THE LOG-NORMAL DISTRIBUTION

J. H. Bigelow

September 1975 WP-75-120

The Log - Normal Distribution

J. H. Bigelow

Sept. 30, 1975

Introduction

The purpose of this brief note is to share some very simple observations concerning the log-normal distribution. I have found this distribution useful chiefly for two reasons. First, it describes a non-negative random variable, such as the time between two events, or the concentration of a pollutant. Second, the log-normal distribution can have any coefficient of variation between zero and infinity. This is not true of other positive distributions, such as the gamma and beta distributions.

Definition

A random variable 'x' is log-normally distributed if 'log x' is normally distributed. Thus 'x' has a cumulative distribution:

$$F(x) = \Phi\left(\frac{\log x - m}{s}\right) \tag{1}$$

The function $\Phi(\cdot)$ is the cumulative normal function that one finds tabulated in every statistical text and handbook. By inspection, therefore, one can see that 'log x' has mean 'm' and standard deviation 's'.

One can obtain the density function of 'x' by differentiating equation (1). After some manipulations, this yields:

$$f(x) = \frac{\exp\left(-m+\frac{s^2}{2}\right)}{s\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\log x - (m-s^2)}{s}\right)^2\right] \tag{2}$$

Moments

The moments about zero of the log-normal distribution can be calculated from the density function using the standard formula:

$$E(x^n) = \int_0^\infty x^n f(x)\, Dx$$

Carrying out this integration is most easily accomplished if one changes variables. Thus let

$$t = \log x$$

so that

$$e^t dt = dx$$

Substituting, one finds that $x^n dx = \exp((n+1)t)\, dt$, and therefore,

$$E(x^n) = \frac{\exp(-m+\frac{s^2}{2})}{s\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(\frac{t - (m-s^2)}{s}\right)^2 + (n+1)t\right] dt$$

To evaluate this integral, one first computes the term that must be subtracted from this argument to make it a perfect square times minus one-half. Then the integrand can be expressed as the exponential of this term times a normal density function. Finally, take advantage of the fact that a normal density function integrates to one to obtain the final result. This is:

$$(3) \qquad E(x^n) = \exp\left(n \cdot m + \frac{n^2 s^2}{2}\right)$$

In particular, and taking advantage of the fact that the variance $V(x) = E(x^2) - E^2(x)$,

$$(4) \qquad \mu = E(x) = \exp\left(m + \frac{s^2}{2}\right)$$

$$(5) \qquad \sigma^2 = V(x) = \exp(2m + s^2)(\exp(s^2)-1)$$

$$(6) \qquad CV = \frac{\sigma}{\mu} = \sqrt{\exp(s^2) - 1}$$

CV is the coefficient of variation. Note that since 's' can be any non-negative number, CV can take on any value between zero and infinity.

Fitting the Distribution to a Sample

Given a sample $(x_1, x_2, .., x_n)$ which has been drawn from a log-normal distribution, one may wish to estimate the parameters 'm' and 's'. The most straightforward way to do this would be to take the logarithm of each sample point, and estimate:

$$(7) \quad \begin{cases} m = \frac{1}{n} \sum_i \log x_i \\ s^2 = \frac{1}{n-1} \sum_i (\log x_i - m)^2 \end{cases}$$

However, for large samples, such as one may encounter in studies of air pollutant concentrations (see [1]), taking the logarithm of each sample can be costly. Further, it is often the case that the device that measures the samples will record zero when the value drops below some critical limit. In this case, using equations (7) is not merely costly, but senseless as well.

A cheaper method for estimating 'm' and 's', that also avoids the problem of zero values, is to take advantage of equation (3). One calculates the first and second sample moments, and equates

them with $E(x)$ and $E(x^2)$ as shown in (3). Taking logarithms of the result yields:

$$(8) \quad \begin{cases} m + \frac{s^2}{2} = \log\left(\frac{\sum x_i}{n}\right) \\ 2m + 2s^2 = \log\left(\frac{\sum x_i^2}{n}\right) \end{cases}$$

Equations (8) are linear in the variables 'm' and 's^2', which can accordingly be calculated easily.

The Truncated Distribution

When I use a log-normal distribution for simulation purposes, I use a truncated version of the distribution, one which extends from zero up to some maximum value, let us say ξ. Of course, the moments of the truncated distribution will differ, sometimes considerably, from the moments of the original distribution. It is important to realize how large this difference may be.

The moments of the truncated distribution are calculated in the same way as the moments of the original distribution. Thus,

$$\hat{E}(x^n|\xi) = \int_0^{\xi} x^n f(x)\,dx$$

Carrying out the usual manipulations yields:

$$(9) \quad \hat{E}(x^n|\xi) = \exp\left(n\cdot m + \frac{n^2 s^2}{2}\right) \cdot \Phi\left(\frac{\log\xi - (m + ns^2)}{s}\right)$$

For n = o, this reduces (as it should) to the cumulative function described by equation (1).

The impact of equation (7) is best seen by comparing it to equation (3). In general one has that

$$(10) \qquad \frac{\hat{E}(x^n \mid \xi)}{E(x^n)} = \Phi\left(\frac{\log\xi - (m+ns^2)}{s}\right)$$

This implies that the relative shortfall in the $n^{\underline{th}}$ moment of the truncated distribution increases with n. Thus one must choose the highest moment whose error one wishes to control, and define ξ to be large enough to make that error tolerable. Then the errors in lower moments will be well within the stated tolerance.

Reference

1. Larsen, R.I., "A Mathematical Model for Relating Air Quality Measurements to Air Quality Standards", U.S. Environmental Protection Agency, Pub. no. AP-89 (1971).