Paper 240

Using SAS® Macros to Develop Confidence Intervals for the Weibull and Extreme Value Distribution Using Type II Censored Data

Scott M. Jordan and Muhammad H. Zaman,
Arkansas Tech University, Math Department, Russellville AR 72801

The Weibull distribution has wide applications, particularly in life distribution. There exists a large body of literature on statistical methods based on the Weibull distribution, but most of the methods require numerical integration or Monte Carlo methods to develop the confidence intervals for model parameters. The reason for the large volume of methods revolves around the fact that there is no two-dimensional sufficient statistic for the shape and scale parameters, and therefore the possibilities of producing estimators are many.

There are two methods of developing confidence intervals for the location, scale, and percentiles associated with the extreme value distribution given in equation (2) that will be considered in this paper. SAS macros will be developed for both methods. The next section will discuss the two methods for developing confidence intervals and the last section will discuss the SAS macros.

The Weibull p.d.f. is

$$f(t/\alpha,\beta) = (\beta/\alpha)^{\beta-1}\exp[-(t/\alpha)^{\beta}] \quad t \geqq 0 \qquad (1)$$

where β and α are the shape and scale parameters, respectively.

If T has the probability density function given in equation (1), then X = ln(T) has the probability density function given in equation (2).

$$f(x/u,b) = (1/b)\exp[(x-u)/b]\exp[-\exp[(x-u)/b]]$$
$$-\infty \leq x \leq \infty \qquad (2)$$

where u ($-\infty < u < \infty$) and $b > 0$ are parameters. Sometimes people prefer to work with the extreme value distribution given in (2), but any results derived can be easily converted back to the Weibull distribution since $u = \ln\alpha$ and $b = \beta^{-1}$.

A Type II censored sample is one for which only the r smallest observations in a random sample of n items are observed $(1 \leq r \leq n)$. Experiments involving Type II censoring are often used in life testing; where a total of n items are placed on a test and the test is terminated at the time of the r^{th} failure. Such tests can save time and money. Clearly when $r = n$, the test is terminated after all n items have failed.

METHODS:

Method 1. A Conditional Method to Develop Confidence Intervals

The following method was developed by Lawless (1972, 1978).

If $\tilde{u}$, $\tilde{b}$, and $w_p = \ln[-\ln(1-p)]$ are equivariant estimators of u, b, and w_p based on a type II censored sample $x_1 \leq \ldots \leq x_r$ from (2), then we have the following results:

(i) $Z_1 = (\tilde{u} - u)/\tilde{b}$, $Z_2 = \tilde{b}/b$, and $Z_p = (\tilde{u} - x_p - w_p b)/\tilde{b}$, are pivotal quantities.

(ii) The quantities $a_i = (x_i - \tilde{u})/\tilde{b}$, $i = 1,\ldots,r$ form a set of ancillary statistics.

The conditional probability density function of Z_2, given a, is of the form

$$h_2(z|a) = \frac{k(a,r,n)z^{r-2}\exp\left((z-1)\sum_{i=1}^{r} a_i\right)}{\left(\frac{1}{r}\sum_{i=1}^{r}{}^{*}\exp(a_i z)\right)^r} \qquad z \geq 0 \qquad (3)$$

where $\sum_{i=1}^{r}{}^{*}\exp(a_i z) = \sum_{i=1}^{r}\exp(a_i z) + (n-r)\exp(a_r z)$.

The conditional p.d.f. of Z_p, given a, is

$$P(Z_p \leq t \mid a) = \int_0^{\infty} h_2(z \mid a) I\left(r, \exp(w_p + tz)\sum_{i=1}^{r}{}^{*}\exp(a_i z)\right) dz \qquad (4)$$

where $I(r,s) = \frac{1}{\Gamma(r)}\int_0^s u^{r-1}e^{-u}du$ for $r>0$, $s>0$ (the incomplete gamma function). The p.d.f. of Z_1, given a, is given by (4) with $w_p = 0$. To construct confidence intervals for u, b, and w_p one needs to obtain the appropriate percentage points for Z_1, Z_2, and Z_p. Numerical integration of equations (3) and (4) is required to find these percentage points. Once this is achieved then the equations given in (5) can be used to construct the confidence intervals.

Method 2. Monte Carlo Method to Produce Confidence Intervals

Since Z_1, Z_2, and Z_p are pivotal quantities, their distributions do not depend on the values of u and b in equation (2). If we set $u = 0$ and $b = 1$ ($\alpha = \beta = 1$), then the pivotal quantities become

$Z_1 = \tilde{u} / \tilde{b}$, $Z_2 = \tilde{b}$,

$Z_p = (\tilde{u} - x_p - w_p)/\tilde{b}$.

The distribution of Z_1 can be estimated by generating samples from an exponential distribution in (1) on the computer, then using PROC LIFEREG to compute $\tilde{u}$ and $\tilde{b}$, (which are used to obtain a value of Z_1). After repeating this process many times (say 100,000 times), the distribution of Z_1 can be approximated fairly well and the percentage points can be obtained from that distribution. The distributions and percentage points of Z_2 and Z_p can be estimated in a similar fashion.

SAS MACROS:

METHOD 1

The input required to process the macro would include the data (r uncensored values and (n-r) censored values from the extreme value distribution given in (2)), $\tilde{u}$, $\tilde{b}$ (the maximum likelihood estimators), and the a_i's.
Then the flow of the macro would proceed as follows:

1. First the macro would have to build the function given in equation (3) for given values of r and n. The major problem is creating the sum ($\sum_{i=1}^{r}{}^{*}\exp(a_i z)$) in the denominator of equation (3) . The following will create the sum:

```
%macro build;
 sumeai = 0;
  %do j = 1 %to &r;
    if &j lt &r then
    eai = exp(a&j*z);
    else eai = (n – r +1)*exp(a&j*z);
 sumeai = sumeai + eai;
 %end;
%mend build;
```

2. The constant $k(a,r,n)$ in equation (3) must be estimated by numerical integration.

The numerical integration technique used in this macro will be a 50 point Gauss-Legendre formula. The following is a brief description of this technique. This technique tries to express the integral as $I \cong c_0 f(x_0) + \ldots + c_{n-1} f(x_{n-1})$, where the c_i's and x_i's are unknown coefficients to be calculated. Values for the *c*'s and *x*'s are summarized in books (Stroud and Secrest 1966) and in many articles. A 50 point Gauss-Legendre formula requires a data set with 50 *c*'s and *x*'s, which can be stored in an external file and brought in when needed using an INCLUDE statement.

The constant

$$k(a,r,n) = \left[\int_0^\infty \frac{z^{r-2} \exp\left((z-1) \sum_{i=1}^{r} a_i \right) dz}{\left(\frac{1}{r} \sum_{i=1}^{r}{}^{*} \exp(a_i z) \right)^r} \right]^{-1}.$$

3. Once *k*(a,r,n) is known, we can determine the appropriate percentage points for our confidence intervals by numerically integrating the appropriate function in equation (3) or (4).

For example:

$P(l \le Z_2 \le m \mid a) = .95$ implies that

$\int_l^m h_2(z \mid a)dz = .95.$

Exact percentage points for l and m that satisfy the probability equation above can be determined by evaluating the above integral numerically.

Now to construct a 95% confidence interval for the scale parameter b, we use $P(l \le Z_2 \le m \mid a) = .95$ which implies that

$P(l \le \tilde{b}/b \le m \mid a) = .95$. This yields

$\tilde{b}/m \le b \le \tilde{b}/l$, a 95% confidence interval for b.

Values for l and m can be determined in a similar fashion for Z_1 and Z_p, and thus confidence intervals for the shape parameter u and percentiles x_p can be found using the results in equation (5).

To determine the limits of integration it is useful to have a graph of the region you are trying to integrate. A macro has been developed using a minimal amount of input and is used in the example to graph the region of interest.

Note: The conditional method produces confidence intervals on the conditional distribution of Z_1, Z_2, and Z_p, given the observed value of the ancillary statistic a. This implies that the percentage points obtained in the confidence interval change each time the data changes. This method is equivalent to a Bayesian posterior probability interval, using the appropriate improper prior distribution.

METHOD 2

The input required to process the macro for this method are the values of r, n, *p*, the lower percentage point, the upper percentage point, and the number of samples to use in the simulation.

Once r and n are known the simulation will be performed as follows.
Since Z_1, Z_2, and Z_p are pivotal quantities, their distributions do not depend on the values of u and b in equation (2) or the values of α and β in equation (1). Therefore, we can choose any value of α and β to conduct the simulation. If we let $\alpha = \beta = 1$ in equation (1) then f(t) = exp(-t) for $t \ge 0$.

To simulate type II censored data we must first simulate n observations from an exponential distribution such that the r smallest observations are uncensored and the remaining $(n-r)$ observations are censored.

Then repeat this process many times (50,000 to 100,000). Then PROC TRANSPOSE can be used to write the *n* observations in each simulated group in row form; for example:

group 1 $W_1\ W_2\ W_3 \ldots\ W_n$
group 2 $W_1\ W_2\ W_3 \ldots\ W_n$
$\vdots$

If the W_i *'s* are in ascending order then the following code will set the $(n-r)$ largest observations equal to the r^{th} largest value. This will make the $(n-r)$ largest values the censored values.

```
data;
        array ww{&n} w1 – w&n;
         do i = &r to &n;
         ww{i} = ww{&r};
         end;
```

Then PROC TRANSPOSE can be used to transpose the data back into column format.

At this point we can use PROC LIFEREG to calculate the maximum likelihood estimates for u and b for each group. PROC LIFEREG by default takes the natural log of the response and then calculates the maximum likelihood estimates for u and b from the extreme value distribution given in equation (2). Now we have many samples from which the distributions of Z_1, Z_2, and Z_p can be simulated. The percentage points can be obtained by using PROC UNIVARIATE on the simulated values of Z_1, Z_2, and Z_p. Once the percentage points are determined, then the confidence intervals can be obtained by the following formulas:

1. $\tilde{b}/m \le b \le \tilde{b}/l$
2. $\tilde{u} - \tilde{b}m < u < \tilde{u} - \tilde{b}l$
3. $\tilde{u} - m\tilde{b} < x_p < \tilde{u} - l\tilde{b}$

   where $x_p = u + w_p b$. (5)

(l and m are the percentage points associated with $P(l < Z_i < m) = C$)

Numerical example: Mann and Fertig (1973) give the failure times of airplane components for a life test in which 13 components were placed on a test and the test was terminated after the tenth failure. The failure times in hours were .22, .50, .88, 1.00, 1.32, 1.33, 1.54, 1.76, 2.50, 3.00. This experiment involves a type II censored sample with n = 13 and r = 10. Suppose we want to construct a 90% confidence interval for b. The maximum likelihood estimate for b is .706.

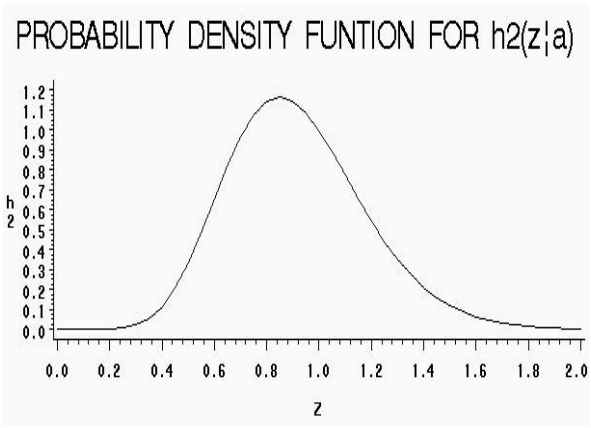
Method 1

The input required to process the macro would include the data (r =10 uncensored values and n-r = 3 censored values from the Weibull distribution), $\tilde{u}$, $\tilde{b}$ (the maximum likelihood estimators), and the a_i's. For example:

```
data mann;
input time cens;
u = .821; b = .706;
lntime = log(time);
a_i  = (lntime – u)/b;
cards;
0.22  0
0.50  0 .
0.88  0
1.0    0
⋮
3.00  0
3.00  1
3.00  1
3.00  1
;
```

At this point the integration macro is called and it will integrate equation (3) to determine the percentage points associated with Z_2. To determine the limits of integration it is useful to see a graph of the region associated with Z_2. Another macro was developed to graph the regions associated with the Z_i's. The

graphing macro output is below and it will help us to get close to the correct percentage points, and with the use of a bisection method we can achieve the accuracy that we want.



Using the integration macro and a bisection method the following results were achieved:

$P(Z_2 \le .52485 \mid a)$ = .05 and
$P(Z_2 \le 1.3833 \mid a)$ = .95; therefore

$P(.52485 \le \tilde{b}/b \le 1.3833 \mid a)$ = .90 which yields .510 ≤ *b* ≤ 1.345 as a 90% confidence interval for *b*.

Method 2

Method 2 uses a macro developed to find the percentage points for Z_1, Z_2, and Z_p by Monte Carlo methods. The input required is the number of simulations, say 100,000, *n* = 13, *r* = 10, the lower percentage point 5, and the upper percentage point 95. The macro then produces the following percentage points for Z_2.

$P(.521 \le Z_2 \le 1.393 \mid a)$ = .90 which yields from equation (5) 0.507 ≤ *b* ≤ 1.355 as a 90% confidence interval for *b*.

Note: There is very little difference between the two methods in this particular problem.

In conclusion, these two methods represent two different schools of thought on developing confidence intervals, a conditional method and classical method. Both methods require a considerable amount of computer work to be useful. The SAS programming language is very useful in dealing with the computer work and is well recognized as a standard in the data analysis community. At present, I do not have any recommendations on which method is better, but with the use of the ever growing arsenal of programs, these methods are more accessible to the general data analysis community and should be examined more closely.

BIBLIOGRAPHY

1. Lawless, J. F. (1972). "Confidence Interval Estimation for the Parameters of the Weibull Distribution." *Utilitas Math.*, 2, 71-87.

2. Lawless, J. F. (1978). "Confidence Interval Estimation for the Weibull and Extreme Value Distributions." *Technometrics*, 20, 355-364.

3. Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, Inc

4. A. H. Stroud, and D. Secrest (1966). *Gaussian Quadrature Formulas*, Prentice-Hall inc.

5. N. R. Mann and K. W.Fertig (1973). "Tables for obtaining confidence bounds and tolerance bounds based on best linear invariant estimates of parameters of the extreme value distribution." *Technometrics*, 15, 87-101.

6. S. C. Chapra, and R. P. Canale (1988). *Numerical Methods for Engineers* 2nd ed., McGraw-Hill.

7. SAS Institute Inc. (1997), *SAS Macro Language*, First edition, Cary, NC: SAS Institute Inc.

This research is supported by the NASA-JOVE program, grant number NAG8-1284.

For information about programs or other general comments you may contact the author at:

Scott M. Jordan
Arkansas Tech University
Math Department
Russellville, AR 72801
Phone: (501) 968-0657
Email: masj@atuvm.atu.edu