On Turing's formula and the estimation of the missing mass

Michael Grabchak

University of North Carolina at Charlotte
Department of Mathematics and Statistics

December 4, 2017

Contents

Part I: Estimation of the Missing Mass

Joint work with Z. Zhang

Bird Example

Consider the bird population in a particular location ...

- Species: $a_1, a_2, \cdots$

Bird Example

Consider the bird population in a particular location ...

- Species: $a_1, a_2, \cdots$
- Unknown species proportions: $P = \{p_1, p_2, \cdots\}$

Bird Example

Consider the bird population in a particular location ...

- Species: $a_1, a_2, \cdots$
- Unknown species proportions: $P = \{p_1, p_2, \cdots\}$
- A sample of $n = 2000$ is taken.

Bird Example

Consider the bird population in a particular location ...

- Species: $a_1, a_2, \cdots$
- Unknown species proportions: $P = \{p_1, p_2, \cdots\}$
- A sample of $n = 2000$ is taken.
- Counts: $x_1 = 300, x_2 = 200, \cdots$

Bird Example

Consider the bird population in a particular location ...

- Species: $a_1, a_2, \cdots$
- Unknown species proportions: $P = \{p_1, p_2, \cdots\}$
- A sample of $n = 2000$ is taken.
- Counts: $x_1 = 300, x_2 = 200, \cdots$
- $\hat{p}_1 = 0.15, \hat{p}_2 = 0.10, \cdots$ estimating $p_1, p_2, \cdots$

Bird Example Questions

- **Q1:** What is the probability that the next observation will be of species a_1?

Bird Example Questions

- **Q1:** What is the probability that the next observation will be of species a_1?
- **A1:** We can estimate it by $\hat{p}_1 = .15$.

Bird Example Questions

- **Q1:** What is the probability that the next observation will be of species a_1?
- **A1:** We can estimate it by $\hat{p}_1 = .15$.
- **Q2:** What is the probability that the next observation will be of a species that we have not observed before?

Bird Example Questions

- **Q1:** What is the probability that the next observation will be of species a_1?
- **A1:** We can estimate it by $\hat{p}_1 = .15$.
- **Q2:** What is the probability that the next observation will be of a species that we have not observed before?
- **A2:** An answer is given by Turing's formula.

Bird Example Questions

- **Q1:** What is the probability that the next observation will be of species a_1?
- **A1:** We can estimate it by $\hat{p}_1 = .15$.
- **Q2:** What is the probability that the next observation will be of a species that we have not observed before?
- **A2:** An answer is given by Turing's formula.

$$T_{0,n} = \frac{\#\text{ of species occurring exactly once in the sample}}{n}.$$

Intuition

Turing's formula estimation the probability of seeing a new species by

$$T_{0,n} = \frac{\#\text{ of species occurring exactly once in the sample}}{n}.$$

Intuition

Turing's formula estimation the probability of seeing a new species by

$$T_{0,n} = \frac{\text{\# of species occurring exactly once in the sample}}{n}.$$

If all observations are of different species then

Intuition

Turing's formula estimation the probability of seeing a new species by

$$T_{0,n} = \frac{\text{\# of species occurring exactly once in the sample}}{n}.$$

If all observations are of different species then

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

Intuition

Turing's formula estimation the probability of seeing a new species by

$$T_{0,n} = \frac{\#\text{ of species occurring exactly once in the sample}}{n}.$$

If all observations are of different species then

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

If all observations are from a few very abundant species then

Intuition

Turing's formula estimation the probability of seeing a new species by

$$T_{0,n} = \frac{\#\text{ of species occurring exactly once in the sample}}{n}.$$

If all observations are of different species then

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

If all observations are from a few very abundant species then

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

Notation

Let

- $\mathcal{A} = \{a_1, a_2, \dots\}$ = an at most countable alphabet.
- $P = \{p_a : a \in \mathcal{A}\}$ the associated probability distribution, where $p_a \in [0, 1]$ and $\sum_{a\in\mathcal{A}} p_a = 1$.
- $\mathcal{S} = \{a \in \mathcal{A} : p_a > 0\}$ = the support of P.

Notation

Let

- $\mathcal{A} = \{a_1, a_2, \dots\}$ = an at most countable alphabet.
- $P = \{p_a : a \in \mathcal{A}\}$ the associated probability distribution, where $p_a \in [0,1]$ and $\sum_{a\in\mathcal{A}} p_a = 1$.
- $\mathcal{S} = \{a \in \mathcal{A} : p_a > 0\}$ = the support of P.

Let $X_1, \dots, X_n$ be independent and identically distributed $\mathcal{A}$-valued random variables, with distribution P.

- $L_n(a) = \sum_{i=1}^{n} \mathbf{1}\{X_i = a\}$ – the sample counts of $a \in \mathcal{A}$;
- $\hat{p}_a = L_n(a)/n$ – the sample proportion of $a \in \mathcal{A}$.
- $K_{r,n} = \sum_{a\in\mathcal{A}} 1_{[L_n(a)=r]}$, $r = 0, 1, \dots, n$.

Missing Mass

The missing mass is the total probability associated with the letters not covered in the sample, it is given by

$$M_{0,n} = \sum_{a \in \mathcal{A}} p_a \mathbf{1}\{L_n(a) = 0\}.$$

Missing Mass

The missing mass is the total probability associated with the letters not covered in the sample, it is given by

$$M_{0,n} = \sum_{a \in \mathcal{A}} p_a \mathbf{1}\{L_n(a) = 0\}.$$

Note that this is not a parameter or a statistic. It depends on both unknown parameters and the sample.

Turing's Formula

A "good" estimator of the missing mass

$$M_{0,n} = \sum_{a \in \mathcal{A}} p_a \mathbf{1}\{L_n(a) = 0\}$$

is provided by Turing's formula

$$T_{0,n} = \frac{\#\text{ of letters occurring exactly once in the sample}}{n} = \frac{K_{1,n}}{n}.$$

Turing's Formula

A "good" estimator of the missing mass

$$M_{0,n} = \sum_{a \in \mathcal{A}} p_a \mathbf{1}\{L_n(a) = 0\}$$

is provided by Turing's formula

$$T_{0,n} = \frac{\#\text{ of letters occurring exactly once in the sample}}{n} = \frac{K_{1,n}}{n}.$$

This formula was first introduced in Good (1953), where the results were largely credited to Alan Turing.

Bias of Turing's Formula

$$\mathbb{E}\left[T_{0,n} - M_{0,n}\right] = \sum_{a \in \mathcal{A}} p_a^2 \left(1 - p_a\right)^{n-1} > 0.$$

Thus, for large n,

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

Consistency of Turing's Formula

We always have

$$T_{0,n} - M_{0,n} \xrightarrow{p} 0.$$

Consistency of Turing's Formula

We always have

$$T_{0,n} - M_{0,n} \xrightarrow{p} 0.$$

However,

$$M_{0,n} \xrightarrow{p} 0, \;\; T_{0,n} \xrightarrow{p} 0, \text{ as } n \to \infty$$

Consistency in Relative Error

Ohannessian and Dahleh (2012) suggested that it is more meaningful to consider consistency in relative error:

$$\frac{T_{0,n} - M_{0,n}}{M_{0,n}} \xrightarrow{p} 0.$$

Consistency in Relative Error

Ohannessian and Dahleh (2012) suggested that it is more meaningful to consider consistency in relative error:

$$\frac{T_{0,n} - M_{0,n}}{M_{0,n}} \xrightarrow{p} 0.$$

This does not hold for all distributions.

Consistency in Relative Error

Ohannessian and Dahleh (2012) suggested that it is more meaningful to consider consistency in relative error:

$$\frac{T_{0,n} - M_{0,n}}{M_{0,n}} \xrightarrow{p} 0.$$

This does not hold for all distributions.

Ben-Hamou et al. (2017) gave sufficient conditions when this holds.

Asymptotic normality

We consider the related problem of asymptotic normality. This allows not just for estimation, but for statistical inference.

Asymptotic normality

We consider the related problem of asymptotic normality. This allows not just for estimation, but for statistical inference.

Asymptotic normality for Turing's formula was considered in Esty (1983), Zhang and Huang (2008), Zhang and Zhang (2009), and Grabchak and Zhang (2017).

Main Theorem

Let g_n be a deterministic sequence of positive numbers with

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

If there are constants $c_1 > 0$ and $c_2 \geq 0$ with

$$\lim_{n\to\infty} \frac{g_n^2}{n}\mathbb{E}[T_{0,n}] = c_1 \text{ and } \lim_{n\to\infty} g_n^2 \sum_{a\in\mathcal{A}} p_a^2 e^{-np_a} = c_2$$

then

Main Theorem

Let g_n be a deterministic sequence of positive numbers with

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

If there are constants $c_1 > 0$ and $c_2 \geq 0$ with

$$\lim_{n\to\infty} \frac{g_n^2}{n} \mathbb{E}[T_{0,n}] = c_1 \text{ and } \lim_{n\to\infty} g_n^2 \sum_{a\in\mathcal{A}} p_a^2 e^{-np_a} = c_2$$

then

$$h_n \left(\frac{T_{0,n} - M_{0,n}}{M_{0,n}} \right) \xrightarrow{d} N(0, c_1 + c_2).$$

where $h_n = \mathbb{E}[M_{0,n}] g_n$

Corollary

In practice, we don't know g_n, but so long as it exists, it and the other parameters can be estimated.

Corollary. If the conditions of the Theorem are satisfied, then

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

where

$$K_{r,n} = \sum_{a \in \mathcal{A}} \mathbf{1}\{L_n(a) = r\}.$$

Confidence Intervals

If $\frac{K_{1,n}}{2K_{2,n}}$ is not very close to 0, then an approximate $(1-\alpha)100\%$ confidence interval

$$\frac{K_{1,n}^2/n}{K_{1,n}+z_{\alpha/2}\sqrt{K_{1,n}+2K_{2,n}}} \le M_{0,n} \le \frac{K_{1,n}^2/n}{K_{1,n}-z_{\alpha/2}\sqrt{K_{1,n}+2K_{2,n}}}$$

where $z_{\alpha/2}$ is a number with $P(Z > z_{\alpha/2}) = \alpha/2$.

Consistency

Corollary. If the conditions of the Theorem are satisfied, then

$$\frac{T_{0,n} - M_{0,n}}{M_{0,n}} \xrightarrow{p} 0.$$

Consistency

Corollary. If the conditions of the Theorem are satisfied, then

$$\frac{T_{0,n} - M_{0,n}}{M_{0,n}} \xrightarrow{p} 0.$$

Our conditions appear to be different from the ones given in Ben-Hamou et al. (2017)

The Counting Function

To talk about tails of distributions on an alphabet, Karlin (1967) introduced the counting function $\nu : [0,1] \to \mathbb{N}$, defined by

$$\nu(\varepsilon) = \sum_{a \in \mathcal{A}} \mathbf{1}\{p_a \geq \varepsilon\}$$

Facts:

1. ν is non-increasing with ε
2. For all $0 < \varepsilon \leq 1$, $\nu(\varepsilon) \leq \varepsilon^{-1}$
3. $\varepsilon\nu(\varepsilon) \to 0$ as $\varepsilon \to 0$

Regularly Varying Distributions

A discrete distribution P is said to be regularly varying with index $\alpha \in [0,1]$ if

$$\nu(\varepsilon) = \varepsilon^{-\alpha}\ell(1/\varepsilon),$$

where ℓ is a slowly varying function, i.e.

$$\lim_{x\to\infty} \frac{\ell(xt)}{\ell(x)} = 1, \text{ for any } t > 0.$$

In this case we write $P \in \mathcal{RV}_\alpha(\ell)$. This definition is due to Karlin (1967), see Gnedin et al. (2007) for a recent review.

Regularly Varying Distributions

Fact: Assume that $\mathcal{A} = \mathbb{N}$. $P \in \mathcal{RV}_\alpha(\ell)$ 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 function, in general, different from ℓ.

Results for Regularly Varying Distributions

Proposition. If $P \in \mathcal{RV}_\alpha(\ell)$ for some $\alpha \in (0,1)$ then, the assumptions of the Theorem hold and

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

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

Results for Regularly Varying Distributions

Proposition. If $P \in \mathcal{RV}_\alpha(\ell)$ for some $\alpha \in (0,1)$ then, the assumptions of the Theorem hold and

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

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

A similar result holds for $\alpha = 1$, but with a somewhat different scaling.

Case $\alpha = 0$

When $\alpha = 0$ the distributions may no longer be heavy tailed and the results of the Theorem need not hold.

Ohannessian and Dahleh (2012) showed that consistency in relative error cannot hold for certain $\mathcal{RV}_0$ distributions.

Extension: rth order Turing Formula

For any $0 \le r \le n-1$ we define the occupancy probabilities by

$$M_{r,n} = \sum_{a \in \mathcal{A}} p_a \mathbf{1}\{L_n(a) = r\}.$$

and the occupancy counts by

$$K_{r,n} = \sum_{a \in \mathcal{A}} \mathbf{1}\{L_n(a) = r\}.$$

Extension: rth order Turing Formula

For any $0 \leq r \leq n-1$ we define the occupancy probabilities by

$$M_{r,n} = \sum_{a \in \mathcal{A}} p_a \mathbf{1}\{L_n(a) = r\}.$$

and the occupancy counts by

$$K_{r,n} \;=\; \sum_{a \in \mathcal{A}} \mathbf{1}\{L_n(a) = r\}.$$

We can estimate $M_{r,n}$ by the rth order Turing's formula

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

and our results can be extended to this case

Part II: Expectation of the Missing Mass

Joint work with G. Decrouez and Q. Paris

Main Objects

For $0 \leq r \leq n$, the occupancy counts $K_{r,n}$ are defined by

$$K_{r,n} = \sum_{a \in \mathcal{A}} \mathbf{1}\{L_n(a) = r\}$$

and the occupancy probabilities $M_{r,n}$ are defined by

$$M_{r,n} = \sum_{a \in \mathcal{A}} p_a \mathbf{1}\{L_n(a) = r\}.$$

Main Objects

For $0 \leq r \leq n$, the occupancy counts $K_{r,n}$ are defined by

$$K_{r,n} = \sum_{a \in \mathcal{A}} \mathbf{1}\{L_n(a) = r\}$$

and the occupancy probabilities $M_{r,n}$ are defined by

$$M_{r,n} = \sum_{a \in \mathcal{A}} p_a \mathbf{1}\{L_n(a) = r\}.$$

The missing mass is $M_{0,n}$.

Statement of Problem

Our Goal: To understand the finite sample properties of $\mathbb{E}K_{r,n}$ and $\mathbb{E}M_{r,n}$.

Statement of Problem

Our Goal: To understand the finite sample properties of $\mathbb{E}K_{r,n}$ and $\mathbb{E}M_{r,n}$.

It can be shown that

$$\mathbb{E}M_{r,n} = \left(\frac{1+r}{1+n}\right)\mathbb{E}K_{r+1,n+1}\,.$$

Statement of Problem

Our Goal: To understand the finite sample properties of $\mathbb{E}K_{r,n}$ and $\mathbb{E}M_{r,n}$.

It can be shown that

$$\mathbb{E}M_{r,n} = \left(\frac{1+r}{1+n}\right)\mathbb{E}K_{r+1,n+1}\,.$$

For this reason we only focus on $\mathbb{E}M_{r,n}$

This problem was previously studied in Ohannessian and Dahleh (2010) and Berend and Kontorovich (2012) for the case $r = 0$.

Statement of Problem

In short, we study the object

$$\mathbb{E}M_{r,n} = \binom{n}{r} \sum_{a \in \mathcal{A}} p_a^{r+1} (1 - p_a)^{n-r}.$$

Upper Bounds

Theorem

For any $n \geq 1$ and any $0 \leq r \leq n-1$, we have

$$\mathbb{E}M_{r,n} \leq \inf_{0 \leq \varepsilon \leq 1} \left\{ \varphi^+_{r,n}(\varepsilon) + \psi^+_{r,n}(\varepsilon) \right\},$$

where

$$\begin{aligned}
\varphi^+_{r,n}(\varepsilon) &= \frac{c(r)\nu(\varepsilon)}{n}, \\
\psi^+_{r,n}(\varepsilon) &= 2^{1+r}\binom{n}{r}\int_0^\varepsilon \nu\left(\frac{u}{2}\right)u^r\left(1-\frac{u}{2}\right)^{n-r}\mathrm{d}u, \\
c(r) &= \begin{cases} e^{-1} & \textit{if } r = 0, \\ \frac{(1+r)^{2+r}}{r!}e^{-\frac{1+r}{2}} & \textit{if } 1 \leq r \leq n-1. \end{cases}
\end{aligned}$$

Upper Bounds

Theorem

For any $n \geq 1$ *and any* $0 \leq r \leq n-1$, *we have*

$$\mathbb{E}M_{r,n} \leq \inf_{0 \leq \varepsilon \leq 1} \left\{ \varphi^{+}_{r,n}(\varepsilon) + \psi^{+}_{r,n}(\varepsilon) \right\},$$

where

$$\begin{aligned} \varphi^{+}_{r,n}(\varepsilon) &= \frac{c(r)\nu(\varepsilon)}{n}, \\ \psi^{+}_{r,n}(\varepsilon) &= 2^{1+r}\binom{n}{r}\int_0^{\varepsilon} \nu\left(\frac{u}{2}\right) u^r \left(1-\frac{u}{2}\right)^{n-r} \mathrm{d}u, \\ c(r) &= \begin{cases} e^{-1} & \textit{if } r = 0, \\ \frac{(1+r)^{2+r}}{r!} e^{-\frac{1+r}{2}} & \textit{if } 1 \leq r \leq n-1. \end{cases} \end{aligned}$$

In many situations, a relevant choice of ε yields explicit and, as far as we know, new bounds.

Finite Support

Corollary. Suppose that $\mathcal{S}$ is finite. Then, for all $n \geq 1$ and all $0 \leq r \leq n-1$,

$$\mathbb{E}M_{r,n} \leq \frac{c(r)|\mathcal{S}|}{n}.$$

Finite Support

Corollary. Suppose that $\mathcal{S}$ is finite. Then, for all $n \geq 1$ and all $0 \leq r \leq n-1$,

$$\mathbb{E}M_{r,n} \leq \frac{c(r)|\mathcal{S}|}{n}.$$

When we take $r = 0$ we recover the bound for the expected missing mass

$$\mathbb{E}M_{0,n} \leq \frac{|\mathcal{S}|}{ne},$$

provided by Berend and Kontorovich (2012).

Regular Variation

Fact : $\nu(\varepsilon) = \varepsilon^{-\alpha}\ell\left(\frac{1}{\varepsilon}\right) \Rightarrow \mathbb{E}M_{r,n} \underset{n\to\infty}{\sim} \frac{\alpha\Gamma(1+r-\alpha)}{r!}\frac{\ell(n)}{n^{1-\alpha}}$

Regular Variation

Fact : $\nu(\varepsilon) = \varepsilon^{-\alpha}\ell\left(\frac{1}{\varepsilon}\right) \Rightarrow \mathbb{E}M_{r,n} \underset{n\to\infty}{\sim} \frac{\alpha\Gamma(1+r-\alpha)}{r!}\frac{\ell(n)}{n^{1-\alpha}}$

Corollary.

$$\nu(\varepsilon) \le \varepsilon^{-\alpha}\ell\left(\frac{1}{\varepsilon}\right) \quad \Rightarrow \quad \mathbb{E}M_{r,n} \le c(\alpha, r)\,\frac{\ell(n)}{n^{1-\alpha}},$$

where

$$c(\alpha, r) = c(r) + \frac{4^{1+r}}{r!}(1+r)^{1+r-\alpha}\int_0^{1/2} u^{r-\alpha}e^{-u}\mathrm{d}u$$

where ℓ is nondecreasing.

Regular Variation: Lower Bounds

Fact : $\nu(\varepsilon) = \varepsilon^{-\alpha}\ell\left(\frac{1}{\varepsilon}\right) \Rightarrow \mathbb{E}M_{r,n} \underset{n\to\infty}{\sim} \frac{\alpha\Gamma(1+r-\alpha)}{r!}\frac{\ell(n)}{n^{1-\alpha}}$

Regular Variation: Lower Bounds

Fact : $\nu(\varepsilon) = \varepsilon^{-\alpha}\ell\left(\frac{1}{\varepsilon}\right) \Rightarrow \mathbb{E}M_{r,n} \underset{n\to\infty}{\sim} \frac{\alpha\Gamma(1+r-\alpha)}{r!}\frac{\ell(n)}{n^{1-\alpha}}$

Corollary.

$$\nu(\varepsilon) \geq \varepsilon^{-\alpha}\ell\left(\frac{1}{\varepsilon}\right) \quad \Rightarrow \quad \mathbb{E}M_{r,n} \geq c_1(\alpha, r)\frac{\ell(n)}{n^{1-\alpha}},$$

where ℓ is nondecreasing.

Application: Bounds in Probability

Concentration inequalities for the missing mass have been studied in McAllester and Ortiz (2003), Ohannessian and Dahleh (2012), and Ben-Hamou et al. (2017).

These can combined with our bounds to get bounds in probability.

Application: Bounds in Probability

Example: Assume that, for some $\alpha \in (0, 1)$,

$$p_k = C_\alpha k^{-1/\alpha}, \; k = 1, 2, \ldots.$$

Application: Bounds in Probability

Example: Assume that, for some $\alpha \in (0,1)$,

$$p_k = C_\alpha k^{-1/\alpha}, \ k = 1, 2, \ldots .$$

For all $t > 0$,

$$\mathbb{P}\left(m^-_{0,n}(t,\alpha) \le M_{0,n} \le m^+_{0,n}(t,\alpha)\right) \ge 1 - 2e^{-t},$$

where

$$m^-_{0,n}(t,\alpha) = \frac{(2^\alpha - 1)\gamma(1-\alpha, 2)}{32}\frac{C_\alpha^\alpha}{n^{1-\alpha}} - \sqrt{\frac{2t}{ne}}$$

$$m^+_{0,n}(t,\alpha) = \left(\frac{1}{e} + 4\gamma\left(1-\alpha, \frac{1}{2}\right)\right)\frac{C_\alpha^\alpha}{n^{1-\alpha}} + \sqrt{\frac{t}{n}}.$$

Part III: Simpson's Indices

Joint work with L. Cao and Z. Zhang

Simpson's Index

Simpson (1949), introduced a bio-diversity index

$$\zeta_1 = \sum_{a \in \mathcal{A}} p_a(1 - p_a)$$

It is sometimes called Simpson's index or the Gini-Simpson index.

Simpson's Index

Simpson (1949), introduced a bio-diversity index

$$\zeta_1 = \sum_{a \in \mathcal{A}} p_a(1 - p_a) = \mathbb{E}\left[M_{0,1}\right].$$

It is sometimes called Simpson's index or the Gini-Simpson index.

Simpson's Index

Simpson (1949), introduced a bio-diversity index

$$\zeta_1 = \sum_{a \in \mathcal{A}} p_a(1 - p_a).$$

To estimate the diversity of an eco-system, we can estimate ζ_1.

Simpson's Index

Simpson (1949), introduced a bio-diversity index

$$\zeta_1 = \sum_{a \in \mathcal{A}} p_a(1 - p_a).$$

A common estimator is the plug-in

$$\sum_{a \in \mathcal{A}} \hat{p}_a(1 - \hat{p}_a),$$

Simpson's Index

Simpson (1949), introduced a bio-diversity index

$$\zeta_1 = \sum_{a \in \mathcal{A}} p_a(1 - p_a).$$

A common estimator is the plug-in

$$\sum_{a \in \mathcal{A}} \hat{p}_a(1 - \hat{p}_a),$$

but this is a biased estimator.

Simpson's Index

Simpson (1949), introduced a bio-diversity index

$$\zeta_1 = \sum_{a \in \mathcal{A}} p_a(1 - p_a).$$

Instead Simpson (1949) suggested the unbiased estimator

$$Z_1 = \frac{n}{n-1} \sum_{a \in \mathcal{A}} \hat{p}_a(1 - \hat{p}_a),$$

Generalized Simpson's Indices

We now introduce a more general class of indices due to Zhang and Zhou (2010).

Generalized Simpson's Indices

We now introduce a more general class of indices due to Zhang and Zhou (2010).

A Generalized Simpson's Index of order $v \in \mathbb{N}$ is

$$\zeta_v = \sum_{a \in \mathcal{A}} p_a (1 - p_a)^v$$

Generalized Simpson's Indices

We now introduce a more general class of indices due to Zhang and Zhou (2010).

A Generalized Simpson's Index of order $v \in \mathbb{N}$ is

$$\zeta_v = \sum_{a \in \mathcal{A}} p_a (1 - p_a)^v = \mathbb{E} M_{0,v}.$$

Generalized Simpson's Indices

We now introduce a more general class of indices due to Zhang and Zhou (2010).

A Generalized Simpson's Index of order $v \in \mathbb{N}$ is

$$\zeta_v = \sum_{a \in \mathcal{A}} p_a (1 - p_a)^v = \mathbb{E} M_{0,v}.$$

Fact: The collection $\{\zeta_v : n = 1, 2, \ldots\}$ determines the distributions $\{p_1, p_2, \ldots\}$ up to permutation.

Generalized Simpson's Indices

Zhang and Zhou (2010) showed that an unbiased estimator of the Generalized Simpson's Indices of order $v = 1, 2, \ldots, (n-1)$

$$\zeta_v = \sum_{a \in \mathcal{A}} p_a (1 - p_a)^v.$$

is given by

$$Z_v = \sum_{a \in \mathcal{A}} \hat{p}_a \prod_{j=1}^{v} \left(1 - \frac{n\hat{p}_a - 1}{n - j}\right).$$

Generalized Simpson's Indices

Zhang and Zhou (2010) showed that an unbiased estimator of the Generalized Simpson's Indices of order $v = 1, 2, \ldots, (n-1)$

$$\zeta_v = \sum_{a \in \mathcal{A}} p_a (1 - p_a)^v.$$

is given by

$$Z_v = \sum_{a \in \mathcal{A}} \hat{p}_a \prod_{j=1}^{v} \left(1 - \frac{n\hat{p}_a - 1}{n - j}\right).$$

Fact: When $v = n - 1$, $Z_{n-1} = T_{0,n}$ reduces to Turing's formula.

Generalized Simpson's Indices

Zhang and Zhou (2010) and Zhang and Grabchak (2016) established the following for $v = 1, 2, \ldots, (n-1)$:

Generalized Simpson's Indices

Zhang and Zhou (2010) and Zhang and Grabchak (2016) established the following for $v = 1, 2, \ldots, (n-1)$:

- Z_v is a UMVUE for ζ_v.

Generalized Simpson's Indices

Zhang and Zhou (2010) and Zhang and Grabchak (2016) established the following for $v = 1, 2, \ldots, (n-1)$:

- Z_v is a UMVUE for ζ_v.
- So long as P is not a uniform distribution

$$\frac{\sqrt{n}(Z_v - \zeta_v)}{\hat{\sigma}_v} \xrightarrow{d} N(0, 1),$$

Generalized Simpson's Indices

Zhang and Zhou (2010) and Zhang and Grabchak (2016) established the following for $v = 1, 2, \ldots, (n-1)$:

- Z_v is a UMVUE for ζ_v.
- So long as P is not a uniform distribution

$$\frac{\sqrt{n}(Z_v - \zeta_v)}{\hat{\sigma}_v} \xrightarrow{d} N(0,1),$$

where

$$\begin{aligned} \hat{\sigma}_v^2 &= \sum_{a \in \mathcal{A}} \hat{p}_a (1-\hat{p}_a)^{2v-2} (1 - v\hat{p}_a - \hat{p}_a)^2 \\ &\quad - \left(\sum_{a \in \mathcal{A}} \hat{p}_a (1-\hat{p}_a)^{v-1} (1 - v\hat{p}_a - \hat{p}_a) \right)^2 \end{aligned}$$

An Application to Linguistics

In 1985 the following poem was discovered. It begins...

In 1985 the following poem was discovered. It begins...

Shall I die? Shall I fly
Lover's baits and deceits
sorrow breeding?
Shall I tend? Shall I send?
Shall I sue, and not rue
my proceeding?
In all duty her beauty
Binds me her servant for ever.
If she scorn, I mourn,
I retire to despair, joining never.
***

Who wrote this poem?

On it the author's name was written: **William Shakespeare.**

In 1986 it was added to the Oxford edition of the complete works of William Shakespeare.

Who wrote this poem?

On it the author's name was written: **William Shakespeare.**

In 1986 it was added to the Oxford edition of the complete works of William Shakespeare.

This was very controversial.

Did Shakespeare really write the poem?

Many literary scholars have debated this question.

Who wrote this poem?

On it the author's name was written: **William Shakespeare.**

In 1986 it was added to the Oxford edition of the complete works of William Shakespeare.

This was very controversial.

Did Shakespeare really write the poem?

Many literary scholars have debated this question.

As have some statisticians, see e.g. Thisted and Efron (1987).

Our Methodology

Let $\mathcal{A} = \{a_1, a_2, \dots\}$ be the words in the English language

Our Methodology

Let $\mathcal{A} = \{a_1, a_2, \dots\}$ be the words in the English language

Let $P = \{p_1, p_2, \dots\}$ be the relative frequencies with which an author uses the words.

Our Methodology

Let $\mathcal{A} = \{a_1, a_2, \dots\}$ be the words in the English language

Let $P = \{p_1, p_2, \dots\}$ be the relative frequencies with which an author uses the words.

We can summarize the information in P by using

$$\zeta_v = \sum_{a \in \mathcal{A}} p_a (1 - p_a)^v$$

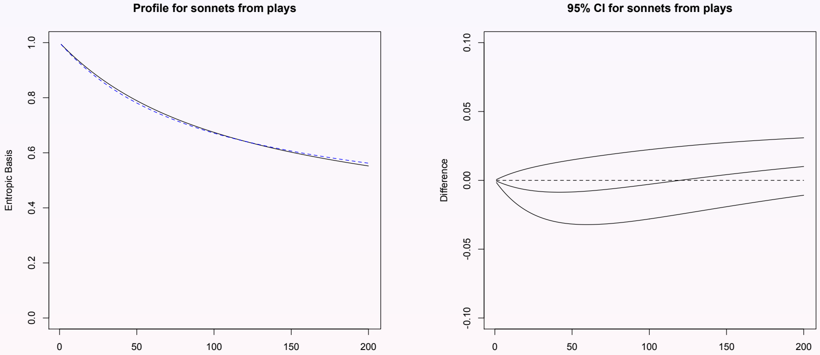
for various values of v.

Our Methodology

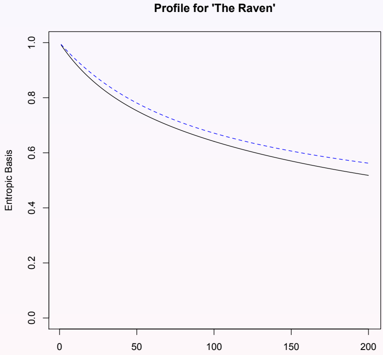
To test the authorship of “Shall I Die?” we:

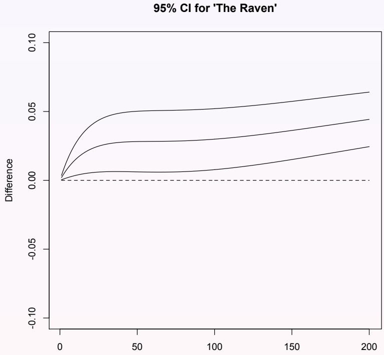
- Estimate $\zeta_1, \ldots, \zeta_{200}$ for the poem
- Estimate $\zeta_1, \ldots, \zeta_{200}$ for a corpus consisting of Shakespeare's sonnets
- Plot the difference and a confidence interval

Comparison of Sonnets and Sonnets From Plays

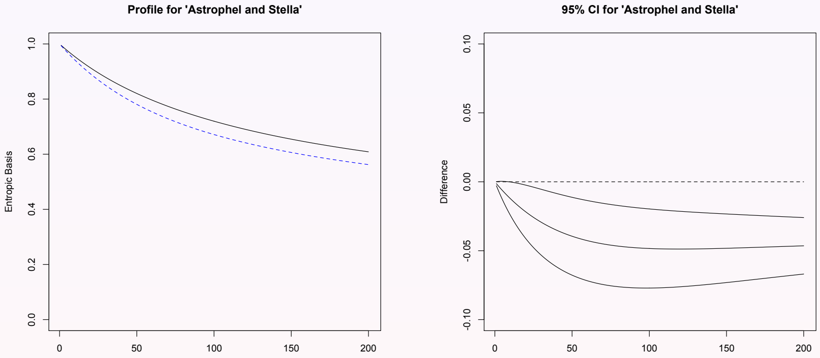


Comparison of Sonnets and The Raven

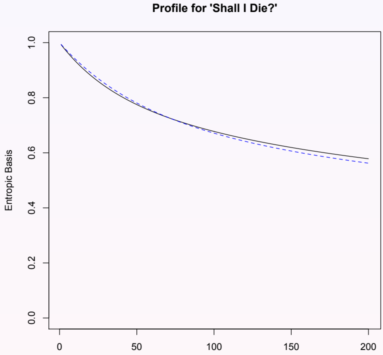


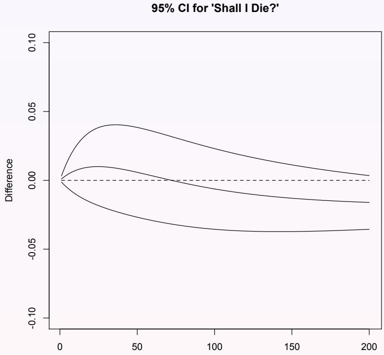


Comparison of Sonnets and Philip Sidney's Astrophel and Stella



Comparison of Sonnets and Shall I Die





Bibliography

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

M. Grabchak, L. Cao, and Z. Zhang (2017). Authorship Attribution Using Diversity Profiles. To appear in *Journal of Quantitative Linguistics*, DOI: 10.1080/09296174.2017.1343268.

M. Grabchak and Z. Zhang (2017). Asymptotic Properties of Turing's Formula in Relative Error. *Machine Learning*, 106(11):1771–1785.

Simulations

To better understand how Turing's formula works, we perform simulations. We measure performance by:

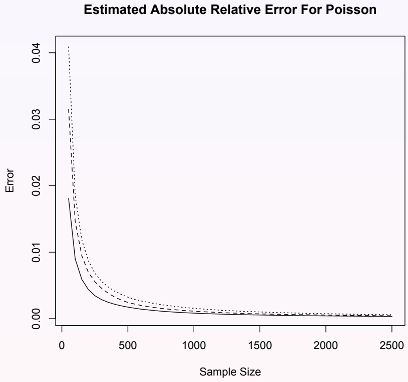
1. Expected absolute error:

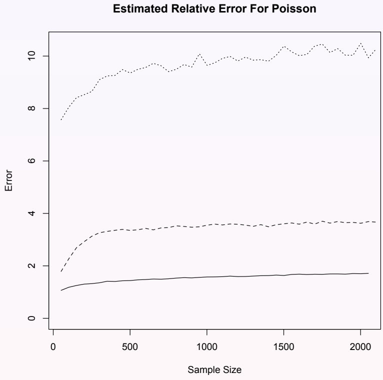
$$\mathbb{E}\left|T_{0,n} - M_{0,n}\right|$$

2. Expected relative error:

$$\mathbb{E}\left|\frac{T_{0,n} - M_{0,n}}{M_0}\right|$$

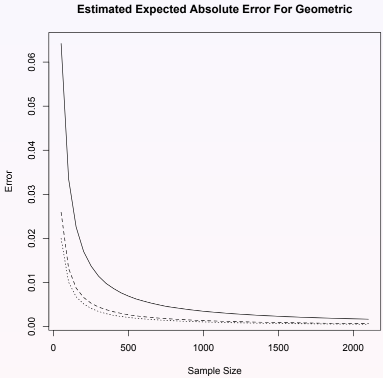
Simulations for Poisson

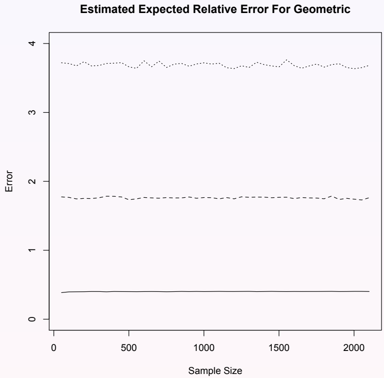




short-dashes: $\lambda = 1$, long-dashes: $\lambda = 5$, solid: $\lambda = 10$

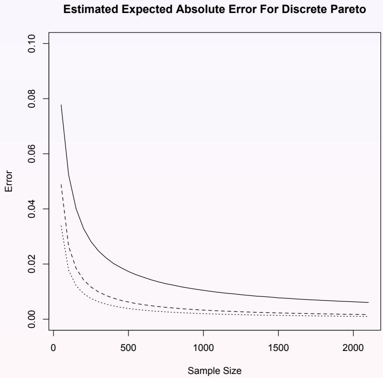
Simulations for Geometric

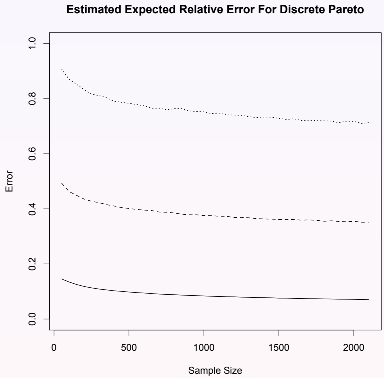




short-dashes: $p = .7$, long-dashes: $p = .5$, solid: $p = .1$

Simulations for Discrete Pareto





short-dashes: $\alpha = 10$, long-dashes: $\alpha = 5$, solid: $\alpha = 1$

Conclusions From Simulations

1. Absolute error decays quickly for all distributions. But, as we have seen, this may not be relevant.

2. Relative error is smaller for heavier tailed distributions. Only goes to zero for heavy tailed distributions.

Extension: Metric Spaces

- (E, d) is a metric space
- P is a probability distribution on E
- $X_1, \ldots, X_n$ are a random sample of E-valued random variables with common distribution P

Extension: Metric Spaces

- (E, d) is a metric space
- P is a probability distribution on E
- $X_1, \ldots, X_n$ are a random sample of E-valued random variables with common distribution P

Since E may not be discrete, we need to define analogues of occupancy probabilities, $M_{r,n}$, and the counting function ν.

Extension: Metric Spaces

Occupancy Probabilities – For $\delta > 0$, $n \geq 1$, and $x \in E$,

$$L_n^{\delta}(x) := \sum_{i=1}^{n} \mathbf{1}\{X_n \in B_{x,\delta}\}$$

Extension: Metric Spaces

Occupancy Probabilities – For $\delta > 0$, $n \geq 1$, and $x \in E$,

$$L_n^{\delta}(x) := \sum_{i=1}^{n} \mathbf{1}\{X_n \in B_{x,\delta}\}$$

Definition. For $n \geq 1$ and $0 \leq r \leq n$,

$$M_{n,r}^{\delta} = \mathbb{P}(L_n^{\delta}(X_{n+1}) = r | X_1, \ldots, X_n) = \int_E \mathbf{1}\{L_n^{\delta}(x) = r\} P(\mathrm{d}x)$$

Extension: Metric Spaces

Occupancy Probabilities – For $\delta > 0$, $n \geq 1$, and $x \in E$,

$$L_n^{\delta}(x) := \sum_{i=1}^{n} \mathbf{1}\{X_n \in B_{x,\delta}\}$$

Definition. For $n \geq 1$ and $0 \leq r \leq n$,

$$M_{n,r}^{\delta} = \mathbb{P}(L_n^{\delta}(X_{n+1}) = r | X_1, \dots, X_n) = \int_E \mathbf{1}\{L_n^{\delta}(x) = r\} P(\mathrm{d}x)$$

Fact: If P has a discrete support with no accumulation point, then

$$M_{n,r}^{\delta} \underset{\delta \to 0+}{\longrightarrow} M_{n,r}$$

Extension: Metric Spaces

δ-Counting Function: For $\delta > 0$ define

$$\mathcal{L}_\delta(\varepsilon) = \{x \in E : P(B_{x,\delta}) \geq \varepsilon\}$$

and

$$\nu_\delta(\varepsilon) = \int_{\mathcal{L}_\delta(\varepsilon)} P(B_{x,\delta})^{-1} P(\mathrm{d}x).$$

Extension: Metric Spaces

δ-Counting Function: For $\delta > 0$ define

$$\mathcal{L}_\delta(\varepsilon) = \{x \in E : P(B_{x,\delta}) \geq \varepsilon\}$$

and

$$\nu_\delta(\varepsilon) = \int_{\mathcal{L}_\delta(\varepsilon)} P(B_{x,\delta})^{-1} P(\mathrm{d}x).$$

Theorem. If P has a discrete support with no accumulation point, then for any $\varepsilon \in (0, 1]$,

$$\nu_\delta(\varepsilon) \underset{\delta \to 0+}{\longrightarrow} \nu(\varepsilon).$$

Extension: Metric Spaces

Fact: In this framework, most of the results from this section still hold. We just need to replace ν by ν_δ.

Extension: Metric Spaces

Fact: In this framework, most of the results from this section still hold. We just need to replace ν by ν_δ.

Future work: Can Turing's formula and concentration inequalities be extended to this framework?