

# Maximum Likelihood Estimates for Alpha and Beta With Zero SAIDI Days

Richard D. Christie  
Department of Electrical Engineering Box 352500  
University of Washington  
Seattle, WA 98195-2500  
christie@ee.washington.edu  
February 10, 2003

## 1. Introduction

[Chr03] explains how substituting the minimum SAIDI value for zero SAIDI days gave the most accurate - or least erroneous - results for Alpha and Beta compared with ignoring zero SAIDI days, or using the median or average SAIDI value as a replacement. [Chr03] also stated that other statistical methods may be available.

This document describes the statistically based maximum likelihood (MLE) method of estimating the values of Alpha and Beta in data sets with zero SAIDI days. Two quantitative examples show that the MLE method is more accurate than minimum value substitution, which in turn is the most accurate of the proposed substitution methods. The MLE method involves the iterative solution of a non-linear equation. This is doable interactively with a spreadsheet in a short time (a few minutes). The Working Group must determine whether the complexity of the method permits its adoption in P1366.

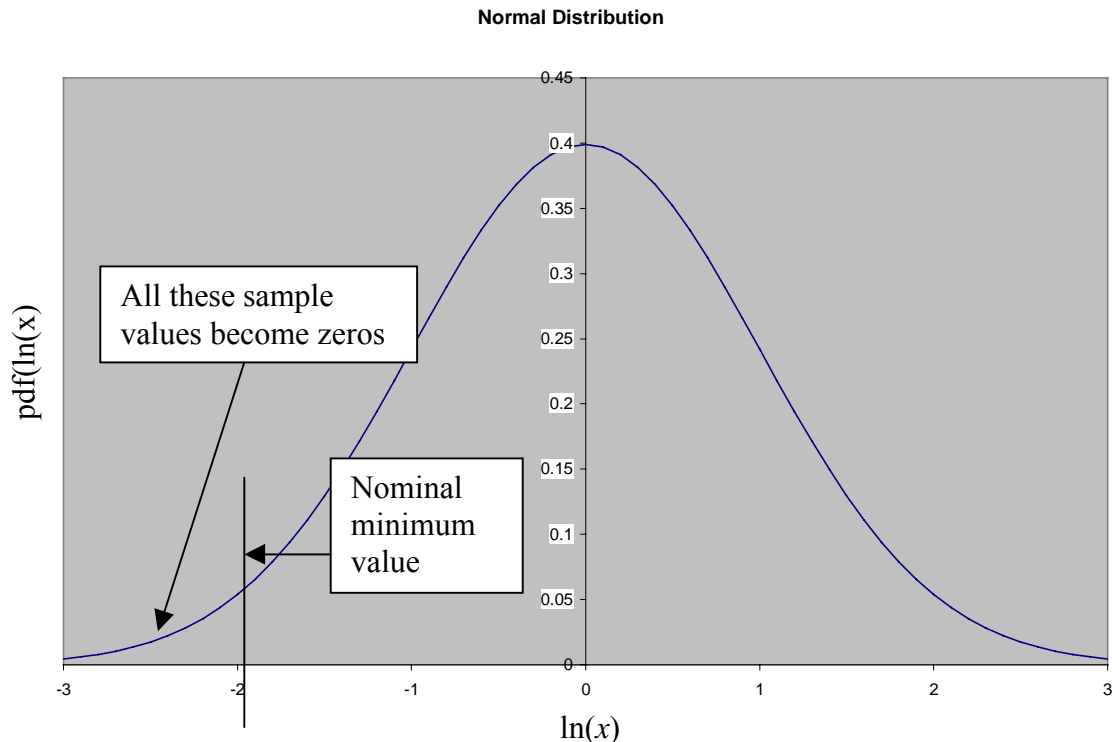


Figure 1 - Normal distribution showing location of zero samples in left hand tail

## II. Background

In theory, an ideal log-normal distribution will never generate samples (daily SAIDI values) with values of zero. In practice, zero values appear because the real process is not exactly log normal - it has some discrete components (faults may or may not occur) as well as continuous - and because of the quantization of time (SAIDI per **day**). It may be useful to think of the sampling process as going through a round-off process in which daily SAIDI values below some minimum are rounded to zero.

These theoretical pre-roundoff sample values (there is no way to measure their actual values) are all less than some minimum but greater than zero and can be thought of as occupying the left hand tail of the normal distribution of the logs of the samples. This is shown in Figure 1. The value of the minimum shown is somewhat large to emphasize the content of the tail of the distribution.

The question, then, is what to do about these zero values, which cannot be used to find the mean and standard deviation of the logs of the data, (Alpha and Beta, respectively) because the logs of the zero sample values are negative infinity.

The objective is to *estimate* values of Alpha and Beta. The actual values of Alpha and Beta are properties of the population of the values of SAIDI for all possible days (i.e. an infinite number of values) for a given utility, and are not knowable. The days are sampled - nominally five years worth of samples - and computations are performed on this sample to estimate values for Alpha and Beta. The computations are called *estimators*. The commonly used and generally preferred estimators are called *maximum likelihood estimators* (MLEs) because the values they estimate have the highest chance of having the least error.

As it happens, for normal (Gaussian) distributions, the maximum likelihood estimator for the population mean is the average of the sample, and the maximum likelihood estimator for the variance is the square of the standard deviation of the sample, and this is the method used to estimate Alpha and Beta from the natural logs of the daily SAIDI values when none of them are zero.

When some daily SAIDI values *are* zero, the problem becomes estimating Alpha and Beta from a *censored* sample, one that is missing values below a certain point. The sample is *singly* censored, because sample values are missing from only one side of the distribution. The maximum likelihood estimators for mean and standard deviation of censored normal distributions are given in [Sch86], found via [Cro88].

## III. Are Utility Reliability Distributions Really Log-Normal?

At this point it may be useful to revisit the issue of whether utility reliability distributions are log-normal, since some Working Group members have claimed they are not and

provided graphical examples where the natural logs of the daily SAIDI values are, for example, somewhat bimodal. The quick answer is that utility daily reliability distributions are not exactly log-normal, but log-normal is close enough to what they really are for all practice purposes.

Just as it is not possible to know the actual values of the mean and variance of the population of all possible daily reliability values, it is also not possible to formally state whether utility daily reliability values are or are not log-normally distributed. When someone makes such a statement, they are speaking informally. It is possible to make a statement about how close a given distribution is to log normal, i.e. "... is log-normal with  $p = \dots$ ". The process that generates daily reliability distributions is sufficiently complex, involving as it does seasonal weather patterns, animal migrations, several independent discrete event processes (fault causes) and continuously distributed response times that include a travel component, that it seems unlikely to be provably log-normal.

What can be said is that all of the utility daily reliability distributions analyzed to date have fit the log-normal distribution better than several other likely distributions, including normal (Gaussian) and Weibull. The common sense test for this is to look at the histogram of the natural logs of the daily reliability data and see if it looks more like a normal (Gaussian) distribution than any other distribution. Even the bimodal distribution offered as evidence of non-log normality can be seen to be Gaussian with some systematic error. If log-normal is the closest distribution to what is actually observed, then methods based on the log-normal distribution can be used that generate results with the least error, even though that error is not zero. This is the case for all of the historical utility data reviewed at present, and this is the basis for assuming log-normality for the rest of the discussion in this paper.

#### IV. Maximum Likelihood Estimators

In [Sch86], Schneider describes maximum likelihood estimators for singly censored normal distributions, reporting they were developed by Cohen in 1950. The presentation in [Sch86] is for right-censored samples while the zero SAIDI day case has left-censored samples, that is, low values are missing instead of high ones. Therefore the equations given here are modified for left censoring. Schneider's notation has also been modified to use Alpha ( $\alpha$ ) for the mean and Beta ( $\beta$ ) for the standard deviation being estimated.

Symbols used are as follows:

- $\alpha$  - Mean of the natural log of daily reliability.
- $\hat{\alpha}$  - Estimate of the mean of the natural log of daily reliability. The Alpha value used to compute the major event day threshold value  $T_{MED}$ .
- $\beta$  - Standard deviation of the natural log of daily reliability.
- $\hat{\beta}$  - Estimate of the standard deviation of the natural log of daily reliability. The Beta value used to compute the major event day threshold value  $T_{MED}$ .
- $\phi$  - Probability density function (pdf) of the standard normal distribution.
- $\Phi$  - Cumulative density function (cdf) of the standard normal distribution.

$h$  - The amount of probability in the censored data.  
 $n$  - The total number of daily SAIDI values  $r_i$ , including zero values.  
 $n_z$  - The number of zero daily SAIDI values  $r_i$ .  
 $r_i$  - The value of SAIDI on day  $i$ .  
 $s^2$  - Sample variance, square of standard deviation of the natural logs of the non-zero daily SAIDI values  $r_i$ .  
 $T_{MED}$  - The major event day threshold value.  
 $u$  - The estimated normalized value of the natural log of the smallest possible non-zero daily SAIDI value.  
 $\bar{x}$  - Average value of the natural logs of the non-zero daily SAIDI values  $r_i$ .  
 $x_{\min}$  - Natural log of the smallest non-zero daily SAIDI value  $r_i$ .

The maximum likelihood estimators are

$$\hat{\alpha} = \bar{x} + \lambda(h, u)(x_{\min} - \bar{x}) \quad (1)$$

$$\hat{\beta} = \sqrt{s^2 + \lambda(h, u)(x_{\min} - \bar{x})^2} \quad (2)$$

where

$$h = \frac{n_z}{n} \quad (3)$$

$$\lambda(h, u) = \frac{Y(h, u)}{Y(h, u) + u} \quad (4)$$

$$Y(h, u) = \frac{h}{1 - h} \tilde{W}(u) \quad (5)$$

$$\tilde{W}(u) = \frac{\phi(u)}{1 - \Phi(u)} \quad (6)$$

and  $u$  is the solution to

$$\frac{1 - Y(h, u)[Y(h, u) + u]}{[Y(h, u) + u]^2} = \frac{s^2}{(x_{\min} - \bar{x})^2} \quad (7)$$

Equation (7) is a non-linear equation that is solved by iteration.

## V. Estimation Process

The maximum likelihood estimators  $\hat{\alpha}$  and  $\hat{\beta}$  can be computed from a set of daily SAIDI values as follows:

1. Sort the sample by value.
2. Count the number of zero values,  $n_z$ .
3. Take the natural log (ln) of all non-zero SAIDI values.

4. Find the average ( $\bar{x}$ ) and standard deviation ( $s$ ) of the values computed in step 3.
5. If there are no zero SAIDI values ( $n_z = 0$ ), then Alpha =  $\bar{x}$  and Beta =  $s$ . Otherwise,
6. Compute  $h = \frac{n_z}{n}$
7. Find  $x_{min}$ , the natural log of the minimum non-zero daily SAIDI value,  $\min(r_i)$ .
8. Solve equation (7) for  $u$ . See the discussion below.
9. Find Alpha and Beta from equations (1) and (2).

Once Alpha and Beta are known, TMED can be computed as usual using the estimates.

$$T_{MED} = \hat{\alpha} + 2.5\hat{\beta} \quad (8)$$

## VI. Solving Equation (7) for $u$

A number of algorithms are available for solving non-linear equations such as equation (7). These could be automated in a spreadsheet macro or programming in to an analysis program. The following spreadsheet-based heuristic interactive iterative process is practical and convenient for spreadsheets that have functions giving the standard normal distribution probability density function (pdf) and cumulative density function (cdf) ( $\phi$  and  $\Phi$ , respectively).

One popular spreadsheet, Excel™, implements these functions as follows:

$$\begin{aligned} \phi(x) & \text{ NORMDIST}(x,0,1,FALSE) \\ \Phi(x) & \text{ NORMDIST}(x,0,1,TRUE) \end{aligned}$$

Using these, after computing the necessary constants like  $h$ ,  $\bar{x}$  and  $s$ , the iterative process can be performed as follows.

1. Select a column in which guesses for  $u$  will be entered.
2. In the next column to the right enter the formula for  $\tilde{W}$  from equation (6) as  $=\text{NORMDIST}(u, 0, 1, FALSE) / (1 - \text{NORMDIST}(u, 0, 1, TRUE))$  where  $u$  is the column selected in step 1.
3. In the next column to the right enter the formula for  $Y$  from equation (5).
4. In the next column to the right enter the formula for the left hand side (LHS) of equation (7)
5. In the next column to the right enter the formula (or copy the value of) the right hand side (RHS) of equation (7). Note that  $\bar{x}$  is the average, and  $s$  is the standard deviation of the natural logs of the non-zero daily SAIDI values.
6. Copy the row several times. Each row will be one iteration. Copy as many times as needed. Alternatively, reenter new values of  $u$  in the same cell.
7. Enter an initial guess for  $u$ . 1.0 is a reasonable value if no other information (like a previous result) is available.
8. Based on the mismatch between the LHS value computed by the spreadsheet for the most recent guess and the constant RHS value, make another guess at  $u$ . In general an increase in  $u$  results in a decrease in the LHS value. The amount of change to make is based on judgement. (Interpolation could be used, but heuristic guessing is faster than

computing the interpolation unless repeating this analysis many times, in which case a macro is recommended.) Repeat until a sufficiently accurate guess is obtained.

## VII. Examples

Alpha and Beta were estimated for two example censored data sets using five different methods.

One data set has five years of simulated daily SAIDI values. The advantage of simulation is that the actual values of Alpha and Beta are known. Daily SAIDI values were found by obtaining a uniform random variable between 0 and 1, finding the value of the normal CDF for the random value, and then exponentiating. This gives an almost ideal log-normal distribution.

The second data set is four years of real world daily SAIDI data for anonymous Utility 2 provided by the Distribution Design Working Group.

Neither data set has zero SAIDI days. Both are censored by assuming that the 110 lowest SAIDI values have been rounded to zero, so that their natural logs are not available. This permits comparison of Alpha and Beta estimates from the uncensored data set with estimates calculated using the censored data set.

The five methods of estimating Alpha and Beta are:

- Ignoring zero SAIDI days.
- Replacing zero SAIDI days with the minimum non-zero SAIDI value.
- Replacing zero SAIDI days with the median SAIDI value.
- Replacing zero SAIDI days with the average SAIDI value.
- Maximum likelihood estimators for censored samples (MLE)

The results for the simulated data set given in Table 1.

Table 1 - Results for Simulated Data Set with 110 Censored Values

Parameter	Actual Values	No Censored Values	Ignore Zero Days	Replace with Minimum	Replace with Average	Replace with Median	Maximum Likelihood Estimates
Alpha	-3.60	-3.59	-3.33	-3.54	-3.23	-3.35	-3.59
Beta	2.00	2.00	1.76	1.89	1.76	1.71	1.99
ln(TMED)	1.40	1.41	1.07	1.18	1.16	0.93	1.39
TMED	4.06	4.08	2.93	3.26	3.20	2.53	4.02

The values of Alpha and Beta estimated by taking the average and standard deviation of the complete set of simulated data (No Censored Values column) are very close to the actual values used to generate the data set. As discussed qualitatively in [Chr03], the values of Alpha and Beta estimated by replacing the zero SAIDI days with the minimum non-zero SAIDI value are closer to the actual values than those found by ignoring zero SAIDI days or replacing with the average or Median SAIDI value. However, the natural

$\log$  of  $T_{MED}$  is significantly lower than the actual value, which would result in classifying more major event days than would be correct.

The Maximum Likelihood Estimates (MLEs) are significantly more accurate than any of the replacement schemes and have about as much error as the uncensored value estimates. The MLE is clearly the preferable estimation technique.

Results for the Utility 2 data set are given in Table 2. In this case, the actual values of the parameters are not known, and the estimates from the censored data must be compared with the estimates (average and standard deviation) from the uncensored data.

Table 2 - Results for the Utility 2 Data Set with 110 Censored Values

Parameter	No Censored Values	Ignore Zero Days	Replace with Minimum	Replace with Average	Replace with Median	Maximum Likelihood Estimates
Alpha	-3.53	-3.19	-3.42	-3.08	-3.20	-3.49
Beta	2.03	1.68	1.81	1.66	1.61	1.94
$\ln(TMED)$	1.55	1.00	1.11	1.06	0.83	1.36
$TMED$	4.71	2.73	3.04	2.89	2.30	3.90

Using real utility data, the qualitative results are the same as before, i.e. MLEs are closer to the non-censored estimates than any of the replacement methods, and replacement with the minimum value is the best replacement method. However, errors are larger. This is probably because the real world data is only close to being log-normally distributed, not exactly log-normally distributed as is the case for the simulated data set. As explained in section III, assuming log-normality permits computation of the MLEs, and the error associated with differences from log-normality is small if the distributions are close to being log normal.

## VIII. Other Estimators

[Sch86] describes a number of other estimators for the mean and standard deviation of censored samples from normally distributed populations. Some of these estimators are described as "simplified" because they do not require an iterative solution of equation (7). However, the closed form solutions provided are more complicated to explain and implement than the use of MLEs. Many involve table look ups, where the table values are more complicated to compute than the solution of (7). Furthermore, all of the additional estimators have lower efficiency than the MLEs, meaning that they produce a wider range of estimates, even if the average estimate is accurate. For these reasons the other estimators in [Sch86] are evaluated as unsuitable for the major event day identification problem.

## IX. Conclusion

The Maximum Likelihood Estimators (MLEs) for censored samples, found in [Sch86], give better estimates of Alpha, Beta and  $T_{MED}$  for data sets with zero SAIDI days than replacement methods, a result backed by theory and illustrated by the examples in section

VII. The MLEs can be computed using a standard spreadsheet. Computation involves solving a non linear equation. Whether the accuracy of the results justifies the complexity of the solution is an issue for the Working Group to resolve. If use of MLEs is deemed too complex, the best replacement method is replacement with the minimum non-zero SAIDI value. A spreadsheet with the example calculations is available at

[www.ee.washington.edu/people/faculty/christie/ZeroDayEst.xls](http://www.ee.washington.edu/people/faculty/christie/ZeroDayEst.xls).

## **X. References**

- [Chr03] R.D.Christie, "Zero SAIDI Days Issue - Response to WMECO", [www.ee.washington.edu/people/faculty/christie/ZeroDayIssue.pdf](http://www.ee.washington.edu/people/faculty/christie/ZeroDayIssue.pdf), January 24, 2003.
- [Cro88] E.L. Crow and K. Shimizu, *Lognormal Distributions: Theory and Applications*, Marcel Dekker, Inc., New York, 1988.
- [Sch86] H. Schneider, *Truncated and Censored Samples from Normal Populations*, Marcel Dekker, Inc, New York, 1986.