The Marshall-Olkin Extended Zipfian Distribution

Marta Pérez-Casany [1] and Aina Casellas Torrentó[2]

[1]Dept. of Applied Mathematics II and DAMA-UPC
Technical University of Catalonia
[2]School of Mathematics and Statistics
Technical University of Catalonia

February 19th 2013

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Contents

1. Zipfian distribution
2. MOEDZipf Distribution
3. Data Analysis

2. Zipfian Distribution

$X \sim Zipf(\alpha)$ with $\alpha > 1$ iff

$$P(X = x) = \frac{x^{-\alpha}}{\zeta(\alpha)} \quad x = 1, 2, 3, 4, \ldots$$

where $\zeta(\alpha)$ is the Riemann zeta function, $\zeta(\alpha) = \sum_{k=1}^{\infty} k^{-\alpha}$.

Main characteristics:

- Large probability at one in most of the parameter space;
- right skewed;
- linear in the log-log scale;
- scale-free distribution.

Suitable to fit **frequencies of frequencies** and **ranking data**.

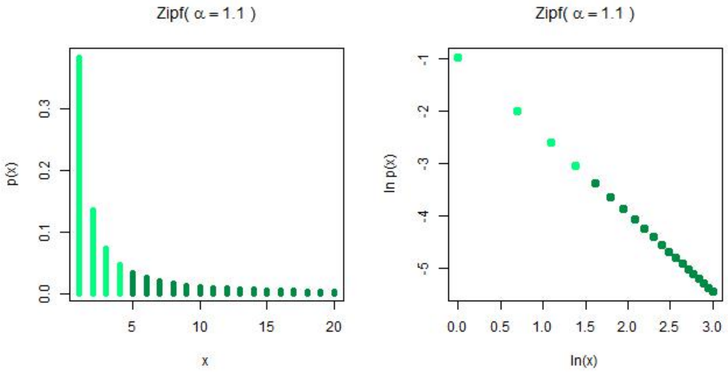


Figura: Zipf. distribution for $\alpha = 1.5$. On the right in the log-log scale.

Nevertheless, quite often appear situations like these ones:

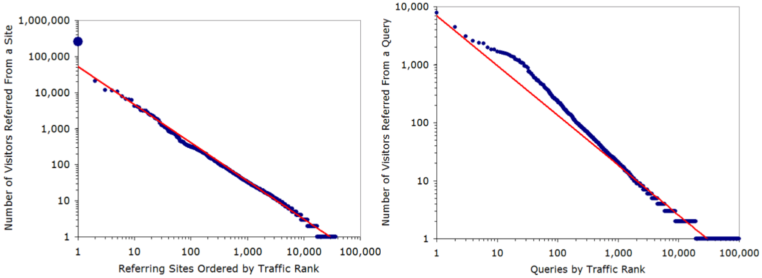


Figura: Ranked data, Traffic referrals and traffic incoming queries

2. MOEZipf Distribution

If $X \sim Zipf(\alpha)$, its SF is $\overline{F}(x) = \frac{\xi(\alpha,x)}{\xi(\alpha)}$, where $\xi(\alpha,x) = \sum_{k=x}^{+\infty} k^{-\alpha}$.

It is said that $Y \sim MOEZipf(\alpha,\beta)$ iff its SF is equal to:

$$\overline{G}(x;\alpha,\beta) = \frac{\beta\,\overline{F}(x)}{1-\overline{\beta}\,\overline{F}(x)} = \frac{\beta\,\zeta(\alpha,x+1)}{\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x+1)}\,.$$

In that case one has that:

$$\begin{aligned} P(Y=x) &= \overline{G}(x-1)-\overline{G}(x) \\ &= \frac{\beta\,\zeta(\alpha)\,x^{-\alpha}}{\left[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x)\right]\left[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x+1)\right]}\,, \end{aligned}$$

where $x \in \mathbb{N}$, $\alpha > 1$ and $0 < \beta < +\infty$.

Marshall, A.W. and Olkin, I. (1997) *A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families.*

Example 1

Number of connections of a total of 225409 electronic addresses. Collected between October 2003 and May 2005.

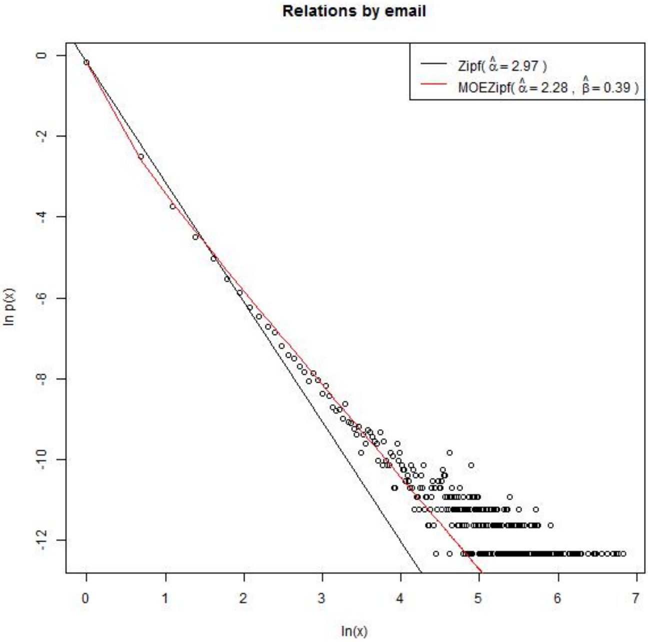
- 85% of the observations are equal to one;
- the three addresses with more contacts have: 854, 871 and 930.

Grouping the values ≥ 65 one has:

Distrib.	Par.	Est.	log-like.	X^2	p-val.	AIC
Zipf	$\hat{\alpha}$	2.968	-156765.21	13714.84	0	313532.42
MOEZipf	$\hat{\alpha}$	2.284	**-154399.82**	**858.27**	0	**308803.64**
(m.l.e.)	$\hat{\beta}$	0.390				

93.74% reduction in the χ^2 goodness of fit statistic.

J. Leskovec, (2008) Dynamics of Large Metworks. *PhD thesis, School of Computer Science, Carnegie Mellon* University. http://snap.stanford.edu/data/email-EuAll.html.


Relations by email
Zipf($\hat{\alpha}$ = 2.97)
MOEZipf($\hat{\alpha}$ = 2.28 , $\hat{\beta}$ = 0.39)
ln p(x)
ln(x)
0
-2
-4
-6
-8
-10
-12
0
1
2
3
4
5
6
7

Example 2

Number of citations of a total of 32158 papers in *High-energy physics*. Published in arxXiv.org between January 1993 and April 2003.

- 26% of the observations correspond to values $\{1, 2\}$;

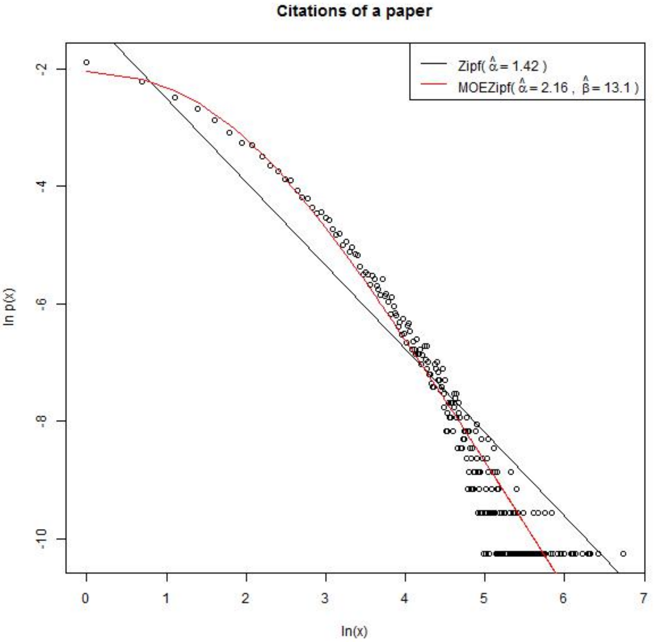
Grouping the values ≥ 119 one has:

Distrib.	Par.	Est.	log-like	X^2	p-val.	AIC
Zipf	$\hat{\alpha}$	1.421	-105839.81	13172.05	0	211681.61
MOEZipf	$\hat{\alpha}$	2.161	**-99197.93**	**816.62**	0	**198399.87**
(m.l.e)	$\hat{\beta}$	13.058				

Taula: Results of fitting the r.v.: *Number of citations of a given paper*.

87.73% reduction in the χ^2 goodness of fit statistic.

J. Leskovec, (2008) Dynamics of Large Metworks. *PhD thesis, School of Computer Science, Carnegie Mellon* University.
http://snap.stanford.edu/data/citHepPh-EuAll.html.


Citations of a paper
Zipf($\hat{\alpha}$ = 1.42)
MOEZipf($\hat{\alpha}$ = 2.16 , $\hat{\beta}$ = 13.1)
ln p(x)
ln(x)
-2
-4
-6
-8
-10
0
1
2
3
4
5
6
7

THANK YOU VERY MUCH FOR YOUR ATTENTION