BREAKING THE RULES to CONTEMPLATE NEW APPLICATIONS
Cornell Engineering

Good-Turing Estimator and its Application to Data Analytics for Real-Time Electricity Markets

Alice Chen xc278@cornell.edu School of Electrical and Computer Engineering, Cornell University

Abstract

Getting accurate electricity price estimates is critical to electricity market participants. However, electricity prices can be highly unpredictable, especially with the rapid development of renewable energy in recent years. Estimating its probability distribution is also hard because not much data is useful to be included in the sample.

This research looks at the Good-Turing estimator and one of its variants, the Simple Good-Turing (SGT) estimator, which works particularly well when estimating the probability distribution of a random event given only sparse data. A detailed (and new) proof of the central formula in the Good-Turing estimator as well as the intuition for the SGT estimator is given, and the performances of the SGT, empirical, and Laplace estimators are compared for samples of different kinds. The fact that the SGT estimator is doing better overall suggests that the Good-Turing estimator and its variants may yield relatively more accurate results when applied to real electricity prices.

Estimators

An estimator is a rule for calculating an estimate of a given quantity based on observed data. The empirical, Laplace, and Good-Turing (see Ref. 2) estimators can be used to estimate the population frequencies of each of the s species given a sample of size N and are defined as follows:

$$\hat{q}_r^e = \frac{r}{N} \tag{1}$$

$$\hat{q}_r^l = \frac{r+1}{N+s} \tag{2}$$

$$\hat{q}_r^{gt} = \begin{cases} \frac{n_1}{N}, & \text{if } r = 0 \\ \frac{(r+1)n_{r+1}}{Nn_r}, & \text{if } r > 0 \end{cases} \tag{3}$$

where r is the number of times a species is represented in the sample, and n_r is the number of species each represented r times in the sample.

Why Good-Turing?

Suppose we want to know the population frequency of each species in a jungle, and all information we have is the snapshot below: 1 rhino, 1 cheetah, 1 orangutan, 1 snake; 2 gazelles, 2 ducks, 2 lions, 2 elephants, 2 giraffes, 2 zebras; 3 parrots, 3 flamingos; and 6 butterflies.

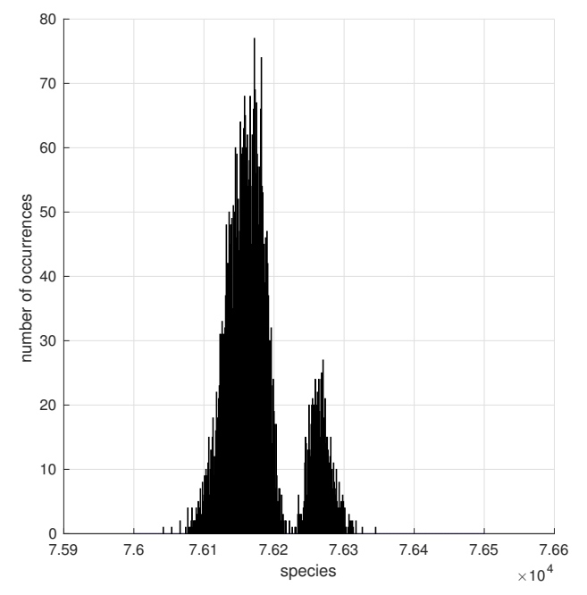
From the snapshot we can calculate the estimated population frequencies of the jungle species referenced by r using the three estimators previously defined:

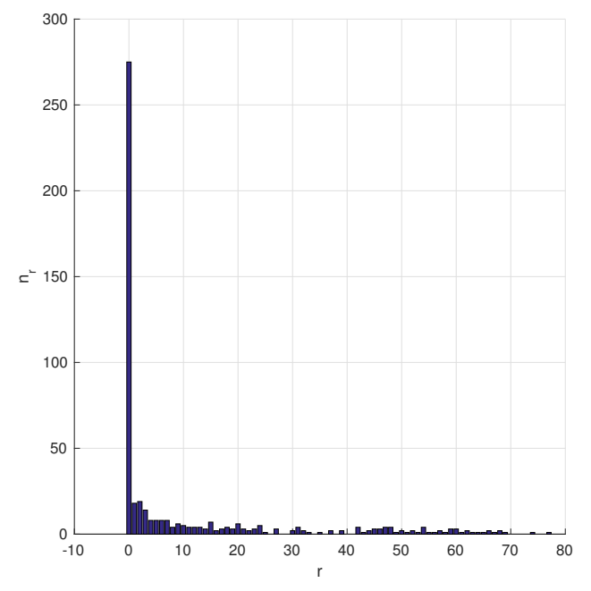
r	n_r	$\hat{q}_r^e$	$\hat{q}_r^l$	$\hat{q}_r^{gt}$
0	(1)	0/28	1/42	(4/28)
1	4	1/28	2/42	3/28
2	6	2/28	3/42	1/28
3	2	3/28	4/42	0/28
5	0	0/28	0/42	(1.2/28)
6	1	6/28	7/42	0/28

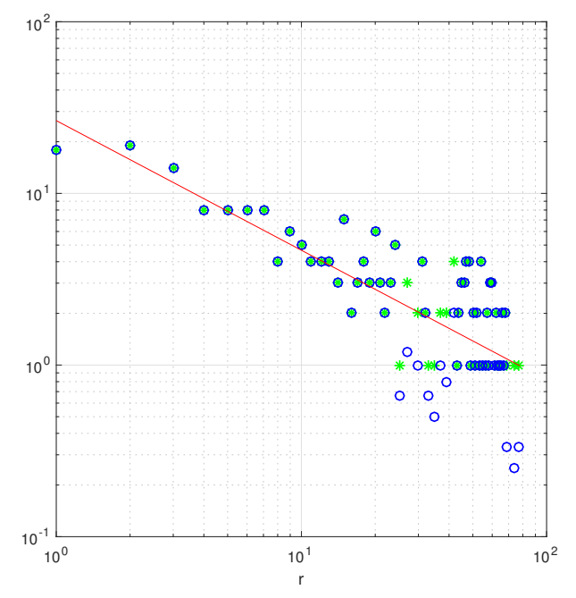
When given sparse data, the empirical estimator simply assigns zero probability to all unseen species, which can be a notable source of inaccuracy. The Laplace estimator tries to avoid assigning zero probability to unseen species by adding one to the number of occurrences of each species (including the unseen ones), but the number of species has to be known, and the decision to add 1 instead of 1/2 or 2 is rather random. With the Good-Turing estimator, both problems are solved: the number of species in the sample is not needed, and the probability of unseen species is connected with the number of species represented once in the sample.

Why Simple Good-Turing?

As can be seen in the jungle example, estimates given by the Good-Turing estimator is not *good* enough for large r's since gaps in r's (where n_r's are zero) can result in "loss" of probability and zero probability assigned to species that are actually represented the most times. That is why smoothing is needed. Various smoothing techniques have been discussed by Irving J. Good himself (see Ref. 2) and other people, among which a simple and relatively good one is the Simple Good-Turing (SGT) estimator proposed by William A. Gale and Geoffrey Sampson (see Ref. 1).







The left figure represents 5000 out of 10000 simulated local marginal prices (LMP's) at one time at one place. The middle figure shows that, as r gradually increments by 1, it is less and less likely to find any species represented exactly r times in the sample at all, and thus the n_r's would alternate between small numbers, most likely 0 and 1, until the species with the greatest number of occurrences in the sample is included.

Plotting n_r against r on a logarithmic scale would show that nonzero n_r's for large r's have a lower limit of 1 and thus need to be rounded down so that the nonzero n_r's can get averaged with the zero n_r's after the best fitting line is found. In the right figure, the green dots represent n_r's before smoothing, the blue circles represent rounded-down n_r's, and the red line represents the smoothed values for all positive integers r.

Results & Conclusion

l_1-Distances for Normal Distribution ($\mu = 0$)

case	σ (SD)	SGT	Empirical	Laplace
1	3	1.5177	2.2364	2.1165
2	4	1.4833	2.2178	2.0649
3	5	1.4629	2.1913	2.0122
4	6	1.4766	2.2105	2.0591

l_1-Distances for Uniform Distribution

case	N (max)	SGT	Empirical	Laplace
1	50	0.4469	0.5370	0.3580
2	100	0.3466	0.7314	0.3657
3	150	0.3031	1.0250	0.4100
4	200	0.2690	1.2111	0.4037

l_1-Distances for Poisson Distribution

case	λ (mean)	SGT	Empirical	Laplace
1	50	0.4414	0.4503	0.4581
2	150	0.6489	0.5944	0.8464
3	350	0.8324	0.7239	0.12024
4	500	0.9225	0.7981	1.3351

l_1-Distances for LMP

case	SGT	Empirical	Laplace
1	0.9817	0.9890	1.1079
2	0.9554	0.9429	1.1621
3	0.8686	0.9290	0.9714
4	1.2048	1.1878	1.4909
Avg.	1.0026	1.0122	1.1581

The SGT estimator performs better than the empirical and Laplace estimators on random samples from two out of three known distributions and on simulated LMP's overall. However, the results here are by no means conclusive, since more extensive and thorough tests are required to further investigate other types of distributions, and if the SGT estimator is only "good" under certain distributions, what those distributions are and why. Also, more can be explored about how to apply the SGT estimator to real LMP's and how the results can help market participants adjust their bidding strategies.

References

[1] William A. Gale and Geoffrey Sampson. Good-Turing Frequency Estimation Without Tears. *Journal of Quantitative Linguistics*, 2(3):217-237, 1995.

[2] Irving J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237-264, 1953.

Acknowledgments

The undergraduate researcher is grateful to Prof. Lang Tong for his guidance and help throughout the research project. Sevi Baltaoglu, Weisi Deng, Yuting Ji, Daniel Munoz-Alvarez, Eugene Zhe Yu, and other members of the Tong Group also helped in many meaningful ways.

This research was funded by the Cornell ECE Early Career Research Scholars Program and administered through Engineering Learning Initiatives at Cornell.

Source Code & Report

MATLAB source code and report are available at the following link:
`https://www.github.com/alicexinyun/2016SummerResearch`