Statistics and Its Interface Volume 7 (2014) 219–233

Modern sample size determination for unordered categorical data

Junheng Ma[†], Jiayang Sun[*,†], and Joe Sedransk

Sample size determination is one of the most important practical tasks for statisticians. In this paper, we study sample size determination for unordered categorical data, with or without a pilot sample. With a pilot sample, we provide a minimal difference method, a first order correction, and bootstrap methods for sample size determination in the comparison of two multinomial distributions using the usual chi-squared test. We also propose a Bayesian approach that uses an extension of a posterior predictive p-value. The performance of these methods is investigated via both a simulation study and a real application to leukoplakia lesion data. We advocate a better performance measure than MSE when the sampling distribution is highly skewed. Practical recommendations are given. Some asymptotic results are also provided.

AMS 2000 subject classifications: Primary 62F10, 62F15; secondary 62F40.

Keywords and phrases: Bootstrap, Calibrated posterior predictive p-value, Multinomial distribution, Pilot data, Power calculation, Practical recommendations.

1. INTRODUCTION

Sample size determination (SSD) is one of the most important practical tasks for statisticians. There has been continuous research to develop appropriate methodology, e.g., methods for sample size determination for continuous or ordered categorical outcome data. However, methodology is relatively limited for sample size determination in comparative studies with unordered categorical data, which has important applications. For example, one of our collaborators was interested in deciding the number of images that he needs radiologists to view in order to evaluate the quality of two contrasting imaging modalities. Other examples include comparing the distributions of the locations of leukoplakia lesions for different smoking or chewing habits [20] and comparing the response patterns for two or more different combinations of treatments in clinical trials [7]. A somewhat different type of example is related to survival data, where the five-year survival rate is often used to evaluate the efficacy of different treatments in clinical trials. However, the difference could also be seen at other important landmark years, such as 1, 2, 3, 10, and 15 years (see Figures 2 and 3 in [15]). Thus, to have a more complete comparison of two survival functions, either simultaneous confidence bands for both survival functions should be provided, or survival rates under the two treatments should be compared for selected time periods, i.e., leading to a comparison of categories.

Unordered categorical data from two populations can be modeled by a two-multinomial model, where $n_{i1}, n_{i2}, \ldots, n_{ik}$ are the observed cell frequencies from a multinomial distribution with parameters $(n_i, \mathbf{p}_i)$, denoted as *Multinomial*$(n_i, \mathbf{p}_i)$, for $i = 1, 2$. Here n_i is the sample size and $\mathbf{p}_i = (p_{i1}, p_{i2}, \ldots, p_{ik})$ is the vector of cell probabilities, satisfying $\sum_{j=1}^{k} n_{ij} = n_i, \sum_{j=1}^{k} p_{ij} = 1$ for $i = 1, 2$. So, to compare these two multinomial populations, one could test if there is a difference between $\mathbf{p}_1$ and $\mathbf{p}_2$, given observed samples $\{n_{ij}\}$. In practice, before observing the two multinomial samples, one might be interested in determining the minimum sample sizes n_1 (for population 1) and n_2 (for population 2) needed to draw the samples to achieve a specified power for a test of the null hypothesis that $\mathbf{p}_1 = \mathbf{p}_2$ against the alternative hypothesis that $\mathbf{p}_1 \neq \mathbf{p}_2$. This SSD problem is the focus of our investigation.

A selective literature review on sample size determination includes the books [5, 6], an excellent review article [2], a representative Bayesian paper [27], and references therein. In detail, [6] presented common methods for determining the sample sizes in experiments and sample surveys. [5] provided a comprehensive and unified presentation of statistical procedures (Bayesian and non-Bayesian) for sample size calculation needed at various phases of clinical research. [27] proposed a simulation-based approach to sample size determination in two situations. The first situation is finding the sample size needed to achieve specified performance with regard to one or more features of a model. The second one is selecting a sample size to achieve a specified separation of two models.

There is limited specific literature on *practical* SSD for comparing differences of two multinomial vectors, but there have been many references on SSD for one multinomial population. [25, 26] considered sample size determination for simultaneously estimating the parameters of a multinomial distribution with a specified confidence interval width; while [1] developed a Bayesian approach to select the sample size such that the parameter of interest is contained in

[*] Corresponding author.

[†] Ma and Sun were supported in part by an NSF grant to Sun.

a tolerance region with specified probability (which holds on average over all possible samples). [11] investigated using information from a previous study to determine the sample size for a chi-squared test. [23] developed two procedures to construct simultaneous confidence intervals for the multinomial proportions and proposed two corresponding sample size determination methods to achieve a specified coverage probability.

In this article, we find the sample sizes needed to detect a difference between two vectors of multinomial proportions using both frequentist (or non-Bayesian) and Bayesian approaches, although the procedures should be applicable to comparisons of I multinomial populations for $I \geq 2$.

Our paper is organized as follows. In Section 2, we briefly set up and review general approaches for sample size determination. In Section 3, we study the chi-squared test for contingency tables, recommend three practical implementation methods for the sample size determination, and investigate some asymptotic properties. In Section 4, we develop bootstrap and other improvements to the basic approach in Section 3. In Section 5, we use the concepts of posterior predictive p-values [9, 10, 17] and calibrated posterior predictive p-values (cppp) [14] to develop a Bayesian approach to sample size determination. More recent applications of cppp can be found in [4, 24]. The simulation studies, a real data application and their results are described and summarized in Sections 6 and 7. This investigation provides a much needed comparative study among the frequentist procedures, and between the frequentist and Bayesian approaches, providing practical advice for SSD for unordered categories. Our conclusions are in Section 8, followed by an Appendix.

2. GENERAL APPROACHES FOR SAMPLE SIZE DETERMINATION

All sample size determination methods require that the sample sizes satisfy a specified target level of accuracy or maximize a specified objective function. The approaches for sample size determination can be classified into two broad groups, Bayesian and non-Bayesian.

In a typical non-Bayesian (i.e. frequentist) approach, we first specify a null and an alternative hypothesis for the parameter of interest, μ, say, $H_0 : \mu \in \omega, \quad H_1 : \mu \in \Omega - \omega$. Second, we specify a desired test size α and power $(1-\beta)$. Then, a critical value for deciding when to reject H_0 is obtained, so that the probability of making a type I error is close to but not larger than α, i.e., $Pr\{\text{reject } H_0|\mu\} \leq \alpha$ for $\mu \in \omega$. Third, based on the critical value, we choose *the smallest sample size* n *to satisfy* $Pr\{\text{reject } H_0|\mu\} \geq 1-\beta$, where μ is a specified value in $\Omega - \omega$. This method usually requires finding an exact or approximate pivotal quantity that is related to the test statistic and the computation is specific to the problem under consideration. Frequentist methods such as this have been used widely in practice and can also serve as a starting point for a Bayesian approach; see our algorithm in Section 5.

In a typical Bayesian approach, we first specify a prior distribution for the model parameter(s). Second, we calculate the posterior distribution of the parameters based on the prior distribution and data. Third, we find a test quantity (also called a discrepancy measure) which is a function of the data and parameters. Fourth, we compute the posterior predictive p-value based on the test quantity. Then we look for the minimum sample size such that

$$P_{H_1}\{p_q(\vec{n}^{obs}) \leq \alpha\} \geq 1-\beta,$$

where $p_q(\vec{n}^{obs})$ is the posterior predictive p-value based on the observed value $\vec{n}^{obs}$ of n, "H_1" denotes a specified value of μ in $\Omega - \omega$, and (α, β) are choices comparable to those for the frequentist method. Note that this approach for SSD differs from the Bayesian model selection approach that involves computing marginal likelihoods for two competing models and hence the corresponding Bayes factors. In Section 5, we replace the posterior predictive p-value, $p_q(\vec{n}^{obs})$, with the calibrated posterior predictive p-value [14] to overcome the possible non-uniformity of the distribution of the posterior predictive p-values under the null (hypothesis) model.

3. BASIC APPROACH

Consider the hypotheses $H_0 : \mathbf{p}_1 = \mathbf{p}_2 = \tilde{\mathbf{p}}, \quad vs \quad H_1 : \mathbf{p}_1 \neq \mathbf{p}_2$, where $\tilde{\mathbf{p}} = (\tilde{p}_1, \ldots, \tilde{p}_k)$ is some unknown parameter vector, s.t. $\sum_{i=1}^{k} \tilde{p}_i = 1$. A reasonable test statistic for testing the null hypothesis is

$$\mathrm{X}^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(n_{ij} - n_i \hat{p}_{0j})^2}{n_i \hat{p}_{0j}}, \tag{3.1}$$

where n_{ij} is the observed cell frequency for cell j from $Multinomial(n_i, \mathbf{p}_i)$, $n_i = n_{i.} = \sum_{j=1}^{k} n_{ij}$ are the row sums, $n_{.j} = n_{1j} + n_{2j}$ are the column sums, $N = n_1 + n_2$ is the total sample size, and $\hat{p}_{0j} = n_{.j}/N$ is the estimated cell probability under H_0 for $j = 1, 2, \ldots, k$ and $i = 1, 2$. It is straightforward to extend this test for the equality of 2 multinomial distributions to that of I multinomials, but the 2-multinomials case is the most important and hence is the focus of this paper. Under the null hypothesis H_0 and reasonable regularity conditions, as $n \longrightarrow \infty$, $\mathrm{X}^2 \overset{appr.}{\sim} \chi^2_{(k-1)}$, by using properties of the generalized likelihood ratio test. Therefore we may reject H_0 at level α if the observed value of $\mathrm{X}^2 > \chi^2_{(k-1)}(\alpha)$, where $\chi^2_{(k-1)}(\alpha)$ is the *upper* α quantile of the chi-squared distribution with $(k-1)$ degrees of freedom. When the alternative hypothesis H_1 is true, i.e., $\mathbf{p}_1 \neq \mathbf{p}_2$, the test statistic X^2 has approximately a noncentral chi-squared distribution with $(k-1)$ degrees of freedom and noncentrality parameter λ as shown by [18].

We advocate rewriting λ in a way that provides a symmetric expression. This will lead to a more accurate approximation to the sample size than the one without symmetrization. Specifically, we denote $\mathbf{p}_1 = \mathbf{p} + \delta_1$ and $\mathbf{p}_2 = \mathbf{p} + \delta_2$, where $\mathbf{p} = (\mathbf{p}_1 + \mathbf{p}_2)/2$ and $\delta_{\mathrm{i}} = \mathbf{p}_{\mathrm{i}} - \mathbf{p}$ for $i = 1, 2$ respectively. A general formula to calculate the noncentrality parameter λ for a $(2 \times k)$ contingency table is then

$$\lambda = N \times q_1 \times q_2 \times \sum_{j=1}^{k} \left\{ \frac{\Delta_j^2}{p_j} \right\},$$

where $q_i = n_i/N$ for $i = 1, 2$; $\Delta_j = p_{1j} - p_{2j}$, and $p_j = (p_{1j} + p_{2j})/2$ for $j = 1, 2, \ldots, k$. In a balanced design, we have $n_1 = n_2 = n$ and then $\lambda = (n/2) \sum_{j=1}^{k} \{\Delta_j^2/p_j\}$. To have large power, we require that $Pr_{H_1}\{\mathrm{X}^2 > \chi^2_{(k-1)}(\alpha)\} \geq 1 - \beta$. Since X^2 has a noncentral chi-squared distribution with $(k-1)$ degrees of freedom and noncentrality parameter λ, then λ satisfies

$$\chi^2_{(k-1),\lambda}(1-\beta) \geq \chi^2_{(k-1)}(\alpha), \tag{3.2}$$

where $\chi^2_{(k-1),\lambda}(1-\beta)$ is the upper $(1-\beta)$ quantile of the noncentral chi-squared distribution with $(k-1)$ degrees of freedom and noncentrality parameter λ.

Given α, β and k, we can find the value of λ_0 *from a noncentral chi-squared table that makes the equality in (3.2) hold*. For example, for $k = 5$ and $\alpha = 0.05, \beta = 0.2$, $\lambda_0 = 11.94$. Also, note that $\chi^2_{(k-1),\lambda}(1-\beta)$ increases in λ, so the minimal n required to have power $1-\beta$ to detect the difference when $\lambda \geq \lambda_0$ is

$$n = \left\lceil 2\lambda_0 \left(\sum_{j=1}^{k} \frac{\Delta_j^2}{p_j} \right)^{-1} \right\rceil, \tag{3.3}$$

where $\lceil x \rceil$ denotes the smallest integer that is greater than or equal to x. The formula in (3.3) provides the **basic approach** to sample size determination by the frequentist method. In the **unbalanced case**, let $r = n_1/n_2$ be the ratio of n_1 to n_2, specified in advance. Then it can be shown easily that

$$n_1 = \left\lceil (r+1)\lambda_0 \left(\sum_{j=1}^{k} \frac{\Delta_j^2}{p_j} \right)^{-1} \right\rceil, \tag{3.4}$$

which is n in (3.3) if $r = 1$, and $n_2 = n_1/r$. The improvements and Bayesian approach described in detail below are for the balanced case, for simplicity of the notation. The data application has unbalanced data.

We notice that n in (3.3) depends on unknown parameters $\Delta = (\Delta_1, \ldots, \Delta_k)$ and $\mathbf{p} = (p_1, \ldots, p_k)$, the differences and averages of $\mathbf{p}_1$ and $\mathbf{p}_2$. Therefore, we recommend the following three *practical implementations to deal with these unknown parameters.*

- Method M1: Specify a minimum average difference d between $\mathbf{p}_1$ and $\mathbf{p}_2$, and a minimum relative difference, r, that we want to detect. In other words, if the average difference D, and relative difference R_j, are defined to be
$$D = \frac{1}{k} \sum_j |\Delta_j|, \quad R_j = \left| \frac{p_{1j} - p_{2j}}{p_j} \right| = \left| \frac{\Delta_j}{p_j} \right|, \ j = 1, 2, \ldots, k,$$
then the sample size needed to detect a difference with $D \geq d, R_j \geq r$ for $j = 1, 2, \ldots, k$ with power at least $1 - \beta$ is
$$\tilde{n} = \lceil 2\lambda_0 (r \cdot kd)^{-1} \rceil. \tag{3.5}$$
- Method M2: Replace Δ and $\mathbf{p}$ by estimates obtained from historical, pilot, proxy, or approximate data. Then
$$\hat{n} = \left\lceil 2\lambda_0 \left(\sum_{j=1}^{k} \frac{\hat{\Delta}_j^2}{\hat{p}_j} \right)^{-1} \right\rceil, \tag{3.6}$$
where $\hat{p}_j = \frac{1}{2}(\hat{p}_{1j} + \hat{p}_{2j})$, and $\hat{\Delta}_j = \hat{p}_{1j} - \hat{p}_{2j}$.
- Method M3: When comparing k categories of two multinomial populations, if we already know that some categories have the same or similar counts for the two populations, it is highly recommended to remove these categories to reduce the problem to one comparing $k'(< k)$ categories. This is not only practical, but also generally results in a smaller n needed to compare these k' categories than n computed based on the original k categories.

Remark 3.1. Methods M1 and M3 are reasonable, as it is common to ask our collaborators for reasonable smallest differences (i.e., the parameters that are analogous to d and r here) that they would like to detect in practice. It is also common to ask for pilot or proxy data if a SSD method depends on some unknown parameters. A proxy data set can be a relevant data set from the literature. If these data do not exist we would use method M1, or suggest conducting a small pilot study first, if such a pilot study is feasible.

In Method M2, the $\hat{\Delta}_j$ and $\hat{\mathbf{p}}$ in (3.6) are estimates. One situation where (3.6) is appropriate is when an investigator observes $\hat{\Delta}_j$ from a pilot sample and thinks that if $\Delta_j = \delta_{1j} - \delta_{2j}$ was, in fact, no less than $\hat{\Delta}_j$, $j = 1, \ldots, J$, he/she would like to be able to reject $H_0 : \mathbf{p}_1 = \mathbf{p}_2 = \mathbf{p}$ with probability $(1-\beta)$ when H_1 is true. Then, the total sample size needed to achieve this objective is given by (3.6). In the following, we provide some asymptotic properties about $\hat{n}$ obtained in (3.6) to show its justification, give a CI for n, and motivate improvements in Section 4.

In a balanced design, suppose the sample size from a pilot sample is m for each multinomial distribution. Let $t_j = (p_{1j}(1-p_{1j}) + p_{2j}(1-p_{2j}))$. For large m, by the Central Limit Theorem, it's easy to see that

$$\hat{p}_j \overset{d}{=} p_j + Z_{j1} \frac{1}{2} \sqrt{\frac{t_j}{m}} + o_p\left(\frac{1}{m}\right); \ \hat{\Delta}_j \overset{d}{=} \Delta_j + Z_{j2} \sqrt{\frac{t_j}{m}} + o_p\left(\frac{1}{m}\right),$$

where "d" represents equality in distribution, and Z_{j1} and Z_{j2} are standard normal random variables. Using this asymptotic expansion, we have the following lemma and proposition. The proof of Lemma 3.2 is straightforward and the proof of Proposition 3.6 is in the Appendix.

Lemma 3.2. *Given a pilot sample of size m from the two multinomial populations, if m is not too small, then*

$$\frac{\hat{n}}{n} \overset{d}{=} 1 - \frac{A}{n} \cdot \left(\sum_{j=1}^{k} \frac{2\Delta_j \sqrt{t_j}}{p_j} Z_{j2} - \sum_{j=1}^{k} \frac{\Delta_j^2 \sqrt{t_j}}{2p_j} Z_{j1} \right) + o\left(\frac{1}{m}\right), \tag{3.7}$$

where

$$A = 2\lambda \cdot \frac{1}{\sqrt{m}} \cdot \left(\sum_{j=1}^{k} \frac{\Delta_j^2}{p_j} \right)^{-2} = O\left(\frac{1}{\sqrt{m}}\right),$$

and Z_{j1} and Z_{j2} are standard normal random variables. Further, as "$m \to \infty$",

$$Var(\hat{n}) = A^2 \cdot B + o\left(\frac{1}{m^2}\right), \tag{3.8}$$

where

$$B = \sum_{j=1}^{k} (a_j^2 + b_j^2) + \sum_{j=1}^{k} \sum_{i \neq j, i=1}^{k} \{a_j b_i (t_i t_j)^{-\frac{1}{2}} (p_{1i} p_{1j} - p_{2i} p_{2j})\}$$
$$- \sum_{j=1}^{k} \sum_{i \neq j, i=1}^{k} \{(b_i b_j + a_i a_j)(t_i t_j)^{-\frac{1}{2}} (p_{1i} p_{1j} + p_{2i} p_{2j})\},$$

and $a_l = 2\Delta_l \sqrt{t_l}/p_l, b_l = \Delta_l^2 \sqrt{t_l}/(2p_l)$ for $l = 1, 2, \ldots, k$.

Remark 3.3. From (3.7), the estimator $\hat{n}$ has the following property: $\hat{n}/n = 1 + O_p(1/\sqrt{m})$.

Remark 3.3 gives a sense of the magnitude of the ratio between $\hat{n}$ and n, which is consistent to what would be expected of a well behaved estimator of n, of course.

Remark 3.4. From (3.8) we can see that $Var(\hat{n}) = O(1/m)$. Hence (3.7) and (3.8) give us some measure of the accuracy of $\hat{n}$ in estimating the fixed n when m is not too small. We shall examine by simulation how good the estimate $\hat{n}$ is in estimating n for finite n and m in Section 6.

Remark 3.5. All these big or little o's should be interpreted as the approximation when the pilot or proxy sample size is not too small so that the pilot estimates of the parameters are not too erratic. Although we can do asymptotics to say that $Var(\hat{n}) = O(1/m) \to 0$ and $\hat{n} \overset{approx}{\sim} N(n, Var(\hat{n}))$ as $m \to \infty$, they would only be "true" if the "pilot" sample of size m is a proxy or a historical sample and is not included in the total sample size n of the current study (otherwise, n would also go to ∞). In practice, it'd be foolish to not reuse the pilot sample if it's from the same experiment under study. This shows the limitation of an asymptotics study in some cases and confirms the need for studying its actual improvements in finite sample situations, as those in Section 6.

Defining the power of the test under H_1 based on a sample of size $\hat{n}$ to be $Power(\hat{n}) = P_{H_1}(X^2 \geq \chi^2_{k-1}(\alpha))$ we have Proposition 3.6 below.

Proposition 3.6. *If $|\hat{n}/n - 1| < \varepsilon$, then $Power(\hat{n}) = 1 - \beta + O(\varepsilon)$.*

This means that the actual power of $\hat{n}$ tends to $(1 - \beta)$ as $\hat{n} \to n$ if the pilot sample gives a good estimate of the unknown parameters or as $m \to \infty$. See the Appendix for the proof and a bound for the $O(\varepsilon)$ term.

For a pilot study with sample size m, we use the $\hat{\Delta}_j$'s and $\hat{p}_j$'s to estimate the Δ_j's and p_j's, and let

$$\sigma_0^2 = \hat{A}^2 \left\{ \sum_{j=1}^{k} (\hat{a}_j^2 + \hat{b}_j^2) + \sum_{j=1}^{k} \sum_{i \neq j, i=1}^{k} \hat{a}_j \hat{b}_i (\hat{t}_i \hat{t}_j)^{-\frac{1}{2}} (\hat{p}_{1i} \hat{p}_{1j} - \hat{p}_{2i} \hat{p}_{2j}) \right\}$$
$$- \sum_{j=1}^{k} \sum_{i \neq j, i=1}^{k} \{(\hat{b}_i \hat{b}_j + \hat{a}_i \hat{a}_j)(\hat{t}_i \hat{t}_j)^{-\frac{1}{2}} (\hat{p}_{1i} \hat{p}_{1j} + \hat{p}_{2i} \hat{p}_{2j})\}$$

which is an ugly, long expression, but is extremely simple to compute. Then, a $100(1 - \alpha)\%$ approximate confidence interval for n is $(\hat{n} - z_{\alpha/2}\sigma_0, \hat{n} + z_{\alpha/2}\sigma_0)$, where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution.

4. IMPROVEMENTS TO THE BASIC APPROACH

Methods M1 and M3 in the basic approach (Section 3) are simple, practical, and straightforward to implement. There is no need to find improvements to M1 and M3 if they are applicable. Method M2 given in (3.6) is based on the estimated parameters from a pilot sample or proxy data. In order to reduce the sampling error of the estimate from the pilot sample or proxy data and also to stabilize the estimated required sample size, we study the following improvements to M2, one borrowing the computing power by bootstrap, one by posing a minimum difference, and another by a simple "ad hoc" correction using asymptotics.

4.1 Bootstrap method (BOOT)

The basic idea of the bootstrap method is to draw bootstrap samples from the pilot data and take, typically, the mean or median of the calculated sample sizes computed from the bootstrap samples as the required sample size, see [5].

Parametric bootstrap-mean and bootstrap-median methods. With these two methods, bootstrap samples are drawn parametrically from multinomial distributions where the parameters of the multinomial distributions are estimated from the pilot data. Then we calculate the sample sizes using (3.6) for each of the bootstrap samples, for $i = 1, 2, \ldots, B$,

$$n^{(*i)} = 2\lambda_0 \left(\sum_{j=1}^{k} \frac{(\hat{\Delta}_{ij}^{(*i)})^2}{\hat{p}_{ij}^{(*i)}} \right)^{-1},$$

where $\hat{\Delta}_{ij}^{(*i)} = \hat{p}_{1j}^{(*i)} - \hat{p}_{2j}^{(*i)}$, $\hat{p}_{ij} = \frac{1}{2}(\hat{p}_{1j}^{(*i)} + \hat{p}_{2j}^{(*i)})$ and the $\hat{p}_{lj}^{(*i)}$, $l = 1, 2$, are the usual estimates but based on the ith bootstrap sample and B is the bootstrap sample size. Then take the mean and median of $\{n^{*i}, i = 1, 2, \ldots, B\}$ as the required sample sizes for the bootstrap-mean method and for the bootstrap-median method, respectively. Preliminary simulations suggested that both bootstrap methods underestimate the true sample size especially when the pilot sample size is either small or moderate. In order to correct for the underestimation issue, we propose using the 75% or 80% quantile of the estimates from the bootstrap samples as the estimated sample size, after experimenting further bias and skewness corrections including a bootstrap BCa method (§6 and §7).

Non-parametric bootstrap-mean and median methods. For a non-parametric bootstrap method, one would proceed as for a parametric bootstrap method but draw the bootstrap samples nonparametrically, i.e., from the pilot data using random sampling with replacement. However, it is important and straightforward to see that the sampling distribution of a non-parametric bootstrap from the pilot sample is also multinomial with cell probabilities set to be the two relative frequencies computed from the pilot data. So in this multinomial case, the parametric and non-parametric bootstrap procedures yield the same result, which is also confirmed by our simulation study. In the following, we will just examine the parametric bootstrap method and call it BOOT.

4.2 Minimum difference method (MIN)

If a scientist is interested only in parameter differences that are at least as large as c, using this information in finding the required sample sizes usually will reduce the sample sizes (relative to not using this information) and can improve the accuracy of the SSD if the true parameter differences are indeed at least as large as c. Doing this the required sample size is

$$\hat{n} = \left\lceil 2\lambda_0 \left(\sum_{j=1}^{k} \frac{\hat{C}_j^2}{\hat{p}_j} \right)^{-1} \right\rceil, \tag{4.1}$$

where $\hat{C}_j = max(c, |\hat{\Delta}_j|)$. Formula (4.1) is the same as that in (3.6) except that $\hat{\Delta}_j$ is replaced by $\hat{C}_j$. The common $\hat{p}_j$ does not change because when Δ_j is changed to $\hat{C}_j$, $\hat{p}_{j1}$ is changed to $\hat{p}_{j1}^* = \hat{p}_{j2} + C_j$ and $\hat{p}_{j2}$ to $\hat{p}_{j2}^* = \hat{p}_{j1} - C_j$, which imply that $\hat{p}_j^* = \frac{1}{2}(\hat{p}_{j1}^* + \hat{p}_{j2}^*) = \frac{1}{2}(\hat{p}_{j1} + \hat{p}_{j2}) = \hat{p}_j$. In practice, c is determined by the scientist as the smallest difference that would be useful and meaningful to detect. That is, the objective of the minimum difference method is to find the sample size so that the power is at least $1-\beta$ in the parameter space where the differences are at least as big as c. This minimal difference method differs from method M1 in that in method M1, one would need to know $d \geq c$ (here) and r. Here, without r, we use both c and the pilot sample to estimate n.

4.3 Correction method (CORR)

From (3.7), we see that the 2nd term in the right hand side

$$A \cdot \left(\sum_{j=1}^{k} \frac{2\Delta_j \sqrt{t_j}}{p_j} Z_{j2} - \sum_{j=1}^{k} \frac{\Delta_j^2 \sqrt{t_j}}{2p_j} Z_{j1} \right)$$

is of the order of $1/\sqrt{m}$. To rigorously correct the bias of $\hat{n}$ in estimating n, one would need to derive an inverse Edgeworth expansion or Cornish-Fisher expansion, and then to evaluate the finite sample performance of this correction numerically. This rigorous approach is left as future research. Here, we opt for a simple, "ad hoc" bagging approach; i.e., draw N pairs of random samples for (Z_{j1}, Z_{j2}) from standard normal distributions and then examine the performance of the "ad hoc" corrected estimate

$$\hat{n}_0 = \hat{n} + \frac{1}{N} \sum_{i=1}^{N} \left[\hat{A} \cdot \left(\sum_{j=1}^{k} \frac{2\hat{\Delta}_j \sqrt{\hat{t}_j}}{\hat{p}_j} Z_{j2}^{(i)} - \sum_{j=1}^{k} \frac{\hat{\Delta}_j^2 \sqrt{\hat{t}_j}}{2\hat{p}_j} Z_{j1}^{(i)} \right) \right] \tag{4.2}$$

numerically. Here $\hat{A}, \hat{\Delta}_j, \hat{t}_j$ and $\hat{p}_j$ are the estimates of A, Δ_j, t_j and p_j respectively. The performance of this simple correction procedure will be evaluated together with the original (M2), the bootstrap (BOOT) and minimal difference (MIN) procedures.

5. BAYESIAN APPROACH

We now describe a Bayesian method that is motivated by and parallel to the frequentist approach. Under the null model, i.e., H_0, the two multinomial distributions have the same probability parameter $\mathbf{p}^* = (p_1^*, p_2^*, \ldots, p_k^*)$ such that $\vec{n}_i = (n_{i1}, n_{i2}, \ldots, n_{ik}) \sim$ multinomial(p^*, n_i), $i = 1, 2$, where $n_i = \sum_{j=1}^{k} n_{ij}$. For a multinomial distribution, the conjugate prior distribution is Dirichlet with density function,

$$f(\mathbf{p}^* | \alpha) \propto \prod_{j=1}^{k} p_j^{*\alpha_j - 1},$$

where the distribution is restricted to nonnegative $\mathbf{p}^* = (p_1^*, \ldots, p_k^*)$ with $\sum_{j=1}^{k} p_j^* = 1, p_j^* \geq 0$ and $\alpha_j > 0$, for $j = 1, \ldots, k$. The resulting posterior distribution of the $\mathbf{p}^*$, given $\vec{n} = (n_{ij}, i = 1, 2; j = 1, 2, \ldots, k)$ under H_0, is

$$f(\mathbf{p}^* | \vec{n}) \propto \prod_{j=1}^{k} p_j^{*(\alpha_j + n_{.j} - 1)}, \tag{5.1}$$

where $n_{\cdot j} = \sum_{i=1}^{2} n_{ij}$. Let $t(n, \mathbf{p})$ be a discrepancy measure and let $\vec{n}^{obs}$ be an observed value of $\vec{n}$. A natural choice of $t(\vec{n}, \mathbf{p})$ is

$$t(\vec{n}, \mathbf{p}) = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(n_{ij} - n_i p_j)^2}{n_i p_j}$$

as in (3.1), where $\vec{n} = (n_{ij} : i = 1, 2; j = 1, 2, \ldots, k)$, $\mathbf{p} = (p_1, p_2, \ldots, p_k)$ and $n_i = \sum_{j=1}^{k} n_{ij}$. Then the *posterior predictive p-value* (ppp) is defined as

$$\text{(5.2)} \quad p_{ppp}\{\vec{n}^{obs}\} = Pr_{H_0}\{t(\vec{n}^{pred}, \mathbf{p}^*) > t(\vec{n}^{obs}, \mathbf{p}^*)|\vec{n}^{obs}\},$$

where $\mathbf{p}^*$ is the common vector of proportions for the two multinomial distributions. The probability in (5.2) is taken over the joint distribution of $(\vec{n}^{pred}, \mathbf{p}^*)$ given $\vec{n}^{obs}$, in which $\mathbf{p}^* \sim f(\mathbf{p}^*|\vec{n}^{obs})$ is Dirichlet in (5.1), and $\vec{n}^{pred}$ is a predictive value of $\vec{n}$, where given $\mathbf{p}^*$, $\vec{n}^{pred}$ is assumed to be independent of $\vec{n}^{obs}$, and $\vec{n}_i^{pred}$ has the multinomial distribution, $\text{mult}(\mathbf{p}^*, n_i^{obs})$. One may reject the null model at level α if $p_{ppp}\{\vec{n}^{obs}\} < \alpha$. However, [14] explained problems with the posterior predictive p-value: "the ppp calculation uses the data twice, first updating the prior to fit the data better and then estimating how surprising the data are relative to the posterior parameter distribution. Thus it is not surprising that its distribution across likely values of y_{obs} is not uniform; we can, in fact, demonstrate various extreme aspects of non-uniformity in several situations. This makes the interpretation and comparison of ppp values a difficult and risky matter."

To correct the possible non-uniformity of the distribution of the posterior predictive p-value, a calibrated posterior predictive p-value was proposed by [14], i.e.,

$$p_{cppp}\{\vec{n}^{obs}\} = Pr_{H_0}\{p_{ppp}\{\vec{n}\} \le p_{ppp}\{\vec{n}^{obs}\}\},$$

where $\vec{n}$ comes from the marginal distribution derived from the joint distribution of $(\mathbf{p}^*, \vec{n})$ in which $\vec{n}$ has the multinomial distribution with parameter $\mathbf{p}^*$, which comes from the prior distribution $\pi(\mathbf{p}^*)$. Another attempt to improve posterior predictive checks was investigated by [16].

Since the calibrated posterior predictive p-value has a uniform distribution under the null model, it can play the role of a classical p-value.

Double-simulation can be used to estimate the value of $p_{cppp}\{\vec{n}^{obs}\}$. In our case, given specified sample sizes $n_1 = n_2 = m$, the value of $p_{ppp}\{\vec{n}^{obs}\}$ can be evaluated by simulation,

$$\text{(5.3)} \quad p_{ppp}\{\vec{n}^{obs}\} \approx \frac{1}{A} \sum_{i=1}^{A} I\big(t(\vec{n}_i^{pred}, p_i^*) \ge t(\vec{n}^{obs}, p_i^*)\big),$$

for a large simulation size A, where each p_i^* is simulated from the posterior distribution $f(p^*|\vec{n}^{obs})$ and $\vec{n}^{pred}$ is simulated from the multinomial distribution with parameters p_i^* and m. The simulation is repeated independently A times.

Then for a large number B, the calibrated posterior predictive p-value can be estimated by simulation,

$$p_{cppp}\{\vec{n}^{obs}\} \approx \frac{1}{B} \sum_{j=1}^{B} I\big(p_{ppp}\{\vec{n}^{(j)}\} \le p_{ppp}\{\vec{n}^{obs}\}\big),$$

where $\mathbf{p}^*$ is simulated from the prior distribution $\pi(\mathbf{p}^*)$ and each $\vec{n}^{(j)}$ is simulated from a multinomial distribution with parameter $\mathbf{p}^*$ and m. For each $\vec{n}^{(j)}$, $p_{ppp}\{\vec{n}^{(j)}\}$ can be approximated using (5.3). The simulation is repeated independently B times.

Given a significance level α and power $1 - \beta$, we want to find the minimum sample size for each multinomial distribution such that

$$Pr_{H_1}\{p_{cppp}(\vec{n}^{obs}) \le \alpha\} \ge 1 - \beta,$$

where "H_1" denotes a specified value in $\Omega - \omega$.

In a balanced design, suppose that the sample size is m for each multinomial. We generate samples from the *two* multinomials N times, $\vec{n}_i^j = (n_{i1}^j, n_{i2}^j, \ldots, n_{ik}^j)$, where $j = 1, 2, \ldots, N$ and $i = 1, 2$.

Let $\vec{n}_j^{obs} = (\vec{n}_1^j, \vec{n}_2^j)$ for $j = 1, 2, \ldots, N$. We need to evaluate the calibrated posterior predictive p-value, $p_{cppp}(\vec{n}_j^{obs})$, for each $\vec{n}_j^{obs}$. Then we compare each calibrated posterior predictive p-value with the specified significance level α and calculate the proportion for which $p_{cppp}(\vec{n}_j^{obs}) \le \alpha$.

If the calculated proportion is less than $1-\beta$, we increase the sample size m and repeat the whole process. Otherwise, we decrease the sample size m and repeat the process. Finally, we can find the minimum sample size m such that the calculated proportion is at least $1 - \beta$.

Since this process is very computationally intensive, a good starting point of m is very important. In practice, we use the hypothesis test method to obtain a reasonable starting point.

The following is our proposed calibrated posterior predictive p-value based procedure to find the desired m:

- *Choose a starting value of m.*
- *Select $\vec{p}_1$ and $\vec{p}_2$ from $\Omega - \omega$. Here, $\vec{p}_1$ and $\vec{p}_2$ are the parameters representing the minimal differences one would like to detect, or, possibly, estimates from pilot data.*
- *Generate samples $\vec{n}_1$ and $\vec{n}_2$ N times, $\vec{n}_i^j \sim$ multi$(m, \vec{p}_i)$, where $j = 1, 2, \ldots, N$ and $i = 1, 2$.*
- *Let $\vec{n}^j = (\vec{n}_1^j, \vec{n}_2^j)$ for $j = 1, 2, \ldots, N$. Evaluate the calibrated posterior predictive p-value $p_{cppp}(\vec{n}^j)$ for each $\vec{n}^j$.*
- *Compare each calibrated posterior predictive p-value with the specified significance level α and calculate the proportion for which $p_{cppp}(\vec{n}^j) \le \alpha$.*

For a specified β, choose alternative values of m until the proportion for which $p_{cppp}(\vec{n}^j) \le \alpha$ is about $1 - \beta$.

6. SIMULATION STUDY

In this simulation study, we evaluate the performance of both the frequentist and Bayesian methods when there is a pilot sample, contrasting these with the basic method in (3.6). In practice, the pilot sample size should depend on the model or problem under consideration. Therefore, we choose three representative experimental settings (Experiments 1, 2 and 3). Under each of the three experiments, we choose pilot sample sizes ranging from 10 to 200. For each parameter setting and fixed pilot sample size we generate pilot data 5000 times; and for each pilot data, we compute $\hat{n}$ by the five frequentist methods, i.e., the original method M2 using (3.6), and the improvements in Sections 4.1–4.3 (bootstrap mean and median, minimum difference and correction methods). For the minimum difference method (Section 4.2), the specified minimum difference between the proportions of the two multinomial distributions is set to be 0.02, i.e., $c = 0.02$. However, we have found that the bootstrap mean and median corrections have severely underestimated the target; see the 3rd and 4th boxplots in each of the panels (for pilot sample sizes 20 to 200) in all Figures 1–4. Then we tried to develop further corrections that would incorporate both bias and skewness of the bootstrap estimates, using the bootstrap BCa idea and an ad-hoc skewness correction idea; unfortunately, these "further corrections" did not help with the underestimation, but had in fact increased the variances from the original bootstrap counterparts. Therefore, in order to uplift the entire set of the bootstrap mean and median estimates, we also examined the performance of bootstrap 75 and 80 quantile estimates (i.e., taking the 75th or 80th percentile instead of the median of bootstrap duplicates). See the 5th and 6th boxplots for the quantile estimates in *each* of the panels in Figures 1–4, where a total of seven methods are compared side-by-side ending with the minimal difference method in the 7th boxplot.

6.1 Experiment 1 (small differences)

We first set the proportions for the two multinomial distributions to be $\vec{p}_1 = (0.10, 0.25, 0.30, 0.20, 0.15)$ and $\vec{p}_2 = (0.15, 0.20, 0.25, 0.30, 0.10)$. Clearly, the true differences between p_{1j} and p_{2j} are small, mostly at 0.05. Thus, the required sample size to detect the difference can be large. Indeed, if these values of $\vec{p}_1$ and $\vec{p}_2$ are known, $\alpha = 0.05$ and $\beta = 0.20$, the true required sample size from (3.3) is 239. Figure 1 is our side-by-side modified boxplot comparison of the estimated sample sizes for the seven frequentist methods when the values of $\vec{p}_1$ and $\vec{p}_2$ are unknown but some pilot data are available. These modified boxplots have all the information provided by the standard boxplots except for the points which are outside of the whisker; the counts of these outside points are presented at the top of each modified boxplot. The horizontal line in the middle of the plot indicates the true sample size, 239. The circle, triangle, and solid line indicate the mean, standard deviation, and median (over the 5000 replicates) of the estimated sample size using each method. This modified boxplot gives a simple, compact, and informative view of the data although other modifications of the standard boxplot are possible.

From the simulation results, in terms of mean (the circle ◦) and standard deviation (the triangle △), we see that the original and correction methods (the 1st and 2nd boxes) have somewhat similar performance although the correction method seems to have corrected the bias slightly, but at the expense of increasing the standard deviation sometimes. The minimum difference method produces comparable means and medians as the original and correction methods do, but with much smaller standard deviations. Comparing these three methods with the two basic bootstrap methods (using bootstrap mean and median estimates), four of the five methods (except bootstrap median) approach the required sample size 239 as the pilot sample size increases while the bootstrap median estimate also tries to reach 239, though very slowly. This confirms our intuition and the asymptotics that we derived. However, for the smallest pilot sample size (30) there is a significant underestimation for all five methods, while the underestimation of the bootstrap mean and median methods are much more severe than the other three methods, although their standard deviations are much smaller than those of the other three methods. Clearly, the mean and standard deviation are not the only possible measures of the performance. The bootstrap median estimates have the smallest standard deviation but 75% of them, i.e., the entire box of bootstrap median estimates, are below the target line even for large pilot sample sizes (150 and 200) and hence they are not good estimates.

On the another hand, the bootstrap 75% and 80% methods, in terms of the performance of median (the bar in each of the boxes) provide great improvements over the other methods, i.e., approaching the target more quickly than others for moderate and large pilot sample sizes (80, 120, 150 and 200).

Up to about 10% of the cases, the calculated sample sizes can be extremely large due to small estimated $\hat{p}_i$ from the pilot data. In a practical situation one would either remove the categories with extremely small probabilities of occurrences or draw another pilot sample before conducting the main study. So, we repeated the simulation study (with 5000 replicates) but required that the calculated sample sizes for all 5000 replicates not exceed 900 which is a large, conservative, upper bound. To have a tighter upper bound, one can use M1 by imposing conservative values of r and d. The results, presented in Figure 2, show that for all the seven methods the means approach the required sample size, 239, as the pilot sample increases though the bootstrap median estimates are the slowest. For the smaller pilot sample size (30), there is also a significant underestimation. The seven

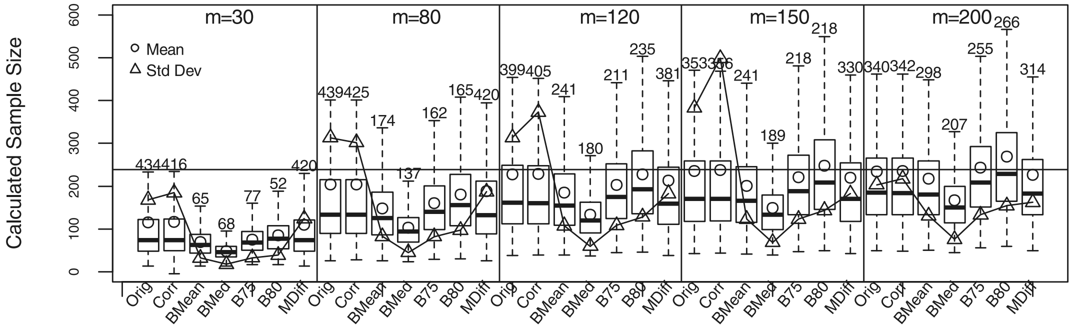

Figure 1. Side-by-side modified Box plot for Experiment 1.

Note: m is the pilot sample size; Orig = Original, Corr = Correction, BMean = Bootstrap Mean, BMed = Bootstrap Median, B75 = Bootstrap 75%, B80 = Bootstrap 80%, MDiff = Mini-Diff. Same abbreviations for Figures 2, 3, and 4.

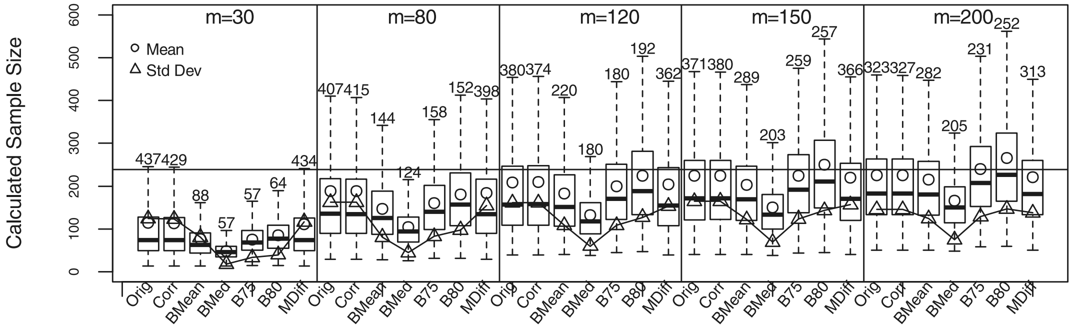

Figure 2. Side-by-side modified Box plot for Experiment 1 with cap 900.

methods have slightly smaller means but much smaller standard deviations than those of the uncapped cases in Figure 1. Again, the bootstrap mean and median methods provide smaller means, medians and standard deviations than the other three methods, i.e., original, correction, and minimum difference methods. Both the bootstrap 75% and 80% methods have better performance in terms of the median for moderate and large pilot sample sizes (80, 120, 150 and 200). The original, correction, and minimum difference methods have similar performance.

Alternative performance measure. Since the bootstrap mean and median methods severely underestimate the target especially for small and moderate pilot sample sizes (30 and 80), mean square error (MSE) could provide misleading information in comparing the performance of the seven methods. To provide a practical recommendation, we rank the methods based on their absolute differences from the target, in two ways.

Let $d_{ij} = |s_{ij} - n_0|$ be the absolute difference of the estimated sample size s_{ij} from the ith replicate and jth method to the target n_0, for $j = 1, \ldots, 7$, and $i = 1, \ldots, 5000$. The first rank measure R_1 is computed by ranking the seven absolute differences d_{ij}, $j = 1, \ldots, 7$, for each replicate, and then averaging 5000 ranks from the replicates for each of seven methods. That is, for method j, $R_{1j} = ave_i(rank_j d_{ij})$. The second rank measure R_2 is computed by averaging absolute differences over 5000 replicates for each method and then ranking the seven averages, i.e., for method j, $R_{2j} = rank_j(ave_i d_{ij})$. We then average these two rank measures to obtain the final combined Rank, $R = (R_1 + R_2)/2$ for the seven methods, where $R_i = (R_{i1}, \ldots, R_{i7})$ is the vector of the ranks of seven methods for $i = 1, 2$.

A summary of these measures when there is no sample size cap is given in Table 1 on page 227, while a summary with the sample size cap of 900 is in Table 2 on page 227. From Tables 1 on page 227 and 2 on page 227 it is clear that for small and moderate pilot sample sizes (30 and 80), the bootstrap 80% performs the best while the bootstrap 75% performs the best for fairly large pilot sample sizes (120 and 150). For large pilot sample size (200), the bootstrap 75% performs better than others for the uncapped case and the bootstrap median and minimum difference methods perform the best for the capped case.

Due to the underestimation issue, we also computed the ranks for trimmed data, where the trimming is done on the original data by a specified percentage from both sides. The trimming percentages we chose were 5%, 10%, 15% and 20%. A summary of the results for these methods by trimming the

Table 1. Rank Distances to the Target (239) under Experiment 1 (without cap)

Pilot Sample Size	30	80	120	150	200
Original Method	3.0	5.6	5.1	5.1	5.1
Correction Method	4.0	5.0	5.6	5.6	5.6
Min-Diff Method	2.6	4.6	4.7	4.6	4.1
Bootstrap Mean	5.7	3.8	3.6	3.1	3.5
Bootstrap Median	6.9	5.2	5.0	4.4	3.2
Bootstrap 75%	4.0	2.4	$\boxed{1.8}$	$\boxed{1.9}$	$\boxed{2.5}$
Bootstrap 80%	$\boxed{1.6}$	$\boxed{1.4}$	2.2	3.4	4.1

Note: Among non-bootstrap procedures, minimum difference method is better. Among all seven methods, bootstrap 80% is the best for small sample and bootstrap 75% is the best for large sample.

Table 2. Rank Distances to the Target (239) under Experiment 1 (with cap)

Pilot Sample Size	30	80	120	150	200
Original Method	3.6	4.5	5.6	5.1	4.0
Correction Method	3.0	5.0	5.1	5.6	5.0
Min-Diff Method	2.7	4.1	4.7	4.6	$\boxed{3.0}$
Bootstrap Mean	5.7	3.8	3.6	3.5	3.5
Bootstrap Median	6.9	6.7	5.0	3.8	3.1
Bootstrap 75%	4.5	2.4	$\boxed{1.8}$	$\boxed{1.9}$	4.2
Bootstrap 80%	$\boxed{1.6}$	$\boxed{1.4}$	2.2	3.4	5.3

Table 3. Rank Distances to the Target (239) under Experiment 1 (trimming 15% and without cap)

Pilot Sample Size	30	80	120	150	200
Original Method	$\boxed{1.8}$	5.0	4.4	4.9	4.8
Correction Method	2.4	3.8	5.0	4.4	4.3
Min-Diff Method	3.2	4.7	4.2	4.1	3.6
Bootstrap Mean	6.0	4.2	4.0	3.9	3.3
Bootstrap Median	7.0	6.9	6.6	6.5	6.4
Bootstrap 75%	4.5	2.3	2.1	$\boxed{1.8}$	$\boxed{2.0}$
Bootstrap 80%	3.1	$\boxed{1.2}$	$\boxed{1.7}$	2.4	3.6

Table 4. Rank Distances to the Target (104) under Experiment 2 (without cap)

Pilot Sample Size	20	30	50	80	100
Original Method	4.6	4.9	5.6	5.1	5.1
Correction Method	3.6	4.3	5.0	5.5	4.5
Min-Diff Method	3.2	3.9	4.6	4.6	4.0
Bootstrap Mean	5.2	4.0	3.8	3.6	3.6
Bootstrap Median	6.9	6.8	5.1	3.8	3.2
Bootstrap 75%	2.9	2.6	$\boxed{1.9}$	$\boxed{1.9}$	$\boxed{2.5}$
Bootstrap 80%	$\boxed{1.6}$	$\boxed{1.4}$	2.1	3.4	5.1

Table 5. Rank Distances to the Target (45) under Experiment 3 (without cap)

Pilot Sample Size	10	20	30	50	80
Original Method	4.1	5.6	5.6	5.0	4.5
Correction Method	2.9	4.8	4.9	4.4	3.4
Min-Diff Method	3.6	4.6	4.6	4.0	4.0
Bootstrap Mean	5.8	3.9	3.2	$\boxed{3.0}$	$\boxed{3.0}$
Bootstrap Median	6.9	5.1	4.4	3.1	4.8
Bootstrap 75%	3.0	2.5	$\boxed{1.9}$	3.1	4.8
Bootstrap 80%	$\boxed{1.7}$	$\boxed{1.6}$	3.4	5.2	5.4

data by 15% in total for uncapped case is given in Table 3 on page 227. For the trimmed data, the original method gives better performance than others for small pilot sample size (30) while the bootstrap 80% performs the best for moderate pilot sample sizes (80 and 120). For fairly large pilot sample sizes (150 and 200), the bootstrap 75% performs better than the other six methods.

6.2 Experiment 2 (medium differences)

In this experiment, the proportions for the two multinomial distributions are set to be $\vec{p}_1 = (0.10, 0.25, 0.30, 0.20, 0.15)$ and $\vec{p}_2 = (0.17, 0.32, 0.36, 0.10, 0.05)$, and the true sample size calculated from (3.3) is 104. The side-by-side boxplots analogous to those in Figure 1 (without cap 900) are shown in Figure 3. The results in Figure 3 lead to conclusions similar to those in Section 6.1 for Experiment 1.

The summarized ranks without the sample size cap are in Table 4 on page 227. These results, and those when there is a sample size cap, are similar to those seen in Experiment 1. The results for the trimmed data show a pattern similar to what we have seen in Experiment 1.

6.3 Experiment 3 (large differences)

In Experiment 3, we consider the case in which the differences between the parameters from the two populations are moderately large, i.e., $\vec{p}_1 = (0.10, 0.25, 0.30, 0.20, 0.15)$ and $\vec{p}_2 = (0.30, 0.10, 0.20, 0.10, 0.30)$. The true required sample size calculated from (3.3) is 45. The side-by-side boxplots analogous to those in Figure 1 (without cap 900) are shown in Figure 4. In this case, i.e., large differences, there are little differences among the original, correction, and minimum difference methods while the bootstrap methods give smaller means and much smaller standard deviations than the other three methods. The summarized ranks are in Table 5 on page 227. Similar to Experiments 1 and 2 for small pilot sample sizes (10 and 20) the bootstrap 80% performs the best while for moderate pilot sample sizes (30 and 40) the bootstrap 75% is preferable. For the large pilot sample size, especially when the pilot sample size is greater than the target, the correction using the bootstrap 75% or 80% methods is no longer necessary and the bootstrap mean has the best performance, though it's not too different from the original and correction method.

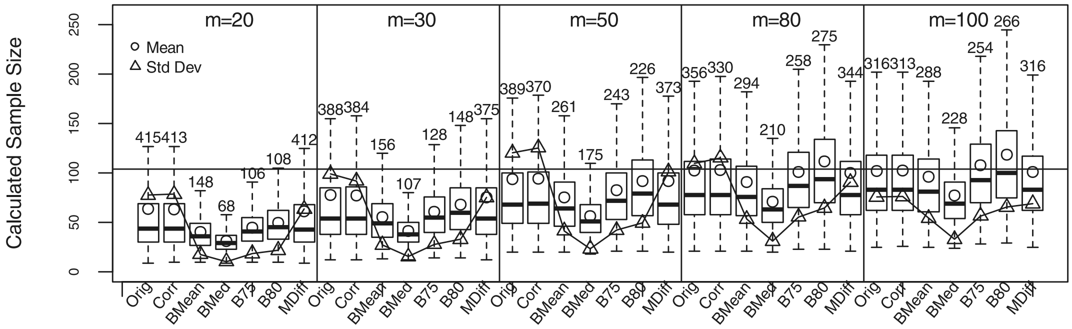

Figure 3. Side-by-side modified Box plot for Experiment 2.

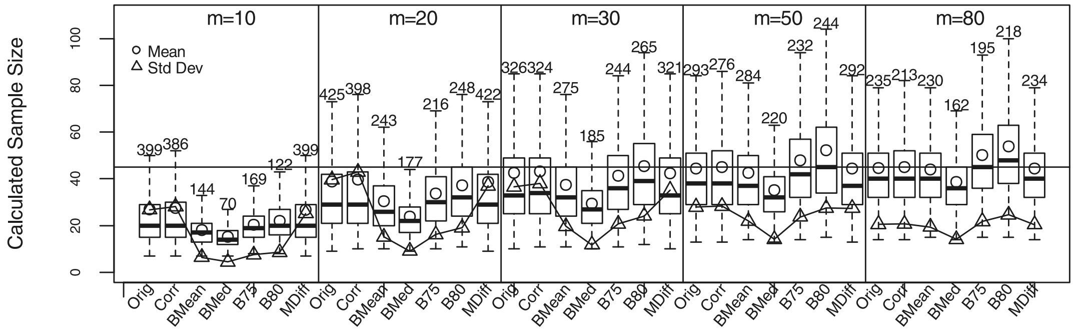

Figure 4. Side-by-side modified Box plot for Experiment 3.

Based on the simulation results, our **recommendation for practice** is provided as follows:

1. If a pilot sample is not available, use methods M1 and M3.
2. If a pilot sample of size $m \geq 10$ is available, use the combined approach below:
    2.1 Compute a cap using M1 with reasonable r and d.
    2.2 If a reliable minimum difference, c, is available, minimum difference method should be used.
    2.3 If c is not available, compute the estimates using the other six methods. If all the estimates are large, remove some categories as suggested in M3 or conduct an additional pilot study.
    2.4 For small and moderate pilot sample sizes, use the estimate from the bootstrap 80% method. For a fairly large pilot sample size, use the estimate from the bootstrap 75% method. For a very large pilot size, use the estimate from the bootstrap mean.

Here we require that a useful pilot sample have a sample size at least 10.

6.4 Experiment 4

In this experiment, we use the same parameter settings as those in Experiments 1 and 2. First we generate 50 pilot data sets. For each of these pilot data sets, we use the original frequentist method to calculate the estimated sample size n_0 using (3.6). Using three different starting values, n_0, $n_0 - 15$, and $n_0 + 15$, we apply our Bayesian method to each of the pilot data sets. Then we compare the performance of the Bayesian method with these three different starting values. Figures 5 and 6 present the simulation results for pilot sample sizes 30 and 80 from Experiment 1. Figures 7 and 8 present the simulation results for pilot sample sizes 20 and 30 from Experiment 2. Each Figure is a scatter plot of indices 1 to 50 versus the 50 sorted n_0, superimposed with the three matching Bayesian estimates at each index.

Overall, these four estimates are very close to each other, indicating some robustness in the Bayesian method. Taking a microscopic look, the original method and Bayesian method with starting value n_0 produce the closest estimates. On the other hand, the calculated sample sizes for the Bayesian method using starting value $n_0 - 15$ ($n_0 + 15$) are usually less (greater) than the calculated sample sizes from the original method. Therefore, it's possible that the Bayesian method starting from $n_0 + 15$ produces more estimates that are close to the target than the original n_0 would, even though the differences are small (see the summary in Table 6 on page 229).

In Table 6 on page 229, we count the number of times that $|\hat{n}_B - n_T| > |\hat{n}_O - n_T|$ for each data set, where $\hat{n}_B$ is the estimated sample size from the Bayesian method, $\hat{n}_O$

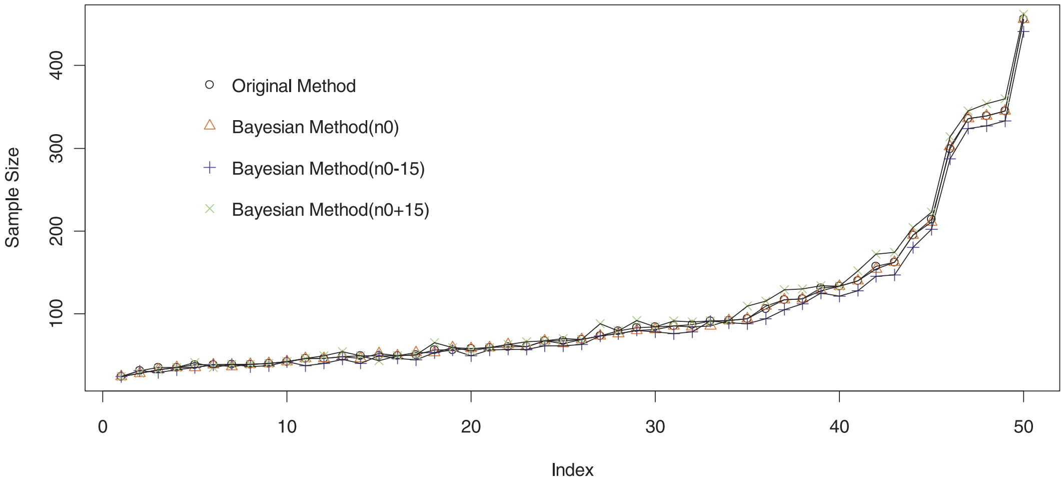

Figure 5. Plot of estimated sample sizes based on pilot sample size 30 for Experiment 1.

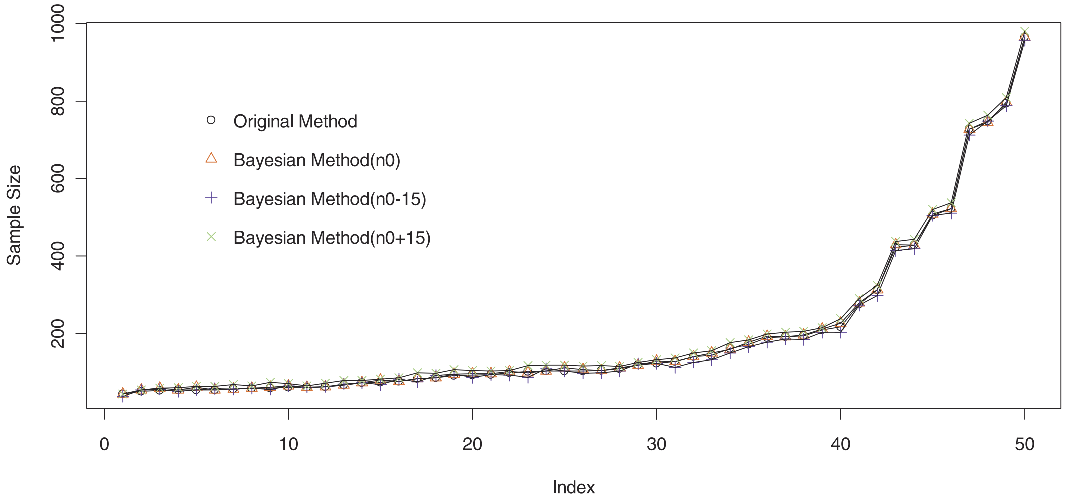

Figure 6. Plot of estimated sample sizes based on pilot sample size 80 for Experiment 1.

Table 6. Percentage that cppp is inferior to the original method

Starting Value	Figure 5	Figure 6	Figure 7	Figure 8
$n_0 - 15$	73%	42%	70%	30%
n_0	32%	12%	40%	20%
$n_0 + 15$	20%	20%	56%	40%

is the estimated sample size from the original method, and n_T is the target sample size. The results are summarized in Table 6 on page 229.

Based on the results in Table 6 on page 229 and Figures 5–8, the Bayesian method with starting value n_0 usually gives better results than the original method, although it is computationally intensive. Note also that with all starting values the performance of the Bayesian method improves with larger pilot sample sizes.

7. APPLICATION TO LEUKOPLAKIA LESIONS

Leukoplakia is associated with several factors such as poor diet, poor oral hygiene, local irritants, alcohol, and tobacco [20]. According to [19], the locations of oral leukoplakia are significantly correlated with the frequency of finding dysplastic or malignant changes at biopsy. The locations of leukoplakia lesions are closely related to different smoking habits. Here, we want to compare the distribution of lesion locations for two different types of smoking, i.e., Bidi smoking and non-Bidi smoking. [20] presented a data set giving 10 locations of lesions for 363 Bidi smoking individuals and for 142 non-Bidi smoking individuals. We test our SSD methods by using them as the pilot data. Instead of examining the potential differences from any of the 10 locations in the original data, we calculate the sample sizes for Bidi smoking

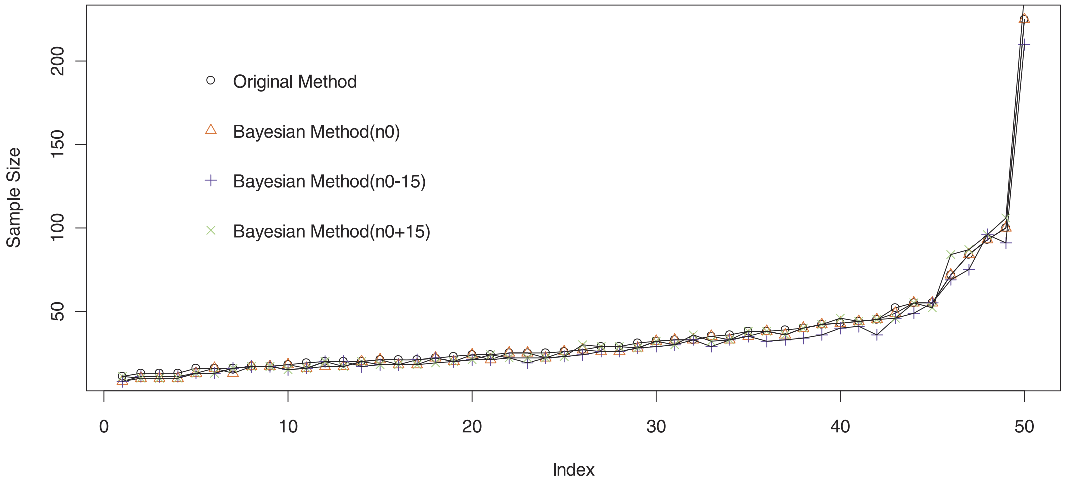

Figure 7. Plot of estimated sample sizes based on pilot sample size 20 for Experiment 2.

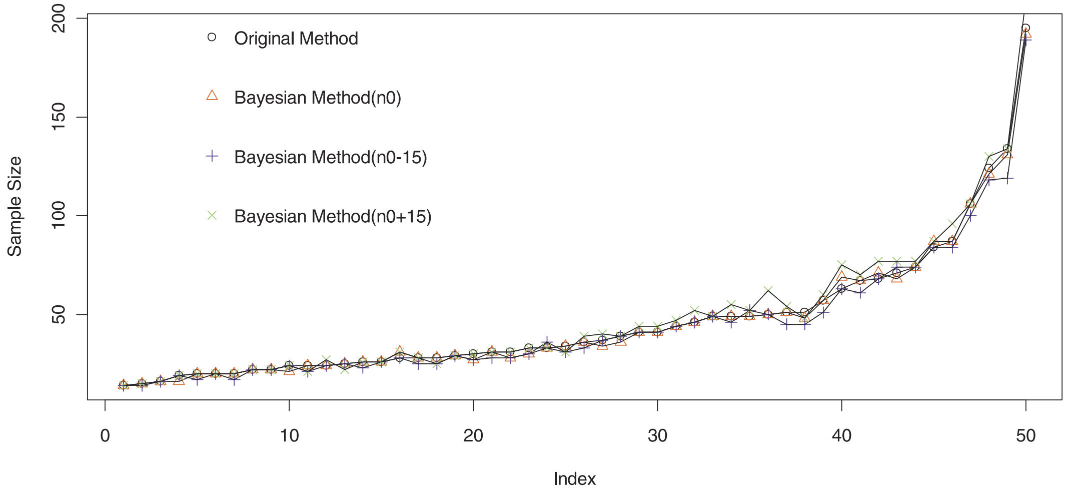

Figure 8. Plot of estimated sample sizes based on pilot sample size 30 for Experiment 2.

Table 7. Leukoplakia lesion locations regarding to Bidi and non-Bidi smoking habits

Location of lesion	Bidi smoking	non-Bidi smoking
LABIAL COMMISSURE		
right	101	24
left	88	25
BUCCAL MUCOSA		
right	70	31
left	70	35
Total	329	115

and non-Bidi smoking in any of its first four lesion locations (see the data Table 7 on page 230). These four locations are the most common leukoplakia lesion locations. The justification for this analysis is that one may be interested in investigating the sample sizes for the major lesion locations instead of all ten locations. Unlike the previous simulation study, the pilot sample sizes in this application are unequal. So, using the pilot sample sizes as our guide, we set the ratio of the two required sample sizes equal to the ratio of the two pilot sample sizes. In this application, the ratio of the two pilot sample sizes, r, is $329/115 = 2.86$. Letting the required sample size for Bidi smoking be n_1, the required sample size for non-Bidi smoking is $n_2 = n_1/2.86$. Then the required sample size for Bidi smoking, n_1, to detect the difference for the two habits can easily be calculated using (3.4) for the unequal sample size situation. To use (3.4), we take $k = 4$ and λ_0 to be the minimum value of λ such that (3.2) holds given α and $(1 - \beta)$ for $\alpha = 0.05$ and $\beta = 0.20$.

The corresponding frequentist improvements for the unbalanced case are similar to those for the balanced case except for the correction method which can not be modified directly. The calculated sample sizes for Bidi smoking and

Table 8. Sample sizes for leukoplakia lesion

Method	Orig.	B-Mean	B-Median	B-75%	B-80%
n_1	454	468	414	536	575
n_2	159	164	145	188	201

non-Bidi smoking, provided in Table 8 on page 231, show similar results for the original and bootstrap mean methods while the estimated sample sizes using the bootstrap median method are slightly smaller than these two methods. The bootstrap 75% and 80% methods provide larger estimated sample sizes than the other three methods.

If we choose $c = 0.02$ for the minimum difference method we get the same results as the original method, i.e., $n_1 = 454$ and $n_2 = 159$. However, this does require that c be specified. In addition, there is always a trade-off between total sample size and the resulting power. If we take $c = 0.02$ and the actual smallest difference is at least as large as 0.02, the power based on the sample sizes of 454 and 159 would be at least 80%.

If each individual patient has only one leukoplakia lesion or his/her primary lesion is of concern, the multinomial models would be reasonable for these data and the sample sizes in Table 8 on page 231 would be the required sample sizes. If there is more than one lesion for some patients (which is often true, but the number of overlaps is unknown a priori) and the primary lesion can not be identified, the sample sizes calculated from our procedures will serve as an upper bound for the true required sample sizes as there will be more occurrences of each lesion (even though the occurrences are likely to be positively correlated). And this upper bound will provide useful information for choosing the sample sizes.

8. DISCUSSION AND CONCLUSION

In this paper, we have systematically studied and developed both frequentist and Bayesian approaches to the calculation of the sample sizes needed to contrast two multinomial populations with and without a pilot/proxy sample. The practical implementation methods M1 and M3 should be used if applicable. The original approach, M2, is based on asymptotic theory while the Bayesian approach is based on computationally intensive simulation. The Bayesian approach starting from the original estimate (or M2) has some advantages over the original one. We also have studied several methods to improve the M2 approach, one using the bootstrap mean or median, one using bootstrap 75 or 80 quantile, one using an ad hoc bias correction, and another specifying the minimum difference between the parameters that the investigator wishes to detect. We have found that both the M2 method and the correction method provide similar choices for the sample sizes while the bootstrap 75% and 80% methods may provide better performance than the others when the pilot sample size is not too large. If the pilot sample size is relatively large, the corrections using the bootstrap 75% and 80% methods are not necessary. In this case, the bootstrap mean method should be used. We also tried bias-corrected and accelerated (BCa) bootstrap methods. While the BCa bootstrap performs better in terms of bias correction than the other methods, it has the largest standard deviations. In the main body of the estimates, the bootstrap 75% and 80% methods are the winners. Therefore, we suggest a combined practical approach at the end of Section 6 with the Bayesian approach added to the bag of choices if the computing power is not an issue. In addition, the correction method may be improved by using a Cornish-Fisher expansion, which is not covered by this paper.

Our methods are developed and studied based on a nonparametric approach to unordered categorical data. Alternatively, parametric solutions that model p_i's as a function of i depending on a finite parameter β, e.g., are possible especially for ordered categories. For ordered categorical data, two representative references are [28] and [21] where a parametric approach is used. Finally, note that although there are other sample size calculations based on similar chi-squared tests, the objectives are different from ours, see, e.g., [8, 12, 13, 22] and [3].

APPENDIX

Proof of Proposition 3.6:

Proof. Let $\hat{\varepsilon}_0 = |\frac{\hat{n}}{n} - 1|$ and

$$c = \frac{1}{2}\sum_{j=1}^{k}\left\{\frac{(\delta_{1j} - \delta_{2j})^2}{p_j}\right\}.$$

Conditioning on the pilot sample, for each fixed numerical value of $\hat{n}$ (i.e. not random value), the power can be evaluated as following:

(A.1)

$$\hat{P}_{H_1}\left(X^2 \geq \chi^2_{k-1}(\alpha)\right) = 1 - \sum_{j=0}^{\infty} e^{-\frac{\hat{\lambda}}{2}} \cdot \frac{(\frac{\hat{\lambda}}{2})^j}{j!} \cdot \frac{\gamma(j + \frac{k-1}{2}, \frac{\chi^2_{k-1}(\alpha)}{2})}{\Gamma(j + \frac{k-1}{2})},$$

where $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$ is the gamma function, $\gamma(a,x) = \int_0^x t^{a-1} \cdot e^{-t}dt$, which is lower incomplete gamma function and

$$\hat{\lambda} = \frac{\hat{n}}{2}\sum_{j=1}^{k}\left\{\frac{(\delta_{1j} - \delta_{2j})^2}{p_j}\right\}.$$

If $\hat{n}$ equals to the true sample size in (3.3), then the power from (A.1) is exactly $(1 - \beta)$ by the construction in (3.2). Next, we will see the effect of the difference between $\hat{n}$ and n, and hence the difference between $\hat{\lambda}$ and λ.

Under H_1, the power is

$$\begin{aligned}
&\hat{P}_{H_1}\big(X^2 \ge \chi^2_{k-1}(\alpha)\big)\\
&\quad = 1 - \sum_{j=0}^{\infty} e^{-\frac{\hat{\lambda}}{2}} \cdot \frac{(\frac{\hat{\lambda}}{2})^j}{j!} \cdot \frac{\gamma(j+\frac{k-1}{2}, \frac{\chi^2_{k-1}(\alpha)}{2})}{\Gamma(j+\frac{k-1}{2})}\\
&\quad = 1 - \sum_{j=0}^{\infty} e^{-\frac{\lambda}{2}} \cdot \frac{(\frac{\lambda}{2})^j}{j!} \cdot \frac{\gamma(j+\frac{k-1}{2}, \frac{\chi^2_{k-1}(\alpha)}{2})}{\Gamma(j+\frac{k-1}{2})} + C\\
&\quad = (1-\beta) + C,
\end{aligned}$$

where

$$C = \sum_{j=0}^{\infty}\Big[e^{-\frac{\lambda}{2}} \cdot \frac{(\frac{\lambda}{2})^j}{j!} - e^{-\frac{\hat{\lambda}}{2}} \cdot \frac{(\frac{\hat{\lambda}}{2})^j}{j!}\Big] \cdot \frac{\gamma(j+\frac{k-1}{2}, \frac{\chi^2_{k-1}(\alpha)}{2})}{\Gamma(j+\frac{k-1}{2})}.$$

Let

$$g(j) = \frac{\gamma(j+\frac{k-1}{2}, \frac{\chi^2_{(k-1)}(\alpha)}{2})}{\Gamma(j+\frac{k-1}{2})}.$$

Notice $g(j) \le 1$. Then

$$\begin{aligned}
\text{(A.2)}\qquad C = &\sum_{j=0}^{M}\Big[e^{-\frac{\lambda}{2}} \cdot \frac{(\frac{\lambda}{2})^j}{j!} - e^{-\frac{\hat{\lambda}}{2}} \cdot \frac{(\frac{\hat{\lambda}}{2})^j}{j!}\Big] \cdot g(i)\\
&+ \sum_{j=M+1}^{\infty}\Big[e^{-\frac{\lambda}{2}} \cdot \frac{(\frac{\lambda}{2})^j}{j!} - e^{-\frac{\hat{\lambda}}{2}} \cdot \frac{(\frac{\hat{\lambda}}{2})^j}{j!}\Big] \cdot g(i)
\end{aligned}$$

For any given $\delta > 0$, we want to find M such that the second part of (A.2) satisfies

$$\text{(A.3)}\qquad \Big|\sum_{j=M+1}^{\infty}\Big[e^{-\frac{\lambda}{2}} \cdot \frac{(\frac{\lambda}{2})^j}{j!} - e^{-\frac{\hat{\lambda}}{2}} \cdot \frac{(\frac{\hat{\lambda}}{2})^j}{j!}\Big] \cdot g(i)\Big| \le \frac{\delta}{4}.$$

Since

$$\begin{aligned}
&\Big|\sum_{j=M+1}^{\infty}\Big[e^{-\frac{\lambda}{2}} \cdot \frac{(\frac{\lambda}{2})^j}{j!} - e^{-\frac{\hat{\lambda}}{2}} \cdot \frac{(\frac{\hat{\lambda}}{2})^j}{j!}\Big] \cdot g(i)\Big|\\
&\le \sum_{j=M+1}^{\infty}\Big[e^{-\frac{\lambda}{2}} \cdot \frac{(\frac{\lambda}{2})^j}{j!} + e^{-\frac{\hat{\lambda}}{2}} \cdot \frac{(\frac{\hat{\lambda}}{2})^j}{j!}\Big]\\
&= \sum_{j=M+1}^{\infty}\Big[p\Big(j;\frac{\lambda}{2}\Big) + p\Big(j;\frac{\hat{\lambda}}{2}\Big)\Big]\\
&= \Big[1 - \frac{\Gamma(M+1,\frac{\lambda}{2})}{M!}\Big] + \Big[1 - \frac{\Gamma(M+1,\frac{\hat{\lambda}}{2})}{M!}\Big]\\
&= \int_0^{\frac{\lambda}{2}} \frac{x^{M-1}e^{-x}}{(M-1)!} \cdot \frac{x}{M}dx + \int_0^{\frac{\hat{\lambda}}{2}} \frac{x^{M-1}e^{-x}}{(M-1)!} \cdot \frac{x}{M}dx\\
&\le \frac{\lambda}{2M}\int_0^{\frac{\lambda}{2}} \frac{x^{M-1}e^{-x}}{\Gamma(M)}dx + \frac{\hat{\lambda}}{2M}\int_0^{\frac{\hat{\lambda}}{2}} \frac{x^{M-1}e^{-x}}{\Gamma(M)}dx\\
&\le \frac{\lambda}{2M} + \frac{\hat{\lambda}}{2M},
\end{aligned}$$

where the $p(j;\lambda)$ is the probability distribution function for Poisson distribution with parameter λ. Hence, let $M = 2(\lambda + \hat{\lambda})/\delta$, then inequality (A.3) holds. For the first part of (A.2), if M is large, we have

$$\begin{aligned}
&\sum_{j=0}^{M}\Big[e^{-\frac{\lambda}{2}} \cdot \frac{(\frac{\lambda}{2})^j}{j!} - e^{-\frac{\hat{\lambda}}{2}} \cdot \frac{(\frac{\hat{\lambda}}{2})^j}{j!}\Big] \cdot g(i)\\
&\le \Big|\sum_{j=0}^{M}\Big[e^{-\frac{\lambda}{2}} \cdot \frac{(\frac{\lambda}{2})^j}{j!} + e^{-\frac{\hat{\lambda}}{2}} \cdot \frac{(\frac{\hat{\lambda}}{2})^j}{j!}\Big]\Big|\\
&= \sum_{j=0}^{M}\Big[e^{-\frac{\lambda}{2}} \cdot \frac{(\frac{\lambda}{2})^j}{j!} + e^{-\frac{\hat{\lambda}}{2}} \cdot \frac{(\frac{\hat{\lambda}}{2})^j}{j!}\Big]\\
&= \sum_{j=0}^{M}\Big[e^{-\frac{\lambda}{2}} \cdot \frac{(\frac{\lambda}{2})^j}{j!}\Big] + \sum_{j=0}^{M}\Big[e^{-\frac{\hat{\lambda}}{2}} \cdot \frac{(\frac{\hat{\lambda}}{2})^j}{j!}\Big]\\
&= \frac{\Gamma(M+1,\frac{\lambda}{2})}{M!} + \frac{\Gamma(M+1,\frac{\hat{\lambda}}{2})}{M!}\\
&= \int_{\frac{\lambda}{2}}^{+\infty} \frac{t^{M+1}\cdot e^{-t}}{\Gamma(M+2)} \cdot \frac{M+1}{t}dt\\
&\quad + \int_{\frac{\hat{\lambda}}{2}}^{+\infty} \frac{t^{M+1}\cdot e^{-t}}{\Gamma(M+2)} \cdot \frac{M+1}{t}dt\\
&\le \frac{2(M+1)}{\lambda} + \frac{2(M+1)}{\hat{\lambda}}\\
&= 2\Big(\frac{1}{\lambda} + \frac{1}{\hat{\lambda}}\Big)\Big(\frac{2}{\delta}(\lambda + \hat{\lambda}) + 1\Big)
\end{aligned}$$

From above calculation, we can see

$$\begin{aligned}
&\sum_{j=0}^{\infty}\big[1 - e^{-\frac{\hat{\varepsilon}_0\lambda}{2}} \cdot (1+\hat{\varepsilon}_0)^j\big] \cdot \Big(e^{-\frac{nc}{2}} \cdot \frac{(\frac{nc}{2})^j}{j!} \cdot g(j)\Big)\\
&\le \frac{\delta}{4} + 2\Big(\frac{1}{\lambda} + \frac{1}{\hat{\lambda}}\Big)\Big(\frac{2}{\delta}(\lambda + \hat{\lambda}) + 1\Big).
\end{aligned}$$

Since the Taylor expansion of $e^{-\frac{\hat{\varepsilon}_0\lambda}{2}}$ at 0 is

$$e^{-\frac{\hat{\varepsilon}_0\lambda}{2}} = 1 - \frac{\hat{\varepsilon}_0\lambda}{2} + \frac{(\hat{\varepsilon}_0\lambda)^2}{8} + O\big(\hat{\varepsilon}_0^3\big).$$

Then

$$\begin{aligned}
&\sum_{j=0}^{\infty}\big[1 - e^{-\frac{\hat{\varepsilon}_0\lambda}{2}} \cdot (1+\hat{\varepsilon}_0)^j\big]\\
&\qquad \cdot \Big(e^{-\frac{nc}{2}} \cdot \frac{(\frac{nc}{2})^j}{j!} \cdot \frac{\gamma(j+\frac{k-1}{2}, \frac{\chi^2_{(k-1)}(\alpha)}{2})}{\Gamma(j+\frac{k-1}{2})}\Big)\\
&= O(\hat{\varepsilon}_0) \cdot \Big(e^{-\frac{nc}{2}} \cdot \frac{(\frac{nc}{2})^j}{j!} \cdot \frac{\gamma(j+\frac{k-1}{2}, \frac{\chi^2_{(k-1)}(\alpha)}{2})}{\Gamma(j+\frac{k-1}{2})}\Big)\\
&= O(\hat{\varepsilon}_0) \cdot \beta\\
&= O(\hat{\varepsilon}_0).
\end{aligned}$$

Now we get $Power(\hat{n}) = 1 - \beta + O(\hat{\varepsilon}_0)$. □

Received 10 December 2013

REFERENCES

[1] Adcock, C. J. (1993). An improved Bayesian procedure for calculating sample sizes in multinomial sampling. *The Statistician* **42** 91–95.

[2] Adcock, C. J. (1997). Sample size determination: A review. *The Statistician* **46** 261–283.

[3] Agresti, A. (1990). *Categorical Data Analysis*. Wiley. MR1044993

[4] Cabras, S., Castellanos, M. E. and Quirós, A. (2011). Goodness-of-fit of conditional regression models for multiple imputation. *Bayesian Anal.* 429–455. MR2843539

[5] Chow, S.-C., Shao, J. and Wang, H. (2007). *Sample Size Calculation in Clinical Research*, Second ed. CRC Press. MR2356591

[6] Desu, M. M. and Raghavarao, D. (1990). *Sample Size Methodology*. Academic Press. MR1162309

[7] Dillman, R. O., Davis, R. B., Green, M. R., Weiss, R. B., Gottlieb, A. J., Caplan, S., Kopel, S., Preisler, H., Mclntyre, O. R. and Schiffer, C. (1991). A comparative study of two different doses of cytarabine for acute myeloid leukemia: A phase I11 trial of cancer and leukemia group B. *Blood* **78** 2520–2526.

[8] Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*, Second ed. Wiley-Interscience. MR0622544

[9] Gelman, A., Meng, X. L. and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica* **6** 733–807. MR1422404

[10] Gelman, A., Carlin, J., Stern, H. and Rubin, D. (2004). *Bayesian Data Analysis*, Second ed. Chapman and Hall. MR2027492

[11] Gillett, R. (1996). Sample size determination in a chi-squared test given information from an earlier study. *Journal of Educational and Behavioral Statistics* **21** 230–246.

[12] Greenland, S. (1983). Tests of interaction in epidemiological studies: A review and study of power. *Statistics in Medicine* **2** 243–251.

[13] Haber, M. (1983). Sample size for the exact test of "no interaction" in a 2 × 2 table. *Biometrics* **39** 493–498. MR0714420

[14] Hjort, N. L., Dahl, F. A. and Steinbakk, G. H. (2006). Post-processing posterior predictive p values. *Journal of the American Statistical Association* **101** 1157–1174. MR2324154

[15] Krstevska, V. and Crvenkova, S. (2006). Altered and conventional fractionated radiotherapy in locoregional control and survival of patients with squamous cell carcinoma of the larynx, oropharynx, and hypopharynx. *Croat. Med. J.* **47** 42–52.

[16] Larsen, M. D. and Lu, L. (2007). Comment: Bayesian checking of the second level of hierarchical models: Cross-validated posterior predictive checks using discrepancy measures. *Statistical Science* **22** 359–362. MR2416812

[17] Meng, X.-L. (1994). Posterior predictive p-values. *Annals of Statistics* **22** 1142–1160. MR1311969

[18] Meng, R. and Chapman, D. (1966). The power of chi square tests for contingency tables. *Journal of the American Statistical Association* **61** 965–975. MR0207109

[19] Neville, B. and Day, T. (2002). Oral cancer and precancerous lesions. *CA Cancer J. Clin.* **52** 195–215.

[20] Pindborg, J. J., Kiaer, J., Gupta, P. C. and Chawla, T. N. (1967). Studies in oral leukoplakias. Prevalence of leukoplakia among 10000 persons in Lucknow, India, with special reference to use of tobacco and betel nut. *Bull. World Health Organ.* **37** 109–116.

[21] Rabbee, N., Coull, B. A., Mehta, C., Patel, N. and Senchaudhuri, P. (2003). Power and sample size for ordered categorical data. *Stat. Methods Med. Res.* **12** 73–84. MR1977236

[22] Rochon, J. (1989). The application of the GSK method to the determination of minimum sample size. *Biometrics* **45** 193–205.

[23] Sison, C. and Glaz, J. (1995). Simultaneous confidence intervals and sample size determination for multinomial proportions. *Journal of the American Statistical Association* **429** 366–369. MR1325142

[24] Steinbakk, G. H. and Storvik, G. O. (2009). Posterior predictive p-values in Bayesian hierarchical models. *Scand. J. Stat.* 320–336. MR2528987

[25] Thompson, S. (1987). Sample size for estimating multinomial proportions. *The American Statistician* **41** 42–46. MR0882768

[26] Tortora, R. (1978). A note on sample size estimation for multinomial populations. *The American Statistician* **32** 100–102. MR0514334

[27] Wang, F. and Gelfand, A. E. (2002). A simulation-based approach to bayesian sample size determination for performance under a given model and for separating models. *Statistical Science* **17** 193–208. MR1925941

[28] Whitehead, J. (1993). Sample size calculations for ordered categorical data. *Statistics in Medicine* **12** 2257–2271.

Junheng Ma
Department of Epidemiology and Biostatistics, SR2c
Case Western Reserve University
10900 Euclid Avenue
Cleveland, OH 44106-4945
USA
E-mail address: jxm216@case.edu

Jiayang Sun
Department of Epidemiology and Biostatistics, SR2c
Case Western Reserve University
10900 Euclid Avenue
Cleveland, OH 44106-4945
USA
E-mail address: jsun@case.edu

Joe Sedransk
Department of Epidemiology and Biostatistics, SR2c
Case Western Reserve University
10900 Euclid Avenue
Cleveland, OH 44106-4945
USA

Joint Program in Survey Methodology
University of Maryland
1218 LeFrak Hall
College Park, MD 20742
USA
E-mail address: jxs123@case.edu