SORT 32 (2) July-December 2008, 141-150

A microbiology application of the skew-Laplace distribution

Olga Julià[1] and Josep Vives-Rego[2]

Universitat de Barcelona

Abstract

Flow cytometry scatter are ofen used in microbiology, and their measures are related to bacteria size and granularity. We present an application of the skew-Laplace distribution to flow cytometry data. The goodness of fit is evaluated both graphically and numerically. We also study skewness and kurtosis values to assess usefulness of the skew-Laplace distribution.

MSC: 62F10, 62GP10

Keywords: skew-Laplace distribution, goodness of fit, bacteria size.

1 Introduction

The counting, sizing and distribution analysis of particles is a task performed in such diverse fields as archaeology, medicine, geology, biology and technology. The distributions most commonly proposed for describing particle size or its logarithm are the normal, the hyperbolic and the skew-Laplace distributions. The normal distribution, widely used in other fields, is unsuitable for the distribution of bacterial size distribution in axenic cultures (Koch *et al.*, 1987; Vives-Rego *et al.*, 1994). Barndorff-Nielsen (1977) and Bagnold (1980) proposed log-hyperbolic as a suitable model for particle size distribution. However this four-parameter model presents some computational difficulties and shows nearly identical hyperbolic distributions for different parameter value combinations (Fieller (1992)). In 1992, Fieller *et al.* presented the skew-Laplace

[1] Departament de Probabilitat, Lògica i Estadística. Universitat de Barcelona. Spain.

[2] Departament de Microbiologia. Universitat de Barcelona. Spain.

Received: November 2007
Accepted: April 2008

distribution as a simple but effective model for particle sizes. It is easily computed and flexible enough, with the flexibility to handle complex data sets. Kotz *et al.*, (1998) have reported several properties, generalizations and applications of the skew-Laplace distribution. More recently, Puig and Stephens (2007) presented two useful goodness-of-fit tests for this distribution.

We applied the skew Laplace to microbiological data in Julià and Vives-Rego (2005), where we reported the suitable skew-Laplace model for the flow cytometry measures (specifically for the side light scatter) of different microorganisms. In the present paper this study is expanded through the introduction of skewness and kurtosis as goodness-of-fit indicators, following the ideas of Puig and Stephens (2007). The flow cytometry data are introduced in Section 2. In Section 3 the skew-Laplace distribution is described along with some properties and maximum likelihood estimators of its parameters. The results of the skew-Laplace and log-skew-Laplace fitting our data are shown in Section 4 together with different ways of assessing the goodness of fit. The biological relevance and potential applications of the fit of cellular parameter to the skew-Laplace distribution is analyzed and discussed in a forthcoming paper (Vives-Rego *et al.* 2008).

2 The forward and scatter flow cytometry data

The data come from flow cytometry which generates two kinds of measure: the forward (FS) and the side scatter (SS). The forward scatter (FS) sensor is a photodiode that collects the laser light scattered at narrow angles (typically 2-11°) from the axis of the laser beam. When light reaches the FS sensor, the sensor generates voltage pulse signals that are proportional to the amount of light that the sensor receives. Sensitivity is enough to detect 0.5 μm particles. The side scatter (SS) is a photodiode sensor that collects the amount of laser light scattered at an angle of about 90° from the axis of the laser beam. The amount of SS is proportional to the granularity of the cell that scatters the laser light. Forward scatter is preferred to side scatter because it shows high signal intensity and is insensitive to sub-cellular structure. Forward scatter is normally assumed to be proportional to bacterial size.

Three microorganisms have been analyzed: strain 31 and strain 41 from the intestinal faeces of laboratory mice Mus musculus, and *Escherichia coli* strain 536. All strains were analyzed after 24 hours of incubation and no treatment was applied. The flow cytometer distributed the forward (or scatter) measures in 1024 channels, giving a number between 1 and 1024 for each cell. Our data are organized in frequency tables. The sample sizes range between 10,000, for *E. coli* and 120,000 for the other two bacteria. For more microbiological details see Julià and Vives-Rego (2005).

3 The skew-Laplace distribution and maximum likelihood estimation

In this Section we introduce the definition of the skew-Laplace distribution and some properties, useful to fit this distribution. An extensive study of this distribution and its properties can be found in Kotz *et. al* (2001).

The skew-Laplace distribution has the following density:

$$f(x;\alpha,\beta,\mu) = \begin{cases} \exp\left(\dfrac{x-\mu}{\alpha}\right)\Big/(\alpha+\beta) & x \le \mu \\ \exp\left(\dfrac{\mu-x}{\beta}\right)\Big/(\alpha+\beta) & x > \mu \end{cases} \tag{1}$$

where α, $\beta > 0$ and $\mu \in \mathbb{R}$. When the logarithm is applied we obtain two straight lines with $1/\alpha$ and $-1/\beta$ slopes that intersect at μ. This fact can be used to check approximately the goodness of fit. The parameter μ is the mode, and in the symmetric case, when $\alpha = \beta$, is also the mean. If X is a random variable skew-Laplace distributed, the mean and variance are:

$$\mathbf{E}[X] = \mu + \beta - \alpha$$

$$\sigma^2 = \alpha^2 + \beta^2.$$

The coefficients of skewness and kurtosis are the following:

$$\gamma_1 = \frac{\mathbf{E}[(X-\mathbf{E}(X))^3]}{\sigma^2} = 2\frac{\beta^3-\alpha^3}{(\alpha^2+\beta^2)^{\frac{3}{2}}}$$

$$\gamma_2 = \frac{\mathbf{E}[(X-\mathbf{E}(X))^4]}{\sigma^4} = 3 + 6\frac{\alpha^4+\beta^4}{(\alpha^2+\beta^2)^2}$$

As it is reported in Puig and Stephens (2007) the skewness value determines the kurtosis value, but not viceversa because the same kurtosis corresponds to γ_1 and $-\gamma_1$. This relationship can be used to assess whether the skew-Laplace is appropriate. The possible values of skewness and kurtosis are $\gamma_1 \in (-2, 2)$ and $\gamma_2 \in [6, 9)$.

The maximum likelihood estimators

The maximum likelihood method is used to estimate the skew-Laplace parameters of (1). The mathematical derivation of those estimators can be found for example in Kotz *et al.* (2001) or Puig and Stephens (2007). The maximum likelihood estimator of μ,

denoted by $\hat{\mu}$, can be obtained by a simple algorithm since $\hat{\alpha}$ and $\hat{\beta}$, maximum likelihood estimators of α and β respectively, have an explicit expression depending on $\hat{\mu}$. Indeed, let $x_1, \ldots, x_n$ be a sample coming from a skew-Laplace distribution: we then consider the following functions:

$$\Delta(\mu) = \frac{1}{n} \sum_{i=1}^{n} |x_i - \mu|$$

$$\psi(\mu) = \Delta(\mu) + \sqrt{\Delta^2(\hat{\mu}) - (\bar{x} - \hat{\mu})^2}.$$

Then, the maximum likelihood estimators are given by:

$$\hat{\mu} = x_j \tag{2}$$

$$\hat{\alpha} = \frac{1}{2}\left(\Delta(\hat{\mu}) - \bar{x} + \hat{\mu} + \sqrt{\Delta^2(\hat{\mu}) - (\bar{x} - \hat{\mu})^2}\right) \tag{3}$$

$$\hat{\beta} = \frac{1}{2}\left(\Delta(\hat{\mu}) + \bar{x} - \hat{\mu} + \sqrt{\Delta^2(\hat{\mu}) - (\bar{x} - \hat{\mu})^2}\right) \tag{4}$$

where x_j is any sample value where the function $\psi(x_j)$ attains its single minimum. Note that ψ could attain its single minimum for two or more x_j sample values.

A simple proof of the derivation of maximum likelihood estimators for the skew-Laplace distribution can be found in Puig and Stephens (2007).

4 Results

In order to see if the flow cytometric scatter data fit the skew-Laplace distribution, we first plot frequencies logarithms versus size values in Figure 1. As we noted in Section 3, two straight lines will appear when the Laplace distribution is appropriate. Even though the parameters of skew-Laplace can be estimated by fitting two straight lines in plots of Figure 1, the maximum likelihood method is preferable. The maximum likelihood estimators are calculated following the steps described in Section 3. In all cases the minimum of function Ψ is reached at only one sample value. Histograms with their estimated skew-Laplace density can be found in Figure 2. Samples with good fits and samples with not such good fits can clearly be seen. Our explanation to the fact that some data sets are not well fitted by the skew-Laplace distribution is that some unknown biological factors are modifying the standard biomass distribution in the culture.

The sample sizes of our data range between 10,000 and 120,000, therefore the p-values of any test of goodness of fit are too small to be useful for comparing goodness of fit. In order to assess the usefulness of the skew-Laplace distribution in Julià and Vives-Rego (2005) we calculated the critical size, N_{crit}. This statistic, proposed by Fieller *et al.*

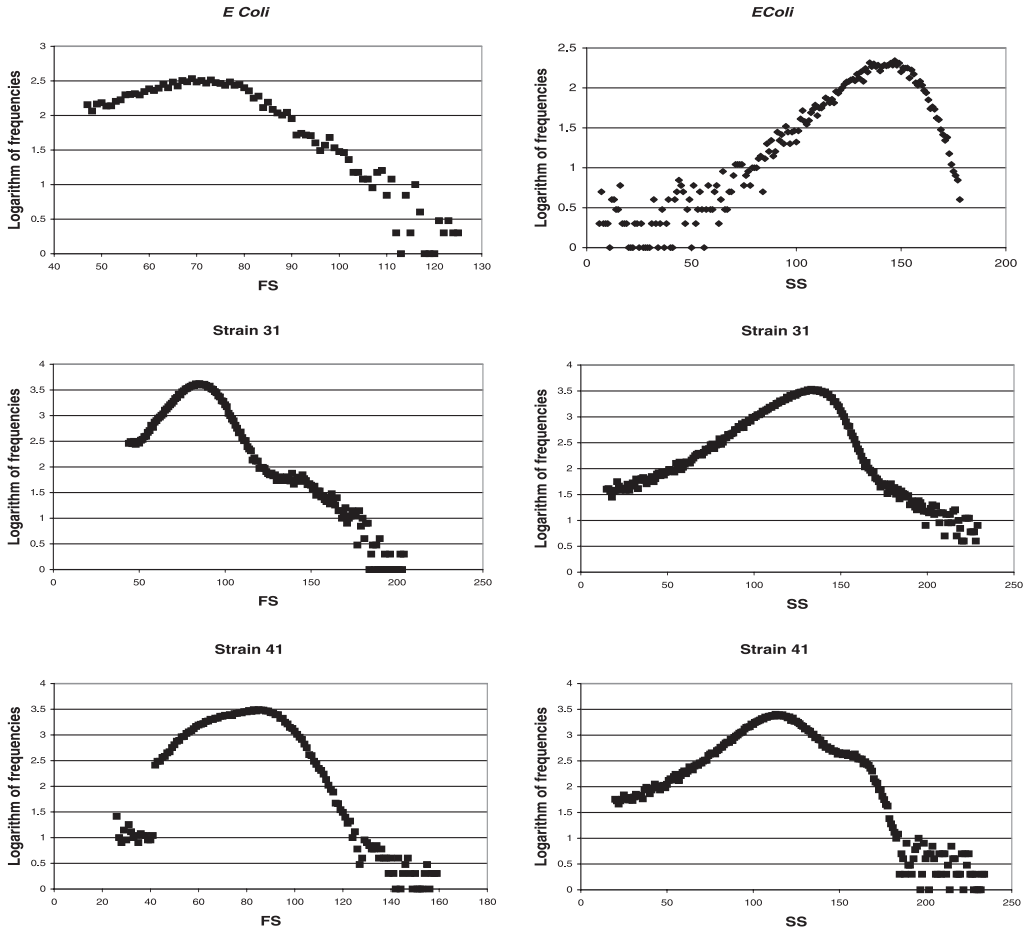

Figure 1: *Logarithm of frequencies versus cytometry FS a SS values.*

Table 1: *Empirical and skew-Laplace theoretical values of skewness and kurtosis.*

Table 1	empirical		estimated skew-Laplace	
	skewness	kurtosis	skewness	kurtosis
E. Coli FS	0.7570	4.9947	0.5236	6.1838
E. Coli SS	−1.5885	7.4991	−1.5655	7.7273
Strain 31 FS	0.9114	7.4101	0.0167	6.0002
Strain 31 SS	−0.8608	3.9256	−1.7929	8.3221
Strain 41 FS	−0.0981	2.8955	−0.9082	6.5588
Strain 41 SS	−0.3166	5.2102	−0.3040	6.0617

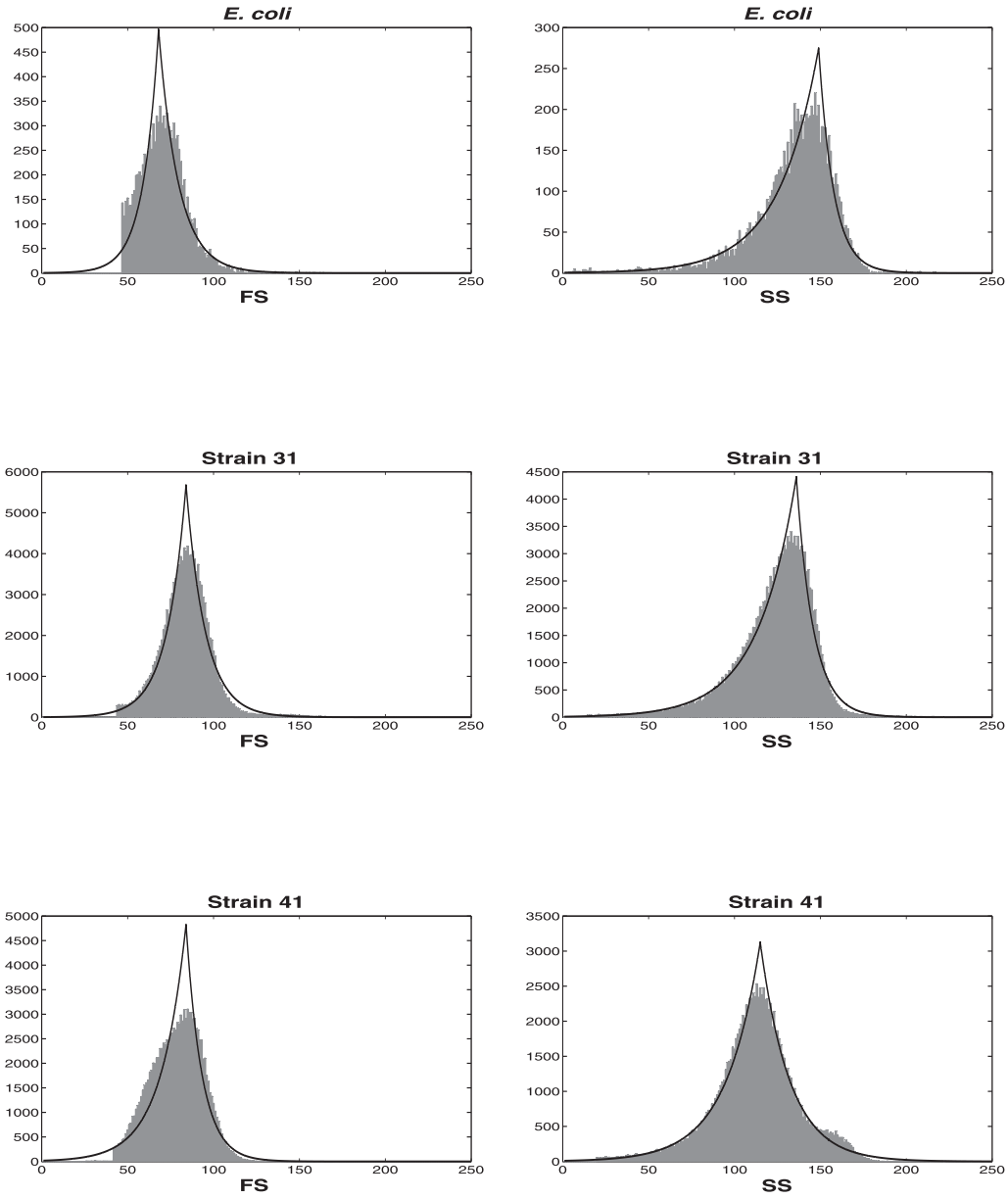

Figure 2: *Skew-Laplace fitting of cytometry FS a SS data. The histogram is shaded in grey and the continuous profile is the estimated skew-Laplace.*

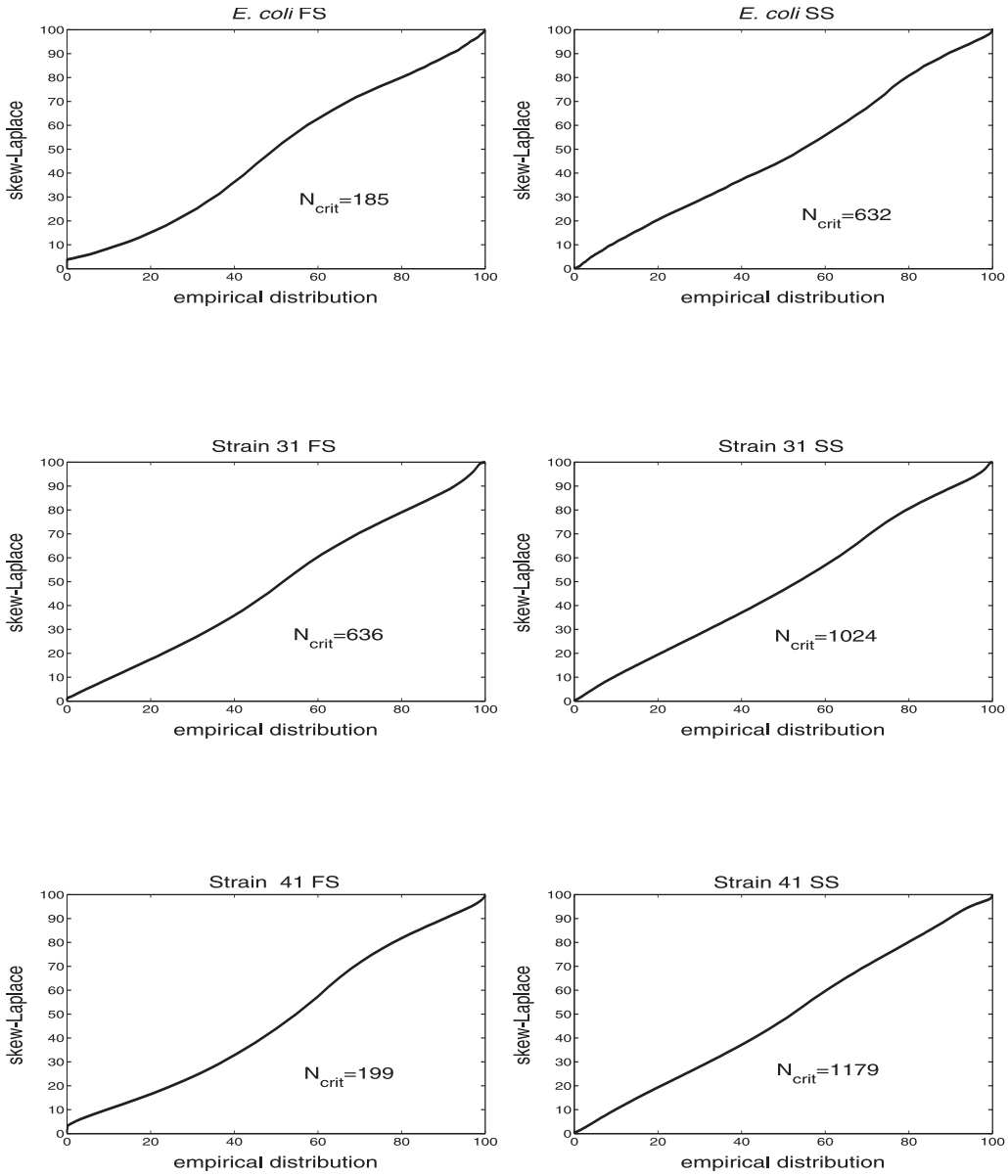

Figure 3: *The q-q plot together N_{crit} values.*

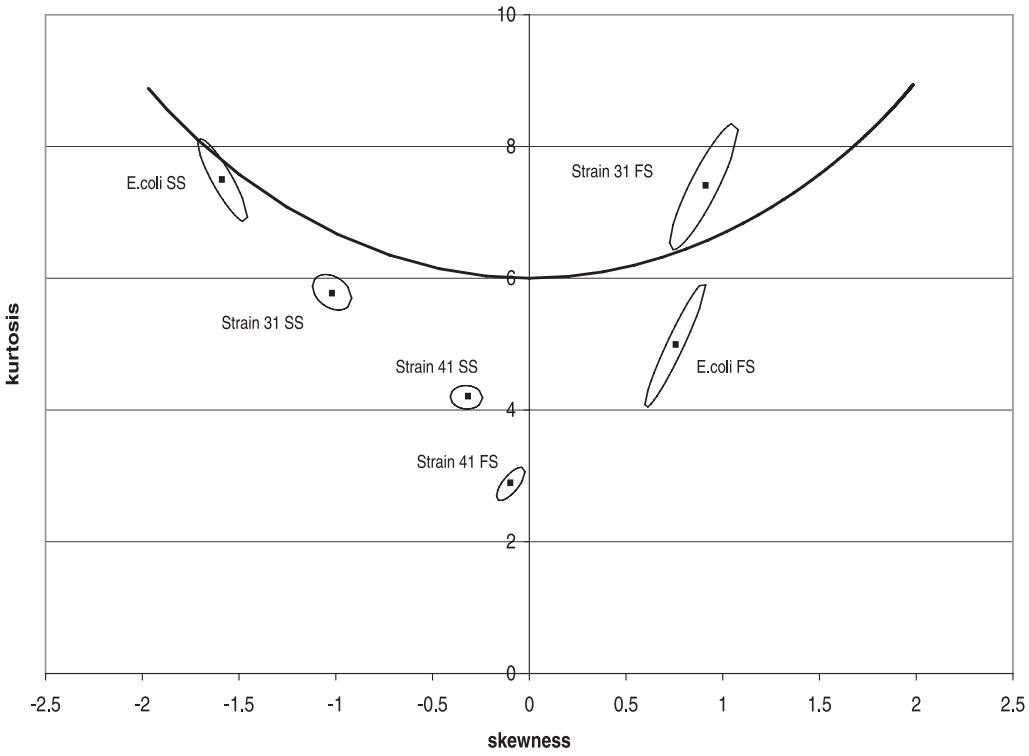

Figure 4: *Skewness and kurtosis of skew-Laplace are represented in solid line, and empirical skewness and kurtosis of cytometry FS and SS values plotted in dots. For each point the 95% confidence region is showed.*

(1992), is based on the chi-square goodness-of-fit test. The N_{crit} statistic is defined as:

$$N_{\text{crit}} = \frac{\chi^2_{k-m-1,\ 0.95}}{\sum_1^k \left(r_i - p_i(\hat{\theta})\right)^2 / p_i(\hat{\theta})}$$

where k represents the number of intervals, m the number of estimated parameters, r_i and $p_i(\hat{\theta})$ are the sample proportion and the estimated skew-Laplace probability of the respective interval. In order to standardize the procedure, we took 40 identical probability intervals for each sample. This statistic can be interpreted as the critical sample size, required just to detect a lack of fit at the 5% level, disregarding the fact that maximum likelihood estimations could be calculated from the grouping data. In Figure 3 the N_{crit} values are shown on each q-q plot. As we can see, greater values of N_{crit} correspond to straighter lines in the q-q plot. In Puig and Stephens (2007) the skewness and kurtosis values are used to build a goodness-of-fit test. Although this test is not appropriate in our case because we have grouped data, we can use this idea to connect the proximity of theoretical skewness and kurtosis values to the empirical values, with

the goodness of fit. The theoretical curve of skew-Laplace skewness and kurtosis and the empirical values are shown in Figure 4. We have also added for each empirical point the 95 % confidence region obtained using the bootstrap method. It can be seen that values near the curve belong to samples with good fit, but it is difficult to assess the goodness of fit according only to their proximity. Table 1 shows the empirical skewness and kurtosis values together with the skew-Laplace values using he estimated parameter.

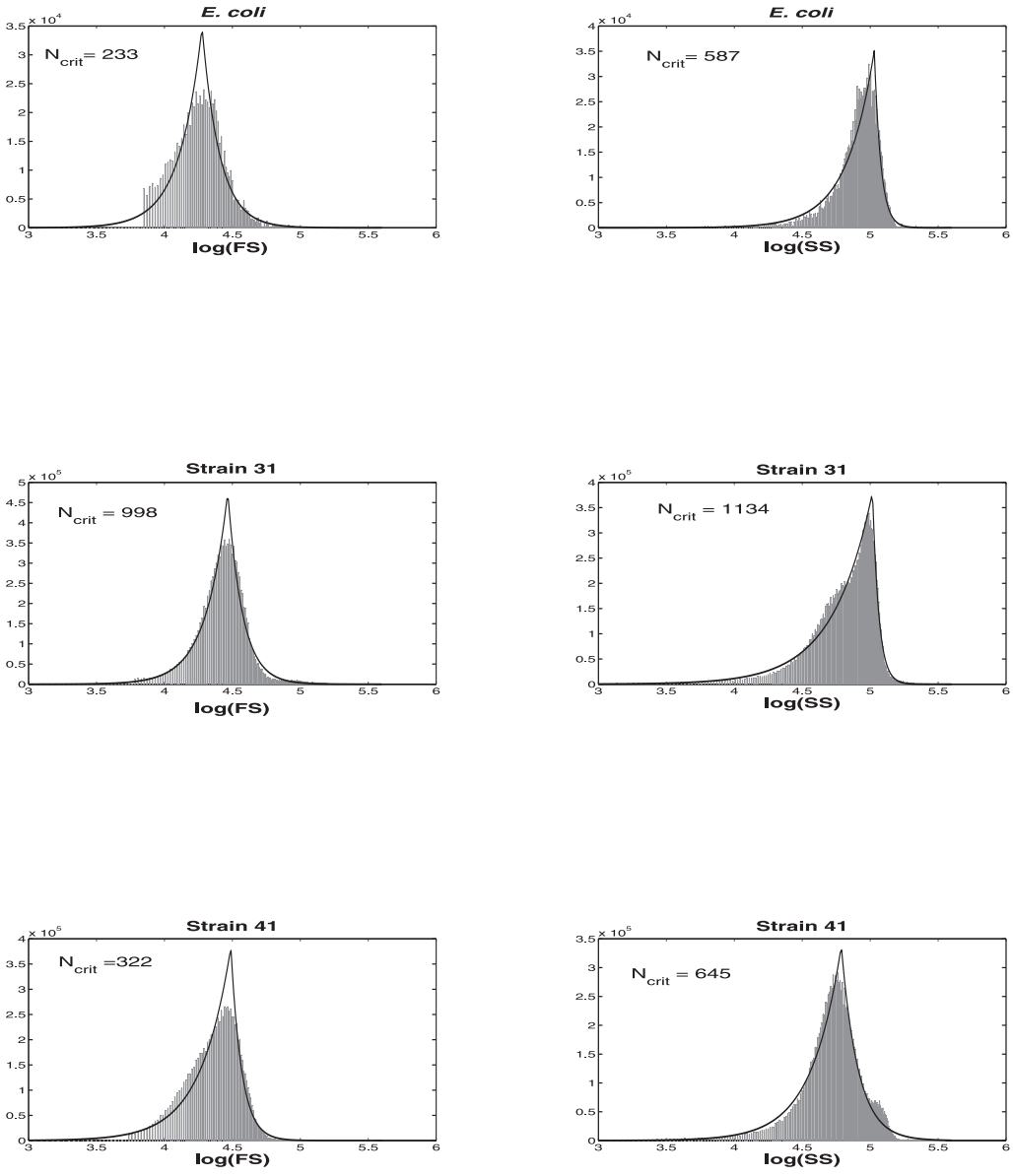

Figure 5: *Log-skew-Laplace fitting of cytometry FS a SS data with N_{crit} values. The histogram is shaded in grey and the continuous profile is the estimated log-skew-Laplace.*

Skew-Laplace versus log-skew-Laplace

According to Fieller (1992) the models based on log-size are more useful not only due to their wide range of particle size, but also because of the multiplicative process of breakage underlying particle production. Even though this multiplicative effect is not clearly applicable to our bacteria size data, we found that in some cases the goodness of fit improves if logarithms are taken. In Figure 5 we can see the histogram and the estimated log-skew-Laplace density. We have also added the respective N_{crit}.

As we can see only in the case of *E. coli* SS and Strain 31 SS no improvement is observable.

Acknowledgments

We thank the referee for their comments and suggestions. This work has been partially supported by grant MTM2005-08886 from Ministerio de Ciencia y Tecnología.

References

Bagnold R. A. and Barndorff-Nielsen, O. (1980). The pattern of natural size distributions. *Sedimentology*, 27, 199-207.

Barndorff-Nielsen, O. (1977). Exponentially decreasing distributions for the logarithm of particle size. *Proceedings of the Royal Society of London* A, 353, 401-419.

Fieller, N. R. J., E. C. Flenley and Olbricht, W. (1992). Statistics of particle size data. *Applied Statistics*, 41, 127-146.

Julià, O. and Vives-Rego, J. (2005). Skew-Laplace distribution in Gram-negative bacterial axenic cultures: new insights into intrinsic cellular heterogeneity. *Microbiology*, 151, 749-755.

Koch, A. L. (1987). The variability and individuality of the bacteria. In: Neidhart, C. (Ed.), *Escherichia coli and Salmonella typhymurium Cellular and Mollecular Biology*. American Society for Microbiology, Washington, DC, 1606-1614.

Kotz, S., T. J. Kozubowski and Podgorski, K. (2001). *The Laplace distribution and generalizations*. Birkhäuser, Berlin.

Puig, P. and Stephens, M. A. (2007). Goodness of fit tests for the skew-Laplace distribution. *Statistics and Operations Research Transactions*, 31, 45-54.

Vives-Rego J, López-Amorós, R. and Comas, J. (1994). Flow cytometric narrow-angle light scatter and cell size during starvation of E.coli in artificial sea water. *Letters in Applied Microbiology*, 19, 374-376.

Vives-Rego, J., Julià, O., Vidal-Mas, J. and Panikov, N. S. (2008). *The flow cytometric scatters of bacterial axenic cultures fit the skew-Laplace distribution pattern: biological consequences*. Submitted, preprint No.400 IMUB (http://www.imub.ub.es/publications/preprints/pdf/PreprintN400.pdf)