NOTES AND CORRESPONDENCE

Fitting an Exponential Distribution

ROBERTO FRAILE AND EDUARDO GARCÍA-ORTEGA

Departamento de Física, Universidad de León, León, Spain

(Manuscript received 1 September 2004, in final form 1 March 2005)

ABSTRACT

Exponential distributions of the type $N = N_0 \exp(-\lambda t)$ occur with a high frequency in a wide range of scientific disciplines. This paper argues against a widely spread method for calculating the λ parameter in this distribution. When the ln function is applied to both members, the equation of a straight line in t is obtained, which may be fit by means of linear regression. However, the paper illustrates that this is equivalent to a least squares fit with a weight function that assigns more importance to the higher values of t. It is argued that the method of maximum likelihood should be applied, because it takes into account all of the data equally. An iterative method for determining λ is proposed, based on the method of moments for cases in which only a truncated distribution is available.

1. Introduction

In physics there are many magnitudes that depend on others in an exponential way. In other cases, the values of a magnitude follow an exponential frequency. An example of the former is the number of radioactive nuclei that persist in time without disintegrating. Drop size distributions are an example of the latter case. The exponential function is the same in both cases, but the two concepts are essentially different.

In the case of exponential dependence, the equation that rules the behavior of radioactive nuclei is

$$N = N_0 \exp(-\lambda t), \tag{1}$$

where N is the number of unstable nuclei remaining after a time span t, with N_0 being the initial number of nuclei. In this equation the parameter of the exponential λ (called the disintegration constant) is related to the half-life of the radioactive isotope. If we have several pairs of data (N_i, t_i), we can determine λ from Eq. (1) merely by taking the natural logarithm of both members:

$$\ln N = \ln N_0 - \lambda t, \tag{2}$$

which is the equation of a straight line, if we consider time t the independent variable and ln N the dependent variable. Thus, a simple least squares fit allows us to obtain an estimation of the slope λ.

In the case of exponential frequency, Eq. (1) is also valid by simply considering that N is the number of drops per unit of volume according to size t. This is the drop size distribution proposed by Marshall and Palmer (1948), which Smith (2003) compares to the probability density function as follows:

$$f(x) = \lambda \exp(-\lambda x). \tag{3}$$

The difference may seem insignificant at first sight, but a controversy arises when it comes to calculating λ.

2. Calculating λ

In the second case, the value of N in Eq. (1) represents the number of drops with a size of approximately t, which means that all of the drops in the sample have to be grouped into classes. Because the distribution is exponential, it may happen that in the larger sizes some of the classes are empty. Fraile et al. (1992) have noted that in this case it is not possible to use Eq. (2) for determining λ by the least squares fit because if $N = 0$, ln$N$ makes no sense.

Corresponding author address: Roberto Fraile, Departamento de Física, Facultad de CC Biológicas y Ambientales, 24071 León, Spain.
E-mail: robertofraile@unileon.es

These authors have also noted that, even if there are no empty size groups, the least squares fitting of the straight line in Eq. (2) places more importance on the larger sizes.

From an experimental perspective, there are instrumental problems when accurately measuring the smallest hydrometeors. These problems are the result of a number of different causes. For instance, the hydrometeors may not reach the lower threshold, or the problems may be the result of the method, the resolution of the equipment, overlapping, etc. Korolev et al. (1998) have pointed out some of these indeterminacies in measuring the number of small hydrometeors with an Optical Array Probe 2D2-C. Fraile et al. (2004) have also noticed these problems in measuring hailstones.

Furthermore, it is well known that in microphysical measurements the largest particles are more important in estimating the liquid water content, because this parameter is proportional to the cubed diameter of the drop. Similarly, the largest hydrometeors are also the ones that contribute more importantly to the reflectivity factor, because this factor depends on the sixth power of the diameter of the drops.

This may award greater importance to the larger sizes of the spectrum. However, if the aim is to calculate size distributions, it is necessary to employ a method that applies equally to the whole size spectrum. If one part of the spectrum needs to be highlighted, an appropriate weight should be established for it, but not the weight that a particular fitting method may introduce.

The aim of this paper is both to demonstrate that the least squares fit of the straight line in Eq. (2) introduces a bias, and to offer an alternative way for calculating λ parameter in an exponential distribution.

a. *Differences in weight factors in the least squares fit*

If we have a set of n data points (x_i, y_i) that follow an approximate relationship of the following type:

$$y_i \cong f(x_i) = Y_i,$$

the form of f is known, even though the k parameters A_j $(j = 1, \ldots, k)$ on which it depends are unknown. Calculating A_j by means of the least squares fit is equivalent to estimating the value of the parameters that minimize the quantity

$$D = \sum_i (Y_i - y_i)^2. \tag{4}$$

Then, the system of k equations that leads to determining k parameters is

$$\frac{d}{dA_j} \sum_i (Y_i - y_i)^2 = 0.$$

In case some of the points are more reliable than others, we may incorporate a weight factor w_i for each pair indicating their reliability. Consequently, instead of Eq. (4) we will have to minimize

$$D_w = \sum_i w_i(Y_i - y_i)^2. \tag{5}$$

If the points (x_i, y_i) still show an approximately exponential relationship, that is, if

$$y_i \cong N_0 \exp(-\lambda x_i) = Y_i, \tag{6}$$

then the least squares fit is equivalent to estimating the value of λ and N_0 that will minimize the quantity

$$D = \sum_i (Y_i - y_i)^2 = \sum_i [N_0 \exp(-\lambda x_i) - y_i]^2, \tag{7}$$

and, if the reliability of the points were different, the expression that would have to be minimized would be

$$D_w = \sum_i w_i(Y_i - y_i)^2 = \sum_i w_i[N_0 \exp(-\lambda x_i) - y_i]^2. \tag{8}$$

It will be illustrated below that the least squares fitting to the straight line in Eq. (2)—built from an exponential—will lead to an expression that is similar to the one in Eq. (8), that is, attributing different weights to the points.

It is certainly the case that if the points (x_i, y_i) comply with Eq. (6), they also confirm that

$$\ln y_i \cong \ln N_0 - \lambda x_i = Z_i,$$

which is a linear equation in x_i. The least squares fitting to that straight line is equivalent to estimating the value of λ and N_0 that minimize the quantity

$$D_l = \sum_i (Z_i - \ln y_i)^2 = \sum_i (\ln N_0 - \lambda x_i - \ln y_i)^2.$$

If we call the relative residual $p_i = [N_0 \exp(-\lambda x_i) - y_i]/y_i$, then

$$D_l = \sum_i [\ln(p_i + 1)]^2.$$

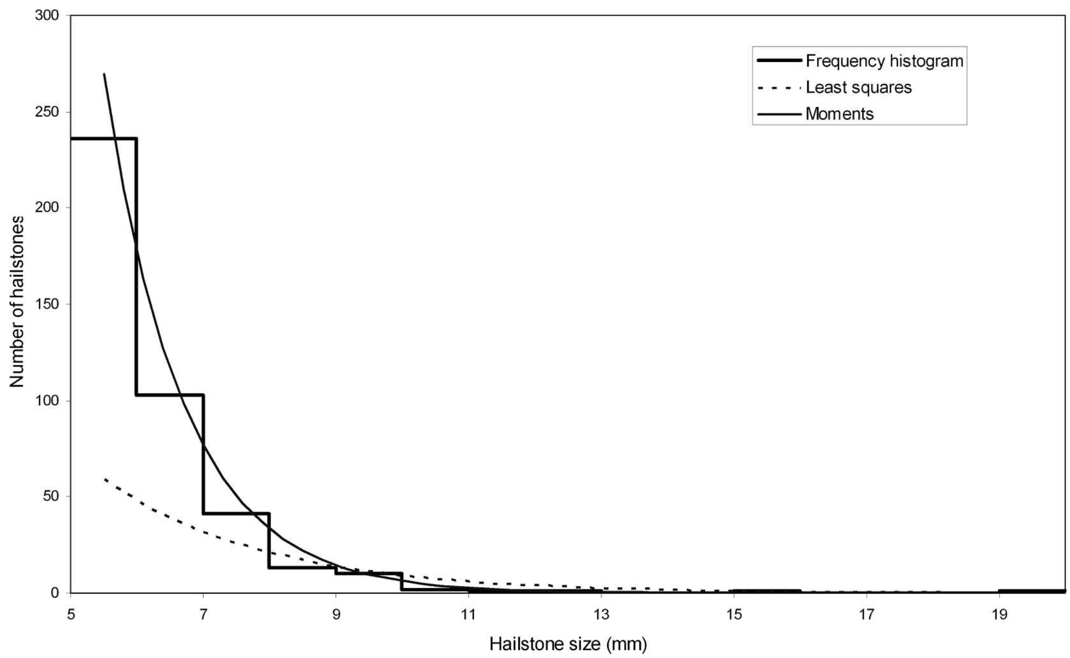

FIG. 1. Histogram of the frequency of hailstone sizes in a hailstorm in León (Spain). The figure also includes the exponential distributions that fit best with the least squares method (with a poor fit in the smaller sizes) and the moment method.

Because, according to Eq. (6), p_i is very small, the Mercator series (a Taylor expansion of the natural logarithm) leads to

$$D_l = \sum_i \left(p_i - \frac{p_i^2}{2} + \frac{p_i^3}{3} - \frac{p_i^4}{4} + \dots\right)^2.$$

As second-order and higher-order terms are neglected, the result is

$$D_l \cong \sum_i p_i^2 = \frac{1}{N_0^2} \sum_i [\exp(2\lambda x_i)][N_0 \exp(-\lambda x_i) - y_i]^2. \tag{9}$$

It is obvious that minimizing Eq. (9) is equivalent to minimizing Eq. (8) if the weight w_i is

$$w_i = \frac{\exp(2\lambda x_i)}{N_0^2}. \tag{10}$$

What does this mean? It simply refers to the fact that transforming an exponential distribution into a linear function to subsequently estimate the parameters of the line by means of the least squares fit is broadly equivalent to applying the least squares fit to the exponential function with a different weight assigned to each point (x_i, y_i). In addition, because w_i is a growing exponential function, more weight is assigned to the higher values of x_i. In consequence, the fit is better for the points with a higher x_i value. This fact can be illustrated graphically when representing an exponential distribution with the value λ calculated in the way described above, together with the data points (x_i, y_i) from which λ has been estimated. These points will be farther away from the exponential curve for the lower values of x_i. This is the usual result when this method is applied to calculate λ (Fraile et al. 1992). In Fig. 1 it can be compared with the exponential function that is obtained from calculating λ with the moment method. In this case, hailstones larger than 5 mm have been measured, and, consequently, an extension of this method was used, as described in section 2b.

In conclusion, it may seem a paradox that with this method the calculated λ depends on a weight function (10), which, in its turn, depends on λ. Therefore, we suggest that the value of λ should be calculated from the probability density function (3) by means of the moment method. This method is generally known to be biased (Wallis et al. 1974). However, in the case of an exponential distribution it is identical to the maximum likelihood method (Sneyers 1990), which is not biased. Other methods may equally be used, for instance, the chi-square minimization method (Cramér 1999). Moreover, if we take into account the fact that the exponential distribution is a particular case of the gamma distribution

$$f(x) = \frac{(x/\beta)^{\alpha-1}\exp(-x/\beta)}{\beta\Gamma(\alpha)}$$

when the shape parameter is $\alpha = 1$, other methods may be used, for instance, the ones proposed by Wilks (1990) for the gamma.

b. A method for fitting truncated exponential distributions

The estimated value of λ according to the method of moments is the inverse of the mean value of the sample. The exponential distribution lies between zero and ∞. But there are situations in which only one range of values can be observed; this is the case of solid precipitation, which is only labeled "hail" if the size surpasses 5 mm in diameter (WMO 1992), or in the case of drops, if the data are provided by equipment that measures only sizes in a particular interval. In both cases, the mean value of the sample does not coincide with the inverse of the expected value of λ. This could be a valid point to insist on using the linear fit, because the method of moments cannot be used in these conditions. However, a simple procedure will be suggested below for calculating λ on the basis of the method of moments.

If we call the minimum threshold x_0 (e.g., $x_0 = 5$ mm in the case of hail), the expected value of x between x_0 and ∞ is

$$E(x) = \frac{\int_{x_0}^{\infty} xf(x)\,dx}{\int_{x_0}^{\infty} f(x)\,dx} = \frac{\int_{x_0}^{\infty} x\lambda\exp(-\lambda x)\,dx}{\int_{x_0}^{\infty} \lambda\exp(-\lambda x)\,dx} = x_0 + \frac{1}{\lambda},$$

and approximating $E(x)$ by means of

$$\bar{x} = \frac{\sum_i x_i}{n} \tag{11}$$

if we have a sample of n data x_i, the value of λ is

$$\lambda = \frac{1}{\bar{x} - x_0} = \frac{1}{\frac{\sum_i x_i}{n} - x_0} = \frac{n}{\sum_i (x_i - x_0)},$$

which is very easy to calculate, because it is a change in the coordinates x_i, taking x_0 as the point of origin.

If the distribution is truncated on both sides and we call the lower threshold x_0 and the upper one x_u, the expected value of x is

$$E(x) = \frac{\int_{x_0}^{x_u} x\lambda\exp(-\lambda x)\,dx}{\int_{x_0}^{x_u} \lambda\exp(-\lambda x)\,dx} = \frac{x_0\exp(-\lambda x_0) - x_u\exp(-\lambda x_u)}{\exp(-\lambda x_0) - \exp(-\lambda x_u)} + \frac{1}{\lambda},$$

and we may reorganize this as follows:

$$\lambda = \frac{1}{E(x) - \frac{x_0\exp(-\lambda x_0) - x_u\exp(-\lambda x_u)}{\exp(-\lambda x_0) - \exp(-\lambda x_u)}}. \tag{12}$$

Consequently, if we want to fit n values of $x_i \in [x_0, x_u]$ to a truncated exponential, the expected value of x will be given by Eq. (11). Equation (12) may be used for calculating a new value of λ, starting from an initial value that is introduced in the member on the right-hand side. In other words, Eq. (12) allows us to determine the parameter of the exponential distribution by means of subsequent iterations: for a given value of λ_k, we obtain the value of $\lambda_{k+1} = f(\lambda_k)$, where

$$f(\lambda_k) = \frac{\exp(-\lambda_k x_0) - \exp(-\lambda_k x_u)}{(\bar{x} - x_0)\exp(-\lambda_k x_0) - (\bar{x} - x_u)\exp(-\lambda_k x_u)}. \tag{13}$$

The λ error can be made as small as required. If we call this error ε, after m iterations the following will result: $\lambda_{m+1} - \lambda_m < \varepsilon$.

Function $f(\lambda_k)$ tends to a fixed point under iteration, that is, Eq. (13) converges, if the slope of the curve Eq. (13) in the fixed point (i.e., the derivative of the function in that point) takes values between -1 and 1 (Kaplan and Glass 1995). The function derived from Eq. (13) with respect to λ_k is

$$f'(\lambda_k) = \frac{(x_u - x_0)^2\exp(-\lambda_k x_0)\exp(-\lambda_k x_u)}{[(\bar{x} - x_0)\exp(-\lambda_k x_0) - (\bar{x} - x_u)\exp(-\lambda_k x_u)]^2},$$

which is always positive, because both the numerator and denominator are always positive. This result indicates that whether the function is convergent or not, the trend is monotonic (not oscillatory). Once the derivative has been seen to be positive, to verify the condition mentioned in the previous paragraph it is only necessary to test that

$$f'(\lambda_k) < 1,$$

which is equivalent to testing that

$$\lambda_k > \frac{2}{x_u - x_0}\ln\frac{x_u - \bar{x}}{\bar{x} - x_0}. \tag{14}$$

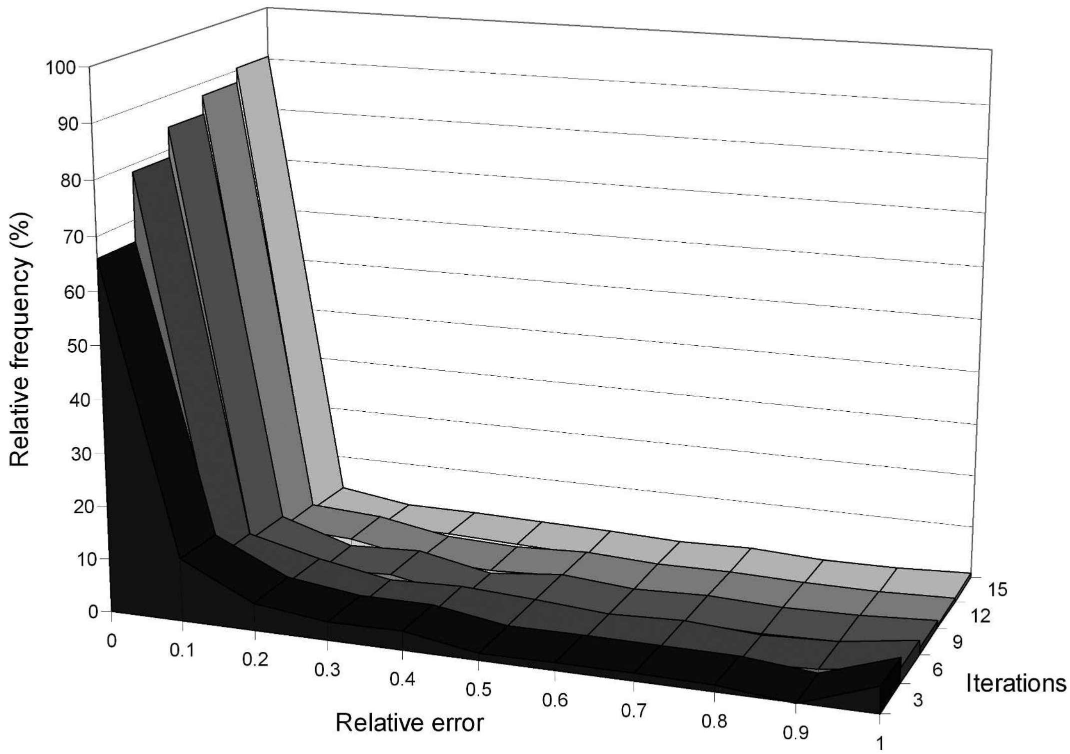

FIG. 2. The frequency of the relative error, understood as $(\lambda - \lambda_{final})/\lambda_{final}$, as the number of iterations for calculating parameter λ increases.

To assess the extent to which the convergence criterion in Eq. (14) applies to the examples suggested (hailstone and drop size), we have applied this equation to hail data that are registered during a summer campaign in the hailpad network of León (Spain) and to the distributions of hydrometeors measured inside convective clouds by means of an optical array probe (OAP) 2D2-C (Sánchez et al. 2001). The result shows that in all cases the inequality persists, thus, demonstrating that the iteration function is useful, at least in the examples described. In addition, not too many iterations are necessary to obtain acceptable results. For example, for the hailpads chosen, with five iterations λ takes a value that differs from the final value (the one it will supposedly reach with infinite iterations), which is less than 5% in 60% of the cases. Moreover, in 60% of the cases the λ that is calculated with 10 iterations approaches the final value with a difference of less than 1%. With 15 iterations this percentage increases to almost 70%. The gradual decrease of the difference with the final result (or relative error) is represented in Fig. 2 according to the number of iterations.

In the paragraphs above we have calculated λ when there is only a lower threshold x_0 and when there are upper and lower thresholds in the sample. The remaining case—an upper threshold only—is equivalent to using Eq. (12) when $x_0 = 0$. As expected (and this may be easily tested), if x_u is very high, the value of λ will be very similar to the inverse of the mean value. In any case, whatever the value of x_u, the new λ will always be lower than the one calculated directly as the inverse of the mean value.

3. Conclusions

- This paper has demonstrated that calculating the λ parameter in an exponential distribution by means of the least squares fitting to the straight line in Eq. (2) incorporates a weight factor that assigns more importance to the higher values of the independent variable.
- In consequence, we suggest that this parameter should be calculated by means of the method of moments or maximum likelihood, whose results are identical in the exponential distribution.
- There are also exponentially distributed data that do not extend to the whole domain of the exponential distribution (from zero to infinity), usually resulting from the sampling techniques employed. In these cases, the sample has to be restricted to a reduced interval, and we suggest the use of a straightforward iterative technique based on the method of moments.

Acknowledgments. The authors are grateful to Dr. Maite Trobajo for her suggestions and to Dr. Noelia Ramón for translating the paper into English. The reviewers have greatly contributed to improving the paper. The study was supported by the Consejería de Educación y Cultura, Junta de Castilla y León (Grant LE53/03).

REFERENCES

Cramér, H., 1999: *Mathematical Methods of Statistics*. Princeton University Press, 575 pp.

Fraile, R., A. Castro, and J. L. Sánchez, 1992: Analysis of hailstone size distributions from a hailpad network. *Atmos. Res.,* **28,** 311–326.

——, L. López, C. Palencia, and A. Castro, 2004: Hailstone size distribution and dent overlap in hailpads. *Proc. 14th Int. Conf. on Clouds and Precipitation*, Bologna, Italy, ICCP, 763–766.

Kaplan, D., and L. Glass, 1995: *Understanding Nonlinear Dynamics.* Springer-Verlag, 440 pp.

Korolev, A. V., J. W. Strapp, and G. A. Isaac, 1998: Evaluation of the accuracy of PMS optical array probes. *J. Atmos. Oceanic Technol.,* **15,** 708–720.

Marshall, J. S., and W. M. Palmer, 1948: The distribution of raindrops with size. *J. Meteor.,* **5,** 165–166.

Sánchez, J. L., E. García-Ortega, and J. L. Marcos, 2001: Characterization of the distribution of cloud spectra for thunderstorms in the western Mediterranean area. Preprints, *Symp. on Precipitation Extremes: Prediction, Impacts, and Responses*, Albuquerque, NM, Amer. Meteor. Soc., 129–132.

Smith, P. L., 2003: Raindrop size distributions: Exponential or gamma—Does the difference matter? *J. Appl. Meteor.,* **42,** 1031–1034.

Sneyers, R., 1990: On the statistical analysis of series of observations. WMO Tech. Note 143, 192 pp.

Wallis, J. R., N. C. Matalas, and J. R. Slack, 1974: Just a moment. *Water Resour. Res.,* **10,** 211–219.

Wilks, D. S., 1990: Maximum-likelihood-estimation for the gamma-distribution using data containing zeros. *J. Climate,* **3,** 1495–1501.

WMO, 1992: *International Meteorological Vocabulary*. World Meteorological Organization, 784 pp.