Tian, Maoshan; Dixon, Huw

Working Paper

The cross-sectional distribution of completed lifetimes: Some new inferences from survival analysis

Cardiff Economics Working Papers, No. E2018/27

Provided in Cooperation with:
Cardiff Business School, Cardiff University

Suggested Citation: Tian, Maoshan; Dixon, Huw (2018) : The cross-sectional distribution of completed lifetimes: Some new inferences from survival analysis, Cardiff Economics Working Papers, No. E2018/27, Cardiff University, Cardiff Business School, Cardiff

This Version is available at:
https://hdl.handle.net/10419/230433

Cardiff Economics Working Papers

Working Paper No. E2018/27

The Cross-sectional Distribution of Completed Lifetimes: Some New Inferences from Survival Analysis

Maoshan Tian and Huw Dixon

December 2018

ISSN 1749-6010

Cardiff Business School
Cardiff University
Colum Drive
Cardiff CF10 3EU
United Kingdom
t: +44 (0)29 2087 4000
f: +44 (0)29 2087 4419
business.cardiff.ac.uk

The Cross-sectional Distribution of Completed Lifetimes: Some New Inferences from the Survival Analysis

Maoshan Tian[1] and Huw Dixon[2‡]

Abstract

The cross-sectional distribution of completed lifetimes (DCL) is a new estimator defined and derived by Dixon (2012) in the general Taylor price model (GTE). DCL can be known as the cross-sectional weighted estimator summing to 1. It is a new statistics applying to describe the data. This paper focuses on the cross-sectional distribution in the survival analysis. The delta method is applied to derive the variance of the of three cumulative distribution functions: the distribution of duration, cross-sectional distribution of age, distribution of duration across firms. The Monte Carlo experiment is applied to do the simulation study. The empirical results show that the asymptotic variance formula of the DCL and distribution of duration performs well when the sample size above 25. With the increasing of the sample size, the bias of the variance is reduced.

JEL Codes: C19, C46
Keywords: Delta Method, Survival Analysis, Kaplan-Meier Estimator

[*1]M. Tian is a PhD candidate in Cardiff Business School, University of Cardiff TianM@cardiff.ac.uk

[†2]H. Dixon is the professor in Cardiff Business School, University of Cardiff dixonh@cardiff.ac.uk

1 Introduction and Literature Review

The best known non-parametric estimator of the survival function was derived by Kaplan and Meier (1958). As Gillesple and Fisher (1979) explained, Kaplan-Meier estimator is the limit of the life table method to calculate the survival function with the increasing of the time intervals thereby tending to be zero. Kaplan and Meier (1958) assumed that lifetime time (the existence of the events or the age of death) was independent with the failure time (hazard rate) and derived the product limit estimator. They derived the variance of the survival function from an alternative way and obtained the same result as Greenwood (1926)'s formula. They also showed the variance formula of the survival function in the large sample size. In addition, the product limit estimators were the consistent estimators. On the other hand, Nelson (1972) applied a graphical method to investigate the failure ratio (hazard rate) rather than the survival ratio. This graphical method was named as "hazard plotting". It plotted the hazard rate and the cumulative hazard rate depending on the distribution of the hazard function. After that, Aalen (1978) investigated the hazard function and the cumulative hazard function from the theoretical part. The counting process theory was applied to derive the cumulative hazard function. Since both Nelson and Aalen derived the cumulative hazard function, there existed the new estimator named as the Nelson-Aalen estimator. Nelson-Aalen estimator was the cumulative hazard function. For the asymptotic properties of KM and NA estimators, see Andersen et al. (1993), Fleming and Harrington (1991), Kalbfleisch and Prentice (2002), Fleming and Harrington (1991), Bohoris (1994) and Colosimo et al. (2002). In terms of the parametric method for estimating the survival function and the hazard function, Cox (1972) derived an exponential function to regress the hazard rate. It assumed that there existed some measurements for each individual. To investigate the null hypothesis test whether two group has the same survival rate depending on the log-rank test, see Mantel (1966). Breslow and Crowley (1974) investigated the life table and the Greenwood formula under the large sample conditions. They also derived the asymptotic normality for the standard life estimators. The estimators were assigned into the vector form which converged to the multivariate normal distribution. The covariance formula for the survival function and cumulative hazard function were derived under large sample conditions. The confidence interval of the KM estimator was introduced by Gillesple and Fisher (1979), Nair (1981), Nair (1984) and Kalbfleisch and Prentice (2002). Kalbfleisch and Prentice

(2002) provided a method called log-log transformation method to guarantee the positive lower bound of the confidence interval of KM estimator.

Both parametric and non-parametric methods of estimation have been well studied and we know the variances of estimators and related statistical properties. In this paper, we want to derive the variances of the three related distributions, two of which are cross-sectional. The age distribution, which gives the proportion of observations at a point in time which have particular ages; the cross-sectional distribution of completed lifetimes which gives the proportion of observations at a point in time which will have a completed lifetime of a particular duration; the distribution of duration, the proportion of observations over the whole period which has completed lifetimes of a particular duration. Our analysis is applicable to a panel of observations, where we observe many agents (people, households, firms, machines) repeatedly over time and also to situations where we observe just a few or even one agent over time. Our framework is one of discrete time, although the analysis easily carries over to continuous time representations.

Suppose we divide time into discrete periods: days, weeks, months and so on. In economic applications, this will often be driven by the data we have. The survival function gives the probability that an event will last for more than i periods, S_i. Clearly, $S_i \in [0, 1]$ and $S_i \geq S_{i+m}$ for $m > 0$. The corresponding hazard function h_i gives the conditional probability that having survived i periods, the event ends (death or failure). There are two classic methods of estimating this process. The Kaplan-Meier estimator (KM) estimates the survival function, whilst the Nelson-Aalen estimator (NA) estimates the cumulative hazard function. The properties of both of these estimators have been well studied and in particular their asymptotic variances. Whilst both KM and NA are general non-parametric estimators, they can also be estimated in parametric forms, such as the Cox proportional hazard model.

Starting from the KM and NA estimators, we are able to construct estimators of the three distributions (durations, ages and completed lifetimes) and to derive their asymptotic variances using both the Taylor expansion and delta methods. Theorem 1 derives the asymptotic variance of the distribution of durations. Theorem 2 and its corollary derive the asymptotic variances for the two cross-sectional distributions. In the simulation part, the Monte Carlo method was applied to explore the performance of the estimators and their sensitivity to censored of observations. We find that whilst there can be small bias for samples as small as 25, for samples 50 or over there is almost

no bias and the results are not sensitive to censored for samples of 50 or over.

The three distributions we estimate we believe may have many useful applications. In economics, the estimated cross-sectional distribution of completed durations can be used to calibrate the Generalized Taylor Model of heterogeneous price and/or wage setting in a macroeconomic setting(See Taylor (1980); Coenen et al. (2008); Dixon and Bihan (2012)). In this setting, the cross-sectional distribution gives the proportion of price or wage setters in the economy who set prices or wages for a particular period of time. However, in demographics, it also gives the distribution of completed lifetimes for those living at a point in time. Whilst it is common to calculate the life-expectancy from life tables, our estimates enable the complete distribution of lifetimes to be estimated. Also, if we are looking at the stock of something at a point in time (unemployed workers, people living in an area, or machines), we can generate the distribution of completed durations (when the unemployed find a job, when people move from an area, when machines will fail). The estimation method can be non-parametric or parametric. The distribution of durations is useful when we want to look at the population over an extended period of time: the distribution of spells of unemployment, the distribution of price spells, the distribution of periods before machines have their first fault and so on.

2 Survival Function and Hazard Function

Kaplan and Meier (1958) provided an estimator for the survival function, the Kaplan-Meier estimators $S_i \in [0,1], i = 1, 2, ...F$, where F is the maximum duration (this can be arbitrarily large, or may have an obvious empirical value such as the length of the dataset). We can imagine that there is a panel of agents. A spell of time is a period when the agent remains in the same state (remains alive, remains ill, sets the same price). *Failure* occurs when that state changes (death, recovery from illness, price changes). When the state changes, this can either be seen as the same agent continuing in the different states (the firm continues but sets the different price) or a new agent replaces the old (the machine fails and is replaced with a new machine).

If we look across the entire data set, we can count the number of spells that last at least k periods as N_k, and the number of failures in the $k-th$ period is D_k. N_0 is the total number of price spells in the sample. The

Kaplan-Meier estimator $\hat{S}_i$ of the survival function S_i can be written as:

$$\hat{S}_i = \prod_{k=1}^{i} \frac{N_k - D_k}{N_k} \tag{1}$$

S_i can be defined as the proportion of spells remaining at $i-th$ period. This formula is also known as the product limited estimator for the survival function. There exists another way to describe the survival function, the *cumulative hazard function* which can be estimated by the Nelson-Aalen estimator. The formula is:

$$\hat{H}_i = \sum_{k=1}^{i} \frac{D_k}{N_k} \tag{2}$$

The cumulative hazard function H_i is the summation of the hazard rate in each period until i period. The (marginal) hazard function can be defined as the proportion of failures amongst spells that have lasted i periods :

$$\hat{h}_i = \frac{D_i}{N_i} \tag{3}$$

By convention, we set $S_0 = 1$ and D_0 and $h_0 = 0$ are equal to zero since all spells last at least 0 periods. Since F is the longest spell, $h_F = 1$ and $S_F = 0$ under the uncensored case. The hazard function can be transformed to the survival function:

$$\hat{S}_i = \prod_{k=1}^{i} (1 - \hat{h}_k) \tag{4}$$

Likewise, the survival function can be transformed into the hazard function:

$$\hat{h}_i = \frac{\hat{S}_{i-1} - \hat{S}_i}{\hat{S}_{i-1}}$$

Before deriving our three distributions, we need to define an additional variable $\bar{h}$:

$$\bar{h} = \frac{1}{\sum_{i=0}^{F} \hat{S}_i} \tag{5}$$

$\bar{h}$ is the reciprocal of the sum of survival probabilities. Intuitively, in a balanced panel, $\bar{h}$ is the proportion of agents that fail every period. To see this, consider some simple examples. First, failures occur in the first period

for all spells. In this case, $S_0 = 1$ and $S_i = 0$[1] for all $i > 0$. All spells last for one period, and $\bar{h} = 1$. Second the example where all spells last for two periods and then fail. In this case we have $S_0 = S_1 = 1, S_i = 0$ for $i \geq 2$. In this case $\bar{h} = 1/2 : 50\%$ of spells fail per period. Hence, with a balanced panel, we can think of $\bar{h}$ as being the proportion of failures each period.

However, if we just have one cohort, we can think of $\bar{h}$ as being the weighted average hazard over $i = 1..F$, where the weights are the proportions surviving to period i divided by the sum of survival probabilities (to ensure the weights add up to 1).

Proposition 1. $\bar{h} = \frac{1}{\sum_{i=0}^{F} \hat{S}_i} \sum_{i=0}^{F-1} \hat{S}_i \hat{h}_{i+1} = \frac{1}{\sum_{i=0}^{F} \hat{S}_i}$

To see why,

$$\begin{aligned}
\bar{h} &= \frac{1}{\sum_{i=0}^{F} \hat{S}_i} \sum_{i=0}^{F-1} \hat{S}_i \hat{h}_{i+1} \\
&= \frac{1}{\sum_{i=0}^{F} \hat{S}_i} \left(\frac{\hat{S}_0 - \hat{S}_1}{\hat{S}_0} + \hat{S}_1 \left(\frac{\hat{S}_1 - \hat{S}_2}{\hat{S}_1} \right) + ...\hat{S}_{F-1} \right) \\
&= \frac{1}{\sum_{i=0}^{F} \hat{S}_i} \left(1 - \hat{S}_1 + \left(\hat{S}_1 - \hat{S}_2 \right) + \left(\hat{S}_2 - \hat{S}_3 \right) + ..(\hat{S}_{F-2} - \hat{S}_{F-1}) + \hat{S}_{F-1} \right) \\
&= \frac{1}{\sum_{i=0}^{F} \hat{S}_i}
\end{aligned}$$

2.1 The Distribution of Durations

The distribution of durations gives the proportion of spells that survive at least $i-1$ periods and change (or "die") in the $i-th$ period. This is sometimes called the unconditional hazard function. The proportion of spells lasting exactly i periods can be defined as:

$$\hat{a}_i^d = \hat{S}_{i-1} \hat{h}_i \tag{6}$$

The product limit estimator of the distribution of duration can be written as:

$$\hat{a}_i^d = \frac{D_i}{N_i} \prod_{k=0}^{i-1} \left(\frac{N_k - D_k}{N_k} \right) \tag{7}$$

[1] It should be mention that S_F equal to zero in the uncensored case; but S_F may not equal to zero when the censored problem is considered

Clearly, $\hat{a}_i^d > 0$ and for $F > 1$, $1 > \hat{a}_i^d$. Note also that

$$\sum_{i=1}^{F} \hat{a}_i^d = 1$$

since

$$\sum_{i=1}^{F} \hat{S}_{i-1}\hat{h}_i = \sum_{i=1}^{F} \left(\hat{S}_{i-1} - \hat{S}_i\right) = \hat{S}_0 = 1$$

The distribution of durations can be treated as applying a particular cohort starting within a specific time frame (as with life tables), or as the distribution of all spells over a possibly long period (as in a balanced panel).

2.2 The Age Distribution

The age distribution can be explained as the ratio between the survival function of the price at the time i as:

$$\hat{a}_i^A = \frac{\hat{S}_i}{\sum_{k=1}^{F} \hat{S}_k} = \hat{S}_i \bar{h} \tag{8}$$

Since the survival function is non-increasing, the age distribution is non-increasing: $\hat{a}_i^A \geq \hat{a}_{i+1}^A$. In addition, it is clear that the summation of the age distribution is equal to 1.

In the case of a balanced panel, we can think of the age distribution as being the cross-sectional distribution of ages across agents at a point in time. However, for a particular cohort, we can also think of it as the proportions of spells from that cohort lasting at least a particular length. This differs from the survival function because the proportions add up to unity (being the survival function pre-multiplied by $\bar{h}$). The survival function does not add up to unity because the events captured are not mutually exclusive. The sum of survival probabilities will exceed one unless all spells last just one period.

2.3 The Distribution of Completed Lifetimes

Next, we introduce the less familiar cross-sectional distribution of the complete lifetimes (DCL) across agents. The new distribution is derived by

Dixon (2012). The DCL can be written as:

$$\hat{a}_i = i\bar{h}\hat{S}_{i-1}\hat{h}_i \tag{9}$$

Therefore, the product limited estimator of the DCL can be written as:

$$\hat{a}_i = i\frac{\hat{S}_{i-1}\hat{h}_i}{\left(\sum_{i=1}^{F}\hat{S}_{i-1}\right)} \tag{10}$$

Clearly, for $i = 1..F,\ 1 > \hat{a}_i > 0$. For $F = 1$, $\hat{a}_1 = 1$.

Proposition 2. $\sum_{i=1}^{F}\hat{a}_i = 1$.

To see why Proposition 2 holds, note that

$$\begin{aligned}
\sum_{i=1}^{F}\hat{a}_i &= \bar{h}\sum_{i=1}^{F} i\hat{S}_{i-1}\hat{h}_i \\
&= \bar{h}\sum_{i=1}^{F} i\left(\hat{S}_{i-1} - \hat{S}_i\right) \\
&= \bar{h}\left[\sum_{i=1}^{F}\left(\hat{S}_{i-1} - \hat{S}_i\right) + \sum_{i=2}^{F}\left(\hat{S}_{i-1} - \hat{S}_i\right) + .. \sum_{i=j}^{F}\left(\hat{S}_{i-1} - \hat{S}_i\right) + \hat{S}_{F-1}\right] \\
&= \bar{h}\left[\hat{S}_0 + \hat{S}_1 + \hat{S}_2 .. + \hat{S}_F\right] \\
&= 1
\end{aligned}$$

If we have a balanced panel, we can think of this as the cross-sectional distribution of completed lifetimes. In the case of a single cohort, we can think of DCL as being the distribution of completed lifetimes where we take an observation over each of the F periods. In the first period, we have all of the spells. In the second period, the one-period spells drop out and we have the spells with a duration of 2 and above and so on. Hence the i period contracts will be counted i times. Thus the distribution of the completed lifetimes for the cohort is given by the distribution of durations $\hat{a}_i^d$, the DCL is given by $a_i = \bar{h}i\hat{a}_i^d$. In effect, we can think of the DCL as weighting the spells by their length, which as was suggested by Baharad and Eden (2004).

2.4 The Three Distributions

The survival function, hazard function, and the three distributions are different ways of describing the same data. They are all linked by identities. These

Table 1: Relationships among different distributions

	S_i	h_i	a_i^d	a_i^A	a_i
S_i	I	$\prod_{j=1}^{i}(1-h_j)$ for $i=1..F.$	$1-\sum_{j=1}^{i} a_j^d$	$\frac{a_{i+1}^A}{a_1^A}$	$1-\frac{1}{\sum_{k=1}^{F}\frac{a_k}{k}}\sum_{j=1}^{i}\frac{a_j}{j}$
h_i	$\frac{S_{i-1}-S_i}{S_{i-1}}$	I	$\prod_{j=1}^{i-1}(1-h_j)$	$\frac{a_i^A-a_{i+1}^A}{a_1^A}$	$\frac{a_i}{i}\left[\sum_{i=1}^{F}\frac{a_i}{i}\right]^{-1}$
a_i^d	$S_{i-1}-S_i$	$h_i\prod_{j=0}^{i-1}(1-h_j)$	I	$\frac{a_i^A-a_{i+1}^A}{a_1^A}$	$\frac{a_i}{i\sum_{j=1}^{F}\frac{a_j}{j}}$
a_i^A	$\left[\sum_{i=0}^{F-1}S_i\right]^{-1}S_{i-1}$	$\left[\sum_{i=1}^{F}\prod_{j=0}^{i-1}(1-h_j)\right]^{-1}\prod_{j=0}^{i-1}(1-h_j)$	$\frac{1-\sum_{j=1}^{i-1}a_j^d}{\sum_{i=1}^{F}ia_i^d}$	I	$\sum_{k=1}^{F}\frac{a_k}{k}-\sum_{j=1}^{i}\frac{a_j}{j}$
a_i	$i\left[\sum_{i=0}^{F-1}S_i\right]^{-1}(S_{i-1}-S_i)$	$i\prod_{j=1}^{i-1}(1-h_j)\,h_i\left[\sum_{i=1}^{F}\prod_{j=0}^{i-1}(1-h_j)\right]^{-1}$	$\frac{a_i^d}{i.\sum_{j=1}^{F}ja_j^d}$	$i.\left(a_i^A-a_{i+1}^A\right)$	I
$\bar{h}$	$\left[\sum_{i=0}^{F-1}S_i\right]^{-1}$	$\left[\sum_{i=1}^{F}\prod_{j=0}^{i-1}(1-h_j)\right]^{-1}$	$\sum_{i=1}^{F}ia_i^d$	a_1^A	$\sum_{i=1}^{F}\frac{a_i}{i}$

identities hold for the estimators as well. If we take a particular survival function, then we can express the hazard function and the three distributions in terms of the survival function. Likewise, if we pick a particular hazard function, we can express the survival function and all three distributions in terms of the particular hazard function.

The full set of relationships is given in table(1). Each column represents the basic function: $\left\{S_i, h_i, a_i^d, a_i^A, a_i\right\}$: each row shows how the element can be written in terms of the element of that column. Thus the first row has the different ways of writing the survival probability S_i in terms of itself (the indicator I), the hazard function h_i, and then the three distributions. The last two expresses the key statistic $\bar{h}$ in terms of all the functions. Note that these identities apply to any and all possible functions. The identities also apply to the estimators if they are unbiased: the estimated values must belong to the set of possible values.

3 Asymptotic Variances of the New Statistics

There are two equivalent ways to derive the variance formulas for the three distributions. The first method is the (multivariate) delta method which can be combined with the first order Taylor expansion to derive the variance of the cumulative functions. Another one is the counting process theory combined with the statistical method to derive the continuous-time version of the variances of the distribution functions. First, we will use the delta method since it fits more easily with the discrete time framework. This method is also introduced by Greenwood (1926) to derive the variance of the survival function. The delta method gives more information with the higher order terms and converges to the true value more quickly than the counting process.

Assume the survival function $\hat{S}_i$ converges to the mean value S_i. It can be expressed as:

$$\sqrt{N_i}[\hat{S}_i - S_i] \to N(0, Var(\hat{S}_i))$$

By Taylor expansion:

$$g(\hat{S}_i) = g(S_i) + g'(S_i)(\hat{S}_i - S_i) + O((\hat{S}_i - S_i)^2)$$

Where $O(.)$ is the asymptotic notation or Bachmann–Landau notation. From Slutsky's theorem[2], there exist the relationships:

$$\sqrt{N_i}[g(\hat{S}_i) - g(S_i)] \to N(0, [g'(S_i)]^2 Var(\hat{S}_i))$$

3.1 Asymptotic Variance for Durations.

We begin with the distribution function of durations. We will then extend this result to cover the distributions of age and DCL. Recall the distribution of durations which can be written as:

$$\hat{a}_i^d = \hat{S}_{i-1}\hat{h}_i$$

with variance:

$$Var(\hat{a}_i^d) = Var(\hat{S}_{i-1}\hat{h}_i)$$

Theorem 1: Assume that we have the hazard function $h_i \in [0,1)^{F-1}$ and the survival function $S_{i-1} \in [0,1)^{F-1}$ for $i = 1, 2...F$, the variance of the estimators $\hat{a}_i^d$ of the distribution of durations are given by:

$$Var(\hat{a}_i^d) = (\hat{S}_{i-1} * \hat{h}_i)^2 * [\sum_{k=1}^{i-1} \frac{D_k}{N_k(N_k - D_k)} + \frac{N_i - D_i}{N_i D_i}] \tag{11}$$

[2]Slutsky's theorem means that there exist two random variables or vectors X_i and Y_i. If those variables or vectors satisfy $X_i \xrightarrow{d.} X$ and $Y_i \xrightarrow{p.} c$, then there exist the relationships:

$$f(X_i, Y_i) \xrightarrow{d.} f(X, c)$$

Where $X_i \xrightarrow{d.} X$ means that X_i converges to the fixed value X in distribution; $Y_i \xrightarrow{p.} c$ means that Y_i converges to the constant point c in probability.

All proofs for theorem 1 are in the appendix.

To derive the variance of DCL, some additional formula should be derived. In equation (10), it can be seen that it is the product of the constant value i, and three random variables $\hat{S}_i$, $\hat{h}_i$ and $\bar{h}$.As Breslow and Crowley (1974) derived the properties of the KM estimator and the hazard function. They consider a very complicated model including both censored and uncensored data. They divided the original data into two groups and derived the distribution of the survival function and the hazard function in large sample size. They found that the survival function and the hazard function followed the normal distribution. The diagonal variance-covariance matrix of the hazard function is the variance formula for the hazard function while the off-diagonal terms are all equal to zero. It means that the hazard function is independent in different periods. On the other hand, they show the covariance for the survival function did not equal to zero.

We adopt a different method to derive the covariance of $\hat{S}_i$ and $\hat{S}_j$ for $i < j$. Recall the Taylor expansion for $\hat{S}_i$ and $\hat{S}_j$:

$$exp(ln\hat{S}_i) = exp(lnS_i) + (ln\hat{S}_i - lnS_i)exp(lnS_i) + O((ln\hat{S}_i - lnS_i)^2)$$

$$exp(ln\hat{S}_j) = exp(lnS_j) + (ln\hat{S}_j - lnS_j)exp(lnS_j) + O((ln\hat{S}_j - lnS_j)^2)$$

Rearrange those two equations:

$$\hat{S}_i - S_i \cong S_i(ln\hat{S}_i - lnS_i)$$

$$\hat{S}_j - S_j \cong S_j(ln\hat{S}_j - lnS_j)$$

Then multiply them and take the expectation:

$$\begin{aligned} Cov(\hat{S}_i, \hat{S}_j) &= E[\hat{S}_i - S_i)(\hat{S}_j - S_j)] \\ &= S_iS_jE[(ln\hat{S}_i - lnS_j)(ln\hat{S}_j - lnS_j)] \\ &= S_iS_jCov(ln\hat{S}_i, ln\hat{S}_j) \\ &= S_iS_jCov(\sum_{k=1}^{i} ln(1-\hat{h}_k), \sum_{k=1}^{j} ln(1-\hat{h}_j)) \\ &= S_iS_jVar(\sum_{k=1}^{i} ln(1-\hat{h}_k) \end{aligned} \tag{12}$$

The delta method is applied to derive the covariance of the KM estimators in equation (12). Since the hazard function h_k follows the binomial distribution

and it is independent in each period, therefore, the $Cov(\hat{h}_m, \hat{h}_n) = 0$ for $m \neq n$. Therefore, $Cov(\sum_{k=1}^{i} ln(1-\hat{h}_k), \sum_{k=1}^{j} ln(1-\hat{h}_j)) = Var(\sum_{k=1}^{i} ln(1-\hat{h}_k))$ when $i < j$. Replace the formula that $Var[\sum_{k=1}^{i} ln(1-\hat{h}_k)] = \sum_{k=1}^{i} \frac{D_k}{N_k(N_k-D_k)}$ which is shown in theorem 1. Applying the large sample properties of the maximum likelihood estimator, the formula of the covariance between $\hat{S}_i$ and $\hat{S}_j$ can be written as:

$$Cov(\hat{S}_i, \hat{S}_j) = \hat{S}_i \hat{S}_j [\sum_{k=1}^{i} \frac{D_k}{N_k(N_k - D_k)}] \; for \; i < j \tag{13}$$

After deriving the covariance, we can use the delta method of ratio variable to derive the formula of the DCL. At this point, the delta method has been applied twice. The first, the delta method is applied to derive the variance of the distribution of the duration. The second step, we treat the a_i as the ratio distribution as $\hat{x}i/\hat{y}$ with $\hat{x}_i = i\hat{S}_{i-1}\hat{h}_i$ and $\hat{y} = \sum_{k=0}^{F} \hat{S}_k$. In other words, the numerator can be known as the $\hat{S}_{i-1}\hat{h}_i$ and the denominator is $1/\bar{h} = \sum_{k=0}^{F} \hat{S}_k$. Using the delta method for the ratio estimator $\hat{x}_i/\hat{y}$ expansion at the mean value x_i and y is:

$$\frac{\hat{x}_i}{\hat{y}} = \frac{x_i}{y} + \frac{\hat{x}_i - x_i}{y} - \frac{x_i}{y^2}(\hat{y} - y) + O((\hat{x}_i - x_i)^2 + (\hat{y} - y)^2)$$

Take the expectation on both sides, it can be seen that:

$$E[\frac{\hat{x}_i}{\hat{y}}] \approx \frac{x_i}{y} \tag{14}$$

Therefore, the variance of the ratio estimator $\hat{a}_i = \hat{x}_i/\hat{y}$ is:

$$Var(\frac{\hat{x}_i}{\hat{y}}) \approx \frac{Var(\hat{x}_i)}{y^2} + \frac{x_i^2}{y^4} Var(\hat{y}) - 2\frac{x_i}{y^3} Cov(\hat{x}_i, \hat{y}) \tag{15}$$

Apply the large sample properties of the maximum likelihood estimator, replace x_i by $\hat{x}_i$ and y by $\hat{y}$ where $\hat{x}_i = i\hat{S}_{i-1}\hat{h}_i$ and $\hat{y} = 1/\bar{h}$.[3] First, note that the variance of $\frac{1}{\bar{h}}$ is:

$$Var(\frac{1}{\bar{h}}) = Var(\sum_{i=0}^{F} \hat{S}_i) = \sum_{i=1}^{F} Var(\hat{S}_i) + 2\sum_{i \neq j} Cov(\hat{S}_i, \hat{S}_j) \tag{16}$$

[3]As Greenwood (1926) showed that the maximum likelihood estimator $\hat{S}_i$ is close to the mean value of S_i in large sample size. the S_i can be replaced by $\hat{S}_i$ in Greenwood formula. At this point, we replace x_i by $\hat{x}_i$ and y by $\hat{y}$ since x_i and y consist of the survival function S_i and the hazard function h_i

Theorem 2 Substitute equation (11)(12),(16) into equation (15), the variance of the DCL can be defined as:

$$Var(\hat{a}_i) = i^2\bar{h}^2 Var(\hat{S}_{i-1}\hat{h}_i) + (i\hat{S}_{i-1}\hat{h}_i)^2\bar{h}^4 Var(\frac{1}{\bar{h}}) - 2i^2\hat{S}_{i-1}\hat{h}_i\bar{h}^3 Cov(\hat{S}_{i-1}\hat{h}_i, \frac{1}{\bar{h}}) \tag{17}$$

For $i = 1, 2, ...F$.

Since the variance of the DCL is derived, we can find out the variance of the age distribution. In terms of the age distribution.

Corollary 1 : Assume that we have the inverse summation of the survival function $\bar{h}$ following the multivariate normal distribution, the survival function $S_{i-1} \in [0,1)^{F-1}$ for $i \in Z_+ = (1, 2, ..., \infty)$, the variance of the age distribution can be derived as:

$$Var(\hat{a}_i^A) = \bar{h}^2 Var(\hat{S}_{i-1}) + \hat{S}_{i-1}^2\bar{h}^4 Var(\frac{1}{\bar{h}}) - 2\hat{S}_{i-1}\bar{h}^3 Cov(\hat{S}_{i-1}, \frac{1}{\bar{h}}) \tag{18}$$

For $i = 1, 2, ..., F$.

3.2 Censored Problem and the Non-parametric Maximum Likelihood Estimation

The KM estimators can be estimated by the non-parametric maximum likelihood estimators (NPMLE). The NPMLE gives the same results as the product limited estimator (PL). NPMLE is the numerical method to solve out the hazard function.The maximum likelihood function for the survival function can be written as:

$$L = \prod_{i=1}^{F} [S_{i-1} - S_i]^{D_i} S_i^{N_i - D_i} \tag{19}$$

Assume the S_0=1, this maximum likelihood formula can be applied to estimate the survival function for F period. This function can be modified if the survival function is replaced by the hazard function:

$$S_i = \prod_{k=0}^{i} (1 - h_k)$$

Define the value N_i to be the total number of the observations at the risk in the period i. Then the NPMLE function can be rewritten as:

$$L = \prod_{i=1}^{F} (1 - h_i)^{N_i - D_i} h_i^{(D_i)} \tag{20}$$

Take the first order derivatives with respect to h_i:

$$\frac{\partial lnL}{\partial h_i} = -\frac{N_i - D_i}{1 - h_i} + \frac{D_i}{h_i} = 0$$

$$h_i = \frac{D_i}{N_i}$$

Since the NPMLE provides the estimator of the hazard function, the survival function can be calculated as $S_i = \sum_{k=1}^{i}(1 - h_k)$. The age distribution, distribution of duration and DCL can be also calculated since they can be expressed as S_i and h_i.

Now we introduce the concept of censored data. The left-censored data are when the starting point cannot be observed, being outside the period of observation, the sample period. However, the endpoint is included in the sample period. Right censored data are where the endpoint cannot be observed but the start is. The KM estimator is applied for the right-censored data. In this section, we will we just consider the implications of right-censored data here. The maximum length of a spell is F periods. N_k means the number of spells that survived up to the $k - th$ period. Define the T_j as the true lifetime of a spell $j \in (1, 2, ...N)$. N is the total number of the sample size in the initial period. The observed lifetime t_i can be defined as:

$$t_j = min(T_j, C_j) \quad and \quad \omega_j = I(T_j \leq C_j) \quad j = 1, 2, ...N_j.$$

Where the C_j means the censored time of the observation for the j-th observation; T_j is the survival time of the j-th observation; The observed lifetime t_j is the minimum of C_j and T_j. ω_j is the uncensored coefficient. If the observation t_j is censored, ω_j is equal to 0. Otherwise, ω_j is equal to 1. If the period of observation is less than the true lifetime, then the data is right censored:

$$C_j < T_j,\ t_j = C_j\ (right\ \ censored) \quad and \quad \omega_j = 0$$

Otherwise, the data is uncensored:

$$T_j \leq C_j,\ t_j = T_j\ (uncensored) \quad and \quad \omega_j = 1$$

4 Monte Carlo Simulation

The variance formulas of the DCL, age distribution and distribution of duration have been derived using the Taylor expansion and the delta method. In this part, we are going to investigate the properties of those formulas. Depending on the simulation, the bias of the asymptotic variances can be evaluated. The data is generated from the exponential distribution function. The sample sizes are chosen to be $N = 25, 50, 100$ and 200. Both the no-censored and censored situations are considered in the simulations. However, we assume that the data is collected in discrete time. We collect the raw continuous time data and transfer them into intervals defined as $(0, r_1], (r_1, r_2], ...(r_{k-1}, r_k], ...(r_{F-1}, r_F]$. For $t_j \in (0, r_1]$ we set duration $t_j = r_1$. For $t_j \in (r_{k-1}, r_k]$ we set $t_j = r_k$ and so on. The simulation process is:

Step 1: assume the observed duration is $t_j = min(T_j, C_j)$. The sample size is chosen to be $N = 25, 50, 100, 200$. The results are reported separately. Both the lifetime time T_j and the censored time C_j follow the exponential distribution. The censored time and the lifetime have the probability density functions (PDF):

$$p(C_j) = 0.5exp(-0.5C_j) \quad p(T_j) = 2exp(-2T_j) \tag{21}$$

Therefore, they have the survival function for each r_i-th period:

$$p(C_j > r_k) = exp(-0.5r_k) \quad p(T_j > r_k) = exp(-2r_k) \tag{22}$$

j is the $j - th$ observation where $j \in (1, 2, ..., N)$. N is the total sample size. There exists the $\frac{2}{2+0.5}$ [4] uncensored proportion of the total observations depending on the PDF. For the uncensored problems, we can just generate the survival time $t_j = T_j$ and assume they are all uncensored with the censored coefficient $\omega_j = 1$ for all the j. In other words, the observation can be written as $(T_j, 1)$ for all the j. For the censored problem, we also need to generate the censored time C_j. If $T_j < C_j$, the $j - th$ observation is non-censored and we assign a parameter $\omega_j = 1$ write the as (T_j, ω_j). If $C_j < T_j$, it means the

[4] Since the parameter of the exponential distribution of Censored time and observed time are 0.5 and 2, separately. The censored proportion of the total sample can be known as $\frac{0.5}{2+0.5}$. The algebra is shown by Efron (1981).

observation is censored. Therefore, we written it as (C_j, ω_j) with $\omega_j = 0$. After that, the survival data are allocated into F group. In other words, different survival periods are transformed into some fixed period group. In this case, F is chose to be 5. We separate them into five regions: $(0, 0.1]$, $(0.1, 0.2]$, $(0.2, 0.3]$, $(0.3, 0.5]$, and $(0.5, \infty)$. This can be known as case 1. In case 2, another five regions are generated: $(0, 0.2]$, $(0.2, 0.4]$, $(0.4, 0.6]$, $(0.6, 0.8]$, and $(0.8, \infty)$

Step 2: The formula (17), (18) and (11) are applied to calculate the variance of the DCL, age distribution and duration of distribution for each period.

Step 3: Repeat step 1 and step 2 by M times. M is chosen to be 10000. One important thing is that there may exist zero observations in one of the 5 intervals in the simulated samples. If there exists this situation, this sample is eliminated and another sample is re-simulated again until we have 10000 samples.

The real value of the variance of DCL can be calculated by:

$$Var(a_i)_{real} = \sum_{m=1}^{M} (i\bar{h}S_{k-1,m}h_{i,m} - \frac{\sum_{m=1}^{M} i\bar{h}S_{i-1,m}h_{i,m}}{M})^2/(M-1) \quad (23)$$

The real value of the variance of *age* distribution can be calculated by:

$$Var(a_i^A)_{real} = \sum_{m=1}^{M} (a_{i,m}^A - \frac{\sum_{m=1}^{M} a_i^A}{M})^2/(M-1) \quad (24)$$

The real value of the variance of *duration* distribution can be calculated by:

$$Var(a_i^d)_{real} = \sum_{m=1}^{M} (a_{i,m}^d - \frac{\sum_{m=1}^{M} a_i^d}{M})^2/(M-1) \quad (25)$$

In other words, we collect M estimators of a_i,a_i^A and a_i^d and calculate the variance of them[5]. Equation (23), (24) and (25) can be known as the real variance of the three distributions depending on the properties of the Monte

[5]In the simulation result, the coefficient i is ignored when the DCL are calculated even it exists in the formula. The reason is that i is a constant parameter for each a_i

Carlo simulations. The benchmark real variances are applied to compare with the analytic variance derived by delta method whether the approximation results are close to the real value. Since the formula $\hat{S}_{i-1}\hat{h}_i = \hat{S}_{i-1} - \hat{S}_i$ is replaced in the variance formula to calculate the true value of the variance of the distribution of duration, so the first period of the variance of DCL can be known as the variance of the age distribution. In addition, the final period variance of DCL is also a special case of the variance of the age distribution when all the observations are uncensored.

Table (2) reports the simulation results for the non-censored data for case 1. Those data are divided into $(0, 0.1]$, $(0.1, 0.2]$, $(0.2, 0.3]$, $(0.3, 0.5]$, and $(0.5, \infty)$. In table (2), all the censored parameters ω_j are equal to 1 for all the j, which means that the observations are all uncensored. As we can see from table (2), when the sample size equal to 25, there exists a slight bias for the variance. When the sample size increased to 50, the asymptotic formula of the variance performs very well for all the regions. When the sample size increased to either n=100 or n=200, the gap between the benchmark value and the analytic value of the variance are reduced. With the increase of the sample size, the approximation value is closer to the true value when the observations are uncensored.

Table 2: The Variance of the Duration of case 1 with Non-censored Simulation. All the Results Are Multiplied by 10^3

True Value					
N	$Var(a^d_{0.1})$	$Var(a^d_{0.2})$	$Var(a^d_{0.3})$	$Var(a^d_{0.5})$	$Var(a^d_{\infty})$
25	5.6497	4.6783	3.7561	5.6901	9.1525
50	2.9913	2.5801	2.1244	2.9637	4.6877
100	1.4844	1.2554	1.0757	1.4705	2.2919
200	0.7451	0.6293	0.5404	0.7406	1.1594
Asymptotic Value					
N	$E[Var(a^d_{0.1})]$	$E[Var(a^d_{0.2})]$	$E[Var(a^d_{0.3})]$	$E[Var(a^d_{0.5})]$	$E[Var(a^d_{\infty})]$
25	5.6880	4.9125	4.2650	5.6862	8.8820
50	2.8987	2.4790	2.0904	2.9041	4.5610
100	1.4677	1.2533	1.0613	1.4640	2.3022
200	0.7367	0.6293	0.5318	0.7355	1.1578

In table (3), the data are simulated by the same process. There still exists a slight bias for the variance of the DCL when the sample size is $N = 25$.

Table 3: The Variance of the DCL of Case 1 with Non-censored Simulation. All the Results Are Multiplied by 10^3

True Value					
N	$Var(a_{0.1})$	$Var(a_{0.2})$	$Var(a_{0.3})$	$Var(a_{0.5})$	$Var(a_{\infty})$
25	0.7378	0.51417	0.3606	0.4858	0.46588
50	0.36525	0.27085	0.19608	0.24835	0.2327
100	0.1777	0.1299	0.0986	0.1228	0.1137
200	0.0879	0.0646	0.0494	0.0616	0.0575
Asymptotic Value					
N	$E[Var(a_{0.1})]$	$E[Var(a_{0.2})]$	$E[Var(a_{0.3})]$	$E[Var(a_{0.5})]$	$E[Var(a_{\infty})]$
25	0.7719	0.5478	0.4084	0.4752	0.4460
50	0.3644	0.2638	0.1951	0.2418	0.2256
100	0.1782	0.1305	0.0982	0.1221	0.1140
200	0.0877	0.0647	0.0489	0.0613	0.0572

When the sample size increase to 50, all the asymptotic results are improved and they are all close to the true value. With respect to $N = 100$ and 200, the asymptotic variance tends to be closer to the real variance. However, it can be found that the asymptotic variances do not always overestimate the true value. Sometimes it underestimates the true variance of the DCL. In conclusion, the asymptotic variance formula of DCL is reduced with the increase of the sample size.

Next, the censored data is considered. Table (4) show the result of the variance of the distribution of duration. Compared with the benchmark value, there exists a slight bias in the variance calculated from the analytic formula when the sample size $N = 25$. When sample size increased to 50, the asymptotic variance formula performs well. When the sample size tends to be a larger (N=100 and 200), the empirical results show that the values of asymptotic variance are nearly the same as the true values. In conclusion, the asymptotic variance formula can capture the true value even the sample size is small(N=25). The asymptotic formula may overestimate or underestimate the true value.

Table (5) show the simulation results of the DCL variance. When the sample size is extremely small (N=25), the asymptotic results are still quite accurate. When the sample size increased to 50, all the asymptotic results are improved. They are all close to the true value. When the sample size increase

Table 4: The Variance of the Duration of case 1 with Censored Simulation. All the Results Are Multiplied by 10^3

True Value

N	$Var(a^d_{0.1})$	$Var(a^d_{0.2})$	$Var(a^d_{0.3})$	$Var(a^d_{0.5})$	$Var(a^d_{\infty})$
25	5.4770	4.7128	4.0016	6.1382	9.5321
50	2.8680	2.5773	2.2555	3.4038	5.0703
100	1.4566	1.3187	1.1746	1.7109	2.5091
200	0.7497	0.6525	0.5807	0.8130	1.2539

Asymptotic Value

N	$E[Var(a^d_{0.1})]$	$E[Var(a^d_{0.2})]$	$E[Var(a^d_{0.3})]$	$E[Var(a^d_{0.5})]$	$E[Var(a^d_{\infty})]$
25	5.6045	5.0796	4.7601	6.4884	9.4990
50	2.8519	2.5617	2.2987	3.3076	4.9242
100	1.4442	1.3024	1.1618	1.6589	2.4841
200	0.7248	0.6537	0.5837	0.8349	1.2496

Table 5: The Variance of the DCL of Case 1 with Censored Simulation. All the Results Are Multiplied by 10^3

True Value

N	$Var(a_{0.1})$	$Var(a_{0.2})$	$Var(a_{0.3})$	$Var(a_{0.5})$	$Var(a_{\infty})$
25	0.6767	0.4910	0.3696	0.5094	0.6005
50	0.3311	0.2565	0.2006	0.2770	0.3102
100	0.1645	0.1290	0.1032	0.1381	0.1545
200	0.0832	0.0631	0.0506	0.0656	0.0772

Asymptotic Value

N	$E[Var(a_{0.1})]$	$E[Var(a_{0.2})]$	$E[Var(a_{0.3})]$	$E[Var(a_{0.5})]$	$E[Var(a_{\infty})]$
25	0.6798	0.5162	0.4174	0.5094	0.5824
50	0.3345	0.2557	0.2032	0.2660	0.2992235
100	0.1648	0.1285	0.1025	0.1345	0.1513
200	0.0811	0.0636	0.0511	0.0676	0.0763

to 100 and 200. They asymptotic results show an accurate approximation value for the true value. Therefore, the asymptotic formula of the variance of the DCL works in the censored data.

Table (6) to (9) present the variance of the duration and DCL under the another category. All the data are assigned into another five regions: $(0, 0.2]$, $(0.2, 0.4]$, $(0.4, 0.6]$, $(0.6, 0.8]$, and $(0.8, \infty)$. This can be known as case 2. In

terms of the uncensored case, all the censored coefficient $\omega_j = 1$. In the censored situation, the $p(C_j) = 0.5exp(-0.5C_j)$ and $p(T_j) = 2exp(-2T_j)$ are generated. This process is the same as the case 1. In table (6) and (7), the asymptotic variance can give a accurate approximation for the true value even in the extremely small sample size (N=25). Both the variance of the DCL and duration are either overestimate or underestimate without a unique conclusion. When the sample size tends to be a large number, they are nearly unbiased from the true value.

With respect to the censored case, the same results can be concluded in table (8) and table (9).

Table 6: The Variance of the Duration of case 2 with Non-censored Simulation.All the Results Are Multiplied by 10^3

True Value					
N	$Var(a^d_{0.2})$	$Var(a^d_{0.4})$	$Var(a^d_{0.6})$	$Var(a^d_{0.8})$	$Var(a^d_{\infty})$
25	8.511	6.4731	4.6370	2.9440	6.0651
50	4.4362	3.4353	2.5537	1.7365	3.2366
100	2.1816	1.6982	1.2514	0.8812	1.5987
200	1.1384	0.8610	0.6392	0.4491	0.8074
Asymptotic Value					
N	$E[Var(a^d_{0.2})]$	$E[Var(a^d_{0.4})]$	$E[Var(a^d_{0.6})]$	$E[Var(a^d_{0.8})]$	$E[Var(a^d_{\infty})]$
25	8.4295	6.5750	4.9049	3.6899	6.1695
50	4.3413	3.3608	2.4783	1.7592	3.1459
100	2.1879	1.7030	1.2538	0.8834	1.5953
200	1.0990	0.8550	0.6298	0.4459	0.8012

Table 7: The Variance of the DCL of Case 2 with Non-censored Simulation. All the Results Are Multiplied by 10^3

True Value					
N	$Var(a_{0.2})$	$Var(a_{0.4})$	$Var(a_{0.6})$	$Var(a_{0.8})$	$Var(a_{\infty})$
25	2.3456	1.1895	0.6760	0.3829	0.5396
50	1.2091	0.6276	0.3794	0.2329	0.2956
100	0.5764	0.3054	0.1860	0.1176	0.1462
200	0.2987	0.1535	0.0938	0.0602	0.0731
Asymptotic Value					
N	$E[Var(a_{0.2})]$	$E[Var(a_{0.4})]$	$E[Var(a_{0.6})]$	$E[Var(a_{0.8})]$	$E[Var(a_{\infty})]$
25	2.4358	1.2334	0.7190	0.4755	0.5430
50	1.2085	0.6222	0.3667	0.2325	0.2828
100	0.5841	0.3095	0.1845	0.1174	0.1440
200	0.2884	0.1540	0.0925	0.0594	0.0727

Table 8: The Variance of the Duration of case 2 with Censored Simulation. All the Results Are Multiplied by 10^3

True Value					
N	$Var(a^d_{0.2})$	$Var(a^d_{0.4})$	$Var(a^d_{0.6})$	$Var(a^d_{0.8})$	$Var(a^d_{\infty})$
25	7.9283	6.9220	5.1855	3.6686	6.3243
50	4.2300	3.7601	3.0929	2.3271	3.9246
100	2.1853	1.8781	1.5767	1.2560	1.9910
200	1.0669	0.9379	0.7799	0.6350	0.9996
Asymptotic Value					
N	$E[Var(a^d_{0.2})]$	$E[Var(a^d_{0.4})]$	$E[Var(a^d_{0.6})]$	$E[Var(a^d_{0.8})]$	$E[Var(a^d_{\infty})]$
25	8.1582	7.0511	6.0630	5.5234	7.5641
50	4.2223	3.6962	3.0363	2.4997	3.8673
100	2.1348	1.8690	1.5411	1.2321	1.9587
200	1.0734	0.9371	0.7749	0.6247	0.9865

5 Conclusion

In this paper, we use the delta method to derive the variance of the distribution of duration, distribution of the age and the distribution of completed lifetimes (DCL). The DCL is cross-sectional distribution among the survival analysis. Depending on the asymptotic approximation of the variance, we

Table 9: The Variance of the DCL of Case 2 with Censored Simulation. All the Results Are Multiplied by 10^3

True Value					
N	$Var(a_{0.2})$	$Var(a_{0.4})$	$Var(a_{0.6})$	$Var(a_{0.8})$	$Var(a_{\infty})$
25	1.9352	1.1768	0.7175	0.4503	0.5984
50	1.0521	0.6580	0.4406	0.2989	0.3947401
100	0.5269	0.3224	0.2233	0.1618	0.2025
200	0.2539	0.1603	0.1108	0.0813	0.1017
Asymptotic Value					
N	$E[Var(a_{0.2})]$	$E[Var(a_{0.4})]$	$E[Var(a_{0.6})]$	$E[Var(a_{0.8})]$	$E[Var(a_{\infty})]$
25	1.8620	1.1101	0.7612	0.5964	0.6855
50	1.0220	0.6266	0.4140	0.3027	0.3806
100	0.5159	0.3197	0.2164	0.1557	0.1970
200	0.2561	0.1594	0.1091	0.0798	0.1000

provide the analytic formula to calculate the variance of the three distributions. The asymptotic variance derived from delta method is straightforward since it is the same way to derived the Greenwood formula. In addition, the covariance between different survival function is derived in a clearer way compared with Breslow and Crowley (1974).

The data is simulated and applied to investigate the accurate the asymptotic variance of the DCL, age distribution and the distribution of the duration. There are two cases of simulation considering in this paper. The observations are assumed to follow the exponential distribution. Depending on the Monte Carlo results, the asymptotic variance of the DCL, age distribution and the distribution of the duration gives more accurate results as the increasing of the sample size. In other words, the bias between the asymptotic results and the true results are reduced as the sample size increased.

For the further study, it is attractive to see the bootstrap performance. Whether the bootstrap corrected variance can provide a better result compared with the asymptotic formula in the small sample size. Another point is the confidence interval for the DCL. It is worth to evaluate whether the delta method provides the accurate confidence interval of DCL when the sample size is small or large.

6 Appendix

6.1 Proof of Theorem 1

The variance formula can be rewritten as:

$$Var(\exp[\ln \hat{a}_i^d]) = Var(\exp[\ln \hat{S}_{i-1} + \ln \hat{h}_i])$$

$$Var(\hat{a}_i^d) = Var(exp(ln\hat{S}_{i-1} + ln\hat{h}_i))$$

Furthermore, we have another Taylor expansion for $\ln S_i$ and $\ln h_i$:

$$\ln \hat{S}_{i-1} = \ln S_{i-1} + \frac{\hat{S}_{i-1} - S_{i-1}}{S_{i-1}} + O((\hat{S}_{i-1} - S_{i-1})^2)$$

$$\ln h_i = \ln h_i + \frac{\hat{h}_i - h_i}{h_i} + O((\hat{h}_i - h_i)^2)$$

Taking the mean value on both sides, we have:

$$E(\ln \hat{S}_{i-1}) \approx \ln S_{i-1}$$

$$E(\ln \hat{h}_i) \approx \ln h_i$$

Hence, the first-order Taylor expansion applied to the equation $exp(\ln \hat{S}_{i-1} + \ln \hat{h}_i)$ and expanded at the mean value of the $E(\ln \hat{S}_{i-1}) \approx \ln S_{i-1}$ and $E(\ln \hat{h}_i) \approx \ln h_i$[6]:

$$\begin{aligned}\exp(\ln \hat{S}_{i-1} + \ln \hat{h}_i) = \exp(\ln S_{i-1} + \ln h_i) + (\ln \hat{S}_{i-1} - \ln S_i)\exp(\ln S_{i-1} + \ln h_i) \\ -(\ln \hat{h}_i - \ln h_i)\exp(\ln S_{i-1} + \ln h_i) + O((\ln \hat{S}_{i-1} - \ln S_i)^2 + (\ln \hat{h}_i - \ln h_i)^2)\end{aligned}$$

[6]The hazard function can be estimated by the maximum likelihood method, and the KM estimator consists of the hazard function. Depending on the properties of the maximum likelihood estimator, it can be known that the KM estimator $\hat{S}_{i-1}$ converges to the true value S_{i-1} and the marginal hazard function $\hat{h}_i$ converges to the true value h_i. At this point, we show that those result can be derived from the delta method

Rearranging this simplifies to:

$$\hat{S}_{i-1}\hat{h}_i - S_{i-1}h_i \approx S_{i-1}h_i[(ln\hat{S}_{i-1} - lnS_{i-1}) + (ln\hat{h}_i - lnh_i)]$$

Hence the variance $Var(a_i^d)$ can be rewritten as:

$$Var(\hat{a}_i^d) = Var(\hat{S}_{i-1}\hat{h}_i) \cong (S_{i-1}h_i)^2[Var(\ln \hat{S}_{i-1}) + Var(\ln \hat{h}_i)]$$

Depending on the large sample property of the maximum likelihood estimator, the KM estimator $\hat{S}_{i-1}$ converges to the true value S_{i-1} and the marginal hazard function $\hat{h}_i$ converges to the true value h_i. Therefore, $Var(\hat{a}_i^d)$ can be written as:

$$Var(\hat{a}_i^d) \cong (\hat{S}_{i-1}\hat{h}_i)^2[Var(ln\hat{S}_{i-1}) + Var(ln\hat{h}_i)]$$

This approximation assumes that $\hat{S}_{i-1}$ is independent with $\hat{h}_i$, the covariance between the $\ln \hat{S}_{i-1}$ and $\ln \hat{h}_i$ is zero.

The logarithm version of the survival function can be written as:

$$\ln \hat{S}_{i-1} = \sum_{k=1}^{i-1} \ln(1 - \hat{h}_k)$$

Assume the D_i follows the binomial distribution with parameters N_i and $\hat{h}_i$. Therefore, $Var(D_i) = N_i\hat{h}_i(1 - \hat{h}_i)$. It can be shown that $Var(\hat{h}_i) = Var(\frac{D_i}{N_i}) = \hat{h}_i(1 - \hat{h}_i)/N_i$. By applying the first-order Taylor expansion:

$$ln(\hat{h}_i) = lnh_i + (\hat{h}_i - h_i)\frac{1}{h_i} + O((\hat{h}_i - h_i)^2)$$

$$ln(1 - \hat{h}_i) = ln(1 - h_i) + (\hat{h}_i - h_i)\frac{1}{1 - h_i} + O((\hat{h}_i - h_i)^2)$$

To rearrange the formula:

$$ln(\hat{h}_i) - lnh_i \cong (\hat{h}_i - h_i)\frac{1}{h_i}$$

$$ln(1 - \hat{h}_i) - ln(1 - h_i) \cong (\hat{h}_i - h_i)\frac{1}{1 - h_i}$$

It can be assumed that the observations are independent Bernoulli distribution and they are independent with each other. Then the variance of $\hat{h}_i$ can be written as:

$$Var(\hat{h}_i) = Var(1 - \hat{h}_i) = \frac{\hat{h}_i(1 - \hat{h}_i)}{N_i}$$

Apply the large sample properties of the maximum likelihood estimator:

$$\begin{aligned} Var(ln(\hat{h}_i)) &\cong \frac{1}{\hat{h}_i^2} \frac{\hat{h}_i(1 - \hat{h}_i)}{N_i} \\ &\cong \frac{N_i - D_i}{N_i D_i} \end{aligned}$$

$$\begin{aligned} Var(ln(1 - \hat{h}_i)) &\cong \frac{1}{(1 - \hat{h}_i)^2} \frac{\hat{h}_i(1 - \hat{h}_i)}{N_i} \\ &\cong \frac{D_i}{N_i(N_i - D_i)} \end{aligned}$$

We have a formula for the exponential function:

$$Var(\hat{S}_{i-1}\hat{h}_i) = (\hat{S}_{i-1}\hat{h}_i)^2[Var(ln\hat{S}_{i-1}) + (ln\hat{h}_i)]$$

Therefore:

$$\begin{aligned} Var(\hat{S}_{i-1}\hat{h}_i)) = (\hat{S}_{i-1} * \hat{h}_i)^2 * [&\sum_{k=1}^{i-1} \frac{D_k}{N_k(N_k - D_k)} \\ &+ \frac{N_i - D_i}{N_i D_i}] \end{aligned} \tag{26}$$

References

Aalen, O. O. (1978). Non parametric inference for a family of counting processes. *Annals of Statistics*, 6:701–726.

Andersen, P. K., Ørnulf Borgan, Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer.

Baharad, E. and Eden, B. (2004). Price rigidity and price dispersion: Evidence from micro data. *Economic Modelling*, 7:613–641.

Bohoris, G. A. (1994). Comparison of the cumulative-hazard and kaplan–meier estimators of the survivor function. *IEEE Transactions on Reliability*, 43(2):230–232.

Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *Annals of Statistics*, 2:437–453.

Coenen, G., Mohr, M., and Straub, R. (2008). Fiscal consolidation in the euro area: Long-run benefits and short-run costs. *Economic Modelling*, 25:912–932.

Colosimo, E., Ferreira, F., Oliveira, M., and Sousa, C. (2002). Empirical comparisons between kaplan-meier and nelson-aalen survival function estimators. *Journal of Statistical Computation and Simulation*, 72(4):299–308.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Dixon, H. (2012). A unified framework for using micro-data to compare dynamic time-dependent price-setting models. *BE Journal of Macroeconomics (Contributions)*, 12:1–43.

Dixon, H. and Bihan, H. L. (2012). Generalised taylor and generalised calvo price and wage setting: Micro-evidence with macro implications. *Economic Journal*, 122(560):532–554.

Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association*, 76(374):312–319.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Process and Survival Analysis*. John Wiley & Sons, Inc.

Gillesple, M. J. and Fisher, L. (1979). Confidence bands for the kaplan-meier survival curve estimate. *Annals of Statistics*, 7:920–924.

Greenwood, M. (1926). The natural duration of cancer. *Reports on Public Health and Medical Subjects*, 33:1–26.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data, 2nd Edition*. John Wiley & Sons, Inc.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50(3):163–170.

Nair, V. N. (1981). Plots and tests for goodness of fit with randomly censored data. *Biometrika*, 68:99–103.

Nair, V. N. (1984). Confidence bands for survival functions with censored data: A comparative study. *Technometrics*, 26:265–275.

Nelson, W. (1972). Theory and application of hazard plotting for censored failure data. *Technometrics*, 14:945–965.

Taylor, J. B. (1980). Aggregate dynamics and staggered contracts. *Journal of Political Economy*, 88(1):1–23.