

APPLICATION BOOTSTRAPPING KAPLAN-MEIER ESTIMATE FOR SURVIVAL CURVE SMOOTHING

MICHAŁ MICHALAK, MAREK POWIDZKI, JERZY A. MOCZKO

*University of Medical Science, Chair and Department of Medical Informatics and Statistics,
Dąbrowskiego 79, 60-529 Poznań, Poland,
e-mail: michal@am.poznan.pl*

Abstract. In this article we wish to present and encourage the other to use bootstrap methods in statistical analysis. We show how to bootstrap Kaplan-Meier estimator and pay attention to its advantage opposite to classical analysis. Then we present simulation study and survival time of second remission of patients suffering for acute leukaemia.

1. INTRODUCTION

One of the most frequently used nonparametric method of estimating survival function is method of Kaplan-Meier.

In medical science we have very often to deal with small sample size. There are several reasons for that. The common one is rarity of illness or difficulty with gathering patients possessing the same biochemical parameters. Furthermore, we have very often censored data.

Therefore, a small sample size either does not let us use classical statistical methods or when they are used, they can give us too general and even false results. Using bootstrap methods can solve some of those problems. With computer simulation we can generate many samples based on original sample data and we can more accurately evaluate parameters determined on bootstrap distribution.

2. ACUTE LEUKAEMIA - SECOND REMISSION STUDY

Let us assume that we have 20 times (in years) of survival times patients with acute leukaemia in second remission and some of them are censored

* - Censored observation

Analysing our sample we get the following results of survival function $S(t)$:

Table I. Table of survival function $S(t)$

Time in years	Number of death cases	Number of censored cases	Risk set n	Probability	$\hat{S}(t)$
1	2	3	4	5	6
2.0	1	0	20	0.9500	0.9500
2.1	0	1	19	1.0000	0.9500
2.2	0	1	18	1.0000	0.9500

1	2	3	4	5	6
3.1	1	0	17	0.9412	0.8941
3.2	0	1	16	1.0000	0.8941
3.3	1	1	15	0.9333	0.8345
4.0	1	0	13	0.9231	0.7703
5.0	0	1	12	1.0000	0.7703
5.1	1	0	11	0.9091	0.7003
5.2	0	1	10	1.0000	0.7003
5.3	1	0	9	0.8889	0.6225
6.0	0	1	8	1.0000	0.6225
6.1	1	0	7	0.8571	0.5336
6.2	1	0	6	0.8333	0.4446
8.0	0	1	5	1.0000	0.4446
8.1	0	1	4	1.0000	0.4446
11.0	1	0	3	0.6667	0.2964
12.0	1	0	2	0.5000	0.1482
12.1	0	1	1	1.0000	0.1482

$\hat{S}(t)$ – estimator of survival function $S(t)$

In order to increase the accuracy of survival function estimator we use bootstrap method.

Let us create two random variables X and Y , where random variable X has the following empirical distribution:

Table II. Empirical distribution of X

x_i	p	x_i	p
2.0	0.05	5.3	0.05
2.1	0.05	6.0	0.05
2.2	0.05	6.1	0.05
3.1	0.05	6.2	0.05
3.2	0.05	8.0	0.05
3.3	0.1	8.1	0.05
4.0	0.05	11.0	0.05
5.0	0.05	12.0	0.05
5.1	0.05	12.1	0.05
5.2	0.05		

While random variable Y , which is a censoring indicator, takes two values 1 if the observation is full and 0 when it is censored (truncated):

Table III. Distribution of Y

y_i	1	0
p	0.5	0.5

We form ordered pairs (x_i, y_i) for $i = 1$ to 20, which are generated according to empirical distribution of random variables X and Y . We use bootstrap resampling method. There are many statistical software packages that offer bootstrap resampling. S-plus 6 for windows (Insightful Corporation) enables jackknife and bootstrap resampling and also calculating bootstrap statistics but they must be a scalar, vector or matrix. SAS offers general purpose jackknife and bootstrap capabilities via two macros available in the file jack-boot.sta at <http://ftp.sas.com/techsup/download/stat/> [1]. We did not find statistical package that provides bootstrap calculating for Kaplan-Meier estimate where outcome is a table. In order to simulate our data we prepared simple generating programme in Pascal 7.0 programming language. That way we receive N bootstrap samples and N Kaplan-Meier estimates. We assume that $N \geq 1000$.

For bootstrap estimator of survival function we take

$$\hat{S}_B(t) = \frac{1}{N} \sum_{j=1}^N \hat{S}(t)$$

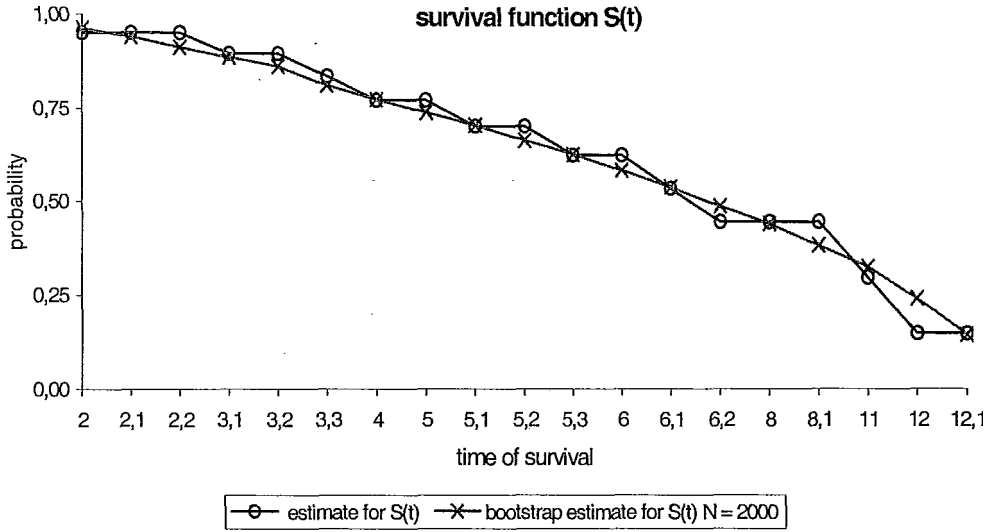
where t - considered month where t takes values:

2 2.1 2.2 3.1 3.2 3.3 3.3 4 5 5.1 5.2 5.3 6 6.1 6.2 8 8.1 11 12 12.1

As a result of $N = 2000$ simulations we get the following outcomes:

Table IV. Comparison $\hat{S}(t)$ and $\hat{S}_B(t)$

Time in years	$\hat{S}(t)$	$\hat{S}_B(t)$ $N = 2000$
2.0	0.9500	0.9618
2.1	0.9500	0.9400
2.2	0.9500	0.9117
3.1	0.8941	0.8844
3.2	0.8941	0.8600
3.3	0.8345	0.8096
4.0	0.7703	0.7723
5.0	0.7703	0.7377
5.1	0.7003	0.7025
5.2	0.7003	0.6628
5.3	0.6225	0.6245
6.0	0.6225	0.5819
6.1	0.5336	0.5372
6.2	0.4446	0.4864
8.0	0.4446	0.4396
8.1	0.4446	0.3811
11.0	0.2964	0.3249
12.0	0.1482	0.2409
12.1	0.1482	0.1446



As it is shown on the graph above, the application of bootstrap method smoothes our survival function.

Let us compare confidence intervals.

Classical confidence interval for Kaplan-Meier estimator is given by formula:

$$\hat{S}(t) - \hat{D}\{\hat{S}(t)\} \cdot u\left(\frac{\alpha}{2}\right); \quad \hat{S}(t) + \hat{D}\{\hat{S}(t)\} \cdot u\left(\frac{\alpha}{2}\right)$$

where $t(k) \leq t < t(k+1)$, $k = 1, 2, \dots, 18$ and $\hat{D}\{\hat{S}(t)\}$ is standard error of survival function and according to Greenwood [2] is given by formula:

$$\hat{D}\{\hat{S}(t)\} = [\hat{S}(t)] \sqrt{\sum_{j=1}^k \frac{d_j}{n_j(n_j - d_j)}} \quad \text{for } t(k) \leq t < t(k+1)$$

While $u(\alpha)$ is the quantile of standard normal distribution of the order of α .

When value of $\hat{S}(t)$ is close to zero or unity, the variance value $\hat{D}\{\hat{S}(t)\}$ may be overestimated and in that case we use formula [2]:

$$\hat{D}\{\hat{S}(t)\} = \frac{\hat{S}(t) \cdot \sqrt{1 - \hat{S}(t)}}{\sqrt{(n_k)}}$$

for $t(k) \leq t < t(k+1)$ where n_k is the number of cases with death risk in time $t(k)$.

For our sample we have the following 95% confidence intervals evaluation:

Table V. Standard errors and 95% confidence intervals for $S(t)$

Time in years	$\hat{S}(t)$	$\hat{D}\{\hat{S}(t)\}$	95% confidence intervals
2.0	0.9500	0.0487	(0.8545; 1.0000)
2.1	0.9500	0.0487	(0.8545; 1.0000)
2.2	0.9500	0.0487	(0.8545; 1.0000)
3.1	0.8941	0.0710	(0.7549; 1.0000)
3.2	0.8941	0.0710	(0.7549; 1.0000)
3.3	0.8345	0.0878	(0.6624; 1.0000)
4.0	0.7703	0.1018	(0.5707; 0.9699)
5.0	0.7703	0.1018	(0.5707; 0.9699)
5.1	0.7003	0.1142	(0.4766; 0.9240)
5.2	0.7003	0.1142	(0.4766; 0.9240)
5.3	0.6225	0.1252	(0.3771; 0.8679)
6.0	0.6225	0.1252	(0.3771; 0.8679)
6.1	0.5336	0.1353	(0.2684; 0.7987)
6.2	0.4446	0.1389	(0.1724; 0.7169)
8.0	0.4446	0.1389	(0.1724; 0.7169)
8.1	0.4446	0.1389	(0.1724; 0.7169)
11.0	0.2964	0.1524	(0.0000; 0.5951)
12.0	0.1482	0.1296	(0.0000; 0.4022)
12.1	0.1482	0.1296	(0.0000; 0.4022)

In order to evaluate bootstrap confidence intervals we use the percentile method. We take the empirical distribution of random variable $\hat{S}_B(t)$ received during N -fold generating of 20 elements sample of ordered pairs (x_i, y_i) which have (X, Y) distribution. Obtained sequences of evaluations $\hat{S}_B(t)$ for

may be used for percentile evaluation.

Let $\hat{S}_B^{(\alpha)}(t)$ denotes the α percentile. Sequence of evaluations $\hat{S}_B(t)$ is organised non-decreasing order. Percentile $\hat{S}_B^{(\alpha)}(t)$ is the value that stands on αN_1 position (N_1 number of evaluation $\hat{S}_B(t)$) for given year provided that this is integral number, in other case the position of percentile is determined by $[(N_1 + 1) \cdot \alpha]$ for $0 < \alpha \leq 0.5$ and $N_1 + 1 - [(N_1 + 1) \cdot (1 - \alpha)]$ for $0.5 < \alpha < 1$ [3-5]. We should point out that number N_1 of evaluations $\hat{S}_B(t)$ is different each month and number of simulations N . In our samples particular times of death may repeat or don't occur at all with the probability that is specified by random variable X .

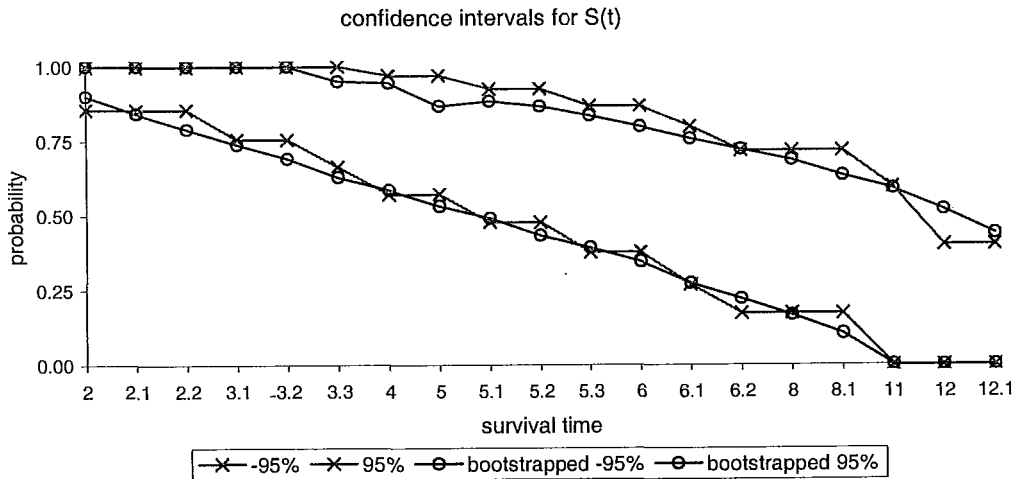
Confidence Interval for $S(t)$ obtained using this method is as follows:

$$P\left(\hat{S}_B^{(\alpha)}(t) < S(t) < \hat{S}_B^{(1-\alpha)}(t)\right) \approx 1 - 2\alpha$$

where $\hat{S}_B^{(\alpha)}(t)$ is α percentile and $\hat{S}_B^{(1-\alpha)}(t)$ is $1 - \alpha$ percentile.

Table VI. 95% bootstrap confidence intervals for $S(t)$

Time in years	95% bootstrap confidence interval for $S(t)$
2.0	(0.9000; 1.0000)
2.1	(0.8421; 1.0000)
2.2	(0.7895; 1.0000)
3.1	(0.7374; 1.0000)
3.2	(0.6906; 1.0000)
3.3	(0.6272; 0.9500)
4.0	(0.5844; 0.9444)
5.0	(0.5304; 0.8972)
5.1	(0.4897; 0.8824)
5.2	(0.4327; 0.8657)
5.3	(0.3914; 0.8345)
6.0	(0.3445; 0.7972)
6.1	(0.2722; 0.7566)
6.2	(0.2210; 0.7205)
8.0	(0.1667; 0.6858)
8.1	(0.1034; 0.6335)
11.0	(0.0000; 0.5891)
12.0	(0.0000; 0.5196)
12.1	(0.0000; 0.4377)



3. CONCLUSION

Using traditional method probability of survival for censored cases remains the level of previous time for uncensored data. Using bootstrap method we obtain both censored and uncensored cases for each time. That is why our survival curve is smoother and we don't have such a rapid jumps for probability of survival. Furthermore bootstrapping denotes confidence intervals, which have the same range. For some observations the range of interval is even smaller. In our study confidence intervals for time of 2 years and from 3.3 to 11.0 years are smaller from obtained in traditional way. This method is useful especially when dealing with small sample observations and helps us to estimate more reliable outcomes.

References

- [1] J. Carpenter, J. Bithell, *Statist. Med.*, **19**, 1141-1164 (2000).
- [2] D. Collett, *Modelling Survival Data in Medical Research*, Chapman & Hall 1994.
- [3] C. Domański, K. Pruska, W. Wagner, *Wnioskowanie statystyczne przy nieklasycznych założeniach*, Wyd. Uniw. Łódzkiego 1998.
- [4] B. Efron, *J. Amer. Statist. Assoc.*, **82**, 171-200 (1987).
- [5] B. Efron, R. J. Tibshirani, *An introduction to the bootstrap*. Chapman & Hall, London 1993.