Data, not only in aerobiology: how normal is the normal distribution?

Eckhard Limpert · James Burke · Carmen Galan · Maria del Mar Trigo · Jonathan S. West · Werner A. Stahel

At the 8th International Congress on Aerobiology (8th ICA) an ad hoc “Working Group Data Analysis” formed as a result of the concern about data analysis and the present use of the normal distribution. The Group was asked to consider these points further. Here we present first results and a recommendation to use, in general, other distributions than the normal since data are more adequately described by them.

Frequencies of allergens and micro-organisms in the air as well as their effects on humans and the environment are of major concern. To characterize and summarize the data, several distributions have been used, among which the normal distribution is most popular. In fact, this model has become the standard of quantitative variation across the sciences.

Over the recent past, considerable doubts have been raised on the suitability of this probability density function in aerobiology, and other methods have been applied for that purpose (Moseholm et al. 1987; Comtois 2000; Galan et al. 2000; Trigo et al. 2000; Angosto et al. 2005; Sánchez-Mesa et al. 2002). Nevertheless, quantitative variation of original data quite often continues to be thought to be symmetrical and normally distributed.

Accordingly, data are again and again summarized, using the arithmetic mean and the standard deviation, by $\bar{x} \pm s$ to indicate an interval which is meant to characterize the statistical variation. However, frequently, the mere summary data immediately demonstrate that the normal distribution would not be a good fit, as in the case of 500 ± 500 further explained below. The description gives a wrong impression of the data and their distribution. Moreover, the sign $\pm$ points to additive effects. By contrast, the processes that affect the variability of the number of air spora are usually multiplicative, as are natural laws in general (Limpert et al. 2001; Limpert and Stahel 2006).
Often, standard errors of the mean, SE, are given instead of standard deviations, resulting, for example, in 500 ± 100. With a sample size of \( n = 25 \), this is equivalent to the case considered above (500 ± 500) since the standard deviation is SE times the square root of sample size. Thus, even though 500 ± 100 appears at first glance to be compatible with the normal distribution and symmetry, the use of the standard error in the form \( \bar{x} \pm SE \) hides the problem of asymmetry.

The frequency of such examples in scientific publications provides empirical evidence that data are often skewed. Thus, adequate ways to characterize them would be beneficial and might help to solve “The trouble of threshold values for allergy forecasts” (Jäger 2006). Presumably, it is the high affinity to symmetry, and to the normal distribution as a standard, that gives rise to such cases where the misuse of the normal distribution leads astray.

(3) Mathematically, the normal distribution is based on the arithmetic scale (1-2-3-4-5) and is justified by the additive form of the Central Limit Theorem, which states that many small additive effects combine to create a normal distribution. The amount of air spora, however, is determined by processes that are above all multiplicative. Bacteria and cells in general multiply (please note the special meaning of the word) and cell divisions create the scale 1-2-4-8-16. Pollen follow the same scale during their genesis. These multiplicative processes are in contrast to the arithmetic scale of the normal distribution.

As a simple example, let us assume a mean concentration of air spora of 1,000. Given favorable conditions, one further division at the relevant stage before dispersal would have led to 2,000 particles, while one division less would have led to 500. Symmetry in such cases would thus not be additive, but multiplicative.

Cell division is but one of many multiplicative effects in aerobiology, as described by Gregory (1973) for instance for processes of diffusion and deposition, or the terminal velocity of sedimentation (\( v_s \)) in Stoke’s law or McCubbins approximate formula: 

\[
    v_s = \frac{\text{length} \times \text{width}}{40}.
\]

Reynold’s number \( Re \) is another example defined as 

\[
    Re = \frac{\text{length} \times \text{wind velocity}}{\text{kinematic velocity}}.
\]

A final, more general example in this respect may be Einstein’s 

\[
    E = mc^2.
\]

In the same way as the normal distribution results from summing many small random effects, multiplying many small random effects leads to the log-normal
distribution which will therefore, synonymously, be called multiplicative normal distribution. It is thus justified by the multiplicative form of the Central Limit Theorem (Aitchison and Brown 1957). Due to recent developments, this distribution can now be handled easily (Limpert and Stahel 1998) based on the geometric mean \( \bar{x} \) which equals the median for log-normal data, and on the multiplicative standard deviation \( s^* \).

Both parameters are obtained by back-transforming \( \bar{x} \) and \( s \) determined at the log-scale, the transformation suggested by John Tukey as the “first aid transformation” for such data (Mosteller and Tukey 1977). The data can be characterized (without introduction of errors and loss of validity) analogously to the common form \( \bar{x} \pm s \) for normal distributions (Limpert et al. 2001; Limpert and Stahel 2006, 2008).

The form for the log-normal distribution is \( \bar{x}^* / s^* \) (\( \bar{x}^* \) read times-divide). Assuming, for example, a geometric mean of \( \bar{x}^* = 100 \) and a multiplicative standard deviation \( s^* = 2 \), the 1s* range extends from \( 100 / 2 = 50 \) to \( 100 \times 2 = 200 \) and covers, again, 68% of the variation, while the 2s* range from \( 100 / 2^2 = 25 \) to \( 100 \times 2^2 = 400 \) contains 95%. In contrast to former ways of handling log-normal distributions (cf. “log-normal” at the web), they are now easy to understand and use.

From the above reasoning, the number of colony forming units per m\(^3\) air in poultry houses (Rinsoz et al. 2006), or results from mutagenicity studies (Vijay et al. 2005) that are based on mutagenicity studies, might be worth re-evaluating and characterizing by log-normal distributions, as appears to be true for the mycotoxin content in contaminated buildings (Naire-Koivisto et al. 2006).

Moseholm et al. (1987) estimated a log-normally-shaped evolution with time for the pollen season, which can now be characterized easily at the level of the original data. In this case, of course, a third parameter is needed for the onset of the season, similar to the description of the age at first marriage of women in Denmark in the 1970s (Limpert et al. 2001, based on Preston), where the third parameter is the age of puberty. Many further applications can be imagined, not only in aerobiology and human medicine. This model of parametric statistics should, of course, also be used for regressions and analyses of variance by using those models at the log-scale.

In aerosol science, the measured number distribution of physical properties of particles (e.g., size, surface area or volume/mass) is commonly highly skewed. Thus, in preference to the normal distribution, the data is routinely fitted to a log-normal distribution and corresponding statistics included as standard in instrument software. More than this, for modeling the formation of aerosol particles, the log-normal probability density function is justified from first principles. Specifically, any proportional size-modifying process (e.g., condensational growth or agglomeration) gives rise to a log-normal particle size distribution, similar to fragmentation which has long been shown to fit log-normality (Kolmogoroff 1941).

(4) To conclude, the normal distribution cannot be adequate to characterize the variation of original data and—with the exception of location in space—skewed distributions should be the rule, not only in aerobiology. This may also be the reason why we did not find any data fitting the normal distribution that did not fit equally well or better to the multiplicative normal distribution (Limpert et al. 2001; Limpert and Stahel 2008, and unpublished). This may be surprising at first sight, but slightly skewed distributions cannot be distinguished from corresponding symmetrical distributions by mere statistical assessment. Therefore, we now believe that for quantitative variation of original data, the multiplicative normal distribution generally is the appropriate and unifying model (for the comprehensive concept and further details, see Limpert et al. 2001).

Why then did the normal distribution become so normal and popular? One explanation can be found with Gaddum (1945) citing a remark of H. Poincaré on the normal law: “Everybody firmly believes in it because the mathematicians imagine it is a fact of observation, and observers that it is a theory of mathematics”. Now, bringing theory and observation together and preserving our respect for the normal distribution in general statistics we conclude that, in contrast, skewed distributions of original data are normal and worth considering even if the normal distribution would fit statistically.

There are several models of skewed distributions for a variety of needs. The Power law described by Gregory (1973, p. 248) as “hollow curve”, the Gamma distribution applied to airborne pollen data in Montréal (Comtois 2000), and the exponential and Weibull distributions appear to be most established in addition to the log-normal, and careful consideration
is recommended to find the most appropriate one for a particular purpose and dataset. For example, the atmospheric transparency of the city of Bratislava was tested against the Power law and the Gamma and log-normal distributions (Kocifaj et al. 2001).

In the end, it is our major aim to stimulate a fundamental discussion on the normal distribution and the origin of variability, which is overdue across the sciences, and which will lead to a higher level of data analysis that is more informative, challenging, and worth the effort.

Acknowledgements We appreciate the considerable interest found for this topic and are grateful for discussions with Siegfried Jäger, Frits Spieksma, Paul Comtois and Branko Sikoparija, as well as for continuous interest and support from Roy Snaydon and Ruth Maria Leuschner.

References


