Development of Population-Variability Distribution Using Maximum Likelihood Method

(Using Interlock Failure on Demand data as a Sample)

(T Cao, 6/6/2014, revised on 6/16/2014)

Following the approach of "The Combined Use of Data and Expert Estimates in Population-Variability Analysis" (Lopez Droguett, 2004), the population-variability distribution (PVD) in this case is chosen to be lognormal $g(x,\nu,\tau)$, where x is the reliability parameter for the component (failure rate or failure probability), and ν and τ, the two unknowns to be determined, are respectively the mean and standard deviation of the normal distribution associated with the lognormal.

The likelihood functions, unconditional on x, for each of the data sources are calculated as follows:

$$L_i(\nu,\tau) = \int_{x1}^{x2} f(x,\nu_i,\tau_i)g(x,\nu,\tau)dx \qquad (1)$$

In the above formula, function $f(x,\nu_i,\tau_i)$ is the likelihood function for data source i. It is also a lognormal distribution function (pdf) in this sample case and ν_i and τ_i are the two known mean and standard deviation of the normal distribution associated with the lognormal for source i. The integration limits $x1$ and $x2$ defines the failure rate range considered (10^{-8}/h to 10^{-2}/h, for example).

The maximum likelihood method is used to calculate ν and τ. The log-likelihood function to be maximized is:

$$L(\nu,\tau) = \sum_{1}^{n} \ln(L_i(\nu,\tau)) \qquad (2)$$

where n is the number of sources.

The key to a successful calculation is that ν and τ cannot be allowed to vary all over the place at the same time because this will allow the PVD function going higher and higher without limit and lead to an infinite L. To avoid this trap, we can preselect a τ value to be large enough so that the PVD will likely to encompass all the likelihood functions for all data sources. With this preselected τ value the solution of ν can be obtain first. Our calculation shows that the solution of ν is independent on the preselected τ value as expected.

There is another issue of following the above formulation. The BSC's (2008ac, attachment H) calculation used the five data sources of interlock failure on demand, which are the five failure rates (1.0e-5, 7.43e-5, 2.75e-5, 1.0e-4, and 2.4e-5) and a common error factor of 5.

They showed that the above maximum likelihood method returns a rate of 8.21e-6 ($\nu = -11.71$), which is even lower than the lowest data point (1.0e-5). BSC (2008ac) had to select input value 2.75e-5 as a more representative value.

This problem is due to the uneven weights applied to each likelihood function in (2), which is inherent in the above formulation. Let us look at the likelihood functions with different median values. It is easy to derive the following expression for the peak value of a lognormal density function:

$$Y_{\max} = \frac{1}{m\sigma\sqrt{2\pi}} \exp(\sigma^2/2) \qquad (3)$$

This relation shows that the peak value is inversely proportional to the median value m. Figure 1 compares two lognormal density functions with the same sigma (=0.6) but different median value m. We see such an inverse relation. So when we do the integration following (1), which multiplying the density function with a common PVD function, the density function with a smaller m will have higher weight than the density function with a higher m value.

We can easily change the above formulation to include a weight correction according to the above inverse relation (3) between the peak height $Y_{\max}$ and m:

$$L(\nu,\tau) = \sum_{1}^{n} m_i * \ln(L_i(\nu,\tau)) \qquad (4)$$

Now all the likelihood function from each data point will have the same weight. Figs. 2 and 3 compare the density functions of five sources without and with the weight corrections respectively. In Fig.2 we see that the density function of m=1.0e-5 (the lowest) is the highest and will dominate the rate of PVD. This explains why BSC (2008ac) got a very low rate (8.21e-6) for PVD. Fig. 3 shows the five data source density functions with weight corrections. We see now all the five density functions have the same height or same weight. The new calculation following equations (1) and (4) produces a ν value of -10.67 or a failure rate of 2.3e-5, which is representative of the data set.

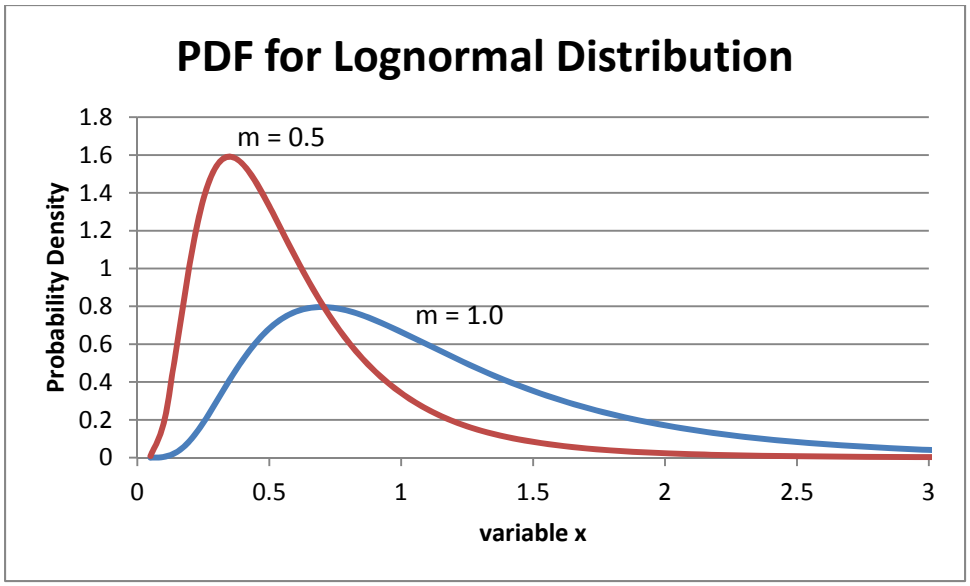

Fig. 1 Two lognormal density distribution functions with same sigma but different median values.

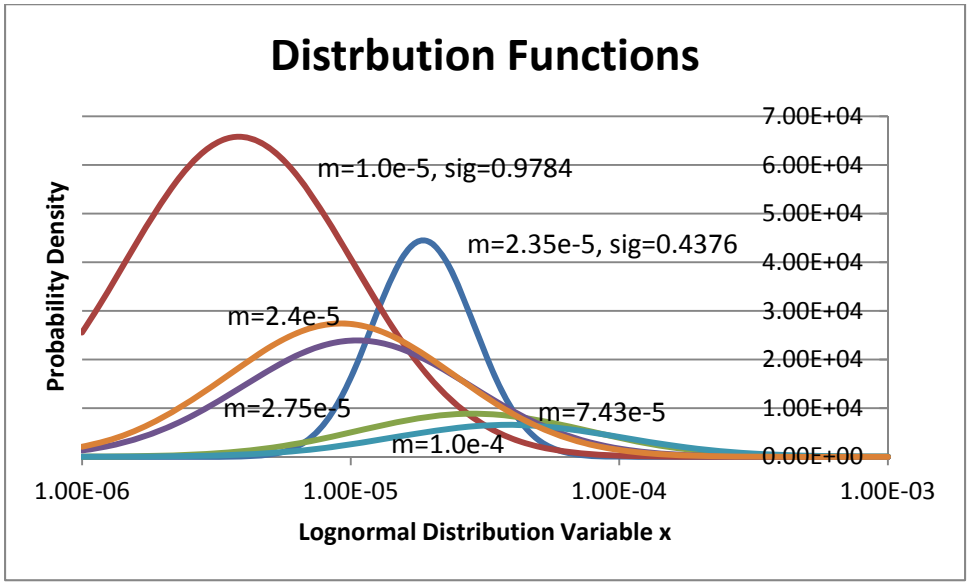

Fig. 2. The lognormal density distribution functions for the five data sources without weight correction. The population-variability distribution (PVD) is the result of using five weight corrected density functions as shown in Fig. 3.

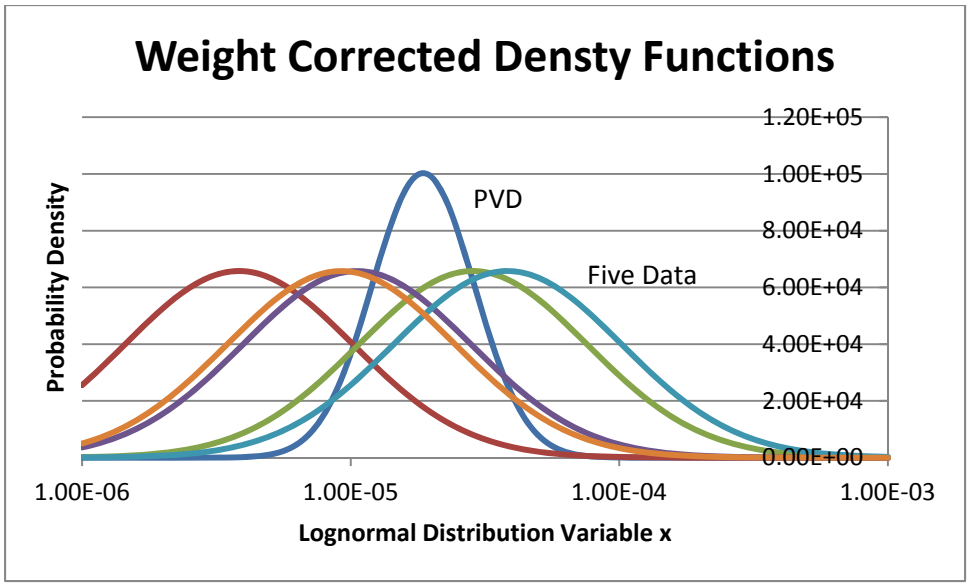

Fig. 3. The weight corrected probability density distribution functions and the PVD distribution.

The attached programs were written in Matlab. Prgram day7.m follows equations (1) and (2), which produces the same result like BSC (2008 ac); program day8.m follows equations (1) and (4), which produces a much representative result. The input data are from Attachment H of the BSC (2008) report. The five failure rates are 1.0e-5, 7.43e-5, 2.75e-5, 1.0e-4, and 2.4e-5. The error factor is 5 for all five rates. The two output files (out7.txt and out8.txt) contain the calculated ν value from the two approaches one without weight correction and the other with the weight correction.

References

Lopez Droguett, E.; Groen, F.; and Mosleh, A. 2004. "The Combined Use of Data and Expert Estimates in Population Variability Analysis." *Reliability Engineering and System Safety, 83*, 311–321. New York, New York. Elsevier. TIC: 259380.

BSC. 2008ac. "Canister Receipt and Closure Facility Reliability and Event Sequence Categorization Analysis." 060–PSA–CR00–00200–000. Rev. 00A (ML090770264). CACN 001 (ML090770367). Rev. 00B. Las Vegas, Nevada: Bechtel SAIC Company, LLC.