The Variances of Non-Parametric Estimates of the Cross-Sectional Distribution of Durations *

Maoshan Tian[1] and Huw Dixon[2‡]

Abstract

This paper focuses on the link between non-parametric survival analysis and three distributions. The delta method is applied to derive the variances of the non-parametric estimators of three distributions: the distribution of durations (DD), the cross-sectional distribution of ages (CSA) and the cross-sectional distribution of (completed) durations (CSD). The non-parametric estimator of the the cross-sectional distribution of durations (CSD) has been defined and derived by Dixon (2012) and used in the generalized Taylor price model (GTE) by Dixon and Le Bihan (2012). The Monte Carlo method is applied to evaluate the variances of the estimators of DD and CSD and how their performance varies with sample size and the censoring of data. We apply those estimators to two data sets: the UK CPI micro-price data and waiting-time data from UK hospitals. Both the estimates of the distributions and their variances are calculated. Depending on the empirical results, the estimated variances indicate that the DD and CSD estimators are all significant.

JEL Codes: C10, C15, C40, E50
Keywords: Delta Method, Survival Analysis, Kaplan-Meier Estimator

*We are grateful for very helpful comments from Patrick Minford, Kul Luintel, Walter Distaso, seminar participants at Cardiff University and also from participants at the 2018 China Meeting of the Econometric Society. We would also like to thank the editor and referee for their comments and advice.

†[1]Dr M. Tian (corresponding author) was a PhD candidate in Cardiff Business School, University of Cardiff and is now an assistant professor at Chongqing University of Post and Telecommunications. tianms@cqupt.edu.cn

‡[2]H. Dixon is Professor at Cardiff Business School, University of Cardiff and Lead Researcher in Economic Measurement at the National Institute of Economic and Social Research (London) dixonh@cardiff.ac.uk

1 Introduction

Survival analysis has a wide range of applications across several disciplines including engineering, medicine and economics. In this paper, three related distributions[1] arising from survival analysis are examined: the distribution of durations (DD)[2], the cross-sectional distribution of ages (incomplete durations) (CSA) and the cross-sectional distribution of (completed) durations (CSD). Dixon (2012) introduces a unified framework for modelling the three distributions, each of which can be written in terms of the survival function and hazard function. They are different ways of describing the same underlying data. The purpose of this paper is to derive the variances of the non-parametric estimates of the three distributions using the the delta method of Greenwood (1926).

Suppose time is divided 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 S_i gives the probability that an event will last for more than i periods. 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 non-parametric methods of estimating this process. The Kaplan-Meier estimator (KM) for the survival function, and the Nelson-Aalen estimator (NA) for the cumulative hazard function (see Kaplan and Meier (1958), Nelson (1972) and Aalen (1978)). The properties of both estimators have been well studied and in particular their asymptotic variances (see Breslow and Crowley (1974)). 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. Our analysis is applicable to a panel of observations, where many agents (people, households, firms and machines) are observed repeatedly over time and also to situations where just a few agents, or even one, are observed over time.

Starting from the KM and NA estimators of the survival and hazard functions, the non-parametric estimators of the three distributions are constructed in section 2. The contribution of the paper is to derives the asymptotic variances of the estimators of the three distributions using the delta method in section 3. Theorem 1 derives the asymptotic variances for the

[1] We use the term *distribution* as short hand for discrete probability density function.

[2] This is also known as the unconditional hazard function.

DD estimator. The main result is found in Theorem 2, which derive the asymptotic variances for the CSD estimator. Corollary 1 follows with the variances for the CSA estimator.

The three distributions we estimate have many useful applications. In economics, CSD 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 Le Bihan (2012) and Taylor (2016)). 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. In demographics, it also gives the cross-sectional distribution of durations for those people living at a point in time. Dixon and Siciliani (2009) estimate the distribution to obtain the completed waiting times of people on hospital waiting lists, moving from CSA and "ages" or incomplete waiting times to completed waiting times given by CSD. Also, if we are looking at the stock of something at a point in time (unemployed workers, people living in an area, or machines), the CSD gives us the distribution across this stock which can be generated if we know either DD or CDA. This allows us to say when the existing unemployed workers find a job, when people will move from an area, when machines will fail. DD 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. The key conceptual difference between CSD and DD is that the cross-sectional distribution weights spells by their duration.

In section 4, the Monte Carlo method is applied to explore the performance of our estimated variances of the DD and CSD estimators in different sample sizes, and also their sensitivity to the presence of censored observations. We find that whilst there can be small biases in the variances for samples as small as 25, for samples 50 or over there are almost no biases. These results are not sensitive to the presence of right-censored observations for sample sizes of 50 or over.[3] In section 5, we also provide illustrative applications of the method to real data on price-spells and hospital waiting times. We show how the estimated variances can be used to evaluate the significance of the estimators of CSD and DD.

[3]Durations are censored if we do not observe their beginning (left-censored) or their end (right-censored). It is common practice in survival analysis not to used left-censored data, which is why we focus on the right-censored data.

We conclude the introduction with a brief review of the literature on the Kaplan-Meier and Nelson-Aalen estimators of the survival and hazard functions.

1.1 Literature Review

Kaplan and Meier (1958) derived the product limit estimator and the variance of the survival function obtaining the same result derived by Greenwood (1926). In addition, the product limit estimators were shown to be consistent. On the other hand, Nelson (1972) applied a graphical method to investigate the hazard rate (failure ratio) and cumulative hazard function. This graphical method was named as "hazard plotting". After that, Aalen (1978) investigated the hazard function and the cumulative hazard function using counting process theory. Since both Nelson and Aalen derived the cumulative hazard function, the new estimator is known as the Nelson-Aalen estimator. 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). With respect to the parametric method for estimating the survival function and the hazard function, see Cox (1972). In terms of comparing the KM estimators in different groups, see Mantel (1966).

Breslow and Crowley (1974) investigated the life table and the Greenwood formula for the survival function in large samples. They also derived the covariance formula for the survival function. 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 the log-log transformation to guarantee the positive lower bound of the confidence interval of the KM estimator.

2 The Survival and Hazard Functions

Kaplan and Meier (1958) provided an estimator for the survival function, the Kaplan-Meier (KM) estimators of the survival probabilities $S_i \in [0, 1]$ for $i = 0, 1, 2, ..., F$, where F is the maximum duration observed in the data set (in purely theoretical work, this can be arbitrarily large and even infinite). 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, machine failure, price change). When the state changes, this can either be seen as the same agent continuing in the different states (the firm continues but sets a different price) or a new agent replaces the old (the machine fails and is replaced with a new machine). In this section and with all of our analytical results, we will assume that all spells are uncensored and observed in their entirety. We will delay the discussion of censored spells until sub-section 3.2 and section 4 below.

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 as D_k. N_0=N is the total number of price spells in the initial period. The Kaplan-Meier estimator $\hat{S}_i$ of the survival function S_i can be written as:

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

$\hat{S}_i$ can be defined as the proportion of spells surviving for longer than i periods. This formula is also known as the product limit estimator for the survival function. We can set $\hat{S}_0 = 1$ (all spells last longer than zero) and $\hat{S}_F$=0 (no spell lasts more than F periods): Hence there remain $F-1$ survival probabilities to be estimated from the data.

The hazard function h_i is estimated as the proportion of failures amongst spells that have lasted i periods:

$$\hat{h}_i = \frac{D_i}{N_i} \tag{2}$$

Again we assume $D_0 = 0$ and $\hat{h}_0 = 0$ because all spells last at least 0 periods. Since F is the longest spell observed, $\hat{h}_F = 1$ and there remain $F-1$ hazards to be estimated. The estimator of hazard function can be transformed into the KM estimator:

$$\hat{S}_i = \prod_{k=1}^{i} (1 - \hat{h}_k) \tag{3}$$

Likewise, the KM estimator can be transformed into the estimator of hazard function:

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

There is thus a one-to-one mapping between the estimated hazard function and survival function. Note that equation (2) is also the maximum likelihood estimator of hazard function[4]. Therefore, KM estimator can be derived from the maximum likelihood estimator of the hazard function.

Before deriving the estimators of our three distributions, we need to define two additional variables. First define the sum of the estimated survival probabilities $\hat{S} = \sum_{k=0}^{F} \hat{S}_k$, and second define $\bar{h}$ as the reciprocal of this sum.

$$\bar{h} = \frac{1}{\sum_{k=0}^{F} \hat{S}_k} = \frac{1}{\hat{S}} \tag{4}$$

Intuitively, in a balanced panel, $\bar{h}$ is the average proportion of agents that fail each period. To see this, consider some simple examples. First, all spells end in the first period. In this case, $F = 1$ and $\hat{S}_0 = 1$ with $\hat{S}_1 = 0$ so that $\bar{h} = 1$. Second, take the example where all spells last for two periods and then fail, so that $F = 2$. In this case, we have $\hat{S}_0 = \hat{S}_1 = 1$ and $\hat{S}_2 = 0$ so that $\bar{h} = 1/2 : 50\%$ of spells fail per period. Hence, with a balanced panel, we can think of $\bar{h}$ as being the average proportion of failures each period. However, if there is only one cohort, $\bar{h}$ can be thought as being the weighted average hazard over F periods for $i = 0, 1, ..., F$, where the weights are the proportions surviving to period i divided by the sum of estimated survival probabilities (to ensure the weights add up to 1). This is summarised in Proposition 1.

Proposition 1:

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

All proofs are given in the appendix. We will now go on to describe the three distributions and how they relate to the survival and hazard functions and to each other. For all three distributions, we are considering the non-parametric probability density functions corresponding to the non-parametric estimates of the survival and hazard functions. As such, they are captured by the estimated proportions belonging to each duration, which are all non-negative and sum to one.

[4]We summarise this derivation in the online appendix.

2.1 The Distribution of Durations DD

We have a population of price-spells with a discrete random duration i ($i = 1, 2, ..., F$), which has a discrete probability density function DD described by the F probabilities a_i^d which we seek to estimate. The estimators of DD are the the proportions of spells in the data lasting for exactly i periods, $i = 1, 2, ..., F$[5]. That is, they survive for at least $i - 1$ periods with estimated survival probability $\hat{S}_{i-1}$ and change (or end) in the i-th period with estimated hazard probability $\hat{h}_i$. Our estimator of the probability of spells lasting exactly i periods can thus be defined as:

$$\hat{a}_i^d = \hat{S}_{i-1}\hat{h}_i \tag{5}$$

Clearly, $\hat{a}_i^d > 0$ and for $F > 1$, $1 > \hat{a}_i^d$. Also note that:

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

Since:

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

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

2.2 The Cross-Sectional Distribution of Ages CSA

The cross-sectional distribution of ages (CSA) is the probability density function of ages (incomplete durations) at a point in time. The obvious example is a census which records the age of people at a particular date. As with DD, the CSA is defined by F probabilities a_i^A ($i = 1, 2, ..., F$) which we seek to estimate. The estimator for age i is the ratio between the estimates of survival probability for duration i divided by the sum of all the estimates of survival probabilities $\hat{S}$ (or equivalently multiplied by $\bar{h}$):

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

[5] DD is NA(not available) when $i = 0$. The reason is that $\hat{a}_i 0^d = \hat{S}_{-1}h_0$.

Since the survival function is non-increasing, the age distribution is also non-increasing: $\hat{a}_i^A \geq \hat{a}_{i+1}^A$. Note that $\hat{a}_1^A = \bar{h}$ since $\hat{S}_0 = 1$. In addition, it is clear that the summation of the estimates of the age distribution is equal to 1.

In the case of a balanced panel, the age distribution can be thought as being the cross-sectional distribution of ages across agents at a (random) point in time. However, for a particular cohort, it can be also thought 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.

The estimates of the CSA and DD are related by the simple equality:

$$\hat{a}_i^A = \hat{a}_i^d \frac{\bar{h}}{h_i}$$

In the case of a constant hazard rate we have $h_i = \bar{h}$ for all i, so that the estimators are equal for all durations $\hat{a}_i^A = \hat{a}_i^d$.[6]

2.3 The Cross-Sectional Distribution of Durations CSD

Next, we consider the less familiar cross-sectional distribution of completed durations (CSD). Unlike the CSA, this can be considered as the completed durations (lifetime of a person, waiting time before the patient is treated) in process at a point in time and is defined by probabilities a_i $(i = 1, ..., F)$[7]. The CSA is purely backward looking; it simply says what the duration is up until a point in time. The CSD not only looks backwards, but also forward to the end of the completed duration. This is a new distribution derived by Dixon (2012). The estimators for CSD can be written as:

$$\hat{a}_i = i\bar{h}\hat{S}_{i-1}\hat{h}_i \tag{7}$$

[6]The cross-section is length biased, so that the probability of observing a spell is proportional to length. The CSA has an interruption bias, since the spells are incomplete. With a constant hazard, the two biases exactly cancel out. This happens when DD follows a Bernoulli distribution with a hazard rate that is constant (in macroeconomics this is used in the discrete-time Calvo model of pricing).

[7]CSD estimator $\hat{a}_0 = 0$ when $i = 0$

Alternatively, the estimators of CSD can be written as:

$$\hat{a}_i = i\frac{\hat{S}_{i-1}\hat{h}_i}{\hat{S}} \tag{8}$$

Furthermore, the sum of these estimators is always unity:

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

The estimates of the CSD and DD are related by the simple equality:

$$\hat{a}_i = i\hat{a}_i^d\bar{h}$$

That is the CSD estimator for i-th period is i times the corresponding estimate for DD multiplied by $\bar{h}$. In the case $F > 1$, we have $\bar{h} < 1$. It follows that the distributions "cross" as $i\bar{h}$ goes from below 1 to greater than 1. If $\bar{h} = 0.25$, $\hat{a}_i < \hat{a}_i^d$ for $i = 1, 2, 3$ and the two distributions cross at $i = 4$, hence $\hat{a}_4 = \hat{a}_4^d$. For $i > 4$, we have $\hat{a}_i < \hat{a}_i^d$. If we have a balanced panel, we can think of this as the cross-sectional distribution which is length weighted. The probability of observing a spell lasting i periods at a random point in time is i times the probability of observing a one period spell.

In the case of a single cohort, the CSD can be thought as being the distribution of durations 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 CSD for the cohort is given by $\hat{a}_i = i\bar{h}\hat{a}_i^d$. In effect, for a single cohort or even a single firm, the CSD can be thought as weighting the spells by their length, as was suggested by Baharad and Eden (2004).

2.4 The Three Distributions

Since the survival and hazard functions and their estimators are well known, the three distributions can be expressed in terms of these functions. However, the survival function, hazard function and the three distributions are just different ways of describing the data. They are all linked by identities: for all unique survival functions there exists a corresponding unique hazard function and unique DD, CSA and CSD. These identities hold for their estimators as well. Likewise, if we pick a particular hazard function, we can express the

Table 1: Relationships Among Different Functions and Distributions

	$\hat{S}_i$	$\hat{h}_i$	$\hat{a}_i^d$	$\hat{a}_i^A$	$\hat{a}_i$
$\hat{S}_i$	I	$\prod_{j=1}^{i}\left(1-\hat{h}_j\right)$ for $i=1,2,...,F.$	$1-\sum_{j=1}^{i}\hat{a}_j^d$	$\frac{\hat{a}_{i+1}^A}{\hat{a}_1^A}$	$1-\frac{1}{\sum_{k=1}^{F}\frac{\hat{a}_k}{k}}\sum_{j=1}^{i}\frac{\hat{a}_j}{j}$
$\hat{h}_i$	$\frac{\hat{S}_{i-1}-\hat{S}_i}{\hat{S}_{i-1}}$	I	$\prod_{j=1}^{i-1}\left(1-\hat{h}_j\right)$	$\frac{\hat{a}_i^A-\hat{a}_{i+1}^A}{\hat{a}_i^A}$	$\frac{\hat{a}_i}{i}\left[\sum_{i=1}^{F}\frac{\hat{a}_i}{i}\right]^{-1}$
$\hat{a}_i^d$	$\hat{S}_{i-1}-\hat{S}_i$	$\hat{h}_i\prod_{j=0}^{i-1}\left(1-\hat{h}_j\right)$	I	$\frac{\hat{a}_i^A-\hat{a}_{i+1}^A}{\hat{a}_1^A}$	$\frac{\hat{a}_i}{i\sum_{j=1}^{F}\frac{\hat{a}_j}{j}}$
$\hat{a}_i^A$	$\left[\sum_{i=0}^{F}\hat{S}_i\right]^{-1}\hat{S}_{i-1}$	$\left[\sum_{i=1}^{F}\prod_{j=0}^{i-1}\left(1-\hat{h}_j\right)\right]^{-1}\prod_{j=0}^{i-1}\left(1-\hat{h}_j\right)$	$\frac{1-\sum_{j=1}^{i-1}\hat{a}_j^d}{\sum_{i=1}^{F}i\hat{a}_i^d}$	I	$\sum_{k=1}^{F}\frac{\hat{a}_k}{k}-\sum_{j=1}^{i}\frac{\hat{a}_j}{j}$
$\hat{a}_i$	$i\left[\sum_{i=0}^{F}S_i\right]^{-1}\left(\hat{S}_{i-1}-\hat{S}_i\right)$	$i\prod_{j=1}^{i-1}\left(1-\hat{h}_j\right)\hat{h}_i\left[\sum_{i=1}^{F}\prod_{j=0}^{i-1}\left(1-\hat{h}_j\right)\right]^{-1}$	$\frac{\hat{a}_i^d}{i.\sum_{j=1}^{F}j\hat{a}_j^d}$	$i.\left(\hat{a}_i^A-\hat{a}_{i+1}^A\right)$	I

survival function and all three distributions in terms of the particular hazard function. However, we can also work in the opposite direction and use one of the three distributions to describe the others.

The full set of relationships is given in table (1). Each column represents the estimators of the functions or distributions: $\left\{\hat{S}_i, \hat{h}_i, \hat{a}_i^d, \hat{a}_i^A, \hat{a}_i\right\}$; each row shows how the elements can be written in terms of the elements of that column. Thus the first row has the different ways of writing the estimator of survival function $\hat{S}_i$ in terms of itself (the indicator I), the estimator of hazard function $\hat{h}_i$, and then the estimators of the three distributions $\hat{a}_i^d$, $\hat{a}_i^A$ and $\hat{a}_i$. The second row shows how we can write the estimators of the survival function, the hazard function (by itself) and the three distributions in terms of the hazard function estimators $\hat{h}_i$. It is these identities that enable us to derive the estimators of the three distributions and their variances by applying the well known estimators of the survival and hazard functions. In the appendix, we show some further relationships between the three distributions, especially at their extreme values such as $i = 0$ and $i = F$.

3 Asymptotic Variances of the Estimators of Three Distributions

In this paper, we employ the delta method to derive the variances of the estimators of DD, CSA and CSD. This method is also applied in Greenwood (1926) to derive the variance of the survival function. Since the survival function has been extensively analysed, we can use this as a natural starting place to derive the variances of the estimators of the three distributions using the identities we have derived in the previous section. We continue to assume that all spells are uncensored for our analytical results.

Assume the estimated survival function $\hat{S}_i$ converges to the mean value

S_i (the underlying survival probability). This can be expressed as:

$$\sqrt{N_i}[\hat{S}_i - S_i] \overset{a.s.}{\sim} N(0, Var(\hat{S}_i))$$

By Taylor expansion we have:

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

where $O_p(.)$ is the *Bachmann-Landau notation*. From Slutsky's theorem[8], there exists the relationship:

$$\sqrt{N_i}[g(\hat{S}_i) - g(S_i)] \overset{a.s.}{\sim} N(0, [g'(S_i)]^2 Var(\hat{S}_i))$$

3.1 Derivation of the Asymptotic Variances of the Estimators

This section contains the main results of the paper, where we derive the variances of the three non-parametric estimators: for DD (Theorem 1), CSD (Theorem 2) and CSA (Corollary 1). Starting with DD, we have the estimator:

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

With variance:

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

Theorem 1: Assume that we have the estimates of the survival function $\hat{\mathbf{S}} = (\hat{S}_1, ...\hat{S}_{F-1})$ and hazard function $\hat{\mathbf{h}} = (\hat{h}_1, \hat{h}_2, ...\hat{h}_F)$. The variances of the DD estimators $\hat{a}_i^d$ are given by:

$$\widehat{Var}(\hat{a}_i^d) = (\hat{S}_{i-1}\hat{h}_i)^2[\frac{N_i - D_i}{N_i D_i} + \sum_{k=1}^{i-1} \frac{D_k}{N_k(N_k - D_k)}] \tag{9}$$

For $i = 2, ..., F$. If $i = 1$, this simplifies to:

[8] Slutsky's theorem states that if there exist two random variables or vectors X_i and Y_i, and those variables or vectors satisfy $X_i \xrightarrow{d.} X$ and $Y_i \xrightarrow{p.} c$, then there exists the relationship:

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

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.

$$\widehat{Var}(\hat{a}_1^d) = (\hat{h}_1)^2[\frac{N_1 - D_1}{N_1 D_1}] \tag{10}$$

Theorem 1 allows us to move from our estimates of $\hat{S}_{i-1}$ and $\hat{h}_i$ to add the terms in the square bracket from the data to give us the variance of $\hat{a}_i^d$. Note also, the variance of $\hat{a}_i^d$ depends only on the data up to and including period i: neither $\hat{a}_i^d$ nor its variance depend on the distribution of spells beyond i periods.

To derive the variance of the CSD estimator, some additional formulae are needed. In equation (8), 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 $\bar{h}$. The constant value i is the length of the duration and hence we say that the CSD is weighted by length.

Breslow and Crowley (1974) showed that both the estimated survival function and the hazard function follow the normal distribution asymptotically. The off-diagonal terms in the variance-covariance matrix of the hazard function are all equal to zero. This means that the estimates $\hat{h}_i$ are asymptotically independent across durations. Note, this does not mean that the underlying true hazards are unrelated, but merely that the errors are unrelated (if one estimate is too high, it has no implications for the errors of the other estimators in large samples).

In contrast, they showed the covariances for the estimators of the survival function do not equal zero. This follows from the fact that to survive to j periods you have to pass through each of the previous i periods ($i < j$). To derive the variances of the CSD estimators, we will need first to derive the covariance of $\hat{S}_i$ and $\hat{S}_j$ for $i < j$. To do this we take the Taylor expansion for $\hat{S}_i$ and $\hat{S}_j$:

$$exp(ln\hat{S}_i) = exp(lnS_i) + (ln\hat{S}_i - lnS_i)exp(lnS_i) + O_p((ln\hat{S}_i - lnS_i)^2) \tag{11}$$

$$exp(ln\hat{S}_j) = exp(lnS_j) + (ln\hat{S}_j - lnS_j)exp(lnS_j) + O_p((ln\hat{S}_j - lnS_j)^2) \tag{12}$$

Rearranging equation (11) and (12):

$$\hat{S}_i - S_i = S_i(ln\hat{S}_i - lnS_i) + O_p((ln\hat{S}_i - lnS_i)^2) \tag{13}$$

$$\hat{S}_j - S_j = S_j(ln\hat{S}_j - lnS_j) + O_p((ln\hat{S}_j - lnS_j)^2) \tag{14}$$

If we multiply equation (13) with (14) and take the expectation:

$$\begin{aligned} Cov(\hat{S}_i, \hat{S}_j) &= E[\hat{S}_i - S_i)(\hat{S}_j - S_j)] \\ &\approx S_i S_j E[(ln\hat{S}_i - lnS_j)(ln\hat{S}_j - lnS_j)] \\ &= S_i S_j Cov(ln\hat{S}_i, ln\hat{S}_j) \\ &= S_i S_j Cov(\sum_{k=1}^{i} ln(1-\hat{h}_k), \sum_{l=1}^{j} ln(1-\hat{h}_l)) \\ &= S_i S_j Var[\sum_{k=1}^{i} ln(1-\hat{h}_k)] \end{aligned} \tag{15}$$

The delta method is applied to derive the covariance of the KM estimators in equation (15). Since $Cov(\hat{h}_k, \hat{h}_l) = 0$ for $k \neq l$, we have:

$$Cov(\sum_{k=1}^{i} ln(1-\hat{h}_k), \sum_{l=1}^{j} ln(1-\hat{h}_l)) = Var[\sum_{k=1}^{i} ln(1-\hat{h}_k)] \; for \; i < j$$

Where $Var[\sum_{k=1}^{i} ln(1-\hat{h}_k)] = \sum_{k=1}^{i} \frac{D_k}{N_k(N_k - D_k)}$, which was shown in the proof of Theorem 1. Applying the large sample properties of the maximum likelihood estimator, the estimated covariance between $\hat{S}_i$ and $\hat{S}_j$ can be written as:

$$\widehat{Cov}(\hat{S}_i, \hat{S}_j) = \hat{S}_i \hat{S}_j [\sum_{k=1}^{i} \frac{D_k}{N_k(N_k - D_k)}] \; for \; i < j \tag{16}$$

The expression above states that the covariance of the survival function estimates from two periods depends on the product of the two survival probabilities and the summation term in square brackets, which includes only data up to the shorter of the two survival durations.

In Theorem 1, we used the delta method to derive the variance of the DD estimator, and have now in addition derived the covariance of the survival functions across time. We now proceed to derive the variance of the CSD estimator, by treating the estimates $\hat{a}_i$ as a ratio distribution $\hat{x}_i/\hat{y}$ with $\hat{x}_i = i\hat{S}_{i-1}\hat{h}_i$ and $\hat{y} = \sum_{k=0}^{F} \hat{S}_k$. We can then apply the delta method for the ratio estimator $\hat{x}_i/\hat{y}$ to approximate $\hat{x}_i$ and $\hat{y}$ at the mean value x_i and y:

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

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

$$E[\frac{\hat{x}_i}{\hat{y}}] \approx \frac{x_i}{y} \tag{17}$$

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

$$Var(\frac{\hat{x}_i}{\hat{y}}) \approx \frac{Var(\hat{x}_i)}{y^2} + \frac{x_i^2}{y^4}Var(\hat{y}) - 2\frac{x_i}{y^3}Cov(\hat{x}_i, \hat{y}) \tag{18}$$

Applying the large sample properties of the maximum likelihood estimator, we can 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 $\hat{y} = \hat{S}$.[9] First, note that the variance of $\hat{S}$ is:

$$Var(\hat{S}) = Var(\sum_{i=0}^{F}\hat{S}_i) = \sum_{i=0}^{F}Var(\hat{S}_i) + 2\sum_{i\neq j}Cov(\hat{S}_i, \hat{S}_j) \tag{19}$$

In addition, the covariance of $iS_{i-1}h_i$ and S_j can be derived as:

$$Cov(i\hat{S}_{i-1}\hat{h}_i, \hat{S}_j) = iCov(\hat{S}_{i-1} - \hat{S}_i, \hat{S}_j) = i[Cov(\hat{S}_{i-1}, \hat{S}_j) - Cov(\hat{S}_i, \hat{S}_j)]$$

Therefore:

$$\begin{aligned} Cov(i\hat{S}_{i-1}\hat{h}_i, \sum_{k=1}^{F}\hat{S}_k) &= i[Cov(\hat{S}_{i-1}, \sum_{k=1}^{F}\hat{S}_k) - Cov(\hat{S}_i, \sum_{k=1}^{F}\hat{S}_k)] \\ &= i[\sum_{k=1}^{F}Cov(\hat{S}_{i-1}, \hat{S}_k) - \sum_{k=1}^{F}Cov(\hat{S}_i, \hat{S}_k)] \end{aligned} \tag{20}$$

Substituting the equation (9), (15), (19) and (20) into equation (18), we are able to state the variances of the CSD estimators:

Theorem 2 The variances of the CSD estimators can be defined as:

$$\widehat{Var}(\hat{a}_i) = i^2\frac{\widehat{Var}(\hat{S}_{i-1}\hat{h}_i)}{\hat{S}^2} + i^2\frac{\hat{S}_{i-1}^2\hat{h}_i^2\widehat{Var}(\hat{S})}{\hat{S}^4} - 2i^2\frac{\hat{S}_{i-1}\hat{h}_i\widehat{Cov}(\hat{S}_{i-1}\hat{h}_i, \hat{S})}{\hat{S}^3} \tag{21}$$

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

[9]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 $\hat{y}$

Note that $\hat{a}_i$ and its variance $\widehat{Var}(\hat{a}_i)$ depend not just on data from periods up to i, but what happens across all durations up to F. This follows because $\hat{S}$, the sum of the survival function, appears in the denominator and the numerator of both $\hat{a}_i$ and its variance $\widehat{Var}(\hat{a}_i)$. This is not only because all of the estimators $\hat{a}_i$ have to add up to unity (Proposition 2), but they are weighted by duration i. When combined, these two factors imply that what happens across the whole distribution influences each $\hat{a}_i$ and its variance $\widehat{Var}(\hat{a}_i)$. This stands in contrast to Theorem 1 for DD, where $\hat{a}_i^d$ and their variances $\widehat{Var}(\hat{a}_i^d)$ only depend on data up to i.

Since the variance of the CSD has been derived, the corresponding variance for CSA follows immediately by using equation (18):

Corollary 1 The variances of the CSA estimators are given by:

$$\widehat{Var}(\hat{a}_i^A) = \frac{\widehat{Var}(\hat{S}_{i-1})}{\hat{S}^2} + \frac{\hat{S}_{i-1}^2 \widehat{Var}(\hat{S})}{\hat{S}^4} - 2\frac{\hat{S}_{i-1}\widehat{Cov}(\hat{S}_{i-1}, \hat{S})}{\hat{S}^3} \qquad (22)$$

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

3.2 Censored and Uncensored Spells

All of the previous results are under the assumption that no spells are censored. In empirical data, we often observe some spells which are not complete. The process of collecting data will be limited in time and there may be errors. An uncensored spell is one that is observed in its entirety, from beginning to end with no break. A left-censored spell occurs when the starting point is not observed or known, being outside the period of observation (the sample period). However, the endpoint is included in the sample period. The right-censored spells are where the endpoint cannot be observed but the start can be. The KM estimator is usually applied after excluding left-censored spells, so here we will just consider the implications of having right-censored and uncensored data in the sample. The maximum length of a spell is assumed to be F periods. N is the total number of the observations. The observed lifetime t_j can be defined as follows:

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

Where the C_j means the censored time for the j-th observation; T_j is the survival time of the j-th observation. The observed lifetime t_j is the minimum

value between C_j and T_j. We also define a dummy variable ω_j indicating whether the spell is censored or not.

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

Otherwise, if the observation is uncensored:

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

In the next section, we will consider the effect of the presence of right-censored data on the estimators we have derived[10] with Monte Carlo simulations using this method.

4 Monte Carlo Simulation

In Theorems 1 and 2 and Corollary 1, the variance formulae for the estimators of CSD, CSA and DD have been derived by the delta method. In this section, we are going to investigate the properties of those formulae using Monte Carlo methods. Depending on the simulation, the accuracy of the analytic variances can be evaluated both when all data is uncensored and when there are some right-censored observations.

The simulation data is generated from the continuous time exponential distribution. The sample sizes used are $N = 25$, $N = 50$, $N = 100$ and $N = 200$. We take the raw continuous time data and then put them into each interval defined as $(0, r_1], (r_1, r_2], ...(r_{i-1}, r_i]$ with $i = 1, 2, ..., F$. [11] For $t_j \in (0, r_1]$ we set the duration at $t_j = r_1$. For $t_j \in (r_{i-1}, r_i]$ we set $t_j = r_i$ and so on. After that, we can count the number of the observations locating in each interval. The number of the observations locating in i-th interval can be defined as D_i if all the observations are uncensored. Therefore, the estimates of the survival functions and the hazard functions can be calculated for each interval. However, we will also allow for the case of right-censoring in each of the simulations, enabling us to evaluate the accuracy of the variances of the estimators of three distributions. The sample sizes used in the simulations are $N = 25$, $N = 50$, $N = 100$ and $N = 200$. The simulation process is:

[10] This method could also be extended to include left-censored data or other data imperfections.

[11] The interval $(0, r_1]$ can be defined as the "first" period, and $(r_{i-1}, r_i]$ is the "i"-th period. At this point, all the formulae are slightly different from previously result. For example, the estimator of CSD is $a_i = \frac{iS_{u_{i-1}}h_{u_i}}{\sum_{k=0}^{u_i} S_k}$

Step 1: The observed duration is $t_j = min(T_j, C_j)$ where $j = 1, 2, ..., N$. Both the lifetime time T_j and the censored time C_j follow the exponential distribution. The censored time and the lifetime have the survival functions for each k-th period:

$$p(C_j > r_i) = exp(-0.5r_i) \quad p(T_j > r_i) = exp(-2r_i) \tag{23}$$

The proportion of uncensored spells is 0.8. [12] For the case without censoring, we can ignore the censored times and just generate the survival time using the lifetimes $t_j = T_j$ and assume they are all uncensored with the right-censored coefficient $\omega_j = 1$ for all j. In other words, the observations can be written as $(T_j, 1)$ for all j.

For the case with right-censoring observations, we also need to generate the censored durations C_j and compare the two values T_j and C_j for each j. If $T_j < C_j$, the j-th observation is uncensored and we assign a parameter $\omega_j = 1$ to the j-th observation. If $C_j < T_j$, it means the observation is right-censored and $\omega_j = 0$.

Once we have generated the raw data with and without censoring, the survival data are allocated into $F = 5$ intervals (periods). We divide up the data in two different ways. In case 1, our five intervals (periods) are defined as: $(0, 0.1]$, $(0.1, 0.2]$, $(0.2, 0.3]$, $(0.3, 0.5]$, and $(0.5, \infty)$. In case 2, our five intervals (periods) are defined as: $(0, 0.2]$, $(0.2, 0.4]$, $(0.4, 0.6]$, $(0.6, 0.8]$, and $(0.8, \infty)$. We consider two cases to evaluate the accuracy of analytic variance formulae of the three distributions.

Step 2: The formulae (9), (21) and (22) are applied to calculate the variance of the estimators for DD, CSD and CSA for each period. When we have right-censored data, we apply the method described in the previous section to the same formulae.[13].

Step 3: Repeat step 1 and step 2 for M times, where we choose $M = 10,000$. If by chance there is a sample with no observations in one of the intervals, this sample is eliminated and another sample is simulated until we

[12] Since the parameter of the exponential distribution of censored time and observed time are 0.5 and 2, separately. The right-censored proportion of the total sample can be known as $0.8 = \frac{0.5}{2+0.5}$. The algebra is shown by Efron (1981).

[13] That is, we include the right censored data in N_i, but have only uncensored data in D_i

have 10,000 samples in which all 5 intervals are non-empty. Following Kiviet and Phillips (2014), we specify the benchmark value[14] of the variances for CSD as:

$$Var(\hat{a}_i)_{benchmark} = \sum_{m=1}^{M}(\hat{a}_{i,m} - \frac{\sum_{m=1}^{M}\hat{a}_{i,m}}{M})^2/(M-1) \tag{24}$$

Where $\hat{a}_{i,m}$ is the estimate from the m-th Monte Carlo simulation.

The benchmark values for CSA[15] are:

$$Var(\hat{a}_i^A)_{benchmark} = \sum_{m=1}^{M}(\hat{a}_{i,m}^A - \frac{\sum_{m=1}^{M}\hat{a}_i^A}{M})^2/(M-1) \tag{25}$$

The benchmark values for DD are:

$$Var(\hat{a}_i^d)_{benchmark} = \sum_{m=1}^{M}(\hat{a}_{i,m}^d - \frac{\sum_{m=1}^{M}\hat{a}_{i,m}^d}{M})^2/(M-1) \tag{26}$$

In other words, we collect M estimators of each of the three distributions $\hat{a}_{i,m}$, $\hat{a}_{i,m}^A$ and $\hat{a}_{i,m}^d$, from which we calculate the estimated variances.[16] Equation (24), (25) and (26) are the benchmark variances of the three distributions based on the properties of the Monte Carlo simulations. The benchmark variances are then compared with the analytic variances derived by the delta method to see whether the theoretical approximation results are close to the benchmark value.

Table (2) reports the simulation results for DD where all data are uncensored using case 1, with interval $(0, 0.1]$, $(0.1, 0.2]$, $(0.2, 0.3]$, $(0.3, 0.5]$, and $(0.5, \infty)$. As we can see from table (2), when the sample size is equal to 25, there exist the small biases for the variances (except for $Var(a_{0.5}^d)$). However, when the sample size is increased to 50, the approximation formulae of the variances perform very well for all the intervals. When the sample size is increased to either N=100 or N=200, the gaps between the benchmark

[14]The benchmark value calculated from the Monte Carlo simulation. It is very close to the true value.

[15]The CSA is the special case of the CSD, so we only provide the empirical results of CSD.

[16]In the simulation results, the coefficient i of equation (24) is ignored in the simulation process. The reason is that i is a constant parameter for each a_i.

values and the analytic values of the variances are further reduced. With the increase of the sample size, the approximation values are closer to the benchmark values when the observations are uncensored.

Table 2: The Variances of DD Estimators for Case 1 When All the Observations Are Uncensored. All the Results Are Multiplied by 10^3

.

Benchmark Values					
N	$Var(\hat{a}^d_{0.1})$	$Var(\hat{a}^d_{0.2})$	$Var(\hat{a}^d_{0.3})$	$Var(\hat{a}^d_{0.5})$	$Var(\hat{a}^d_{\infty})$
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
Approximation Values					
N	$E[\widehat{Var}(\hat{a}^d_{0.1})]$	$E[\widehat{Var}(\hat{a}^d_{0.2})]$	$E[\widehat{Var}(\hat{a}^d_{0.3})]$	$E[\widehat{Var}(\hat{a}^d_{0.5})]$	$E[\widehat{Var}(\hat{a}^d_{\infty})]$
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

Note: $Var(\hat{a}^d_i)$ is the benchmark value calculated from formula (25); $E[\widehat{Var}(\hat{a}^d_i)]$ is the variance calculated from formula (9).

In table (3), the variances for CSD are simulated by the same process. There still exist the small biases for the variances for CSD when the sample size is $N = 25$. When the sample size is increased to 50, all the approximated results are improved and they are all close to the benchmark values. With respect to N=100 and N=200, the approximations of the variances tend to be closer to the benchmark variances. However, it can be found that the approximations of the variances do not always overestimate the benchmark values. In conclusion, the biases of the approximated variances of CSD estimators are reduced with the increase of the sample size.

Next, we consider the case with right-censored observations. Table (4) shows the the variances of the DD estimators. Compared with the benchmark values, there exist the small biases in the variances calculated from the analytic formulae when the sample size N=25. When sample size is increased to 50, the analytic variances perform well. When the sample size tends to be larger (N=100 and N=200), the empirical results show that the

Table 3: The Variances of the CSD Estimators for Case 1 When All the Observations Are Uncensored. All the Results Are Multiplied by 10^3

Benchmark Values

N	$Var(\hat{a}_{0.1})$	$Var(\hat{a}_{0.2})$	$Var(\hat{a}_{0.3})$	$Var(\hat{a}_{0.5})$	$Var(\hat{a}_{\infty})$
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

Approximation Values

N	$E[\widehat{Var}(\hat{a}^d_{0.1})]$	$E[\widehat{Var}(\hat{a}^d_{0.2})]$	$E[\widehat{Var}(\hat{a}^d_{0.3})]$	$E[\widehat{Var}(\hat{a}^d_{0.5})]$	$E[\widehat{Var}(\hat{a}^d_{\infty})]$
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

Note: $Var(\hat{a}_i)$ is the benchmark value calculated from formula (24); $E[\widehat{Var}(\hat{a}_i)]$ is the variance calculated from formula (21).

approximations of the variances are nearly the same as the benchmark values. In conclusion, the estimated variances form the analytic formulae are close to the benchmark values even when the sample size is small(N=25). The approximated variances may overestimate or underestimate the benchmark variances.

Table (5) shows the simulation results of the CSD variances with right-censored data. When the sample size is extremely small (N=25), the approximations of the variances are still quite accurate. When the sample size is increased to 50, all the analytic variances are improved. They are all close to the benchmark values. When the sample size is larger (N=100 and N=200), the analytic variances are very close to the benchmark values. Therefore, the analytic formulae of the variances of the CSD perform well even when there are right-censored observations.

In tables (6) to (9), the variances for DD and CSD are presented under the alternative assumption of case 2 where all the data are assigned into the five intervals: $(0, 0.2]$, $(0.2, 0.4]$, $(0.4, 0.6]$, $(0.6, 0.8]$, and $(0.8, \infty)$. Otherwise, the simulations are carried out as before.

From tables (6) and (7), we can see that the analytic formulae of the variances can give the accurate approximations for the benchmark values

Table 4: The Variances of the DD Estimators for Case 1 When the Right-Censored Observations Exist in the Samples. All the Results Are Multiplied by 10^3

Benchmark Values

N	$Var(\hat{a}^d_{0.1})$	$Var(\hat{a}^d_{0.2})$	$Var(\hat{a}^d_{0.3})$	$Var(\hat{a}^d_{0.5})$	$Var(\hat{a}^d_{\infty})$
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

Approximation Values

N	$E[\widehat{Var}(\hat{a}^d_{0.1})]$	$E[\widehat{Var}(\hat{a}^d_{0.2})]$	$E[\widehat{Var}(\hat{a}^d_{0.3})]$	$E[\widehat{Var}(\hat{a}^d_{0.5})]$	$E[\widehat{Var}(\hat{a}^d_{\infty})]$
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

Note: $Var(a^d_i)$ is the benchmark value calculated from formula (25); $E[\widehat{Var}(a^d_i)]$ is the variance calculated from formula (9).

even in the extremely small sample size (N=25). Both the variances of the DD and CSD estimators are either overestimated or underestimated without a systematic bias. When the sample size tends to be a large number, they are nearly unbiased from the benchmark values. When there are right-censored observations in the samples, the same conclusion can be seen to hold in table (8) and table (9) as we saw in case 1.

Table 5: The Variances of the CSD Estimators for Case 1 When the Right-Censored Observations Exist in the Samples. All the Results Are Multiplied by 10^3

Benchmark Values

N	$Var(\hat{a}_{0.1})$	$Var(\hat{a}_{0.2})$	$Var(\hat{a}_{0.3})$	$Var(\hat{a}_{0.5})$	$Var(\hat{a}_{\infty})$
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

Approximation Values

N	$E[\widehat{Var}(\hat{a}^d_{0.1})]$	$E[\widehat{Var}(\hat{a}^d_{0.2})]$	$E[\widehat{Var}(\hat{a}^d_{0.3})]$	$E[\widehat{Var}(\hat{a}^d_{0.5})]$	$E[\widehat{Var}(\hat{a}^d_{\infty})]$
25	0.6798	0.5162	0.4174	0.5094	0.5824
50	0.3345	0.2557	0.2032	0.2660	0.2992
100	0.1648	0.1285	0.1025	0.1345	0.1513
200	0.0811	0.0636	0.0511	0.0676	0.0763

Note: $Var(\hat{a}_i)$ is the benchmark value calculated from formula (24); $E[\widehat{Var}(\hat{a}_i)]$ is the variance calculated from formula (21).

Table 6: The Variances of the DD Estimators for Case 2 When All the Observations Are Uncensored. All the Results Are Multiplied by 10^3

Benchmark Values

N	$Var(\hat{a}^d_{0.2})$	$Var(\hat{a}^d_{0.4})$	$Var(\hat{a}^d_{0.6})$	$Var(\hat{a}^d_{0.8})$	$Var(\hat{a}^d_{\infty})$
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

Approximation Values

N	$E[\widehat{Var}(\hat{a}^d_{0.2})]$	$E[\widehat{Var}(\hat{a}^d_{0.4})]$	$E[\widehat{Var}(\hat{a}^d_{0.6})]$	$E[\widehat{Var}(\hat{a}^d_{0.8})]$	$E[\widehat{Var}(\hat{a}^d_{\infty})]$
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

Note: $Var(\hat{a}^d_i)$ is the benchmark value calculated from formula (25); $E[\widehat{Var}(\hat{a}^d_i)]$ is the variance calculated from formula (9).

Table 7: The Variances of the CSD Estimators for Case 2 When All the Observations Are Uncensored. All the Results Are Multiplied by 10^3

Benchmark Values

N	$Var(\hat{a}_{0.2})$	$Var(\hat{a}_{0.4})$	$Var(\hat{a}_{0.6})$	$Var(\hat{a}_{0.8})$	$Var(\hat{a}_{\infty})$
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

Approximation Values

N	$E[\widehat{Var}(\hat{a}^d_{0.2})]$	$E[\widehat{Var}(\hat{a}^d_{0.4})]$	$E[\widehat{Var}(\hat{a}^d_{0.6})]$	$E[\widehat{Var}(\hat{a}^d_{0.8})]$	$E[\widehat{Var}(\hat{a}^d_{\infty})]$
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

Note: $Var(\hat{a}_i)$ is the benchmark value calculated from formula (24); $E[\widehat{Var}(\hat{a}_i)]$ is the variance calculated from formula (21).

Table 8: The Variances of the DD Estimators for Case 2 When the Right-Censored Observations Exist in the Samples. All the Results Are Multiplied by 10^3

Benchmark Values

N	$Var(\hat{a}^d_{0.2})$	$Var(\hat{a}^d_{0.4})$	$Var(\hat{a}^d_{0.6})$	$Var(\hat{a}^d_{0.8})$	$Var(\hat{a}^d_{\infty})$
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

Approximation Values

N	$E[\widehat{Var}(\hat{a}^d_{0.2})]$	$E[\widehat{Var}(\hat{a}^d_{0.4})]$	$E[\widehat{Var}(\hat{a}^d_{0.6})]$	$E[\widehat{Var}(\hat{a}^d_{0.8})]$	$E[\widehat{Var}(\hat{a}^d_{\infty})]$
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

Note: $Var(\hat{a}^d_i)$ is the benchmark value calculated from formula (25); $E[\widehat{Var}(\hat{a}^d_i)]$ is the variance calculated from formula (9).

Table 9: The Variances of the CSD Estimators for Case 2 When the Right-Censored Observations Exist in the Samples. All the Results Are Multiplied by 10^3

Benchmark Values

N	$Var(\hat{a}_{0.2})$	$Var(\hat{a}_{0.4})$	$Var(\hat{a}_{0.6})$	$Var(\hat{a}_{0.8})$	$Var(\hat{a}_{\infty})$
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

Approximation Values

N	$E[\widehat{Var}(\hat{a}^d_{0.2})]$	$E[\widehat{Var}(\hat{a}^d_{0.4})]$	$E[\widehat{Var}(\hat{a}^d_{0.6})]$	$E[\widehat{Var}(\hat{a}^d_{0.8})]$	$E[\widehat{Var}(\hat{a}^d_{\infty})]$
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

Note: $Var(a_i)$ is the benchmark value calculated from formula (24); $E[\widehat{Var}(a_i)]$ is the variance calculated from formula (21).

5 Two Applications to Data

Having explored the properties of the estimators using Monte Carlo simulations, it can be seen that the analytic variances of the non-parametric estimators are quite accurate. In this section, we apply our methods to real UK data: the first application is price-quote data from the Office for National Statistics (ONS); the second application is waiting-time data from the National Health Service (NHS). These examples show how we can use different types of duration data to estimate the two distributions DD and CSD, and corresponding variances. With the price-quote data, we start with the raw data of completed durations to estimate the variances of the estimators for DD and CSD. With the waiting time data, we start with data on incomplete durations to estimate the same statistics.

5.1 Price-Quote Data

Our first application is using the underlying UK micro price data which is used to construct the CPI inflation statistics. This data set gives price-quotes each month across over 700 items sampled from different sellers across the UK in order to measure CPI inflation: there are over 100,000 price-quotes collected each month. The period we have chosen is 1996-2007 (inclusive), the great moderation prior to the Great Financial Crisis, which includes over 20 million price quotes (for a more detailed description of the data see Dixon and Tian (2017) and Dixon et al. (2020)).

From the price-quotes, we can construct price-spells: these are the sequences of monthly quotes where an individual price-setter sets the same price each month. Our data includes 319,784 price-spells, which are sorted into durations of 1-48 months. For $i = 1, 2, ..., 47, 48$, a price-spell has duration i if it lasts exactly i months. For example if the duration is 12 months, that means we observe twelve consecutive months where the seller set the same price and in both the preceding and following month set a different price. The only exception is 48 months, where all spells of 48 months or longer are counted. It is common practice to truncate the distribution in this way. For applications of this type of data and distributions in dynamic macroeconomic models, see for example Dixon and Le Bihan (2012) using French CPI data and Dixon (2012) using the same UK data. Following these two papers, we do not use left-censored spells and assume that right-censored spells end with price-changes. In this paper, the first 5 months of first year and fourth year of those distributions and their variance are reported.

Table 10: The DD and CSD Estimators and Their Variances for UK Micro-CPI Data in the First 5 Months of the First Year

DD				
$\hat{a}_1^d$	$\hat{a}_2^d$	$\hat{a}_3^d$	$\hat{a}_4^d$	$\hat{a}_5^d$
0.4113	0.1703	0.0979	0.0682	0.0462
$\widehat{Var}(\hat{a}_1^d)$	$\widehat{Var}(\hat{a}_2^d)$	$\widehat{Var}(\hat{a}_3^d)$	$\widehat{Var}(\hat{a}_4^d)$	$\widehat{Var}(\hat{a}_5^d)$
$7.6270 * 10^{-8}$	$4.4513 * 10^{-8}$	$2.7821 * 10^{-7}$	$2.0019 * 10^{-8}$	$1.3878 * 10^{-8}$
$std.(\hat{a}_1^d)$	$std.(\hat{a}_2^d)$	$std.(\hat{a}_3^d)$	$std.(\hat{a}_4^d)$	$std.(\hat{a}_5^d)$
0.0002762	0.0002110	0.0001668	0.0001415	0.0001178
CSD				
$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_4$	$\hat{a}_5$
0.1036	0.0858	0.0740	0.0687	0.0582
$\widehat{Var}(\hat{a}_1)$	$\widehat{Var}(\hat{a}_2)$	$\widehat{Var}(\hat{a}_3)$	$\widehat{Var}(\hat{a}_4)$	$\widehat{Var}(\hat{a}_5)$
$1.5661 * 10^{-8}$	$1.7549 * 10^{-8}$	$1.9664 * 10^{-8}$	$2.2830 * 10^{-8}$	$2.3271 * 10^{-8}$
$std.(\hat{a}_1)$	$std.(\hat{a}_2)$	$std.(\hat{a}_3)$	$std.(\hat{a}_4)$	$std.(\hat{a}_5)$
0.0001251	0.0001325	0.0001402	0.0001511	0.0001525

Table 11: The DD and CSD Estimators and Their Variances for UK Micro-CPI Data in the First 5 Months of the Fourth Year

DD				
$\hat{a}_{37}^d$	$\hat{a}_{38}^d$	$\hat{a}_{39}^d$	$\hat{a}_{40}^d$	$\hat{a}_{41}^d$
0.0003	0.0003	0.0002	0.0003	0.0002
$(\hat{a}_{37}^d)$	$\widehat{Var}(\hat{a}_{38}^d)$	$\widehat{Var}(\hat{a}_{39}^d)$	$\widehat{Var}(\hat{a}_{40}^d)$	$\widehat{Var}(\hat{a}_{41}^d)$
$9.4508 * 10^{-11}$	$9.4569 * 10^{-11}$	$6.2925 * 10^{-11}$	$9.4413 * 10^{-11}$	$6.3025 * 10^{-11}$
$std.(\hat{a}_{37}^d)$	$std.(\hat{a}_{38}^d)$	$std.(\hat{a}_{39}^d)$	$std.(\hat{a}_{40}^d)$	$std.(\hat{a}_{41}^d)$
0.000009722	0.000009725	0.000007933	0.000009717	0.000007939
CSD				
$\hat{a}_{37}$	$\hat{a}_{38}$	$\hat{a}_{39}$	$\hat{a}_{40}$	$\hat{a}_{41}$
0.002795	0.002871	0.001964	0.003022	0.002065
$\widehat{Var}(\hat{a}_{37})$	$\widehat{Var}(\hat{a}_{38})$	$\widehat{Var}(\hat{a}_{39})$	$\widehat{Var}(\hat{a}_{40})$	$\widehat{Var}(\hat{a}_{41})$
$8.1690 * 10^{-9}$	$8.6207 * 10^{-9}$	$6.0507 * 10^{-9}$	$9.5334 * 10^{-9}$	$6.6966 * 10^{-9}$
$std.(\hat{a}_{37})$	$std.(\hat{a}_{38})$	$std.(\hat{a}_{39})$	$std.(\hat{a}_{40})$	$std.(\hat{a}_{41})$
0.00009036	0.00009280	0.00007782	0.00009767	0.00008181

Table (10) shows the results for the first 5 months of the two distributions (DD and CSD). In the first row we have the estimates of the probabilities, in the second row their variances and in the third the implied standard deviations. Table (11) gives the results for another 5 months (37-41) towards the end. The small variance observed is because of the large sample size. In the

first 5 months, the sample size is very large, as over 60 per cent last no more than 4 months. However, even in the later months 37-41 when the estimated coefficients of DD and CSD are small, their variances are also small. The values of the estimators of DD and CSD are more than twice the standard deviation implied by the estimated variances and hence significantly different from zero.[17]

5.2 Hospital Waiting Times

The second data set is about hospital waiting times. This is generated from a census taken at particular dates and gives the length of time patients have been waiting for a particular type of procedure up to that date. It is in effect data on the age distribution, since the time waited is an incomplete duration. We can the use this age distribution to estimate the two other distributions: DD and CSD. As discussed in Dixon and Siciliani (2009) and Siciliani et al. (2013), although the raw data collected is in the form of the age distribution (non-completed waiting times at a point in time), it is often more useful to interpret the data in terms of completed waiting times (both DD and CSD) for health policy.

The data used is the same as used in Dixon and Siciliani (2009). It covers three financial years: 2004-5, 2005-6 and 2007-8 with more than 800,000 patients involved each year. The data is collected weekly by the NHS (hospitals report the data to the NHS who then published it). We use the weekly data of 2007-8 for this exercise. We were able to use 625,960 observations in our estimation. From the waiting time data, we estimate both DD and CSD for weeks 6-10 and 21-25 and the corresponding variances of these estimators. The empirical results are reported in table (12) and (13). As before, the standard deviation reported is here implied by the estimated variance so that we can easily interpret the significance of the DD and CSD estimators.

[17] As Cox (1990) and Franz (2007) have shown, the delta method is a robust method for calculating the confidence interval for the ratio variable if the coefficient of variation, CV, of denominator of the ratio variable is a small value, where $CV = \sigma/\mu$. In the CPI micro-data, $CV = 0.0061795$. Since we use the delta method, we can interpret the ratio of the estimator to its standard deviation as a t-statistic, demonstrating that it is significantly different from zero. In Tian and Dixon (2022), we evaluate the empirical size for the delta approximation of CSD estimator. The empirical results indicate that delta method is valid to do the null hypothesis test and construct the confidence interval for CSD estimator by using the critical value from student-t test.

Table 12: The DD and CSD Estimators and Their Variances for UK Waiting Times in Weeks 5-10

DD				
$\hat{a}_6^d$	$\hat{a}_7^d$	$\hat{a}_8^d$	$\hat{a}_9^d$	$\hat{a}_{10}^d$
0.08686	0.09918	0.03515	0.07538	0.05516
$\widehat{Var}(\hat{a}_6^d)$	$\widehat{Var}(\hat{a}_7^d)$	$\widehat{Var}(\hat{a}_8^d)$	$\widehat{Var}(\hat{a}_9^d)$	$\widehat{Var}(\hat{a}_{10}^d)$
$1.2671 * 10^{-7}$	$1.4273 * 10^{-7}$	$5.4181 * 10^{-8}$	$1.1134 * 10^{-7}$	$8.3260 * 10^{-8}$
$std.(\hat{a}_6^d)$	$std.(\hat{a}_7^d)$	$std.(\hat{a}_8^d)$	$std.(\hat{a}_9^d)$	$std.(\hat{a}_{10}^d)$
0.0003560	0.0003778	0.0002328	0.0003337	0.0002885
CSD				
$\hat{a}_6$	$\hat{a}_7$	$\hat{a}_8$	$\hat{a}_9$	$\hat{a}_{10}$
0.05592	0.07450	0.03017	0.07280	0.05919
$\widehat{Var}(\hat{a}_6)$	$\widehat{Var}(\hat{a}_7)$	$\widehat{Var}(\hat{a}_8)$	$\widehat{Var}(\hat{a}_9)$	$\widehat{Var}(\hat{a}_{10})$
$5.7707 * 10^{-8}$	$8.7831 * 10^{-8}$	$4.0812 * 10^{-8}$	$1.0718 * 10^{-7}$	$9.6872 * 10^{-8}$
$std.(\hat{a}_6)$	$std.(\hat{a}_7)$	$std.(\hat{a}_8)$	$std.(\hat{a}_9)$	$std.(\hat{a}_{10})$
0.0002402	0.0002964	0.0002020	0.0003274	0.0003112

Table 13: The DD and CSD Estimators and Their Variances for UK Waiting Times in Weeks 21-25

DD				
$\hat{a}_{21}^d$	$\hat{a}_{22}^d$	$\hat{a}_{23}^d$	$\hat{a}_{24}^d$	$\hat{a}_{25}^d$
0.0003126	0.0005612	0.003826	0.005195	0.004257
$\widehat{Var}(\hat{a}_{21}^d)$	$\widehat{Var}(\hat{a}_{22}^d)$	$\widehat{Var}(\hat{a}_{23}^d)$	$\widehat{Var}(\hat{a}_{24}^d)$	$\widehat{Var}(\hat{a}_{25}^d)$
$4.9790 * 10^{-9}$	$8.9154 * 10^{-9}$	$6.0890 * 10^{-9}$	$8.2565 * 10^{-9}$	$6.7725 * 10^{-9}$
$std.(\hat{a}_{21}^d)$	$std.(\hat{a}_{22}^d)$	$std.(\hat{a}_{23}^d)$	$std.(\hat{a}_{24}^d)$	$std.(\hat{a}_{25}^d)$
0.00007056	0.00009442	0.00007803	0.00009087	0.00008230
CSD				
$\hat{a}_{21}$	$\hat{a}_{22}$	$\hat{a}_{23}$	$\hat{a}_{24}$	$\hat{a}_{25}$
0.007045	0.013252	0.009443	0.01338	0.01142
$\widehat{Var}(\hat{a}_{21})$	$\widehat{Var}(\hat{a}_{22})$	$\widehat{Var}(\hat{a}_{23})$	$\widehat{Var}(\hat{a}_{24})$	$\widehat{Var}(\hat{a}_{25})$
$2.5109 * 10^{-8}$	$4.9012 * 10^{-8}$	$3.6716 * 10^{-8}$	$5.3950 * 10^{-8}$	$4.8104 * 10^{-8}$
$std.(\hat{a}_{21})$	$std.(\hat{a}_{22})$	$std.(\hat{a}_{23})$	$std.(\hat{a}_{24})$	$std.(\hat{a}_{25})$
0.0001585	0.0002214	0.0001916	0.0002323	0.0002193

As can be seen, all of the DD and CSD estimators are significant if we use the standard deviations implied by our estimated variances.[18] Whilst the

[18] As with the CPI data, the $CV = 0.01071$ is small.

DD are declining in weeks 6-10, it is not so in weeks 21-25. The estimated variances are not monotonic in weeks 6-10 or 21-25 for either distribution.

Both of the examples used of price-spells and hospital waiting times were from large data sets, with low estimated variances relative to the corresponding DD and CSD estimators. We could also consider small sample size and explore the issues of estimating the variances in that context (as in the previous section with Monte Carlo simulations). However, that is a matter that lies beyond this paper as it would be specific to the type of data used and how it was collected. In an online appendix we show the complete range of the DD and CSD estimators and their variances in both table form and graphs.

6 Conclusion

In this paper, we use the delta method to derive the variances of the estimators of the distribution of durations DD, and the two cross-sectional distributions of ages CSA and (completed) durations CSD. Whilst both CSD and CSA are cross-sectional distributions, they can be also be applied to the case of a single cohort of data rather than a panel. Depending on the asymptotic approximations of the variances, we provide the analytic formulae to calculate the variances of the estimators of the three distributions. The asymptotic variances derived from the delta method are easy to understand since they are the same derivations as the Greenwood formula. In addition, in this paper we derive the covariance between different estimated survival probabilities in a simpler way than Breslow and Crowley (1974).

We used Monte Carlo simulations to investigate the accuracy of the asymptotic variances for the three distributions. The Monte Carlo results show that the analytic formulae of the variances of the DD and CSD estimators become more accurate as the sample size increases. In other words, the bias between the approximations and the benchmark values are reduced as the sample size increases. This is true even in the presence of censored data.

In addition, the variance formulae are applied in two data sets: micro price-quote data and hospital waiting times. The variances and implied standard deviations of DD and CSD estimators are reported from the two data sets. Furthermore, we show how the estimated variances can be used to judge whether the estimated coefficients of the distributions are significant.

For further study, it will be useful to see if the bootstrap corrected variances can provide a better result compared with the asymptotic formulae in the case of a small sample size. Another extension is to derive the confidence intervals for the estimators of the three distributions.

7 Appendix.

Recall that all of the theoretical results proven in this appendix are derived under the assumption that all spells are uncensored.

7.1 Proof of Proposition 1

To see why, note that:

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

7.2 Proof of Proposition 2

To see why Proposition 2 holds, note that

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

7.3 Proof of Theorem 1

The first-order Taylor expansion can be applied to derive the variance of the estimators of DD. Recall the DD formula $\hat{a}_i^d = \hat{S}_{i-1}\hat{h}_i$. The Taylor series for $\hat{a}_i^d$ at $\hat{S}_{i-1}$=S_{i-1} and $\hat{h}_i$=h_i is:

$$\hat{S}_{i-1}\hat{h}_i \approx S_{i-1}h_i + h_i(\hat{S}_{i-1} - S_{i-1}) + S_{i-1}(\hat{h}_i - h_i) \tag{27}$$

Recalling equations (11) and (12), we can use the same method to obtain the following equations:

$$\hat{S}_{i-1} - S_{i-1} = S_i(ln\hat{S}_{i-1} - lnS_{i-1}) + O_p((ln\hat{S}_{i-1} - lnS_{i-1})^2) \tag{28}$$

$$\hat{h}_i - h_i = h_i(ln\hat{h}_i - lnh_i) + O_p((ln\hat{h}_i - lnh_i)^2) \tag{29}$$

From equations (28) and (29), we can rewrite equation (27) as:

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

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

$$Var(\hat{a}_i^d) = Var(\hat{S}_{i-1}\hat{h}_i) \cong (S_{i-1}h_i)^2[Var(\ln \hat{S}_{i-1}) + Var(\ln \hat{h}_i)] \tag{30}$$

From 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.[19] Therefore, $Var(\hat{a}_i^d)$ can be written as:

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

Note that this approximation assumes that $\hat{S}_{i-1}$ and $\hat{h}_i$ are independent and that the covariance between $(\ln \hat{S}_{i-1})$ and $(\ln \hat{h}_i)$ is zero.

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

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

[19] As shown in the online appendix, the KM estimator is a maximum likelihood estimator so that $\hat{S}_{i-1}$ 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. In the online appendix, we show how the KM estimator can be derived as an MLE.

If we interpret the hazard probability as a Bernoulli probability, we can assume that the "failures" D_i follow the binomial distribution with parameters N_i and $\hat{h}_i$, so that $Var(D_i) = N_i\hat{h}_i(1-\hat{h}_i)$. It can be shown that $Var(\hat{h}_i) = Var(\frac{D_i}{N_i}) = \hat{h}_i(1-\hat{h}_i)/N_i$. Taking the first-order Taylor expansion we can obtain the following two equations:

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

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

We can rewrite these two formulae as:

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

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

Under the assumption that the observations D_i are independent with each other, the variance of $\hat{h}_i$ can be written as:

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

Applying the large sample properties of the maximum likelihood estimator this becomes:

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

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

We have a formula for the exponential function:

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

Substitute equation (31) and (31) into equation (30):

$$\widehat{Var}(\hat{a}_i^d) = (\hat{S}_{i-1}\hat{h}_i)^2[\frac{N_i - D_i}{N_i D_i} + \sum_{k=1}^{i-1}\frac{D_k}{N_k(N_k-D_k)}] \tag{31}$$

7.4 Identities of The Three Distributions

In this section we point out some identities linking the the three distributions. Whilst these are not used explicitly in the paper, these identities will also apply to the estimators. As in section 2, we present the results under the assumption that all spells are uncensored. First, recall the estimators (a_i^d, a_i^A, a_i) $i = 1...F$ from section 2:

$$a_i^d = S_{i-1}h_i$$

$$a_i^A = S_{i-1}\bar{h}$$

$$a_i = i\bar{h}S_{i-1}h_i$$

Hence for $i = 1...F$

$$a_i = ih_i a_i^A = i\bar{h}a_i^d$$

Turning first to the shortest duration $i = 1$, since $S_0 = 1$ we have:

$$a_1^d = h_1$$

$$a_1^A = \bar{h}$$

Hence:

$$a_1 = \bar{h}h_1 = a_1^A a_1^d$$

Turning next to the longest duration, we have $h_F = 1$, so that:

$$a_F^d = S_{F-1}$$

$$a_F^A = \bar{h}S_{F-1} = \bar{h}a_F^d$$

$$a_F = F\bar{h}S_{F-1} = i\bar{h}a_F^d = Fa_F^A$$

Notice that $\bar{h} < 1$ except in the degenerate case of $F = 1$. Assuming $F > 1$, for durations $i = 1, 2, ..., F-1$, we have the following relationship between DD and CSD:

$$i < \bar{h}^{-1} \rightarrow a_i < a_i^d$$

$$i > \bar{h}^{-1} \rightarrow a_i > a_i^d$$

$$i = \bar{h}^{-1} \rightarrow a_i = a_i^d$$

This implies that there is a unique "cross-over" point for the two distributions a_i and a_i^d. If we compare a_i and a_i^A, it is slightly more complicated, since the hazard function h_i varies with i and its reciprocal can be arbitrarily large (since we can have h_i at or close to zero for any $i < F$).

References

Aalen, O. O. (1978). Nonparametric 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, C. (1990). Fieller theorem, the likelihood and the delta method. *Oxford Bulletin of Economics*, 46(3):709–718.

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 Le Bihan, H. (2012). Generalised Taylor and generalised Calvo price and wage setting: micro-evidence with macro implications. *Economic Journal*, 122(560):532–554.

Dixon, H., Luintel, K., and Tian, K. (2020). The impact of the 2008 crisis on UK prices: what we can learn from the CPI microdata. *Oxford Bulletin of Economics*, 82(6):1322–1341.

Dixon, H. and Siciliani, L. (2009). Waiting-time targets in the healthcare sector: how long are we waiting ? *Journal of Health Economics*, 28:1081–1098.

Dixon, H. and Tian, K. (2017). What we can learn about the behavior of firms from the average monthly frequency of price-changes: an application to the UK CPI data. *Oxford Bulletin of Economics*, 79(6):907–932.

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.

Franz, V. (2007). Ratios: A short guide to confidence limits and proper use. *Working Paper*.

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.

Kiviet, J. F. and Phillips, G. D. A. (2014). Improved variance estimation of maximum likelihood estimators in stable first-order dynamic regression models. *Computational Statistics and Data Analysis*, 76:424–448.

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.

Siciliani, L., Borowitz, M., and Moran, V. (2013). *Waiting Time Policies in the Health Sector: What Works?* OECD Health Policy Studies, OECD Publishing, Paris.

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

Taylor, J. B. (2016). *The Staying Power of Staggered Wage and Price Setting Models in Macroeconomics*, volume 2. Elsevier.

Tian, M. and Dixon, H. (2022). The confidence interval of cross-sectional distribution of durations. *Working Paper*.