Journal of Statistical Theory and Practice (2021) 15:24
https://doi.org/10.1007/s42519-020-00163-y

ORIGINAL ARTICLE

Estimation of the Minimum Probability of a Multinomial Distribution

Ali Mahzarnia[1] · Michael Grabchak[1] · Jiancheng Jiang[1]

Accepted: 23 December 2020 / Published online: 20 January 2021

Abstract
The estimation of the minimum probability of a multinomial distribution is important for a variety of application areas. However, standard estimators such as the maximum likelihood estimator and the Laplace smoothing estimator fail to function reasonably in many situations as, for small sample sizes, these estimators are fully deterministic and completely ignore the data. Inspired by a smooth approximation of the minimum used in optimization theory, we introduce a new estimator, which takes advantage of the entire data set. We consider both the cases with a known and an unknown number of categories. We categorize the asymptotic distributions of the proposed estimator and conduct a small-scale simulation study to better understand its finite sample performance.

Keywords Minimum probability · Multinomial distribution · Smooth minimum

1 Introduction

Consider the multinomial distribution $\mathbf{P} = (p_1, p_2, \dots, p_k)$, where $k \geq 2$ is the number of categories and $p_i > 0$ is the probability of seeing an observation from category i. We are interested in estimating the minimum probability

$$p_0 = \min\{p_i; i = 1, \dots, k\}$$

in both the cases where k is known and where it is unknown.

Given an independent and identically distributed random sample $X_1, X_2, \dots, X_n$ of size n from $\mathbf{P}$, let $y_i = \sum_{j=1}^{n} 1(X_j = i)$ be the number of observations of category i. Here and throughout, we write $1(\cdot)$ to denote the indicator function. The maximum likelihood estimator (MLE) of p_i is $\hat{p}_i = y_i/n$ and the MLE of p_0 is

✉ Michael Grabchak
mgrabcha@uncc.edu

[1] Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC 28223, USA

$$\hat{p}_0 = \min\{\hat{p}_i; i = 1, \ldots, k\}.$$

The MLE has the obvious drawback that $\hat{p}_0$ is zero when we do not have at least one observation from each category. To deal with this issue, one generally uses a modification of the MLE. Perhaps the most prominent modification is the so-called Laplace smoothing estimator (LSE). This estimator was introduced by Pierre-Simon Laplace in the late 1700s to estimate the probability that the sun will rise tomorrow, see, e.g., [4]. The LSE of p_0 is given by

$$\hat{p}_0^{\mathrm{LS}} = \min\{(y_i + 1)/(n + k); i = 1, \cdots, k\}.$$

Note that both $\hat{p}_0$ and $\hat{p}_0^{\mathrm{LS}}$ are based only on the smallest y_i. Note further that, in situations where we have not seen all of the categories in the sample, we always have $\hat{p}_0 = 0$ and $\hat{p}_0^{\mathrm{LS}} = 1/(n + k)$. This holds, in particular, whenever $n < k$. Thus, in these cases, the estimators are fully deterministic and completely ignore the data.

In this article, we introduce a new estimator for p_0, which is based on a smooth approximation of the minimum. It uses information from all of the categories and thus avoids becoming deterministic for small sample sizes. We consider both the cases when the number of categories is known and when it is unknown. We show consistency of this estimator and characterize its asymptotic distributions. We also perform a small-scale simulation study to better understand its finite sample performance. Our numerical results show that, in certain situations, it outperforms both the MLE and the LSE.

The rest of the paper is organized as follows: In Sect.2, we introduce our estimator for the case where the number of categories k is known and derive its asymptotic distributions. Then, in Sect. 3 we consider the case where k is unknown, and in Sect. 4 we consider the related problem of estimating the maximum probability. In Sect. 5 we give our simulation results, and in Sect. 6 we give some conclusions and directions for future work. Finally, the proofs are given in "Appendix". Before proceeding, we briefly describe a few applications:

1. One often needs to estimate the probability of a category that is not observed in a random sample. This is often estimated using the LSE, which always gives the deterministic value of $1/(n + k)$. On the other hand, a data-driven estimate would be more reasonable. When the sample size is relatively large, it is reasonable to assume that the unobserved category has the smallest probability and our estimator could be used in this case. This situation comes up in a variety of applications including language processing, computer vision, and linguistics, see, e.g., [6, 14], or [15].
2. In the context of ecology, we may be interested in the probability of finding the rarest species in an ecosystem. Aside for the intrinsic interest in this question, this probability may be useful as a diversity index. In ecology, diversity indices are metrics used to measure and compare the diversity of species in different ecosystems, see, e.g., [7, 8], and the references therein. Generally one works with several indices at once as they give different information about the ecosystem. In particular, the probability of the rarest species may be especially useful when

combined with the index of species richness, which is the total number of species in the ecosystem.

3. Consider the problem of internet ad placement. There are generally multiple ads that are shown on the same webpage, and at most one of these will be clicked. Thus, if there are $k-1$ ads, then there are k possible outcomes, with the last outcome being that no ad is clicked. In this context, the probability of a click on a given ad is called the click through rate or CTR. Assume that there are $k-1$ ads that have been displayed together on the same page and that we have data on these. Now, the ad company wants to replace one of these with a new ad, for which there are no data. In this case, the minimum probability of the original $k-1$ ads may give a baseline for the CTR of the new ad. This may be useful for pricing.

2 The Estimator When *k* Is Known

We begin with the case where the number of categories k is known. Let $\mathbf{p} = (p_1, \ldots, p_{k-1})$, $\hat{\mathbf{p}} = (\hat{p}_1, \ldots, \hat{p}_{k-1})$, and note that $p_k = 1 - \sum_{i=1}^{k-1} p_i$ and $\hat{p}_k = 1 - \sum_{i=1}^{k-1} \hat{p}_i$. Since $p_0 = g(\mathbf{p})$, where $g(\mathbf{p}) = \min\{p_1, p_2, \ldots, p_k\}$, a natural estimator of p_0 is given by

$$\hat{p}_0 = g(\hat{\mathbf{p}}) = \min\{\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k\},$$

which is the MLE. However, this estimator takes the value of zero whenever there is a category that has not been observed. To deal with this issue, we propose approximating g with a smoother function. Such approximations, which are sometimes called smooth minimums, are often used in optimization theory, see, e.g., [1, 9, 10], or [11]. Specifically, we introduce the function

$$g_n(\mathbf{p}) = w^{-1} \sum_{i=1}^{k} p_i e^{-n^\alpha p_i}, \tag{1}$$

where $w = w(\mathbf{p}) = \sum_{j=1}^{k} e^{-n^\alpha p_j}$ and $\alpha > 0$ is a tuning parameter. Note that

$$\lim_{n\to\infty} g_n(\mathbf{p}) = g(\mathbf{p}) = p_0. \tag{2}$$

This leads to the estimator

$$\hat{p}_0^* = g_n(\hat{\mathbf{p}}) = \hat{w}^{-1} \sum_{i=1}^{k} \hat{p}_i e^{-n^\alpha \hat{p}_i}, \tag{3}$$

where $\hat{w} = w(\hat{\mathbf{p}}) = \sum_{j=1}^{k} e^{-n^\alpha \hat{p}_j}$.

We now study the asymptotic distributions of $\hat{p}_0^*$. Let $\nabla g_n(\mathbf{p}) = \left(\frac{\partial g_n(\mathbf{p})}{\partial p_1}, \ldots, \frac{\partial g_n(\mathbf{p})}{\partial p_{k-1}} \right)^T$. It is straightforward to check that, for $1 \le i \le k-1$,

$$\frac{\partial g_n(\mathbf{p})}{\partial p_i} = e^{-n^\alpha p_i} w^{-1}[1 + n^\alpha \big(g_n(\mathbf{p}) - p_i\big)] - e^{-n^\alpha p_k} w^{-1}[1 + n^\alpha \big(g_n(\mathbf{p}) - p_k\big)]. \quad (4)$$

Let r be the cardinality of the set $\{j : p_j = p_0,\ j = 1, \ldots, k\}$, i.e., r is the number of categories that attain the minimum probability. Note that $r \geq 1$ and that we have a uniform distribution if and only if $r = k$. With this notation, we give the following result.

Theorem 2.1 *Assume that* $0 < \alpha < 1/2$ *and let* $\hat{\sigma}_n = \{\nabla g_n(\hat{\mathbf{p}})^T \hat{\Sigma} \nabla g_n(\hat{\mathbf{p}})\}^{1/2}$, *where* $\hat{\Sigma} = \text{ diag } (\hat{\mathbf{p}}) - \hat{\mathbf{p}}\hat{\mathbf{p}}^T$.

(i) *If* $r \neq k$, *then*

$$\sqrt{n}\hat{\sigma}_n^{-1}\{\hat{p}_0^* - p_0\} \xrightarrow{D} \mathcal{N}(0, 1).$$

(ii) *If* $r = k$, *then*

$$k^2 n^{1-\alpha}\{p_0 - \hat{p}_0^*\} \xrightarrow{D} \chi^2_{(k-1)}.$$

Clearly, Theorem 2.1 both proves consistency and characterizes the asymptotic distributions. Further, it allows us to construct asymptotic confidence intervals for p_0. If $r \neq k$, then an approximate $100(1-\gamma)\%$ confidence interval is

$$\hat{p}_0^* \pm n^{-1/2}\hat{\sigma}_n z_{1-\gamma/2},$$

where $z_{1-\gamma/2}$ is the $100(1-\gamma/2)$th percentile of the standard normal distribution. If $r = k$, then the corresponding confidence interval is

$$[\hat{p}_0^*, \hat{p}_0^* + k^{-2} n^{\alpha-1} \chi^2_{k-1,1-\gamma}],$$

where $\chi^2_{k-1,1-\gamma}$ is the $100(1-\gamma)$th percentile of a Chi-squared distribution with $k-1$ degrees of freedom.

As far as we know, these are the first confidence interval for the minimum to appear in the literature. In fact, to the best of our knowledge, the asymptotic distributions of the MLE and the LSE have not been established. One might think that a version of Theorem 2.1 for the MLE could be proved using the asymptotic normality of $\hat{\mathbf{p}}$ and the delta method. However, the delta method cannot be applied since the minimum function g is not differentiable. Even in the case of the proposed estimator $\hat{p}_0^*$, where we use a smooth minimum, the delta method cannot be applied directly since the function g_n depends on the sample size n. Instead, a subtler approach is needed. The detailed proof is given in "Appendix".

3 The Estimator When *k* Is Unknown

In this section, we consider the situation where the number of categories k is unknown. In this case, one cannot evaluate the estimator $\hat{p}_0^*$. The difficulty lies in the need to evaluate $\hat{w}$. Let $\ell = \sum_{j=1}^{k} 1\left(y_j = 0\right)$ be the number of categories that are not observed in the sample and note that

$$\hat{w} = \sum_{j=1}^{k} e^{-n^{\alpha}\hat{p}_j} = \sum_{j=1}^{k} e^{-n^{\alpha}\hat{p}_j} 1\left(y_j > 0\right) + \ell.$$

If we have an estimator $\hat{\ell}$ of ℓ, then we can take

$$\hat{w}^{\sharp} = \sum_{j=1}^{k} e^{-n^{\alpha}\hat{p}_j} 1\left(y_j > 0\right) + \hat{\ell}$$

and define the estimator

$$\hat{p}_0^{\sharp} = \frac{1}{\hat{w}^{\sharp}} \sum_{i=1}^{k} \hat{p}_i e^{-n^{\alpha}\hat{p}_i}. \tag{5}$$

Note that $\hat{p}_0^{\sharp}$ can be evaluated without knowledge of k since $\hat{p}_i = 0$ for any category i that does not appear in the sample.

Now, assume that we have observed $k^{\sharp}$ categories in our sample and note that $k^{\sharp} \le k$. Without loss of generality, assume that these are categories $1, 2, \ldots, k^{\sharp}$. Assume that $k^{\sharp} \ge 2$, let $\hat{\mathbf{p}}^{\sharp} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_{k^{\sharp}-1})$, and note that $\hat{p}_{k^{\sharp}} = 1 - \sum_{i=1}^{k^{\sharp}-1} \hat{p}_i$. For $i = 1, 2, \ldots, (k^{\sharp} - 1)$ let

$$h_i = e^{-n^{\alpha}\hat{p}_i} \frac{1}{\hat{w}^{\sharp}} \left[1 - n^{\alpha}\{\hat{p}_i - \hat{p}_0^{\sharp}\}\right] - e^{-n^{\alpha}\hat{p}_{k^{\sharp}}} \frac{1}{\hat{w}^{\sharp}} [1 - n^{\alpha}\{\hat{p}_{k^{\sharp}} - \hat{p}_0^{\sharp}\}]$$

and let $\mathbf{h} = (h_1, h_2, \ldots, h_{k^{\sharp}-1})$. Note that we can evaluate $\mathbf{h}$ without knowing k.

Theorem 3.1 *Assume that $\hat{\ell}$ is such that, with probability* 1, *we eventually have $\hat{\ell} = 0$. When $k^{\sharp} \ge 2$, let $\hat{\sigma}_n^{\sharp} = \{\mathbf{h}^T \hat{\Sigma}^{\sharp} \mathbf{h}\}^{1/2}$, where $\Sigma^{\sharp} = \mathrm{diag}(\hat{\mathbf{p}}^{\sharp}) - \hat{\mathbf{p}}^{\sharp}\left(\hat{\mathbf{p}}^{\sharp}\right)^T$. When $k^{\sharp} = 1$, let $\hat{\sigma}_n^{\sharp} = 1$. If the assumptions of Theorem 2.1 hold, then the results of Theorem 2.1 hold with $\hat{p}_0^{\sharp}$ in place of $\hat{p}_0^*$ and $\hat{\sigma}_n^{\sharp}$ in place of $\hat{\sigma}_n$.*

Proof Since k is finite and we eventually have $\hat{\ell} = 0$, there exists an almost surely finite random variable N such that if the sample size $n \ge N$, then $\hat{\ell} = 0$, and we have observed each category at least once. For such n, we have $k^{\sharp} = k$, $\hat{w}^{\sharp} = \hat{w}$, $\hat{\mathbf{p}}^{\sharp} = \hat{\mathbf{p}}$, and $\nabla g_n(\hat{\mathbf{p}}) = \mathbf{h}$. If follows that, for such n, $\hat{\sigma}_n^{\sharp} = \hat{\sigma}_n$ and $\hat{p}_0^{\sharp} = \hat{p}_0$. Hence $\hat{\sigma}_n^{\sharp}/\hat{\sigma}_n \xrightarrow{p} 1$ and $\sqrt{n}\hat{\sigma}_n^{-1}\{\hat{p}_0^* - \hat{p}_0^{\sharp}\} \xrightarrow{p} 0$. From here the case $r \ne k$ follows by Theorem 2.1 and two applications of Slutsky's theorem. The case $r = k$ is similar and is thus omitted. □

There are a number of estimators for ℓ available in the literature, see, e.g., [2, 3, 5], or [16] and the references therein. One of the most popular is the so-called Chao2 estimator [3, 5], which is given by

$$\hat{\ell} = \begin{cases} \frac{n-1}{n}\frac{f_1^2}{2f_2} & \text{if } f_2 > 0 \\ \frac{n-1}{n}\frac{f_1(f_1-1)}{2} & \text{if } f_2 = 0, \end{cases} \tag{6}$$

where $f_i = \sum_{j=1}^{k} 1\left(y_j = i\right)$ is the number of categories that were observed exactly i times in the sample. Since k is finite, we will, with probability 1, eventually observe each category at least three times. Thus, we will eventually have $f_1 = f_2 = 0$ and $\hat{\ell} = 0$. Thus, this estimator satisfies the assumptions of Theorem 3.1. In the rest of the paper, when we use the notation $\hat{p}_0^{\sharp}$ we will mean the estimator where $\hat{\ell}$ is given by (6).

4 Estimation of the Maximum

The problem of estimating the maximum probability is generally easier than that of estimating the minimum. Nevertheless, it may be interesting to note that our methodology can be modified to estimate the maximum. Let

$$p_{\vee} = \max\{p_i : i = 1, \cdots, k\}.$$

We begin with the case where the number of categories k is known. We can approximate the maximum function with a smooth maximum given by

$$g_n^{\vee}(\mathbf{p}) = w_{\vee}^{-1} \sum_{i=1}^{k} p_i e^{n^{\alpha} p_i}, \tag{7}$$

where $w_{\vee} = w_{\vee}(\mathbf{p}) = \sum_{i=1}^{k} e^{n^{\alpha} p_i}$. Note that

$$g_n^{\vee}(\mathbf{p}) = -g_n(-\mathbf{p}),$$

where g_n is given by (1). It is not difficult to verify that $g_n^{\vee}(\mathbf{p}) \to p_{\vee}$ as $n \to \infty$. This suggests that we can estimate $p_{\vee}$ by

$$\hat{p}_{\vee}^{*} = g_n^{\vee}(\hat{\mathbf{p}}) = \hat{w}_{\vee}^{-1} \sum_{i=1}^{k} \hat{p}_i e^{n^{\alpha} \hat{p}_i}, \tag{8}$$

where $\hat{w}_{\vee} = w_{\vee}(\hat{\mathbf{p}}) = \sum_{i=1}^{k} e^{n^{\alpha} \hat{p}_i}$.

Let $r_{\vee}$ be the cardinality of the set $\{j : p_j = p_{\vee}, j = 1, \ldots, k\}$ and let $\nabla g_n^{\vee}(\mathbf{p}) = \left(\frac{\partial g_n^{\vee}(\mathbf{p})}{\partial p_1}, \ldots, \frac{\partial g_n^{\vee}(\mathbf{p})}{\partial p_{k-1}}\right)^T$. It is easily verified that, for $1 \le i \le k-1$,

$$\frac{\partial g_n^{\vee}(\mathbf{p})}{\partial p_i} = \frac{\partial g_n(-\mathbf{p})}{\partial p_i} = e^{n^{\alpha} p_i} w_{\vee}^{-1}[1 + n^{\alpha}\{p_i - g_n^{\vee}(\mathbf{p})\}] - e^{n^{\alpha} p_k} w_{\vee}^{-1}[1 + n^{\alpha}\{p_k - g_n^{\vee}(\mathbf{p})\}]. \tag{9}$$

We now characterize the asymptotic distributions of $\hat{p}_{\vee}$.

Theorem 4.1 *Assume that* $0 < \alpha < 1/2$ *and let* $\hat{\sigma}_n^{\vee} = \{\nabla g_n^{\vee}(\hat{\mathbf{p}})^T \hat{\Sigma} \nabla g_n^{\vee}(\hat{\mathbf{p}})\}^{1/2}$, *where* $\hat{\Sigma} = \text{diag}\,(\hat{\mathbf{p}}) - \hat{\mathbf{p}}\hat{\mathbf{p}}^T$.

(i) *If* $r_{\vee} \neq k$, *then*

$$\frac{\sqrt{n}}{\hat{\sigma}_n^{\vee}}\{\hat{p}_{\vee}^* - p_{\vee}\} \xrightarrow{D} \mathcal{N}(0, 1).$$

(ii) *If* $r_{\vee} = k$, *then*

$$k^2 n^{1-\alpha}\{\hat{p}_{\vee}^* - p_{\vee}\} \xrightarrow{D} \chi^2_{(k-1)}.$$

As with the minimum, we can consider the case where the number of categories k is unknown. In this case, we replace $\hat{w}_{\vee}$ with

$$\hat{w}_{\vee}^{\sharp} = \sum_{i=1}^{k} e^{n^{\alpha}\hat{p}_i} 1\big(y_i > 0\big) + \hat{\ell},$$

for some estimator $\hat{\ell}$ of ℓ. Under the assumptions of Theorem 3.1 on $\hat{\ell}$, a version of that theorem for the maximum can be verified.

5 Simulations

In this section, we perform a small-scale simulation study to better understand the finite sample performance of the proposed estimator. We consider both the cases where the number of categories is known and where it is unknown. When the number of categories is known, we will compare the finite sample performance of our estimator p_0^* with that of the MLE $\hat{p}_0$ and the LSE $\hat{p}_0^{\text{LS}}$. When the number of categories is unknown, we will compare the performance of $\hat{p}_0^{\sharp}$ with modifications of the MLE and the LSE that do not require knowledge of k. Specifically, we will compare with

$$\hat{p}_{0,u} = \frac{y_0^{\sharp}}{n} \text{ and } \hat{p}_{0,u}^{\text{LS}} = \frac{y_0^{\sharp} + 1}{n + k^{\sharp}}, \tag{10}$$

where $y_0^{\sharp} = \min\{y_i : y_i > 0,\ i = 1, 2, \ldots, k\}$ and $k^{\sharp} = \sum_{i=1}^{k} 1(y_i > 0)$. Clearly, both $\hat{p}_{0,u}$ and $\hat{p}_{0,u}^{\text{LS}}$ can be evaluated without knowledge of k. Throughout this section,

when evaluating p_0^* and $p_0^\sharp$, we set the tuning parameter to be $\alpha = 0.49$. We chose this value because it tends to work well in practice and it is neither too large nor too small. If we take α to be large, then (2) implies that the estimator will be almost indistinguishable from the MLE. On the other hand, if we take α to be small, then the estimator will not work well because it will be too far from convergence.

In our simulations, we consider two distributions. These are the uniform distribution on k categories, denoted by $U(k)$, and the so-called square-root distribution on k categories, denoted by $S(k)$. The $S(k)$ distribution has a probability mass function (pmf) given by

$$p(i) = C\frac{1}{\sqrt{i}}, \quad i = 1, 2, \ldots, k,$$

where C is a normalizing constant. For each distribution, we will consider the case where $k = 10$ and $k = 20$. The true minimums for these distributions are given in Table 1.

The simulations were performed as follows. For each of the four distributions and each sample size n ranging from 1 to 200, we simulated $R = 10000$ random samples of size n. For each of these random samples, we evaluated our estimator. This gave us the values $\hat{p}_{0,1}^*, \hat{p}_{0,2}^*, \ldots, \hat{p}_{0,R}^*$. We used these to estimate the relative root-mean-square error (relative RMSE) as follows:

$$\text{Relative RMSE} = \frac{1}{p_0}\sqrt{\frac{1}{R}\sum_{i=1}^{R}\left(\hat{p}_{0,i}^* - p_0\right)^2} = \sqrt{\frac{1}{R}\sum_{i=1}^{R}\left(\frac{\hat{p}_{0,i}^*}{p_0} - 1\right)^2},$$

where p_0 is the true minimum. We repeated this procedure with each of the estimators. Plots of the resulting relative RMSEs for the various distributions and estimators are given in Fig. 1 for the case where the number of categories k is known and in Fig. 2 for the case where k is unknown. We can see that the proposed estimator works very well for the uniform distributions in all cases. For the square-root distribution, it also beats the other estimators for a wide range of sample sizes.

It may be interesting to note that, in the case where k is known, the relative RMSE of the MLE $\hat{p}_0$ is exactly 1 for smaller sample sizes. This is because, when we have not seen all of the categories in our sample, the MLE is exactly 0. In particular, this holds for any sample size $n < k$. When the MLE is 0, then the LSE $\hat{p}_0^{\text{LS}}$ is exactly $1/(n + k)$. Thus, when k is known and $n < k$, both $\hat{p}_0$ and $\hat{p}_0^{\text{LS}}$ are fully deterministic functions that ignore the data entirely. This is not the case with $\hat{p}_0^*$, which is always based on the data.

When k is unknown, we notice an interesting pattern in the errors of the MLE and the LSE. There is a dip at the beginning, where the errors decrease quickly before increasing just as quickly. After this, they level off and eventually begin to decrease

Table 1 True minimums for the distributions considered

Distribution	$U(10)$	$U(20)$	$S(10)$	$S(20)$
Minimum	0.100	0.050	0.063	0.029

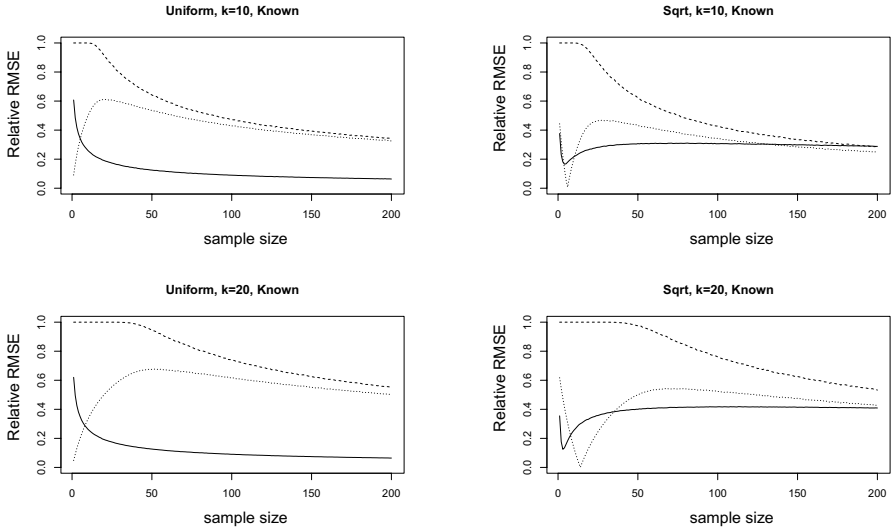

Fig. 1 Plots for the relative RMSE in the case where the sample size k is known. The solid line is for the proposed estimator $\hat{p}_0^*$, the dashed line is for the MLE $\hat{p}_0$, and dotted line is for the LSE $\hat{p}_0^{\mathrm{LS}}$

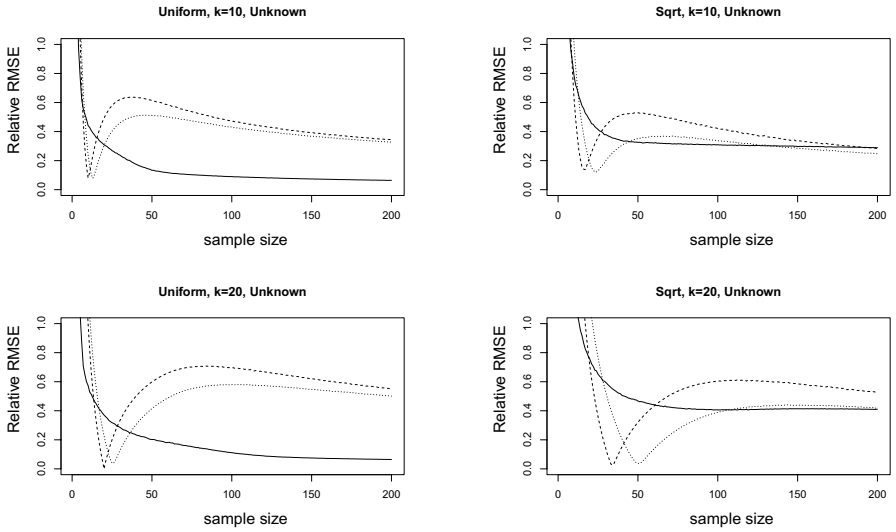

Fig. 2 Plots for the relative RMSE in the case where the sample size k is unknown. The solid line is for the proposed estimator $\hat{p}_0^{\sharp}$, the dashed line is for the MLE $\hat{p}_{0,u}$, and dotted line is the LSE $\hat{p}_{0,u}^{\mathrm{LS}}$

slowly. While it is not clear what causes this, an explanation may be as follows. From (10), we can see that, for relatively small sample sizes, the numerators of both estimators are likely to be small as we would have only seen very few observations from the rarest category. As n begins to increase, the numerators should stay small, while the

denominators increase. This would make the estimators decrease and thus get closer to the value of p_0. However, once n becomes relatively large, the numerators should begin to increase, and thus, the errors would increase as well. It would not be until n gets even larger that it would be large enough for the errors to begin to come down due to the statistical properties of the estimators. If this is correct, then the dip is just an artifact of the deterministic nature of these estimators. For comparison, in most cases the error of p_0^* just decreases as the sample size increases. The one exception is under the square-root distribution, when the number of categories is known. It is not clear what causes the dip in this case, but it may be a similar issue.

6 Conclusions

In this paper, we have introduced a new method for estimating the minimum probability in a multinomial distribution. The proposed approach is based on a smooth approximation of the minimum function. We have considered the cases where the number of categories is known and where it is unknown. The approach is justified by our theoretical results, which verify consistency and categorize the asymptotic distributions. Further, a small-scale simulation study has shown that the method performs better than several baseline estimators for a wide range of sample sizes, although not for all sample sizes. A potential extension would be to prove asymptotic results in the situation where the number of categories increases with the sample size. This would be useful for studying the problem when there are a very large number of categories. Other directions for future research include obtaining theoretical results about the finite sample performance of the estimator and proposing modifications of the estimator with the aim of reducing the bias using, for instance, a jackknife approach.

Acknowledgements This paper was inspired by the question of Dr. Zhiyi Zhang (UNC Charlotte): How to estimate the minimum probability of a multinomial distribution? We thank Ann Marie Stewart for her editorial help. The authors wish to thank two anonymous referees whose comments have improved the presentation of this paper. The second author's work was funded, in part, by the Russian Science Foundation (Project No. 17-11-01098).

Compliance with Ethical Standards

Conflict of interest. On behalf of all authors, the corresponding author states that there is no conflict of interest.

Appendix: Proofs

Throughout the section, let $\Sigma = \mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T$, $\sigma_n = \sqrt{\nabla g_n(\mathbf{p})^T \Sigma \nabla g_n(\mathbf{p})}$, $\Lambda = \lim_{n\to\infty} \nabla g_n(\mathbf{p})$, and $\sigma = \sqrt{\Lambda^T \Sigma \Lambda}$. It is well known that Σ is a positive definite matrix, see, e.g., [12]. For simplicity, we use the standard notation $O(\cdot)$, $o(\cdot)$, $O_p(\cdot)$, and $o_p(\cdot)$, see, e.g., [13] for the definitions. In the case of matrices and vectors, this notation should be interpreted as component wise.

It may, at first, appear that Theorem 2.1 can be proved using the delta method. However, the difficulty lies in the fact that the function $g_n(\cdot)$ depends on n. For this reason, the proof requires a more subtle approach. We begin with several lemmas.

Lemma A.1

1 *There is a constant* $\epsilon > 0$ *such that* $p_0 \le g_n(\mathbf{p}) \le p_0 + (k-r)e^{-n^\alpha \epsilon}$. *When* $r \ne k$, *we can take* $\epsilon = \min_{j:p_j > p_0} (p_j - p_0)$

2. *For any constant* $\beta \in \mathbb{R}$

$$n^\beta \{g_n(\mathbf{p}) - p_0\} \to 0 \text{ as } n \to \infty.$$

3 *For any* $1 \le j \le k$ *and any constant* $\beta \in \mathbb{R}$

$$n^\beta e^{-n^\alpha p_j} w^{-1} \{g_n(\mathbf{p}) - p_j\} \to 0 \text{ as } n \to \infty.$$

Proof We begin with the first part. First, assume that $r = k$. In this case, it is immediate that $g_n(\mathbf{p}) = k^{-1} = p_0$ and the result holds with any $\epsilon > 0$. Now assume $r \ne k$. In this case,

$$p_0 = p_0 \sum_{i=1}^{k} e^{-n^\alpha p_i} w^{-1} \le \sum_{i=1}^{k} p_i e^{-n^\alpha p_i} w^{-1} = g_n(\mathbf{p}).$$

To show the other inequality, note that

$$e^{-n^\alpha p_0} w^{-1} = \left\{ \sum_{i=1}^{k} e^{-n^\alpha (p_i - p_0)} \right\}^{-1} \le (re^0)^{-1} = r^{-1} \tag{11}$$

and that, for any $p_i > p_0$, we have

$$e^{n^\alpha p_i} w = \sum_{j=1}^{k} e^{-n^\alpha (p_j - p_i)} \ge e^{-n^\alpha (p_0 - p_i)} = e^{n^\alpha (p_i - p_0)} \ge \exp\left\{ n^\alpha \min_{j:p_j > p_0} (p_j - p_0) \right\}.$$

Setting $\epsilon = \min_{j:p_j > p_0} (p_j - p_0) > 0$, it follows that, for $p_i > p_0$,

$$e^{-n^\alpha p_i} w^{-1} \le e^{-n^\alpha \epsilon}. \tag{12}$$

We thus get

$$g_n(\mathbf{p}) = \sum_{i:p_i = p_0} p_i e^{-n^\alpha p_i} w^{-1} + \sum_{i:p_i > p_0} p_i e^{-n^\alpha p_i} w^{-1} \le r p_0 (r)^{-1} + (k-r) e^{-n^\alpha \epsilon}.$$

The second part follows immediately from the first. We now turn to the third part. When $p_j = p_0$ Eq. (11) and Part 1 imply that $e^{-n^\alpha p_j} w^{-1} \le r^{-1}$ and that there is an $\epsilon > 0$ such that

$$0 \le g_n(\mathbf{p}) - p_j \le (k-r)e^{-n^\alpha \epsilon}.$$

It follows that when $p_j = p_0$

$$0 \le n^\beta e^{-n^\alpha p_j} w^{-1}\{g_n(\mathbf{P}) - p_j\} \le (k-r)r^{-1}n^\beta e^{-n^\alpha \epsilon} \to 0 \text{ as } n \to \infty.$$

On the other hand, when $p_j > p_0$, by Part 1 there is an $\epsilon > 0$ such that

$$0 \le |g_n(\mathbf{p}) - p_j| \le p_j - p_0 + (k-r)e^{-n^\alpha \epsilon}.$$

Using this and Eq. (12) gives

$$0 \le |n^\beta e^{-n^\alpha p_j} w^{-1}(g_n(\mathbf{p}) - p_j)| \le (p_j - p_0)n^\beta e^{-n^\alpha \epsilon} + (k-r)n^\beta e^{-n^\alpha (2\epsilon)} \to 0,$$

as $n \to \infty$. □

We now consider the case when the probabilities are estimated.

Lemma A.2 *Let* $\mathbf{p}_n^* = \mathbf{p}^* = (p_1^*, \ldots, p_{k-1}^*)$ *be a sequence of random vectors with* $p_i^* \ge 0$ *and* $\sum_{i=1}^{k-1} p_i^* \le 1$. *Let* $p_k = 1 - \sum_{i=1}^{k-1} p_i^*$, $w^* = \sum_{i=1}^{k} e^{-n^\alpha p_i^*}$, *and assume that* $\mathbf{p}_n^* \to \mathbf{p}$ *a.s. and* $n^\alpha(\mathbf{p}_n^* - \mathbf{p}) \xrightarrow{p} 0$. *For every* $j = 1, 2, \ldots, k$, *we have*

$$n^\alpha \left(p_j^* - p_0\right) e^{-n^\alpha p_j^*} \frac{1}{w^*} \xrightarrow{p} 0$$

and

$$n^\alpha e^{-n^\alpha p_j^*} \frac{1}{w^*}\{g_n(\mathbf{p}_n^*) - p_j^*\} \xrightarrow{p} 0 \text{ as } n \to \infty.$$

Proof First note that, from the definition of w^*, we have

$$0 \le e^{-n^\alpha p_j^*} \frac{1}{w^*} \le 1. \tag{13}$$

Assume that $p_j = p_0$. In this case, the first equation follows from (13) and the fact that $n^\alpha \left(p_j^* - p_0\right) = n^\alpha \left(p_j^* - p_j\right) \xrightarrow{p} 0$. In particular, this completes the proof of the first equation in the case where $k = r$.

Now assume that $k \ne r$. Let $p_0^* = \min\{p_i^* : i = 1, 2, \ldots, k\}$, $\epsilon = \min_{i: p_i \ne p_0} \{p_i - p_0\}$, and $\epsilon_n^* = \min_{i: p_i \ne p_0} \{p_i^* - p_0^*\}$. Since $\mathbf{p}_n^* \to \mathbf{p}$ a.s., it follows that $\epsilon_n^* \to \epsilon$ a.s. Further, by arguments similar to the proof of Eq. (12), we can show that, if $p_j \ne p_0$ then there is a random variable N, which is finite a.s., such that for $n \ge N$

$$e^{-n^\alpha p_j^*} \frac{1}{w^*} \le e^{-n^\alpha \epsilon_n^*} \le e^{-n^\alpha \epsilon/2}.$$

It follows that for such j and $n \ge N$

$$n^{\alpha}\left|p_j^* - p_0\right| e^{-n^{\alpha}p_j^*}\frac{1}{w^*} \le 2n^{\alpha}e^{-n^{\alpha}\epsilon/2} \to 0.$$

This completes the proof of the first limit.

Now assume either $k = r$ or $k \neq r$. For the second limit, note that

$$\begin{aligned} & n^{\alpha}e^{-n^{\alpha}p_j^*}\frac{1}{w^*}(g_n(\mathbf{p}_n^*) - p_j^*) \\ &= n^{\alpha}e^{-n^{\alpha}p_j^*}\frac{1}{w^*}(g_n(\mathbf{p}^*) - p_0) + n^{\alpha}e^{-n^{\alpha}p_j^*}\frac{1}{w^*}(p_0 - p_j^*) \\ &= n^{\alpha}e^{-n^{\alpha}p_j^*}\frac{1}{w^*}\sum_{i=1}^{k}(p_i^* - p_0)e^{-n^{\alpha}p_i^*}\frac{1}{w^*} + n^{\alpha}e^{-n^{\alpha}p_j^*}\frac{1}{w^*}(p_0 - p_j^*). \end{aligned}$$

From here the result follows by the first limit and (13). □

Lemma A.3 1. *If $r = k$, then for each $i = 1, 2, \ldots, k$*

$$\frac{\partial g_n(\mathbf{p})}{\partial p_i} = 0.$$

2. *If $r \neq k$, then for each $i = 1, 2, \ldots, k$*

$$\lim_{n\to\infty}\frac{\partial g_n(\mathbf{p})}{\partial p_i} = \begin{cases} r^{-1}, & \text{if } p_k \neq p_0 \text{ and } p_i = p_0 \\ -r^{-1}, & \text{if } p_k = p_0 \text{ and } p_i \neq p_0 \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

Proof When $r = k$, the result is immediate from (4). Now assume that $r \neq k$. We can rearrange equation (4) as

$$\frac{\partial g_n(\mathbf{p})}{\partial p_i} = w^{-1}\left(e^{-n^{\alpha}p_i} - e^{-n^{\alpha}p_k}\right) + r_n, \tag{15}$$

where $r_n = n^{\alpha}e^{-n^{\alpha}p_i}w^{-1}\{g_n(\mathbf{p}) - p_i\} - n^{\alpha}e^{-n^{\alpha}p_k}w^{-1}\{g_n(\mathbf{p}) - p_k\}$. Note that Lemma A.1 implies that $r_n \to 0$ as $n \to \infty$. It follows that

$$\begin{aligned} \lim_{n\to\infty}\frac{\partial g_n(\mathbf{p})}{\partial p_i} &= \lim_{n\to\infty} e^{-n^{\alpha}p_i}w^{-1} - \lim_{n\to\infty} e^{-n^{\alpha}p_k}w^{-1} \\ &= \lim_{n\to\infty}\left\{\sum_{j=1}^{k} e^{-n^{\alpha}(p_j - p_i)}\right\}^{-1} - \lim_{n\to\infty}\left\{\sum_{j=1}^{k} e^{-n^{\alpha}(p_j - p_k)}\right\}^{-1}. \end{aligned}$$

Consider the case where $p_k \neq p_0$ and $p_i = p_0$. In this case, the first part has r component(s) in the denominator that are equal to one (e^0) and the remaining $k - r$ terms go to zero individually. However, since $p_k \neq p_0$, the denominator of the second fraction has r terms of the form $e^{-n^{\alpha}(p_0 - p_k)}$, which go to $+\infty$, while the other terms go to 0, 1, or $+\infty$. Thus, in this case, the limit is $r^{-1} - 0 = r^{-1}$. The arguments in the other cases are similar and are thus omitted. □

Lemma A.4 *Assume that* $r \neq k$ *and let* $\mathbf{p}_n^*$ *be as in Lemma A.2. In this case,* $\frac{\partial(g_n(\mathbf{p}))}{\partial p_i} = O(1)$, $\frac{\partial(g_n(\mathbf{p}_\mathbf{n}^*))}{\partial p_i} = O_p(1)$, $\frac{\partial^2(g_n(\mathbf{p}))}{\partial p_i \partial p_j} = O(n^\alpha)$, $\frac{\partial^2(g_n(\mathbf{p}_\mathbf{n}^*))}{\partial p_i \partial p_j} = O_p(n^\alpha)$, $\frac{\partial^3(g_n(\mathbf{p}))}{\partial p_\ell \partial p_i \partial p_j} = O(n^{2\alpha})$, *and* $\frac{\partial^3(g_n(\mathbf{p}_\mathbf{n}^*))}{\partial p_\ell \partial p_i \partial p_j} = O_p(n^{2\alpha})$.

Proof The results for the first derivatives follow immediately from (4), (13), Lemma A.2, and Lemma A.3. Now let δ_{ij} be 1 if $i = j$ and zero otherwise. It is straightforward to verify that

$$\begin{aligned}\frac{\partial^2 g_n(\mathbf{p})}{\partial p_j \partial p_i} =& n^\alpha w^{-1}\left(e^{-n^\alpha p_i} - e^{-n^\alpha p_k}\right)\frac{\partial g_n(\mathbf{p})}{\partial p_j} \\ &+ n^\alpha w^{-1}\left(e^{-n^\alpha p_j} - e^{-n^\alpha p_k}\right)\frac{\partial g_n(\mathbf{p})}{\partial p_i} \\ &- n^\alpha e^{-n^\alpha p_k} w^{-1}\left[n^\alpha\left(g_n(\mathbf{p}) - p_k\right) + 2\right] \\ &- \delta_{ij} n^\alpha e^{-n^\alpha p_i} w^{-1}\left[n^\alpha\left(g_n(\mathbf{p}) - p_i\right) + 2\right],\end{aligned} \tag{16}$$

that for $\ell \neq i$ and $\ell \neq j$ we have

$$\begin{aligned}\frac{\partial^3 g_n(\mathbf{p})}{\partial p_\ell \partial p_j \partial p_i} =& n^\alpha w^{-1}\left(e^{-n^\alpha p_\ell} - e^{-n^\alpha p_k}\right)\frac{\partial^2 g_n(\mathbf{p})}{\partial p_j \partial p_i} \\ &+ n^\alpha w^{-1}\left(e^{-n^\alpha p_i} - e^{-n^\alpha p_k}\right)\frac{\partial^2 g_n(\mathbf{p})}{\partial p_\ell \partial p_j} \\ &+ n^\alpha w^{-1}\left(e^{-n^\alpha p_j} - e^{-n^\alpha p_k}\right)\frac{\partial^2 g_n(\mathbf{p})}{\partial p_\ell \partial p_i} \\ &- n^{2\alpha} e^{-n^\alpha p_k} w^{-1}\left(\frac{g_n(\mathbf{p})}{\partial p_\ell} + \frac{\partial g_n(\mathbf{p})}{\partial p_j} + \frac{\partial g_n(\mathbf{p})}{\partial p_i} + 1\right) \\ &- n^{2\alpha} e^{-n^\alpha p_k} w^{-1}\left[n^\alpha\left(g_n(\mathbf{p}) - p_k\right) + 2\right] \\ &- \delta_{ij} n^{2\alpha} e^{-n^\alpha p_i} w^{-1}\frac{\partial g_n(\mathbf{p})}{\partial p_\ell},\end{aligned} \tag{17}$$

and that for $i = j = \ell$ we have

$$\begin{aligned}\frac{\partial^3 g_n(\mathbf{p})}{\partial p_i^3} =& n^\alpha w^{-1}\left(e^{-n^\alpha p_i} - e^{-n^\alpha p_k}\right)\left(3\frac{\partial^2 g_n(\mathbf{p})}{\partial p_i^2} + 2n^\alpha\right) \\ &+ n^{2\alpha}\frac{\partial g_n(\mathbf{p})}{\partial p_i}\left[1 - 3w^{-1}\left(e^{-n^\alpha p_i} + e^{-n^\alpha p_k}\right)\right].\end{aligned} \tag{18}$$

Combining this with Lemma A.2 and the fact that $0 \leq w^{-1}e^{-n^\alpha p_s} \leq 1$ for any $1 \leq s \leq k$ gives the result. □

Lemma A.5 *Assume* $r \neq k$ *and* $0 < \alpha < 0.5$, *then* $\nabla g_n(\hat{\mathbf{p}}) - \nabla g_n(\mathbf{p}) = O_p(n^{\alpha - \frac{1}{2}})$.

Proof By the mean value theorem, we have

$$n^{\frac{1}{2}-\alpha}\nabla g_n(\hat{\mathbf{p}}) = n^{\frac{1}{2}-\alpha}\nabla g_n(\mathbf{p}) + n^{-\alpha}\nabla^2 g_n(\mathbf{p}^*)\sqrt{n}(\hat{\mathbf{p}}-\mathbf{p}), \tag{19}$$

where $\mathbf{p}^* = \mathbf{p} + \text{diag}(\boldsymbol{\omega})(\hat{\mathbf{p}}-\mathbf{p})$ for some $\boldsymbol{\omega} \in [0,1]^{k-1}$. Note that by the strong law of large numbers $\hat{\mathbf{p}} \to \mathbf{p}$ a.s., which implies that $\mathbf{p}^* - \mathbf{p} \to 0$ a.s. Similarly, by the multivariate central limit theorem and Slutsky's theorem $n^{\alpha}(\hat{\mathbf{p}}-\mathbf{p}) \xrightarrow{p} 0$ implies that $n^{\alpha}(\mathbf{p}^*-\mathbf{p}) \xrightarrow{p} 0$. Thus, the assumptions of Lemma A.4 are satisfied and that lemma gives

$$n^{-\alpha}\nabla^2 g_n(\mathbf{p}^*)\sqrt{n}(\hat{\mathbf{p}}-\mathbf{p}) = n^{-\alpha}O_p(n^{\alpha})O_p(1).$$

From here, the result is immediate. □

Lemma A.6 *Assume that $r \neq k$. In this case, $\sigma > 0$ and $\lim_{n\to\infty} \sigma_n^{-1}\sigma = 1$. Further, if $0 < \alpha < 0.5$, then $\hat{\sigma}_n^{-1}\sigma_n \xrightarrow{p} 1$.*

Proof Since Σ is a positive definite matrix, and by Lemma A.3, $\Lambda \neq 0$, it follows that $\sigma > 0$. From here, the fact that $\lim_{n\to\infty} \sigma_n = \sigma$ gives the first result. Now assume that $0 < \alpha < 0.5$. It is easy to see that $\hat{p}_i\hat{p}_j - p_ip_j = \hat{p}_j(\hat{p}_i - p_i) + p_i(\hat{p}_j - p_j) = O_p(n^{-1/2})$ and $\hat{p}_i(1-\hat{p}_i) - p_i(1-p_i) = (\hat{p}_i - p_i)(1-p_i-\hat{p}_i) = O_p(n^{-1/2})$. Thus, $\hat{\Sigma} = \Sigma + O_p(n^{-1/2})$. This together with Lemma A.3 and Lemma A.5 leads to

$$\begin{aligned}
\frac{\hat{\sigma}_n^2}{\sigma_n^2} &= \frac{\nabla g_n(\hat{\mathbf{p}})^T\hat{\Sigma}\nabla g_n(\hat{\mathbf{p}})}{\nabla g_n(\mathbf{p})^T\Sigma\nabla g_n(\mathbf{p})} \\
&= \frac{(\nabla g_n(\mathbf{p}) + O_p(n^{\alpha-\frac{1}{2}}))^T(\Sigma + O_P(n^{-\frac{1}{2}}))(\nabla g_n(\mathbf{p}) + O_p(n^{\alpha-\frac{1}{2}}))}{\nabla g_n(\mathbf{p})^T\Sigma\nabla g_n(\mathbf{p})} \\
&= 1 + O_p(n^{\alpha-\frac{1}{2}}) + O_p(n^{\alpha-1}) + O_p(n^{2\alpha-\frac{3}{2}}) + O_p(n^{-\frac{1}{2}}) + O_p(n^{2\alpha-1}) \xrightarrow{p} 1,
\end{aligned}$$

which completes the proof. □

Lemma A.7 *If $r \neq k$ and $0 < \alpha < 0.5$, then $\sqrt{n}\hat{\sigma}_n^{-1}\{g_n(\hat{\mathbf{p}}) - g_n(\mathbf{p})\} \xrightarrow{D} \mathcal{N}(0,1)$.*

Proof Taylor's theorem implies that

$$\begin{aligned}
\sqrt{n}(g_n(\hat{\mathbf{p}}) - g_n(\mathbf{p})) = &\sqrt{n}(\hat{\mathbf{p}}-\mathbf{p})^T\nabla g_n(\mathbf{p}) \\
&+ 0.5\sqrt{n}(\hat{\mathbf{p}}-\mathbf{p})^T n^{-\alpha}\nabla^2 g_n(\mathbf{p}^*)n^{\alpha}(\hat{\mathbf{p}}-\mathbf{p}),
\end{aligned}$$

where $\mathbf{p}^* = \mathbf{p} + \text{ diag } (\boldsymbol{\omega})(\hat{\mathbf{p}}-\mathbf{p})$ for some $\boldsymbol{\omega} \in [0,1]^{k-1}$. Using Lemma A.4 and arguments similar to those used in the proof of Lemma A.5 gives $n^{-\alpha}\nabla^2 g_n(\mathbf{p}^*) = O_p(1)$, $\sqrt{n}(\hat{\mathbf{p}}-\mathbf{p}) = O_p(1)$, and $n^{\alpha}(\hat{\mathbf{p}}-\mathbf{p}) = o_p(1)$. It follows that the second term on the RHS above is $o_p(1)$ and hence that

$$\sqrt{n}(g_n(\hat{\mathbf{p}}) - g_n(\mathbf{p})) = \sqrt{n}(\hat{\mathbf{p}}-\mathbf{p})^T\nabla g_n(\mathbf{p}) + o_p(1).$$

It is well known that $\sqrt{n}(\hat{\mathbf{p}}-\mathbf{p})^T \xrightarrow{D} \mathcal{N}(0,\Sigma)$. Hence

$$\sqrt{n}(\hat{\mathbf{p}}-\mathbf{p})^T\Lambda \xrightarrow{D} \mathcal{N}(0,\Lambda^T\Sigma\Lambda)$$

and, by Slutsky's theorem,

$$\sqrt{n}(g_n(\hat{\mathbf{p}})-g_n(\mathbf{p})) \xrightarrow{D} \mathcal{N}(0,\Lambda^T\Sigma\Lambda).$$

By Lemma A.6, $\sigma_n^{-1}\sigma \to 1$, and $\hat{\sigma}_n^{-1}\sigma_n \xrightarrow{p} 1$. Hence, the result follows by another application of Slutsky's theorem. □

Lemma A.8 *Let* $\mathbf{A} = -n^{-\alpha}\nabla^2 g_n(\mathbf{p})$ *and let* $\mathbf{I}_{k-1}$ *be the* $(k-1)\times(k-1)$ *identity matrix. If* $r = k$, *then* $\Sigma^{\frac{1}{2}}\mathbf{A}\Sigma^{\frac{1}{2}} = 2k^{-2}\mathbf{I}_{k-1}$.

Proof Let $\mathbf{1}$ be the column vector in $\mathbb{R}^{k-1}$ with all entries equal to 1. By Eq. (16), we have

$$\mathbf{A} = -n^{-\alpha}\nabla^2 g_n(\mathbf{p}) = 2k^{-1}[\mathbf{1}\mathbf{1}^T + \mathbf{I}_{k-1}]. \tag{20}$$

Note that $\Sigma = \text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T = k^{-2}[k\mathbf{I}_{k-1} - \mathbf{1}\mathbf{1}^T]$. It follows that

$$\begin{aligned}\mathbf{A}\Sigma =& 2k^{-1}[\mathbf{1}\mathbf{1}^T + \mathbf{I}_{k-1}]k^{-2}[k\mathbf{I}_{k-1} - \mathbf{1}\mathbf{1}^T]\\ =& 2k^{-3}[k\mathbf{1}\mathbf{1}^T - \mathbf{1}\mathbf{1}^T\mathbf{1}\mathbf{1}^T + k\mathbf{I}_{k-1} - \mathbf{1}\mathbf{1}^T]\\ =& 2k^{-3}[k\mathbf{1}\mathbf{1}^T - (k-1)\mathbf{1}\mathbf{1}^T + k\mathbf{I}_{k-1} - \mathbf{1}\mathbf{1}^T] = 2k^{-2}\mathbf{I}_{k-1}.\end{aligned}$$

Now multiplying both sides by $\Sigma^{1/2}$ on the left and $\Sigma^{-1/2}$ on the right gives the result. □

Proof of Theorem 2.1 (i) If $r \neq k$, then

$$\sqrt{n}\hat{\sigma}_n^{-1}\{\hat{p}_0^* - p_0\} = \sqrt{n}\hat{\sigma}_n^{-1}\{g_n(\hat{\mathbf{p}}) - g_n(\mathbf{p})\} + \sqrt{n}\hat{\sigma}_n^{-1}\{g_n(\mathbf{p}) - p_0\}. \tag{21}$$

The first part approaches a $\mathcal{N}(0,1)$ distribution by Lemma A.7, and the second part approaches zero in probability by Lemmas A.6 and A.1. From there, the first part of the theorem follows by Slutsky's theorem.

(ii) Assume that $r = k$. In this case, $g_n(\mathbf{p}) = p_0 = k^{-1}$, and by Lemma A.3, $\nabla g_n(\mathbf{p}) = 0$. Thus, Taylor's theorem gives

$$\begin{aligned}n^{1-\alpha}\{p_0 - \hat{p}_0^*\} =& n^{1-\alpha}\{g_n(\mathbf{p}) - g_n(\hat{\mathbf{p}})\}\\ =& 0.5\sqrt{n}(\hat{\mathbf{p}}-\mathbf{p})^T(-n^{-\alpha})\nabla^2 g_n(\mathbf{p})\sqrt{n}(\hat{\mathbf{p}}-\mathbf{p}) + r_n,\end{aligned} \tag{22}$$

where $r_n = -6^{-1}\sum_{q=1}^{k-1}\sum_{r=1}^{k-1}\sum_{s=1}^{k-1}\sqrt{n}(\hat{p}_q - p_q)\sqrt{n}(\hat{p}_r - p_r)n^{\alpha}(\hat{p}_s - p_s)n^{-2\alpha}\frac{\partial^3 g_n(\mathbf{p}^*)}{\partial p_q \partial p_r \partial p_s}$, and $\mathbf{p}^* = \mathbf{p} + \text{diag}\,(\boldsymbol{\omega})(\hat{\mathbf{p}} - \mathbf{p})$ for some $\boldsymbol{\omega} \in [0,1]^{k-1}$. Lemma A.4 implies that $n^{-2\alpha}\frac{\partial^3 g_n(\mathbf{p}^*)}{\partial p_q \partial p_r \partial p_s} = O_p(1)$. Combining this with the facts that $\sqrt{n}(\hat{p}_q - p_q)$ and $\sqrt{n}(\hat{p}_r - p_r)$ are $O_p(1)$ and that, for $\alpha \in (0, 0.5)$, $n^{\alpha}(\hat{p}_s - p_s) = o_p(1)$, it follows that $r_n \xrightarrow{p} 0$.

Let $\mathbf{x}_n = \sqrt{n}(\hat{\mathbf{p}} - \mathbf{p})$, $\mathbf{T}_n = \Sigma^{-\frac{1}{2}}\mathbf{x}_n$, and $\mathbf{A} = -n^{\alpha}\nabla^2 g_n(\mathbf{p})$. Lemma A.8 implies that

$$\begin{aligned}\mathbf{x}_n^T\mathbf{A}\mathbf{x}_n =&(\Sigma^{-\frac{1}{2}}\mathbf{x}_n)^T\Sigma^{\frac{1}{2}}\mathbf{A}\Sigma^{\frac{1}{2}}(\Sigma^{-\frac{1}{2}}\mathbf{x}_n)\\ =&\mathbf{T}_n^T(2k^{-2}\mathbf{I}_{k-1})\mathbf{T}_n.\end{aligned}$$

Since $\mathbf{x}_n \xrightarrow{D} \mathcal{N}(0,\Sigma)$, we have $\mathbf{T}_n \xrightarrow{D} \mathbf{T}$, where $\mathbf{T} \sim \mathcal{N}(0,\mathbf{I}_{k-1})$. Let T_i be the ith component of vector $\mathbf{T}$. Applying the continuous mapping theorem, we obtain

$$\mathbf{x}_n^T\mathbf{A}\mathbf{x}_n \xrightarrow{D} \mathbf{T}^T(2k^{-2}\mathbf{I}_{k-1})\mathbf{T} = 2k^{-2}\sum_{i=1}^{k-1} T_i^2.$$

Thus, Eq. (22) becomes

$$n^{1-\alpha}\{p_0 - g_n(\hat{\mathbf{p}})\} = 0.5\mathbf{x}_n^T\mathbf{A}\mathbf{x}_n + o_p(1) \xrightarrow{D} k^{-2}\sum_{i=1}^{k-1} T_i^2.$$

The result follows from the fact that the T_i^2 are independent and identically distributed random variables, each following the Chi-square distribution with 1 degree of freedom. □

The proof of Theorem 4.1 is very similar to that of Theorem 2.1 and is thus omitted. However, to help the reader to reconstruct the proof, we note that the partial derivatives of $g_n^\vee$ can be calculated using the facts that

$$\frac{\partial g_n^\vee(\mathbf{p})}{\partial p_j} = \frac{\partial g_n(-\mathbf{p})}{\partial p_j} \text{ and } \frac{\partial^2 g_n^\vee(\mathbf{p})}{\partial p_i \partial p_j} = -\frac{\partial^2 g_n(-\mathbf{p})}{\partial p_i \partial p_j}.$$

Further, we formulate a version of Lemmas A.1 and A.2 for the maximum.

Lemma A.9

1. *There is a constant* $\epsilon > 0$ *such that* $p_\vee - (k - r_\vee)e^{-n^\alpha\epsilon} \le g_n^\vee(\mathbf{p}) \le p_\vee$. *When* $r_\vee \ne k$, *we can take* $\epsilon = \min_{j:p_j<p_\vee}(p_\vee - p_j)$.
2. *For any constant* $\beta \in \mathbb{R}$

$$n^\beta\{g_n^\vee(\mathbf{p}) - p_\vee\} \to 0 \text{ as } n \to \infty.$$

3. *For any* $1 \le j \le k$ *and any constant* $\beta \in \mathbb{R}$

$$n^\beta \frac{e^{n^\alpha p_j}}{w^\vee}\{g_n^\vee(\mathbf{p}) - p_j\} \to 0 \text{ as } n \to \infty.$$

4. *If* $\mathbf{p}_n^*$ *is as in Lemma A.2 and* $w^{\vee *} = \sum_{i=1}^{k} e^{n^\alpha p_i^*}$, *then for every* $j = 1, 2, \ldots, k$ *we have*

$$n^\alpha\left(p_j^* - p_\vee\right)e^{n^\alpha p_j^*}\frac{1}{w^{\vee *}} \xrightarrow{p} 0$$

and

$$n^{\alpha}e^{n^{\alpha}p_j^*}\frac{1}{w^{\vee *}}\{g_n^{\vee}(\mathbf{p}_n^*)-p_j^*\}\xrightarrow{p}0 \text{ as } n\to\infty.$$

Proof We only prove the first part, as proofs of the rest are similar to those of Lemmas A.1 and A.2. If $r_{\vee}=k$, then $g_n^{\vee}(\mathbf{p})=1/k=p_{\vee}$ and the result holds with any $\epsilon>0$. Now, assume that $k\neq r_{\vee}$ and let ϵ be as defined above. First note that

$$g_n^{\vee}(\mathbf{p})=\sum_{j=1}^{k}p_j e^{n^{\alpha}p_j}\frac{1}{w^{\vee}}\leq p_{\vee}\sum_{j=1}^{k}e^{n^{\alpha}p_j}\frac{1}{w^{\vee}}=p_{\vee}.$$

Note further that for $p_i<p_{\vee}$

$$\begin{aligned}\frac{e^{n^{\alpha}p_j}}{w^{\vee}}&=\left\{\sum_{i=1}^{k}e^{n^{\alpha}(p_i-p_j)}\right\}^{-1}\leq\left\{\sum_{i:p_i=p_{\vee}}e^{n^{\alpha}(p_{\vee}-p_j)}\right\}^{-1}\\&=\frac{1}{r_{\vee}}e^{-n^{\alpha}(p_{\vee}-p_j)}\leq\frac{1}{r_{\vee}}e^{-n^{\alpha}\epsilon}.\end{aligned}$$

It follows that

$$\begin{aligned}g_n^{\vee}(\mathbf{p})&\geq\sum_{i:p_i=p_{\vee}}p_i e^{n^{\alpha}p_i}\frac{1}{w^{\vee}}=p_{\vee}\frac{r_{\vee}e^{n^{\alpha}p_{\vee}}}{w^{\vee}}=p_{\vee}+p_{\vee}\left(\frac{r_{\vee}e^{n^{\alpha}p_{\vee}}}{w^{\vee}}-1\right)\\&=p_{\vee}+\frac{p_{\vee}}{w^{\vee}}\left(r_{\vee}e^{n^{\alpha}p_{\vee}}-\sum_{i:p_i=p_{\vee}}e^{n^{\alpha}p_{\vee}}-\sum_{i:p_i<p_{\vee}}e^{-n^{\alpha}p_i}\right)\\&=p_{\vee}-\frac{p_{\vee}}{w^{\vee}}\sum_{i:p_i<p_{\vee}}e^{-n^{\alpha}p_i}\geq p_{\vee}-\frac{p_{\vee}}{r_{\vee}}(k-r_{\vee})e^{-n^{\alpha}\epsilon}.\end{aligned}$$

From here the result follows. □

References

1. Boyd S, Vandenberghe L (2004) Convex Optimization. Cambridge University Press, Cambridge
2. Chao A (1984) Nonparametric estimation of the number of classes in a population. Scandinavian J Stat 11:265–270
3. Chao A (1987) Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43:783–791
4. Chung K, AitSahlia F (2003) Elementary Probability Theory with Stochastic Processes and an Introduction to Mathematical Finance, 4th edn. Springer, New York
5. Colwell C (1994) Estimating terrestrial biodiversity through extrapolation. Philos Trans Biol Sci 345:101–118
6. Csurka G, Dance CR, Fan L, Willamowski J, Bray C (2004) 'Visual categorization with bags of keypoints', In: Workshop on Statistical Learning in Computer Vision, ECCV, pp. 1–22

7. Grabchak M, Marcon E, Lang G, Zhang Z (2017) The generalized Simpson's entropy is a measure of biodiversity. PLOS ONE 12:e0173305
8. Grabchak M, Zhang Z (2018) Asymptotic normality for plug-in estimators of diversity indices on countable alphabets. J Nonparam Stat 30:774–795
9. Gu Z, Shao M, Li L, Fu Y (2012) 'Discriminative metric: Schatten norm vs. vector norm', In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), pp. 1213–1216
10. Haykin S (1994) Neural networks: a comprehensive foundation. Pearson Prentice Hall, New York
11. Lange M, Zühlke D, Holz T, Villmann O (2014) 'Applications of l_p-norms and their smooth approximations for gradient based learning vector quantization', In: ESANN 2014: Proceedings of the 22nd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, pp. 271–276
12. May WL, Johnson WD (1998) On the singularity of the covariance matrix for estimates of multinomial proportions. J Biopharmaceut Stat 8:329–336
13. Shao J (2003) Mathematical Statistics, 2nd edn. Springer, New York
14. Turney P, Littman ML (2003) Measuring praise and criticism: inference of semantic orientation from association. ACM Trans Inf Syst 21:315–346
15. Zhai C, Lafferty J (2017) A study of smoothing methods for language models applied to ad hoc information retrieval. ACM SIGIR Forum 51:268–276
16. Zhang Z, Chen C, Zhang J (2020) Estimation of population size in entropic perspective. Commun Stat Theory Methods 49:307–324