From sand grains to tomatoes: applications of size distributions.

N. R. J. Fieller

Department of Probability & Statistics
University of Sheffield
Sheffield S3 7RH, U.K.

e-mail: n.fieller@sheffield.ac.uk

22 February 2002

1. Introduction

Particle size data are collected to determine the overall size distribution of a collection of particles since it is this which characterizes the formation processes of the particles. Additionally, this distribution may influence further behaviour of the particles. For example, geologists measure sizes of sand grains since the size distributions from water deposited samples and wind deposited samples can be distinguished and so samples from unknown environments can be classified, of interest in oil exploration and in past climate studies. Medical researchers are interested in the sizes of blood cells since this reflects the health of the coronary system and influences the prognosis for coronary disease. Materials engineers are interested in the size distribution of silicon carbide partilces added to metals (aluminium) to make high-performance materials since this can influence the mechanical properties of the materials (crack formation, resilience etc.). Fuel technologists measure droplet size of fuels to determine their combustion properties. The grain size of feedstuffs fed to farm animals influences the digestibility and energy conversion of the feed.

Inevitably, such a wide variety of disciplines has generated a wide variety of *ad hoc* techniques for processing particle size data. These include simple calculation of sample moments, principal component and factor analysis and modelling of the size distributions by normal, log-normal and Weibull distributions. Much of this work has been developed by specialists on fields other than statistics and it is only comparatively recently that statisticians have given attention to the wide variety of problems that arise.

This presentation will concentrate on various applications of modelling size distributions using log-skew-Laplace distributions in particular. These combine computational ease and the flexibility to handle more complex extensions. One of the by-products of using such models is that they can be used to answer other questions of interest not directly related to the size distributions themselves. These include problems of average shapes of soil grains, 'thresholding' in image analysis and the relationship between size and weight of tomatoes (of major commercial importance to horticulturalists).

2. Outline

The presentation begins with a brief discussion of *size* of an object and how different measuring techniques measure different apsects of size. The relationship between these aspects is a property of the shape of the object. A brief justification is given of the use of log-normal, log-hyperbolic and log-skew-Laplace models. The latter of these is used to answer an archæological problem on the past environment of a mesolithic site. The use of mixture models is illustrated on a problem concerning the influence of vegetation and sand transportation and on problems of separating types of starch grains in palynological analysis.

Problems of combining, or marrying, measurements made by different techniques on different size ranges of particles are described and it is shown that information on shapes of particles is obtained as a by-product of the analysis; illustrations are on data from cave sediments.

A brief account is given of how modelling of particle sizes measured automatically by an image processor can be used to determine satisfactory threshold values for segmentation of the image.

The final section is concerned with applications in horticulture where data on both weights and numbers of tomatoes in various size classes are available. Here the problems are firstly on the relationship between size and weight and how this is influenced by growing conditions and how it changes during the growing season. This requires the extension of the modelling techniques to incorporate covariate information and this is work currently in progress.

Emphasis will be given to applications but some mathematical details and references are given below.

3. Some mathematical details

Most commonly used practical measurement techniques yield a grouped size distribution, often determined inherently by the instruments (for example sieve sizes or reading times of a hydrometer). Further, in most practical situations it is usually only possible to determine the proportion *by weight* of particles within a size class. In the rare cases where particles are counted within size classes or are measured individually the data can be converted to proportions (whether by weight or by number) in size classes. Estimation of model parameters is then essentially via a multinomial distribution with cell probabilities determined as the integral of the underlying density over the interval corresponding to the size class. If proportions are determined only by weight then the resulting distribution is a *mass-size* distribution rather than a *frequency-size* one.

It is convenient to consider models based on log size, not least because of the wide range of sizes encountered in practice, but also because of the multiplicative process of breakage underlying particle production.

Denote by c_i ; $i = 1, \ldots, k{+}1$ the (natural) logarithms of the class boundaries in increasing order of magnitude. Let $I_i = [c_i,\ c_{i+1})$ be the ith size class. Let w_i be the weight of particles in I_i with $w = \sum_1^k w_i$ the total weight of particles. Let $r_i = w_i/w$.

Let the density function of the logarithm of particle size X be $f(.;\theta)$, where θ is a vector of parameters. Typically, $f(.;\theta)$ will be a truncated version of a density $g(.;\theta)$ defined on the whole real line, so

$$f(x;\theta) = \frac{g(x;\theta)}{\int_{c_1}^{c_{k+1}} g(t;\theta)dt} \qquad (c_1 \le x \le c_{k+1}).$$

To estimate θ from the observed data w_i and c_i define

$$p_i(\theta) = \int_{c_i}^{c_{i+1}} f(t;\theta)dt, \quad q_i(\theta) = \int_{c_i}^{c_{i+1}} g(t;\theta)dt, \quad L(\theta) = \sum_{i=1}^{k} r_i \ln(p_i(\theta)),$$

and maximize $L(\theta)$ with respect to θ. $L(\theta)$ is termed the *likeness function* and is essentially the log likelihood function for θ, up to a multiplier of the unknown sample size and an additive constant, on the basis of a multinomial distribution over the k size classes.

Note that

$$L(\theta) = \sum_{i=1}^{k} r_i \ln\{q_i(\theta)\} - \ln\{T(\theta)\}, \quad \text{where } T(\theta) = \int_{c_1}^{c_{k+1}} g(t;\theta)dt$$

so $L'(\theta) = \sum_{i=1}^{k} r_i q_i'(\theta)/q_i(\theta) - T'(\theta)/T(\theta)$. This latter equation is generally preferable to work with rather than one derived directly from the actual truncated density $f(.;\theta)$.

The three densities most commonly proposed for $g(.;\theta)$ are the normal, hyperbolic and skew-Laplace, the latter two with, respectively, densities:

$$\begin{aligned} g(x;\phi,\gamma,\delta,\mu) &= \frac{\sqrt{\phi\gamma}}{\delta(\phi+\gamma)\mathcal{K}_1(\delta\sqrt{\phi\gamma})} \times \\ &\quad \exp\left\{-[(\phi+\gamma)\sqrt{(\delta^2+(x-\mu)^2)} + (\phi-\gamma)(x-\mu)]/2\right\} \end{aligned}$$

(where $\mathcal{K}_1(.)$ is the modified Bessel function of the third kind and $\phi,\ \gamma,\ \delta > 0$) and

$$g(x;\alpha,\beta,\mu) = \begin{cases} (\alpha+\beta)^{-1}\exp\{(x-\mu)/\alpha\} & x \le \mu \\ (\alpha+\beta)^{-1}\exp\{(\mu-x)/\beta\} & x > \mu \end{cases}$$

(where $\alpha,\ \beta > 0$). Likeness functions derived from all of these densities are continuously differentiable with respect to all parameters since they occur in grouped form. The restrictions on some of the parameters being positive are not strictly necessary when dealing with truncated densities but experience suggests severe computational problems without such restrictions.

Software for fitting all of these models as well as mixture log-skew-Laplace models, with graphical presentations, are available as a Windows application **sheffSize** which can be downloaded from the web site http://www.shef.ac.uk/nickfieller

Acknowledgements: Much of the work referred to here is joint work with many others, notably Dr Eleanor Stillman (née Flenley) and Andrew Lynch (University of Sheffield), Dr Walter Olbricht (Universität Bayreuth). Data have been provided by Professor David Gilbertson (Institute of Earth Sciences, University of Aberystwyth), Dr Robin Torrence (Australian National Museum), Carol Lentfer, (Southern Cross University, NSW, Australia) Dr Helen Atkinson (Department of Materials Engineering, University of Sheffield), Dr John Fenlon (Horticultural Research International, Wellesbourne).

References

The selected references below have not been specifically cited in the text above but provide many of the details which have been omitted.

Barndorff-Nielsen, O. (1977). Exponentially decreasing distributions for the logarithm of particle size. *Proc. R. Soc. A*, **353**, 401–419.

Barndorff-Nielsen, O., Blæsild, P., Jensen, J.L., & Sørensen, M. (1985). The Fascination of Sand. *In* Atkinson, A.C. & Fienberg, S.E. (Eds.) *A Celebration of Statistics: The ISI Centenary Volume*. Springer-Verlag, New York, 57–87.

Fieller, N.R.J. and Flenley, E.C. (*In preparation*), The marrying of particle size data: a new technique for shape analysis.

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

Fieller, N.R.J., Gilbertson, D.D. & Olbricht, W. (1984). A new method for environmental analysis of particle size distribution data from shoreline sediments. *Nature*, **311**, 648–651.

Flenley, E.C., Fieller, N.R.J. and Gilbertson, D.D. (1987). The statistical analysis of 'mixed' grain size distributions from æolian sands in the Libyan Pre-Desert using log skew Laplace models. In Frostick, L. and Reid, I. (Eds). *Desert Sediments: Ancient and Modern.* Geological Society of London, Special Publications, **35**, 271-280.

Folk, R.L. & Ward, W.C. (1957). Brazos river bar: a study in the significance of grain size parameters. *J. sedim. Petrol.*, **27**, 3–26.

Kolmogorov, A.N. (1941). Über das logarithmisch Normale Verteilungsgesetz der Dimensionen der Teilchen bei Zerstuckelung. *Comptes Rendus (Doklady) de L'Academie des Sciences de l'URSS*, **31**, 99–101.

Olbricht, W. (1982). *Modern Statistical Analysis of Ancient Sand.* University of Sheffield. Unpublished M.Sc. Thesis.

Scalon, J.D., (1998). *Spatial and size distributions of particles in composite materials.* Ph.D. thesis, University of Sheffield.