Inside *zipfR*

Stefan Evert
`https://zipfR.R-forge.R-project.org/`

typeset on September 2, 2020

Contents

1 The mathematics of LNRE modelling

1.1 Notation

1.2 Sampling distribution

1.3 LNRE models

1.4 Parameter estimation

1.5 Posterior distribution & Good-Turing

2 Productivity & lexical diversity

2.1 Overview of productivity measures

$$\text{TTR} = \frac{V}{N} \tag{1}$$

$$\text{Guiraud's } R = \frac{V}{\sqrt{N}} = \sqrt{N} \cdot \text{TTR} \tag{2}$$

Carroll's CTTR $= V/\sqrt{2N} = R/\sqrt{2}$ is fully equivalent.

$$\text{Herdan's } C = \frac{\log V}{\log N} = \frac{\log \text{TTR}}{\log N} + 1 \tag{3}$$

Herdan assumes the general power law $V \sim N^{\alpha}$, with $C \to \alpha$ for $N \to \infty$. The assumption is met approximately by any infinite Zipf-Mandelbrot population and $C \to 1/a$.

$$\text{Dugast's } k = \frac{\log V}{\log \log N} = \frac{\log(N \cdot \text{TTR})}{\log \log N} \tag{4}$$

$$\text{Dugast's } U = \frac{(\log N)^2}{\log N - \log V} = \frac{\log N}{1 - C} \tag{5}$$

Maas's $a^2 = (\log N - \log V)/(\log N)^2 = 1/U$ is fully equivalent, but formulated as a measure of "lexical poverty", i.e. low values indicate high productivity.

$$\text{Brunet's } W = N^{V^{-a}} \text{ with } a = 0.172 \tag{6}$$

which looks less ridiculous in the form $\log W = V^{-a} \cdot \log N$

$$\text{Summer's } S = \frac{\log \log V}{\log \log N} \tag{7}$$

is not implemented in *zipfR* because the symbol clashes with Sichel's S and because it's taking the double logs to absurdity.

Baayen (1992) proposes the productivity index

$$\mathscr{P} = \frac{V_1}{N}$$

(originially from his PhD thesis, Baayen 1989), which corresponds to the slope of the vocabulary growth curve.

$$\text{Honoré's } H = 100\frac{\log N}{1 - V_1/V} \tag{8}$$

$$\text{Sichel's } S = \frac{V_2}{V} \tag{9}$$

Michéa's $M = V/V_2 = 1/S$ is a measure of lexical poverty and fully equivalent to Sichel's S.

For an infinite Zipf-Mandelbrot population, the slope parameter $\alpha = 1/a$ can directly be estimated from the proportions of hapax (Evert 2004: 130) and dis legomena (Evert 2004: 127), which are independent of sample size under certain simplifying assumptions and large-sample approximations:

$$\text{Stefan's } \alpha_1 = \frac{V_1}{V} \quad \text{and} \quad \alpha_2 = 1 - 2\frac{V_2}{V_1}$$

First experiments show that α_1 and α_2 work well for random samples from a ZM population, but can be very sensitive to deviations, esp. in the form of a finite population. Of particular interest is the measure α_1, which is simply the proportion of hapaxes (in analogy to S), and which has also been suggested as an estimator for the Zipf slope parameter by Rouault (1978: 172).

Also note that Sichel's S can be seen as a combination of the two measures:

$$S = \alpha_1 \frac{1 - \alpha_2}{2}$$

Entropy H needs to be distinguished from Honoré's H:

$$H = -\sum_{i=1}^{\infty} \frac{f_i}{N} \log_2 \frac{f_i}{N} = -\sum_{m=1}^{\infty} V_m \frac{m}{N} \log_2 \frac{m}{N} \tag{10}$$

Maximum value depends on V, viz. $H \leq \log_2 V$. Hence compute normalized entropy (aka evenness or efficiency)

$$\eta = \frac{H}{\log_2 V} \tag{11}$$

with $0 \leq \eta \leq 1$. (Or better η^{-1} as productivity measure?) η may be problematic due to counter-intuitive scaling with sample size, and a quick Web search suggests that unbiased estimation of H is really difficult, so we will not pursue H as a productivity measure (but may implement it do demonstrate problematic issues).

Yule suggested a measure based on statistical moments of the frequency spectrum (which sounds quite absurd but could possibly be motivated in terms of sampling random types), leading to

$$\text{Yule's } K = 10^4 \left(-\frac{1}{N} + \sum_{m=1}^{\infty} V_m \left(\frac{m}{N}\right)^2 \right) = 10^4 \cdot \frac{\sum_{m=1}^{\infty} m^2 \cdot V_m - N}{N^2} = 10^4 \cdot \sum_{i=1}^{\infty} \frac{f_i(f_i - 1)}{N^2} \tag{12}$$

Herdan proposed a very similar measure $v_m \approx \sqrt{K}$ based on a different mathematical derviation:

$$\text{Herdan's } v_m = \sqrt{-\frac{1}{V} + \sum_{i=1}^{\infty} V_i \left(\frac{i}{N}\right)^2} \tag{13}$$

Simpson proposed a closely related measure that can be interpreted as an unbiased estimator of a population coefficient $\delta = \sum_i \pi_i^2$, i.e. the probability of drawing the same type twice from the population.

$$\text{Simpson's } D = \sum_{m=1}^{\infty} V_m \frac{m}{N} \frac{m-1}{N-1} = \sum_{i=1}^{\infty} \frac{f_i}{N} \frac{f_i - 1}{N-1} \tag{14}$$

For a LNRE population, the coefficient δ corresponds to the second moment of type density function:

$$\delta = \int_0^\infty \pi^2 g(\pi)\, d\pi \tag{15}$$

Baayen *et al.* (1996: 124) already note that "[t]he values of both D and K are primarily determined by the high end of the frequency distribution structure".

2.2 Expected values and sampling distribution

Linear transformations of V or V_m Expectations and full sampling distributions can directly be obtained for TTR and other measures that are linear transformations of V or a single spectrum element V_m. In particular:

$$E[\mathrm{TTR}] = \frac{E[V]}{N} \tag{16}$$

$$E[R] = \frac{E[V]}{\sqrt{N}} \tag{17}$$

$$E[\mathscr{P}] = \frac{E[V_1]}{N} \tag{18}$$

Nonlinear transformations of V For a nonlinear transformation $f(V)$, approximate expectations can be obtained if the distribution of V covers a region in which f is approximately linear. In this case,

$$E[f(V)] \approx f(E[V]) \tag{19}$$

As long as f is monotonic, the corresponding equality for the median is always exact. Provided that V has a symmetric distribution (e.g. by being approximately normal), we obtain

$$\mathrm{med}[f(V)] = f(\mathrm{med}[V]) \approx f(E[V]) \tag{20}$$

We assume that these conditions are met for all productivity measures of this form, but may want to check later whether linearity of the transformation is always a plausible assumption.

$$E[C] = \frac{\log E[V]}{\log N} \tag{21}$$

$$E[k] = \frac{\log E[V]}{\log \log N} \tag{22}$$

$$E[U] = \frac{(\log N)^2}{\log N - \log E[V]} \tag{23}$$

$$E[\log W] = E[V]^{-a} \cdot \log N \text{ with } a = 0.172 \tag{24}$$

W seems to have (only) slightly larger curvature than $\log W$, so both forms are equally viable.

Measures involving ratios of V and V_m For such measures, approximate expectations can be obtained from normal approximations to V and V_m (which should be very good in the range of samples where the measures are reasonable) and the fact that the ratio of two independet normal variables $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$ can itself be approximated by a normal distribution

$$\frac{X}{Y} \sim N(\mu, \sigma^2) \quad \text{with} \quad \mu = \frac{\mu_1}{\mu_2}, \;\; \sigma^2 = \mu^2 \left(\frac{\sigma_1^2}{\mu_1^2} + \frac{\sigma_2^2}{\mu_2^2} \right) \tag{25}$$

provided that $\mu_i / \sigma_i \gg 1$ (Díaz-Francés and Rubio 2013: 313). This should also hold for correlated variables X and Y with the same μ but different σ^2, as demonstrated for the special case $X/(X+Y)$ by Evert (2004: Lemma A.8). For nonlinear transformations of such ratios, we assume

$$E\left[f\left(\frac{X}{Y}\right)\right] \approx f\left(E\left[\frac{X}{Y}\right]\right) \approx f\left(\frac{E[X]}{E[Y]}\right) \tag{26}$$

This yields expectations for:

$$E[S] = \frac{E[V_2]}{E[V]} \tag{27}$$

$$E[\alpha_1] = \frac{E[V_1]}{E[V]} \tag{28}$$

$$E[\alpha_2] = 1 - 2\frac{E[V_2]}{E[V_1]} \tag{29}$$

$$E[H] = 100\frac{\log N}{1 - E[V_1]/E[V]} \tag{30}$$

Variance and sampling distribution Except in the case of linear transformations of V or V_m, the full sampling distributions are much harder to compute. Approximate variances could be determined from normal approximations to V and V_m together with (25) and a linear approximation

$$f(X) \approx f(E[X]) + f'(E[X]) \cdot (X - E[X])$$

Variances would have to be worked out in detail and require a version of (25) for correlated variables, possibly based on the proof by Evert (2004: Lemma A.8).

For the time being, variances and confidence intervals will be determined empirically by parametric boostrapping.

Measures based on the full frequency spectrum Expectations for Simpson D and Yule K are derived from the individual binomial distributions of $f_i \sim B(N, \pi_i)$; additivity of expected values does not required independence of the random variables. Since $E[f_i] = N\pi_i$ and $\mathrm{Var}[f_i] = N\pi_i(1-\pi_i)$, we find that

$$E[f_i^2] = \mathrm{Var}[f_i^2] + E[f_i]^2 = N\pi_i(1-\pi_i) + (N\pi_i)^2 \tag{31}$$

Application to Simpson's D yields

$$\begin{aligned} E[D] &= \frac{1}{N(N-1)} \sum_{i=1}^{\infty} \left(E[f_i^2] - E[f_i]\right) \\ &= \frac{1}{N(N-1)} \sum_{i=1}^{\infty} \left(N\pi_i(1-\pi_i) + (N\pi_i)^2 - N\pi_i\right) \\ &= \frac{1}{N(N-1)} \sum_{i=1}^{\infty} \pi_i^2 (N^2 - N) = \sum_{i=1}^{\infty} \pi_i^2 = \delta \end{aligned}$$

proving the claim that D is an unbiased estimator of the population coefficient δ (Simpson 1949: 688).

If we approximate the binomial distributions of f_i with Poisson distributions (as in the simplified Poisson sampling approach to LNRE models), we have $E[f_i] = N\pi_i$ and $E[f_i^2] = N\pi_i + (N\pi_i)^2$.

Under these assumptions, Yule's K becomes an unbiased estimator:

$$\begin{aligned} E\bigl[K\bigr] &= \frac{10^4}{N^2} \cdot \sum_{i=1}^{\infty} \bigl(E\bigl[f_i^2\bigr] - E\bigl[f_i\bigr]\bigr) \\ &= \frac{10^4}{N^2} \cdot \sum_{i=1}^{\infty} N^2 \pi_i^2 = 10^4 \cdot \sum_{i=1}^{\infty} \pi_i^2 = 10^4 \cdot \delta \end{aligned}$$

With binomial sampling, the expectation is $E\bigl[K\bigr] = 10^4 \frac{N-1}{N} \delta$.

An equation for the variance of D is provided by Simpson (1949: 688) and may be accepted without further proof, given that his claim about $E\bigl[D\bigr]$ was correct.

3 Specific LNRE models

3.1 Zipf-Mandelbrot (ZM)

The second moment δ of the tdf (15) is easily obtained from (33) by setting $A = 0$ (which also applies to the normalizing constant C):

$$\delta = \frac{C}{2-\alpha} B^{2-\alpha} = \frac{(1-\alpha) B^{2-\alpha}}{(2-\alpha) B^{1-\alpha}} \tag{32}$$

3.2 Finite Zipf-Mandelbrot (fZM)

The second moment δ of the tdf (15) is given by

$$\begin{aligned} \delta &= \int_0^{\infty} \pi^2 g(\pi)\, d\pi = C \int_A^B \pi^{1-\alpha}\, d\pi \\ &= C \left[\frac{\pi^{2-\alpha}}{2-\alpha} \right]_A^B = \frac{C}{2-\alpha} \left(B^{2-\alpha} - A^{2-\alpha} \right) \\ &= \frac{1-\alpha}{2-\alpha} \cdot \frac{B^{2-\alpha} - A^{2-\alpha}}{B^{1-\alpha} - A^{1-\alpha}} \end{aligned} \tag{33}$$

3.3 Generalised Zipf-Mandelbrot (gZM)

3.4 Generalised Inverse Gauss-Poisson (GIGP)

3.5 Montemurro / Tsallis

Based on original research by Tsallis, Montemurro (2001) proposes the following type density function:

Insert references to Tsallis

$$g(\pi) = C \left(\mu \pi^R + (\lambda - \mu) \pi^Q \right)^{-1} \tag{34}$$

with parameters $1 < R < Q$ and $\mu, \lambda \in \mathbb{R}$, normalising constant C and possible restriction to a suitable region $A \le \pi \le B$. Eq. (34) is derived from a differential equation for the rank-frequency relationship, which does not have a closed-form rank-frequency solution in the general case. See Sec. 4.1.1 for motivation and details.

It will be difficult to obtain closed-form solutions for $E\bigl[V\bigr], E\bigl[V_m\bigr]$ and other relevant quantities from Eq. (34), and numerical integration may often be required.

Montemurro's LNRE model can be thought of as a smooth interpolation between two power laws with different slopes that hold in different frequency ranges. Therefore, a two-segment gZM model (Sec. 3.3) or a mixture of two ZM/gZM models (Sec. 3.6) should give an approximation to Eq. (34), but will be much easier to handle mathematically (closed-form solutions, numerical accuracy).

3.6 Mixture models

Write up mathematics and purpose of mixture models

4 Literature notes

4.1 Extensions of Zipf's law

4.1.1 Montemurro (2001)

Montemurro (2001) proposes an extension to Zipf's law that results in a better empirical fit for higher ranks (i.e., low-frequency data), while Zipf-Mandelbrot only improves the fit for low ranks. Based on all-words Zipf rankings for various literary corpora compiled from Gutenberg e-texts, he observes that the original form of Zipf's law (with $a \approx 1$) only seems to hold for a "middle range" of frequency ranks, $r \approx 100 \ldots 2000/6000$ (depending on corpus size). He further claims that higher ranks follow a similar power law with steeper slope ($a \approx 2 \ldots 3$).

In the following, notation has been adjusted to *zipfR* conventions:

- r = Zipf rank (original: s)
- p_r = relative frequency of r-th type in Zipf ranking (original: $f(s)$)
- Q, R = Zipf slopes (original: q, r)

Noting that the Zipf-Mandelbrot law can be derived from a differential equation

$$\frac{dp}{dr} = -\lambda p^Q \tag{35}$$

(Eq. (3), p. 572), he derives a generalisation from the differential equation

$$\frac{dp}{dr} = -\lambda p^R - (\lambda - \mu) p^Q \tag{36}$$

(Eq. (4), p. 572). There are closed-form solutions for the special cases $R = Q = 1$ (Zipf-Mandelbrot) and $R = 1, Q > 1$ (Zipf's law in middle range, steeper slope for higher ranks), but not for the general case $1 < R < Q$ (p. 573).

Empirically, a good fit is obtained for literary corpora from single authors, using the closed form solution with $R = 1, Q > 1$:

$$p_r = \left(1 - \frac{\lambda}{\mu} + \frac{\lambda}{\mu} e^{(Q-1)\mu r}\right)^{-\frac{1}{Q-1}} \tag{37}$$

(Eq. (6), p. 573). For larger, composite corpora, the general form $1 < R < Q$ seems to be required.

The differential equation (36) and its various closed-form or implicit solutions are attributed to Constantino Tsallis:

Find papers by Tsallis & Denisov

- Tsallis/Bemski/Mendes (1999), *Phys Lett A* **257**, 93
- Tsallis (1988), *J Stat Phys* **52**, 479 — underlying framework of statistical mechanics
- Montemurro says that Tsallis has suggested application to linguistic data in "private communication"
- Denisov (1997), *Phys Lett A* **235**, 447 — relates Zipf-Mandelbrot law to "fractal structure of symbolic sequences with long-range correlations" (p. 572)

Montemurro notes that with some approximations, the general case can be expressed in closed form as a type density function (which he awkwardly refers to as a “probability density”), resulting in the LNRE model:

$$g(\pi) \propto \left(\mu\pi^R + (\lambda - \mu)\pi^Q\right)^{-1} \tag{38}$$

See Sec. 3.5 for more information on this LNRE model.

Montemurro postulates that the two power laws may correspond to “general” and “specialised” vocabulary without further evidence: “This suggests … the vocabularies can be divided into two parts of distinct nature: one of basic usage …, and a second part containing more specific words with a less flexible syntactic function.” (p. 571, citing then unpublished work by Ferrer & Solé). If we accept this claim, a linear mixture model (Sec. 3.6) would be much more appropriate than a hard split at rank $r \approx 2000 \ldots 6000$.

There is a completely unfounded claim at the end of the paper that “it seems quite plausible that there may be a deep connection between differential equation (4) and the actual processes underlying the generation of syntactic language.” (p. 577).

5 Notes and ideas

5.1 Mathematics and implementation

- LNRE model fit may be affected by a small number of very high-frequency types, esp. “echo” tokens (Baroni and Evert 2007) or the “other” type when modelling vocabulary growth wrt. all tokens (*external productivity*). It would probably be useful to separate the most frequent types, estimate their occurrence probabilities directly (MLE are reliable barring non-randomness effects), and apply the LNRE model only to the remaining vocabulary. This should present no major mathematical obstacles, but will have to be taken into account throughout the implementation (expectations, variances, chi-squared statistics, etc.).

- Standard LNRE fitting uses only the low end of the frequency spectrum and may produce an unsatisfactory fit for the “middle range” of the Zipf ranking. If we want to account for these data as well – esp. in connection with mixture models (Sec. 3.6) and gZM (Sec. 3.3) – a different goodness-of-fit goal function will be needed for parameter estimation.

  One possibility is to pool multiple frequency classes together, e.g. on a logarithmic scale: $m = 1, \ldots, 10$, $m = 11 \ldots 14$, $m = 15 \ldots 20$, $m = 21 \ldots 50$, $m = 51 \ldots 100$, $m = 101 \ldots 1000$, etc.; granularity will have to be adjusted to the available data, of course. Assuming the usual multivariate Gaussian joint distriubtion for the original frequency spectrum, the pooled frequency spectrum should also be multivariate Gaussian as a linear map of the original spectrum. Expectations, variances and covariances should be straightforward, although care has to be taken to avoid performance and/or numerical accuracy issues.

  Are there simplified equations for expectations and (co)variances of a pooled frequency spectrum?

- Sometimes it would be useful to fit a LNRE model to multiple frequency spectra. E.g. for Gordon Pipa’s neural spiking data, where it is plausible that trials for the same condition follow the same Zipfian distribution, but data cannot be pooled directly; or to avoid overfitting of non-random data by parallel parameter estimation from frequency spectra at different sample sizes. Such *co-estimation* should be relatively straightforward to implement by adding up cost functions (perhaps with suitable scaling to account for different sample sizes), but custom estimators available for some of the models can no longer be used.

  Is “co-estimation” an appropriate term?

- One problem of the fZM implementation may be numerical accuracy due to cancellation when “short” Gamma integrals are calculated as differences between incomplete Gamma functions, esp. on very small or otherwise extreme samples. This will become much more virulent for gZM models with many components. Suggest a two-step strategy:

1. Encapsulate finite Gamma integrals into a helper function, which estimates cancellation errors and collects statistics. This should be controlled by a global `debug` option for *zipfR* (set with `zipfR.par()`).
2. Implement more accurate algorithm for finite Gamma integrals. So far, the only solution seems to be numeric integration, which is easy and accurate for monotonic functions (possibly splitting integrand into monotonic parts). Code might be implemented in R (using standard numeric integration functions) for preliminary testing.
3. Reimplement numeric integration in C; for better efficiency and accuracy on "long" Gamma integrals, might compute incomplete Gamma function first and run numeric integration only when estimated cancellation error exceeds a pre-defined threshold (also set with `zipfR.par()`).

5.2 Thoughts on goals and applications of LNRE modelling

- Most research on Zipf's law (both Zipf himself and more recent work by physicists) focuses on middle-range frequency ranks, which are highlighted in a logarithmic rank-frequency graph. By contrast, LNRE models (Khmaladze 1987; Baayen 2001) based on truncated frequency spectra are only interested in the lowest-frequency types. Note that for typical applications – productivity, vocabulary growth, estimation of vocabulary diversity, adjusted significance tests – only such lowest-frequency types are of major concern, as probability estimates for middle- and high-frequency data can be obtained directly from any sizable corpus. This is explains why most LNRE models find Zipf slows $a \gg 1$ rather than $a \approx 1$ as observed by Zipf and related work.

References

Baayen, Harald (1992). Quantitative aspects of morphological productivity. *Yearbook of Morphology 1991*, pages 109–149.

Baayen, Harald; van Halteren, Hans; Tweedie, Fiona (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, **11**(3), 121–132.

Baayen, R. Harald (2001). *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht.

Baroni, Marco and Evert, Stefan (2007). Words and echoes: Assessing and mitigating the non-randomness problem in word frequency distribution modeling. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 904–911, Prague, Czech Republic.

Díaz-Francés, Eloísa and Rubio, Francisco J. (2013). On the existence of a normal approximation to the distribution of the ratio of two independent normal random variables. *Statistical Papers*, **54**(2), 309–323.

Evert, Stefan (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Published in 2005, URN urn:nbn:de:bsz:93-opus-23714. Available from `http://www.collocations.de/phd.html`.

Khmaladze, E. V. (1987). The statistical analysis of large number of rare events. Technical Report MS-R8804, Department of Mathematical Statistics, CWI, Amsterdam, Netherlands.

Montemurro, Marcelo A. (2001). Beyond the Zipf-Mandelbrot law in quantitative linguistics. *Physica A*, **300**, 567–578.

Rouault, Alain (1978). Lois de Zipf et sources markoviennes. *Annales de l'Institut H. Poincaré (B)*, **14**, 169–188.

Simpson, E. H. (1949). Measurement of diversity. *Nature*, **163**, 688.

Todo list