Good-Bootstrap: Simultaneous Confidence Intervals for Large Alphabet Distributions

Daniel Marton (danielmarton@mail.tau.ac.il)
Tel Aviv University
Amichai Painsky
Tel Aviv University

Research Article

Keywords: Simultaneous Confidence Intervals, Multinomial Distribution, Good-Turing, large Alphabet, Count Data

Posted Date: September 7th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2005957/v1

Good-Bootstrap: Simultaneous Confidence Intervals for Large Alphabet Distributions

Daniel Marton[1*] and Amichai Painsky[2]

[1]School of Mathematical Sciences, Tel Aviv University, Tel Aviv, Israel.
[2]Department of Industrial Engineering, Tel Aviv University, Tel Aviv, Israel.

*Corresponding author(s). E-mail(s): danielmarton@mail.tau.ac.il;
Contributing authors: amichaip@tauex.tau.ac.il;

Abstract

Simultaneous confidence intervals (SCI) for multinomial proportions are a corner stone in count data analysis and a key component in many applications. A variety of schemes were introduced over the years, mostly focusing on an asymptotic regime (where the sample is large), or a small sample regime, where the alphabet size is relatively small. In this work we introduce a new SCI framework which considers a large alphabet setup. Our proposed framework utilizes bootstrap sampling with the Good-Turing probability estimator as a plug-in distribution. We demonstrate the favorable performance of our proposed method in synthetic and real-world experiments. Importantly, we provide an exact analytical expression for the bootstraped statistic, which replaces the computationally costly sampling routine. Our proposed framework is publicly available at the first author's webpage.

Keywords: Simultaneous Confidence Intervals, Multinomial Distribution, Good-Turing, large Alphabet, Count Data

1 Introduction

Consider a multinomial distribution p over an alphabet size m. Let $X^n \triangleq \{X_1, \ldots, X_n\}$ be a collection of n independent and identically distributed samples from p. In this work we study simultaneous confidence intervals (SCIs) for p. SCIs for multinomial proportions are a corner stone of statistical inference. This problem was extensively studied over the years, with several notable contributions such as Quesenberry and Hurst [26], Goodman [12], Fitzpatrick and Scott [7] and Sison and Glaz [28]. Current methodologies focus on two basic setups. The first considers an asymptotic regime, where the sample size is very large [12, 26]. The second addresses a fixed sample size, where the alphabet size is relatively small [28]. In this work we study the complementary case, where the alphabet size is large, or at least comparable to the sample size. This setup is also known as the *large alphabet* regime. Large alphabet modeling is of special interest in many applications such as language processing and bioinformatics [22]. One of the typical challenges in this setup is the inference of rare events. Specifically, the probability of symbols that do not appear in the sample (unseen symbols). Here, classical tools tend to underestimate the desired parameters [21] and more sophisticated schemes are required.

The problem of estimating p in the large alphabet regime was extensively studied in the context of point estimation. Probably the first to address

this problem was Pierre-Simon Laplace [30]. In his work, Laplace suggested adding a single count to every symbol in the alphabet. This way, unseen symbols are inferred as if they have a single count in the sample. A major milestone to large alphabet point estimation was achieved in the work of I.J. Good and A.M. Turing during world war II, while deciphering the Enigma Code [10]. Good and Turing proposed a surprisingly efficient and unintuitive framework which assigns symbols with t appearance a probability proportional to the number of events that appear $t+1$ times. Since its appearance the Good-Turing (GT) estimator has gained popularity in a variety of fields. To this day, GT estimators are perhaps the most commonly used methods in practical setups [22].

In this paper, we study interval estimation for the multinomial proportions over large alphabets. We introduce a new approach which applies bootstrap sampling with the GT estimator as the plug-in distribution. To the best of our knowledge, our method is the first to directly address multinomial interval estimation in the large alphabet regime. We demonstrate the performance of our suggested scheme in a series of synthetic and real-world experiments. We show it outperforms popular alternatives, as it attains significantly smaller SCIs while maintaining the prescribed coverage rate.

2 Related Work

We distinguish between two setups of interest. The first, considers the case where n is large (asymptotic), compared to the alphabet size, while the second addresses a fixed n and a relatively small m. In the first regime, Quesenbeny and Hurst [26] suggested joint confidence intervals based on large sample properties of the sample proportions and the inversion of Pearson's chi-square goodness-of-fit test. The intervals control the joint coverage probability for all possible linear combinations of the parameters. Therefore, these intervals are necessarily conservative, often wider than necessary, with larger coverage than required [18]. Goodman [12] proposed an adaptation for Quesenbeny and Hurst by using the Bonferroni inequality, making them less conservative and thus shorter for the same confidence level. Fitzpatrick and Scott [7] introduced an adaptation of the binomial distribution to the multinomial scheme by finding a lower bound for the simultaneous coverage probability of all binomial symbols CIs together.

All the methods above are popular and well-established SCI schemes. Unfortunately, they all assume an asymptotic regime in their derivation. Consequently, they perform quite poorly in cases of small sample size, small number of observed symbol frequencies or sampling zeros [18].

The second regime focuses on the case where n is fixed and the alphabet size m is relatively small. Here, the most popular SCI scheme is arguably of Sison and Glaz [28]. In their work, Sison and Glaz (SG) proposed a method that is not based on large sample properties. They used the relationship between the Poisson, truncated Poisson and multinomial distributions to derive an alternative formulation for the joint multinomial probability, which is then approximated by Edgeworth expansions. Through extensive simulations, they demonstrated their method leads to smaller SCIs while maintaining a coverage rate closer to the desired level, compared to known methods at the time. Unfortunately, the SG algorithm does not perform well in cases where the expected symbol counts are disparate [18].

In their review survey, May and Johnson [18] performed a simulation study and compared different SCIs methods. They recommended the Goodman intervals for cases where the sample size supports the large sample theory, m is small and the expected counts are at least ten per symbol. For cases where the expected symbol counts are small and nearly equal across all symbols, they recommended using SG intervals. No method was uniformly superior for every examined setup. Moreover, typically one cannot assess the expected symbol counts prior to the experiment.

It is important to emphasise that SCIs may also be obtained from a binomial viewpoint, where each symbol is treated in turn against all other symbols. That is, one may construct a binomial CI for each symbol independently (for example, by using Clopper-Pearson intervals [3]) and correct for multiplicity (for example, by applying a Bonferroni correction [4]). Henceforth, the SCIs are just a collection of binomial CIs (of confidence level α/m), for every symbol in the alphabet. Naturally, this approach controls the prescribed confidence level (for every n and m), but may be over pessimistic and result in large CIs. Notice that such an approach also applies for unobserved

symbols. Specifically, the binomial CI for symbols with zero counts is obtained by the *rule-of-three*, which suggests $[0, -\log(\alpha/m)/n]$ as a CI for this case [6].

Additional SCI methods were proposed more recently. For example, Hou et al. [14] based their method on inverting power-divergence statistics. Chafaï and Concordet [2] proposed a method that consists of the inversion of the *covering collection* associated with level-sets of the likelihood. Withers and Nadarajah [29] proposed a modification for the well known Pearson test statistic using a Laplace correction making it usable also for cases of unsampled symbols. However, their results are still based on asymptotic large sample properties. To the best of our knowledge, no method directly addresses the construction of SCIs for multinomial proportions over for large alphabets.

The problem of large alphabet probability estimation is of great interest and was extensively studied over the years, mostly in the context of point estimation. One of the major challenges when dealing with large alphabets is under sampled symbols. This problem is mostly evident in cases where some symbols are not sampled at all (for example, consider $n < m$). Historically, Laplace [15] was perhaps the first to address this problem [30]. In his estimation scheme, a single count is added to every symbol in the alphabet, followed by maximum likelihood estimation. This solution guarantees that no symbols are missing from the sample. Laplace's scheme was later generalized to a family of add-constant estimators, where instead of adding a single count, a constant number c is added. The add-constant estimators are very simple and practical. Unfortunately, they preform quite poorly is cases where the alphabet size is much larger than the sample [21].

During World War II, in an effort to decipher the Enigma Code, Good and Turing [10] developed an alternative method for estimating the proportions of large multinomial distributions. The basic Good-Turing (GT) framework considers an estimator for symbols that appear t times in the sample. Formally,

$$\hat{p}_t = \frac{\varphi_{t+1}}{\varphi_t} \cdot \frac{t+1}{n}, \tag{1}$$

where φ_t is the number of symbols appearing t times in the sample. In words, the probability of the symbols that appear t times in the sample is proportional to the number of symbols that appear $t+1$ times. Hence, non-sampled probabilities are assigned a probability proportional to the number of events with a single appearance in the sample. Notice that as opposed to the add-constant estimators, the GT estimator is oblivious to the alphabet size, which makes it more robust.

The GT estimator has gained a great popularity and was applied to a variety of fields. Perhaps its most common application is in language modeling, where it is used to estimate the probability distribution of words [22]. On the theoretical side, interpretations of the GT estimator have been proposed [11, 20] and its favorable properties were studied [19, 22–25].

In practice, it was shown that the original GT scheme (1) is sub-optimal for symbols that appear more frequently [8]. This problem is addressed by several adaptations. Gale [8] proposed a smoothed version of the algorithm which utilizes linear regression to smooth the erratic values. Another approach is to use an hybrid scheme, which applies GT for low frequency symbols and maximum likelihood for symbols that appear many times [25]. An additional example, which we utilize later in this paper, was introduced in [22],

$$\hat{p}_t = \begin{cases} \frac{t}{n} & \text{if } t > \varphi_{t+1} \\ \frac{\varphi_{t+1}+1}{\varphi_t} \cdot \frac{t+1}{n^*} & \text{else,} \end{cases} \tag{2}$$

where n^* is a normalization factor. Notice that the term φ_{t+1} in the original GT formulation (1) is replaced with $\varphi_{t+1}+1$ to ensure that every symbol is assigned a non-zero probability.

3 Problem Statement

Let $\mathcal{X}$ be a finite alphabet of size m. Let $p = p_1, \ldots, p_m$ be an unknown probability distribution over $\mathcal{X}$. Let $X \sim p$ be a random variable taking values over $\mathcal{X}$. Let $X^n = \{X_1, ..., X_n\}$ be a collection of n independent and identically distributed samples from $\mathcal{X}$. In this work we study rectangular confidence region (RCR) for the probability distribution p. An RCR of level $100(1-\alpha)\%$ for p is defined as a region $S(X^n)$ such that

$$P(p \in S(X^n)) \geq 1 - \alpha$$

where $S(X^n) = S_1(X^n), S_2(X^n), \ldots, S_m(X^n)$, and $S_i(X^n) = [a_i, b_i]$ for $i = 1, \ldots, m$. Further we naturally have that, $0 \leq a_i \leq b_i \leq 1$. In other words, an RCR for p in defined by a collection SCIs such that,

$$P(p \in S(X^n)) = P\left[\bigcap_{i=1}^{m} \{p_i \in S_i(X^n)\}\right] \geq 1 - \alpha$$

SCIs are very popular in multinomial inference, as they are intuitive and easy to interpret. In fact, almost all the inference schemes discussed in Section 2 are in fact SCIs. Obviously, there are many ways to construct SCIs that satisfy the above (for example, $a_i = 0, b_i = 1$ for every i). We are interested in minimal volume SCIs that satisfy the prescribed coverage rate. As mentioned above, we focus on the large alphabet regime, where m is relatively large, compared to n.

4 Methodology

Our proposed method utilizes bootstrap sampling to construct SCIs. Specifically, we focus on the plug-in principle, with several adaptations that account for the large alphabet regime. We begin this section with a brief overview of the bootstrap paradigm for inference problems.

4.1 Bootstrap Confidence Intervals

Bootstrap sampling is a popular approach for inference problems, especially in cases where the underlying distribution is involved or too complicated to analyze. By using the bootstrap principle we typically avoid the assumptions needed in classical analytical inference [1]. Bootstrap sampling was introduced by Efron in the early 1980s [5] and is considered a favorable framework with many applications.

Statistical inference problems often involve estimating a statistic of the underlying probability distribution $\phi(p)$. In some cases, the statistic of interest cannot be estimated directly from the sample. A typical example is the variance of the average of a sample, as later discussed. The bootstrap principle suggests that different statistics may be estimated by repeated sampling from the given sample. Specifically, the classical bootstrap *plug-in principle* utilizes an estimate $\hat{p}$ of the underlying distribution for this purpose. The plug-in principle suggests the following framework. Given a sample of n observations from p, we evaluate an estimate $\hat{p}$, and repeatedly sample n observations from it. Then, we numerically evaluate $\phi(\hat{p})$ from the bootstrap samples. For example, assume we are interested in the variance of the average of a sample. We collect n observations from p and evaluate $\hat{p}$. Then, we draw n observations from $\hat{p}$. We repeat this process B times and evaluate the average of each drawn sample. Finally, we compute the variance of these averages over the B bootstrap samples.

We distinguish between two main bootstrap schemes. The first is the nonparametric bootstrap, where a sample of size n is sampled with replacement from the data. Notice that this is equivalent to sampling n observations from the empirical distribution (or alternatively, treating the empirical distribution as a plug-in). The second approach is the parametric bootstrap. Here, we assume that the underlying distribution is parametric, with unknown parameters. We estimate the parameters from the sample and consider the resulting distribution as a plug-in [1].

In this work we study SCIs for multinomial proportions. Here, we briefly review several well known bootstrap schemes for constructing CI of different parameters. Let p be an unknown distribution and denote θ as a parameter of interest. Let x be a drawn sample of n observations from p (notice we omit the upper-script n for brevity). Denote $\hat{p}(x)$ as the empirical distribution of the sample. Let x^* be a bootstrap sample (of size n) from x. The *percentile method* is perhaps the simplest scheme for constructing a bootstrap CI for θ. Specifically, for every bootstrap sample x^* we evaluate a corresponding estimate of the unknown parameter. Then, we evaluate a distribution of these estimators. Finally, the CI is defined by the quantiles of this bootstrap distribution, denoted $[\theta^*_{\alpha/2}, \theta^*_{1-\alpha/2}]$. The *reverse percentile* is a similar approach to the percentile scheme discussed above, which introduces several favorable properties [13]. The reverse percentile utilizes the bootstrap quantiles to construct the interval $[2\hat{\theta} - \theta^*_{1-\alpha/2}, 2\hat{\theta} - \theta^*_{\alpha/2}]$ where $\hat{\theta}$ is an estimate of θ from the original sample x.

Both the percentile and the reverse percentile CIs are easy to implement and generally work well, but tend to fail in cases where the bootstrap distribution is asymmetric. To account for

the asymmetry, Efron proposed the *bias-corrected and accelerated* (BCa) method that adjusts for both the bias and the skewness in the bootstrap distribution [5].

An additional bootstrap CI scheme is the *studentized bootstrap* method, also known as the *bootstrap-t*. Bootstrap-t CIs are evaluated similarly to the standard Student's CIs. They are formally defined as $[\widehat{\theta} - t^*_{(1-\alpha/2)} \cdot \widehat{\text{se}}_\theta, \widehat{\theta} - t^*_{(\alpha/2)} \cdot \widehat{\text{se}}_\theta]$ where t^*_a denotes the a percentile of the bootstrapped Student's test, $t^* = (\widehat{\theta}^* - \widehat{\theta})/\widehat{\text{se}}_{\widehat{\theta}^*}$, while $\widehat{\theta}^*$ is an estimate of θ from the bootstrap sample and $\widehat{\text{se}}_\theta$ is the estimated standard error of θ in the original model [5]. The studentized bootstrap procedure is a useful generalization of the classical Student-t CI. However, as stated by Efron [5], it might result in somewhat erratic results and can be heavily influenced by a few outlying data points. On the other hand, the percentile based methods are considered more reliable [5].

Several variations which consider SCI for multiple parameters were also considered over the years. For example, Mandel and Betensky [16] derived an algorithm which constructs SCI by assigning ranks to the bootstrap samples and basing the SCI on the quantiles of the ranks with the percentile method. It is important to emphasize this scheme considers an arbitrary set of parameters whereas our problem of interest focuses on a set of parameters over the unit simplex (multinomial parameters).

Constructing bootstrap CIs for multinomial proportions is not an immediate task. This problem becomes more complicated in the large alphabet regime, where many symbols are not sampled at all. Notice that in this case, a naive plug-in approach would assign zero probability to unobserved symbols and a corresponding zero length CI. In this work we introduce a new CI estimation scheme which utilizes a parametric bootstrap sample, using the GT probability estimation as the plug-in distribution.

4.2 Good–Turing as a Plug-in

The Good-Turing scheme is perhaps the most popular approach for estimating large alphabet probability distributions [22]. Therefore, it is a natural choice as a bootstrap plug-in distribution. Specifically, given a sample of n observations from the multinomial distribution p, we apply the GT estimator (for example, following (2)) to obtain $\hat{p}_{GT}$. Then, we sample B bootstrap samples of size n from $\hat{p}_{GT}$ and construct corresponding SCIs following one of the methods discussed above (for example, the percentile method).

Unfortunately, this somewhat direct approach fails to obtain the prescribed coverage rate. Specifically, we observe that the obtained SCIs mostly fail to cover the symbols that do not appear in the sample. Therefore, we propose a different bootstrap approach, which allows a special treatment to the unobserved symbols. Our suggested framework, which we discussed in detail in the following section, distinguishes between two sets of symbols. Specifically, we construct two types of CIs. The first is a bootstrap CI for symbols that do not appear in the sample. The second is a Bonferroni corrected (analytical) CI for all the symbols that do appear in the sample. We show that by controlling these CIs simultaneously, we obtain an RCR that controls the prescribed coverage rate while introducing smaller volume than alternatives.

4.3 The Good-Bootstrap Algorithm

As mentioned above, unobserved symbols pose an inherent challenge. Therefore, we treat these symbols separately from the observed symbols. Let $N_i(X^n)$ be the number of appearances of the i^{th} symbol in the sample. Let

$$p_{max}(X^n) = \max_i \{p(i) \mid N_i = 0\}. \tag{3}$$

be the maximal probability among all unobserved symbols. Our first goal is to provide a CI for $p_{max}(X^n)$. Namely, we are interested in $T(X^n)$ such that

$$P(p_{max}(X^n) \leq T(X^n)) \geq 1 - \delta, \tag{4}$$

for a prescribed confidence level δ. Unfortunately, the distribution of $p_{max}(X^n)$ is quite difficult to analyze. Hence, we turn to a precentile bootstrap CI, using GT as a plug-in distribution. Our suggested scheme works as follows. Given a sample X^n, we evaluate the GT estimator, $\hat{p}_{GT}$. (for example, following (2)). Then, we sample n symbols from $\hat{p}_{GT}$ and evaluate p^*_{max}, the maximal probability over all the unobserved symbols in the bootstrap sample. We repeat this process B times and obtain a collection of p^*_{max} values. Finally,

we set the desired CI as $[0, T(X^n)]$ where $T(X^n)$ is the $1-\delta$ percentile of the bootstrapped p^*_{max}. Notice that the above scheme also applies to other bootstrap CI methods (as discussed in Section 2). Here, we focus on the simple percentile method for simplicity. Next, we proceed to the observed symbols. Here, we construct a simple binomial CI (for example, using Clopper-Pearson [3] interval) for every symbol that appear in the sample.

Finally, we apply a Bonferonni correction to obtain the desired confidence level simultaneously. Specifically, let $S_i(X^n)$ be a binomial CI with a confidence level of $\alpha/(m+1)$. For the simplicity of the derivation, we set $S_i(X^n) = [0, 1]$ for all the unobserved symbols as well. Further, let $T(X^n)$ be a (bootstrap) CI for p_{max}, with a confidence level of $\alpha/(m+1)$. Then,

$$P\left(\{\cup_i\{p_i \notin S_i(X^n)\}\} \cup \{p_{max} \geq T(X^n)\}\right) \leq$$
$$P(\cup_i\{p_i \notin S_i(X^n)\}) + P(p_{max} \geq T(X^n)) \leq$$
$$\frac{m}{m+1}\alpha + \frac{1}{m+1}\alpha = \alpha.$$

This means that with probability $1-\alpha$, we simultaneously have that

- The parameters of the observed symbols are covered by their corresponding binomial CI.
- The parameters of the unobserved symbols are covered by $[0, T(X^n)]$.

We denote our suggested scheme as the *Good-Bootstrap* SCI. Our proposed scheme is summarized in Algorithm 1 below.

Importantly, we further show that the bootstrapped $T(X^n)$ may be obtain analytically, without any repeated bootstrap sampling. That is, given a distribution p and a collection of samples $X^n \sim p$, we may obtain an analytical form for the distribution of $p_{max}(X^n)$ and henceforth (4). The crux of our analysis is that $p_{max}(X^n)$ takes values over a finite set, p. Assume without loss of generality that the m symbols are sorted in a ascending order, according to their corresponding probabilities. Then,

$$P(p_{max}(X^n) = p(i)) = \quad (5)$$
$$P(N_i = 0, N_{i+1} > 0, \ldots, N_m > 0).$$

This expression may be evaluated by recursively applying the Bayes rule, as shown in Appendix A.

Algorithm 1 Good-Bootstrap SCIs

Input: A sample X^n, alphabet size m and a confidence level α.

1: Set the GT estimator $\hat{p}_{GT}(X^n)$.
2: **for** $b = 1, \ldots, B$ **do**
3: Sample n observations from $\hat{p}_{GT}(X^n)$ and evaluate p^*_{max} (see (3)).
4: **end for**
5: Define $T(X^n)$ as the $1-\alpha/(m+1)$ percentile of the collection p^*_{max}, obtained in Step 2.
6: Set $[0, T(X^n)]$ as a CI for every unobserved symbol in X^n.
7: Set a Clopper-Pearson binomial CI of level $\alpha/(m+1)$, for every observed symbol in X^n.

5 Experiments

We now illustrate the performance of our proposed SCI in synthetic and real-world experiments. Figure 1 describes three synthetic example distributions, which are common benchmarks for related problems [22]. The Zipf's law distribution is a typical benchmark in large alphabet probability estimation; it is a commonly used heavy-tailed distribution, mostly for modeling natural (real-world) quantities in physical and social sciences, linguistics, economics and others fields [27]. The Zipf's law distribution follows $p(i;\ s, m) = i^{-s}/\sum_{i=1}^{m} i^{-s}$ for alphabet size m, where s is a skewness parameter. In our experiments we consider two different values of s, namely $s = 1.01$ and $s = 1.5$. Additional example distributions are the uniform, $p(i) = 1/m$, and the step distribution, with half of the symbols proportional to $1/2m$ while the other half are proportional to $3/2m$. Our simulations consider alphabet sizes m in the range of $[100, 7000]$ and sample sizes n within $[100, 5000]$ observations. We set $\alpha = 0.05$ in all of our experiments.

We compare the performance of the Good-Bootstrap scheme with two alternative methods. The first is Sison-Glaz (SG) which is discussed in Section 2. The second is a Bonferroni corrected SCI which utilizes the rule-of-three (ROT) for unobserved symbols (as described in detail in Section 2). Although there exist additional SCI scheme (Section 2), they are omitted from our report as being non competitive [12, 26] or computationally infeasible for large alphabets [2].

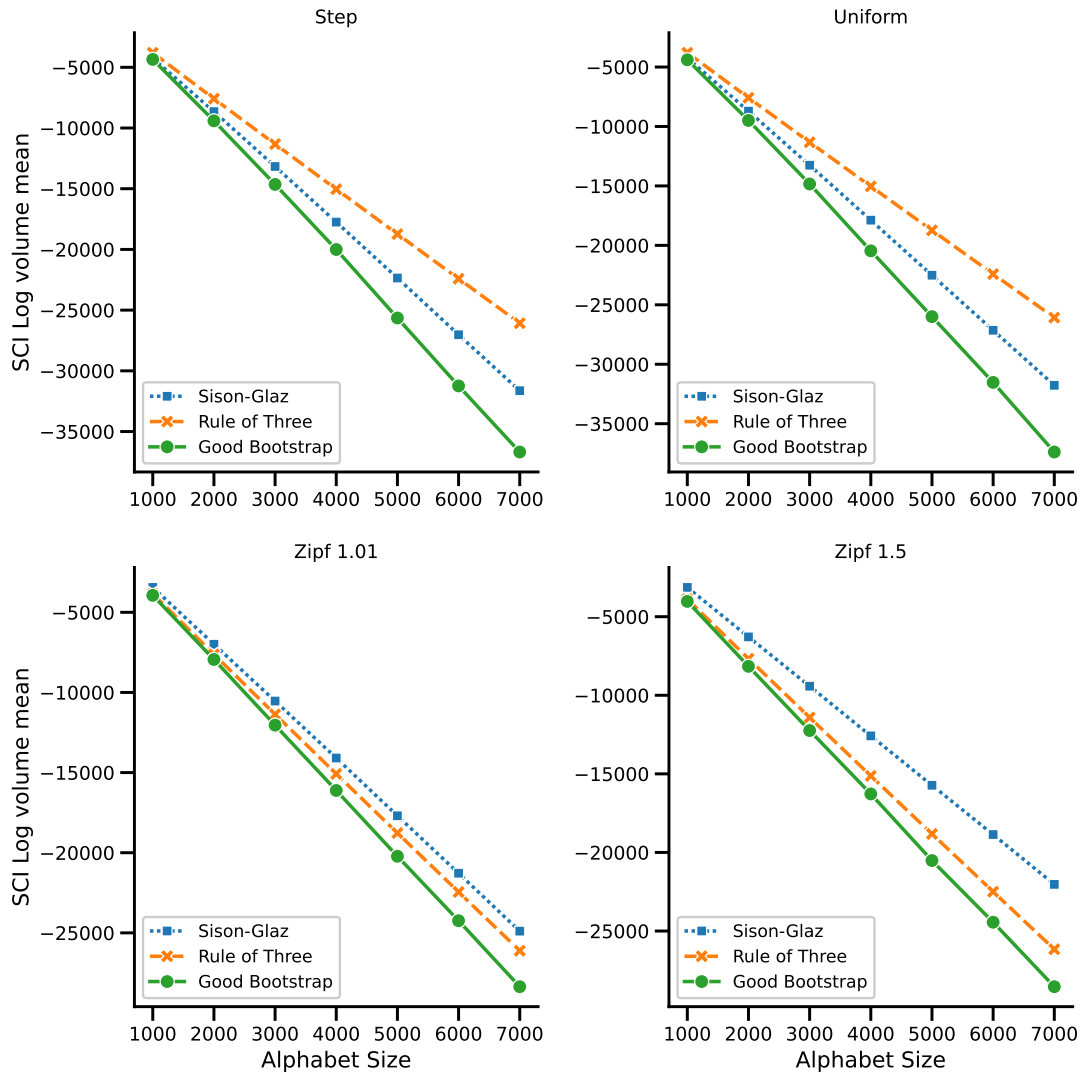

Fig. 1 Simulation results for sample size of 500 and an increasing alphabet size, averaged over 100 trials. Each plot represents different distribution, Step distribution (upper left), Uniform (upper right), Zipf's Law with $s = 1.01$ (lower left), and Zipf's Law with $s = 1.5$ (lower right)

Figure 1 demonstrates the log-volume of the examined SCIs for a sample size of $n = 500$ and an increasing alphabet size. As we can see, our proposed Good-Bootstrap scheme outperforms both alternative methods, in all the examined distributions and alphabet sizes. In addition, Figure 2 demonstrates the coverage of the there SCI schemes. As we can see, both the Good-Bootstrap and the Bonferroni corrected ROT methods maintain the desired confidence level, while SG fails to do so. This is not quite surprising, as discussed in Section 2. Additional examination of the SCIs coverage for larger experimental setting is provided in Appendix B.

Next, we examine the obtained CI for different symbols in the alphabet. Figure 3 demonstrates the length of the CI for every symbol in the Zipf's Law experiment ($s = 1.01$). Importantly, notice that the grayed area corresponds to the collection of symbols with 95% of the probability mass. As we can see, the Good-Bootstrap scheme outperforms both alternatives in this region.

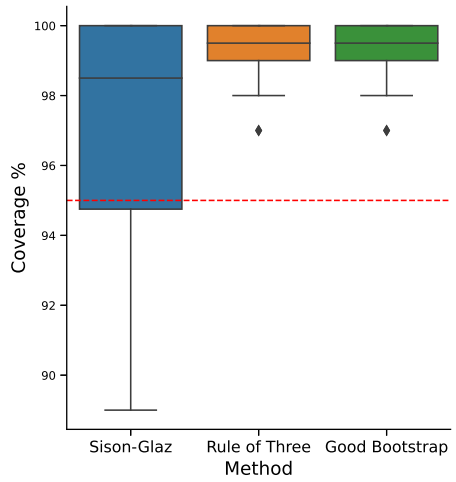

Fig. 2 Coverage rate of the different methods out of 100 trials. The red dashed line indicates the desired level

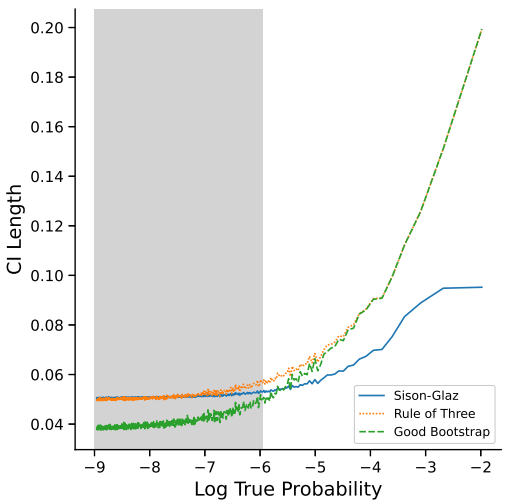

Fig. 3 CI length of the different methods by log true probability of symbols in a zipf's Law experiment ($s = 1.01$). The Gray area corresponds to 95% percent of the total probability mass

Finally, we proceed to real-world experiments. The Biota data-set is a computational biology sample which considers the forearm skin biota of six subjects. It contains a total of 1221 clones consisting of 182 different species-level operational taxonomic units (SLOTUs) [9]. The Hamlet data-set is a linguistic collection which considers the frequency of approximately 5000 distinct words in the classical Hamlet play. Notice that in these real-world settings, the true underlying probability is unknown. Hence, the underlying distribution p refers to the total frequency of symbols, in the full data-set. We examine the three SCI schemes for different sample sizes. Similarly to the above, we compare their corresponding confidence region volumes and evaluate their coverage rate. Figure 4 summarizes the results we achieve. As above, the Good-Bootstrap scheme demonstrate improved performance compared to the alternatives.

6 Discussion and Conclusions

In this work we introduce a new SCI scheme for large alphabet multinomial proportions. The proposed Good-Bootstrap scheme is based on a bootstrap statistic, $p_{max}(X^n) = \max_i\{p(i) \mid N_i = 0\}$, which corresponds to the maximal probability over all the symbols that do not appear in the sample. This statistic is utilized to construct a CI for all the symbols that do not appear in the sample, while the remaining symbols are treated with a Bonferroni corrected binomial CI. We examine our proposed method on synthetic and real-world data, showing it outperforms popular alternatives in large alphabet regimes. We further show that the Good-Bootstrap scheme maintains the desired coverage level, as expected. To the best of our knowledge, our method is the first to address multinomial intervals estimation in a large alphabet regime. A Python implementation of the Good-Bootstrap scheme is publicly available at [17].

Our proposed algorithm provides a special treatment to unobserved symbols. However, it could be generalized to consider symbols that appear once, twice or more times in the sample, in a similar manner. This may further reduce the volume of the obtained SCIs in the studied regime. An additional improvement may be obtained in the binomial treatment of the sampled symbols. Currently, we apply the Clopper-Pearson method, which is considered conservative. Using a less conservative method may result in smaller CIs while still maintaining the desired confidence level. We consider these directions for our future work.

Appendix A

Given a sample X^n and a distribution p, we introduce an exact analytical expression for (4). Assume without loss of generality that the m symbols are sorted in a ascending order, according to

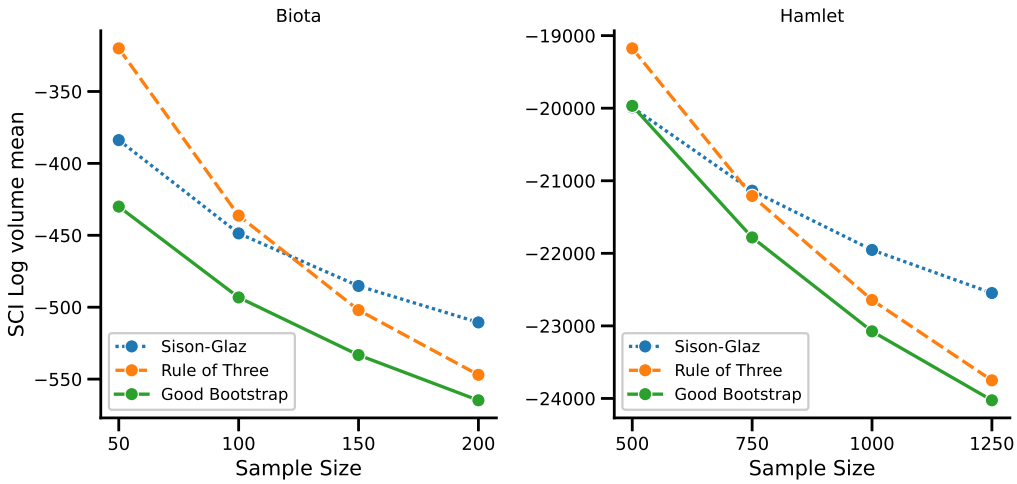

Fig. 4 Real-world experiments. The Biota data-set (left) and the Hamlet play (right)

their corresponding probabilities $p(i)$. Then, we evaluate the distribution of $p_{max}(X^n)$ as follows:

$$P(p_{max}(X^n) = p(i)) = \tag{A1}$$
$$P(N_i = 0, N_{i+1} > 0, \ldots, N_m > 0) =$$
$$P(N_i = 0)P(N_{i+1} > 0 | N_i = 0) \cdot$$
$$P(N_{i+2} > 0 | N_{i+1} > 0, N_i = 0) \cdot \ldots$$
$$P(N_m > 0 | N_{m-1} > 0, \ldots, N_i = 0)$$

Let us derive each of the terms above. First we have $P(N_i = 0) = (1 - p(i))^n$. Next,

$$P(N_j > 0 | N_i = 0) = 1 - P(N_j = 0 | N_i = 0) =$$
$$1 - \left(1 - \frac{p(j)}{1 - p(i)}\right)^n, \tag{A2}$$

for every $j \neq i$. Similarly, we have

$$P(N_j > 0 | N_i = 0, N_t = 0) =$$
$$1 - P(N_j = 0 | N_i = 0, N_t = 0) =$$
$$1 - \left(1 - \frac{p(j)}{1 - p(i) - p(t)}\right)^n, \tag{A3}$$

which we use for our following derivations. Further,

$$P(N_w > 0 | N_j > 0, N_i = 0) = \tag{A4}$$
$$\sum_{k=1}^{n} P(N_w > 0 | N_j > 0, N_i = 0, N_j = k) \cdot$$
$$P(N_j = k | N_j > 0, N_i = 0) =$$
$$\sum_{k=1}^{n} P(N_w > 0 | N_i = 0, N_j = k) \cdot$$
$$\frac{P(N_j > 0 | N_j = k, N_i = 0)}{P(N_j > 0 | N_i = 0)} \cdot P(N_j = k | N_i = 0) =$$
$$\frac{1}{P(N_j > 0 | N_i = 0)} \cdot$$
$$\sum_{k=1}^{n} P(N_w > 0 | N_i = 0, N_j = k) P(N_j = k | N_i = 0) =$$
$$\frac{1}{P(N_j > 0 | N_i = 0)} (P(N_w > 0 | N_i = 0) -$$
$$P(N_w > 0 | N_i = 0, N_j = 0) P(N_j = 0 | N_i = 0))$$

where all the terms that are used in (A4) may be evaluated according to (A3) and (A2). This means we may compute each of the terms in (A1) recursively, one after the other. Specifically,

$$P(N_t > 0 | N_{t-1}, \ldots, N_{t-l} > 0, N_{t-l-1} = 0) =$$
$$\sum_{k=1}^{n} P(N_t > 0 | N_{t-1} = k, N_{t-2} > 0, \ldots,$$
$$N_{t-l} > 0, N_{t-l-1} = 0) \cdot$$
$$P(N_{t-1} = k | N_{t-2} > 0, .., N_{t-l} > 0, N_{t-l-1} = 0) =$$
$$\sum_{k=1}^{n} P(N_t > 0 | N_{t-1} = k, N_{t-2} > 0, \ldots,$$
$$N_{t-l} > 0, N_{t-l-1} = 0) \cdot$$
$$P(N_{t-1} > 0 | N_{t-1} = k, N_{t-2} > 0, \ldots,$$
$$N_{t-l-1} = 0) \cdot$$

$$\frac{P(N_{t-1}=k|N_{t-2}>0,\ldots,N_{t-l-1}=0)}{P(N_{t-1}>0|N_{t-2}>0,\ldots,N_{t-l-1}=0)} = \frac{1}{P(N_{t-1}>0|N_{t-2}>0,\ldots,N_{t-l-1}=0)}\cdot \sum_{k=1}^{n} P(N_t>0|N_{t-1}=k,N_{t-2}>0,\ldots, N_{t-l-1}=0)\cdot P(N_{t-1}=k|N_{t-2}>0,\ldots,N_{t-l-1}=0) = \frac{1}{P(N_{t-1}>0|N_{t-2}>0,\ldots,N_{t-l-1}=0)}\cdot \Big(P(N_t>0|N_{t-2}>0,\ldots,N_{t-l-1}=0) - P(N_t>0|N_{t-1}=0,N_{t-2}>0,\ldots,N_{t-l-1}=0)\cdot P(N_{t-1}=0|N_{t-2}>0,\ldots,N_{t-l-1}=0)\Big) \quad \text{(A5)}$$

where:

- $P(N_{t-1}>0|N_{t-2}>0,\ldots,N_{t-l-1}=0)$ is evaluated according to the previous step.
- $P(N_t>0|N_{t-1}=0,N_{t-2}>0,N_{t-l-1}=0)$ is similar to the previous expression, but since we are given that $N_{t-1}=0$, the conditional probability satisfies $1-\left(1-\frac{p(t)}{1-p(t-1)}\right)^n$.

Appendix B

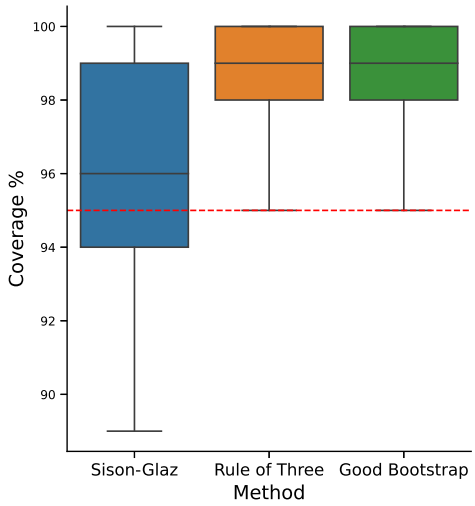

Fig. B1 Coverage rate of the different methods, out of 100 trials. Alphabet sizes of 100, 500, 1000, 5000, and sample sizes of 100, 500, 1000, 5000.

Acknowledgements

This research is supported by the Israel Science Foundation grant number 963/21.

Compliance with Ethical Standards

The authors declare having no competing interests as defined by Springer, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

References

[1] Carpenter, J. and J. Bithell. 2000. Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. *Statistics in medicine 19*(9): 1141–1164 .

[2] ChafaÃ¯, D. and D. Concordet. 2009. Confidence regions for the multinomial parameter with small sample size. *Journal of the American Statistical Association 104*(487): 1071–1079 .

[3] Clopper, C.J. and E.S. Pearson. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika 26*(4): 404–413 .

[4] Dunn, O.J. 1961. Multiple comparisons among means. *Journal of the American statistical association 56*(293): 52–64 .

[5] Efron, B. and R.J. Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

[6] Eypasch, E., R. Lefering, C. Kum, and H. Troidl. 1995. Probability of adverse events that have not yet occurred: a statistical reminder. *Bmj 311*(7005): 619–620 .

[7] Fitzpatrick, S. and A. Scott. 1987. Quick simultaneous confidence intervals for multinomial proportions. *Journal of the American Statistical Association 82*(399): 875–878 .

[8] Gale, W.A. and G. Sampson. 1995. Good-Turing frequency estimation without tears. *Journal of quantitative linguistics 2*(3): 217–237 .

[9] Gao, Z., C.h. Tseng, Z. Pei, and M.J. Blaser. 2007. Molecular analysis of human forearm superficial skin bacterial biota. *Proceedings of the National Academy of Sciences 104*(8): 2927–2932 .

[10] Good, I.J. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika* 40: 237–264 .

[11] Good, I.J. 2000. Turing's anticipation of empirical bayes in connection with the cryptanalysis of the naval enigma. *Journal of Statistical Computation and Simulation 66*(2): 101–111 .

[12] Goodman, L.A. et al. 1964. Simultaneous confidence intervals for contrasts among multinomial populations. *The Annals of Mathematical Statistics 35*(2): 716–725 .

[13] Hesterberg, T.C. 2015. What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The american statistician 69*(4): 371–386 .

[14] Hou, C.D., J. Chiang, and J.J. Tai. 2003. A family of simultaneous confidence intervals for multinomial proportions. *Computational statistics & data analysis 43*(1): 29–45 .

[15] Laplace, P.S. 1825. *Pierre-Simon Laplace Philosophical Essay on Probabilities: Translated from the fifth French edition of 1825 With Notes by the Translator*, Volume 13. Springer Science.

[16] Mandel, M. and R.A. Betensky. 2008. Simultaneous confidence intervals based on the percentile bootstrap approach. *Computational statistics & data analysis 52*(4): 2158–2165 .

[17] Marton, D. https://github.com/DanielMarton/Good-Bootstrap.

[18] May, W.L. and W.D. Johnson. 1997. Properties of simultaneous confidence intervals for multinomial proportions. *Communications in Statistics 26*(2): 495–518 .

[19] McAllester, D.A. and R.E. Schapire 2000. On the convergence rate of Good-Turing estimators. In *COLT*, pp. 1–6.

[20] Nadas, A. 1985. On turing's formula for word probabilities. *IEEE Transactions on Acoustics, Speech, and Signal Processing 33*(6): 1414–1416 .

[21] Orlitsky, A., N.P. Santhanam, and J. Zhang. 2003. Always Good Turing: Asymptotically optimal probability estimation. *Science 302*(5644): 427–431 .

[22] Orlitsky, A. and A.T. Suresh 2015. Competitive distribution estimation: Why is Good-Turing good. In *Advances in Neural Information Processing Systems*, pp. 2143–2151.

[23] Painsky, A. 2021. Refined convergence rates of the good-turing estimator. In *2021 IEEE Information Theory Workshop (ITW)*, pp. 1–5.

[24] Painsky, A. 2022a. A data-driven missing mass estimation framework. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pp. 2991–2995. IEEE.

[25] Painsky, A. 2022b. Generalized good-turing improves missing mass estimation. *Journal of the American Statistical Association*: 1–10 .

[26] Quesenberry, C.P. and D. Hurst. 1964. Large sample simultaneous confidence intervals for multinomial proportions. *Technometrics 6*(2): 191–195 .

[27] Saichev, A.I., Y. Malevergne, and D. Sornette. 2009. *Theory of Zipf's law and beyond*, Volume 632. Springer Science.

[28] Sison, C.P. and J. Glaz. 1995. Simultaneous confidence intervals and sample size determination for multinomial proportions. *Journal of the American Statistical Association 90*(429): 366–369 .

[29] Withers, C.S. and S. Nadarajah. 2017. A new confidence region for the multinomial distribution. *Communications in Statistics-Simulation and Computation 46*(5): 4113–4126 .

[30] Zabell, S.L. 1989. The rule of succession. *Erkenntnis 31*(2): 283–321 .