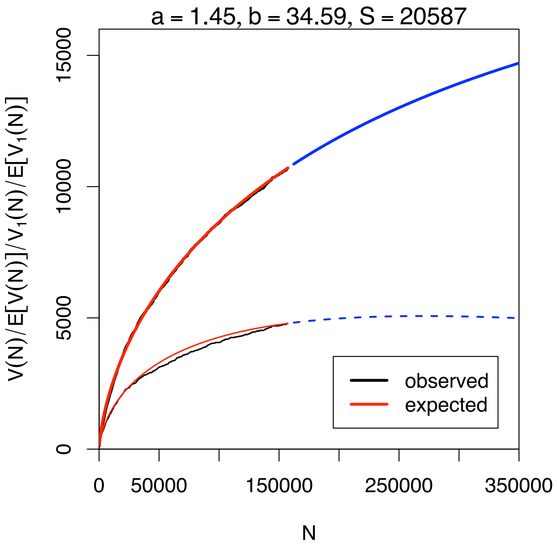
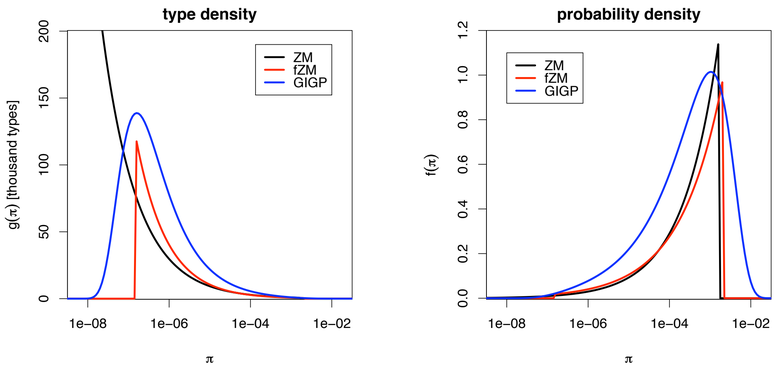
## Word frequency distributions & LNRE models

- type-token statistics for any type-rich population with Zipf-like probability distribution (LNRE = Large Number of Rare Events, Baayen 2001)
- extrapolation of vocabulary growth & frequency spectrum to larger samples (➠ morphological productivity, vocabulary richness, stylometry, data sparseness, etc.)
- estimation of vocabulary size from small samples (e.g. sentence patterns or word senses)
- prior distribution in Bayesian inference & population model for Good-Turing smoothing
- reliability of statistical inference for low-frequency data from Zipfian population (Evert 2004)




- random sampling from LNRE models
- distribution and density functions
- expectations and variance/covariance for V(N) and frequency spectrum $V_k(N)$




$$g(\pi) := C \cdot \pi^{\gamma-1} \cdot e^{-\frac{\pi}{c} - \frac{b^2 c}{4\pi}} \qquad g(\pi) := \begin{cases} C \cdot \pi^{-\alpha-1} & A \le \pi \le B \\ 0 & \text{otherwise} \end{cases}$$

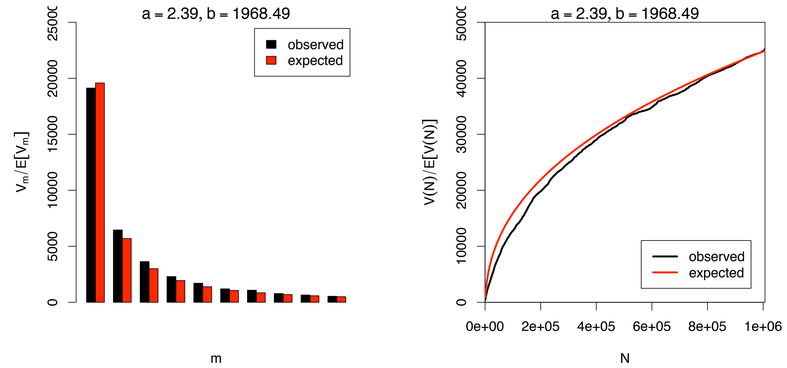
**GIGP** **ZM** (A=0) & **fZM**

*user-friendly LNRE modelling*

zipfR

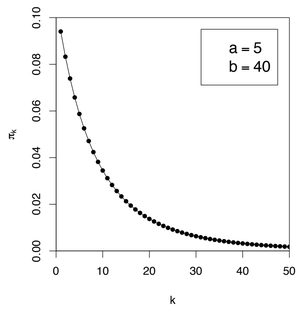
available on CRAN

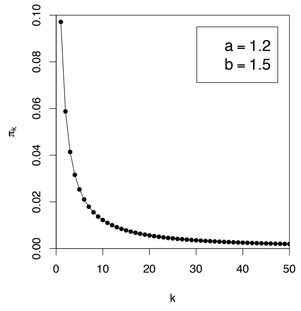
*http://purl.org/stefan.evert/zipfR*




## Parameter estimation

- robust & flexible parameter estimation
- 4 minimisation techniques
- 5 customisable cost functions
- user-specified model parameters
- goodness-of-fit (multivariate chi-squared)





**References**

- 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*, Prague.
- Evert, Stefan (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations.* Dissertation, IMS, University of Stuttgart. Published in 2005, URN urn:nbn:de:bsz:93-opus-23714.
- Evert, Stefan (2004a). A simple LNRE model for random character sequences. In *Proceedings of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles*, 411–422. Louvain-la-Neuve, Belgium.
- Evert, Stefan and Baroni, Marco (2007). zipfR: Word frequency distributions in R. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Posters & Demos Session.
- Sichel, H. S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association*, **70**, 542–547.



**Marco Baroni** (CIMeC, U Trento)
*marco.baroni@unitn.it*

**Stefan Evert** (IKW, U Osnabrück)
*stefan.evert@uos.de*