distributions with parameters n and $p_1,\ldots,p_k$. 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 $\hat{p}_1=x_1/n,\ldots,\hat{p}_k=x_k/n$. Then as $n\to\infty$ the distribution of

$$\frac{X_1-np_1}{\sqrt{n}},\ldots,\frac{X_k-np_k}{\sqrt{n}}$$

and

$$\frac{X_1^*-n\hat{p}_1}{\sqrt{n}},\ldots,\frac{X_k^*-n\hat{p}_k}{\sqrt{n}}$$

converge to the same multivariate normal distribution.

To obtain an approximate solution to the sample equation (2.5) Sison and Glaz (1995a) investigated two methods. The first method is based on the Edgeworth expansion approximation for the multinomial distribution. The second method is based on the product-type approximation that utilizes the inherent negative dependence structure of the multinomial distribution. Since both approximations produce equally accurate results and the approximation based on the Edgeworth expansion is significantly faster, in this article we will review only that approach. To get the integer value of c that approximately solves the sample equation (2.5) we have to evaluate rectangular probabilities for the random vector $\boldsymbol{X}^*=(X_1^*,\ldots,X_k^*)$ that has a multinomial distribution with parameters n and $\hat{p}_1,\ldots,\hat{p}_k$. For a moderate or large values of k it becomes computationally impractical to evaluate these probabilities. The discrete nature of the multinomial distribution also complicates the matter of obtaining an accurate approximation. The rectangular probabilities for $\boldsymbol{X}^*=(X_1^*,\ldots,X_k^*)$ in the sample equation (2.5) are approximated via the Edgeworth expansion for the rectangular multinomial probabilities (Sison and Glaz, 1995a, Theorem 2.1):

$$P\{b_i\leqslant X_i^*\leqslant a_i;\ 1\leqslant i\leqslant k\}$$

$$\approx \frac{n!}{n^n e^{-n}(\sum_{i=1}^k \sigma_i^2)^{1/2}}\prod_{i=1}^k P(b_i\leqslant V_i\leqslant a_i)\ f_e\left(\left(n-\sum_{i=1}^k\mu_i\right)\Big/\left(\sum_{i=1}^k\sigma_i^2\right)^{1/2}\right), \tag{2.6}$$

where V_i are independent Poisson random variables with mean x_i, $\mu_{r,i}$ and σ_i are the rth central moments and the variance, respectively, of W_i which are independent truncated Poisson random variables on the integers $b_i\leqslant\cdots\leqslant a_i$,

$$f_e(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}[1+\gamma_1(x^3-3x)/6+\gamma_2(x^4-6x^2+3)/24$$

$$+\gamma_1^2(x^6 - 15x^4 + 45x^2 - 15)/72],$$

$$\gamma_1 = \frac{\sum_{i=1}^{k} \mu_{3,i}}{(\sum_{i=1}^{k} \sigma_i^2)^{3/2}} \quad \text{and} \quad \gamma_2 = \frac{\sum_{i=1}^{k} (\mu_{4,i} - 3\sigma_i^4)}{(\sum_{i=1}^{k} \sigma_i^2)^2}.$$

Set

$$v(c) = P\{x_i - c \leqslant X_i^* \leqslant x_i + c;\ 1 \leqslant i \leqslant k\}.$$

Let c be the integer such that $v(c) < 1 - \alpha < v(c+1)$ and define $\rho = [(1-\alpha) - v(c)]/[v(c+1) - v(c)]$. Sison and Glaz (1995a, Section 2) recommend the following rectangular region:

$$\left(\hat{p}_i - \frac{c}{n} \leqslant p_i \leqslant \hat{p}_i + \frac{c+2\rho}{n};\ 1 \leqslant i \leqslant k \right), \tag{2.7}$$

as an approximate $(1-\alpha)100\%$ confidence region for $p_1, \ldots, p_k$. The adjustment using the quantity ρ was employed to compensate for the discrete and skewed nature of the multinomial distribution. Note that if $\hat{p}_i - c/n$ is less than 0 it is set to be equal to 0, and if $\hat{p}_i + (c+2\rho)/n$ is greater than 1 then it is set to be equal to 1. It follows from Eqs. (2.1)–(2.5) that this rectangular region is an approximate $(1-\alpha)100\%$ parametric bootstrap confidence region for $p_1, \ldots, p_k$.

The performance of the rectangular confidence region for the multinomial proportions given in Eq. (2.7) was investigated in Sison and Glaz (1995a) and extensive numerical studies were recorded in Sison and Glaz (1993) and Sison (1995). It was shown there that it outperformed the established rectangular confidence regions in the statistical literature studied in Quesenberry and Hurst (1964), Goodman (1965) and Fitzpatrick and Scott (1987). It had the lowest volume and at the same time its coverage probability was the closest to the targeted value. In particular its superior performance was evident for data with sparse cell counts. The advantage in using the confidence region in (2.7) with $\rho = 0$ was especially evident in the sample size determination problem.

Suppose we are given the required value for the volume $V = (2h)^k$, where $2h$ is the length of each of the intervals, and a specified confidence level $1-\alpha$. At each iteration of the algorithm we evaluate

$$\eta(n) = P([x_i - nh + 0.5] \leqslant X_i^* \leqslant [x_i + nh];\ 1 \leqslant i \leqslant k),$$

where $X_i^*, \ldots, X_k^*$ have a multinomial distribution F_1 and $[x]$ is the integer part of x. The parametric bootstrap estimate of the required sample size is given by

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

where n_0 is a starting value for our algorithm. In applications the starting value is determined from a previous experiment using the sample size that produced a confidence region with a volume that has to be reduced. It is evident from Sison and Glaz (1995a) and a more extensive numerical study in Sison (1995) that the parametric bootstrap algorithm for sample size determination is more accurate than the algorithms that have been used in the statistical literature. In some cases the savings in the sample size can amount to as much as 30% (Sison and Glaz, 1995a, Table 4).

3. Parametric bootstrap confidence intervals for $p_{\max}$ and $p_{\min}$

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 $p_i \geqslant 0$ and $\sum_{i=1}^{k} p_i = 1$. Denote this multinomial distribution by F_0. Let F_1 and F_2 be the multinomial distributions defined in Section 2. In this section two types of parametric bootstrap equal-tailed confidence intervals for $p_{\max}=\max\{p_1,\ldots,p_k\}$ and $p_{\min}=\min\{p_1,\ldots,p_k\}$ will be derived. Numerical results will be presented to evaluate the performance of these confidence interval procedures.

Following the parametric bootstrap formulation of Hall (1992, Chapter 1), consider the functional

$$f_t = I\{\theta(F_1) - t_{01} \leqslant \theta(F_0) \leqslant \theta(F_1) + t_{02} \mid F_0\} - (1-\alpha), \tag{3.1}$$

where $\theta(F_0)=p_{\max}$ or $p_{\min}$ and $\theta(F_1)=\hat{P}_{\max}=X_{(k)}/n$ or $\hat{P}_{\min}=X_{(1)}/n$, respectively, where $X_{(k)}=\max\{X_1,\ldots,X_k\}$ and $X_{(1)}=\min\{X_1,\ldots,X_k\}$. We will denote by $\hat{p}_{\max}=x_{(k)}/n$ and $\hat{p}_{\min}=x_{(1)}/n$ the maximum likelihood estimates of $p_{\max}$ and $p_{\min}$, respectively. Also, we will assume that conditional on F_1, $\theta(F_1)=\hat{p}_{\max}$ or $\hat{p}_{\min}$, respectively, and $\theta(F_2)=\hat{P}^*_{\max}=X^*_{(k)}/n$ or $\hat{P}^*_{\min}=X^*_{(1)}/n$, respectively, where $X^*_{(k)}=\max\{X^*_1,\ldots,X^*_k\}$ and $X^*_{(1)}=\min\{X^*_1,\ldots,X^*_k\}$ are the largest and smallest order statistics from a sample of size n from F_1. It follows from Eq. (3.1) that the population equation is given by

$$P\{\theta(F_1) - t^{(j)}_{01} \leqslant \theta(F_0) \leqslant \theta(F_1) + t^{(j)}_{02} \mid F_0\} = 1-\alpha,$$

which is equivalent to

$$P\{\theta(F_0) - t^{(j)}_{02} \leqslant \theta(F_1) \leqslant \theta(F_0) + t^{(j)}_{01} \mid F_0\} = 1-\alpha, \tag{3.2}$$

where $j=1,2$ refers to $p_{\max}$ and $p_{\min}$, respectively. The corresponding sample equation is given by

$$P\{\theta(F_2) - t^{(j)}_{01} \leqslant \theta(F_1) \leqslant \theta(F_2) + t^{(j)}_{02} \mid F_1\} = 1-\alpha,$$

which is equivalent to

$$P\{\theta(F_1) - t^{(j)}_{02} \leqslant \theta(F_2) \leqslant \theta(F_1) + t^{(j)}_{01} \mid F_1\} = 1-\alpha. \tag{3.3}$$

The parametric bootstrap estimates for the constants $t^{(j)}_{01}$ and $t^{(j)}_{02}$ in the population equation (3.2) are obtained by solving the following two equations:

$$P\{\theta(F_2) \leqslant \theta(F_1) + t^{(j)}_{01} \mid F_1\} = 1-\alpha/2 \quad \text{and} \quad P\{\theta(F_2) \leqslant \theta(F_1) - t^{(j)}_{02} \mid F_1\} = \alpha/2. \tag{3.4}$$

Therefore, an approximate $(1-\alpha)100\%$ type I equal-tailed parametric confidence interval for $\theta(F_0)$ is given by

$$(\theta(F_1) - \hat{t}^{(j)}_{01},\ \theta(F_1) + \hat{t}^{(j)}_{02}), \tag{3.5}$$

where $\hat{t}^{(j)}_{01}$ and $\hat{t}^{(j)}_{02}$ are the approximate solutions for $t^{(j)}_{01}$ and $t^{(j)}_{02}$ in Eqs. (3.4), respectively. If the lower confidence limit is negative it is set to be equal to 0, and if the upper confidence limit exceeds the value 1 it is set to be equal to 1.

We now proceed to derive an approximate $(1-\alpha)100\%$ type I equal-tailed parametric confidence interval for $\theta(F_0)=p_{\max}$. Let $c_1=nt_{01}^{(1)}$ and $c_2=nt_{02}^{(1)}$ be positive integers. It follows from Eqs. (3.4) that we have to solve for c_1 and c_2

$$P\{X_{(k)}^{*}\leqslant x_{(k)}+c_1\}=1-\alpha/2 \quad \text{and} \quad P\{X_{(k)}^{*}\leqslant x_{(k)}-c_2\}=\alpha/2, \tag{3.6}$$

respectively. For an integer $c>0$, let

$$\eta_1(c)=P\{0\leqslant X_1^{*}\leqslant x_{(k)}+c; 1\leqslant i\leqslant k\}. \tag{3.7}$$

Solving Eq. (3.6) is equivalent to solving

$$\eta_1(c_1)=1-\alpha/2 \quad \text{and} \quad \eta_1(-c_2-1)=\alpha/2,$$

respectively. The rectangular multinomial probabilities $\eta_1(c)$ in Eq. (3.7) are approximated via the Edgeworth expansion in Eq. (2.6) with $b_i=0$ for $1\leqslant i\leqslant k$. For this special case of rectangular multinomial probabilities the Edgeworth approximation has been studied in Levin (1981). Since $\eta_1(c)$ is an increasing function we choose c_1 and c_2 to be the smallest integers such that $\eta_1(c_1)\geqslant 1-\alpha/2$ and $\eta_1(-c_2-1)\leqslant\alpha/2$. This yields an approximate $(1-\alpha)100\%$ type I equal-tailed parametric confidence interval for $p_{\max}$ given by

$$\left(\hat{P}_{\max}-\frac{c_1}{n},\hat{P}_{\max}+\frac{c_2}{n}\right). \tag{3.8}$$

Note that $\hat{P}_{\max}$, the maximum likelihood estimator of $p_{\max}$, has a positive bias, since $E(\hat{P}_{\max})\geqslant p_{\max}$. For small values of n or when the number of cells is large this will have an effect on the coverage probability of the confidence interval procedures. The modified cell estimators $\hat{P}_{i,a}$ given by

$$\hat{P}_{i,a}=\frac{X_i+a/k}{n+a}=\frac{n}{n+a}\hat{P}_i+\frac{a}{n+a}\frac{1}{k} \tag{3.9}$$

are considered for replacing the cell estimators in $\hat{P}_{\max}$ and $\hat{P}_{\min}$ to compensate for bias and sparse cell counts. The adjustment value, a, is sort of ad hoc choice, the most popular ones in the statistical literature being $a=k/2$ and k (Bishop et al., 1975). For a recent article on these adjustments see Kunte and Upadhya (1996). Based on an extensive numerical study in Sison (1995) we recommend for the equal-tailed confidence interval in Eq. (3.8) to replace $\hat{P}_{\max}$ with the estimator $\hat{P}_{\max,k/2}=\max\{\hat{P}_{i,k/2}; 1\leqslant i\leqslant k\}$.

We now proceed to derive a parametric bootstrap type I equal-tailed confidence interval for $p_{\min}$ based on the functional given in Eq. (3.1). It follows from Eq. (3.4) that we have to solve for c_3 and c_4 in the following equations, respectively:

$$P\{X_{(1)}^{*}>x_{(1)}+c_3\}=\alpha/2 \quad \text{and} \quad P\{X_{(1)}^{*}\geqslant x_{(1)}-c_4\}=1-\alpha/2, \tag{3.10}$$

where $c_3=nt_{01}^{(2)}$ and $c_4=nt_{02}^{(2)}$. Let

$$\eta_2(c)=P\{x_{(1)}+c\leqslant X_1^{*}\leqslant n; 1\leqslant i\leqslant k\,|\,F_1\}. \tag{3.11}$$

Solving Eq. (3.10) is equivalent to solving

$$\eta_2(-c_4)=1-\alpha/2 \quad \text{and} \quad \eta_2(c_3+1)=\alpha/2.$$

Approximate solutions for c_3 and c_4 are obtained from the Edgeworth expansion approximation given in Eq. (2.6) with $a_i = n$ and $b_i = -c_4$ or $c_3 + 1$, respectively. Since $\eta_2(c)$ is a decreasing function of c, the approximate solutions are obtained by selecting the smallest integers c_3 and c_4 such that $\eta_2(c_3 + 1) \leqslant \alpha/2$ and $\eta_2(-c_4) \geqslant 1 - \alpha/2$. This yields an approximate $(1 - \alpha)100\%$ type I equal-tailed parametric confidence interval for $p_{\min}$ given by

$$\left(\hat{P}_{\min} - \frac{c_3}{n}, \hat{P}_{\min} + \frac{c_4}{n}\right). \tag{3.12}$$

Note that $\hat{P}_{\min}$, the maximum likelihood estimator of $p_{\min}$, has a negative bias, since $E(\hat{P}_{\min}) \leqslant p_{\min}$. For small values of n or when the number of cells is large this will have an effect on the coverage probability of the confidence interval procedures. Based on an extensive numerical study in Sison (1995) we recommend for the equal-tailed confidence intervals in Eq. (3.12) to replace $\hat{P}_{\min}$ with the modified estimator $\hat{P}_{\min,a} = \min\{\hat{P}_{i,a}; 1 \leqslant i \leqslant k\}$, where the modified cell estimators $\hat{P}_{i,a}$ is given in Eq. (3.9). For the problem at hand we recommend for most cases to use $a = k$ and in some cases adjusted values as high as $a = n$. We will say more about these adjustments when we present the numerical results.

We now proceed to discuss a different approach for deriving equal-tailed parametric bootstrap confidence intervals for $p_{\max}$ and $p_{\min}$. This approach for a different class of problems had been studied in Hall (1992, Chapters 1 and 3). The following motivation for the development of these intervals are given in Hall (1992, p. 12). Let F_0, F_1, and F_2 be the multinomial distributions defined in Section 2. Recall that F_1 and F_2 are the parametric bootstrap estimates of F_0 and F_1, respectively. Define the distribution function

$$\hat{H}(x) = P\{\theta(F_2) \leqslant x \mid F_1\}. \tag{3.13}$$

For a specified confidence level of $1 - \alpha$ the equal-tailed confidence interval

$$(\hat{H}^{-1}(\alpha/2), \hat{H}^{-1}(1 - \alpha/2)), \tag{3.14}$$

where $\hat{H}^{-1}(v) = \inf\{x; \hat{H}(x) \geqslant v\}$, will be referred to as a parametric bootstrap type II equal-tailed confidence interval for $\theta(F_0)$. It follows from Eqs. (3.1) – (3.10) that the parametric bootstrap type II equal-tailed confidence intervals for $p_{\max}$ and $p_{\min}$ are given by

$$\left(\hat{P}_{\max} - \frac{c_2}{n}, \hat{P}_{\max} + \frac{c_1}{n}\right) \tag{3.15}$$

and

$$\left(\hat{P}_{\min} - \frac{c_4}{n}, \hat{P}_{\min} + \frac{c_3}{n}\right), \tag{3.16}$$

respectively. The constants c_1, c_2 and c_3, c_4 are obtained following the steps of the algorithms that were described above for the type I equal-tailed parametric bootstrap confidence intervals for $p_{\max}$ and $p_{\min}$, respectively. For the parametric bootstrap type II equal-tailed confidence intervals the methods described above for adjusting the estimators to compensate for bias and possible sparse cell counts have been employed.

Table 1
Parametric bootstrap equal-tailed confidence intervals for $p_{\max}$, $1-\alpha=0.95$

Multinomial Probabilities	Average length	Average coverage	Multinomial probabilities	Average length	Average coverage
$\theta_0^{(1)}$	0.356	0.791	$\theta_0^{(5)}$	0.183	0.922
	0.370	0.981		0.245	1.000
$\theta_0^{(2)}$	0.391	0.929	$\theta_0^{(6)}$	0.254	0.897
	0.395	0.953		0.279	1.000
$\theta_0^{(3)}$	0.210	0.823	$\theta_0^{(7)}$	0.234	0.597
	0.296	0.985		0.272	1.000
$\theta_0^{(4)}$	0.216	0.962	$\theta_0^{(8)}$	0.160	0.971
	0.290	0.494		0.242	0.657

$\hat{P}_{\max,k/2}$ is used to estimate $p_{\max}$. The upper and lower values are for types I and II equal-tailed intervals, respectively.

For these confidence intervals too if the lower confidence limits are negative they are set to be equal to 0, and if the upper confidence limits exceed the value 1 they are set to be equal to 1. Before presenting the numerical results for the parametric bootstrap equal-tailed confidence intervals derived in this section we would like to note that Sison and Glaz (1995b) and Sison (1995) investigated the performance of symmetric confidence intervals and equal-tailed confidence intervals that are based on an approximate pivot for $p_{\max}$ and $p_{\min}$. Since these methods yielded confidence intervals that are inferior to the equal-tailed confidence intervals recommended in this article we will not present these results here.

We now present numerical results for evaluating the performance of the two types of equal-tailed confidence intervals for $p_{\max}$ and $p_{\min}$. Eight numerical examples were selected. Each example contains a hypothetical vector of multinomial probabilities along with a specified sample size, n, which is equal to 20. In examples 1–4, $k=5$ and in examples 5–8, $k=10$. For each of the eight multinomial distributions 10,000 random samples are generated using the IMSL subroutine RNMTN. For each of the generated samples we construct the confidence intervals discussed above, note their respective lengh and verify if they cover the hypothesized probability vector. In Tables 1 and 2 we report the average length and the average coverage probability. Let $\theta_0^{(i)}=\{p_1^{(i)},\ldots,p_k^{(i)}\}, 1\leqslant i\leqslant 8$, denote the multinomial probability vector for each of the examples. The eight multinomial probability vectors are:

$$\theta_0^{(1)}=\{0.50,0.15,0.15,0.10,0.10\},$$

$$\theta_0^{(2)}=\{0.70,0.075,0.075,0.075,0.075\},$$

$$\theta_0^{(3)}=\{0.30,0.175,0.175,0.175,0.175\},$$

$$\theta_0^{(4)}=\{0.24,0.24,0.24,0.24,0.04\},$$

$$\theta_0^{(5)}=\{0.20,0.10,0.10,0.10,0.10,0.10,0.10,0.10,0.05,0.05\},$$

Table 2
Parametric bootstrap equal-tailed confidence intervals for $p_{\min}$, $1-\alpha=0.95$

Multinomial probabilities	Average length	Average coverage	Multinomial probabilities	Average length	Average coverage
$\theta_0^{(1)}$	0.108	0.696	$\theta_0^{(5)}$	0.083	1.000
	0.130	0.978		0.079	1.000
$\theta_0^{(2)}$	0.094	1.000	$\theta_0^{(6)}$	0.074	1.000
	0.112	1.000		0.062	1.000
$\theta_0^{(3)}$	0.116	0.921	$\theta_0^{(7)}$	0.077	0.597
	0.146	1.000		0.064	1.000
$\theta_0^{(4)}$	0.110	0.848	$\theta_0^{(8)}$	0.095	1.000
	0.144	1.000		0.095	1.000

In examples 1–2, 4 and 5–7 $\hat{P}_{\min,k}$ is used to estimate $p_{\min}$, in examples 3 and 8 $\hat{P}_{\min,n}$ is used to estimate $p_{\min}$. The upper and lower values are for types I and II equal-tailed intervals, respectively.

$$\theta_0^{(6)}=\{0.25,0.25,0.10,0.10,0.05,0.05,0.05,0.05,0.05,0.05\},$$

$$\theta_0^{(7)}=\{0.28,0.08,0.08,0.08,0.08,0.08,0.08,0.08,0.08,0.08\},$$

$$\theta_0^{(8)}=\{0.145,0.095,0.095,0.095,0.095,0.095,0.095,0.095,0.095,0.095\}.$$

It follows from Table 1 that the parametric bootstrap type II equal-tailed confidence interval for $p_{\max}$ is more accurate as long as the cell probabilities and hence $p_{\max}$ are not too close to $1/k$. Examples 4 and 8 are the most difficult ones to handle since in these examples most of the cell probabilities and $p_{\max}$ are close to $1/k$. For this case the parametric bootstrap type I equal-tailed confidence interval performed well. The modified estimator $\hat{P}_{\max,k/2}$ enhanced the performance of these confidence intervals.

From Table 2 we observe that the parametric bootstrap type II equal-tailed confidence interval for $p_{\min}$ performs best. Here examples 3 and 8 are the most difficult ones since $p_{\min}$ is close to $1/k$. In situations where $p_{\min}$ is not close to $1/k$ the modified estimator $\hat{P}_{\min,k}$ enhances the performance of these confidence interval procedures. In examples 3 and 8 a more extreme adjustment is needed. Adding a constant of n/k to each cell enhanced the performance in these examples. The reason we have to add a larger constant to each of the cell estimators in these examples follows from the fact that the modified cell estimator in Eq. (3.9) $\hat{P}_{i,a}\to 1/k$ as $a\to\infty$.

4. Concluding remarks and future work

In this article approximate parametric bootstrap confidence regions have been discussed for several functions of multinomial proportions. The implementations of

these confidence regions are based on the Edgeworth expansion approximation for rectangular multinomial probabilities. From the numerical results one can conclude that this approach yielded quite accurate confidence regions. The confidence region for the multinomial cell probabilities discussed in Section 2 also yielded an efficient algorithm for the sample size determination problem in multinomial experiments. We would like to note that the same methods can be applied to derive confidence regions, studied in Sections 2 and 3, for a subset of the multinomial probability vector.

We now proceed to describe several inference problems for multinomial proportions for which the parametric bootstrap approach discussed in this article might be useful. In Section 3 equal-tailed approximate parametric bootstrap intervals were derived for $p_{\min}$ and $p_{\max}$. A more challenging problem is to derive a confidence interval for $p_{(j)}$, the jth ordered multinomial proportion. The main difficulty here is to derive an accurate Edgeworth expansion approximation (or any other approximation) for the distribution of $X_{(j)}$, the jth order statistic in a multinomial vector.

Let $r = p_{\max} - p_{\min}$ be the range for the multinomial cell probabilities. One can utilize the Edgeworth expansion for the rectangular multinomial probabilities to implement an approximate parametric bootstrap confidence interval procedures for r.

A related problem is deriving an approximate confidence interval for $\zeta = p_{\max}/p_{\min}$. Here too one can utilize the Edgeworth expansion for the rectangular multinomial probabilities to implement the parametric bootstrap confidence interval for ζ. Confidence intervals for r and ζ are of interest in investigations of departures from the equal cell probability model in multinomial experiments.

Assume now that we have two independent multinomial populations with parameters $n_1, \theta_{01} = (p_{11}, \ldots, p_{k1})$ and $n_2, \theta_{02} = (p_{12}, \ldots, p_{k2})$, respectively. We are interested in simultaneous confidence intervals for $p_{i2} - p_{i1}, 1 \leqslant i \leqslant k$. Fitzpatrick and Scott (1987) derived a conservative rectangular confidence region for this problem. From Sison and Glaz (1995a) it is evident that this type of a confidence region tends to be too conservative. It will be interesting to investigate the performance of an approximate parametric bootstrap confidence region for $p_{i2} - p_{i1},\ 1 \leqslant i \leqslant k$.

Let $\boldsymbol{X} = (X_1, \ldots, X_k)$ be an observation from a multinomial distribution with parameters n_1 and $\theta_0 = (p_1, \ldots, p_k)$. We are interested in simultaneous confidence intervals for $p_i - p_j,\ 1 \leqslant i < j \leqslant k$. Goodman (1965), and Fitzpatrick and Scott (1987) have derived conservative asymptotic rectangular confidence regions for this problem. It will be interesting to derive an approximate parametric bootstrap confidence region based on the Edgeworth expansion approximation and compare its performance with the two procedures mentioned above.

Acknowledgements

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 the referees for helpful comments.

References

Angers, C., 1984. Large sample sizes for the estimation of multinomial frequencies from simulations studies. Simulation 27, 175–178.

Bishop, Y., Feinberg, S., Holland, P., 1975. Discrete Multivariate Analysis: Theory and Practice. MIT Press, Cambridge, MA.

Cochran, W.G., 1963. Sampling Techniques, second ed. Wiley, New York.

DiCiccio, T., Efron, B., 1992. More accurate confidence intervals in exponential families. Biometrika 79, 231–245.

Efron, B., 1987. Better bootstrap confidence intervals (with discussion). J. Amer. Statist. Assoc. 82, 171–200.

Ethier, S.N., 1982. Testing for favorable numbers on a roulette wheel. J. Amer. Statist. Assoc. 77, 660–665.

Fitzpatrick, S., Scott, A., 1987. Quick simultaneous confidence intervals for multinomial proportions. J. Amer. Statist. Assoc. 82, 875–878.

Gelfand, A.E., Glaz, J., Kuo, L., Lee, T.M., 1992. Inference for the maximum cell probability under multinomial sampling. Naval Res. Logist. 39, 97–114.

Goodman, L.A., 1965. On simultaneous confidence intervals for multinomial proportions. Technometrics 7, 247–254.

Hall, P., 1992. The Bootstrap and Edgeworth Expansion. Springer, New York.

Hochberg, Y., Tamhane, A.C., 1988. Multiple Comparison Procedures. Wiley, New York.

Hurtubise, R., 1969. Sample size and confidence intervals associated with a Monte Carlo simulation model possessing a multinomial output. Simulation 12, 71–77.

Kunte, S., Upadhya, K.S., 1996. Estimating multinomial probabilities. Amer. Statist. 50, 214–216.

Laird, N., Louis, T.A., 1987. Empirical Bayesian confidence intervals based on bootstrap sampling. J. Amer. Statist. Assoc. 82, 253–263.

Levin, B., 1981. A representation for multinomial cumulative distribution functions. Ann. Statist. 9, 1123–1126.

Quesenberry, C.P., Hurst, D.C., 1964. Large-sample simultaneous confidence intervals for multinomial proportions. Technometrics 6, 191–195.

Sison, C.P., 1995. Simultaneous confidence intervals, sample size determination and testing procedures for multinomial proportions. Unpublished Ph.D. Thesis, Department of Statistics, University of Connecticut, Storrs.

Sison, C.P., Glaz, J., 1993. Simultaneous confidence intervals and sample size determination for multinomial proportions. Technical Report No. 93-5, Department of Statistics, University of Connecticut, Storrs.

Sison, C.P., Glaz, J., 1995a. Simultaneous confidence intervals and sample size determination for multinomial proportions. J. Amer. Statist. Assoc. 90, 366–369.

Sison, C.P., Glaz, J., 1995b. Confidence intervals for the maximum and minimum multinomial cell proportions. Technical Report No. 95-15, Department of Statistics, University of Connecticut, Storrs.

Thompson, S.K., 1987. Sample size for estimating multinomial proportions. Amer. Statist. 41, 42–46.

Tortora, R.D., 1978. A note on sample size estimation for multinomial populations. Amer. Statist. 32, 100–102.