Journal of Exposure Science and Environmental Epidemiology (2007) 17, 499–500

www.nature.com/jes

Letter to the Editor

The correct application of Mage's Johnson S_B procedure for fitting exposure data

DAVID T. MAGE

USEPA (retired) and WHO (retired), Newark, Delaware, USA
E-mail: magedonner@aol.com

Journal of Exposure Science and Environmental Epidemiology (2007) **17,** 499–500. doi:10.1038/sj.jes.7500592

Flynn (2006) made an excellent review of different methods of applying the Johnson S_B distribution, also known as the four-parameter lognormal distribution, to exposure data. He applied the Mage (1980a) percentile solution procedure for estimating S_B distribution parameters (Johnson, 1949) and compared it with four other methods for fitting the S_B model. However, he concluded that Mage's percentile procedure did not always return values for the four S_B parameters, and that "the quantile and method-of-moments fitting procedures *may* provide performance superior to that of the percentile method."

For $y=(x–X_{min})/(X_{max}–x)$, where x is an exposure variable bounded between a minimum value (X_{min}) and a maximum value (X_{max}), the S_B probability density function $p(x)$ is written as Eq. (1), where μ and σ are the mean and SD of $\log(y)$, respectively:

$$p(x) = \sqrt{(2\pi\sigma^2)^{-1}}[(X_{max}-x)^{-1} + (x-X_{min})^{-1}]\exp(-0.5[(\log(y)-\mu)/\sigma]^2); \quad (1)$$
$$X_{min} < x < X_{max}$$

The classical maximum likelihood method for the Johnson S_B distribution often fails because the log likelihood function approaches infinity as X_{min} is attracted up to the minimum observation or X_{max} is attracted down to the maximum observation (Hill, 1963). Cheng and Traylor (1995) developed modified likelihood functions for several special cases that might be applicable here. Thus, Flynn's default condition setting $X_{min}=0$, to apply "conditional" maximum likelihood estimation, might not be appropriate.

Given that an exposure data set consists of N-independent random samples from the same unknown stationary distribution, one can then plot these data, ranked from lowest ($i=1$) to highest ($i=N$), as $\log(x)$ *versus* a probability (p) scale of standard normal deviates (z) with plotting position z_i corresponding to $p=i/(N+1)$ (Taylor, 1990). In this coordinate system, a Johnson S_B plot must be concave downward at high concentrations as x goes to $X_{max}<\infty$, as the integral of $p(x)\,dx$, from X_{min} to X_{max}, goes to 1. At low concentrations, the curve may be either concave upward ($p(x)$ goes to 0 as x goes to $X_{min}>0$), or concave downward if x goes to 0 at a finite value of z. This latter concave downward condition implies a finite probability that true 0 values are possible (Ott and Mage, 1976).

The data analyst then applies such a spline curve to these plotted data with, at most, only one point of inflection. This curve does not need to go through any datum point, but should seek to minimize the maximum difference in probability between the predicted frequency and the observed frequency $i/(N+1)$ for all x_i (Mage, 1982), which minimizes the Kolmogorov–Smirnov statistic (Mage, 1980b). Then, one can take any four points from this spline curve that are equidistant in z (e.g., $z_4–z_3=z_3–z_2=z_2–z_1$) and compute the set of four parameters of the Johnson S_B distribution that by definition must pass through these very same four points.

A reason why Flynn may have reached his incorrect conclusion is that in some cases, his choice of four paired (x_i, z_i) values was not in the S_B region, but actually in the Johnson S_U (unbounded) region with $mn/p^2>1$, defined by his Eq. (9), where $m=x_4–x_3$, $n=x_2–x_1$, and $p=x_3–x_2$. By chance, 5 of his 20 sets of random samples from an S_B distribution must have been interpolated to four values ($x_1<x_2<x_3<x_4$, with $mn/p^2>1$) at $z=\pm0.5384$ and $z=\pm1.6152$, which caused both his Eqs. (6) and (8) to return square roots of negative numbers. Had he merely chosen a different set of four paired (x_i, z_i) values from a curve meeting all the above conditions, both the Mage (1980a) and the equivalent Slifker and Shapiro (1980) percentile method would have returned four identical S_B parameters without any imaginary values involving $\sqrt{(-1)}$.

Finally, Flynn's choice of data may have violated the primal requirement that any data set being fit should be an independent random sample from a stationary distribution.

The benzene–air exposure data of Tolentino *et al*. (2003) were from 11 different workers who were sampled from 1 to 4 times each for a total of $N = 33$ samples. One sample at 1.4 p.p.m. tested as an outlier and was removed by them from consideration. They recognized that workers' day-to-day exposures may be autocorrelated, which violates the independence assumption, so they analyzed the 10 mean exposure values of the 10 workers who had more than one sample, so that a reciprocal variance could be calculated to use for the sample weight for each worker. However, Flynn analyzed all 33 observations together as if they were all independent random samples from the same distribution, when perhaps the 10 mean exposure values, properly weighted, would have been more appropriate for analysis.

An outlier will often occur when one inadvertently samples from two or more different distributions and combines the samples for analysis. This situation occurs if members of the general public are sampled for exposure, when one or more of them happen to be employed where the target chemical is used, and these subjects register a very high value. Such high values, if not recognized, can cause the higher concentration region of a probability plot to have the typical S_U curvature concave upwards, even though the rest of the sample was from an S_B distribution.

An example of a sample from a complex mixture of distributions is the Jacobs and Smith (1988) CO_2 data modeled by Flynn, which caused "no fit" by his conditional maximum likelihood procedure (which assumed a global background CO_2 of 0 p.p.m. instead of the observed 1987 value of 350 p.p.m. (Gore, 2006)). The carbon dioxide occupational exposure data ($N = 13$) in this case were from proximity to "dry ice", where workers were sampled at three different poultry processing plants. Furthermore, at these plants, consistently lower measurements of CO_2 *outside* the holding coolers at the palletizing line were combined with consistently higher measurements made *inside* the holding coolers (Table 2 of Jacobs and Smith, 1988), which is a violation of the requirement that all analyzed data should be generated from a single distribution.

In conclusion, the Mage (1980a) percentile procedure can always return real S_B parameters if applied correctly, which could not be said of the quantile, moments, and conditional maximum-likelihood methods. Thus, the reader is encouraged to reconsider the Mage (1980a) percentile method, described above, as a viable alternative for estimating the parameters of the Johnson S_B distribution.

References

Cheng R.C.H., and Traylor L. Non-regular maximum likelihood problems. *J. R. Stat. Soc. B* 1995: 57: 3–44.

Flynn M.R. Fitting human exposure data with the Johnson S_B distribution. *J. Expos. Sci. Environ. Epidemiol.* 2006: 16: 56–62, and 385 (corrigenda).

Gore A. *An Inconvenient Truth* 2006, Rodale, Emmaus, PA.

Hill B.M. The three-parameter lognormal distribution and Bayesian analysis of a point source epidemic. *J. Am. Stat. Assoc.* 1963: 58: 72–84.

Jacobs D.S., and Smith M.S. Exposures to carbon dioxide in the poultry processing industry. *Am. Ind. Hyg. Assoc. J.* 1988: 49: 624–629.

Johnson N.L. Systems of frequency curves generated by methods of translation. *Biometrika* 1949: 36: 148–176.

Mage D.T. An explicit solution for S_B parameters using four percentile points. *Technometrics* 1980a: 22: 247–251.

Mage D.T. An empirical model for the Kolmogorov–Smirnov statistic. *J. Environ. Sci. Health* 1980b: A15: 139–147.

Mage D.T. An objective graphical method for testing normal distributional assumptions using probability plots. *Am. Statist.* 1982: 36: 116–120.

Ott W.R., and Mage D.T. A general purpose univariate probability model for environmental data. *Comput. Oper. Res.* 1976: 3: 209–216.

Slifker J.F., and Shapiro S.S. The Johnson system: selection and parameter estimates. *Technometrics* 1980: 22: 239–246.

Taylor J.K. *Statistical Techniques for Data Analysis* 1990, Lewis Publishers, Chelsea, MI.

Tolentino D., Zenari E., Dall'Olio M., Ruani G., Gelormini A., and Mirone G. Applications of statistical models to estimate the correlation between urinary benzene as a biomarker of exposure and air concentrations determined by personal monitoring. *Am. Ind. Hyg. Assoc. J.* 2003: 64: 625–629.