Asymptotic Properties of Turing's Formula in Relative Error

Michael Grabchak*and Zhiyi Zhang
University of North Carolina Charlotte

November 24, 2016

Abstract

Turing's formula allows one to estimate the total probability associated with letters from an alphabet, which are not observed in a random sample. In this paper we give conditions for the consistency and asymptotic normality of the relative error of Turing's formula of any order. We then show that these conditions always hold when the distribution is regularly varying with index $\alpha \in (0, 1]$.

AMS 2010 subject classifications: 62G20; 62G32

Keywords: asymptotic normality; consistency; distributions on alphabets; missing mass; regular variation; Turing's formula

1 Introduction

In many situations one works with data that has no natural ordering and is categorical in nature. In such cases, an important problem is to estimate the probability of seeing a new category that has not been observed before. This probability is called the missing mass. See [23, 3, 2, 9] and the references therein for a discussion of its properties. The problem of estimating the missing mass arises in many applications, including ecology [16, 5, 6], genomics [21], speech recognition [18, 7], authorship attribution [11, 28, 32], and computer networks [29]. Perhaps the most famous estimator of the missing mass is Turing's formula, sometimes also called the Good-Turing formula. This formula was first published by I. J. Good in [15], where the idea is credited, largely, to Alan Turing. To discuss this estimator and how it works, we begin by formally defining our framework.

Let $\mathcal{A} = (a_1, a_2, \dots)$ be a countable alphabet and let $\mathcal{P} = (p_1, p_2, \dots)$ be a probability distribution on $\mathcal{A}$. We refer to the elements in $\mathcal{A}$ as letters.

*Email address: mgrabcha@uncc.edu

These represent the various categories of our data. Assume that $X_1, \ldots, X_n$ is a random sample of size n from $\mathcal{A}$ according to $\mathcal{P}$, let

$$y_{k,n} = \sum_{i=1}^{n} 1_{[X_i = a_k]}$$

be the number of times that letter a_k appears in the sample, and let

$$N_{r,n} = \sum_{k} 1_{[y_{k,n} = r]}$$

be the number of letters observed exactly r times.

Define

$$\pi_{0,n} = \sum_{k} p_k 1_{[y_{k,n}=0]}.$$

This is the missing mass, i.e. the probability that the next observation will be of a letter that has not yet been observed. Turing's formula is an estimator of $\pi_{0,n}$ given by

$$T_{0,n} = \frac{N_{1,n}}{n}.$$

One way to see how well this estimator works is through simulations; for a recent simulation study see [17]. Other approaches, based on decision theory and Bayesian inference, are given in [8] and [13]. A simpler approach is to consider the bias of Turing's formula and to see when

$$\mathrm{E}\left[T_{0,n} - \pi_{0,n}\right] \approx 0.$$

Many discussions are given in, e.g., [27, 22, 32]. Along similar lines, it can be easily shown that

$$(T_{0,n} - \pi_{0,n}) \xrightarrow{p} 0 \text{ as } n \to \infty. \tag{1}$$

While this may seem to resolve the issue of consistency, it is not as informative as it first appears. This is because, as is easy to see, we have both $T_{0,n} \xrightarrow{p} 0$ and $\pi_{0,n} \xrightarrow{p} 0$ as $n \to \infty$. Thus, we cannot know if (1) tells us that $T_{0,n}$ is estimating $\pi_{0,n}$ well, or if both are just very small. This is an issue with the studies of bias as well.

A different and, arguably, more meaningful way to think about consistency was introduced in [25]. There, the question under consideration was when does the relative error of Turing's formula approach zero, i.e. when does

$$\frac{T_{0,n} - \pi_{0,n}}{\pi_{0,n}} \xrightarrow{p} 0 \text{ as } n \to \infty \tag{2}$$

hold. Specifically, [25] and [2] showed that a sufficient condition for (2) is that the underlying distribution, $\mathcal{P}$, is regularly varying.

Another approach to understanding and using Turing's formula is to ask when asymptotic normality holds. In particular, conditions under which there exists a deterministic sequence g_n with

$$g_n \left(T_{0,n} - \pi_{0,n}\right) \xrightarrow{d} N(0,1) \text{ as } n \to \infty \tag{3}$$

are given in [12, 33, 30, 31]. Results of this type are important for constructing statistical tests and confidence intervals. See [32] for an application to authorship attribution.

In this paper we give conditions for the asymptotic normality of the relative error of Turing's formula. Specifically, we give conditions under which there exists a deterministic sequence h_n with

$$h_n \frac{T_{0,n} - \pi_{0,n}}{\pi_{0,n}} \xrightarrow{d} N(0,1) \text{ as } n \to \infty. \tag{4}$$

These conditions also imply consistency of the relative error in the sense of (2). We note that (4) leads to confidence intervals and hypothesis tests quite different from those determined by (3). The nature of this difference will be studied in a future work. All of our results are presented not just for Turing's formula, but also for higher order Turing's formulae, which we discuss in the next section.

We will prove (4) under a new sufficient condition. Interestingly, this condition turn out to be more restrictive than the one for (3). This is likely due to the fact that, since we are now dividing by $\pi_{0,n}$, we must make sure that it does not approach zero too quickly. We note that our approach is quite different from the one used in [25] to prove (2). That approach uses tools specific to regularly varying distributions, which we will not need.

The remainder of the paper is organized as follows. In Section 2 we recall Turing's formulae of higher orders and give conditions for the asymptotic normality and consistency of their relative errors. In Section 3, we give some comments on the assumptions of the main results and give alternate ways of checking them. Then, in Section 4, we show that the assumptions always hold when the distribution, $\mathcal{P}$, is regularly varying with index $\alpha \in (0, 1]$. For $\alpha \neq 1$ this is the condition under which the consistency results of [25] were obtained. Finally, proofs of the main results are given in Section 5.

Before proceeding we introduce some notation. For $x > 0$ we denote the gamma function by $\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} \mathrm{d}t$. For real valued functions f and g, we write $f(x) \sim g(x)$ as $x \to c$ to mean $\lim_{x \to c} \frac{f(x)}{g(x)} = 1$. For sequences a_n and b_n, we write $a_n \sim b_n$ to mean $\lim_{n \to \infty} \frac{a_n}{b_n} = 1$. We write $N(\mu, \sigma^2)$ to refer to a normal distribution with mean μ and variance σ^2. We write $\xrightarrow{d}$ to refer to convergence in distribution and $\xrightarrow{p}$ to refer to convergence in probability.

2 Main Results

In this section we give our main results about the asymptotic normality and consistency of the relative error of Turing's formula. We begin by recalling higher order Turing's formulae, which were introduced in [15]. For any $r = 0, 1, \ldots, n-1$ let

$$\pi_{r,n} = \sum_k p_k 1_{[y_{k,n}=r]}$$

be the probability that the next observation will be of a letter that has been observed exactly r times, and let

$$\mu_{r,n} := \mathrm{E}[\pi_{r,n}] = \binom{n}{r} \sum_k p_k^{r+1}(1-p_k)^{n-r}.$$

An estimator of $\pi_{r,n}$ is given by

$$T_{r,n} = \frac{r+1}{n-r} N_{r+1,n},$$

which is Turing's formula of order r. Turing's formula of order 0 is just called Turing's formula and is the most useful in applications as it estimates the probability of seeing a letter that has never been observed before. We now recall the results about asymptotic normality given in [33] and [31].

Let g_n be a deterministic sequence of positive numbers such that

$$\limsup_{n\to\infty} \frac{g_n}{n^{1-\beta}} < \infty \text{ for some } \beta \in (0, 1/2). \tag{5}$$

For an integer s, we say that Condition A_s is satisfied if

$$\lim_{n\to\infty} g_n^2 n^{s-2} \sum_k p_k^s (1-p_k)^{n-s} = c_s \tag{6}$$

for some $c_s \geq 0$. The following result is given in [31] and for the case $r = 0$ in [33].

Lemma 1. *Fix any integer $r \geq 0$ and let g_n be a deterministic sequence of positive numbers satisfying (5). If Conditions A_{r+1} and A_{r+2} hold with $c_{r+1} + c_{r+2} > 0$, then*

$$g_n \left(T_{r,n} - \pi_{r,n}\right) \xrightarrow{d} N\left(0, \frac{(r+1)c_{r+1} + c_{r+2}}{r!}\right) \text{ as } n \to \infty.$$

We are now ready to state our main result.

Theorem 1. *Fix any integer $r \geq 0$ and let g_n be a deterministic sequence of positive numbers satisfying (5). If $r \geq 2$ assume that*

$$\limsup_{n\to\infty} g_n^2 \sum_k p_k^2 (1-p_k)^{n-2} < \infty. \tag{7}$$

If Conditions A_{r+1} and A_{r+2} hold with $c_{r+1} > 0$ and $c_{r+2} \geq 0$, then

$$\mu_{r,n} g_n \left(\frac{T_{r,n} - \pi_{r,n}}{\pi_{r,n}} \right) \xrightarrow{d} N\left(0, \frac{(r+1)c_{r+1} + c_{r+2}}{r!}\right) \text{ as } n \to \infty.$$

Proof. The proof is given in Section 5. □

Remark 1. *Note that Theorem 1 does not, in general, give $\sqrt{n}$-convergence. In fact, the rate of convergence is different for different distributions. In Section 4 we will characterize the rates for the case of regularly varying distributions.*

Since the most important case is when $r = 0$, we restate Theorem 1 for this case.

Corollary 1. *Let g_n be a deterministic sequence of positive numbers satisfying* (5)*. If Conditions A_1 and A_2 hold with $c_1 > 0$ and $c_2 \geq 0$, then*

$$\mu_{0,n} g_n \left(\frac{T_{0,n} - \pi_{0,n}}{\pi_{0,n}} \right) \xrightarrow{d} N(0, c_1 + c_2) \text{ as } n \to \infty.$$

The results of Theorem 1 may not appear to be of practical use since we generally do not know the values of g_n, $\mu_{r,n}$, c_{r+1}, or c_{r+2}. However, it turns out that we do not need to know these quantities. So long as a sequence g_n satisfying the assumptions exists, it and everything else can be estimated.

Corollary 2. *If the conditions of Theorem 1 are satisfied, then*

$$\frac{\sqrt{r+1}\mathrm{E}[N_{r+1,n}]}{\sqrt{(r+1)\mathrm{E}[N_{r+1,n}] + (r+2)\mathrm{E}[N_{r+2,n}]}} \left(\frac{T_{r,n} - \pi_{r,n}}{\pi_{r,n}} \right) \xrightarrow{d} N(0,1)$$

and

$$\frac{\sqrt{r+1}N_{r+1,n}}{\sqrt{(r+1)N_{r+1,n} + (r+2)N_{r+2,n}}} \left(\frac{T_{r,n} - \pi_{r,n}}{\pi_{r,n}} \right) \xrightarrow{d} N(0,1).$$

Proof. The proof is given in Section 5. □

In the proof of Theorem 1, it is shown that, under the assumptions of that theorem, $\mu_{r,n} g_n \to \infty$. This means that we can immediately get consistency.

Corollary 3. *If the conditions of Theorem 1 are satisfied, then*

$$\frac{T_{r,n} - \pi_{r,n}}{\pi_{r,n}} \xrightarrow{p} 0. \tag{8}$$

The assumptions of Corollary 3 are quite general. Different conditions are given in Corollary 5.3 of [2]. The most general possible conditions for (8) are not known, but it is known is that some conditions are necessary. In fact, [24] showed that there cannot exist an estimator of $\pi_{0,n}$ for which (8) holds for every distribution.

3 Discussion

In this section we discuss the assumptions of Theorem 1 and give alternate ways to verify them. We begin by giving several equivalent ways to check that Condition A_s holds. Toward this end, we introduce the notation

$$\Phi_s(n) = \frac{n^s}{s!} \sum_k p_k^s e^{-np_k}.$$

Lemma 2. *For any integer $s \geq 1$ the following are equivalent:*

$$\text{a)} \quad \lim_{n\to\infty} g_n^2 n^{s-2} \sum_k p_k^s (1-p_k)^{n-s} = c_s,$$

$$\text{b)} \quad \lim_{n\to\infty} (s-1)! \frac{g_n^2}{n} \mu_{s-1,n-1} = c_s,$$

$$\text{c)} \quad \lim_{n\to\infty} s! \frac{g_n^2}{n^2} \mathrm{E}[N_{s,n}] = c_s,$$

$$\text{d)} \quad \lim_{n\to\infty} s! \frac{g_n^2}{n^2} \Phi_s(n) = c_s.$$

Proof. The proof is given in Section 5. □

Remark 2. *An intuitive interpretation of $\Phi_s(n)$ is as follows. Consider the case where the sample size is not fixed at n, but is a random variable n^*, where n^* follows a Poisson distribution with mean n. In this case y_{k,n^*} follows a Poisson distribution with mean np_k and $\Phi_s(n) = \mathrm{E}[N_{s,n^*}]$. In this sense, Condition d) can be thought of as a Poissonization of Condition c). Poissonization is a useful tool when studying the occupancy problem and is discussed, at length, in [14].*

We now turn to the effects of Condition A_s.

Lemma 3. *1. Let $s \geq 1$ be an integer. If Condition A_s holds with $c_s > 0$ then*

$$\lim_{n\to\infty} \frac{g_n}{n^{1/2}} = \infty \tag{9}$$

and

$$\lim_{n\to\infty} \frac{g_n}{g_{n+1}} = 1. \tag{10}$$

2. When $r \geq 2$ and Condition A_{r+1} holds, (7) is equivalent to

$$\limsup_{n\to\infty} \sum_{k:p_k<1/n} (g_n p_k)^2 < \infty.$$

Proof. The proof is given in Section 5. □

It is important to note that (9) is implicitly used in the proof of Lemma 1 as given in [33] and [31], although it is not directly mentioned there. Further, (9) implies that the assumption in (5) that $\beta \in (0, 1/2)$ is not much of a restriction. It also tells us that g_n must approach infinity quickly, but (5) tell us that it should not do so too quickly. On the other hand, (10) is a smoothness assumption. It looks like a regular variation condition, but is a bit weaker, see Theorem 1.9.8 in [4].

Remark 3. *Lemmas 2 and 3 help to explain what kind of distributions satisfy the assumptions of Theorem 1. Specifically, the two lemmas imply that* $r!n^{-1}g_n^2\mu_{r,n} \to c_{r+1} > 0$. *In light of* (5)*, this means that* $\mu_{r,n}$ *cannot approach zero too quickly. Thus,* $\pi_{r,n}$ *cannot approach zero quickly either. This condition means that the distribution must have heavy tails of some kind. Arguably, the best known distributions with heavy tails are those that are regularly varying, which we focus on in the next section.*

4 Regular Variation

In this section we show that the assumptions of Theorem 1 are always satisfied when $\mathcal{P}$ is regularly varying. The concept of regular variation of a probability measure on an alphabet seems to have originated in the classical paper [19], see also [14] for a recent review. We begin by introducing the measure

$$\nu(\mathrm{d}x) = \sum_k \delta_{p_k}(\mathrm{d}x),$$

where δ_y denotes the Dirac mass at y, and the function

$$\bar{\nu}(x) = \nu([x, 1]) = \sum_k 1_{[p_k \geq x]}.$$

We say that $\mathcal{P}$ is regularly varying with index $\alpha \in [0, 1]$ if

$$\bar{\nu}(x) \sim \ell(1/x)x^{-\alpha} \text{ as } x \downarrow 0, \tag{11}$$

where ℓ is slowly varying at infinity, i.e. for any $t > 0$ it satisfies

$$\lim_{x\to\infty} \frac{\ell(xt)}{\ell(x)} = 1.$$

In this case we write $\mathcal{P} \in \mathcal{RV}_\alpha(\ell)$. When $\alpha = 0$ we say that $\mathcal{P}$ is slowly varying and when $\alpha = 1$ we say that it is rapidly varying. These cases have different behavior from the others and we will discuss them separately.

To better understand the meaning of regular variation, we recall the following result from [14]. It says that $\mathcal{P} \in RV_\alpha(\ell)$ with $\alpha \in (0, 1)$ if and only if

$$p_k \sim \ell^*(k)k^{-1/\alpha} \text{ as } k \to \infty,$$

where ℓ^* is a slowly varying at infinity function, in general, different from ℓ. We now state the main results of this section.

Proposition 1. *If $\mathcal{P} \in \mathcal{RV}_\alpha(\ell)$ for some $\alpha \in (0,1)$ then, for any integer $r \geq 0$,*

$$\kappa_\alpha n^{\alpha/2}[\ell(n)]^{1/2}\left(\frac{T_{r,n}-\pi_{r,n}}{\pi_{r,n}}\right) \xrightarrow{d} N(0,1) \text{ as } n \to \infty,$$

where $\kappa_\alpha = \sqrt{\frac{\alpha\Gamma(r+1-\alpha)}{r!(2r+2-\alpha)}}$.

Proof. Let $g_n = n^{1-\alpha/2}[\ell(n)]^{-1/2}$. Proposition 17 in [14] implies that for every integer $s \geq 1$

$$\lim_{n\to\infty} s!\frac{g_n^2}{n^2}\Phi_s(n) \quad = \quad \alpha\Gamma(s-\alpha).$$

By Lemmas 2 and 3 this implies that

$$\mu_{r,n} \sim \frac{\alpha}{r!}\Gamma(r+1-\alpha)n^{-(1-\alpha)}\ell(n).$$

Thus

$$\mu_{r,n}g_n \sim \frac{\alpha}{r!}\Gamma(r+1-\alpha)n^{\alpha/2}[\ell(n)]^{1/2}$$

and $(r+1)c_{r+1}+c_{r+2} = (r+1)\alpha\Gamma(r+1-\alpha)+\alpha\Gamma(r+2-\alpha) = \alpha(2r+2-\alpha)\Gamma(r+1-\alpha)$. From here the result follows by Theorem 1. □

Next, we turn to the case when $\alpha = 1$.

Proposition 2. *Assume that $\mathcal{P} \in \mathcal{RV}_1(\ell)$ and let*

$$\ell_1(y) = \int_y^\infty u^{-1}\ell(u)\mathrm{d}u.$$

If $r = 0$ then

$$n^{1/2}[\ell_1(n)]^{1/2}\left(\frac{T_{0,n}-\pi_{0,n}}{\pi_{0,n}}\right) \xrightarrow{d} N(0,1) \text{ as } n \to \infty,$$

and if $r \geq 1$ then

$$\kappa_1 n^{1/2}[\ell(n)]^{1/2}\left(\frac{T_{r,n}-\pi_{r,n}}{\pi_{r,n}}\right) \xrightarrow{d} N(0,1) \text{ as } n \to \infty,$$

where $\kappa_1 = [r(2r+1)]^{-1/2}$.

We note that the integral in the definition of ℓ_1 converges, see the proof of Proposition 14 in [14]. Further, by Karamata's Theorem (see e.g. Theorem 2.1 in [26]), ℓ_1 is slowly varying at infinity.

Proof. We begin with the case $r = 0$. In this case we let $g_n = n^{1/2}[\ell_1(n)]^{-1/2}$. Proposition 18 in [14] implies that

$$\lim_{n\to\infty} \frac{g_n^2}{n^2}\Phi_1(n) = 1 \text{ and } 2\Phi_2(n) \sim n\ell(n).$$

This means that $\mu_{0,n} \sim \ell_1(n)$ and that

$$2\frac{g_n^2}{n^2}\Phi_2(n) \sim \frac{\ell(n)}{\ell_1(n)} \to 0, \tag{12}$$

where the convergence follows by Karamata's Theorem, see e.g. Theorem 2.1 in [26]. Thus $\mu_{0,n}g_n \sim n^{1/2}[\ell_1(n)]^{1/2}$ and $c_1 + c_2 = 1$. From here the first part follows by Corollary 1.

Now assume that $r \geq 1$. In this case we let $g_n = n^{1/2}[\ell(n)]^{-1/2}$. Proposition 18 in [14] says that for $s \geq 2$

$$\lim_{n\to\infty} s!\frac{g_n^2}{n^2}\Phi_s(n) = (s-2)!,$$

which means that

$$\mu_{r,n} \sim \frac{\ell(n)}{r}.$$

This implies that $\mu_{r,n}g_n \sim r^{-1}\sqrt{n\ell(n)}$ and that $(r+1)c_{r+1} + c_{r+2} = (2r+1)(r-1)!$. Putting everything together and applying Theorem 1 gives the result. □

Remark 4. *From the proof of Proposition 2 we see that, when $\alpha = 1$ and $r \geq 1$, we have $g_n = n^{1/2}[\ell(n)]^{-1/2}$. Further, $\Phi_1(n) \sim n\ell_1(n)$, which means that*

$$\frac{g_n^2}{n^2}\Phi_1(n) \sim \frac{\ell_1(n)}{\ell(n)} \to \infty,$$

where the convergence follows by (12)*. Thus condition A_1 fails to hold. However, Conditions A_s for $s \geq 2$ hold.*

Remark 5. *When $\alpha = 0$ the distributions may no longer be heavy tailed and the results of Theorem 1 need not hold. In fact, while all geometric distributions are regularly varying with $\alpha = 0$, [25] showed that for some of them* (8) *does not hold, and thus neither does the result of Theorem 1.*

We can also show that, under a mild additional condition, the assumptions of Theorem 1 do not hold for $\alpha = 0$. Specifically, assume that $\mathcal{P} \in \mathcal{RV}_0(\ell)$ and that there is a slowly varying at infinity function ℓ_0 such that

$$\sum_k p_k 1_{[p_k \leq x]} \sim x\ell_0(1/x) \text{ as } x \downarrow 0.$$

In this case, Proposition 19 in [14] implies that

$$\ell(x) \sim \int_1^x u^{-1}\ell_0(u)\mathrm{d}u \text{ as } x \to \infty,$$

and that for each $s \geq 1$

$$\Phi_s(n) \sim \frac{1}{s}\ell_0(n).$$

Thus, to get $s!\frac{g_n^2}{n^2}\Phi_s(n)$ *to converge to a positive constant* c_s*, we must take* $g_n \sim n\sqrt{\frac{c_s}{(s-1)!\ell_0(n)}}$*. However, since* ℓ_0 *is slowly varying at infinity,* g_n *does not satisfy* (5) *for any* $\beta \in (0, 1/2)$*, and the assumptions of Theorem 1 do not hold. Thus, the question of when and if asymptotic normality holds in this case cannot be answered using Theorem 1.*

Combining Corollaries 2 and 3 with Propositions 1 and 2 gives the following.

Corollary 4. *If* $\mathcal{P} \in \mathcal{RV}_\alpha(\ell)$ *with* $\alpha \in (0, 1]$ *then for any integer* $r \geq 0$ *we have*

$$\frac{\sqrt{r+1}N_{r+1,n}}{\sqrt{(r+1)N_{r+1,n} + (r+2)N_{r+2,n}}}\left(\frac{T_{r,n} - \pi_{r,n}}{\pi_{r,n}}\right) \xrightarrow{d} N(0,1)$$

and

$$\left(\frac{T_{r,n} - \pi_{r,n}}{\pi_{r,n}}\right) \xrightarrow{p} 0.$$

Note that, in the above, we do not need to know what α and ℓ are, only that they exist.

Remark 6. *Using a different approach, [25] showed that, when* $\alpha \in (0, 1)$*, the second convergence in Corollary 4 can be replaced by almost sure convergence. This was extended to the case* $\alpha = 1$ *and* $r = 0$ *by Corollary 5.3 of [2].*

5 Proofs

In this section we give the proofs of our results. We begin by introducing some notation and giving several lemmas that may be of independent interest. For any integer $r \geq 0$ let

$$\Pi_{r,n} = \sum_{i=0}^{r} \pi_{i,n} = \sum_k p_k 1_{[y_{k,n} \leq r]}$$

be the total probability of all letters observed *at most* r times, and let

$$M_{r,n} = \mathrm{E}[\Pi_{r,n}].$$

Note that, for $r \geq 1$, we have $\pi_{r,n} = \Pi_{r,n} - \Pi_{r-1,n}$ and $\mu_{r,n} = M_{r,n} - M_{r-1,n}$.

Lemma 4. *For any integer $r \geq 1$ and any $\epsilon > 0$*

$$\begin{aligned} P(|\pi_{r,n} - \mu_{r,n}| > \epsilon) &\leq P(|\Pi_{r,n} - M_{r,n}| > \epsilon/2) + P(|\Pi_{r-1,n} - M_{r-1,n}| > \epsilon/2) \\ &\leq 4\epsilon^{-2} \left[\mathrm{Var}(\Pi_{r,n}) + \mathrm{Var}(\Pi_{r-1,n}) \right]. \end{aligned}$$

The proof is similar to that of Lemma 20 in [25]. We include it for completeness.

Proof. Define the events $A = [-\epsilon/2 < \Pi_{r,n} - M_{r,n} < \epsilon/2]$, $B = [-\epsilon/2 < M_{r-1,n} - \Pi_{r-1,n} < \epsilon/2]$, and $C = [-\epsilon < \pi_{r,n} - \mu_{r,n} < \epsilon]$. Since $A \cap B \subset C$ it follows that $P(C^c) \leq P(A^c \cup B^c) \leq P(A^c) + P(B^c)$, which gives the first inequality. The second follows by Chebyshev's inequality. □

We will need bounds on the variances in the above lemma.

Lemma 5. *For any integer $r \geq 0$ we have*

$$\mathrm{Var}(\Pi_{r,n}) \leq \sum_{i=2}^{r+2} n^{i-2} \sum_k p_k^i (1 - p_k)^{n-i}.$$

This result follows from the fact that the random variables $\{y_{k,n} : k = 1, 2, \dots\}$ are negatively associated, see [10]. For completeness we give a detailed proof.

Proof. Note that

$$\begin{aligned} \mathrm{Var}(\Pi_{r,n}) &= \mathrm{E}[\Pi_{r,n}^2] - (\mathrm{E}[\Pi_{r,n}])^2 \\ &= \mathrm{E}\left[\left(\sum_k p_k 1_{[y_{k,n} \leq r]} \right)^2 \right] - \left(\sum_k p_k P(y_{k,n} \leq r) \right)^2 \\ &= \sum_k p_k^2 P(y_{k,n} \leq r) + \sum_{k \neq \ell} p_k p_\ell P(y_{k,n} \leq r, y_{\ell,n} \leq r) \\ &\qquad - \sum_k p_k^2 \left[P(y_{k,n} \leq r) \right]^2 - \sum_{k \neq \ell} p_k p_\ell P(y_{k,n} \leq r) P(y_{\ell,n} \leq r) \\ &\leq \sum_k p_k^2 P(y_{k,n} \leq r) \\ &\qquad + \sum_{k \neq \ell} p_k p_\ell \left[P(y_{k,n} \leq r, y_{\ell,n} \leq r) - P(y_{k,n} \leq r) P(y_{\ell,n} \leq r) \right] \\ &\leq \sum_k p_k^2 P(y_{k,n} \leq r), \end{aligned}$$

where the last inequality follows by the well-known fact that, for a multinomial distribution, $[P(y_{k,n} \leq r, y_{\ell,n} \leq r) - P(y_{k,n} \leq r) P(y_{\ell,n} \leq r)] \leq 0$, see

e.g. [20]. Combining the above with the fact that $y_{k,n}$ has a binomial distribution with parameters n and p_k gives

$$\begin{aligned}
\mathrm{Var}(\Pi_{r,n}) &\leq \sum_k p_k^2 \sum_{i=0}^{r} \binom{n}{i} p_k^i (1-p_k)^{n-i} \\
&= \sum_{i=0}^{r} \binom{n}{i} \sum_k p_k^{i+2} (1-p_k)^{n-i} \\
&\leq \sum_{i=0}^{r} n^i \sum_k p_k^{i+2} (1-p_k)^{n-i} \\
&= \sum_{i=2}^{r+2} n^{i-2} \sum_k p_k^i (1-p_k)^{n+2-i} \\
&\leq \sum_{i=2}^{r+2} n^{i-2} \sum_k p_k^i (1-p_k)^{n-i},
\end{aligned}$$

which completes the proof. □

To help simplify the above bound, we give the following result.

Lemma 6. *If $1 \leq s \leq t \leq u < \infty$ then*

$$\begin{aligned}
n^{t-2} \sum_k p_k^t (1-p_k)^{n-t} &\leq n^{s-2} \sum_k p_k^s (1-p_k)^{n-s} \\
&\quad + n^{u-2} \sum_k p_k^u (1-p_k)^{n-u}.
\end{aligned}$$

Proof. Observing that

$$\begin{aligned}
& n^{-2} \sum_k \left(\frac{np_k}{1-p_k} \right)^t (1-p_k)^n \\
&\leq n^{-2} \sum_k \max \left\{ \left(\frac{np_k}{1-p_k} \right)^s, \left(\frac{np_k}{1-p_k} \right)^u \right\} (1-p_k)^n \\
&\leq n^{s-2} \sum_k p_k^s (1-p_k)^{n-s} + n^{u-2} \sum_k p_k^u (1-p_k)^{n-u}
\end{aligned}$$

gives the result. □

Proof of Lemma 2. Observing that

$$\begin{aligned}
(s-1)! \frac{g_n^2}{n} \mu_{s-1,n-1} &= (s-1)! \frac{g_n^2}{n} \binom{n-1}{s-1} \sum_k p_k^s (1-p_k)^{n-s} \\
&\sim g_n^2 n^{s-2} \sum_k p_k^s (1-p_k)^{n-s}
\end{aligned}$$

and

$$\begin{aligned} s!\frac{g_n^2}{n^2}\mathrm{E}[N_{s,n}] &= s!\frac{g_n^2}{n^2}\binom{n}{s}\sum_k p_k^s(1-p_k)^{n-s} \\ &\sim g_n^2 n^{s-2}\sum_k p_k^s(1-p_k)^{n-s} \end{aligned}$$

gives the equivalence between a), b), and c). The equivalence between c) and d) is shown in [31]. □

Proof of Lemma 3. First note that for $s \geq 1$

$$\begin{aligned} \lim_{n\to\infty} n^{s-1}\sum_k p_k^s(1-p_k)^{n-s} &\leq \lim_{n\to\infty}\sum_k p_k(np_k)^{s-1}e^{-p_k(n-s)} \\ &= e^s\sum_k p_k \lim_{n\to\infty}(np_k)^{s-1}e^{-p_k n} = 0, \end{aligned}$$

where we use the well-known fact that $(1-x) \leq e^{-x}$ for $x \geq 0$ (see, e.g., 4.2.29 in [1]) and we interchange limit and summation by dominated convergence and the fact that the function $f(x) = x^{s-1}e^{-x}$ is bounded for $x \geq 0$. Combining the above with Condition A_s and the assumption that $c_s > 0$ gives $g_n^2/n \to \infty$, which implies (9).

We now turn to (10). Throughout the proof of this part we use the formulation of Condition A_s given in d) of Lemma 2. We have

$$g_{n+1}^2(n+1)^{s-2}\sum_k p_k^s e^{-(n+1)p_k} \to c_s,$$

and hence

$$\begin{aligned} g_n^2 n^{s-2}\sum_k p_k^s e^{-(n+1)p_k} &= \frac{g_n^2}{g_{n+1}^2}\frac{n^{s-2}}{(n+1)^{s-2}}g_{n+1}^2(n+1)^{s-2}\sum_k p_k e^{-(n+1)p_k} \\ &\sim c_s\frac{g_n^2}{g_{n+1}^2}. \end{aligned}$$

Combining this with the fact that

$$g_n^2 n^{s-2}\sum_k p_k^s e^{-(n+1)p_k} \leq g_n^2 n^{s-2}\sum_k p_k^s e^{-np_k} \to c_s$$

gives

$$\limsup_{n\to\infty}\frac{g_n}{g_{n+1}} \leq 1.$$

Now, let $A_n = \{k : p_k \leq n^{-1/2}\}$ and let $B_n = A_n^c = \{k : p_k > n^{-1/2}\}$. Note that the cardinality of B_n is bounded by $n^{1/2}$. Using the facts that the

function $f(t) = t^s e^{-nt}$ is decreasing on $(s/n, 1]$ and that $sn^{-1} < n^{-1/2}$ for large enough n gives

$$\begin{aligned} 0 \leq \limsup_{n\to\infty} g_n^2 n^{s-2} \sum_{k\in B_n} p_k^s e^{-np_k} &\leq \limsup_{n\to\infty} g_n^2 n^{s-2} \sum_{k\in B_n} n^{-s/2} e^{-n^{1/2}} \\ &\leq \lim_{n\to\infty} \left(\frac{g_n}{n^{1-\beta}}\right)^2 n^{.5(s+1)-2\beta} e^{-n^{1/2}} = 0, \end{aligned}$$

where the convergence follows by (5). This implies that

$$\begin{aligned} \liminf_{n\to\infty} g_n^2 n^{s-2} \sum_k p_k^s e^{-(n+1)p_k} &\geq \liminf_{n\to\infty} g_n^2 n^{s-2} \sum_{k\in A_n} p_k^s e^{-np_k} e^{-p_k} \\ &\geq \liminf_{n\to\infty} g_n^2 n^{s-2} e^{-n^{-1/2}} \sum_{k\in A_n} p_k^s e^{-np_k} \\ &= \lim_{n\to\infty} g_n^2 n^{s-2} \sum_k p_k^s e^{-np_k} = c_s. \end{aligned}$$

Now note that

$$\begin{aligned} g_n^2 n^{s-2} \sum_k p_k^s e^{-(n+1)p_k} &= \frac{g_n^2}{g_{n+1}^2} \frac{n^{s-2}}{(n+1)^{s-2}} g_{n+1}^2 (n+1)^{s-2} \sum_k p_k e^{-(n+1)p_k} \\ &\sim c_s \frac{g_n^2}{g_{n+1}^2}. \end{aligned}$$

This implies that

$$\liminf_{n\to\infty} \frac{g_n}{g_{n+1}} \geq 1,$$

which completes the proof of the first part.

We now turn to the second part. Note that when A_{r+1} holds for $r \geq 2$ we have

$$\begin{aligned} &\limsup_{n\to\infty} g_n^2 \sum_{k:p_k\geq 1/n} p_k^2 (1-p_k)^{n-2} \\ &\quad = \limsup_{n\to\infty} \frac{g_n^2}{n^2} \sum_{k:p_k\geq 1/n} (np_k)^2 (1-p_k)^{n-2} \\ &\quad \leq \limsup_{n\to\infty} \frac{g_n^2}{n^2} \sum_k (np_k)^{r+1} (1-p_k)^{n-(r+1)} = c_{r+1} \end{aligned}$$

and that for $n \geq 2$

$$\left(1-\frac{1}{n}\right)^{n-2} g_n^2 \sum_{k:p_k<1/n} p_k^2 \leq g_n^2 \sum_{k:p_k<1/n} p_k^2 (1-p_k)^{n-2} \leq g_n^2 \sum_{k:p_k<1/n} p_k^2.$$

From here, the fact that $\left(1-\frac{1}{n}\right)^{n-2} \to e^{-1}$ gives the result. □

Proof of Theorem 1. We have

$$\mu_{r,n} g_n \left(\frac{T_{r,n} - \pi_{r,n}}{\pi_{r,n}} \right) = \frac{\mu_{r,n}}{\pi_{r,n}} g_n \left(T_{r,n} - \pi_{r,n} \right).$$

By Lemma 1 and Slutsky's Theorem it suffices to show that

$$\frac{\pi_{r,n}}{\mu_{r,n}} \xrightarrow{p} 1.$$

When $r = 0$, Chebyshev's inequality implies that for any $\epsilon > 0$

$$P\left(\left| \frac{\pi_{0,n}}{\mu_{0,n}} - 1 \right| > \epsilon \right) = P\left(|\pi_{0,n} - \mu_{0,n}| > \mu_{0,n}\epsilon \right) \leq \epsilon^{-2} \frac{\mathrm{Var}\left(\pi_0\right)}{\mu_{0,n}^2}.$$

Similarly, when $r \geq 1$, Lemma 4 implies that for any $\epsilon > 0$

$$P\left(\left| \frac{\pi_{r,n}}{\mu_{r,n}} - 1 \right| > \epsilon \right) \leq 4\epsilon^{-2} \left[\frac{\mathrm{Var}(\Pi_{r,n})}{\mu_{r,n}^2} + \frac{\mathrm{Var}(\Pi_{r-1,n})}{\mu_{r,n}^2} \right].$$

In both cases we can combine the above with Lemma 5 to show that

$$\begin{aligned} P\left(\left| \frac{\pi_{r,n}}{\mu_{r,n}} - 1 \right| > \epsilon \right) &\leq 8\epsilon^{-2} \frac{\sum_{i=2}^{r+2} n^{i-2} \sum_k p_k^i (1-p_k)^{n-i}}{\mu_{r,n}^2} \\ &= 8\epsilon^{-2} \frac{\sum_{i=2}^{r+2} g_n^2 n^{i-2} \sum_k p_k^i (1-p_k)^{n-i}}{g_n^2 \mu_{r,n}^2}. \end{aligned}$$

We must now show that this approaches zero. By (7), the fact that Condition A_{r+2} holds, and Lemma 6, the limsup of the numerator is bounded and it suffices to show that $[g_n \mu_{r,n}]^{-1} \to 0$. To see this note that by (5) and Lemmas 2 and 3

$$\frac{1}{g_n \mu_{r,n}} = \frac{g_n}{n} \frac{n}{g_n^2 \mu_{r,n}} \sim \frac{r!}{c_{r+1}} \frac{g_n}{n} = \frac{r!}{c_{r+1}} \frac{g_n}{n^{1-\beta}} n^{-\beta} \to 0,$$

which concludes the proof. □

Proof of Corollary 2. We begin with the first part. Note that Lemma 2 combined with Lemma 3 implies that

$$\mu_{r,n} g_n \sim \frac{c_{r+1}}{r!} \frac{n}{g_n} \sim (r+1) \frac{g_n}{n} \mathrm{E}[N_{r+1,n}]$$

and

$$(r+1)^2 \frac{g_n^2}{n^2} \mathrm{E}[N_{r+1,n}] + (r+2)(r+1) \frac{g_n^2}{n^2} \mathrm{E}[N_{r+2,n}] \to \frac{(r+1)c_{r+1} + c_{r+2}}{r!}.$$

Putting everything together and applying Slutsky's Theorem gives the first part. For the second part, we note that in the proof of Theorem 3.3 in [31] it was shown that

$$(r+1)! \frac{g_n^2}{n^2} N_{r+1,n} \xrightarrow{p} c_{r+1} \text{ and } (r+2)! \frac{g_n^2}{n^2} N_{r+2,n} \xrightarrow{p} c_{r+2}.$$

From here the result follows as in the previous part. □

Acknowledgments

The authors wish to thank the anonymous referees whose detailed comments led to improvements in the presentation of this paper.

References

[1] M. Abramowitz and I. A. Stegun (1972). *Handbook of Mathematical Functions* 10th ed. Dover Publications, New York.

[2] A. Ben-Hamou, S. Boucheron, and M. I. Ohannessian (2017). Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli*, 23(1):249–287.

[3] D. Berend and A. Kontorovich (2013). On the concentration of the missing mass. *Electronic Communications in Probability*, 18(3):1–7.

[4] N. H. Bingham, C. M. Goldie, and J. L. Teugels (1987). *Regular Variation*. Encyclopedia of Mathematics And Its Applications. Cambridge University Press, Cambridge.

[5] A. Chao (1981). On estimating the probability of discovering a new species. *The Annals of Statistics*, 9(6):1339–1342.

[6] A. Chao, T. C. Hsieh, R. L. Chazdon, R. K. Colwell, and N. J. Gotelli (2015). Unveiling the species-rank abundance distribution by generalizing the Good–Turing sample coverage theory. *Ecology*, 96(5):1189–1201.

[7] S. F. Chen and J. Goodman (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394.

[8] A. Cohen and H. B. Sackrowitz (1990). Admissibility of estimators of the probability of unobserved outcomes. *Annals of the Institute of Statistical Mathematics*, 42(4):623–636, 1990.

[9] G. Decrouez, M. Grabchak, and Q. Paris (2016). Finite sample properties of the mean occupancy counts and probabilities. To appear in *Bernoulli*.

[10] D. Dubhashi and D. Ranjan (1998). Balls and bins: A study in negative dependence. *Random Structures & Algorithms*, 13(2):99–124.

[11] B. Efron and R. Thisted (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63(3):435–447.

[12] W. W. Esty (1983). A normal limit law for a nonparametric estimator of the coverage of a random sample. *Annals of Statistics*, 11(3):905–912.

[13] S. Favaro, B. Nipoti, and Y. W. Teh (2016). Rediscovery of Good-Turing estimators via Bayesian nonparametrics. *Biometrics*, 72(1):136–145.

[14] A. Gnedin, B. Hansen, and J. Pitman (2007). Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probability Surveys*, 4:146–171.

[15] I. J. Good (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3/4):237–264.

[16] I. J. Good and G. H. Toulmin (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43(1–2):45–63.

[17] M. Grabchak and V. Cosme (2015). On The Performance of Turing's Formula: A simulation study. *Communication in Statistics – Simulation and Computation*, DOI: 10.1080/03610918.2015.1109658.

[18] V. Gupta, M. Lennig, and P. Mermelstein (1992). A language model for very large-vocabulary speech recognition. *Computer Speech & Language*, 6(4):331–344.

[19] S. Karlin (1967). Central limit theorems for certain infinite urn schemes. *Journal of Mathematical Mechanics*, 17:373–401.

[20] C. L. Mallows (1968). An inequality involving multinomial probabilities. *Biometrika*, 55(2):422–424.

[21] C. X. Mao and B. G. Lindsay (2002). A Poisson model for the coverage problem with a genomic application. *Biometrika*, 89(3):669–681.

[22] D. A. McAllester and R. E. Schapire (2000). On the convergence rate of Good-Turing estimators. In *COLT '00: Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, pp. 1–6.

[23] D. A. McAllester and L. E. Ortiz (2003). Concentration inequalities for the missing mass and for histogram rule error. *Journal of Machine Learning Research*, 4(Oct):895–911.

[24] E. Mossel and M. I. Ohannessian (2015). On the Impossibility of Learning the Missing Mass. *arXiv*:1503.03613v1.

[25] M. I. Ohannessian and M. A. Dahleh (2012). Rare probability estimation under regularly varying heavy tails. *JMLR Workshop and Conference Proceedings*, 23:21.1–21.24.

[26] S. I. Resnick (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer, New York.

[27] H. E. Robbins (1968). Estimating the total probability of the unobserved outcomes of an experiment. *Annals of Mathematical Statistics*, 39(1):256–257.

[28] R. Thisted and B. Efron (1987). Did Shakespeare write a newly discovered poem. *Biometrika*, 74(3):445–455.

[29] C. H. Zhang (2005). Estimation of sums of random variables: Examples and information bounds. *The Annals of Statistics*, 33(5):2022–2041.

[30] C. H. Zhang and Z. Zhang (2009). Asymptotic normality of a nonparametric estimator of sample coverage. *Annals of Statistics*, 37(5A):2582–2595.

[31] Z. Zhang (2013). A multivariate normal law for Turing's formulae. *Sankhya A*, 75(1):51–73.

[32] Z. Zhang and H. Huang (2007). Turing's formula revisited. *Journal of Quantitative Linguistics*, 14(2-3):222–241.

[33] Z. Zhang and H. Huang (2008). A sufficient normality condition for Turing's formula. *Journal of Nonparametric Statistics*, 20(5):431–446.