CURE RATE ESTIMATION UNDER CASE-1 INTERVAL CENSORING VIA SMOOTHING

Mehrnoosh Malekiha

A Thesis

in

the Department

of

Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Science (Mathematics) at

Concordia University

Montreal, Quebec, Canada

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Mehrnoosh Malekiha**

Entitled: **Cure rate estimation under case-1 interval censoring via smoothing**

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Mathematics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Examiner

Dr. Yogendra P. Chaubey

_____ Examiner

Dr. Xiaowen Zhou

_____ Thesis Supervisor

Dr. Arusharka Sen

Approved by _____

Chair of Department or Graduate Program Director

_____

Dean of Faculty

Date _____

Abstract

Estimating cure-rate is a popular research subject in testing the reliability of treatment for terminal diseases such as cancers and HIV. So far, the publications mainly proposed estimation techniques based on parametric methods, with a few exceptions.

In this thesis, under case-1 interval censoring, we develop and propose two novel non-parametric estimators that improve upon previously proposed estimation techniques (Sen and Tan, 2008), using smoothing. We show our estimators are strongly consistent. In addition, their asymptotic normality is studied and we applied proposed estimator to estimate cure-rate on data collected for lung tumor in mice (Finkelstein and Wolfe, 1985). Finally, the smoothing parameter for optimum estimation has been determined by using Jackknife method.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my adviser Dr. Arusharka Sen for his kindness, guidance and patience throughout my research which has been enormously helpful.

I would also like to thank my committee members Dr. Chaubey and Dr.Zhou for their careful reading and suggestion especially Dr. Chaubey because of his valuable work which was my light throughout this study.

Special gratitude is given to my dear husband Mahdi who helped and encouraged me along the way.

I can not end without thanking my parents on whose constant love I have relied throughout all my time. I am indebted to all of you for your help.

Notation and abbreviations

a.s.	Almost surely
BCH	Bounded cumulative hazard
CDF	Cumulative distribution function
CLT	Central limit theorem
iid	Independent and identically distributed
MSE	Mean square error
NPMLE	Non-parametric maximum-likelihood estimator

Contents

List of Figures

Chapter 1

Introduction and Preliminaries

Cure-rate plays an important role in determining reliability of treatment for terminal diseases such as cancers, human immunodeficiency virus (HIV), etc and often needs to be considered in analyzing medical data. However, there are some technical issues left in analyzing of cure-rate such as censoring which refers to a subject who does not experience the event during monitoring time, and we may never observe cure due to the finite monitoring time. Estimating the probability of cure can help scientists to shed light on currently unknown factors relating terminal diseases. In this thesis we develop two estimators for cure-rate under case-1 interval censoring (Gu *et al.*, 2011).

1.1 Survival Analysis

Survival function is the probability of surviving beyond some time point, x. Traditionally in survival analysis, it is assumed that every case of study is susceptible to the event of interest and eventually experiences it (Maller and Zhou, 1996). Survival function is an essential tool to work with medical data (Vij, 2014) and can be define

as:

$$S(x) = P\{X \geq x\} \tag{1.1}$$

1.2 Cure-Rate

Cure-rate defined as the probability that an individual survives or is immune to a terminal disease. In clinical trials, estimation of cure-rate becomes important when a significant proportion of individuals is assumed not to experience the event of interest. Cure-rate is usually defined by (Vij, 2014):

$$p \stackrel{\Delta}{=} P(X = \infty) = S(\infty) \tag{1.2}$$

where S and X are a survival function and time-to-event of interest, respectively.

There are two well known models that have been used to model cure rate: 1) mixture model (Berkson and Gage, 1952) and 2) bounded cumulative hazard (BCH) model (Tsodikov, 1998).

1.2.1 Mixture Model

Mixture model is a non-parametric two component (binary) model that is being used to analyze cure-rate in clinical trails (Sen and Tan, 2008). In this model, population is divided in to two groups. In the first group, probability of cure is equal to p but the others are susceptible to the event and at some point will experience the event with probability $1 - p$ and proper survival function $S_0(t)$

$$S_0(t) = Pr\{X \geq t | X < \infty\}. \tag{1.3}$$

where X is a non-negative random variable denoting the lifetime of an individual.

If F_0 and F_p denote the cumulative distribution function (CDF) for uncured patients and whole population respectively, survival function can be calculated as follow

$$S_p(t) = 1 - F_p(t) = p + (1-p)(1 - F_0(t)) \tag{1.4}$$

where

$$p = P\{X = \infty\} = lim_{t\to\infty} P\{X > t\} = 1 - lim_{t\to\infty} F_p(t) \tag{1.5}$$

and

$$F_p(t) = (1-p)F_0(t) \Leftrightarrow F_0(t) = \frac{F_p(t)}{1-p} = P\{X \leqslant t \mid X < \infty\}. \tag{1.6}$$

This model has few drawbacks when it involves with covariates such that not being able to have proportional hazard structure which is required in many asymptotic and computational result.

1.2.2 Bounded Cumulative Hazard (BCH) Model

BCH Model is another general representation of an improper survival time distribution (i.e. survival function with probability of cure) (Tsodikov, 1998). The motivation of proposing BCH model was biological application and drawback which is mentioned regarding to mixture model(Chen *et al.*, 1999). In this model, it has been assumed that for an individual in a population, the number of cancer cells, which are still active after first treatment, is N. Assuming N has a Poisson distribution with mean

θ and the random time related to the ith cancer cell is Z_i, $i = 1, 2, \cdots$ which are iid with common distribution

$$F(t) = 1 - S(t) \tag{1.7}$$

also $Z_i \perp N$.

Furthermore, the time of deteriorate of cancer can be determined by the random variable

$$X = \min\{Z_i, 0 \leq i \leq N\} \tag{1.8}$$

where $P(Z_0 = \infty) = 1$ and $Z_i \perp N$.

Therefore, the survival function for X and population can be derived as follows

$$\begin{aligned} S_p(t) &= P(\text{no cancer by time t}) \\ &= P\{N = 0\} + P(Z_1 > t, Z_2 > t, \cdots, Z_n > t, N \geq 1) \\ &= P(-\theta) + \sum_{n=1}^{\infty} S(t)\theta^k/k!e^{-\theta} \\ &= exp(-\theta + \theta S(t)) = exp(-\theta F(t)) \end{aligned} \tag{1.9}$$

since $S_p(\infty) = exp\{-\theta\} > 0$ is not a proper survival function (Chen *et al.*, 1999).

Finally, we want to point out that both mixture and BCH model are equal in non-parametric frame work to estimate cure-rate. However, the situation is different when $S_0(x)$ is defined parametrically (Sen and Tan, 2008).

1.3 Censoring

Censoring is one of the analytical problems in modeling cure-rate. It corresponds to cases that the information about exact survival-time is incomplete (Kleinbaum and

Klein, 2006). The following common forms of censoring:

1) Right censoring occurs when a subject leaves the study or its information is lost before the occurrence of the event interest or end of study which is illustrated in Fig. 1.1.

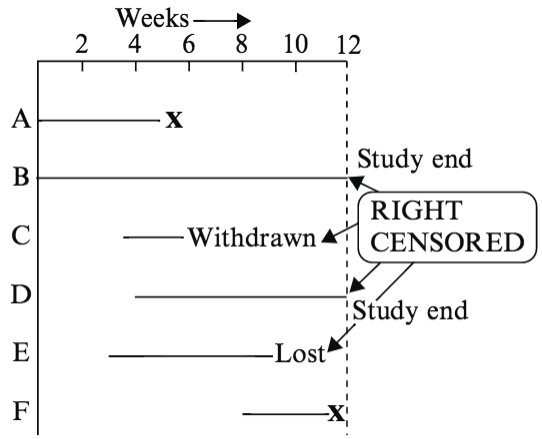

Figure 1.1: Right censoring

2) Interval censoring is corresponds to the case when the event happened in an interval of time but its exact time is not known. It is shown in Fig. 1.2.

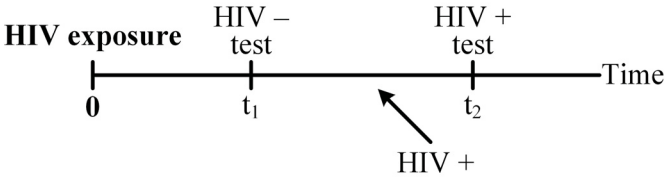

Figure 1.2: Interval censoring (Kleinbaum and Klein, 2006)

3) Case-1 interval censoring is a special case of interval censoring when the event happened before or after the first follow up which is demonstrated in Fig. 1.3.

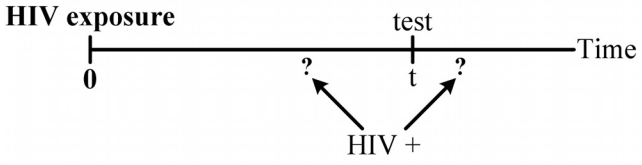

Figure 1.3: Case-1 interval censoring

In most medical research, It is not possible to monitor the patients continuously over a long period. Therefore, status of event is only available at random inspection times and the exact time of the event is unknown. This refers to case-1 interval censoring where the event was observed before or after some specific inspection time. In this thesis, we work on a cure-rate estimation method under case-1 interval censoring using mixture model.

1.4 Non-Parametric Case-1 Interval Cure-rate Estimation For Censored Data

We now discuss the analysis of case-1 interval censored survival data without parametric assumptions about the form of the distribution. If X_i is time of the event of interest such as HIV infection and Y_i is time of a check-up. Assume X_i's are independent and identically distributed (iid) random variables with distribution function F. Here Y_i are censoring variables with iid distribution function G, so the 'current

statue' of i^{th} individual under case-1 interval censoring is

$$(\delta_i, Y_i) \quad \text{where} \quad \delta_i = I(X_i \leq Y_i) \tag{1.10}$$

It has been shown that the non-parametric maximum likelihood for estimator F can be calculated by solving (Groeneboom and Wellner, 1992)

$$\max_F L(F_1, ..., F_n) \quad \text{subject to} \quad 0 \leq F_1 \leq ... \leq F_n \leq 1, \tag{1.11}$$

where

$$L(F_1, ..., F_n) = \sum_{i=1}^{n} (\delta_{[i]} \log(F_i) + (1 - \delta_{[i]}) \log(1 - F_i)), \tag{1.12}$$

and $F_i = F(Y_{(i)})$, $Y_{(i)}$ is order-statistic for Y_i, $\delta_{[i]}$ is concomitant of $Y_{(i)}$. The solution is given by a "max-min formula". (Groeneboom and Wellner, 1992).

$$\hat{F}_i = \max_{h \leq i} \min_{k \geq i} \frac{\sum_{j=h}^{k} \delta_{[j]}}{k - h + 1} \tag{1.13}$$

Sen and Tan (2008) showed that nonparametric maximum-likelihood estimator (NPMLE) of cure-rate in Eq. (1.13) is non-unique and inconsistent (Sen and Tan, 2008) by stating following theorems:

Theorem 1 The likelihood function for p is given by
$L^c(p) = \max_{0 \leq F_1 \leq ... \leq F_n \leq 1-p} L(F_1, ..., F_n)$. Further,

$$L^c(p) = L(\hat{F}_1 \wedge (1 - p), ..., \hat{F}_n \wedge (1 - p)), \tag{1.14}$$

where $\wedge$ is minimum operator and $\hat{F}_i$ are given by Eq. (1.13). Proof has been provided in (Sen and Tan, 2008). This theorem proves the inconsistency and non-uniqueness the NPMLE of p as following:

1) Non-uniqueness of NPMLE: $L^c(p)$ is nonincreasing in interval $0 \leq p \leq 1$, and

$$L^c(\hat{p}) = \sup_{0 \leq p \leq 1} L^c(p) = L(\hat{F}_1, ..., \hat{F}_n) \quad \text{for any} \quad 0 \leq \hat{p} \leq (1 - \hat{F}_n) \tag{1.15}$$

Therefore, $\hat{p}$ is unique if and only if

$$\hat{p} = (1 - \hat{F}_n) = 0. \tag{1.16}$$

2) Inconsistency of NPMLE: Based on Eq. (1.13), we have

$$\hat{F}_n = \max_{i \leq n} \frac{\sum_{j=i}^{n} \delta_{[j]}}{n - i + 1} \tag{1.17}$$

so that $\hat{F}_n = 1$ if and only if $\delta_{[n]} = 1$. Therefore, for $0 < p < 1$ and any $0 < \epsilon < p$,

$$\begin{aligned} P\{\mid \hat{F}_n - (1-p) \mid > \epsilon\} &\geq P\{\hat{F}_n = 1\} = P\{\delta_{[n]} = 1\} \\ &= (1-p)E((F(Y_{(n)}))) \rightarrow (1-p)F(\sup\{y \mid G(y) = 1\}) \end{aligned} \tag{1.18}$$

So, $\hat{F}_n$ is not a consistent estimator of $1 - p$ which is in contrast of random censoring (Sen and Tan, 2008).

Sen and Tan proposed two novel estimators by modifying $\hat{F}_n$. Note that $\hat{F}_n$ may be written as

$$\hat{F}_n = \max_{y \le Y_{(n)}} \frac{\sum_{i=1}^{n} \delta_i I(Y_i \ge y)}{\sum_{i=1}^{n} I(Y_i \ge y)} \tag{1.19}$$

Their proposed estimators are:

1) Note that

$$p_{1n}(x) = \frac{\sum\limits_{i=1}^{n} \delta_i I(Y_i \ge x)}{\sum\limits_{i=1}^{n} I(Y_i \ge x)} \to (1-p)\frac{\int_x^\infty F dG}{\int_x^\infty dG} \quad \text{as} \quad n \to \infty \tag{1.20}$$

Thus we may choose $p_{1n}(x_n)$, x_n large, as an estimator for $(1-p)$

2) Motivated by F_n, we may propose an alternative estimator as

$$p_{2n}(x_n) := max_{0 \le y \le x_n} \ p_{1n}(y), \tag{1.21}$$

which is a partial maximum of tail-averages.

Furthermore, They used cross validation technique to choose a proper cut-off point. Finally, they provided limiting distribution of estimators based on extreme-value theorem (Sen and Tan, 2008).

However, these estimators are sensitive to choose of x_n, their convergence is very slow. In this thesis, we extend their work and propose two consistent estimators based on smoothing. Furthermore, their asymptotic properties have been investigated. Finally, smoothing parameter selection has been done by jackknife-based (leave-one-out) cross validation.

1.5 Smoothing

Smoothing creates a connection between non-parametric approach which does not contains any assumptions and a parametric approach that makes strong assumptions. Smoothing methods in combination with non parametric method can help to extract more information from data in comparison to non parametric method per se. Also, this method provides reasonable assumption of smoothness which makes analysis flexible and robust. Furthermore, it clears important structure of data. Therefore, it is better to involve non-parametric methods with some kind of approximation or smoothing method. (Simonoff, 2012).

1.5.1 Poisson Smoothing

In 1996, Chaubey and Sen proposed to use Poisson smoothing with survival functions. Their estimation technique solved issues which were existed in previously published kernel smoothing scheme for estimating function with non-negative random variable such as survival data. In summary, issues can be listed as 1) Positive mass outside support (Silverman, 1986) and 2) Failure to estimate discontinuity at boundary (Chaubey *et al.*, 2010). Alternatively, (Marron and Ruppert, 1994) and (Bagai and Rao, 1995) proposed transformation method and replacing kernel function by non-negative density function to deal with mentioned issues, respectively. However, it is still interesting to find a method close to kernel smoothing not transforming data. Also, (Bagai and Rao, 1995) method was not able to use all of the data.

Chaubey and Sen proposed an alternative formulation for kernel smoothing which incorporated generalized Hille's smoothing lemma (Chaubey and Sen, 1996; Chaubey *et al.*, 2010).

Following lemma (Hilles'Lemma) (Feller, 1968) is the key of Chaubey and Sen's work.

Lemma 1.1 Let u be any bounded and continuous function and $G_{x,n}, n = 1, 2, \ldots$ be a family of distribution with mean μ_n and variance $h_n^2(x)$, then we have as $\mu \to x$ and $h_n(x) \to 0$

$$\tilde{u}(x) = \int_{-\infty}^{\infty} u(t) dG_{x,n}(t) \longrightarrow u(x) \tag{1.22}$$

The convergence is uniform in every sub interval in which $h_n(x) \to 0$ and u is uniformly continues.

This lemma can be adapted to replace $u(x)$ by, e.g., an empirical distribution function $F_n(x)$ for the smooth estimator as follow

$$\tilde{F}_n(x) = \int_{-\infty}^{\infty} F_n(x) dG_{x,n}(t) \tag{1.23}$$

Strong convergence of the empirical distribution function translates to the strong convergence of $\tilde{F}_n(x)$ which is stated in the following theorem.

Lemma 1.2 If $h \equiv h_n(x) \to 0$ for every fix x as $n \to \infty$ we have

$$sup_x |\tilde{F}_n(x) - F(x)| \xrightarrow{a.s} 0 \tag{1.24}$$

as $n \to \infty$

For their estimator, they proposed using following nonnegative array $\{W_{nk}(t, y); 0 \leq k \leq n; n \geq 1\}$ where

$$W_{nk}(t, y) = \frac{(ty)^k / k!}{\sum_{i=0}^{n} (ty)^i / i!} \tag{1.25}$$

and

$$\sum_{k=0}^{n} W_{nk}(t,y) = 1 \quad \forall t, y \in R^{+} \tag{1.26}$$

Assume $\{\lambda_n; n \geq 1\}$ is a sequence of positive numbers such that $\lambda_n \to \infty$ and $n \to \infty$ almost surely (a.s.) but $\lambda_n/n \to 0$ (a.s.). Selection of λ_n enables them to use Hill theorem and propose a smooth estimator for survival function as (Chaubey and Sen, 1996).

$$\hat{S}_n(t) = \sum_{k=0}^{n} W_{nk}(t,\lambda_n) S_n(\frac{k}{\lambda_n}) \tag{1.27}$$

They have shown that when n increases for all fixed $t \in R^{+}$ Eq.(1.25) behaves similar to $e^{-t\lambda_n}(t\lambda_n)^k(k!)^{-1}$ for $k \leq n$. As a result, for large n Eq. (1.25) can be approximated as a Poisson mixture of survival function with parameter $t\lambda_n$. Also, because of the right tiltedness of Poisson distribution result in the smooth and monotone estimator (Chaubey and Sen, 1996).

In next chapters we will propose two smooth estimations of cure-rate based on motivation from Hills lemma. Also, their consistency analysis and asymptotic behavior will be studied. Finally, we will calculate the optimum smoothing parameter by simulation.

Chapter 2

Poisson Smooth Estimator of Cure-Rate

In this chapter, we will propose two estimators for cure-rate via Poisson smoothing. Afterward, we demonstrate the consistency and asymptotic normality of the estimators.

2.1 Estimator 1.

This section demonstrates the basis the smooth estimation of Eq. (1.20) using Poisson process characteristics and the Hille's Lemma which was reviewed in chapter one.

Assuming, we have a Poisson random variable $N_t \sim Poisson(t)$, $t > 0$ then $E(\frac{N_t}{t}) = 1$, therefore

$$E(\frac{N_t}{t} - 1)^2 \to 0 \tag{2.1}$$

since,

$$Var(\frac{N_t}{t}) = \frac{t}{t^2} = \frac{1}{t} \to 0 \quad \text{as} \quad t \to \infty. \tag{2.2}$$

Now consider $N_{tx} \sim Poisson(tx)$, $t > 0$, $x \geq 0$ then, we have

$$\frac{N_{tx}}{t} \xrightarrow{P} x \quad \text{as} \quad t \to \infty. \tag{2.3}$$

So, for any continuous function such as f:

$$f(\frac{N_{tx}}{t}) \longrightarrow f(x) \tag{2.4}$$

and

$$E[f(\frac{N_{tx}}{t})] = \sum_{k=0}^{\infty} f(\frac{k}{t}) e^{-tx} \frac{(tx)^k}{k!} \longrightarrow f(x) \tag{2.5}$$

Chaubey and Sen (1996) replaced $f(\frac{k}{t})$ with empirical survival function to create smooth estimation of survival function as

$$\hat{\bar{F}}(x) = \frac{1}{n} \sum_{k=0}^{\infty} \sum_{j=1}^{n} I(Y_j \geq \frac{k}{t}) \frac{e^{-tx}(tx)^k}{k!} \tag{2.6}$$

Now for the smooth the estimation of cure-rate p_{1n} in Eq. (1.20), We propose Poison smoothing of numerator and denominator in Eq. (1.20) separately. The numerator of first Sen and Tan cure-rate estimator, p_{1n}, is equal to (Sen and Tan, 2008)

$$\frac{1}{n} \sum_{i=1}^{n} \delta_i I(Y_i \geq x) \tag{2.7}$$

Therefore, the smooth version of Eq. (2.7) when t_n is large enough is

$$N(t_n) := \frac{1}{n}\sum_{k=0}^{\infty}\sum_{i=1}^{n}\delta_i I(Y_i \geq k)e^{-t_n}t_n^k/k! \tag{2.8}$$

Similarly, for the denominator of p_{1n} is equal to

$$\frac{1}{n}\sum_{i=1}^{n} I(Y_i \geq x) \tag{2.9}$$

and its corresponding smooth version when t_n is large enough is

$$D(t_n) := \frac{1}{n}\sum_{k=0}^{\infty}\sum_{i=1}^{n} I(Y_i \geq k)e^{-t_n}t_n^k/k! \tag{2.10}$$

Therefore, we introduce our smooth cure-rate estimator as p_{3n} as following

$$p_{3n} = \frac{\sum_{k=0}^{\infty}\frac{1}{n}(\sum_{i=1}^{n}\delta_i I(Y_i \geq k))e^{-t_n}\frac{t_n^k}{k!}}{\sum_{k=0}^{\infty}\frac{1}{n}(\sum_{i=1}^{n} I(Y_i \geq k))e^{-t_n}\frac{t_n^k}{k!}} = \frac{N(t_n)}{D(t_n)}, \quad \text{for large } t_n > 0 \tag{2.11}$$

Our proposed estimator, in contrast to Chaubey and Sen (1996) estimator, does not estimate a function at a point; our goal is to estimate cure-rate in the limit as $t_n \to \infty$.

Fig. 2.1 is a sample plot of Eq. (2.11). We assumed cure-rate is equal to 0.3 and generated time to the event and the check-up times using $exp(0.4)$ and $exp(0.2)$ distribution functions, respectively. Fig. 2.1 shows that the new estimator has better performance in comparison to cure-rate estimated by Eq. (1.20) and Eq. (1.21), respectively.

The green, blue, red and black curves in figure 2.1 demonstrates assumed cure-rate, p_{1n}, p_{2n} and p_{3n}, respectively.

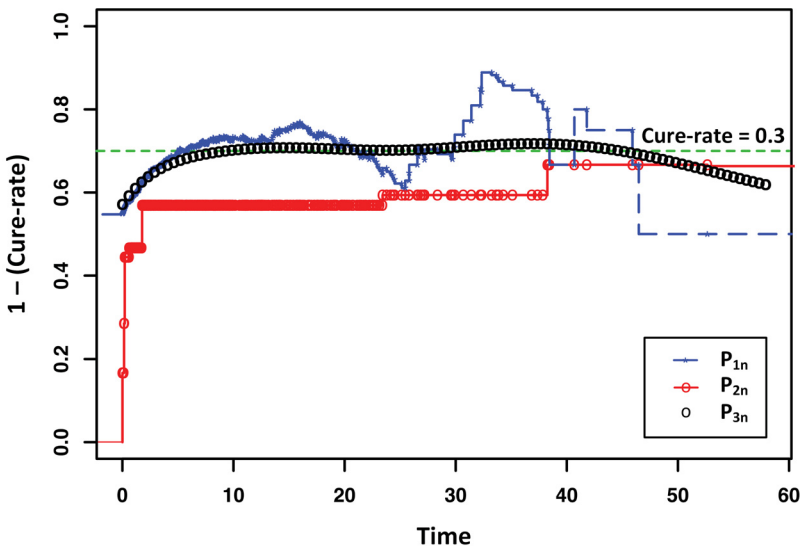

Figure 2.1: Sample-plot of cure rate estimation versus time. The green, blue, red and black curves are assumed cure-rate, p_{n1}, p_{n2} and p_{n3}, respectively. ($F(y) = exp(0.4)$, $G(y) = exp(0.2)$, $n = 500$ and $1 - p = 0.7$)

The main problems that we are dealing with them for this chapter are the following

1. Prove the consistency of p_{3n}.

2. Reach limiting distribution of p_{3n}.

2.1.1 Consistency of the 1st Estimator

In this section, we study the consistency of our proposed smooth estimator as follows

$$p_{3n} \xrightarrow{P} (1-p) \quad \text{as} \quad t_n \to \infty \quad \text{and} \quad n \to \infty \tag{2.12}$$

Definition 2.1: Let X be a random variable with cdf $F(x,\theta)$ where $\theta \in \Omega$. Let $X_1, \cdots, X_n$ be a sample from the distribution of X and let T_n denote a statistics. Then, T_n is a consistent estimate of θ if

$$T_n \xrightarrow{P} \theta \tag{2.13}$$

$$P_\theta(|T_n - \theta| > \varepsilon) \longrightarrow 0 \quad as \quad n \to \infty \quad \text{for all} \quad \varepsilon > 0 \tag{2.14}$$

Sufficient condition of consistency: Suppose that T_n is an estimator for θ. If

$$\begin{aligned} &1) \quad \text{bias}(T_n) \to 0 \quad \text{as} \quad n \to \infty \\ &2) \quad \text{var}(T_n) \to 0 \quad \text{as} \quad n \to \infty \end{aligned} \tag{2.15}$$

then T_n is a consistent estimator for θ (Hogg, 2012).

We can apply the result of definition 2.1 on Eq. (2.11) to prove the consistency of the proposed cure-rate smooth estimator, specifically

$$p_{3n} = \frac{\frac{1}{n}\sum_{k=0}^{\infty}(\sum_{i=1}^{n}\delta_i I(Y_i \geq k))e^{-t_n}\frac{t_n^k}{k!}}{\frac{1}{n}\sum_{k=0}^{\infty}(\sum_{i=1}^{n} I(Y_i \geq k))e^{-t_n}\frac{t_n^k}{k!}} \longrightarrow 1-p \tag{2.16}$$

when t_n approaching to ∞.

Proving consistency by finding the variance of Eq. (2.11) is quite challenging since it is a fraction. In this thesis, in order to overcome this difficulty, we will use the result of Eq. (2.15) separately for numerator and denominator of p_{3n}, via telescoping. Let

$$N(t_n) = \frac{1}{n}\sum_{k=0}^{\infty}(\sum_{i=1}^{n}\delta_i I(Y_i \geq k))e^{-t_n}\frac{t_n^k}{k!} \tag{2.17}$$

and

$$D(t_n) = \frac{1}{n}\sum_{k=0}^{\infty}(\sum_{i=1}^{n} I(Y_i \geq k))e^{-t_n}\frac{t_n^k}{k!}. \tag{2.18}$$

Based on definition 2.1, first, we need to prove that the bias of estimator $P_{3n}(t) = N(t_n)/D(t_n)$ goes to zero for large t_n. Expectation of $N(t_n)$ can be calculated as

$$\begin{aligned} n(t_n) = E(N(t_n)) &= E\left(\sum_{k=0}^{\infty}\delta_1 I(Y_1 \geq k)e^{-t_n}\frac{t_n^k}{k!}\right) \\ &= \sum_{k=0}^{\infty} E\left(\delta_1 I(Y_1 \geq k)\right)e^{-t_n}\frac{t_n^k}{k!} \end{aligned} \tag{2.19}$$

where

$$\delta_1 = I(X_1 \leq Y_1) = \begin{cases} 1 & \text{if } (X_1 \leq Y_1) \\ 0 & \text{if otherwise} \end{cases}$$

and

$$\begin{aligned} E(\delta_1 I(Y_1 \geq k)) &= P((X_1 \leq Y_1),(Y_1 \geq k)) \\ &= \int_0^{\infty} P(X \leq y, y \geq k)P(Y = y)\mathrm{d}y \\ &= (1-p)\int_k^{\infty} F(y)\mathrm{d}G(y) \end{aligned} \tag{2.20}$$

Therefore,

$$n(t_n) = (1-p)\sum_{k=0}^{\infty}\int_k^{\infty} F(y)\mathrm{d}G(y)\frac{e^{-t_n}t_n^k}{k!} \tag{2.21}$$

Similarly

$$d(t_n) = E(D(t_n)) = \sum_{k=0}^{\infty} E(I(Y_1 \geq k)) = \sum_{k=0}^{\infty} P(Y \geq k) = \sum_{k=0}^{\infty} \int_k^{\infty} \mathrm{d}G(y) \frac{e^{-t_n} t_n^k}{k!} \tag{2.22}$$

Therefore, the bias term is;

$$\text{Bias} = \frac{n(t_n)}{d(t_n)} - (1-p) \tag{2.23}$$

Following Lemma is used to prove

$$\text{bias}(p_{3n}) \to 0 \quad \text{as} \quad t_n \to \infty \tag{2.24}$$

which is necessary for consistency of estimator.

Lemma 2.1: If $G(.)$ is absolutely continuous with density $g(.)$, then

$$\frac{n(t_n)}{d(t_n)} \to 1-p \quad \text{if} \quad t_n \to \infty \tag{2.25}$$

Proof:

$$\frac{E(N(t_n))}{E(D(t_n))} = \frac{(1-p) \sum_{k=0}^{\infty} \int_k^{\infty} \mathrm{F}(y)\, \mathrm{d}G(y) e^{-t_n} \frac{t_n^k}{k!}}{\sum_{k=0}^{\infty} \int_k^{\infty} \mathrm{d}G(y) e^{-t_n} \frac{t_n^k}{k!}} \tag{2.26}$$

let

$$\begin{aligned} a_k &= \int_k^{\infty} \mathrm{F}(y)\, \mathrm{d}G(y) \\ b_k &= \int_k^{\infty} \mathrm{d}G(y) \end{aligned} \tag{2.27}$$

it can be seen that when $k \to \infty$,

$$a_k \to 0 \quad \text{and} \quad b_k \to 0 \tag{2.28}$$

Using L'Hopital rule when $k \to \infty$

$$\lim_{k\to\infty} \frac{a_k}{b_k} = \lim_{k\to\infty} \frac{\frac{\partial}{\partial k}\int_k^\infty \mathrm{F}(y)\,\mathrm{d}G(y)}{\frac{\partial}{\partial k}\int_k^\infty \mathrm{d}G(y)} = \frac{-F(k)g(k)}{-g(k)} = 1 \tag{2.29}$$

Therefor, the limit of $n(t_n)/d(t_n)$ when $t_n \to \infty$ can be calculated as follow

$$\lim_{t_n\to\infty} \frac{n(t_n)}{d(t_n)} = \lim_{t_n\to\infty} \frac{(1-p)\sum_{k=0}^\infty \frac{a_k t_n^k}{k!}}{\sum_{k=0}^\infty \frac{b_k t_n^k}{k!}} \tag{2.30}$$

Using repeated L'Hopital rule

$$\lim_{t_n\to\infty} \frac{n(t_n)}{d(t_n)} = \lim_{t_n\to\infty} \frac{(1-p)\sum_{k=0}^\infty a_{k+m}\frac{t_n^k}{k!}}{\sum_{k=0}^\infty b_{k+m}\frac{t_n^k}{k!}} \tag{2.31}$$

based on Eq. (2.29) and for all $k \geq 0$ and given $\epsilon > 0$, for large m

$$\left|1 - \frac{a_{k+m}}{b_{k+m}}\right| < \epsilon \tag{2.32}$$

It follows that

$$\left|(1-p) - \frac{(1-p)\sum_{k=0}^\infty a_{k+m}\frac{t^k}{k!}}{\sum_{k=0}^\infty b_{k+m}\frac{t^k}{k!}}\right| = \left|\frac{(1-p)\sum_{k=0}^\infty (1-\frac{a_{k+m}}{b_{k+m}})b_{k+m}\frac{t^k}{k!}}{\sum_{k=0}^\infty b_{k+m}\frac{t^k}{k!}}\right| < (1-p)\epsilon \tag{2.33}$$

Therefore, bias goes to zero when $t_n \to \infty$ based on Lemma 2.1.

$$\text{bias}(p_{3n}) = \frac{n(t_n)}{d(t_n)} - (1-p) \to 0. \tag{2.34}$$

Now, to prove consistency of estimator, we only need to show that

$$\frac{N(t_n)}{D(t_n)} - \frac{n(t_n)}{d(t_n)} \xrightarrow{p} 0 \quad \text{as} \quad n \to \infty \tag{2.35}$$

Using telescopic method

$$\begin{aligned} p_{3n} - (1-p) &= \frac{N(t_n)}{D(t_n)} - \frac{n(t_n)}{d(t_n)} + \frac{n(t_n)}{d(t_n)} - (1-p) \\ &= \frac{N(t_n)}{D(t_n)} - \frac{n(t_n)}{d(t_n)}\frac{D(t_n)-d(t_n)}{D(t_n)} + \frac{n(t_n)}{d(t_n)} - (1-p) \\ &= \frac{\frac{N(t_n)-n(t_n)}{d(t_n)}}{\frac{D(t_n)}{d(t_n)}} - \frac{n(t_n)}{d(t_n)}\frac{\frac{D(t_n)-d(t_n)}{d(t_n)}}{\frac{D(t_n)n(t_n)}{d(t_n)}} + \frac{n(t_n)}{d(t_n)} - (1-p) \\ &= \frac{n(t_n)}{d(t_n)}\frac{\frac{N(t_n)-n(t_n)}{n(t_n)}}{\frac{D(t_n)}{d(t_n)}} - \frac{n(t_n)}{d(t_n)}\frac{\frac{D(t_n)-d(t_n)}{d(t_n)}}{\frac{D(t_n)}{d(t_n)}} + \frac{n(t_n)}{d(t_n)} - (1-p) \end{aligned} \tag{2.36}$$

Lemma 2.2: Suppose $nd(t_n) \to \infty$ as $n \to \infty$

a) $\frac{D(t_n)}{d(t_n)} \xrightarrow{p} 1$ when, $n \to \infty$

b) $\frac{N_n(t)}{n(t)} \xrightarrow{p} 1$ when, $n \to \infty$

Proof:

a) Based on Eq. (2.22)

$$\frac{D(t_n)}{d(t_n)} = \frac{\frac{1}{n}\sum_{k=0}^{\infty}(\sum_{i=1}^{n} I(Y_i \geq k))e^{-t_n}\frac{t_n^k}{k!}}{\sum_{k=0}^{\infty}\frac{e^{-t_n}t_n^k}{k!}\int_k^{\infty} \mathrm{d}G(y)} \tag{2.37}$$

obviously,

$$E(\frac{D(t_n)}{d(t_n)}) = 1 \tag{2.38}$$

Further

$$\begin{aligned} var(\frac{D(t_n)}{d(t_n)}) &= \frac{E(D(t_n)^2) - E(D(t_n))^2}{d^2(t_n)} \\ &= \frac{1}{n}\cdot\frac{\sum_{k=0}^{\infty}\sum_{l=0}^{\infty} P(Y \geq k \vee l)\frac{e^{-2t_n}t_n^{k+l}}{k!l!} - d^2(t_n)}{d^2(t_n)} \\ &= \frac{1}{nd(t_n)}\left(\frac{\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\bar{G}(k \vee l)\frac{e^{-2t_n}t_n^{k+l}}{k!l!}}{d(t_n)} - d(t_n)\right) \end{aligned} \tag{2.39}$$

where

$$\bar{G}(k) = \int_k^{\infty} \mathrm{d}G(y). \tag{2.40}$$

Note that

1) $d(t) = d(t_n) \to 0$ if $t_n \to \infty$

2) $nd(t) = nd(t_n) \to \infty$, by our assumption

Also, we know that $\bar{G}$ is a decreasing function, therefore

$$\bar{G}(k \vee l) \leq \bar{G}(k) \tag{2.41}$$

and

$$\frac{\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\bar{G}(k\vee l)\frac{e^{-2t_n}t_n^{k+l}}{k!l!}}{\sum_{k=0}^{\infty}\bar{G}(k)\frac{t_n^k}{k!}}\leq 1. \tag{2.42}$$

We have

$$var(\frac{D(t_n)}{d(t_n)})\to 0 \tag{2.43}$$

as $n\to\infty$ and $t_n\to\infty$.

b) Based on Eq. (2.19), we have

$$\frac{N(t_n)}{n(t_n)}=\frac{\frac{1}{n}\sum_{k=0}^{\infty}(\sum_{i=1}^{n}\delta_i I(Y_i\geq k))e^{-t_n}\frac{t_n^k}{k!}}{\sum_{k=0}^{\infty}\frac{e^{-t_n}t_n^k}{k!}\int_k^{\infty}(1-p)F(y)\mathrm{d}G(y)} \tag{2.44}$$

Again,

$$E(\frac{N(t_n)}{n(t_n)})\to 1 \tag{2.45}$$

and

$$\begin{aligned} var(\frac{N(t_n)}{n(t_n)}) &= \frac{E(N(t_n)^2)-E(N(t_n))^2}{n^2(t_n)} \\ &= \frac{\frac{1}{n}\left((1-p)\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\frac{e^{-2t_n}t_n^{k+l}}{k!l!}\int_{max(k,l)}^{\infty}\mathrm{F}(y)\,\mathrm{d}G(y)-n^2(t_n)\right)}{n^2(t_n)} \\ &= \frac{1}{n\cdot n(t)}\left(\frac{(1-p)\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\frac{e^{-2t_n}t_n^{k+l}}{k!l!}\int_{max(k,l)}^{\infty}\mathrm{F}(y)\,\mathrm{d}G(y)}{n(t_n)}-n(t_n)\right) \end{aligned} \tag{2.46}$$

Again using

$$1) \quad d(t) = d(t_n) \to 0 \quad \text{if } t_n \to \infty$$

$$2) \quad nd(t) = nd(t_n) \to \infty$$

Also, we know that $\bar{G}$ and $\int_0^\infty FdG$ are decreasing functions, therefore

$$\frac{\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\frac{e^{-2t_n}t_n^{k+l}}{k!l!}(1-p)\int_{max(k,l)}^{\infty}\mathrm{F}(y)\,\mathrm{d}G(y)}{\sum_{k=0}^{\infty}\bar{G}(k)\frac{e^{-t}t_n^k}{k!}} \leq 1. \tag{2.47}$$

We have

$$var(\frac{N(t_n)}{n(t_n)}) \to 0. \tag{2.48}$$

as $n \to \infty$ and $t_n \to \infty$. As a result, based on Lemma 2.1 and Lemma 2.2, it can be shown that

$$p_{3n} \xrightarrow{p} 1-p \tag{2.49}$$

Therefore, based on Eqs (2.49) and (2.35) p_{3n} is a consistence estimator.

Example 2.1: If $\bar{G}(k)$ corresponds to the exponential distribution with parameter k

$$\bar{G}(k) = e^{-k}, \tag{2.50}$$

then, if $t_n \to \infty$, $ne^{-t_n} \to c$, where $0 < c \leq \infty$, we have

$$nd(t_n) = ne^{-t_n}\sum_{k=0}^{\infty}\bar{G}(k)\frac{t_n^k}{k!} = ne^{-t_n}\sum_{k=0}^{\infty}e^{-k}\frac{t_n^k}{k!} = ne^{-t_n}e^{t_n/e} \to \infty \tag{2.51}$$

Again if $F(k)$ have the exponential distribution with parameter k

$$\text{if} \quad \bar{G}(k) = e^{-k} \quad \text{and} \quad F(k) = (1 - e^{-k}), \tag{2.52}$$

then, if $t_n \to \infty$ and, $ne^{-t_n} \to c$, $0 < c \leq \infty$. we have

$$\begin{aligned} n \cdot n(t_n) &= (1-p)ne^{-t_n}(1-p)\int_k^\infty F(k)\mathrm{d}G(k) \\ &= (1-p)ne^{-t_n}\sum_{k=0}^{\infty}\int_k^\infty (1-e^{-k})e^{-k}\frac{t_n^k}{k!} \\ &= (1-p)[ne^{-t_n}e^{t_n/e} - \frac{n}{2}e^{-t_n}e^{t_n/e^2}] \to \infty \end{aligned} \tag{2.53}$$

2.1.2 Limiting Distribution of 1st Estimator

In this section, we start with definition of limiting distribution. Afterward, we study the limiting distribution for our new smooth estimator p_{3n} and previously proposed estimator p_{1n}.

Definition: Consider a sequence of random variable X_1, X_2, ... and corresponding sequence of cdfs F_{X_1}, F_{X_2}, ... so that for $n = 1,\ 2,\ \ldots$ we have $F_{X_n}(x) = P[X_n \leq x]$. Suppose that there exists a cdf such that for all X at which F_X is continuous

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \tag{2.54}$$

Then $X_1,\ X_2,\ \ldots$ converge in distribution to random variable X with cdf F_X denoted

$$X_n \xrightarrow{d} X \tag{2.55}$$

F_X is called **limiting distribution** (Severini, 2005).

Degenerate distributions is a special case of limiting distribution. The sequence of random variables $X_1, \ldots, X_n$ converge in distribution to constant c if the limiting distribution of $X_1, \ldots, X_n$ is degenerate at c, that is $X_n \xrightarrow{d} X$ and $P[X = c] = 1$ Severini (2005), so that

$$F_X(x) = \begin{cases} 0 & \text{if} \quad (x < c) \\ 1 & \text{if} \quad (x \geq c) \end{cases} \tag{2.56}$$

In order to determine the limiting distribution, we need to use Slutsky theorem and central limit theorem (CLT).

Theorem (2.1) (Slutsky Theorem): Suppose that X_n and Y_n are random variables and let c be a constant

$$\begin{aligned} X_n &\xrightarrow{d} X \\ Y_n &\xrightarrow{p} c \end{aligned} \tag{2.57}$$

then (Hogg, 2012)

$$\begin{aligned} &X_n + Y_n \xrightarrow{d} X + c \\ &Y_n \cdot A_n \xrightarrow{d} cX \\ &X_n/Y_n \xrightarrow{d} X/c \quad \text{provided} \quad c \neq 0 \end{aligned} \tag{2.58}$$

Theorem 2.2 (Central Limit Theorem): If $X_1, \cdots, X_n$ are iid with $E(X_1^2) < \infty$ then

$$\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} X_i - E(X_1)) \tag{2.59}$$

has Normal$(0, \sigma^2 \equiv var(X_1))$ limiting distribution. Equivalently

$$\frac{(\sum_{i=1}^{n} X_i - nE(X_1))}{\sqrt{n}\sigma} \xrightarrow{P} N(0,1) \tag{2.60}$$

We shall also need;

Theorem 2.3(Dominated Converge Theorem): If X_n, $n \geq 1$, is a sequence of random variable such that $X_n \xrightarrow{p} X$ and $|X_n| \leq Y$ for some Y

$$E|Y| < \infty, \quad \text{then} \quad E|X_n - X| \to 0 \quad \text{as} \quad n \to \infty \tag{2.61}$$

Asymptotic behavior of p_{1n}

Sen and Tan (2008) used asymptotic theory of sample extremes to obtain limiting distribution of

$$p_{1n}(x_n) = \frac{\sum_{j=1}^{n} \delta_j I(Y_j \geq x_n)}{\sum_{j=1}^{n} I(Y_j \geq x_n)} \tag{2.62}$$

They defined

$$Z_{1n} = \frac{\sqrt{\sum_{j=1}^{n} I(Y_j \geq x_n)}(p_{1n}(x_n) - (1-p))}{\sqrt{p(1-p)}}. \tag{2.63}$$

and showed that if $n\bar{G}(x_n) \to \infty$ and $x_n \to \infty$,

$$\lim_{n\to\infty} Z_{1n} = N(0,1). \tag{2.64}$$

Here we have proposed another method for calculating limiting distribution of p_{1n} which can be easily modified to calculate limiting distribution for p_{3n}. Lets start by considering

$$p_{1n}(x_n) - \frac{(1-p)\int_{x_n}^{\infty} F dG(y)}{\int_{x_n}^{\infty} \mathrm{d}G(y)} = \frac{\frac{1}{n}\sum_{j=1}^{n} \delta_j I(Y_j \geq x_n)}{\frac{1}{n}\sum_{j=1}^{n} I(Y_j \geq x_n)} - \frac{(1-p)\int_{x_n}^{\infty} F(y)\mathrm{d}G(y)}{\int_{x_n}^{\infty} \mathrm{d}G(y)}$$
$$= Z_{1n}^1 - \frac{(1-p)\int_{x_n}^{\infty} F(y)\mathrm{d}G(y)}{\int_{x_n}^{\infty} \mathrm{d}G(y)} \cdot Z_{1n}^2 \tag{2.65}$$

where

$$Z_{1n}^1 = \frac{\frac{1}{n}\sum_{j=1}^{n} \delta_j I(Y_j \geq x_n) - (1-p)\int_{x_n}^{\infty} F(y)\mathrm{d}G(y)}{\frac{1}{n}\sum_{j=1}^{n} I(Y_j \geq x_n)} \tag{2.66}$$

and

$$Z_{1n}^2 = \frac{\frac{1}{n}\sum_{j=1}^{n} I(Y_j \geq x_n) - \int_{x_n}^{\infty} \mathrm{d}G(y)}{\frac{1}{n}\sum_{j=1}^{n} I(Y_j \geq x_n)}. \tag{2.67}$$

The bivariate limiting distribution can be calculated as follow. For Z_{1n}^1 we have

$$\begin{aligned} Z_{1n}^1 &= \frac{1/\bar{G}(x_n)}{1/\bar{G}(x_n)} \cdot \frac{\frac{1}{n}\sum_{j=1}^{n} \delta_j I(Y_j \geq x_n) - (1-p)\int_{x_n}^{\infty} F(y)\mathrm{d}G(y)}{\frac{1}{n}\sum_{j=1}^{n} I(Y_j \geq x_n)} \\ &= \frac{Z_{1n}^{11} - Z_{1n}^{12}}{Z_{1n}^{13}} \end{aligned} \tag{2.68}$$

where

$$Z_{1n}^{11} = \frac{\frac{1}{n}\sum_{j=1}^{n} \delta_j I(Y_j \geq x_n)}{\bar{G}(x_n)} \tag{2.69}$$

$$Z_{1n}^{12} = \frac{(1-p)\int_{x_n}^{\infty} F(y)\mathrm{d}G(y)}{\bar{G}(x_n)} \tag{2.70}$$

$$Z_{1n}^{13} = \frac{\frac{1}{n}\sum_{j=1}^{n} I(Y_j \geq x_n)}{\bar{G}(x_n)} \tag{2.71}$$

$$\bar{G}(x_n) = \int_{x_n}^{\infty} \mathrm{d}G(y). \tag{2.72}$$

For limiting distribution, first we show that if $x_n \to \infty$ and $n\bar{G}(x_n) \to \infty$

$$\begin{aligned} &1) \lim_{x_n \to \infty} Z_{1n}^{12} = 1 - p \\ &2) \lim_{x_n \to \infty} Z_{1n}^{13} = 1 \end{aligned} \tag{2.73}$$

Part one of Eq. (2.73) has been proved in Lemma 2.1. For the second part based on result of consistency, we have

$$E(\frac{\frac{1}{n}\sum_{i=1}^{n} I(Y_j \geq x_n)}{\bar{G}(x_n)}) - 1 = 0 \tag{2.74}$$

and

$$var(\frac{\frac{1}{n}\sum_{i=1}^{n} I(Y_j \geq x_n)}{\bar{G}(x_n)}) = \frac{\bar{G}(x_n)(1-\bar{G}(x_n))}{n\bar{G}(x_n)^2} \to 0 \tag{2.75}$$

when $x_n \to \infty$ and $n.\bar{G}(x_n) \to \infty$, So, we have

$$\lim_{x_n \to \infty} Z_{1n}^{13} = \lim_{x_n \to \infty} \frac{\frac{1}{n}\sum_{i=1}^{n} I(Y_j \geq x_n)}{\bar{G}(x_n)} = 1 \tag{2.76}$$

Therefore

$$\begin{aligned} var(Z_{1n}^{11}) &= var(\frac{\frac{1}{n}\sum_{j=1}^{n} \delta_j I(Y_j \geq x_n)}{\bar{G}(x_n)}) \\ &= \frac{1}{n \cdot \bar{G}^2(x_n)} \cdot ((1-p)\int_{x_n}^{\infty} F(y)\mathrm{d}G(y))(1-(1-p)\int_{x_n}^{\infty} F(y)\mathrm{d}G(y)) \\ &= \frac{1}{n \cdot \bar{G}(x_n)} \frac{(1-p)\int_{x_n}^{\infty} F(y)\mathrm{d}G(y)}{\bar{G}(x_n)} \cdot (1-(1-p)\int_{x_n}^{\infty} F(y)\mathrm{d}G(y)) \end{aligned} \tag{2.77}$$

We chose $r_n = \sqrt{n \cdot \bar{G}(x_n)}$ as a normalizing constant and based on CLT theorem, we have

$$\lim_{x_n\to\infty}\sqrt{n\bar{G}(x_n)}\cdot Z_{1n}^1 = N(0,1-p) \tag{2.78}$$

Similarly, the limiting distribution for Z_{1n}^2 in Eg. 2.67 can be calculated as

$$\begin{aligned} Z_{1n}^2 &= \frac{1/\bar{G}(x_n)}{1/\bar{G}(x_n)}\cdot\frac{\frac{1}{n}\sum_{j=1}^n I(Y_j\geq x_n)-(1-p)\int_{x_n}^\infty F(y)\mathrm{d}G(y)}{\frac{1}{n}\sum_{j=1}^n I(Y_j\geq x_n)} \\ &= \frac{Z_{1n}^{21}-Z_{1n}^{22}}{Z_{1n}^{23}} \end{aligned} \tag{2.79}$$

where

$$Z_{1n}^{21} = \frac{\frac{1}{n}\sum_{j=1}^n I(Y_j\geq x_n)}{\bar{G}(x_n)} \tag{2.80}$$

$$Z_{1n}^{22} = \frac{(1-p)\int_{x_n}^\infty F(y)\mathrm{d}G(y)}{\bar{G}(x_n)} \tag{2.81}$$

$$Z_{1n}^{23} = \frac{\frac{1}{n}\sum_{j=1}^n I(Y_j\geq x_n)}{\bar{G}(x_n)} \tag{2.82}$$

Therefore

$$\begin{aligned} var(Z_{1n}^2) &= var(\cdot\frac{\frac{1}{n}\sum_{j=1}^n I(Y_j\geq x_n)}{\bar{G}(x_n)}) \\ &= \frac{1}{n\cdot\bar{G}^2(x_n)}\cdot(\bar{G}(x_n)(1-\bar{G}(x_n))) \\ &= \frac{1}{n\cdot\bar{G}(x_n)}\cdot(1-\bar{G}(x_n)) \end{aligned} \tag{2.83}$$

We chose $r_n=\sqrt{n\cdot\bar{G}(x_n)}$ as a normalizing constant and based on CLT theorem, we have

$$\lim_{x_n\to\infty}\sqrt{n\bar{G}(x_n)}\cdot Z_{1n}^2 = N(0,1) \tag{2.84}$$

Finally,

$$\begin{aligned} cov(Z_{1n}^{22}, Z_{1n}^{11}) =& \frac{(1-p)}{n\bar{G}(x_n)} cov\left(\sum_{j=1}^{n} \delta_j I(Y_j \geq x_n), \sum_{j=1}^{n} I(Y_j \geq x_n)\right) \\ &= \frac{1}{n\bar{G}(x_n)}\{(1-p)\int_{x_n}^{\infty} F(y)\mathrm{d}G(y) - (1-p)\bar{G}(x_n)\int_{x_n}^{\infty} F(y)\mathrm{d}G(y)\} \\ &= \frac{(1-p)\int_{x_n}^{\infty} F(y)\mathrm{d}G(y)}{n\bar{G}(x_n)}\{1-\bar{G}(x_n)\} \end{aligned} \tag{2.85}$$

If we use normalizing constant as $r_n = \sqrt{n \cdot \bar{G}(x_n)}$ for Z_{1n}^1 and Z_{1n}^2, the normalized limiting covariance will be equal to

$$\lim_{x_n \to \infty} n\bar{G}(x_n) \cdot cov(Z_{1n}^1, Z_{1n}^2) = (1-p) \tag{2.86}$$

As a result, the limiting distribution of $\sqrt{n\bar{G}(x_n)}(Z_{1n}^1 - (1-p)Z_{1n}^2)$ is equal to $N(0, ((1-p) + (1-p)^2 - 2(1-p)^2) = p(1-p))$. Since

$$\sqrt{n\bar{G}(x_n)} \begin{bmatrix} Z_{1n}^1 \\ \\ Z_{1n}^2 \end{bmatrix} \to \left(\begin{bmatrix} 0 \\ \\ 0 \end{bmatrix}, \begin{bmatrix} 1-p & 1-p \\ \\ 1-p & 1 \end{bmatrix} \right) \tag{2.87}$$

provided $x_n \to \infty$, $n\bar{G}(x_n) \to \infty$

Asymptotic behavior of p_{3n}

In this section, we derive the limiting distribution of p_{3n} by following the same approach which was p_{1n}.

Lets start by considering

$$\begin{aligned} p_{3n} - \frac{n(t_n)}{d(t_n)} &= \frac{\sum_{k=0}^{\infty} \frac{1}{n} (\sum_{i=1}^{n} \delta_i I(Y_i \geq k)) e^{-t_n} \frac{t_n^k}{k!}}{\sum_{k=0}^{\infty} \frac{1}{n} (\sum_{i=1}^{n} I(Y_i \geq k)) e^{-t_n} \frac{t_n^k}{k!}} \\ &- \frac{(1-p) \sum_{k=0}^{\infty} \int_{k=0}^{\infty} F(y) \mathrm{d}G(y) \frac{e^{-t_n} t^k}{k!}}{\sum_{k=0}^{\infty} \int_{k=0}^{\infty} \mathrm{d}G(y) \frac{e^{-t_n} t_n^k}{k!}} \\ &= \frac{N(t_n) - n(t_n)}{D(t_n)} - \frac{n(t_n)}{d(t_n)} \cdot \frac{D(t_n) - d(t_n)}{D(t_n)} \\ &= Z_{3n}^1 - \frac{n(t_n)}{d(t_n)} Z_{3n}^2 \end{aligned} \tag{2.88}$$

where

$$N(t_n) = \frac{1}{n} \sum_{k=0}^{\infty} (\sum_{i=1}^{n} \delta_i I(Y_i \geq k)) e^{-t_n} \frac{t_n^k}{k!} \tag{2.89}$$

$$D(t_n) = \frac{1}{n} \sum_{k=0}^{\infty} (\sum_{i=1}^{n} I(Y_i \geq k)) e^{-t_n} \frac{t_n^k}{k!} \tag{2.90}$$

$$n(t_n) = (1-p) \sum_{k=0}^{\infty} \int_{k=0}^{\infty} F(y) \mathrm{d}G(y) \frac{e^{-t_n} t_n^k}{k!} \tag{2.91}$$

$$d(t_n) = \sum_{k=0}^{\infty} \int_{k=0}^{\infty} \mathrm{d}G(y) \frac{e^{-t_n} t_n^k}{k!} \tag{2.92}$$

and

$$Z_{3n}^1 = \frac{N(t_n) - n(t_n)}{D(t_n)} \tag{2.93}$$

$$Z_{3n}^2 = \frac{n(t_n)}{d(t_n)} \cdot \frac{D(t_n) - d(t_n)}{D(t_n)} \tag{2.94}$$

The bivariate limiting distribution can be calculated as follow. For Z_{3n}^1 we have

$$Z_{3n}^{1} = \frac{1/d(t_n)}{1/d(t_n)} \frac{N(t_n) - n(t_n)}{D(t_n)} \quad = \frac{Z_{3n}^{11} - Z_{3n}^{12}}{Z_{3n}^{13}} \tag{2.95}$$

where

$$Z_{3n}^{11} = \frac{N(t_n)}{d(t_n)} \tag{2.96}$$

$$Z_{3n}^{12} = \frac{n(t_n)}{d(t_n)} \tag{2.97}$$

$$Z_{3n}^{13} = \frac{D(t_n)}{d(t_n)} \tag{2.98}$$

similar to previous section and based on Lemma 2.1 and 2.2

$$\begin{aligned} &1) \lim_{t_n \to \infty} Z_{3n}^{12} = 1 - p \\ &2) \lim_{t_n \to \infty} Z_{3n}^{13} = 1 \\ &3) \lim_{t_n \to \infty} n(t_n) = 0 \\ &4) \lim_{t_n \to \infty} d(t_n) = 0 \end{aligned} \tag{2.99}$$

Next

$$\begin{aligned} var(\sqrt{nd(t_n)} Z_{3n}^{1}) &= \frac{\sum_{k=0}^{\infty} \sum_{l=0}^{\infty} E(\delta I(Y \geq k \vee l) e^{-t_n} \frac{t_n^k}{k!} e^{-t_n} \frac{t_n^l}{l!} - n(t_n)^2}{d(t_n)} \\ &= \frac{E[(1-p) \int_{M_1(t_n) \vee M_2(t_n)} F(y) \mathrm{d}G(y)]}{E[\bar{G}(M_1(t_n))]} - \frac{n(t_n)}{d(t_n)} \cdot n(t_n) \end{aligned} \tag{2.100}$$

where $M_1(t_n)$ and $M_2(t_n)$ are iid Poisson random variable with mean t. Therefore,

$$\lim_{t_n \to \infty} var(\sqrt{nd(t_n)} Z_{3n}^{1}) = \lim_{t \to \infty} \frac{E[(1-p) \int_{M_1(t_n) \vee M_2(t_n)} F(y) \mathrm{d}G(y)]}{E[\bar{G}(M_1(t_n))]} \tag{2.101}$$

Next, the limiting distribution for Z^2_{3n} can be calculated as follow

$$Z^2_{3n} = \frac{1/d(t_n)}{1/d(t_n)} \cdot \frac{n(t_n)}{d(t_n)} \cdot \frac{D(t_n) - d(t_n)}{D(t_n)} \tag{2.102}$$

Similarly to Z^1_{3n}, If we we define

$$Z^{22}_{3n} = \frac{D(t_n)}{d(t_n)} \tag{2.103}$$

$$Z^{21}_{3n} = \frac{n(t_n)}{d(t_n)} \tag{2.104}$$

$$Z^{23}_{3n} = \frac{D(t_n)}{d(t_n)} \tag{2.105}$$

The limiting variance of Z^2_{3n}, with the normalizing constant $r_n = \sqrt{nd(t_n)}$, is as following;

$$\begin{aligned}
\lim_{t_n\to\infty} var(\sqrt{nd_t}Z^2_{3n}) &= \frac{\sum_{k=0}^{\infty}\sum_{l=0}^{\infty} E(I(Y \geq k \vee l)e^{-t_n}\frac{t_n^k}{k!}e^{-t_n}\frac{t_n^l}{l!} - d(t_n)^2}{d(t_n)} \\
&= \lim_{t_n\to\infty} \frac{E[\bar{G}(M_1(t_n) \vee M_2(t_n))]}{E[\bar{G}(M_1(t_n))]} - d_t \\
&= \lim_{t_n\to\infty} \frac{E[\bar{G}(M_1(t_n) \vee M_2(t_n))]}{E[\bar{G}(M_1(t_n))]}
\end{aligned} \tag{2.106}$$

Finally, the limiting covariance between Z^1_{3n} and Z^2_{3n} with normalizing constant is as below

$$\begin{aligned}
\lim_{t_n\to\infty} cov(\sqrt{nd(t_n)}Z^1_{3n}, \sqrt{nd(t_n)}Z^2_{3n}) &= \lim_{t_n\to\infty} \frac{E[(1-p)\int_{M_1(t_n)\vee M_2(t_n)} F(y)\mathrm{d}G(y)]}{E[\bar{G}(M_1(t_n))]} - n_t \\
&= \lim_{t_n\to\infty} \frac{E[(1-p)\int_{M_1(t_n)\vee M_2(t_n)} F(y)\mathrm{d}G(y)]}{E[\bar{G}(M_1(t_n))]}
\end{aligned} \tag{2.107}$$

In the section (2.2), it has been shown based on Eq. (2.41) that

$$\frac{E[\int_{M_1(t_n)\vee M_2(t_n)} F(y)\mathrm{d}G(y)]}{E[\bar{G}(M_1(t_n))]} \leq 1 \tag{2.108}$$

and

$$\frac{E[\bar{G}(M_1(t_n) \vee M_2(t_n))}{E[\bar{G}(M_1(t_n))]} \leq 1 \tag{2.109}$$

Currently, for exponential distribution we were not able to calculate the limit of Eqs. (2.108) and (2.109) analytically and we only found the upper bound for them. This issue will be addressed in future publication. At this point, we used simulation technique to study Eqs. (2.108) and (2.109) to get an understanding how variance of our estimator behave.

Fig. 2.2 demonstrates variance histogram for limiting distribution p_{3n} when F and G have exponential distribution with normalizing constant, $\sqrt{nd(t_n)}$. It can be seen that in the all of the variance are almost zero $(6.8e-4)$ for 1000 different simulations cases. As a result, we conclude that the limiting variance goes to zero for large enough t_n. In this case, we have a degenerate distribution where its mass is placed on a single point.

In the next section, we assumed that F and G with Pareto distribution and we were able to find an analytical solution for normalized limiting distribution of p_{3n}. We need to use following Lemmas and concept of Regularly Varying Function for analysis.

Lemma 2.3: If $M_1(t)$ and $M_2(t)$ are independent Poisson with mean of t,

1. $\frac{M_1(t)}{t} \xrightarrow{P} 1$ as $t \to \infty$.
2. $\frac{M_1(t)\vee M_2(t)}{t} \xrightarrow{P} 1$ as $t \to \infty$.

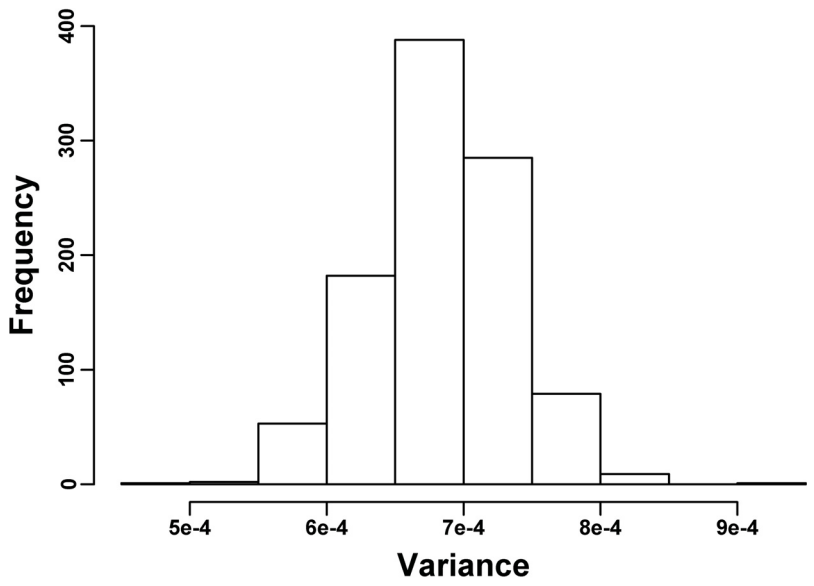

Figure 2.2: Histogram plot For variance of p_{3n} when F and G are exponential distribution.

3. $E|\frac{M_1(t)}{t} - 1| \to 0$ converge in mean.

Proof:

1)

$$\begin{aligned} &E(M(t)/t) = 1 \\ &var(M(t)/t) = \frac{1}{t} \to 0 \quad \text{when t} \to \infty \end{aligned} \tag{2.110}$$

2)

$$\frac{M_1(t) \vee M_2(t)}{t} = \frac{M_1(t) + M_2(t)}{2t} + \frac{\mid M_1(t) - M_2(t) \mid}{2t} \to 1$$

When t $\to \infty$ (2.111)

Based on proof of part 1

Lemma 2.4: If $M_1(t)$ and $M_2(t)$are independent Poisson with mean t, and $\bar{G}(t) = \frac{1}{t^\alpha}$

for $t > 0$ and some $\alpha > 0$ (i. e. , G is Pareto)

$$\lim_{t\to\infty} \frac{\int_{M_1(t)\vee M_2(t)} F(y)\mathrm{d}G(y)}{\bar{G}(M_1(t)\vee M_2(t))} \to 1 \tag{2.112}$$

Proof:

$$\begin{aligned}\lim_{t\to\infty} \frac{\int_{M_1(t)\vee M_2(t)} F(y)\mathrm{d}G(y)}{\bar{G}(M_1(t)\vee M_2(t))} &= \lim_{t\to\infty} \frac{-F(M_1(t)\vee M_1(t))g(M_1(t)\vee M_1(t))}{-g(M_1(t)\vee M_1(t))} \\ &= 1\end{aligned} \tag{2.113}$$

Then the variances of Z_{3n}^{11} is calculated as following

$$\begin{aligned}\lim_{t_n\to\infty} var(\sqrt{nd(t_n)}Z_{3n}^{11}) &= \lim_{t\to\infty} \frac{E[(1-p)\int_{M_1(t_n)\vee M_2(t_n)} F(y)\mathrm{d}G(y)]}{E[\bar{G}(M_1(t_n))]} \\ &= \lim_{t\to\infty} \frac{E[(1-p)\int_{M_1(t_n)\vee M_2(t_n)} F(y)\mathrm{d}G(y)]}{E[\bar{G}(M_1(t_n))]} \\ &= \lim_{t\to\infty} \frac{(1-p)E[\bar{G}(M_1(t_n)\vee M_2(t_n))\cdot \frac{-F(M_1(t_n)\vee M_1(t_n))g(M_1(t_n)\vee M_1(t_n))}{-g(M_1(t_n)\vee M_1(t_n)}]}{E(\bar{G}(M_1(t_n)))}\end{aligned} \tag{2.114}$$

Using L'Hospital's rule and Dominate Converge Theorem, and assuming $\bar{G}$ has a Pareto distribution (i.e. $\bar{G}(y) = \frac{1}{y^\alpha} = y^{-\alpha}$ and $\alpha > 0$),

$$\lim_{t_n\to\infty} var(\sqrt{nd(t_n)}Z_{3n}^{11}) = \lim_{t\to\infty} \frac{(1-p)E[\frac{1}{((M_1(t_n)\vee M_2(t_n)))^\alpha}]}{E[\frac{1}{(M_1(t_n))^\alpha}]} \tag{2.115}$$

To find limiting distribution of Eq. (2.115), we need the following two Lemmas based on it.

Lemma 2.5: If $M_1(t)$ and $M_2(t)$ are independent process with mean t,

1. $\frac{M_1^\alpha(t)}{t^\alpha} \xrightarrow{P} 1$ as $t \to \infty$.

2. $\frac{(M_1(t)\vee M_2(t))^\alpha}{t^\alpha} \xrightarrow{P} 1$ as $t \to \infty$.

Lemma 2.6: If $M_1(t)$ is a Poisson variable with mean t,

$$E[\frac{t^\alpha}{M_1^\alpha(t)}] \xrightarrow{P} 1 \quad \text{as} \quad t \to \infty \tag{2.116}$$

Proof: We have

$$\frac{M^\alpha(t)}{t^\alpha} \xrightarrow{P} 1 \Rightarrow \frac{t^\alpha}{M^\alpha(t)} \xrightarrow{P} 1 \quad \text{as} \quad t \to \infty \tag{2.117}$$

Based on dominated convergence theorem (Bartle, 2014)for proving Lemma, it is enough to check that

$$\max_{t\geq 0} E(\frac{t^\alpha}{M^\alpha(t)})^{1+\epsilon} < \infty \quad \text{for some} \quad \epsilon > 0. \tag{2.118}$$

We can show that $\max_{t\geq 0} E(\frac{t^k}{M^k(t)}) < \infty$ for any $k \geq 1$ when k is integer.

$$\begin{aligned} E(\frac{t^k}{M^k(t)}) &= \sum_{r=1}^{\infty} \frac{t^k}{r^k} e^{-t} \cdot \frac{t^r}{r!} \\ &= \sum_{r=1}^{\infty} \frac{1}{r^k} e^{-t} \cdot \frac{t^{r+k}}{r!} \\ &= \sum_{r=1}^{\infty} \frac{(r+k)!}{r!} \cdot \frac{1}{r^k} e^{-t} \cdot \frac{t^{r+k}}{(r+k)!} \\ &= \sum (1+\frac{1}{r})(1+\frac{2}{r}) \cdots (1+\frac{k}{r}) e^{-t} \frac{t^{r+k}}{(r+k)!} \\ &\leq (k+1)! \sum_{r=1}^{\infty} e^{-t} \frac{t^{r+k}}{(r+k)!} = (k+1)!(1 - \sum_{S=1}^{k} e^{-t} \frac{t^S}{S!}) \leq (k+1)! \end{aligned} \tag{2.119}$$

So based on Lemma 2.5 and 2.6 we can conclude that;

$$\lim_{t_n\to\infty} var(\sqrt{nd(t_n)}Z_{3n}^{11}) = \lim_{t\to\infty} \frac{(1-p)E[\frac{1}{((M_1(t_n)\vee M_2(t_n)))^{\alpha}}]}{E[\frac{1}{(M_1(t_n))^{\alpha}}]} = 1-p \tag{2.120}$$

The limiting variance of Z_{3n}^2 when assuming $\bar{G}$ has a Pareto distribution (i.e. $\bar{G}(y) = \frac{1}{y^{\alpha}} = y^{-\alpha}$ and $\alpha > 0$),

$$\lim_{t_n\to\infty} var(\sqrt{nd(t_n)}Z_{3n}^{2}) = \lim_{t\to\infty} \frac{E[((M_1(t_n)\vee M_2(t_n)))^{-\alpha}]}{E[(M_1(t_n))^{-\alpha}]} = 1 \tag{2.121}$$

Finally, the normalized limiting covariance between Z_{3n}^1 and Z_{3n}^2

$$\lim_{t_n\to\infty} cov(\sqrt{nd(t_n)}Z_{3n}^{1}, \sqrt{nd(t_n)}Z_{3n}^{2}) = \lim_{t\to\infty} \frac{(1-p)E[\frac{1}{((M_1(t_n)\vee M_2(t_n)))^{\alpha}}]}{E[\frac{1}{(M_1(t_n))^{\alpha}}]} = 1-p \tag{2.122}$$

Therefore, p_{3n} assuming $\bar{G}$ has a Pareto distribution (i.e. $\bar{G}(y) = \frac{1}{y^{\alpha}} = y^{-\alpha}$ and $\alpha > 0$) has a limiting bivariate normal distribution as below;

$$\sqrt{nd(t_n)} \begin{bmatrix} Z_{3n}^1 \\ \\ Z_{3n}^2 \end{bmatrix} \to N\left(\begin{bmatrix} 0 \\ \\ 0 \end{bmatrix}, \begin{bmatrix} 1-p & 1-p \\ \\ 1-p & 1 \end{bmatrix} \right) \tag{2.123}$$

2.2 Estimator 2.

As we have seen in the previous section, we were not able to calculate analytically the limiting distribution for P_{3n} when time of event and check ups (i.e. F and G) had exponential distribution. To overcome this problem a new smooth estimator proposed as follows

$$P_{4n} = \frac{\frac{1}{n}\sum_{i=1}^{n}\delta_i \sum_{k=1}^{\infty} I(Y_i \geq \ln(k)) e^{-t_n}\frac{t_n^k}{k!}}{\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{\infty} I(Y_i \geq \ln(k)) e^{-t_n}\frac{t_n^k}{k!}} \tag{2.124}$$

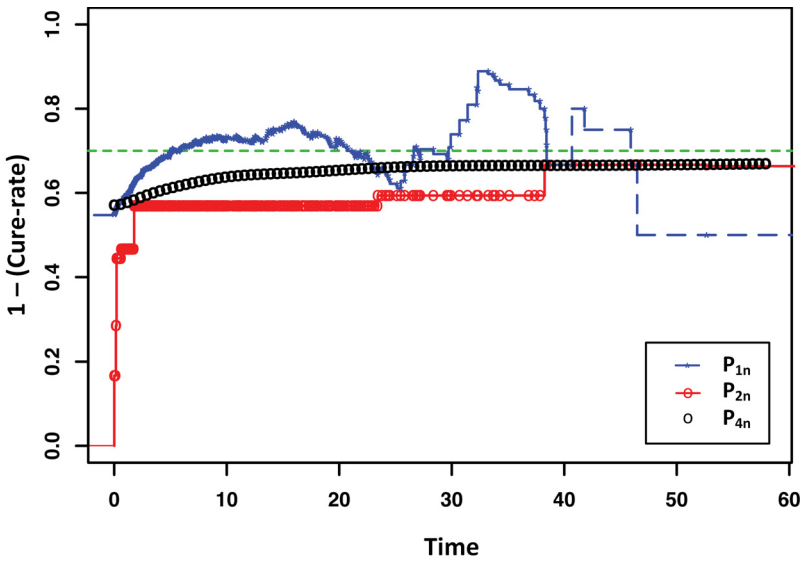

Figure 2.3: Sample-plot of cure rate estimation versus time. The green, blue, red and black curves are assumed cure-rate, p_{n1}, p_{n2} and p_{n4}, respectively. ($F(y) = exp(0.4)$, $G(y) = exp(0.1)$, $n = 500$ and $1 - p = 0.7$)

Fig. 2.3 demonstrates sample plot of cure rate estimation versus time for Eq. (2.124). We assumed that cure-rate is equal to 0.7 and generated time to the event and the check ups using $exp(0.4)$ and $exp(0.2)$ distribution functions, respectively. Fig. 2.3 shows that the new modified smooth estimator p_{4n} estimator has better performance

in comparison to cure-rate estimated by Eq. (1.20) and Eq.(1.21); however, it coverages slower than p_{3n}.

Regularly Varying Function

Roughly speaking, regularly varying functions are the function such that their asymptotic behavior is like power functions.

Definition: $f(x)$ is of regular variation of order α ($f \in RV_\alpha$) $-\infty \leq \alpha \leq \infty$ if

$$\lim_{t\to\infty} \frac{f(tx)}{f(t)} = x^\alpha \text{for all} \qquad x > 0 \tag{2.125}$$

A function which is satisfying Eq. (2.125) with $\alpha = 0$ in called slowly varying function (De Haan and Ferreira, 2007).

Theorem 2.3: Suppose $\{N_n, n \geq 1\}$ is a sequence of nonnegative random variable such that (Resnick, 2007)

$$\frac{N_n}{n} \xrightarrow{P} 1 \tag{2.126}$$

If $a(.) \in RV_\rho$ and $P\{N > 0\} = 1$, therefore, we have;

$$\frac{a(N_n)}{a(n)} \xrightarrow{P} 1^\rho \tag{2.127}$$

Lemma 2.5: If $M_1(t)$ and $M_2(t)$ are independent process with mean t,

1. $\ln(\frac{M_1(t)}{t}) \xrightarrow{P} 0 \quad \& \quad \ln(\frac{M_1(t)\vee M_2(t)}{t}) \xrightarrow{P} 0 \quad t \to \infty$
2. $\frac{\ln M_1(t)}{\ln t} = \frac{\ln(\frac{M_1(t)}{t})+\ln t}{\ln t} \xrightarrow{P} 1 \quad t \to \infty$

3. $E|\frac{\ln M_1(t)}{\ln t} - 1| \to 0$ converge in mean.

4. $\frac{\ln M_1(t) \vee M_2(t)}{\ln t} = \frac{\ln(\frac{M_1(t)\vee M_2(t)}{t}) + \ln t}{\ln t} \xrightarrow{P} 1 \quad t \to \infty$

They can be proved simply based on Lemma 2.3 proof.

Lemma 2.6: Assuming $\bar{G}(x)$ is of regular variation of order α ($f \in RV_\alpha$) $-\infty \leq \alpha \leq \infty$.

$$\lim_{t\to\infty} \frac{\bar{G}(tx)}{\bar{G}(t)} \to x^{-\alpha} \tag{2.128}$$

therefor

$$\lim_{t\to\infty} \frac{\bar{G}(\ln tx)}{\bar{G}(\ln t)} \to 1 \tag{2.129}$$

Proof:

$$\frac{f(tx)}{f(t)} = \frac{\bar{G}(\ln t + \ln x))}{\bar{G}(\ln t)} = \frac{\bar{G}(\ln t(1 + \frac{\ln x}{\ln t})))}{\bar{G}(\ln t)} \to 1 \tag{2.130}$$

2.2.1 Consistency of 2nd Estimator

Similarly as what we proved for p_{3n}, we will apply the result of definition 2.1, specifically

$$p_{4n} = \frac{1/n \sum_{i=1}^{n} \delta_i \sum_{k=1}^{\infty} I(Y_i \geq \ln k) e^{-t_n} \frac{t_n^k}{k!}}{1/n \sum_{i=1}^{n} \sum_{k=1}^{\infty} I(Y_i \geq \ln k) e^{-t_n} \frac{t_n^k}{k!}} \to 1 - p \tag{2.131}$$

when t_n are approaching to ∞.

Therefore, we need to prove that

$$\text{Bias} = \frac{n_t^{\ln}}{d_t^{\ln}} - (1 - p) \to 0 \tag{2.132}$$

and

$$\frac{N_t^{\ln}}{D_t^{\ln}} - \frac{n_t^{\ln}}{d_t^{\ln}} \to 0 \tag{2.133}$$

Where

$$N_t^{\ln} = 1/n \sum_{i=1}^{n} \delta_i \sum_{k=1}^{\infty} I(Y_i \geq \ln k) e^{-t_n} \frac{t_n^k}{k!} \tag{2.134}$$

and

$$D_t^{\ln} = 1/n \sum_{i=1}^{n} \sum_{k=1}^{\infty} I(Y_i \geq \ln k) e^{-t_n} \frac{t_n^k}{k!} \tag{2.135}$$

Therefore, by repeating the same procedure that we have done to calculate expectation p_{3n}, we will have expectation of p_{4n}

$$n_t^{\ln} = (1-p) \sum_{k=1}^{\infty} \int_{\ln k}^{\infty} \mathrm{F}(y)\, \mathrm{d}G(y) e^{-t_n} \frac{t_n^k}{k!} \tag{2.136}$$

and

$$d_t^{\ln} = \sum_{k=1}^{\infty} \int_{\ln k}^{\infty} \mathrm{d}G(y) e^{-t_n} \frac{t_n^k}{k!} \tag{2.137}$$

As a result, based on lemma 2.1

$$\lim_{t\to\infty} \frac{n_t^{\ln}}{d_t^{\ln}} - (1-p) = 0 \tag{2.138}$$

Next, for calculating variance

$$\begin{aligned} p_{4n} - (1-p) &= \frac{N_t^{\ln}}{D_t^{\ln}} - \frac{n_t^{\ln}}{d_t^{\ln}} + \frac{n_t^{\ln}}{d_t^{\ln}} - (1-p) \\ &= \frac{N_t^{\ln} - n_t^{\ln}}{D_t^{\ln}} - \frac{n_t^{\ln}}{d_t^{\ln}} \frac{D_t^{\ln} - d_t^{\ln}}{D_t^{\ln}} + \frac{n_t^{\ln}}{d_t^{\ln}} - (1-p) \\ &= \frac{n_t^{\ln}}{d_t^{\ln}} \frac{\frac{N_t^{\ln} - n_t^{\ln}}{n_{\ln t}}}{\frac{D_t^{\ln}}{d_t^{\ln}}} - \frac{n_t^{\ln}}{d_{\ln t}} \frac{\frac{D_t^{\ln} - d_t^{\ln}}{d_{\ln t}}}{\frac{D_t^{\ln}}{d_t^{\ln}}} + \frac{n_t^{\ln}}{d_t^{\ln}} - (1-p) \end{aligned} \tag{2.139}$$

Therefore, the limiting variance of p_{4n} based on Lemma 2.1 and 2.2,

$$p_{4n} - \frac{n_t^{\ln}}{d_t^{\ln}} \xrightarrow{P} 0 \tag{2.140}$$

2.2.2 Limiting Distribution of 2nd Estimator

In this section we will derive limiting distribution of p_{4n} by following the same approach which was p_{3n} and p_{1n}, but assuming that $\bar{G}(\ln x)$ is $RV_{-\alpha}$ for some $\alpha > 0$

$$\begin{aligned} p_{4n} - \frac{n_t^{\ln}}{d_t^{\ln}} &= \frac{1/n \sum_{i=1}^{n} \delta_i \sum_{k=1}^{\infty} I(Y_i \geq \ln k) e^{-t_n} \frac{t_n^k}{k!}}{1/n \sum_{i=1}^{n} \sum_{k=1}^{\infty} I(Y_i \geq \ln k) e^{-t_n} \frac{t_n^k}{k!}} \\ &\quad - \frac{(1-p) \sum_{k=1}^{\infty} \int_{\ln k}^{\infty} \mathrm{F}(y)\, \mathrm{d}G(y) e^{-t_n} \frac{t_n^k}{k!}}{\sum_{k=1}^{\infty} \int_{\ln k}^{\infty} \mathrm{d}G(y) e^{-t_n} \frac{t_n^k}{k!}} \\ &= \frac{N_t^{\ln} - n_t^{\ln}}{D_t^{\ln}} - \frac{n_t^{\ln}}{d_t^{\ln}} \cdot \frac{D_t^{\ln} - d_t^{\ln}}{{}_t^{\ln}} \\ &= Z_{4n}^1 - \frac{n_t^{\ln}}{d_t^{\ln}} Z_{4n}^2 \end{aligned} \tag{2.141}$$

Where

$$Z_{4n}^1 = \frac{N_t^{\ln} - n_t^{\ln}}{D_t^{\ln}} \tag{2.142}$$

and

$$Z_{4n}^2 = \frac{D_t^{\ln} - d_t^{\ln}}{D_t^{\ln}} \tag{2.143}$$

The limiting variance of Z_{4n}^1 and Z_{4n}^2 have been calculated similar to preview section as follow

$$\lim_{t_n \to \infty} var(\sqrt{n d_{\ln t}} Z_{4n}^1) = \lim_{t_n \to \infty} \frac{E[(1-p) \int_{\ln(M_1(t_n) \vee M_2(t_n))} F(y) \mathrm{d}G(y)]}{E[\bar{G}(\ln(M_1(t_n)))]} \tag{2.144}$$

and

$$\begin{aligned}\lim_{t_n\to\infty} var(\sqrt{nd_t^{\ln}}Z_{4n}^2) &= \lim_{t_n\to\infty} \frac{E[\bar{G}\ln(M_1(t_n)\vee M_2(t_n))]}{E[\bar{G}\ln(M_1(t_n))]} \\ &= \lim_{t_n\to\infty} \frac{E[(M_1(t_n)\vee M_2(t_n))^{-\alpha}]}{E[(M_1(t_n))^{-\alpha}]}\end{aligned} \tag{2.145}$$

by our assumption that $\bar{G}(\ln x)$ is $RV_{-\alpha}$ and using theorem 2.4.

Finally, the limiting covariance between Z_{4n}^1 and Z_{4n}^2 with normalizing constant is as below

$$\lim_{t_n\to\infty} cov(\sqrt{nd_t^{\ln}}Z_{4n}^1, \sqrt{nd_t^{\ln}}Z_{4n}^2) = \lim_{t_n\to\infty} \frac{E[(1-p)\int_{\ln(M_1(t_n)\vee M_2(t_n))} F(y)\mathrm{d}G(y)]}{E[\bar{G}(\ln(M_1(t_n)))]} \tag{2.146}$$

As a result, the limiting distribution of modified estimator is the same as limiting distribution of p_{3n} with Pareto distribution for $\bar{G}(y)$ that has been provided in section 2.1.2 which is as follow

$$\sqrt{nd_t^{\ln}}\begin{bmatrix} Z_{4n}^1 \\ Z_{4n}^2 \end{bmatrix} \to N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1-p & 1-p \\ 1-p & 1 \end{bmatrix}\right) \tag{2.147}$$

example 2.2: if we had assumed Pareto distribution for $\bar{G}$, it would not have changed the fact that we could found the limiting distribution.

The calculation is as following;

Suppose $\bar{G}(y) = \frac{1}{y^\alpha}$ with Pareto distribution, so

$$\bar{G}(\ln(y)) = \frac{1}{\ln(y)^\alpha} \tag{2.148}$$

As a result $\bar{G}(\ln(y))$ is a regular variation of order 0, based on Lemma 2.6.

For finding the liming distribution, we will follow the same steps that we did for p_{3n},

and p_{4n} .

$$\sqrt{nd_t^{\ln}}\begin{bmatrix} Z_{4n}^1 \\ \\ Z_{4n}^2 \end{bmatrix} \to \left(\begin{bmatrix} 0 \\ \\ 0 \end{bmatrix}, \begin{bmatrix} 1-p & 1-p \\ \\ 1-p & 1 \end{bmatrix} \right) \tag{2.149}$$

if $t_n \to \infty$, $d_t^{\ln} \to \infty \to$.

Assume $\bar{G}(y) = e^{-\alpha y}$ has an exponential distribution, so $\bar{G}(\ln(y)) = \frac{1}{y^\alpha}$, so the result is the same as what we got for limiting distribution of p_{3n} with Pareto distribution in Eq. 2.123.

Remark: In the above we showed that

$$\begin{aligned} \sqrt{nd(t_n)}(p_{3n} - \frac{n_(t_n)}{d_(t_n)}) &\to N(0, p(1-p)) \\ \sqrt{nd_t^{\ln}}(p_{4n} - \frac{n_t^{\ln}}{d_t^{\ln}}) &\to N(0, p(1-p)) \end{aligned} \tag{2.150}$$

if $t_n \to \infty$ and $nd(t_n) \to \infty$ and $nd_t^{\ln} \to \infty$, respectively.

It follows that $\sqrt{nd(t_n)}(p_{3n} - (1-p))$ and $\sqrt{nd_t^{\ln}}(p_{4n} - (1-p))$ also have the same limiting distributions, provided

$$\begin{aligned} \sqrt{nd(t_n)}(\frac{n(t_n)}{d_(t_n)} - (1-p)) &\to N(0, p(1-p)) \\ \sqrt{nd_t^{\ln}}(\frac{n_t^{\ln}}{d_t^{\ln}} - (1-p)) &\to N(0, p(1-p)) \end{aligned} \tag{2.151}$$

respectively.

At this time we do not have any simple sufficient conditions for 2.151.

Chapter 3

Selecting Optimum Smoothing Parameter

There are many techniques for finding the optimum value that we have the best estimate at it such as least squared error and likelihood based cross-validation. In this chapter we applied jackknife (i.e. leave one out) least-squared-error cross-validation method to find the Poisson parameter with the optimum estimated cure-rate.

3.1 Variance-Bias Trade-Off

We have used mean-square-error (MSE) scheme and we decompose the error term of our estimation to bias and variance. Then the trade-off between them helps us to choose the optimal estimated value.

$$\mathrm{MSE}(p_{3n}) = E(p_{3n} - (1-p))^2 \tag{3.1}$$

where p_{3n} and p are proposed estimator and cure-rate, respectively and decomposition form of MSE to variance and bias is as follow

$$\text{MSE}(p_{3n}) = var(p_{3n}) + \text{bias}(p_{3n})^2 \tag{3.2}$$

In this study, we have used Jackknife method to estimate the bias. Also, variance can be calculated directly based on (Sen and Tan, 2008) as

$$var(p_{3n}) = \frac{p_{3n}(t_n)(1 - p_{3n}(t_n))}{nD(t_n)} \tag{3.3}$$

Example 3.1: If $\bar{G}(y)$ and $F(y)$ has exponential distribution (i.e. $F(y) = 1 - e^{-y}$ and $\bar{G}(y) = e^{-y}$). Therefor,

$$\begin{aligned} d(t_n) &= e^{-t_n} \sum_{k=0}^{\infty} \bar{G}(k) \frac{t_n^k}{k!} \\ &= e^{-t_n} \sum_{k=0}^{\infty} e^{-k} \frac{t_n^k}{k!} \\ &= e^{-t_n} e^{t_n/e} \end{aligned} \tag{3.4}$$

and

$$\begin{aligned} n(t_n) &= (1-p)e^{-t_n} \sum_{k=0}^{\infty} \int_k^{\infty} F(y)\bar{G}(k) \frac{t_n^k}{k!} \\ &= (1-p)e^{-t_n} \sum_{k=0}^{\infty} \int_k^{\infty} (1 - e^{-k}) e^{-k} \frac{t_n^k}{k!} \\ &= (1-p)[e^{-t_n} e^{t_n/e} - 1/2 e^{-t_n} e^{t_n/e^2}] \end{aligned} \tag{3.5}$$

As a result, the bias is

$$\begin{aligned} \text{Bias} &= \frac{(1-p)[e^{-t_n}e^{t_n/e} - 1/2e^{-t_n}e^{t_n/e}]}{e^{-t_n}e^{t_n/e}} - (1-p) \\ &= (1-p)\left[-\frac{1/2e^{-t_n}e^{t_n/e^2}}{e^{-t_n}e^{t_n/e}}\right] \end{aligned} \tag{3.6}$$

and the variance is

$$\text{var}(p_{3n}) = \frac{p(1-p)}{nD(t_n)} \tag{3.7}$$

where

$$D(t_n) = 1/n \sum_{i=1}^{n} \sum_{k=0}^{\infty} e^{-k} e^{-t_n} \frac{t_n^k}{k!} = e^{-t_n} \tag{3.8}$$

Therefor, the MSE($p_{3n}(t_n)$) is as follows

$$\text{MSE} = (1-p)^2 \frac{e^{-2t_n e^{-1}(1-e^{-1})}}{4} + \frac{p.(1-p)}{ne^{-t_n}} \tag{3.9}$$

and the optimum smoothing parameter, t_n, can be obtain as follows

$$\frac{\partial}{\partial t_n}\text{MSE} = -\frac{(1-p)^2}{2}(e^{-1})(1-(e^{-1}))e^{-2t_n(e^{-1})(1-(e^{-1})} + \frac{p(1-p)}{n}e^{t_n} \tag{3.10}$$

as a result, the optimum t_n is

$$t_n = \frac{\ln\left(\frac{n(1-p)}{p}(e^{-1})(1-(e^{-1})\right)}{1+2(e^{-1})(1-(e^{-1})} \tag{3.11}$$

3.2 Jackknife Estimation: Illustration

Jackknife procedure has been proposed by Quenouille (1949) for bias estimation. Chaubey and Sen (2009) proposed Jackknife method to select smoothing parameter.

In this section we also use Jackknife technique to estimate the bias. Basically, the jackknife is a resampling technique where the bias is estimated by removing one of n observations and aggregating the estimates of each $n-1$ remaining observations. Therefore, the bias can be calculated as

$$\text{Bias}(p_{3n}(t_n)) = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{N_j(t_n)}{D_j(t_n)} - \frac{N(t_n)}{D(t_n)}\right) \tag{3.12}$$

where $N(t_n)$ and $D(t_n)$ are defined Eqs. (2.89) and (2.90) and

$$N_j(t_n) = \frac{1}{n-1}\sum_{k=0}^{\infty}\left(\sum_{i=1,i\neq j}^{n}\delta_i I(Y_i \geq k)\right)e^{-t_n}\frac{t_n^k}{k!} \tag{3.13}$$

$$D_j(t_n) = \frac{1}{n-1}\sum_{k=0}^{\infty}\left(\sum_{i=1,i\neq j}^{n} I(Y_i \geq k)\right)e^{-t_n}\frac{t_n^k}{k!} \tag{3.14}$$

As a result, Based on Eq. (3.3) and (3.12), the MSE is given by

$$\text{MSE}(p_{3n}(t_n)) = \frac{p_{3n}(t_n)(1-p_{3n}(t_n))}{nD(t_n)} + \left(\frac{1}{n}\sum_{j=1}^{n}(\frac{N_j(t_n)}{D_j(t_n)} - \frac{N(t_n)}{D(t_n)})\right)^2 \tag{3.15}$$

Optimum t_n corresponds to the time with minimum mean square error and analytical analysis of finding time with minimum mean square error was postponed to the future work. Here, we numerical calculated the optimum t_n for the example provided in Chapter 2 where we generated observation using exponential distribution and assuming cure-rate equal 0.3. Figure 3.1 demonstrates square error versus time for p_{3n}. It can be seen that at 8.29, MSE is minimum and we found that estimated curate is 0.31.

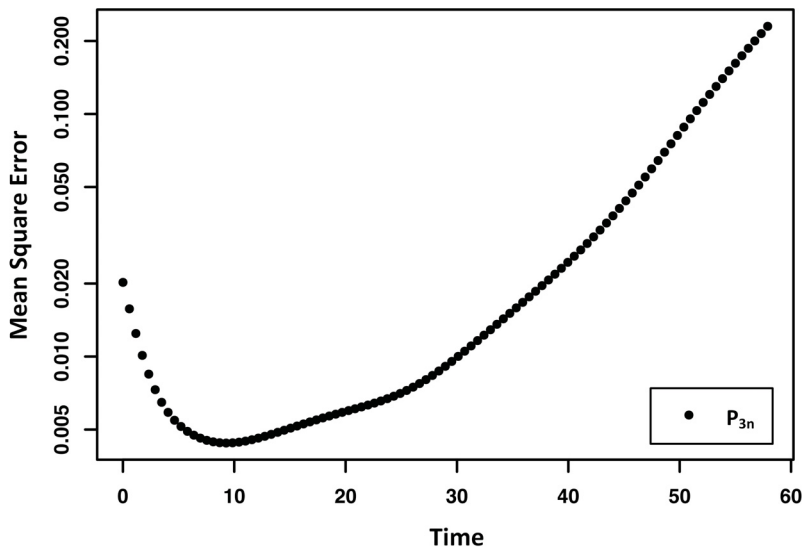

Figure 3.1: Sample mean square error for p_{3n} versus t_n, ($F(y) = \text{Exp}(0.4)$, $G(y) = \text{Exp}(0.2)$, $n = 500$ and cure-rate $= 0.3$)

Similarly, the bias for p_{4n} can be calculated as

$$\text{Bias}(p_{4n}(t_n)) = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{N_j^{\ln}(t_n)}{D_j^{\ln}(t_n)} - \frac{N^{\ln}(t_n)}{D^{\ln}(t_n)}\right) \tag{3.16}$$

where $N^{\ln}(t_n)$ and $D^{\ln}(t_n)$ are defined Eqs. (2.134) and (2.135) and

$$N_j^{\ln}(t_n) = \frac{1}{n-1}\sum_{k=0}^{\infty}\left(\sum_{i=1,i\neq j}^{n}\delta_i I(Y_i \geq \ln(k))\right)e^{-t_n}\frac{t_n^k}{k!} \tag{3.17}$$

$$D_j^{\ln}(t_n) = \frac{1}{n-1}\sum_{k=0}^{\infty}\left(\sum_{i=1,i\neq j}^{n} I(Y_i \geq \ln(k))\right)e^{-t_n}\frac{t_n^k}{k!} \tag{3.18}$$

As a result, the MSE for $p_4 n$ is;

$$\text{MSE}(p_{4n}(t_n)) = \frac{p_{4n}(t_n)(1-p_{4n}(t_n))}{nD^{\ln}(t_n)} + \left(\frac{1}{n}\sum_{j=1}^{n}(\frac{N_j^{\ln}(t_n)}{D_j^{\ln}(t_n)} - \frac{N^{\ln}(t_n)}{D^{\ln}(t_n)})\right)^2 \tag{3.19}$$

Here, we numerical calculated the optimum t_n for the example provided in Chapter 2 where we generated observation using exponential distribution and assuming cure-rate equal 0.3. Figure 3.2 demonstrates square error versus time for p_{4n}. It can be seen that at 17.95, MSE is minimum and we found that estimated curate is 0.36.

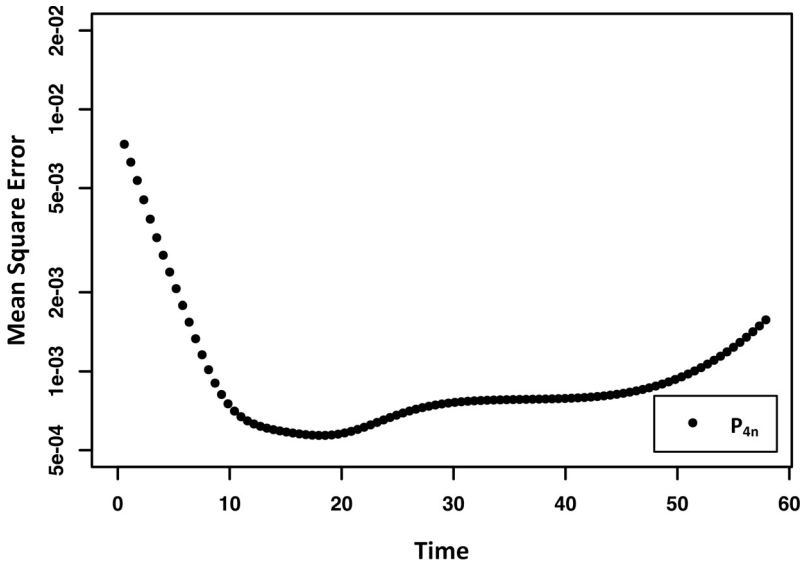

Figure 3.2: Sample mean square error for p_{4n} versus t_n, ($F(y) = \text{Exp}(0.4)$, $G(y) = \text{Exp}(0.2)$, $n = 500$ and cure-rate $= 0.3$)

3.3 Example with Real Data

In this section, we used our proposed estimators with data reported in (Finkelstein and Wolfe, 1985). The data was related to development lung tumor in mice germ-free and conventional environments in presence of carcinogens. Table 3.3 summarizes the data.

Necropsy finding	Indivial age at death (days)
A. Convetional mice (96)	
Lung tumor	381,477,485,515,539,563,565,582,603,616,624,650,651,656,659,672, 679,698,702,709,723,731,775,779,795,811,839
No lung tumor	45,198,215,217,257,262,266,371,431,447,454,459,475,479,484,500, 502,503,505,508,516,531,541,553,556,570,572,575,577,585,588,594, 600,601,608,614,616,632,632,638,642,642,642,644,644,647,647,653, 659,660,662,663,667,667,673,673,677,689,693,718,720,721,728,760, 762,773,777,815,886
B. Germfree mice (48)	
Lung tumor	546,609,692,692,710,752,753,781,782,789,808,810,814,842,846,851, 871,873,876,888,888,890,894,896,911,913,914,914,916,921,921,926, 936,945,1008
No lung tumor	412,524,647,648,695,785,814,817,851,880,913,942,986

Table 3.1: Ages at death of untreated male mice dying with lung cancer

We determined optimal t_n numerically. It can be seen that in Fig 3.3, at 496.8 and 627.57 MSE is minimum for conventional and germ-free environments respectively.

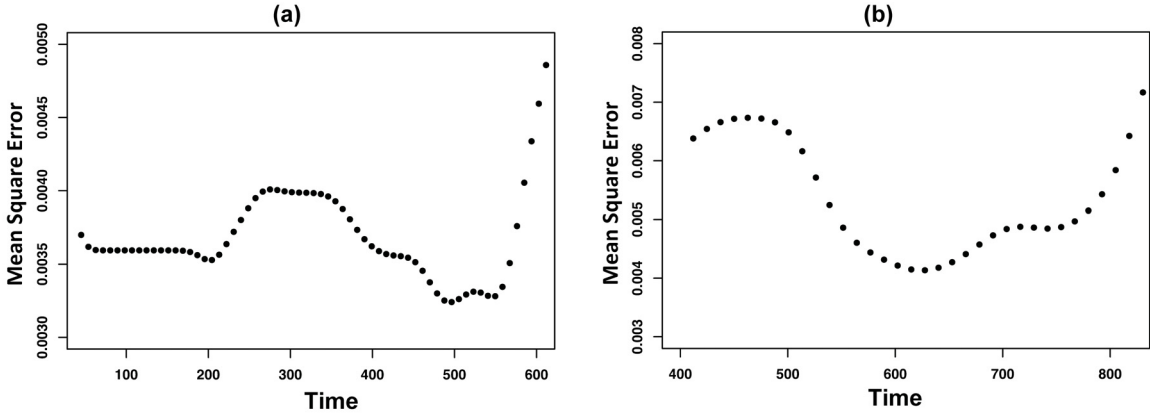

Figure 3.3: MSE of p_{3n} for Mice in (a) conventional and (b) germ-free environments

Figure 3.4 demonstrates results of various estimator versus time. It can be seen that our new estimator p_{3n} has fast convergence and much smoother than frivolously proposed estimation techniques (i.e. p_{1n} and p_{2n}).

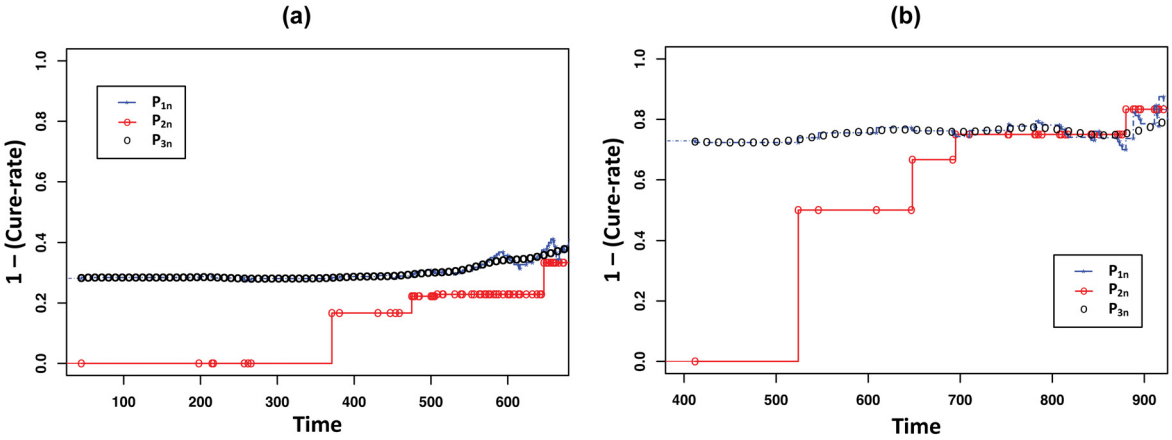

Figure 3.4: Cure-rate estimation for mice in (a) conventional and (b) germ-free environments.

Chapter 4

Conclusion and Future Work

4.1 Conclusion

Estimating cure-rate for diseases with high rate of mortality helps to develop and test the current treatments. In this work, we proposed two new estimators of cure-rate under case-1 interval censoring via Poisson smoothing. Our first estimator is Eq. 2.11 ,p_{3n}, is the smooth version of Eq. 1.20 ,p_{1n}, which was proposed by (Sen and Tan, 2008). The non-smooth estimator is noisy and there was not a clear cut-off-point for it. In comparison to p_{1n}, p_{3n} is less noisy and the choice of cut-of-point is easier.

The fraction structure of our estimators always makes the analysis challenging, therefore, we decided to write estimator via telescoping which let us to treat denominator and numerator separately. However,we found out that the estimator has a degenerate limit distribution if we assume exponential distribution for time of check-up and time to the event; after trying Pareto distribution for the time of check-up and time to the event we obtain limiting normal distribution.

The form of Pareto distribution which is a case of regularly varying function

motivate us to proposed another estimator Eq. 2.124, p_{4n} by modifying p_{3n}. The new estimator based on regularly varying functions and Poisson distribution characteristic has limiting normal distribution.

Finally, in the last chapter we found the smoothing parameter by jackknife technique and we test our estimator on real data.

4.2 Future Work

For future work we can extend our study to

1. Obtain conditions for Eq. 2.151

2. Obtain limiting distribution for p_{3n} under exponential distribution for F and G, possibly we can choose different normalization.

3. Obtain theoretical analysis of cross-validation function, i. e. , optimal order of the smoothing parameter, t_n.

4. Estimate based on

$$\begin{aligned} \tilde{F}_n(x) &= \sum_{k=0}^{N} F_n(\frac{k}{\lambda_n}) p_k(\lambda_n x) \\ N : \frac{N}{\lambda_n} &= X_{(n)} \end{aligned} \tag{4.1}$$

Appendix A

appendix

1) Pareto distribution: If X is a random variable with Pareto distribution;

$$\bar{F}(x) = Pr(X > x) = \begin{cases} (\frac{x_m}{x})^{\alpha} & \text{if } x \geq x_m \\ 1 & \text{if } \quad x < x_m \end{cases} \tag{A.1}$$

where x_m is necessary positive, minimum value of x, and a is a positive parameter.

$$f(x) = \begin{cases} (\frac{\alpha x_m^{\alpha}}{x^{\alpha}}) & \text{if } \quad x \geq x_m \\ 0 & \text{if } \quad x < x_m \end{cases} \tag{A.2}$$

Bibliography

Bagai, I. and Rao, B. P. (1995). Kernel type density estimates for positive valued random variables. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 56–67.

Bartle, R. G. (2014). *The elements of integration and Lebesgue measure*. John Wiley & Sons.

Berkson, J. and Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **47**(259), 501–515.

Chaubey, Y. P. and Sen, P. K. (1996). On smooth estimation of survival and density functions. *Statistics & Risk Modeling*, **14**(1), 1–22.

Chaubey, Y. P. and Sen, P. K. (2009). On the selection of the smoothing parameter in poisson smoothing of histogram estimator: Computational aspects. *Pakistan Journal of Statistics*.

Chaubey, Y. P., Sen, A., and Sen, P. K. (2010). A new smooth density estimator for non-negative random variables. *Journal of the Indian Statistical Association*.

Chen, M.-H., Ibrahim, J. G., and Sinha, D. (1999). A new bayesian model for survival

data with a surviving fraction. *Journal of the American Statistical Association*, **94**(447), 909–919.

De Haan, L. and Ferreira, A. (2007). *Extreme value theory: an introduction*. Springer Science & Business Media.

Feller, W. (1968). *An introduction to probability theory and its applications: volume I*, volume 3. John Wiley & Sons London-New York-Sydney-Toronto.

Finkelstein, D. M. and Wolfe, R. A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics*, pages 933–945.

Groeneboom, P. and Wellner, J. A. (1992). *Information bounds and nonparametric maximum likelihood estimation*, volume 19. Springer Science & Business Media.

Gu, Y., Sinha, D., and Banerjee, S. (2011). Analysis of cure rate survival data under proportional odds model. *Lifetime data analysis*, pages 123–134.

Hogg, Joeseph McKean, T. C. (2012). *Introduction to Mathematical Statistics*. Pearson.

Kleinbaum, D. G. and Klein, M. (2006). *Survival analysis: a self-learning text*. Springer Science & Business Media.

Maller, R. A. and Zhou, X. (1996). *Survival analysis with long-term survivors*. Wiley New York.

Marron, J. S. and Ruppert, D. (1994). Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 653–671.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series 3. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 45, pages 483–484. Cambridge Univ Press.

Resnick, S. I. (2007). Crash course i: Regular variation. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*, pages 17–38.

Sen, A. and Tan, F. (2008). Cure-rate estimation under case-1 interval censoring. *Statistical Methodology*, **5**(2), 106–118.

Severini, T. A. (2005). *Elements of distribution theory*, volume 17. Cambridge University Press.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Simonoff, J. S. (2012). *Smoothing methods in statistics*. Springer Science & Business Media.

Tsodikov, A. (1998). A proportional hazards model taking account of long-term survivors. *Biometrics*, pages 1508–1516.

Vij, K. (2014). *Textbook of Forensic Medicine & Toxicology: Principles & Practice*. Elsevier Health Sciences.