Spline Restoration Method for Nonparametric Estimation of Reliability on Incomplete Data

I.V. Gadolina, R.I. Zainetdinov (Dept. of Mechanical Design)

Key Words: Reliability, Incomplete Data, Nonparametric Estimation, Right Censoring, Left Truncation, Spline Restoration Method, Lognormal Distribution, Gongola Car Body.

ABSTRACT : We propose the Spline Restoration method (SRM) for nonparametric estimation of a sample cumulative distribution function *Cdf{t}* in the case when the data are right censored and left truncated. An interactive procedure is based on the division of the sample to the number of age groups (subsets). Each subset contains the items of the same manufacturing year. For the approbation of method a simulation study was done. The estimation of reliability of load-carrying welded structures of the gondola car body was performed on the base of data that were assembled during single inspections of technical state.

Nomenclature

j	: Order number of age group (subset), $j=1, 2,..., j_{max}$
JO	: Estimator based on Johnson method
m_j	: Number of censored items in age group (subset) j
N	: Total number of the items in the sample
N^*	: Number of the examined items during the inspection
N_j	: Total number of the items in the age group (subset) j
N^*_j	: Number of the examined items in the age group j during the inspection
q_j	: Number of items in age group (subset) j, failed before the year of inspection
r_j	: Number of failed items in the age group j in the year of inspection (observer value)
SRM	: Spline Restoration Method
t	: Time to failure
t_j	: Approximation for the time to failure; integer number
Var	: Variance of logarithm t
W	: The inspection year
$x_j=W-Y_j+1$	: Approximation for the service time of items in the age group j; integer number
Y_j	: Manufacturing year of items in the age group (subset) j

Greek Symbol

μ	: Mean (expected) value of logarithm t

1. INTRODUCTION

Different types of incomplete sets are known: right censored, left censored, and left truncated. In these cases estimation of the cumulative distribution function *Cdf{t}* can create problems. In the case of the reliability studies of the gondola car's body a special case of set truncation takes place. Usually these machines are observed only once during the annual repair and a lot of information is being lost. The only information available is about the condition of machine parts in the year of inspection. They could be in working condition or they just had failed down during that year. Most likely these items will be never observed again. The initial size of the sample is also unknown –- that is the main problem.

Another important feature of the data is their presentation in a grouped form. Each age group (that is, subset of original sample) consists of the items of the same manufacturing year and we know the time to failure only approximately within one year of accuracy. Pieces of *Cdf{t}* are estimated on the data about each subset (we assumed that each subset was responsible for its own part on the plot) and they are put together using an iterative formula. So, the name of the method is "Spline restoration method". An example of a simulation study has to elucidate the formation of these specific sets and to explain the estimation procedure. Since the case of multiple right censored samples often takes place in reliability studies, a comparison of the results to those obtained from the Johnson formula, which is good for right-censored samples, has been done.

2. REVIEW

In machine reliability studies *t*, time to failure, is one of the most important values. Therefore it has a statistical nature due to many reasons, the cumulative distribution function *Cdf{t}* should be investigated.

The problem of *Cdf{t}* estimation is a common one in statistical analysis no matter which branch of science is being considered. In medicine, in sociology, in education it appears as well as in technique. If the sample of *N* observations is exactly observed, the natural estimator is the sample distribution function *Cdf{t},* which is the proportion of the number of items failed with the times to failure less or equal to the argument *t*. Often not all the items are exactly observed. This fact creates many problems in statistical analysis and many studies have been done. However often a new statistical investigation brings a new type of censoring. It leads to the appearance of some new approaches for the *Cdf{t}* estimation.

So far the terminology is not too definite. Mostly a distinction is made between the samples, which are right censored, left censored and left truncated. The case of right censoring is best known. Kaplan-Meier, [(1)] Johnson [(2)] and Nelson [(3)] estimators are the maximum-likelihood estimators and are shown to be asymptotically the same. [(4)] The case of left censored data often takes place in education studies, when, for example, a researcher deals with a group of children and studies their ability to perform a definite task. So, if there are some children, who already know, how to perform the task, he deals with a left-censored sample, because the number of these children is known. Turnbull [(5)] proposed the self-consistent estimators for such samples. In [(6)] the idea of self-consistency was extended to the case of double right- and left-censored data. In machine reliability studies, however, often we do not know the number of machines broken before the investigation, so we have left truncated samples. This case had been studied among the others in. [(7)] The

author proposed a parametric model, where the parameters of the Weibull distribution depend on the environment (namely stress in this case).

3. ORIGIN OF LEFT TRUNCATED SAMPLES IN RELIABILITY STUDIES

In the practice of single inspections of nonrestorable[1] items during the annual inspection of the gondola cars a specific case of the multiple randomly censored sample often takes place. The samples are right censored, that means that some of observable items are efficient at the moment of inspection. Since the censoring times are different, the obtained samples are multiple right censored. Moreover since there were no observations done before the single inspections, the information about early failures was lost. Since the items which failed before have been just replaced, their number is also lost and the total size of the initial sample is unknown. This makes the sample also left truncated.

Another important peculiarity of these samples is their representation in a grouped form. That means an interval censoring. At single inspections the researcher does not know the exact service life of the item. Only the manufacturing year is known. This fact allows representing the entire sample as consisting from a few small samples, each of which represents a definite subset j, namely the age group of items of one manufacturing year.

The restoration method is based on this division to age groups (subsets). Each item had been observed only once and x_j is its service time at the moment of inspection, where *j* is the order number of the age group that the item belongs to. Since we do not take into account the time differences within one age group we can use an approximation $x_j = j$ for the all items and $t_j = j$ for the failed items.

4. SIMULATION STUDY

To reveal the differences in the *Cdf{t}* estimation obtained by using different methods a simulation study has been done. Although the method is non-parametric, to have data in a simulation study a lognormal distribution has been used. This distribution describes the reliability as well as Weibull distribution. In the case of the lognormal distribution time to failure *t=exp(r.v.)* where *r.v.* is *gau(μ,Var).* In modelling each item *i (I = 1, 2, ..., N)* should belong to a definite age group *j (j = 1, ..., j_{max})* with the "censoring window" *[0; j-1).* For random t_i one of the following outcomes is possible: [2]

1) $j-1 < t_i \leq j$. It means failure, which had been scored during the inspection;

2) $t_i > j$. It means right censoring;

---

[1] We can call it "nonrestorable" because in according to the aim of the study only reliability before appearing the first crack was investigated. Therefore the secondary cracks were not under investigation.

[2] We can call it "censoring window" using the term "window" like in. (5) The difference is that here we have a number of windows with the different width.

3) $t_i \le j-1$. It means left truncation and, hence, this information is inaccessible.

For an expert, who conducts the inspection of the technical state, only observation at points 1) and 2) are accessible. He would never know about 3). The number of observed items N^* would be less than the initial volume N of the sample.

An example of a simulation with $\mu = 1$ and $Var = 0.64$ is shown in Fig. 1, à), in triangular Matrix 1. Here the total numbers of failed and censored items are shown. Each row of the matrix corresponds to a group of the items of the same manufacturing year. The matrix has a triangular form because in group j $x_j = j$ and the time to failure t could not exceed service time ($t \le x_j$). The total numbers of failures in the year of inspection are in the darkened cells. The numbers of the failures in previous years are shown in the cells that are l.h.s. of the main diagonal. Numbers of right censored items are shown r.h.s. in column separately from a matrix. If we follow the idea of single inspection, we can say the information laying l.h.s. of the main diagonal is inaccessible, and the matrix achieves a form shown in Fig. 1, b), Matrix 2.

5. SRM ESTIMATION

If only there were no censoring the nonparametric estimation of the distribution function would be

$$Cdf\{t_j\} = \frac{\sum_{k=1}^{j} r_k}{N + 1} \qquad (1)$$

In our situation we could not do that, because we know neither r_k nor N.

Let us accept that the jth age group is crucial for point estimation of $Cdf\{t, t = j\}$ at the end of the jth time interval ($t_{j-1} < t \le t_j$). It enables to apply an iterative procedure.

Let us assume the non-zero number of failures in the first age group, that is $r_1 \ne 0$. Now we can say that as a result of single inspection it has been revealed, that in subset $j=1$ with a service time $x=1$ year, r_1 items had failed and m_1 items had been right censored. Here the number of examined items N_1^* is equal to the initial age group (subset) size

$$N_1^* = N_1 = r_1 + m_1. \qquad (2)$$

The unbiased estimator is defined as in Eq.(1):

$$Cdf(t_1) = \frac{r_1}{N_1 + 1} = \frac{r_1}{n_1 + m_1 + 1} \qquad (3)$$

For subset $j=2$ $(x=2)$ the total number of examined items

$$N_2^* = r_2 + m_2 \qquad (4)$$

is not equal to the initial volume of the age group N_2

$$N_2^* = N_2 + q_2$$

(5)

Here q_2 is the unknown total number of items in the second age group which failed during the first year of service life. We can say the information was lost because of left censoring. Let us use the estimator $Cdf(1)$ for the restoration of unknown number q_2:

$$q_2 = Cdf(t_1)(N_2 + 1) \quad (6)$$

where $N_2 = r_2 + m_2 + q_2$ (7)

The equations (6) and (7) form a system with two unknown parameters N_2 and q_2. After calculating N_2 and q_2, it is possible to estimate $Cdf\{2\}$ using the data from the second age group (subset):

$$Cdf\{2\} = \frac{q_2 + n_2}{N_2 + 1} \quad (8)$$

In general the estimators for N_2 and q_2 are not the integer numbers.

The estimator $Cdf\{2\}$ is used for restoration of the total number of the left censored items q_3 for the third age group and the total number of items in this age group. Next, the unknown parameters for the fourth subset are estimated, etc.

The iterative formula for the estimation of $Cdf\{j\}$ at the end of jth time interval

$$Cdf\{j\} = Cdf\{j-1\} + \frac{r_j[1 - Cdf\{j-1\}]}{1 + r_i + m_i} \quad (9)$$

The total size of the initial sample can be estimated using estimators $Cdf\{1\}, Cdf\{2\}, \ldots Cdf\{j_{max}\}$

$$N = n_1 + m_1 + \sum_{j=2}^{j_{max}} \frac{Cdf\{j-1\} + n_j + m_j}{1 - Cdf\{j-1\}} \quad (10)$$

In Fig. 2 along with the initial function of lognormal distribution (the line 1) SRM-estimator of *Cdf{t}* is shown (line 3). It has a smaller bias than in the case of JO-estimation.

Results of the simulation for six modeled (initial) samples with different parameters μ and *Var* are shown in the table. SRM-estimators show some bias to conservative area, which however is less than non-conservative bias, which has place in using JO method.

6. EXAMPLE

Now we shall consider an example, which created an initial problem with the left truncated data. With the application of the proposed method an estimation of the reliability parameters for the load-carrying welded structure of a body and a frame of the four-axle gondola cars has been done. The data concerning to

their technical state were received during single field inspections. [7, 8] Criterion of a failure in a load-carrying structure was detection of a first fatigue crack in welded connections. The total number of items in an examined sample of the gondola cars' frames were 1440 (N^*) with unknown initial sample size N. The estimator of N with the application of SRM method, was 1590. Estimators of $Cdf\{t\}$ for a pin section of the gondola car frame by SRM and JO methods are shown (Fig. 3). The presence of conservative bias in using the SRM-method and non-conservative in using JO-method allows us to assume that the true distribution function dislocates between lines 1 and 2.

For reliability parameters it is also important to estimate bias and confidence intervals. In the special case of the grouped multiple right censored and left truncated data the procedure of construction of confidence intervals is very complicated. For the effective solution of this problem bootstrap-modelling could be helpful, as it was shown in. [9-11]

ACKNOWLEDGMENT

The authors are grateful to Prof. Wanda I. Griffith from Yong-In Technical College for reading the manuscript and giving valuable comments.

REFERENCES

(1) Kaplan, E. L., Meier, P., 1958, "Nonparametric estimation from incomplete observations", *J. Amer. Statist. Ass*., Vol 53, pp 457-481.

(2) Johnson, L. G., 1964, *Theory and Technique of Variation Research*, Elsevier, N.Y.

(3) Nelson, W., 1982, *Applied Life Data Analysis*, Wiley, N.Y.

(4) Blagoveschensky, U. N., 1979, "Of asymptotically normality of one class statistics for randomly censored sets", *Teorija Verojatnostey i eje Primemnenija*, No. 3. (In Russian).

(5) Turnbull, B. V., 1974, "Nonparametric Estimation of a Survivorship Function with Doubly Censored Data", *J. Amer. Statist. Ass*., Vol. 69, No. 3, pp. 169 - 173 .

(6) Tsai, W. Y., Crowley, J., 1985, "A Large Sample Study of Generalized Maximum Likelihood Estimators from Incomplete Data via Self-Consistency", *The Annals of Statistics*, Vol. 13, No. 4, pp. 1317 - 1334.

(7) Glaser, R. E., 1995, "Weibull Accelerated Life Testing With Unreported Early Failures", *IEEE Transactions on Reliability*, Vol. 44, pp. 31 - 36.

(8) Kiselev, S. N., Faerstein, Y. O., Zainetdinov, R. I., 1984, "Reliability of Welded Structures of Freight Wagons", *Jeleznodorojny Transport*, No.11, pp. 35-37. (In Russian).

(9) Zainetdinov, R.I., 1996, "Using Bootstrap Modelling for Statistical Evaluation of the Reliability Parameters of Load-Carrying Welded Structures in Freight Wagons, *Welding International*, Vol. 10, No. 6, pp. 491 - 495.

(10) Gadolina, I. V., 1986, "Application of Bootstrap-Modelling at Construction of Confidence Intervals on Censored Samples," *Nadejnost i Kontrol Kachestva*, No. 6, pp. 53 -57. (In Russian).

(11) Adler, Y. P., Gadolina, I. V., Ljandres, M. N., 1987, "Bootstrap-Modelling at Construction of Confidence Intervals on Censored Samples," *Zavodskaja Laboratorija*, No. 10, pp. 90 - 94. (In Russian).