The Missing Mass Problem: Number of Unseen Species, Enigma and NLP

Joshua Dall'Acqua
Supervised by Prof. M. Asgharian
Dept of Math and Stat, McGill University
joshua.dallacqua@mail.mcgill.ca

***Abstract*— It is conceivable that a random sample of a population consisting of various species likely fail to include a representative from all the species in the population. For example, if our sample is a piece of text, and the population is a language, it is quite likely that the text won't contain every word in the language. It is natural to wonder about the parts of the population that are not seen by the sample and question how representative the sample is. Our aim in this manuscript is to present methods for estimation of the proportion of unseen species and various application of this general problem along with their historical context. The most striking case for these techniques took place during the second world war in the British effort to crack the German's enigma code. An estimator, which we will look at deeply, known as the Good-Turing estimator was discovered out of necessity by Allen Turing and proved to be invaluable in breaking the code. It may seem like some of the estimators we will discuss are designed to solve the same problem. We will attempt to shed some light on their differences through experiments on a curated data set.**

I. The Missing Mass Problem

Suppose we take a sample of size n from a population with s mutually exclusive categories where s is unknown. In other words, we have a sample $\boldsymbol{X} = X_1, ..., X_N$ where $X_i \in C_j$, for $i = 1...N$ and $j = 1, ...s$. In the sequel, we will refer to the observations as specimens and the categories as species. A general question we often try to address in statistical and data science is "what can we learn about the population as a whole?" For example, it might be desirable to know how many total species there are. And more specifically, if each observation is an animal from a specific region, we may want to know how many different species living in that region. We may also want to predict the category to which the next observation belong or estimate the probability that the next observation belongs to a given class of species, perhaps species not seen before. Various approaches to answering these, and similar, questions will be explored in this paper.

To answer the first question, we need to have an estimate of the number of unseen species. To mathematically formalize, and slightly generalize this question, we can express the first question as finding an estimator T such that

$$\mathbb{E}[T(\boldsymbol{X})] = \boldsymbol{p} \tag{1}$$

where $\boldsymbol{p} = [p_0, ..., p_l]$ is a vector representing the population frequency for each species frequency. This means each p_i represents the probability that if we make another observation, it will be of a species which has already appeared i times in our sample and that

$$\sum_{i=0}^{l} p_i = 1 \tag{2}$$

p_0 represents the probability that the next observation has not already been seen in the sample; this is referred to as the "Missing Mass". The answer to this question is the Good-Turing estimator [1] which we will look at in depth later.

Related to the above formulation, we may wish to find an estimator T_k such that

$$\mathbb{E}[T_k(\boldsymbol{X})] = U \tag{3}$$

where U represents the number of new species we can expect to observe if we take another sample of size $k \cdot N$ from the same population. There exists a unique unbiased estimator for this problem found by Good and Toulmin [2] also to be examined in later sections.

II. Origins

In the 1930s, Corbet and Williams were both conducting experiments which involved collecting bug specimens and classifying them by species. Before 1942, there had been no attempt to find a relationship between the number of individuals and the number of species in a sample. In 1943 Corbet, Williams, and Fisher came out with a paper titled "The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population" [3]. Corbet found he could relate the number of individuals to the number of species in a Mylan Butterfly population using the formula

$$S = \frac{C}{N^m} \tag{4}$$

where S is the number of species, C and m are constants, and N is the sample size. Fisher extended this idea finding a model which only depends on one constant α to represent the richness or diversity of a species and obtained the formula

$$S = \alpha \ln\left(1 + \frac{N}{\alpha}\right) \tag{5}$$

The constant α can be found graphically. A higher value of α indicates a more diverse population. We will later compare (5) to more modern approaches by Good, Toulmin and Turing.

III. Enigma

A. What was Enigma?

During Word War II, the Germans used a device called an "Enigma Machine" to encode messages containing sensitive information. The machine had over 150 trillion possible settings, each providing a unique encryption/decryption mapping. The code was notorious for being unbreakable but that didn't stop the British from trying. They assembled a team of code breakers at a mansion in the countryside which housed a cryptography institute known as Bletchley Park. Amongst these code breakers was Allen Turing. Turing is widely credited as being the keystone in the effort of cracking Enigma and became the creator of the first computer or "Turing Machine" in the process. The team finally broke the code in 1941 and was able to provide intelligence to the British forces instrumental to the success of the allies. It is said that without this information, the war would have gone on for years longer than it did and there would have been no guarantee of the allies winning [4].

B. The Enigma Machine

The Enigma machine was used for both encoding and decoding messages. The machine itself was of no use in decoding a message unless the recipient knew what settings were used to encode the message. This machine was not kept secret by the Germans, and the allies got their hands on one very early in the war. The Germans had so much confidence in Enigma that they did not try to send the encrypted messages secretly and many were intercepted. The machine itself consists of 3 rotors (which can be chosen out of a set of 8 total rotors) to scramble the letters and a pegboard which performs additional scrambling. The machine seen in Figure 1 is an Enigma I machine. There are other versions of Enigma but this was the one which the team at Bletchley Park was focused on cracking. It has a total of 103,325,660,891,587,134,000,000 possible encryptions [5].

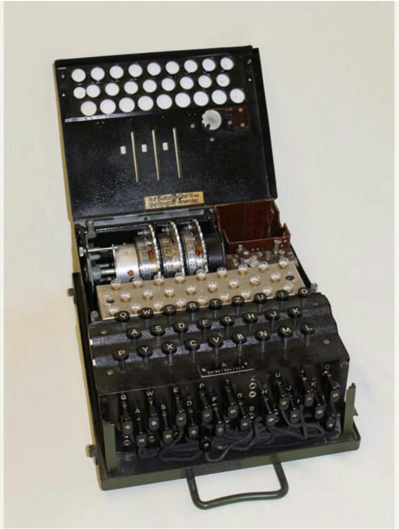

Figure 1: Three rotor Enigma I Machine on display at the Bletchley Park museum

C. Tri-gram Frequency Estimation

To decode Enigma, the recipient needs to apply the same settings to their Enigma Machine that the sender used to encrypt the message. Each message includes a tri-gram (three letters) at the beginning which tells the recipient how to set the rotors on their Enigma Machine. More settings must be applied and more steps must be taken to decode a message but this rotor setting is what is relevant to us. The trigrams used are supposed to be random, and taken from a book provided by German intelligence. Turing realized that this was not entirely true and some tri-grams, like those appearing at the tops of the pages in the book, were more likely to be used [6]. We think of each tri-gram as a species, and every possible tri-gram in the German book as being the population. Another random letter was added to each tri-gram before it was sent in the message, so the tri-grams were slightly obfuscated. The team at Bletchley Park realized it would be very helpful to know which tri-grams had a high probability of appearing next. It would serve as a starting point for guessing the tri-gram in each message. This is where Turing first came up with his method for unknown species frequency estimation. The estimator is called the "Good-Turing estimator" and will be discussed in more details in the next section.

D. Good-Turing NLP

The Good-Turing method was developed to study the frequency of tri-grams and continues to be useful for similar tasks. in 2013, Huang et al published a paper [7] looking at the efficacy in using the Good-Turing estimator to solve the zero weights problem in Chinese language models. The zero weights problem arises while training language models when the training corpus does not contain an example of every token in the population. These unseen tokens are assigned a weight of zero which inaccurately represents their probability of appearing outside the training corpus. The smoothing methods discussed later in Section V.C are designed to provide a non zero estimate for these unseen values. The paper by Huang et al [7] looks at Chinese n-gram models. Due to the large number of characters in the Chinese alphabet, many tokens fall victim to the zero weight problem. The paper finds Good-Turing smoothing provides an effective remedy to this issue.

IV. Other Notable Applications

A. Palomar Green Survey

In 1988 Peter Thejll and H.L. Shipman [8] wrote a paper which uses Fisher's method [3] to gain insight on data collected in the Palomar Green astronomical survey. Considering classes of astronomical objects as species, one can see the direct applicability of using this type of estimator. Their goal was to provide an estimate for the size of survey which would be needed to discover new objects in each class. They concluded that if the Palomar Green Survey would be repeated at the same scale on another part of the sky, it is likely that one or two types of new objects would be discovered. Having access to this information plays a role in discussions on where to allocate resources for future experiments. In their paper, Thejll and Shipman also analyzed observations from the McGraw Transit Telescope. They found that it was unlikely for any new classes of white dwarfs would appear, but expected to see roughly twelve unseen before galaxy like objects.

B. Efron Thisted + Shakespeare

B. Efron and R. Thisted used this idea to give an estimate for the total number of words types that Shakespeare knew [9]. They took the sample to be his collected works, and were able to find a lower bound that Shakespeare knew at least 35 000 more word types than the 31 543 that originally appeared in his works. This means that the number of unique species in the population (number of words Shakespeare knew) was over 67 000. In this paper, Efron and Thisted also re-designed the Good-Toulmin estimator which we will talk about later in Section VI.B.1 and were able to greatly improve its efficacy.

V. Good-Turing Estimator

A. A Naive estimator

If we are given a sample of N species, intuitively, we might guess that the expected value of observing a species which appears r times in the sample can be estimated as

$$\mathbb{E}[q_r] = \frac{r}{N} \tag{6}$$

where q_r is the event that we observe a species appearing r times in the sample. It makes sense that the probability of observing a species should be proportional to how common it is. Unfortunately, this simple estimate does not tell the whole story. Unless the sample is sufficiently large, there will be species that do not appear at all. The "naive" estimator in (6) assigns these unseen species a probability of zero and misrepresents the underlying distribution.

B. An Improved Estimator

In 1953 I.J. Good published a paper [1] presenting Turing's solution to the problem. He realized that the number of species which appear r times, say n_r, should differ from the number of species which appear $r+1$ times, n_{r+1}, by an amount proportional to r. His improved estimate is that

$$\mathbb{E}[q_r] = \frac{r^*}{N} \tag{7}$$

where

$$r^* = (r+1)\frac{n_{r+1}}{n_r} \tag{8}$$

We can see that r^* is essentially a proxy for r which takes into account the value of n_{r+1}.

C. Smoothing

The Good-Turing estimator which we have discussed so far makes an estimate for q_r depending only on r, n_{r+1}, and n_r. All of the information we use is taken from a two data points in the observed distribution of n_r's. If the sample size is small, the data at this point may be noisy making it a poor representation of the actual distribution. The solution to this is to smooth the observed n_r's before feeding them into the estimator. Each n_r is the cardinality of a bucket containing all the species observed r times. The idea of smoothing is that we can move some species in bucket r to its neighboring buckets $(r-t)$ and $(r+t)$, for reasonable values of t, with the goal of removing noise from the distribution. There are lots of techniques that can be used to smooth the observed n_r's but many can be computationally expensive or difficult to estimate. The most common smoothing method is called linear smoothing [10]. It uses the assumption that the frequency of species frequencies is modeled by exponential decay. This allows for a linear fit when the n_r's are plotted on a log-log scale. There is not enough data for this trend to be visible at higher values of r so the data used for the linear fit is truncated to only include values of r which are sufficiently represented. Once the fit has been obtained, the buckets can be rebalanced and the linear trend can be extrapolated to include the larger values of r which are often less represented.

D. *Properties*

1) *Bias:* in 1994 B. H. Juang and S. H. Lo published an analysis of bias [11] of the Good-Turing estimator. They found that the order of the bias is

$$O\left(\frac{1}{N}\right) \tag{9}$$

where N is the sample size. For this reason, the estimator is considered to be "nearly unbiased", and unbiased for large values of N.

2) *Variance:* In the original paper by Good [1], it is proven that the variance of the estimator is given by

$$\mathrm{Var}(T_r) = \frac{(r+1)(r+2)}{N^2}\frac{n_{r+2}}{n_r} - \left(\frac{r+1}{N}\frac{n_{r+1}}{n_r}\right)^2 \tag{10}$$

where T_r is the estimate of the rth population frequency, n_r is the number of species which are observed r times in the sample, and N is the sample size.

3) *Consistency:* An estimator T_n which estimates a parameter θ is strongly consistent if

$$\Pr\left\{\lim_{n\to\infty} T_n = \theta\right\} = 1 \tag{11}$$

Meaning that the estimate converges almost surely to the actual value. Proving the consistency of the Good-Turing estimator is difficult as the distribution of the species frequencies depends on the sample size. If the sample is small, a large number of species will have been unseen and the missing mass p_0 will be large. As more observations are made, p_0 will shrink because an increasing number of species will have at least one representative in the sample. Aaron B. Wagner et al were able to prove that the estimator is strongly consistent in 2006 [12].

4) *Rate of Convergence:* In 2000 David McAllester and Robert E. Schapire [13] (2000) found PAC (Probably Approximately Correct) confidence intervals for the population frequencies (p_i's) and the missing mass. Let T_r be the Good-Turing estimate for the population frequency of r, they found that for any sample of size N, with probability greater than or equal to $1 - \delta$, the following holds.

$$\begin{aligned} |T_r - p_r| &< \frac{r+2}{N-r} + \sqrt{\frac{2\ln\left(\frac{3}{\delta}\right)}{N}} \\ &\times \left[\frac{r+1}{1-\frac{r}{N}} + r + \ln\left(2r\ln\left(\frac{3N}{\delta}\right)\right) + 2\ln\left(\frac{3N}{\delta}\right)\right] \end{aligned} \tag{12}$$

and that the missing mass has a tighter upper bound of

$$p_0 \le T_0 + \left(2\sqrt{2}+\sqrt{3}\right)\sqrt{\frac{\ln\left(\frac{3}{\delta}\right)}{N}} \tag{13}$$

E. *Implementation*

The implementation used for the experiments to come later in this paper is based on the method described in the paper titled "Good-Turing Frequency Estimation Without Tears" [10]. It uses the method of linear smoothing and provides a guide to implement this technique on modern hardware.

VI. Good-Toulmin Estimator

A. *Original Estimator*

We have discussed the Good-Turing estimator as a means for estimating the population frequencies of each species. We find the probability that the next observation will be from a specific species. Another question we can ask is if we increase the sample size by a factor of k, what proportion of the new sample will have been unseen in the old sample. This is the purpose of the Good-Toulmin estimator [2] Introduced in 1956. This paper builds on many of the ideas in Good's earlier paper on the Good-Turing estimator [1].

The main idea of the estimator is very simple and elegant: If we count up all the species which appear an even number of times or an odd number of times, in the absence of any other knowledge about the frequencies, it is reasonable, according to the Laplace's Principle of Insufficient Reason (also called Principle of Indifference), to expect the same value. Formally, this can be expressed as

$$U_k^{\mathrm{GT}} = -\sum_{i=1}^{R} (-k)^i \cdot n_i \tag{14}$$

where U_k^{GT} is the number of new species that will be observed if the sample size is increased by a factor of k, and R is the largest r for which n_r is non-zero.

B. *Improvments*

The above estimator found good results when $k \le 1$. The MSE of this estimator is given by

$$\mathbb{E}\left[U - U_k^{\mathrm{GT}}\right]^2 \le nk^2 \tag{15}$$

The problem is, when $k > 1$, the error grows very quickly or "super-linearly". The original paper [2] discusses possibilities to rectify this divergence but doesn't provide any formal results.

1) *Efron-Thisted:* In 1976, Efron and Thisted [9] Improved on the work of Good and Toulmin and found an estimator which could make good predictions in some cases where $k > 1$ using the Euler transform. While their method had good results in practice, it proved difficult to bound theoretically.

2) *Orlitsky:* In 2016, Orlitsky et al [14] published a result which found an estimator with provably good performance for values of $k \ge 1$. The paper provides a class of estimators which give provably good predictions for values of k proportional to the logarithm of the sample size.

Orlitsky's Improved estimator can be expressed as the following linear combination of prevalences

$$U_k = -\sum_{i\geq 1} (-k)^i \Pr(L_k > i) \cdot n_i \tag{16}$$

where L is a tail distribution. This estimator accomplishes the same thing as randomly truncating the alternating series so that the last term does not dominate the sign. Orlitsky also found that by taking $L_k \sim \text{Bin}\left(k, \frac{1}{1+k}\right)$, the resulting estimator will be equivalent to Efron and Thisted's estimator. He discovered the optimal tail distribution to be $L_k \sim \text{Bin}\left(\left\lfloor \frac{1}{2}\log_3\left(\frac{nk^2}{k-1}\right)\right\rfloor, \frac{2}{k+2}\right)$. We will refer to this version of the estimator as Smoothed Good-Toulmin (SGT). The appendix of Orlitsky's paper [14] also contains the proof that the original Good-Toulmin estimator is the unique unbiased estimator for U_k.

VII. Experiments: Estimating the Number of Unseen Species

A. Benchmark Data

The data-set which will be used to compare performance of the estimators will be drawn from the beta-binomial distribution with parameters $\alpha = 1, \beta = 6.5$ for reasons to be discussed in this section. If we want to accurately test the estimators, it is important to know the data generating process. It enables us to calculate the expected values of species frequencies and compare them against those found by the estimators. The beta-binomial distribution allows for the binomial assumption to be made which is a condition for using the Good-Turing estimator [1]. It also has two shape parameters enabling us to control the shape of the data. The parameters have been chosen to mimic natural language data which is the most common application of the Good-Turing and Good-Toulmin estimators. This is illustrated in Figure 2 and Figure 3 where we look at the normalized frequency frequencies (n_r) for our test data set compared to frequency frequencies found in Shakespeare's Macbeth. This is a useful representation of the data as the estimators discussed use the species frequency frequencies to make their estimates.

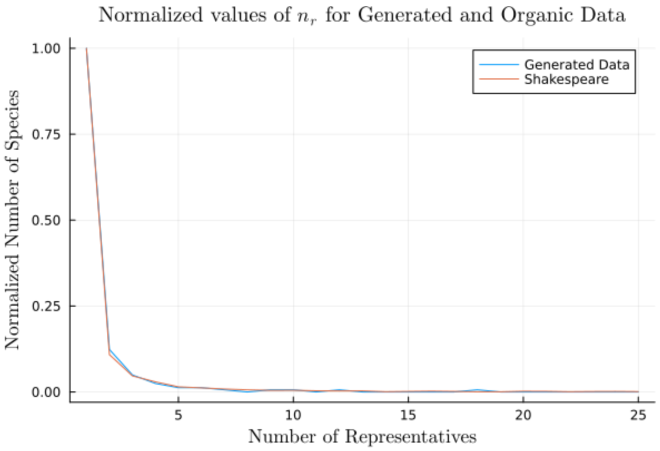

Figure 2: Comparison of the normalized number of species with r representatives coming from BetaBinomial($\alpha = 1, \beta = 6.5$) and Shakespeare's Macbeth

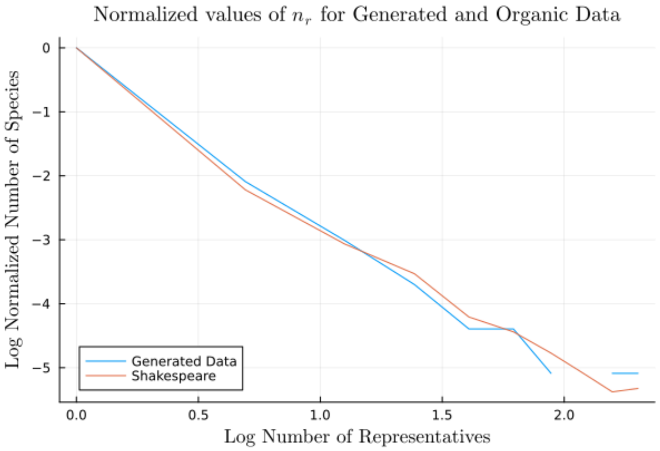

Figure 3: A log-log plot of normalized number of species with r representatives from each dataset

B. Toulmin, Turing and Fisher

1) *Estimators:* The Good-Toulmin estimator answers the question of what proportion of a sample of size $k \cdot N$ will have been unseen by the existing sample (of size N). The Good-Turing estimator gives the probability of seeing a specific species which has already been seen some number of times in the sample (or not at all). It is possible to use the probabilities found by Good-Turing to estimate the number of unseen species in a sample of size $k \cdot N$.

Recall that the Good-Turing estimator outputs values p_i which is the probability that the next observation will belong to a species which has been seen i times in the original sample. The value p_0 corresponds to the probability that the next observation will have not yet been seen in the sample. We can estimate U, the number of new species that will be found in a new sample of size N as follows

$$U = p_0 \cdot N \tag{17}$$

This is equivalent to setting $k = 1$ when using the Good-Toulmin estimator.

Fishers method gives equation (5) which also relates the size of the sample to the number of unique individuals S. Using a sample of size N we can find a value for α and predict S for a sample of size $2 \cdot N$.

2) *Comparison:* Figure 4 shows the estimates made by each method on our benchmark dataset for $k = 1$. The values plotted are the predicted numbers of unique species seen by a sample of half the size. For example, at $x = 1000$, the y -value corresponds to a prediction for the number of unique species in a sample of size 500 made by an estimator with access to 250 data points. A Monte-Carlo simulation was carried out in the following manner: each estimator was given 1000 randomly generated samples of the same size from the benchmark data and the outputs were averaged to get the values seen in Figure 4 and Figure 6.

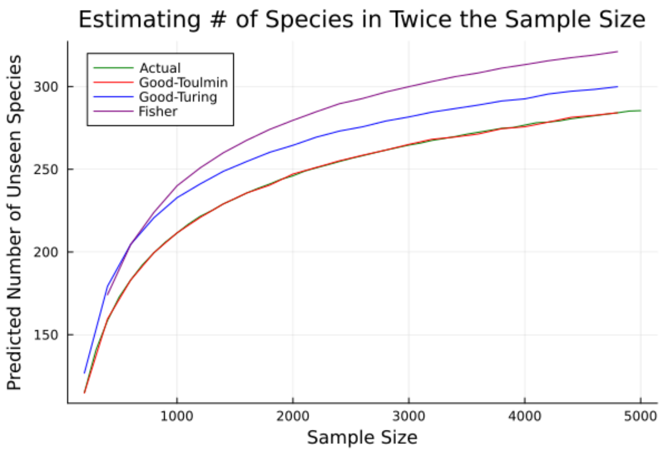

Figure 4: Estimator predictions for number of unique species seen using a sample of size $\lfloor N \div 2 \rfloor$

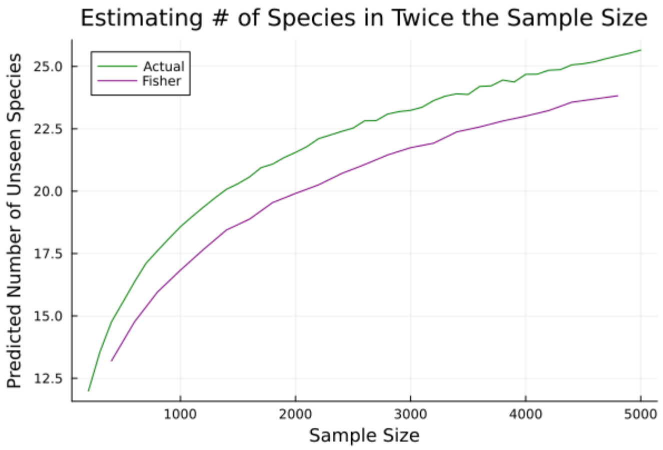

Figure 5: Estimator predictions for number of unique species seen using a sample of size $\lfloor N \div 2 \rfloor$ with data from $\mathrm{BetaBinomial}(\alpha = 0.2, \beta = 100)$

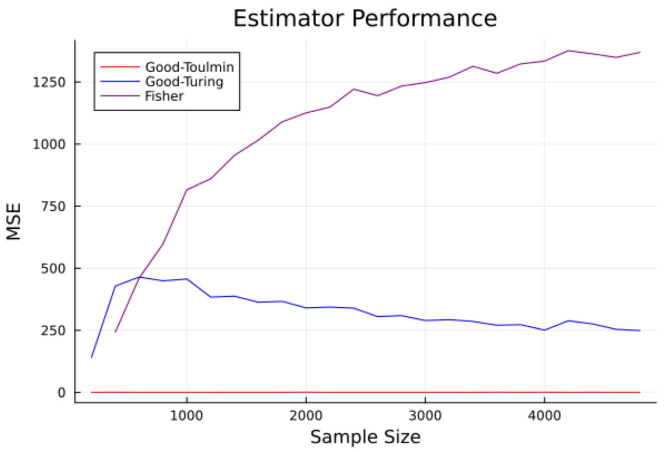

Figure 6: MSE of estimators seen in Figure 4

3) *Discussion:* Figure 4 shows the Good-Toulmin estimator as being indistinguishable from the actual value. In Section VI.B.2 we saw that Orlitsky [14] found Good-Toulmin to be the unique unbiased estimator. Therefore, over a long run averaging simulation like this, it makes sense that the estimate should approach the actual value.

In Figure 4, the Good-Turing estimator Consistently overestimates the number of new species that will appear in a sample of twice the size. This is due to an inaccuracy in the way which we are using the estimator outlined in Section VII.B.1. The probability p_0 given by the Good-Turing estimator is the probability that the **next** observation will be previously unseen. We assumed that this value would stay constant for the next N observations when in reality it would change as we observe more information about the population. For this reason, the Good-Toulmin estimator should be used in place of the Good-Turing estimator when estimating the number of unseen species in a larger sample.

Fisher's estimator is more primitive than the other two and based on a very different method. We see in Figure 4 it consistently over estimates the actual value. This however, depends on the distribution. Looking at Figure 5, we can see that when the beta-binomial parameters are set to $\alpha = 0.2$, $\beta = 100$, the estimator underestimates the actual value. For some distributions, Fisher's estimator will be very close to the actual but is not as reliable as Good and Toulmin's.

C. *Improved Good-Toulmin Estimators*

1) *Efron Thisted and Orlitsky:* We have seen that the obvious choice for estimating how many unique species there will be in a larger sample is the Good-Toulmin estimator. However, as discussed in Section VI.B the standard Good-Toulmin estimator is still only capable of making good estimates for samples at most twice the size of the original. Here we will compare the original estimator to the improved versions by Efron and Thisted [9] and Orlitsky [14] who claim

to be able to make accurate estimates for larger multiples of the sample size.

2) *Experiment:*

Using the benchmark dataset, a sample of size 1000 was generated and used by each estimator to predict the number of new species that would be seen in a sample of size $N \cdot (1+k)$. Efron and Thisted's estimator is labled as ET and Orlitsky's smoothed estimator is labled as SGT.

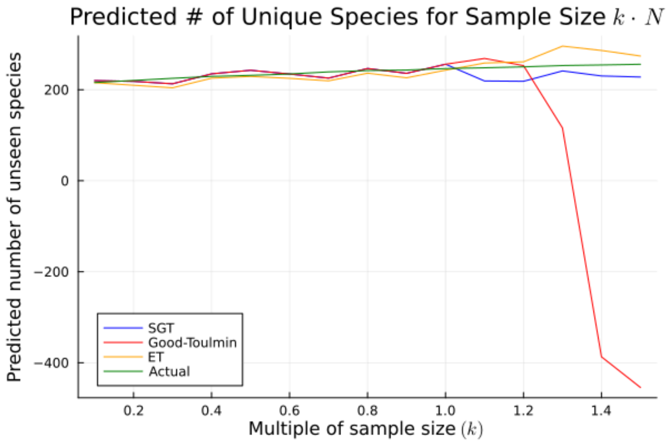

Figure 7: Comparison of Good-Toulmin type estimators estimating the number of unique species in a sample of size $N \cdot (1+k)$

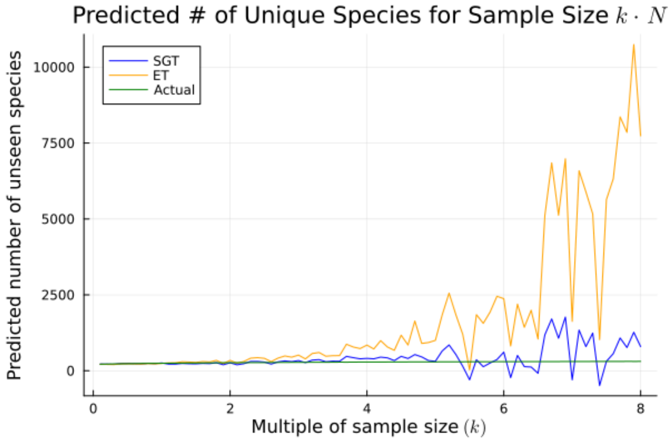

Figure 8: Comparison of improved Good-Toulmin type estimators estimating the number of unique species in a sample of size $N \cdot (1+k)$

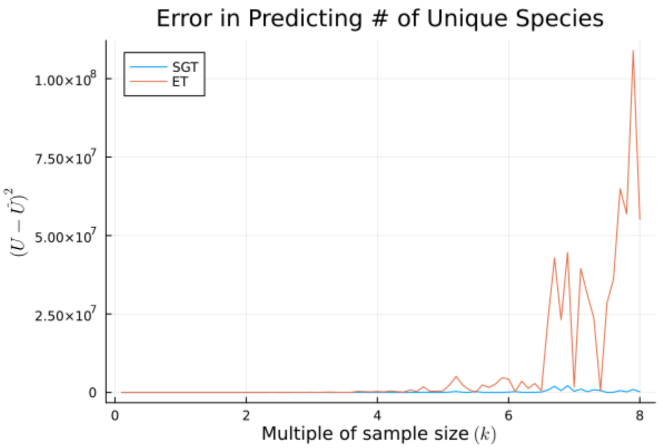

Figure 9: Error in improved Good-Toulmin type estimators

3) *Discussion:* As discussed in section Section VI.B, the standard Good-Toulmin error has super-linear MSE when $k > 1$. This is exactly what we see in Figure 7 where the Good-Toulmin estimator's estimate becomes unusable for $k > 1.2$.

Looking at Figure 8 and Figure 9, we see that both ET and SGT do significantly better and are able to make usable estimates up until around $k = 4$. After this point, ET begins to consistently over estimate and is no longer usable. SGT is able to stay relatively close to the actual value right up to $k = 8$ but begins to vary more for values of $k > 6$. Orlitsky's paper [14] claims that SGT should be able to make good estimates up to $k = \log(N)$. For a sample of size 1000, we get $\log(1000) \approx 6.9$. When $4 < k < 6.9$, we can see in Figure 9 that the errors for SGT are lower than ET but more importantly, they does not appear to be diverging from the actual value.

The purpose of this experiment has been to illustrate the performance differences between different implementations of Good-Toulmin estimators, not to formally verify any results. Formal proofs can be found in Orlitsky's paper [14].

VIII. Conclusion

We have seen how the wish of building better picture of a butterfly population from a small sample has grown into a rich area of study. Though the initial ideas presented by fisher were rudimentary, they provided an entry point for others to study the Missing Mass Problem. Turing made the next major leap in the effort to crack Enigma during WWII. Good not only formalized and presented the work of Turing, but also worked with Toulmin to expand its utility coming up with a new class of "Good-Toulmin" estimators. In the following years improvements were made to this class of estimators, most notably by Efron and Thisted who greatly improved on Good and Turings work, but still, many properties of the estimators in this class had no theoretical backing. Orlitsky brought us formal proofs as well as an optimal estimator.

Finally, we saw some experiments that show why the above advancements were necessary. While Fisher's model was capable of making viable predictions, it was susceptible to being biased. Good-Turing was not designed to estimate the number of unseen species, and naively adapting it for this purpose was not enough. These shortcomings gave us the Good-Toulmin estimator. Still, it was only capable of making predictions for samples up to twice the size. This issue was later solved in the work of Orlitsky. Building on results from Efron and Thisted, he lets us predict the number of unknown species in a larger, theoretical sample with size up to $N \cdot \log(N)$ where N is the size of the original sample.

References

[1] I. J. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, no. 3–4, pp. 237–264, 1953.

[2] I. J. Good and G. H. Toulmin, "The number of new species, and the increase in population coverage, when a sample is increased," *Biometrika*, vol. 43, no. 1–2, pp. 45–63, 1956.

[3] R. A. Fisher, A. S. Corbet, and C. B. Williams, "The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population," *Journal of Animal Ecology*, vol. 12, no. 1, pp. 42–58, 1943, Accessed: Jun. 20, 2024. [Online]. Available: http://www.jstor.org/stable/1411

[4] Wikipedia, "Bletchley Park — Wikipedia, The Free Encyclopedia." 2024.

[5] B. Park, "Enigmas of Bletchley Park." 2021.

[6] Y. K. Fadoua Balabdaoui, "The Enigma behind the Good-Turing formula," 2021.

[7] F.-L. Huang, M.-S. Yu, and C.-Y. Hwang, "An Empirical Study of Good-Turing Smoothing for Language Models on Different Size Corpora of Chinese," *Journal of Computer and Communications*, vol. 1, pp. 14–19, 2013, doi: 10.4236/jcc.2013.15003.

[8] P. Thejll and H. L. Shipman, "BUTTERFLY STATISTICS, ASTRONOMICAL SURVEYS, AND THE DISCOVERY OF NEW CLASSES OF OBJECTS," *Publications of the Astronomical Society of the Pacific*, vol. 100, no. 625, p. 398–399, Mar. 1988, doi: 10.1086/132184.

[9] B. Efron and R. Thisted, "Estimating the Number of Unsen Species: How Many Words Did Shakespeare Know?," *Biometrika*, vol. 63, no. 3, pp. 435–447, 1976, Accessed: Jun. 29, 2024. [Online]. Available: http://www.jstor.org/stable/2335721

[10] W. A. Gale and G. Sampson, "Good-turing frequency estimation without tears," *Journal of quantitative linguistics*, vol. 2, no. 3, pp. 217–237, 1995.

[11] B. Juang and S. Lo, "On the bias of the Turing-Good estimate of probabilities," *IEEE Transactions on Signal Processing*, vol. 42, no. 2, pp. 496–498, 1994, doi: 10.1109/78.275640.

[12] A. B. Wagner, P. Viswanath, and S. R. Kulkarni, "Strong Consistency of the Good-Turing Estimator," vol. 0, no. , pp. 2526–2530, 2006, doi: 10.1109/ISIT.2006.262066.

[13] D. A. McAllester and R. E. Schapire, "On the Convergence Rate of Good-Turing Estimators.," *COLT*, pp. 1–6, 2000.

[14] A. Orlitsky, A. T. Suresh, and Y. Wu, "Optimal prediction of the number of unseen species," *Proceedings of the National Academy of Sciences*, vol. 113, no. 47, pp. 13283–13288, 2016, doi: 10.1073/pnas.1607774113.