Nonparametric estimation of the survival probability function when some observations are incomplete

Terotechnica, 1 (1979) 39–45

NONPARAMETRIC ESTIMATION OF THE SURVIVAL PROBABILITY FUNCTION WHEN SOME OBSERVATIONS ARE INCOMPLETE

J.H.J. Geurts

Eindhoven University of Technology, Department of Industrial Engineering, P.O. Box 513, Eindhoven (The Netherlands)

(Received September 15, 1978; accepted January 13, 1979)

Abstract

Empirical survival data sometimes consist of a number of observations of running times to failure, interspersed with a number of running times that have not yet been completed by a failure but of which no further observation is available. After a brief discussion of another approach a method is described for estimating the survival probability function from such data, making full use of the information that is contained in all the observed running times, whether completed by a failure or not. An approximative method, long known to actuarial science, is discussed. Extensions and a related estimation method are mentioned.

1. INTRODUCTION

Most practitioners in the field of Maintenance or, more specifically, Reliability have at some time needed to estimate from experimental or field data the survival probability function (often also referred to as the reliability function) of an item, for instance in order to determine whether a preventive maintenance policy might be useful. They are familiar enough with estimation methods that are applicable to a sample of times to failure. However, sometimes the data are in the form of a number of running times to failure interspersed with a number of running times where failure has not yet occurred but where further observation is not available. In this latter case estimation of the survival probability function is still no problem if all the incomplete observations are of running times that exceed all the observed times to failure. But if some of the observed running times where failure has not yet occurred are smaller than some of the times that were terminated by a failure, an estimation problem arises. It is to this problem that the present article addresses itself.

A simplified but typical example of the data structure is given in Table 1, where the observations are listed in order of increasing registered running time, which is not necessarily the chronological order in which they became available.

It should be kept in mind that the unit in which "running time" is measured ought to be the unit in which the life of the part is most appropriately expressed, e.g. miles travelled, number of operations, dissipated KW-hours, or of course calendar time.

It is not uncommon to encounter data of

TABLE 1

Observations on 10 units

Unit no.	Registered running time (weeks)	Failed at that time	Not failed at that time (end of observation)
1	9	√	
2	13		√
3	30		√
4	35	√	
5	55	√	
6	75		√
7	92	√	
8	120		√
9	127	√	
10	139		√

this form. They may arise

(1) from field surveys, when units have been installed at different times or used with different intensities;

(2) from laboratory tests, when some units have been accidentally damaged or removed from test before failure;

(3) generally, when units are subject to different, mutually independent, failure modes and failure due to one mode implies the impossibility of observation of time to failure due to another mode.

The question is how to incorporate, if at all, the information on running times of the non-failed sample units into the estimate of the population survival probability function.

It would be a waste of precious data to exclude the running times registered by the non-failed items for clearly, although such items have not told us the complete story of their life, they do yield information about it. Consider for instance an estimate, from the data of Table 1, of the survival probability for the first five weeks of running time. Whatever the true (population) value of the survival probability, it would be covered in 95% of the cases by the (Clopper–Pearson) confidence interval 0.48–1.00 when estimated from the 5 complete observations alone. However, when estimated from the full sample of 10 registered running times, the true survival probability would be covered in 95% of the cases by the much shorter interval 0.69–1.00.

What is needed, then, is a method of estimation that makes the fullest use possible of the information contained in the sample and that does not make any assumptions about the form (e.g. Weibull, lognormal) of the survival probability function of the population of which the sample is supposed to be representative, this form being in many cases the very subject of inquiry.

In Section 3 a method is described that does just this. In Section 4 an approximating method is described that has long been known in actuarial science. As mentioned before, neither method is new, see for instance [1]. In Sections 5 and 6 extensions and a related method are briefly discussed. But first we shall review and briefly comment upon another method that incorporates information from the non-failed items.

2. RANK METHOD

A method in fairly common use for dealing with this problem is that described by, among others, Johnson [2] and which, for convenience, we shall refer to as the rank method. To explain this method we make use of an example taken from this latter text. A sample of six running times (measured in hours) consists of three items to failure, $F_1 = 112$, $F_2 = 250$ and $F_3 = 572$ respectively, and three times when observations on items still running were suspended at times $S_1 = 213$, $S_2 = 484$ and $S_3 = 500$, respectively. Estimates of points on the survival probability function can be calculated if the final order numbers (f.o.n.) of the failed items can be estimated, i.e. the order numbers that the failed items would have occupied in the sample of six if the suspended items had been observed until failure.

The f.o.n. of F_1 obviously is 1. If S_1 were to fail before 250 h, F_2 would receive f.o.n. 3,

regardless of the eventual but not yet observed times to failure of S_2 and S_3; if S_1 were to fail after more than 250 h running time, F_2 would receive f.o.n. 2, again regardless of what the eventual times to failure of S_2 and S_3 would turn out to be. (It is assumed that failure of S_1 at exactly 250 h is impossible.) Of the two possible final order numbers for F_2 one may be more likely than the other; it seems clear that the population failure rate in the neighbourhood of 250 h running time has something to do with it. If the failure rate between 213 and 250 h is decreasing, survival of S_1 to some time after 250 h – and thus a f.o.n. of 2 for F_2 – would seem more likely than if the reverse were the case, when a f.o.n. 3 for F_2 would seem more likely. The rank method resolves this problem as follows.

The various possible failure times of the suspended items can only lead to a total of 30 different possible final orderings of the six items in the sample. The method assumes all these orderings to be equally likely and uses the number of orderings (24) that result in a f.o.n. of 2 for F_2 and the number of orderings (6) that result in a f.o.n. of 3 to arrive at an estimate of the f.o.n. for F_2 of $(24 \times 2 + 6 \times 3)/30 = 2.2$. This last figure is then used to make a median or mean estimation of the corresponding population percentage rank which is used as the cumulative percentage failed on, say, a Weibull plot.

Contrary to the requirements set forth at the end of the previous section, the equiprobability assumption implies an assumption about the form of the population survival probability function. This may be demonstrated as follows.

The probability that F_2 should receive f.o.n. 3 is identical to the probability that S_1, having survived 213 h, fails between 213 and 250 h. This probability is

$$\Pr\{\text{f.o.n. } F_2 = 3\} = [P(213) - P(250)]/P(213)$$

where $P(t)$ stands for the value of the *population* survival probability function P at time t. Similarly the probability that F_2 should receive f.o.n. 2 is $\Pr\{\text{f.o.n. } F_2 = 2\} = P(250)/P(213)$. The equiprobability assumption puts $\Pr\{\text{f.o.n. } F_2 = 2\}/\Pr\{\text{f.o.n. } F_2 = 3\} = 24/6 = 4$; after substitution this reduces to $P(213) = (5/4)P(250)$, and this is clearly an assumption about the form of the population survival probability function. Further implicit assump-

TABLE 2

Calculation scheme, PL method (in last column: estimates of survival probabilities at the beginning of the corresponding interval according to Johnson's method; left based on mean ranks, right based on median ranks)

Interval no.	Interval (weeks)		No. of units			Estimate of probability of survival			
	>	⩽	Entering	Lost, due to		During this interval	Up to and including this interval	Johnson's	
				Failure	End of observation			Mean	Median
1	0	9	10	0	0	10/10	1		
2	9	13	10	1	1	9/10	9/10 = 0.90	0.91	0.93
3	13	30	8	0	1	8/8	9/10 = 0.90	–	–
4	30	35	7	0	0	7/7	9/10 = 0.90	–	–
5	35	55	7	1	0	6/7	27/35 = 0.77	0.80	0.81
6	55	75	6	1	1	5/6	27/42 = 0.64	0.68	0.69
7	75	92	4	0	0	4/4	27/42 = 0.64	–	–
8	92	120	4	1	1	3/4	27/56 = 0.48	0.55	0.55
9	120	127	2	0	0	1/2	27/56 = 0.48	–	–
10	127	139	2	1	1	1/1	27/112 = 0.24	0.36	0.36
11	139	?							

tions about the form of P follow from the estimation formulae for the f.o.n. of F_3.

It should also be noted that, for the above example, the method results in an estimate of 33% of the population surviving 572 running hours. It would appear that this is rather optimistic in the face of a sample of 6 items, of which not one has given any indication that there is life at all beyond 572 h. The PL method, to be described in the next section, would lead to an estimate of $P(572) = 0$. However, if the implied assumptions about the shape of the population survival probability function are viewed as approximations, the rank method may be satisfactory, though not always reliable. The practical effects of such approximations are not immediately clear and may depend, for instance, on the relative number of suspensions. In Table 2 estimates resulting from the application of the rank method are compared with those from the PL method. A comparison of the two methods on the basis of the errors of the estimated probabilities is not possible at present: for the PL-method only asymptotic variance formulae are available, see [1], while the 90% confidence intervals provided for use with the rank method [2], are only strictly valid for complete samples. Further investigations in this area are to be conducted.

3. THE PL METHOD

The method to be described in this section is known as the Product Limit (PL) method or Kaplan–Meier method. For a thorough and rigorous treatment, see [1].

A basic condition for the validity of this method (and its approximation) is that the times beyond which observation is not possible and the times to failure are independent. This condition is for instance violated if test units are removed from the test rig because they seem to be deteriorating. It is also violated if, in the case of different failure modes, failure due to some modes changes the probabilities of failure due to the modes of interest.

The scheme of calculations is summarised in Table 2. Note that the interval limits are constituted by the times at which failure or loss from observation occurred; it is only at those times that changes, if any, in the estimated survival probability function can occur.

All 10 units have been in service during 9 weeks or more (interval 1). One failed after 9 weeks; therefore the probability of survival until the end of the 9th week is estimated as $10/10 = 1$.

Of the 10 units that entered the second time interval of 9 to 13 weeks, one failed – immediately after the start of the interval; that is how we have chosen our interval limits – and 9 survived.

We conclude: conditional upon a unit's surviving its first 9 weeks (for which the estimated probability was 1) the probability that it will reach the end of the next 4 week interval is estimated as 9/10. Therefore the probability that it will survive unconditionally until the end of the 13th week is estimated as $1 \times 9/10 = 9/10$. At this moment (the end of the second time interval, i.e. 13 weeks) we lose another unit from our sample. It has been operating until its 13th week and was still operating at that time, but we have no observation of its life (or failure) beyond that time. So we cannot learn anything from it about its chances of survival beyond that time; therefore we exclude it from further consideration.

So we enter the third interval (13 to 30 weeks) with 8 surviving units. Not one of them fails during these 17 weeks. We conclude: if a unit reaches its 13th week (the corresponding estimated probability was 9/10), then its estimated probability also to survive the next 17 week interval is $8/8 = 1$. The unconditional estimated probability to reach the 30th week then is $9/10 \times 1 = 9/10$. Here again we lose a unit from observation and we continue with 7 units.

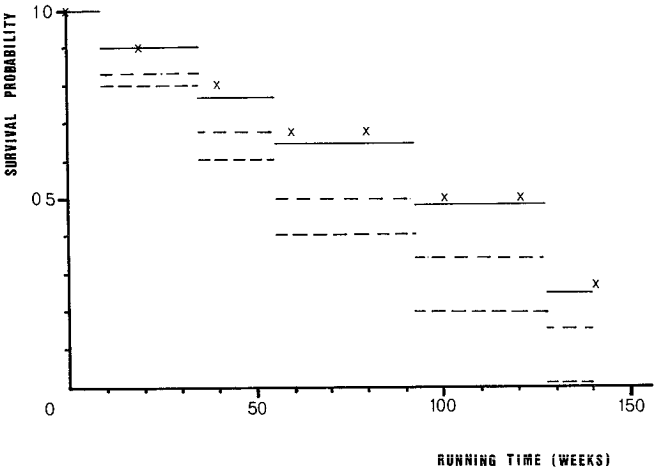

Fig. 1. Estimated survival probability functions. ————, PL method; - - - - - -, P_a; - · - · - · -, P_b; ×, actuarial method (values at end points of 20 week intervals).

Performing similar calculations for intervals 4 through 10 we arrive at the "interval" after the 139th week. Its length is indeterminate and we enter it with 0 units. Thus we reach the end of our estimation procedure.

We have found, in the last but one column of Table 2, an estimate of the (unconditional) survival probability function; no assumptions had to be made about its general form. The estimate made use of all the information contained in the data, up to every last observed unit of running time. Moreover, the estimates according to the PL method are virtually unbiased and have other agreeable statistical properties; for more details, see [1]. The gain achieved by allowing also the non-failed units to contribute to the estimate of the survival probability function is clearly demonstrated in Fig. 1. This figure shows for comparison two commonly encountered reliability estimates derived only from complete running times to failure i.e.

(a) $P_a(t_i) = 1 - i/n$,

(b) $P_b(t_i) = 1 - i/(n+1)$,

where t_i = time to failure of the ith failed unit; $t_0 = 0$; $i = 1, 2, \ldots, n$; n = total number of failed units in the sample (in our example n = 5); and $P_a(t_i)$, $P_b(t_i)$ = estimate of the survival probability between times t_{i-1} and t_i.

4. PRACTICAL USE OF THE METHOD

4.1. The exact PL method

The exact PL method is especially practicable in cases where relatively few data are available. For larger amounts of data and when the method is to be used more frequently, use of a computer may be considered. Some work can be eliminated by evaluating the estimates of the survival probability function only at the times of failure, because it turns out that the function changes only at those points. Perusal of Table 2 will demonstrate that to do this, it is sufficient to keep track of the number of failures and the number of losses from observation since time zero.

4.2. An approximation

When larger amounts of data are available it is common practice to start by making a table of the number of failures and "losses" occurring in successive equal time intervals;

TABLE 3

Calculation scheme, actuarial method

Interval (weeks)		No. of units			Probability of survival	
>	⩽	Entering	Lost, due to		During this interval	Up to and including this interval
			Failure	End of observation		
0	20	10	1	1	9/10	9/10
20	40	8	1	1	7/8	63/80
40	60	6	1	0	5/6	63/96
60	80	5	0	2	5/5	63/96
80	100	4	1	0	3/4	189/384
100	120	3	0	1	3/3	189/384
120	140	2	1	1	1/2	189/768
140	160	0				

often this is the way in which the data have been gathered to begin with. The scheme of calculations for an arbitrary interval length of 20 weeks is then according to Table 3, where implicitly the assumption is used that all failures have occurred at the start of the interval and all losses to observation at the end of the interval. This assumption clearly disregards some of the information available in the data.

Other negative aspects of this method are the arbitrary choice of the constant length of the time interval and its variable influence on the shape of the estimate of the function. However, the severity of these disadvantages diminishes as the length of the intervals decreases; in fact, this approximation to the PL method is used in the construction of actuarial life tables and has long been known in actuarial science.

As an illustration and comparison, the results of Table 3 have also been plotted in Fig. 1.

5. EXTENSIONS

In the case that has been treated in the previous sections the complication arose from the fact that for the non-failed units only lower bounds for their running times to failure were available. In practice the situation may be even more complicated. Consider a laboratory endurance test where some units were accidentally removed from the test before they had failed; also some of the recorders for the time to failure had failed to record that time for some failed units so that upon discovery of this fact the recorder reading is only an upper limit for the actual time to failure. In such a case we would have a number of exact times to failure, a number of lower limits of time to failure and a number of upper limits of time to failure. A method for estimating the survival probability function is available for this case also [3] and for even more complicated cases [4].

6. CONCLUSIONS

The PL method and its approximation appear to be more appropriate than the rank method in cases where one is not willing to assume more about the population survival probability function than its continuity. The methods described provide a more accurate and precise estimate of points on the survival probability function of a population than would be obtained when incomplete observa-

tions are discarded from a sample of running times. Also, although slightly more complicated they are particularly appropriate and useful when analyses are performed only occasionally.

When estimates of survival probability functions are required frequently, a different method, Hazard Plotting, should be considered. This also allows for the utilisation of incomplete observations; moreover, it yields information about possible forms of the function and estimates of their parameters. However, it does require a variety of special graph papers. For excellent presentations of the Hazard Plotting method see [5] or [6].

REFERENCES

1 Kaplan E.L. and Meier P., 1958. Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc., 53 (June 1958): 457–481.

2 Johnson L.A., 1964. Theory and Technique of Variation Research, Elsevier, Amsterdam.

3 Turnbull B.W., 1974. Nonparametric estimation of a survivorship function with doubly censored data. J. Am. Stat. Assoc., 69 (March 1974): 169–173.

4 Turnbull B.W., 1976. The empirical distribution function with arbitrarily grouped, censored and truncated data. J. R. Stat. Soc., Series B, 38(3): 290–295.

5 Nelson W., 1969. Hazard plotting for incomplete failure data. J. Qual. Technol., 1(1): 27–52.

6 Nelson W., 1972. Theory and applications of hazard plotting for censored failure data. Technometrics, 14(4): 945–966.