THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

Kaplan-Meier Approach to Censored Data in Clinical Trials

Eric Boahen
Department of Statistics, Faculty of Mathematical Sciences
University for Development Studies, Tamale, Ghana
Richard Puubalanta
Department of Statistics, Faculty of Mathematical Sciences
University for Development Studies, Tamale, Ghana

Abstract
In clinical studies, patients survival time are subject to paucity of patients who report for treatment. For this reason, clinical data are mostly censored. Estimating mean, variance and standard deviation in censored data are different from those in uncensored data. Censored data sometime do not follow a specific distribution, hence using parametric method in estimating parameters in censored data calls for specific assumptions. KM method is a nonparametric approach that does not depend on any assumption. KM approach also known as product limit approach has been shown to provide more robust estimate of the mean and standard deviation for both left and right censored observations than other methods. In this paper, the method has been used to estimate mean survival time of patients in multiple censored data.

Keywords: *Product-limit, censored, substitution, imputation, clinical trial and mean*

1. Introduction

In clinical trial, as well as censored data, Kaplan – Meier method in estimating parameters is the most widely used. This technique is popularly referred to as product limit. This method had been in use for many years before Kaplan- Meier gave a theoretical justification that , the method behaves as non-parametric maximum likelihood estimate that makes no distributional assumptions about the population. In the Kaplan –Meier (1958) approach, ordered observations are used instead of grouped data. This method has the advantage of yielding results that are not dependent on the choice of the event time interval. To define the product limit $\hat{p}(t)$, suppose $t_1 \leq t_2 \leq \ . \ . \ .t_i$ are the time to survive in particular clinical trial and n_i is the number of patients still under observation at time t_i if a death and a loss occur at the same time in this time frame, the censored observations are recorded. In a situation when clinical data contains censored values, Kaplan – Meier estimator is simple to operate and the product limit

$$\hat{p}(t) = \prod_{j=1}^{i} \frac{n_j - 1}{n_j} \qquad (1)$$

reduces to the ordinary binomial estimate. This is because the product limit function $p(t)$ is the probability that an event time is greater than t, where t can be any nonnegative number. In case there is no censoring, the Kaplan- Meier product estimator $\hat{p}(t)$ is just the sample proportion of observations with event time greater than t. KM estimator is quite simple in the case of single right censoring, that is when all censored cases are censored at the same time a and all observed event time are less than a. when this happen, $\iota \leq a$. $\hat{p}(\iota)$ is still the sample proportion of observation with event time greater than ι. For $\iota \geq a$, $\hat{p}(\iota)$ is undefined. A more complex and complicated situation is when there is censoring and all the censored evnts are less than event time. This is because, knowledge about the event is not known and information or covariates needed for proper estimation is unavailable. In this case using only the observed proportion of cases greater than t downward because those cases are not known.

Another Kaplan –Meier estimators can be obtained using software for the more typical right censored survival data. This is done by flipping the data by subtracting all values, from a large constant. The general idea behind this is to convert all lift censored to right censored so that software can conveniently be used to estimate the parameters. This is possible because most of the implementations of the right censored KM estimators are designed for values that correspond to lifetimes; extreme values are liable to occur. The constant (C) should be larger than the largest value in the data so that all flipped observations can be maintained positive. However, the problem with flipping on censored data is that, flipped estimates are truly biased on small samples. This problem is that, the cumulative distribution function (cdf) of left censored observation is defined as $\hat{p}(t) = p[T \leq t]$. Since flipping is followed by estimation of parameters and thereafter followed by a flip back, it should be

noted that the flipped estimates no longer take the form $p[T \leq t]$ but correspond to $p[T < t]$. This is because the flipping does not take into account the event time t or censored time. The probability of an event $p[T = t]$ in the sample is compromised and put in a wrong place (Gillespie et al., 2010). Kaplan-Meier estimators are very useful in situations where the smallest value is uncensored. In situation where the smallest values are also censored values, product- limit $\hat{p}(t)$ is undefined. This is because, the probability mass associated with a censored observation is redistributed onto the smallest uncensored value as indicated in equation (1). In singly left censored, there is no smallest uncensored value below event time t and hence Kaplan –Meier estimators are usually not appropriate good estimators. Kaplan –Meier estimators are more prominent in multiple censoring. In view of this, treating singly censored clinical trial data with KM estimators or product limit method is equivalent to substitution as an imputation technique.

If t_{i-1} and t_i are two consecutive times to death then

$$\hat{p}(t) = \hat{p}(t_{i\ 1})(\frac{n_j - 1}{n_j}) \quad (2)$$

holds if there is no lost between the interval t_{i-1} and t_i

$$var[\hat{p}(\iota)] - [\hat{p}(\iota)]^2 \Sigma_{i=1}^{n} \frac{1}{n_i(n_{i-1})} \quad (3)$$

$$\hat{\sigma}(\hat{p}(t)) = \sqrt{[\hat{p}(t)]^2 \Sigma_{i=1}^{n} \frac{1}{n_i(n_{i-1})}} \quad (4)$$

At the ordered time to death $t_1, < t_2, <, \ . \ . < t_n$

$$\hat{\mu} = 1.0t_1^{'} + p(t_1^{'})(t_2^{'} - t_1^{'}) + \ . \ . \ . \ p(t_{n-1}^{'})(t_n^{'} - t_{n-1}^{'}) \quad (5)$$

$$\hat{v}(\hat{\mu}) = \Sigma_{i=1}^{m} \frac{k_i^2}{(n-i)(n-i+1)} \quad (6)$$

where m is the number of uncensored values and k_i is the cumulative of $\hat{p}(t)$, but

$$A_i = p(t_{m-1}^{'})(t_m^{'} - t_{m-1}^{'}) + \ . \ . \ , + p((t_{n-1}^{'})(t_n^{'} - t_{n-1}^{'}) \quad (7)$$

KAPLAN –MEIER APPLICATION

In a clinical trial in which eight patients are observed to have the following survival pattern in months 2 , $2^+, 3, 4, 4^+, 5, 6, 7.$ where 2^+ and 4^+ are individuals lost to follow-up at 2 and 4 months respectively. Applying the product limit,

$$\hat{p}(2) = \left(\frac{8-1}{8}\right) = 0.875$$

$$\hat{p}(3) - \left(\frac{8-1}{8}\right)\left(\frac{6-1}{6}\right) - 0.729$$

$$\hat{p}(4) = \left(\frac{8-1}{8}\right)\left(\frac{6-1}{6}\right)\left(\frac{5-1}{5}\right) = 0.583$$

$$\hat{p}(5) = \left(\frac{8-1}{8}\right)\left(\frac{6-1}{6}\right)\left(\frac{5-1}{5}\right)\left(\frac{3-1}{3}\right) = 0.389$$

$$\hat{p}(6) - \left(\frac{8-1}{8}\right)\left(\frac{6-1}{6}\right)\left(\frac{5-1}{5}\right)\left(\frac{3-1}{3}\right)\left(\frac{2-1}{2}\right) - 0.194$$

$$\hat{p}(7) = 0$$

$$\hat{V}ar[\hat{p}(2)] = (0.875)^2\left[\frac{1}{7\times8}\right] = 0.01367$$

$$\hat{V}ar[\hat{p}(3)] = (0.729)^2\left[\frac{1}{7\times8} + \frac{1}{5\times6}\right] = 0.0272$$

$$\hat{V}ar[\hat{p}(4)] = (0.583)^2\left[\frac{1}{7\times8} + \frac{1}{5\times6} + \frac{1}{4\times5}\right] = 0.0344$$

$$\hat{V}ar[\hat{p}(5)] = (0.389)^2\left[\frac{1}{7\times8} + \frac{1}{5\times6} + \frac{1}{4\times5} + \frac{1}{3\times2}\right] = 0.0405$$

$$\hat{V}ar[\hat{p}(6)] = (0.194)^2\left[\frac{1}{7\times8} + \frac{1}{5\times6} + \frac{1}{4\times5} + \frac{1}{3\times2} + \frac{1}{2\times1}\right] = 0.0376$$

$$\hat{V}ar[\hat{p}(7)] = 0$$

To compute an estimate of mean survival time, the time must be conform to order statistics. The time to death for the patients are $\iota_1, \leq \iota_2, \leq \iota_3 \leq, \ . \ . \ . \ \leq \iota_d$, it is shown by Kaplan –Meier that the estimated mean survival time is

$$\hat{\mu} = [1.00t_1 + p(t_1)(t_2 - t_1) + \ . \ . \ . \ + p(t_{d-1})(t_d - t_{d-1})] \quad (8)$$

$\hat{\mu} - (1.00)(2) + (0.875)(3-2) + (0.729)(4-3) + (0.583)(5-4) + (0.389)(6-5) + (0.194)(7-6) = 4.770$ months. This is the estimated mean survival time for the patients.

The estimate of the variance of $\hat{\mu}$, $\hat{v}(\hat{\mu})$ is

$$\hat{v}(\hat{\mu}) = \sum_r \frac{k_r^2}{(n-r)(n-r+1)} \quad (9)$$

Here r is summed over those integers for which t_r reflect the death.

Then. $k_r = p(t_{r-1})(t_r - t_{r-1}) + \ . \ . \ . + p(t_{d-1})(t_d - t_{d-1})$.

$k_7 = 0.389 + 0.194 = 0.583$

$k_6 = 0.583 + 0.389 + 0.194 = 1.166$

$k_4 = 0.729 + 0.583 + 0.389 + 0.194 = 1.895$

$k_3 = 0.875 + 0.729 + 0.583 + 0.389 + 0.194 = 2.770$

$k_1 = 2.00 + 0.875 + 0.729 + 0.583 + 0.389 + 0.194 = 4.770$

$\hat{v}(\hat{\mu}) = \frac{(4.770)^2}{7\times8} + \frac{(2.770)^2}{5\times6} + \frac{(1.895)^2}{4\times5} + \frac{(1.166)^2}{2\times3} + \frac{(0.583)^2}{1\times2} = 1.2382$

The standard deviation is $\sqrt{1.2382}$ =1.1127 months

Method	Mean	Variance	Standard Deviation	Standard Error Of The Mean
Kaplan-Meier	4.770	1.2382	1.1127	0.3940
Mean imputation	4.566	2.2011	1.4836	0.5245
Substitution with zero	3.375	5.9844	2.4463	0.8649

Table 1: KM Estimator Compared with Other Methods

2. Conclusion

From table 1, the KM mean of 4.770 months is seen as the maximization of the mean survival time of patients as compare to mean imputation and zero substitution. This is seen in the standard errors of the means. The standard error of the mean for KM estimate of 0.3940 is minimum as compare to that of mean imputation and zero substitution. In this case, the survival time of the patients from KM estimate is very close to the true mean survival time, even though part of the data is censored the available known information has been accounted for. It is seen from the paper that substitution in all forms including imputation distort censored information and consequently lead to bias. It is not surprising that Helsel (1990) and other publications consistently shown that using substitution and any other form of imputation in censored observations unnecessary introduce bias in summary statistics.

3. References

1. Kaplan, E. L and O. Meier, 1958 nonparametric estimation from incomplete observations Journal of the American statistical association vol. 53, 457-481
2. Armitage, P. 1959 The comparison of survival curves Journal of the Royal Statistical Society, vol.A122, 279-300
3. Berkson J. and R. P Gage 1952 Survival curve for cancer patients following treatment Journal of the American Statistical Association vol. 47, 501-515