Applied Mathematics E-Notes, 14(2014), 151-160 © ISSN 1607-2510
Available free at mirror sites of http://www.math.nthu.edu.tw/∼amen/

Some Statistical Inferences For Two Frequency Distributions Arising In Bioinformatics*

Davood Farbod†

Received 28 February 2014

Abstract

Discrete Distribution Generated by Levy's Density (DLD) and some Pareto-like frequency Distribution (PD) are considered. First, as examples, we will examine the DLD and the PD with two real data sets in bioinformatics. Second, regression models for the parameters of the DLD and PD are built based on two methods. Consistency, asymptotic normality and optimality of the Least Square (LS) estimators are verified, respectively. Some Corollaries, Remarks and numerical examples are also given.

1 Introduction

Several frequency distributions have been proposed for description phenomena arising in large-scale biomolecular systems (see [1]). In this paper, we consider two DLD and PD models. One of the most important problems for the DLD and PD is to investigate the statistical analysis of parameters estimators. This paper is organized as follows. Subsections 1.1 and 1.2 briefly give information about Levy Distribution and then introduce distribution generated by Levy's Law and also the PD model. Sections 2 and 3 contain main results of the paper. Conclusion is given in Section 4.

1.1 The DLD

The Levy Distribution is one of the few distributions that is Stable and has probability density function which is analytically expressible. On the other hand, the Levy Distribution is a sub-family of Stable Laws (*Stable Laws form a four-parametric class of probability distributions allowing skewness, heavy tails and have other useful mathematical properties. The class was determined by Paul Levy in the 1920's. For more details see [13]*) with the following density ([13])

$$s(x;\gamma,\delta)=\sqrt{\frac{\gamma}{2\pi}}(x-\delta)^{-\frac{3}{2}}\exp(-\frac{\gamma}{2(x-\delta)}),\quad \gamma\in\mathbb{R}^+,\ \delta\in\mathbb{R},\ x>\delta, \tag{1}$$

where γ is the scale parameter and δ is the location parameter.

[*] Mathematics Subject Classifications: 62F12, 62F10, 62P10, 62J05.

[†] Department of Mathematics, Quchan University of Advanced Technology, Quchan, Iran

Let us have the following probability function constructed from (1) when $\delta = 0$ (see, for example, [3]):

$$p(x;\gamma) = \frac{s(x;\gamma,0)}{d_\gamma}, \qquad x = 1, 2, ..., \tag{2}$$

where $d_\gamma = \sum_{y=1}^{\infty} s(y;\gamma,0)$ is the *normalization* factor. Some statistical analysis of the parameters estimators of the model (2) have been considered in [3], [4] and [5].

NOTE: If random variable ξ has the DLD with the probability function (2), then $E(\xi)^j = \infty$ when $j \geq \frac{1}{2}$, and $E(\xi)^j < \infty$ when $j < \frac{1}{2}$.

1.2 The PD

A two-parametric frequency distribution so-called PD was introduced by V. A. Kuznetsov in 2001 (see [11]) for biomolecular needs. Its probability mass function is:

$$f(x;\rho,b) = \frac{(x+b)^{-\rho}}{\sum_{y=1}^{\infty}(y+b)^{-\rho}}, \quad x = 1, 2, ..., \tag{3}$$

where $1 < \rho < \infty$ is the shape parameter and $-1 < b < \infty$ is the location parameter and shows the deviation of the PD from a simple power law. It appears as a distribution associated with stochastic processes of gene expression in eukaryotic cells (see [11]).

2 Fitting of the DLD

Let us note that the model (2) has been constructed using *discretization*. But, the model (2) has not been fitted with real data sets by now. In this Section, we shall attempt to propose two real data sets in order to fit the model (2) and also compare to the PD (3). Comparing to Farbod and Gasparian (see [6]), for applying the probability function (2) to the data, *truncated* DLD is considered. Namely, random variable is restricted to maximum observed in each data set. Some plots of the distribution (2) for some different values of the scale parameter are presented as well.

EXAMPLE 1. We consider the number of amino acids in the protein chain (see [7]) as a real data set in the following Table:

Table 1

36	153	146	97	83	46	150	43
29	30	71	58	26	40	70	138

Based on Kolmogorov-Smirnov (K-S) test the *p-value* is 0.5896, which does not reject the adequacy of the DLD for the number of amino acids. In order to an informal goodness of fit test, we plot the empirical cumulative distribution function (ecdf) and fitted cumulative distribution function (cdf) for the number of amino acids in Figure 1. Moreover, the ML estimation is $\hat{\gamma}_{ML} = 122.05$.

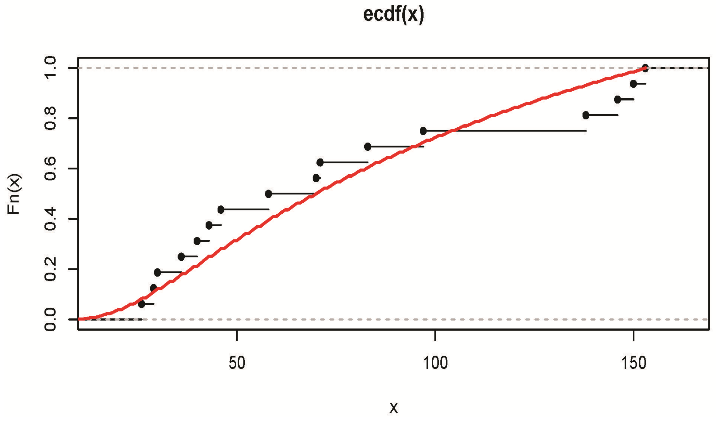

Figure 1: Fitting of the truncated DLD to the data of Table 1. The dashed line is the ecdf of data and the solid line is the fitted cdf.

EXAMPLE 2. Let us collect the number of residues for 12 electron transports in globular proteins (see [6,10]) as a real data set in the following Table:

Table 2

85	103	103	112	134	82
54	98	138	54	125	99

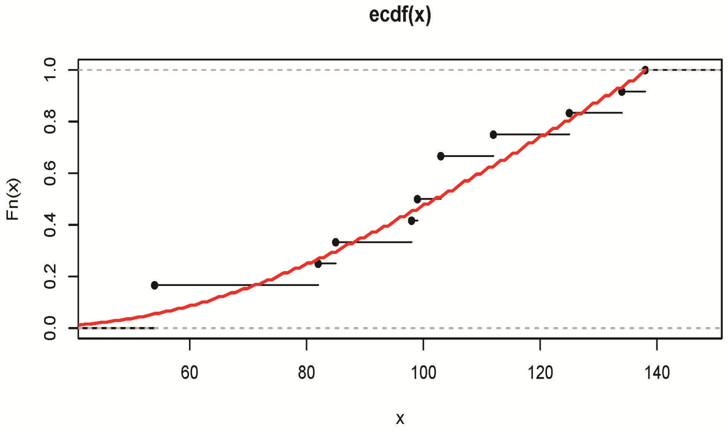

Figure 2: Fitting of the truncated DLD to the data of Table 2. The dashed line is the ecdf of data and the solid line is the fitted cdf.

Again by K-S test, the *p-value* is 0.9544, which does not reject the adequacy of the DLD for the number of residues. To do an informal goodness of fit test, let us plot the ecdf and fitted cdf of the number of residues in Figure 2. Meanwhile, the ML estimation is $\hat{\gamma}_{ML} = 450.315$.

2.1 Figure of the Model

We present some Plots of the truncated DLD for different values of the scale parameter γ in Figure 3.

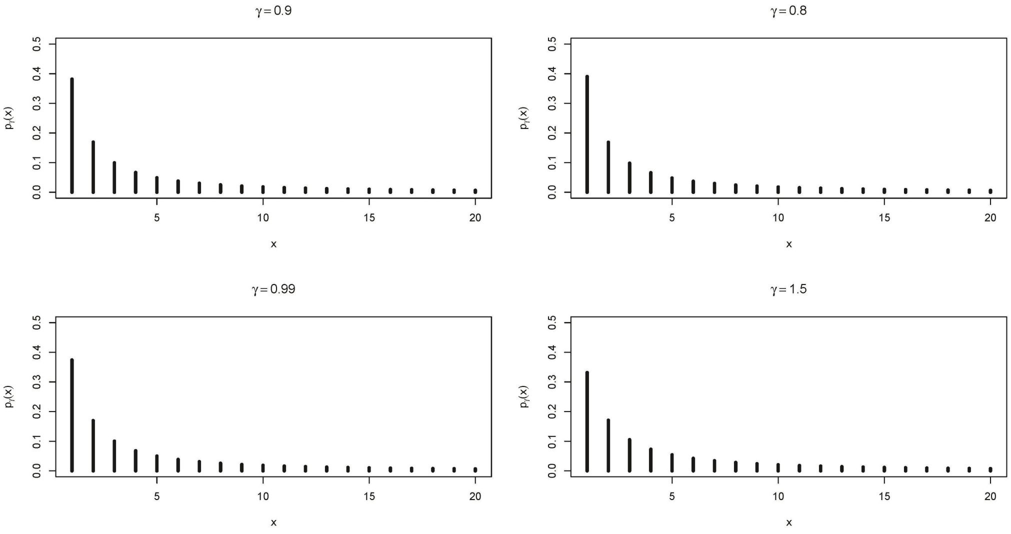

Figure 3: Some Plots of the truncated DLD (2) for different values of the parameter γ.

2.2 Compare to the PD

We shall fit the data, in Examples 1-2, with the PD and also compare to the DLD from biomolecular applications. For doing this, we consider the PD when $b = 0$.

EXAMPLE 3. Consider the data in Table 1. Then, using K-S test the *p-value* is 0.4775. Fitting of the truncated PD to the data of Table 1 is proposed in the Figure 4. Also, the ML estimation equals -0.12 which is *not* an acceptable estimation.

EXAMPLE 4. Let us have the data in Table 2. Then, using K-S test the *p-value* is 0.6938. Fitting of the truncated PD to the data of Table 1 is given in the Figure 5. The ML estimation equals -1.65 which is *not* acceptable.

COROLLARY 1. It is easily seen that the DLD fits data well with respect to the PD. It seems that the DLD fits large data better than the PD. We notice that the tails of the DLD are much heavy (more than the PD).

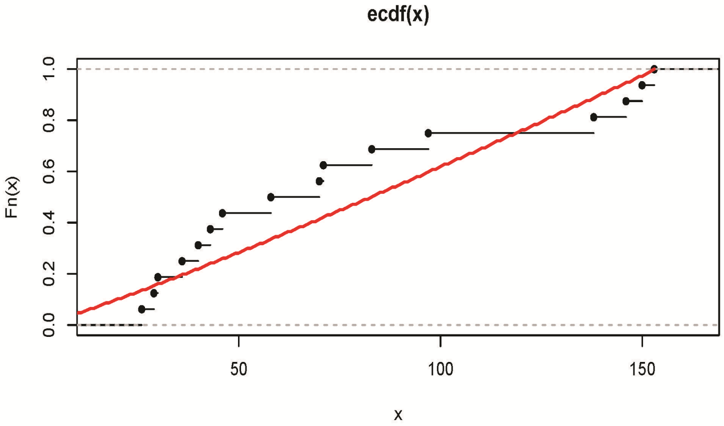

Figure 4: Fitting of the truncated PD to the data of Table 1. The dashed line is the ecdf of data and the solid line is the fitted cdf.

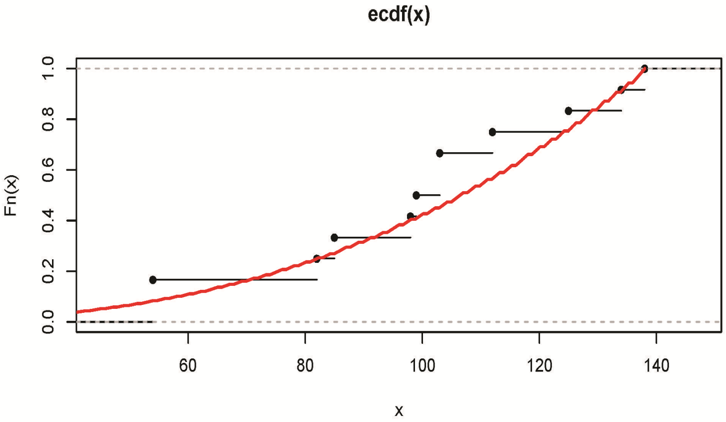

Figure 5: Fitting of the truncated PD to the data of Table 2. The dashed line is the ecdf of data and the solid line is the fitted cdf.

3 Regression Model

In this Section, we are going to consider regression model for the models (2) and (3), and then investigate some properties for them.

3.1 DLD

Without loss of generality, we consider the following distribution received from (2):

$$p(x;\gamma) = \frac{x^{-\frac{3}{2}}\exp(-\frac{\gamma}{2x})}{c_\gamma}, \qquad x = 1, 2, ..., \tag{4}$$

where $c_\gamma = \sum_{y=1}^{\infty} y^{-\frac{3}{2}} \exp\left(-\frac{\gamma}{2y}\right)$. It is well-known that the left-side of the (4) may be written as follows (if $x = x_i$):

$$p(x_i;\gamma) = F(x_i, \gamma) - F(x_{i-1}, \gamma), \quad i = 1, 2, ..., n, \tag{5}$$

where $F(.,.)$ is the theoretical cdf.

Taking logarithm from both sides of (4) and with regard to (5), we get

$$\ln\Big(F(x_i, \gamma) - F(x_{i-1}, \gamma)\Big) = -\frac{3}{2}\ln x_i - \Big(\frac{1}{2x_i}\Big)\gamma - \ln c_\gamma. \tag{6}$$

The left-side of (6), that is $\ln(F(x_i, \gamma) - F(x_{i-1}, \gamma))$, depends on unknown parameter γ and hence the relation (6) *may not* be used for a regression model. To overcome this problem (*compare with the used method based on sample characteristic function by Koutrouvelis [9]*), assuming $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x)$ is the ecdf, $I(.)$ is indicator function, then for large n we have,

$$Var(F_n(x_i) - F_n(x_{i-1})) = \frac{1}{n}[F(x_i) - F(x_{i-1})] \cdot [1 - F(x_i) + F(x_{i-1})], \quad i = 1, ..., n, \tag{7}$$

which turns out the mean square consistency of $(F_n(x_i) - F_n(x_{i-1}))$ for $(F(x_i, \gamma) - F(x_{i-1}, \gamma))$. Note that (compare to [9])

$$F_n(x_i) = \frac{1}{n}(\nu_1 + \nu_2 + ... + \nu_i),$$

where $\nu_i = \sum_{j=1}^{n} I(x_{i-1} < X_j \le x_i), \ i = 1, ..., n$. By (6) and (7), we conclude that

$$u_i = \ln\Big(F_n(x_i) - F_n(x_{i-1})\Big) + \frac{3}{2}\ln x_i = -\Big(\frac{1}{2x_i}\Big)\gamma + \theta,$$

where $\theta = -\ln c_\gamma$.

Now it is possible ([9,12]) to suggest the estimation γ by regressing $u_i = \theta - \frac{1}{2x_i}\gamma$ on $\frac{1}{2x_i}$ for the following model

$$u_i = \theta - \frac{1}{2x_i}\gamma + \varepsilon_i, \tag{8}$$

where $\varepsilon_i,\ i = 1, 2, ..., n$, are independent identically distributed with $N(0, \sigma^2)$ and also $x = (x_1, ..., x_n)$ is non-random sample (regressor). Using (8) the parameter γ can be estimated (without loss of generality and compare to [9], one of the parameter depends to the other) by regressing u_i on $\frac{1}{2x_i}$.

Now, let us consider the LS estimator for the model (8). It is readily seen that the unbiased LS estimator $\widehat{\gamma}_{LS}$ of the parameter γ in the model (8) is as follows:

$$\widehat{\gamma}_{LS} = -\frac{\sum_{i=1}^{n}\left(\frac{1}{2x_i} - \overline{\frac{1}{2x}}\right)\cdot\left(u_i - \overline{u}\right)}{\sum_{i=1}^{n}\left(\frac{1}{2x_i} - \overline{\frac{1}{2x}}\right)^2}. \tag{9}$$

From (9) and based on (8), we obtain the following corollary.

COROLLARY 2.

$$\widehat{\gamma}_{LS} = -\frac{\frac{3}{2}\cdot\sum_{i=1}^{n}\left(\frac{1}{2x_i} - \overline{\frac{1}{2x}}\right)\cdot\left(\ln x_i - \overline{\ln x}\right)}{\sum_{i=1}^{n}\left(\frac{1}{2x_i} - \overline{\frac{1}{2x}}\right)^2}.$$

EXAMPLE 5. Let us have the real data set in Table 1 and 2, then the LS estimations are, respectively,

$$\widehat{\gamma}_{LS} = 169.73, \quad \widehat{\gamma}_{LS} = 240.57.$$

Now, we have the following theorem.

THEOREM 1. The LS estimator $\widehat{\gamma}_{LS}$ of the parameter γ is

(i) asymptotically normal;

(ii) consistent in a weak sense, i.e. $\widehat{\gamma}_{LS} \xrightarrow{P} \gamma$, as $n \longrightarrow \infty$;

(iii) best unbiased linear (by observations) estimator.

PROOF. To demonstrate asymptotic normality and consistency of the LS estimator $\widehat{\gamma}_{LS}$, it suffices to show that $Var_\gamma[\widehat{\gamma}_{LS}] < \infty$ and $Var_\gamma[\widehat{\gamma}_{LS}] \longrightarrow 0$ when $n \longrightarrow \infty$, respectively, which are met obviously.

In order to establish that $\widehat{\gamma}_{LS}$ is the best unbiased linear estimator, firstly, it is readily seen that $\widehat{\gamma}_{LS}$ is an unbiased estimator, that is $E_\gamma[\widehat{\gamma}_{LS}] = \gamma$. Then, optimality (minimal variance in the class of all linear unbiased estimators) follows by the well-known Gauss-Markov Theorem (see, for example, [8]). In other words, under the following conditions (are satisfied obviously):

$*$ $E_\gamma(\varepsilon_i) = 0$ for all observations.

$*$ $Var_\gamma(\varepsilon_i) = \sigma^2 < \infty,$ so-called "homoskedasticity" condition.

$*$ $Cov_\gamma(\varepsilon_i, \varepsilon_j) = 0,\ \forall i \neq j$ the error terms are uncorrelated.

$*$ X_i is deterministic constant.

It turns out the LS estimator $\widehat{\gamma}_{LS}$ is the best unbiased linear estimator for the parameter γ (ε_i, $i = 1, 2, ..., n$, are residuals or errors).

COROLLARY 3. For the regression model (8), the statistic

$$\widehat{\sigma}_n^2 = \frac{1}{n-2} \sum_{i=1}^{n} \varepsilon_i^2$$

is unbiased estimator of σ^2 and $\chi^2_{n-2} \overset{d}{=} (n-2)\frac{\widehat{\sigma}_n^2}{\sigma^2}$, i.e. the statistic χ^2_{n-2} has χ^2-distribution with $n-2$ degree of freedom.

REMARK 1. This results (in Section 3) may be used for interval estimations and statistical hypothesis testing with regard to the parameters of the model.

REMARK 2. The above mentioned regression model can *not* be built for the PD model. On the other hand, if we consider this method for the PD then it turns out that the LS estimator $\widehat{\rho}_{LS}$ always is equal to 0.

3.2 PD

As we said in Remark 2, the above regression model can not be formed for the PD. So, we need to propose other method (say *second method*) for constructing a regression model. To do this, let $\widehat{\rho}_{ML}$ be the ML estimator of the ρ. It is well-known that the ML estimator is a consistent estimator, i.e. $\widehat{\rho}_{ML} \xrightarrow{P} \rho$. From the well-known property (see, for example, [2]) it follows that for large n,

$$\ln f(x; \widehat{\rho}_{ML}) \xrightarrow{P} \ln f(x; \rho).$$

Therefore (again compare to used method by Koutrouvelis [9]) the regression model is constructed when $b = 0$ as follows

$$z_i = \eta - (\ln x_i)\rho + \varepsilon_i,$$

where $z_i = \ln f(x; \widehat{\rho}_{ML})$, $\eta = -\ln \sum_{y=1}^{\infty} y^{-\rho}$ and $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, 2, ..., n$.

Based on *second method*, the LS estimator is:

$$\widehat{\rho}_{LS} = -\frac{\sum_{i=1}^{n} (\ln x_i - \overline{\ln x})(z_i - \overline{z})}{\sum_{i=1}^{n} (\ln x_i - \overline{\ln x})^2}. \qquad (10)$$

By (10), we obtain that

$$\widehat{\rho}_{LS} = \widehat{\rho}_{ML}.$$

COROLLARY 4. The LS estimator $\widehat{\rho}_{LS}$ is consistent, asymptotically normal and best unbiased linear estimator for the shape parameter ρ.

As we saw in Examples 3 and 4, the data in Table 1 and 2 do *not* give us acceptable ML estimations for the shape parameter ρ. So, it is needed to propose another data for the model PD. We have the following.

EXAMPLE 6. Consider the data $x = 7, 8, 2, 2, 4, 3, 4, 9, 1, 15, 10, 1, 11, 200, 21$, then $\widehat{\rho}_{LS} = \widehat{\rho}_{ML} = 1.21$ (here, the p-value is 0.2505).

COROLLARY 5. The *second method* can be also considered for the scale parameter γ of the DLD. In other words, based on *second method* we have $\widehat{\gamma}_{LS} = \widehat{\gamma}_{ML}$.

4 Conclusions

In this paper we have considered the DLD and PD models. Two real data sets have given for fitting of this frequency distributions in order to model phenomena arising in bioinformatics. It has been seen that the DLD fits such data well with respect to the PD. Note that all of computations and fitting of the models have been done by statistical software "R".

In Section 3, we have proposed two methods for constructing linear regression models with respect to the corresponding parameters and followed by the LS estimators have been obtained. We notice that in the *second method* the LS and ML estimators are the same.

Acknowledgment. The author would like to thank the referees for their valuable comments and suggestions to improve the quality of the paper.

References

[1] J. Astola and E. Danielian, Frequency Distributions in Biomolecular Systems and Growing Networks, Tampere International Center for Signal Processing (TICSP), Series no. 31, Tampere, Finland, 2007.

[2] A. A. Borovkov, Mathematical Statistics. Translated from the Russian by A. Moullagaliev and revised by the author. Gordon and Breach Science Publishers, Amsterdam, 1998.

[3] D. Farbod, The asymptotic properties of some discrete distributions generated by Levy's law, Far East J. Theor. Stat., 26(2008), 121–128.

[4] D. Farbod and K. Arzideh, Asymptotic properties of moment estimators for distribution generated by Levy's Law, Int. J. Appl. Math. Stat., 20(2011), 55–59.

[5] D. Farbod and K. Arzideh, On the properties of a parametric function for distribution generated by Levy's Law, Int. J. Math. Comput., 20(2013), 52–59.

[6] D. Farbod and K. Gasparian, On the maximum likelihood estimators for some generalized Pareto-like frequency distribution, J. Iran. Stat. Soc. (JIRSS) 12(2013), 211–233.

[7] U. Hobohm, M. Scharf, R. Schneider, and C. Sander, Selection of representative protein data sets, Protein Sci. Mar, 1(1992), 409–417.

[8] G. I. Ivchenko and Yu. I. Medvedev, Mathematical Statistics. Translated from the 1984 Russian edition by Elena Troshneva. “Mir”, Moscow, 1990.

[9] I. A. Koutrouvelis, Regression type estimation of the parameters of stable laws, J. Amer. Statist. Assoc. 75 (1980), 918–928.

[10] W. Kabsch and C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers, 22(1983), 2577–2637.

[11] V. A. Kuznetsov, Distribution associated with stochastic processes of gene expression in a single Eukaryotic cell, EURASIP J. Appl. Signal Proc., 4(2001), 285–296.

[12] P. Oliveras and L. Seco, Stable distribution: A survey on simulation and calibration methodologies, Technical Report, (2003), www.risklab.ca/Stableproject.pdf.

[13] V. M. Zolotarev, One-Dimensional Stable Distributions. Translated from the Russian by H. H. McFaden. Translation edited by Ben Silver. Translations of Mathematical Monographs, 65. American Mathematical Society, Providence, RI, 1986.