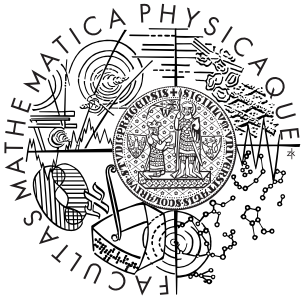
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Matej Ďurčík

Odhad momentů při intervalovém cenzorování typu I

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Arnošt Komárek Ph.D.

Studijní program: Matematika
Studijní obor: Finanční a pojistná matematika

Praha 2012

Na tomto mieste by som sa chcel poďakovať vedúcemu mojej diplomovej práce RNDr. Arnoštovi Komárekovi Ph.D. za odbornú pomoc a cenné rady pri dokončení diplomovej práce.

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V Praze dne 5.4.2012 Matej Ďurčík

Názov práce: Odhad momentů při intervalovém cenzorování typu I
Autor: Matej Ďurčík
Katedra (ústav): Katedra pravdepodobnosti a matematickej štatistiky
Vedúci bakalárskej práce: RNDr. Arnošt Komárek Ph.D.

Abstrakt: V tejto práci sa zameriame na aplikáciu rovnomerného konvolučného modelu na problém intervalového cenzorovania. Obmedzíme sa výhradne na intervalové cenzorovanie typu 1. Ukážeme ako aplikovať rovnomerný konvolučný model na odhad charakteristík pravdepodobnostných rozdelení pri intervalovom cenzorovaní typu 1. Okrem toho sa zameriame na vyvodenie limitných rozdelení odhadov strednej hodnoty a rozptylu. Potom porovnáme tieto odhady s asymptoticky efektívnymi odhadmi vychádzajúcich z neparametrických maximálne vierohodných odhadov, pomocou simulácií na určitých pravdepodobnostných rozdeleniach náhodných veličín.

Kľúčové slová: Intervalové cenzorovanie typu 1, rovnomerná konvolúcia, neparametrický maximálne vierohodný odhad, asymptoticky efektívny odhad

Title: Moments Estimation under Type I Interval Censoring
Author: Matej Ďurčík
Department: Faculty of Probability and Mathematical Statistics
Supervisor: RNDr. Arnošt Komárek Ph.D.

Abstract: In this thesis we apply the uniform deconvolution model to the interval censoring problem. We restrict ourselves only on interval censoring case 1. We show how to apply uniform deconvolution model in estimating the probability distribution characteristics in the interval censoring case 1. Moreover we derive limit distributions of the estimators of mean and variance. Then we compare these estimators to the asymptotically efficient estimators based on the nonparametric maximum likelihood estimation by simulation studies under some certain distributions of the random variables.

Keywords: Interval censoring case 1, uniform deconvolution, nonparametric maximum likelihood estimation, asymptotically efficient estimator

Obsah

Chapter 1

Introduction

Missing and incomplete data problems have been an important field of research in mathematical statistics during the past decades. In particular, a lot of theory has been developed for the analysis of the interval censored data or in other words, the interval censoring problem. Here we will consider the uniform deconvolution problem and its application to the interval censoring problem. In the following sections of this chapter we will mathematically describe these two problems separately, as well as the relation between them, and we will mention several examples of an application of this particular problem in real world. We will also show how to transform interval censored data to uniform deconvolution data. In Chapter 2. we will focus only on the uniform deconvolution model. Here we will derive estimators of the mean and variance, and their limit distributions. At the end of this chapter we will give some simulation results for different distributions of the random variables to show how the theoretical values, from our limit theorems, are close to the simulated ones. In Chapter 3. we will transform the uniform deconvolution model into the interval censoring problem and derive related estimators of the mean and variance. Here we will derive, also as in the previous chapter, the limit distributions of these estimators. This is crucial for the thesis in order to investigate how these relatively simple estimators, based on the random variable moments estimation, behave with respect to the asymptotically efficient estimators based on the nonparametric maximum likelihood estimation, i.e. the NPMLE, of the distribution function. The NPMLE method will be described in Chapter 4. In Chapter 5. we will present and concentrate on the simulation results from both methods in order to compare the asymptotic efficiency of the related estimators. Chapter 6. will summarize the results gained in Chapter 5.

Let us first give a general description of a statistical estimation problem. The random quantity X takes values in the measurable space $(\chi, \mathcal{A})$ and its distribution F is unknown. This distribution belongs to the known class $\mathcal{F}$. The map $\nu : \mathcal{F} \to \mathbb{R}^m$ defines a Euclidean parameter. We will study estimation of the unknown value of $\nu(F)$. In particular we will focus on $\nu(F) = \int t dF(t)$, the mean of F, and $\nu(F) = \int (t - \int s dF(s))^2 dF(t)$, the variance of F, in two models, interval censoring case I. and uniform deconvolution. Any measurable map $t : \chi \to \mathbb{R}^m$ defines an estimator $T = t(X)$ of $\nu(F)$.

1.1 Interval censoring case 1.

Let us consider the interval censoring problem case 1, or in other words the current status data problem. In this estimation problem one observes at *i.i.d.* random time instants $T_1, \ldots, T_n$ whether unobservable variables of interest $X_1, \ldots, X_n$ are smaller or larger than the corresponding T_i. Writing $\Delta_i = 1_{[X_i \leq T_i]}$, the observations are denoted by $(T_1, \Delta_1), \ldots, (T_n, \Delta_n)$. The exact value of X_i is not measured. Based on these data, under some regularity conditions, one can estimate the distribution of the X_i and its mean and the variance, see Chapter 3 and 4.

Before we start with the mathematical computations let us first give some examples of a possible application of such a problem.

i) *The maximum price of a good that a consumer is willing to pay:* Let us consider the issue of looking for a maximum price, which we can represent by the variable X. Assume that one can observe a supply price of a certain specified good, which would be in our case represented by the variable T. It is important to emphasize that in this special case we consider just one observation, our supply price T, so we can omit the subscript i. It implies that we can observe, whether the variable X representing the maximum price is smaller or bigger than the supply price T. In other words we can observe whether the consumer is willing to pay more or less for a certain good than the supply price. The exact maximum price X is unknown and we are loooking for the distribution of this price and its characteristics.

ii) *The time of infection of a certain disease that causes antibodies at infection:* Let us represent the times when the persons blood is checked with variables T_i and the exact time of the infection with variables X_i. During a check of the persons blood, antibodies are present or not. This means that the person has been infected before the time of the check ($X_i \leq T_i$) or after ($X_i > T_i$). In both cases the exact time of infection X_i is unknown.

iii) *The time of arrival of a new mail to the mailbox:* Let us represent the times when the mailbox is checked by the owner with variables T_i and the exact time of arrival of a new mail with variables X_i. During a check of the mailbox, new mail is already inside or not. This means that a new mail was already received ($X_i \leq T_i$) or not ($X_i > T_i$). In both cases the exact time X_i of arrival of a new mail is unknown.

The probabilities for Δ_i, given the value of T_i, are given by

$$\begin{aligned} P(\Delta_i = 0 | T_i = t_i) &= 1 - F(t_i), \\ P(\Delta_i = 1 | T_i = t_i) &= F(t_i). \end{aligned} \tag{1.1}$$

The following transformation of the points T_i plays a crucial role. Let

$$V_i = \begin{cases} T_i + 1 & \text{if } \Delta_i = 0, \\ T_i & \text{if } \Delta_i = 1. \end{cases} \tag{1.2}$$

Lemma 1.1.1. *Assume that the distribution of X_i is concentrated on $[0,1]$. Let us denote the density of V_i by g and the density of T_i by q. If T_i is supported on $[0,1]$, then the density g of V_i is given by*

$$g(v) = \bar{q}(v)\left(F(v) - F(v-1)\right), \tag{1.3}$$

with the function $\bar{q}$ defined by

$$\bar{q}(v) = q(v) + q(v-1). \tag{1.4}$$

Proof. We omit the subscript i. For $v \in [0,1]$ we have

$$\begin{aligned}
&P(V \le v)\\
&= P(V \le v, \Delta = 0) + P(V \le v, \Delta = 1)\\
&= P(V \le v, \Delta = 1)\\
&= P(T \le v, \Delta = 1)\\
&= \int_0^1 P(T \le v, \Delta = 1 | T = t) q(t) dt\\
&= \int_0^v P(\Delta = 1 | T = t) q(t) dt\\
&= \int_0^v F(t) q(t) dt
\end{aligned}$$

By the support restrictions on the distribution induced by F and on q this confirms (1.3) for $v \in [0,1]$.

Let us also check the claim on the interval $[1,2]$. For $v \in [0,1]$ we have

$$\begin{aligned}
&P(1 \le V \le 1+v)\\
&= P(1 \le V \le 1+v, \Delta = 0) + P(1 \le V \le 1+v, \Delta = 1)\\
&= P(1 \le V \le 1+v, \Delta = 0)\\
&= \int_0^1 P(1 \le V \le 1+v, \Delta = 0 | T = t) q(t) dt\\
&= \int_0^1 P(T \le v, \Delta = 0 | T = t) q(t) dt\\
&= \int_0^v P(\Delta = 0 | T = t) q(t) dt\\
&= \int_0^v (1 - F(t)) q(t) dt\\
&= \int_1^{1+v} (1 - F(t-1)) q(t-1) dt.
\end{aligned}$$

Hence the density of V at $1+v$ equals $(1 - F(t-1))q(t-1)$, which confirms (1.3) for t in $[1,2]$. □

Note that V takes values in $[0,2)$. On the interval $[0,1)$ the function $\bar{q}(v)$ equals $q(v)$ and on the interval $[1,2)$ it equals $q(v-1)$, because we consider the density of the Uniform distribution on $[0,1]$ as an indicator $1_{[0,1)}(v)$.

1.2 Deconvolution model

Let us now consider the general deconvolution model. Let $V_1, \ldots, V_n$ be *i.i.d.* observations, where $V_i = X_i + U_i$ and X_i and U_i are independent. Assume that the unobservable X_i have distribution function F and density f. Also assume that the unobservable random variables U_i have a known density k. Note that the density g of the V_i is equal to the convolution of f and k, so $g = k * f$, where $*$ denotes convolution. So we have

$$g(v) = \int_{-\infty}^{\infty} k(v-u)f(u)du. \tag{1.5}$$

The deconvolution problem is the problem of estimating f or F, or its moments, from the observations V_i. In Chapter 2. we will restrict ourselves to *uniform deconvolution* where we require the distribution of the U_i to be uniformly distributed on $[0, 1]$.

1.3 Relation between the uniform deconvolution model and interval censoring case 1.

In the uniform deconvolution problem the error U is Uniform$[0, 1]$ distributed. So in this particular deconvolution problem we assume to have *i.i.d.* observations from the density

$$g(v) = \int_{-\infty}^{\infty} I_{[0,1)}(v-u)f(u)du = \int_{v-1}^{v} f(u)du = F(v) - F(v-1). \tag{1.6}$$

Note that the density of V in the uniform deconvolution model is exactly the same as the density of V in the interval censoring case 1., see (1.3), when the function $\bar{q} \equiv 1$ on $[0, 2]$, i.e. when the observation times T_i are uniformly distributed in $[0, 1]$. This shows that in this case the transformation (1.2) transforms interval censored data to uniform deconvolution data.

Chapter 2

Uniform deconvolution model

We consider random variables $V_i = X_i + U_i$ in the uniform deconvolution model and we denote $\mathrm{E}\, X_i = \mu$ and $\mathrm{Var}\, X_i = \sigma^2$, for all i. Note that, $\mathrm{E}\, V_i = \mu + \frac{1}{2}$, so apparently we have

$$\int_{-\infty}^{\infty} v(F(v) - F(v-1))dv = \mu + \frac{1}{2}. \tag{2.1}$$

The previous equality can be shown directly by

$$\begin{aligned}
\int_{-\infty}^{\infty} v(F(v) - F(v-1))dv &= \int_{-\infty}^{\infty} v \left(\int_{v-1}^{v} f(t)dt \right) dv \\
&= \int_{-\infty}^{\infty} \left(\int_{t}^{t+1} vf(t)dv \right) dt \\
&= \int_{-\infty}^{\infty} \frac{1}{2} \left((t+1)^2 - t^2 \right) f(t)dt \\
&= \int_{-\infty}^{\infty} \left(t + \frac{1}{2} \right) f(t)dt \\
&= \mu + \frac{1}{2}.
\end{aligned}$$

Also note that $\mathrm{E}\, V_i^2 = \mathrm{E}\, X_i^2 + 2\,\mathrm{E}(X_i U_i) + \mathrm{E}\, U_i^2 = \sigma^2 + \mu^2 + \mu + \frac{1}{3}$, so we have

$$\int_{-\infty}^{\infty} v^2(F(v) - F(v-1))dv = \sigma^2 + \mu^2 + \mu + \frac{1}{3}. \tag{2.2}$$

To show that the equation (2.2) is true we will use the same approach as we used by showing (2.1). So we can write

$$\begin{aligned}\int_{-\infty}^{\infty} v^2(F(v) - F(v-1))dv &= \int_{-\infty}^{\infty} v^2 \left(\int_{v-1}^{v} f(t)dt \right) dv \\ &= \int_{-\infty}^{\infty} \left(\int_{t}^{t+1} v^2 f(t)dv \right) dt \\ &= \int_{-\infty}^{\infty} \frac{1}{3} \left((t+1)^3 - t^3 \right) f(t)dt \\ &= \int_{-\infty}^{\infty} \left(t^2 + t + \frac{1}{3} \right) f(t)dt \\ &= \sigma^2 + \mu^2 + \mu + \frac{1}{3}.\end{aligned}$$

Note that $\sigma^2 + \mu^2$ is the second moment of the random variable X i.e., $\mathrm{E}\, X^2$. To show the previous equalities we used Fubini's theorem, see W. Rudin (1986).

2.1 Estimators of mean and variance

In this section we will derive estimators of the mean and variance in the uniform deconvolution model. Recall

$$\mathrm{E}\, V_i = \mathrm{E}\, X_i + \mathrm{E}\, U_i = \mu + \frac{1}{2}. \tag{2.3}$$

Now we can estimate $\mu_V = \mathrm{E}\, V_i$ by the sample mean $\bar{V}_n$ and we get

$$\bar{V}_n \approx \mu + \frac{1}{2}. \tag{2.4}$$

So our estimator for the mean μ in the uniform deconvolution model, denoted by $M_{X,n}$ will be

$$M_{X,n} = \frac{1}{n} \sum_{i=1}^{n} V_i - \frac{1}{2}. \tag{2.5}$$

Note that $M_{X,n}$ is an unbiased estimator of μ. To derive an estimator for the variance we proceed analogically. Note that because of the independence of X_i and U_i we can write

$$\mathrm{Var}\, V_i = \mathrm{Var}\, X_i + \mathrm{Var}\, U_i = \sigma^2 + \frac{1}{12}. \tag{2.6}$$

We can estimate $\sigma_V^2 = \mathrm{Var}\, V_i$ by $S_{V,n}^2$ and we get

$$S_{V,n}^2 \approx \sigma^2 + \frac{1}{12}, \tag{2.7}$$

where

$$S_{V,n}^2 = \frac{1}{n} \sum_{i=1}^{n} (V_i - \bar{V}_n)^2 = \frac{1}{n} \sum_{i=1}^{n} V_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} V_i \right)^2. \tag{2.8}$$

So our estimator for the variance σ^2 in the uniform deconvolution model, denoted by $S^2_{X,n}$, will be

$$S^2_{X,n} = S^2_{V,n} - \frac{1}{12} = \frac{1}{n}\sum_{i=1}^{n}(V_i - \bar{V}_n)^2 - \frac{1}{12}. \tag{2.9}$$

This estimator would be unbiased if we had used $\frac{1}{n-1}$ instead of $\frac{1}{n}$, but for mathematical convenience we have chosen $\frac{1}{n}$.

2.2 Limit distribution of the estimators of mean and variance

In this section we derive the limit distributions of the estimators $M_{X,n}$ and $S^2_{X,n}$ in the following two theorems.

Theorem 2.2.1. *Assume that* $\mathrm{E}\,X^2 < \infty$. *As* n *approaches infinity, the random variable* $\sqrt{n}(M_{X,n} - \mu)$, *where* $M_{X,n}$ *is defined by* (2.5), *converges to a normal* $\mathcal{N}(0, \sigma_V^2)$ *distribution,*

$$\sqrt{n}\,(M_{X,n} - \mu) \xrightarrow{D} \mathcal{N}(0, \sigma_V^2) \tag{2.10}$$

with $\sigma_V^2 = \sigma^2 + \frac{1}{12}$.

Proof. To prove (2.10) we must show that the requirements of the Central Limit Theorem (CLT), see Serfling (1980) are satisfied. First we will derive the mean and the variance of the estimator $M_{X,n}$. We have by (2.3) and (2.6)

$$\mathrm{E}\,M_{X,n} = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} V_i - \frac{1}{2}\right) = \mu$$

$$\mathrm{Var}(M_{X,n}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} V_i\right) = \frac{1}{n}\,\mathrm{Var}(V_i) = \frac{\sigma_V^2}{n}$$

Now by applying the CLT to the estimator $M_{X,n}$ we get the required result. □

The limit distribution of $S^2_{X,n}$ is given by the following theorem.

Theorem 2.2.2. *Assume that* $\mathrm{E}\,X^4 < \infty$. *As* n *approaches infinity, the random variable* $\sqrt{n}(S^2_{X,n} - \sigma^2)$, *where* $S^2_{X,n}$ *is defined by* (2.9), *converges to a normal* $\mathcal{N}(0, \sigma_{1s}^2)$ *distribution,*

$$\sqrt{n}\left(S^2_{X,n} - \sigma^2\right) \xrightarrow{D} \mathcal{N}(0, \sigma_{1s}^2) \tag{2.11}$$

with $\sigma_{1s}^2 = \mathrm{Var}\left((V - \mu_V)^2\right)$.

Proof. To prove (2.11) we will apply CLT to $Y_i = (V_i - \mu_V)^2$. We have

$$\mathrm{E}\,\bar{Y}_n = \mathrm{E}\,Y_i = \mathrm{Var}(V_i) = \sigma_V^2 = \sigma^2 + \frac{1}{12},$$

$$\mathrm{Var}(\bar{Y}_n) = \frac{1}{n}\,\mathrm{Var}(Y_1) = \frac{1}{n}\,\mathrm{Var}((V_1 - \mu_V)^2) = \frac{\sigma_{1s}^2}{n}.$$

The CLT says, when n approaches infinity, that we have

$$\sqrt{n}\left(\frac{\bar{Y}_n - \sigma^2 - 1/12}{\sigma_{1s}}\right) \xrightarrow{D} \mathcal{N}(0,1).$$

Note that $S^2_{V,n} \neq \bar{Y}_n$, so we must show that the following expression is true for large n

$$\sqrt{n}\left(\frac{\bar{Y}_n - \sigma_V^2}{\sigma_{1s}}\right) = \sqrt{n}\left(\frac{S^2_{V,n} - \sigma^2 - 1/12}{\sigma_{1s}}\right) + o_p(1) = \sqrt{n}\left(\frac{S^2_{X,n} - \sigma^2}{\sigma_{1s}}\right) + o_p(1), \qquad (2.12)$$

and we will proceed as follows. Note that we have $S^2_{V,n} = \frac{1}{n}\sum_{i=1}^n (V_i - \bar{V}_n)^2$, thus we can write

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n (V_i - \mu_V)^2 &= \frac{1}{n}\sum_{i=1}^n (V_i - \bar{V}_n + \bar{V}_n - \mu_V)^2 \\
&= \frac{1}{n}\sum_{i=1}^n (V_i - \bar{V}_n)^2 + 2\frac{1}{n}\sum_{i=1}^n (V_i - \bar{V}_n)(\bar{V}_n - \mu_V) + \frac{1}{n}\sum_{i=1}^n (\bar{V}_n - \mu_V)^2 \\
&= S^2_{V,n} + (\bar{V}_n - \mu_V)^2,
\end{aligned}$$

since

$$2\frac{1}{n}\sum_{i=1}^n (V_i - \bar{V}_n)(\bar{V}_n - \mu_V) = \frac{2}{n}\left(\bar{V}_n - \mu_V\right)\underbrace{\sum_{i=1}^n \left(V_i - \bar{V}_n\right)}_{=0} = 0.$$

So we have

$$S^2_{V,n} = \frac{1}{n}\sum_{i=1}^n (V_i - \mu_V)^2 - (\bar{V}_n - \mu_V)^2$$

and we can rewrite the previous expression to the form

$$\begin{aligned}
\sqrt{n}\left(S^2_{V,n} - \sigma_V^2\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \left((V_i - \mu_V)^2 - \sigma_V^2\right) - \sqrt{n}(\bar{V}_n - \mu_V)^2 \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(Y_i - \sigma_V^2\right) - \sqrt{n}(\bar{V}_n - \mu_V)^2 \\
&= \sqrt{n}\left(\bar{Y}_n - \sigma_V^2\right) - \sqrt{n}(\bar{V}_n - \mu_V)^2.
\end{aligned}$$

Now

$$\sqrt{n}(\bar{V}_n - \mu_V)^2 = \sqrt{n}(\bar{V}_n - \mu_V)(\bar{V}_n - \mu_V) \xrightarrow{P} 0.$$

In the previous expression we used Slutsky's Theorem, see Serfling(1980), and

$$\begin{aligned}
\sqrt{n}(\bar{V}_n - \mu_V) &\xrightarrow{D} \mathcal{N}(0, \sigma_V^2), \\
(\bar{V}_n - \mu_V) &\xrightarrow{P} 0.
\end{aligned}$$

This proves the first equality in (2.12). The second equality follows by definition of $S^2_{X,n}$. $\square$

Remark 2.2.1. *Note that σ_{1s}^2 is related to the kurtosis $\gamma(V)$ of V. Recall that the kurtosis is defined as*

$$\gamma(V) = \frac{\mathrm{E}((V-\mu_V)^4)}{\sigma_V^4} - 3 = \frac{\mathrm{Var}((V-\mu_V)^2)}{\sigma_V^4} - 2.$$

We find that $\sigma_{1s}^2 = (\gamma(V)+2)\sigma_V^4$.

Example 2.2.1. *For $V_1, \ldots, V_n$ a sample from the normal distribution $\mathcal{N}(\mu_V, \sigma_V^2)$ there are no problems. In this case $\gamma(V) = 0$, thus we can write $\sigma_{1s}^2 = 2\sigma_V^4$, see previous Remark. As n approaches infinity, we have*

$$\sqrt{n}\left(S_{X,n}^2 - \sigma^2\right) \xrightarrow{D} \mathcal{N}(0, 2\sigma_V^4), \tag{2.13}$$

see (2.11).

2.3 Simulation of the estimators

In this Section we generate a sequence of random variables $X_i, i = 1, \ldots, n$, from different distributions and construct $V_i = X_i + U_i$, where the random variables U_i are from the Uniform distribution on $[0, 1]$, and are independent of X_i for all $i = 1, \ldots, n$. Then we compute our estimates of the mean (2.5) and the variance (2.9), and compare them to our limit theorems (2.10) and (2.11), for different values of n.

i) The distribution of random variable X is Uniform on $[0, 1]$.

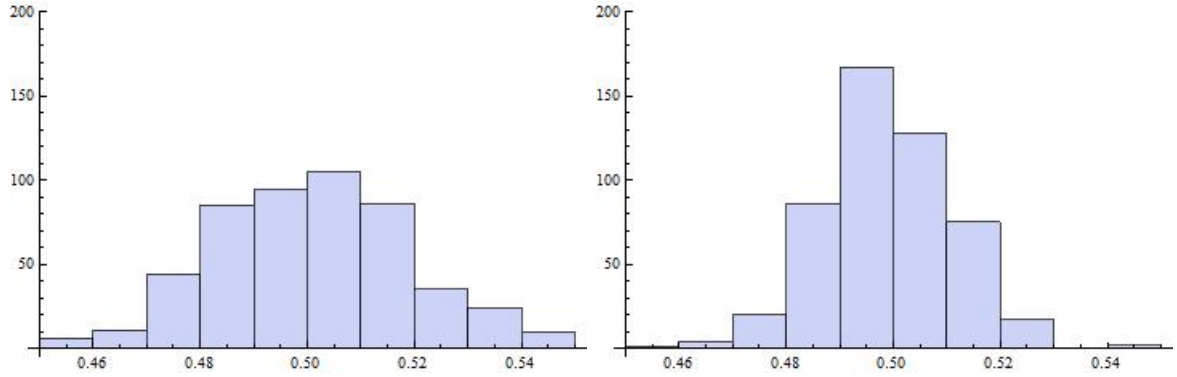


Figure 2.1: Histogram of 500 simulations of the estimator $M_{X,n}$, where the random variable X has an Uniform distribution on $[0, 1]$. Left: For $n = 500$ observations, the mean of the $M_{X,n}$ samples is 0.5009 and the sample variance is equal to 3.29599×10^{-4}. Right: For $n = 1000$ observations, the mean of the $M_{X,n}$ samples is 0.49905 and the sample variance is equal to 1.49041×10^{-4}.

Now we will check the result of our Theorem 2.2.1., where $M_{X,n}$ has an $\mathcal{AN}\left(\mu, \frac{\sigma_V^2}{n}\right)$ distribution. When X has an Uniform distribution on $[0, 1]$, then $\mu = \frac{1}{2}$ and the variance $\sigma^2 = \frac{1}{12} = 8.33 \times 10^{-2}$, and we can write in general that $M_{X,n}$ has an $\mathcal{AN}\left(\frac{1}{2}, \frac{1}{6n}\right)$ distribution.

For $n = 500$ observations, $M_{X,500}$ approximately has a $\mathcal{N}\left(0.5, 3.33 \times 10^{-4}\right)$ distribution and for $n = 1000$ observations, $M_{X,1000}$ approximately has a $\mathcal{N}\left(0.5, 1.67 \times 10^{-4}\right)$ distribution.

Simulations of the estimator $M_{X,n}$ gave us values, which are close to the real values. The simulation results support the theorem.

Let us check also Theorem 2.2.2., where $S^2_{X,n}$ has an $\mathcal{AN}\left(\sigma^2, \frac{\sigma^2_{1s}}{n}\right)$ distribution. When X has an Uniform distribution on $[0, 1]$, then we can write in general that $S^2_{X,n}$ has an $\mathcal{AN}\left(\frac{1}{12}, \frac{7}{180n}\right)$ distribution.

For $n = 500$ observations, $S^2_{X,500}$ approximately has a $\mathcal{N}\left(8.33 \times 10^{-2}, 7.78 \times 10^{-5}\right)$ distribution and for $n = 1000$ observations, $S^2_{X,1000}$ approximately has a $\mathcal{N}\left(8.33 \times 10^{-2}, 3.89 \times 10^{-5}\right)$ distribution.

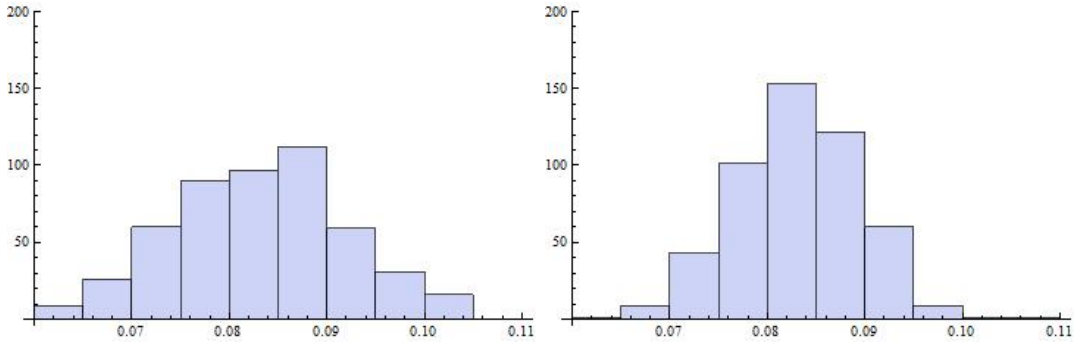


Figure 2.2: Histogram of 500 simulations of the estimator $S^2_{X,n}$, where the random variable X has an Uniform distribution on $[0, 1]$. Left: For $n = 500$ observations, the mean of the $S^2_{X,n}$ samples is 8.31318×10^{-2} and the sample variance is equal to 7.7228×10^{-5}. Right: For $n = 1000$ observations, the mean of the $S^2_{X,n}$ samples is 8.31943×10^{-2} and the sample variance is equal to 4.07688×10^{-5}.

Simulations of the estimator $S^2_{X,n}$ gave us values, which confirm our Theorem 2.2.2..

ii) The distribution of random variable X is standard normal.

Now we will check the result of our Theorem 2.2.1., where $M_{X,n}$ has an $\mathcal{AN}\left(\mu, \frac{\sigma^2_V}{n}\right)$ distribution. When X has a standard normal distribution, then $\mu = 0$ and the variance $\sigma^2 = 1$. Here we can write in general that $M_{X,n}$ has an $\mathcal{AN}\left(0, \frac{13}{12n}\right)$ distribution.

For $n = 500$ observations, $M_{X,500}$ approximately has a $\mathcal{N}\left(0, 2.17 \times 10^{-3}\right)$ distribution and for $n = 1000$ observations, $M_{X,1000}$ approximately has a $\mathcal{N}\left(0, 1.08 \times 10^{-3}\right)$ distribution.

Simulations of the estimator $M_{X,n}$ gave us values, which are close to the real values. The simulation results support the theorem.

Let us check also Theorem 2.2.2., where $S^2_{X,n}$ has an $\mathcal{AN}\left(\sigma^2, \frac{\sigma^2_{1s}}{n}\right)$ distribution. When X has a standard normal distribution, then we can write in general that $S^2_{X,n}$ has an $\mathcal{AN}\left(1, \frac{421}{180n}\right)$ distribution.

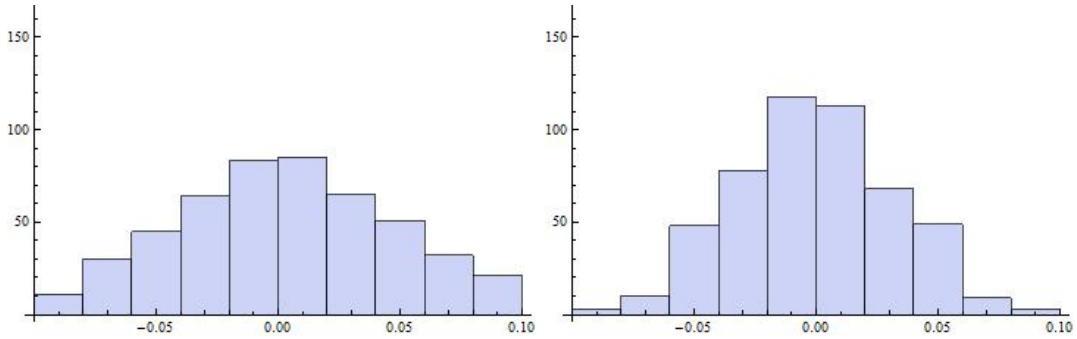


Figure 2.3: Histogram of 500 simulations of the estimator $M_{X,n}$, where the random variable X has a standard normal distribution. Left: For $n = 500$ observations, the mean of the $M_{X,n}$ samples is 2.56367×10^{-2} and the sample variance is equal to 2.22689×10^{-3}. Right: For $n = 1000$ observations, the mean of the $M_{X,n}$ samples is -9.06578×10^{-4} and the sample variance is equal to 1.04319×10^{-3}.

For $n = 500$ observations, $S^2_{X,500}$ approximately has a $\mathcal{N}\left(1, 4.68 \times 10^{-3}\right)$ distribution and for $n = 1000$ observations, $S^2_{X,1000}$ approximately has a $\mathcal{N}\left(1, 2.34 \times 10^{-3}\right)$ distribution.

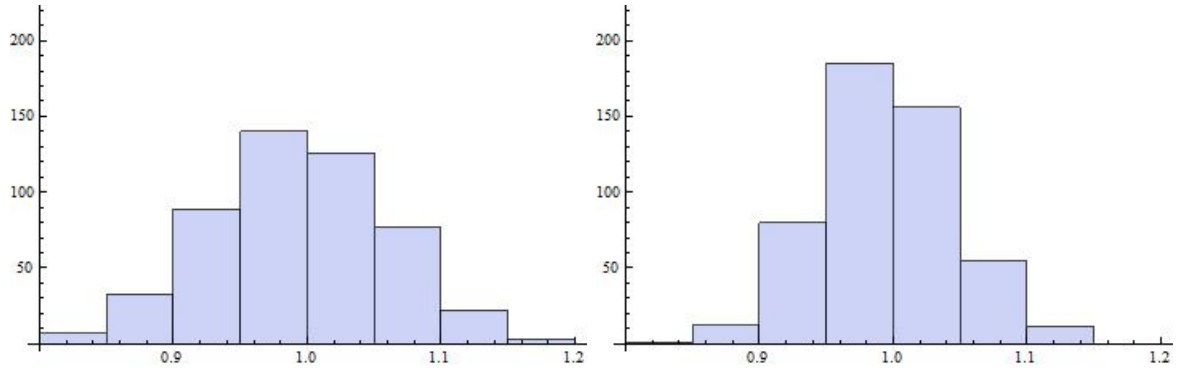


Figure 2.4: Histogram of 500 simulations of the estimator $S^2_{X,n}$, where the random variable X has a standard normal distribution. Left: For $n = 500$ observations, the mean of the $S^2_{X,n}$ samples is 0.99509 and the sample variance is equal to 4.51769×10^{-3}. Right: For $n = 1000$ observations, the mean of the $S^2_{X,n}$ samples is 0.994162 and the sample variance is equal to 2.47613×10^{-3}.

Simulations of the estimator $S^2_{X,n}$ gave us values, which confirm our Theorem 2.2.2..

iii) The distribution of random variable X is squared Uniform on $[0, 1]$.
Now we will check the result of our Theorem 2.2.1., where $M_{X,n}$ has an $\mathcal{AN}\left(\mu, \frac{\sigma_V^2}{n}\right)$ distribution. X has a squared Uniform distribution on $[0, 1]$, so here $X = U^2$. Then $\mu = \frac{1}{3}$ and the variance $\sigma^2 = \frac{4}{45} = 8.89 \times 10^{-2}$. We can write in general that $M_{X,n}$ has an $\mathcal{AN}\left(\frac{1}{3}, \frac{31}{180n}\right)$ distribution.
For $n = 500$ observations, $M_{X,500}$ approximately has a $\mathcal{N}\left(0.33, 3.44 \times 10^{-4}\right)$ distribu-

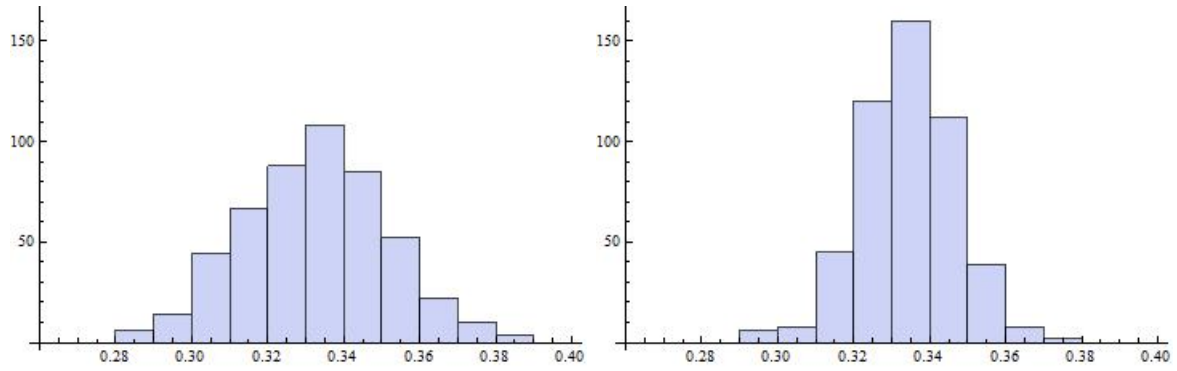


Figure 2.5: Histogram of 500 simulations of the estimator $M_{X,n}$, where the random variable X has a squared Uniform distribution on $[0, 1]$. Left: For $n = 500$ observations, the mean of the $M_{X,n}$ samples is 0.332464 and the sample variance is equal to 3.59296×10^{-4}. Right: For $n = 1000$ observations, the mean of the $M_{X,n}$ samples is 0.33408 and the sample variance is equal to 1.58466×10^{-4}.

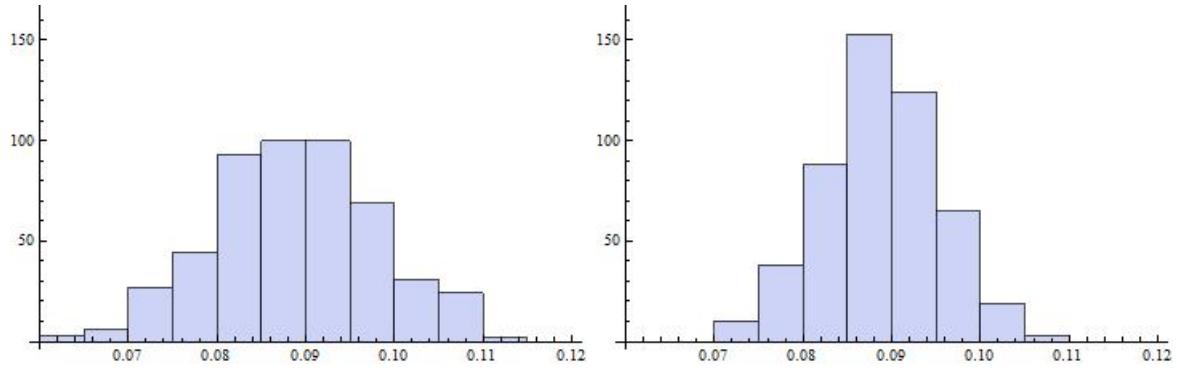


Figure 2.6: Histogram of 500 simulations of the estimator $S^2_{X,n}$, where the random variable X has a squared Uniform distribution on $[0, 1]$. Left: For $n = 500$ observations, the mean of the $S^2_{X,n}$ samples is 8.89417×10^{-2} and the sample variance is equal to 8.91079×10^{-5}. Right: For $n = 1000$ observations, the mean of the $S^2_{X,n}$ samples is 8.87355×10^{-2} and the sample variance is equal to 4.28718×10^{-5}.

tion and for $n = 1000$ observations, $M_{X,1000}$ approximately has a $\mathcal{N}\left(0.33, 1.72 \times 10^{-4}\right)$ distribution.

Simulations of the estimator $M_{X,n}$ gave us values, which are close to the real values. The simulation results support the theorem.

Let us check also Theorem 2.2.2., where $S^2_{X,n}$ has an $\mathcal{AN}\left(\sigma^2, \frac{\sigma^2_{1s}}{n}\right)$ distribution. When X has a squared Uniform distribution on $[0, 1]$, then we can write in general that $S^2_{X,n}$ has an $\mathcal{AN}\left(\frac{4}{45}, \frac{2507}{56700n}\right)$ distribution.

For $n = 500$ observations, $S^2_{X,500}$ approximately has a $\mathcal{N}\left(8.89 \times 10^{-2}, 8.84 \times 10^{-5}\right)$ distribution and for $n = 1000$ observations, $S^2_{X,1000}$ approximately has a $\mathcal{N}\left(8.89 \times 10^{-2}, 4.42 \times 10^{-5}\right)$ distribution.

Simulations of the estimator $S^2_{X,n}$ gave us values, which confirm our Theorem 2.2.2..

iv) The distribution of random variable X is square root Uniform on $[0, 1]$.

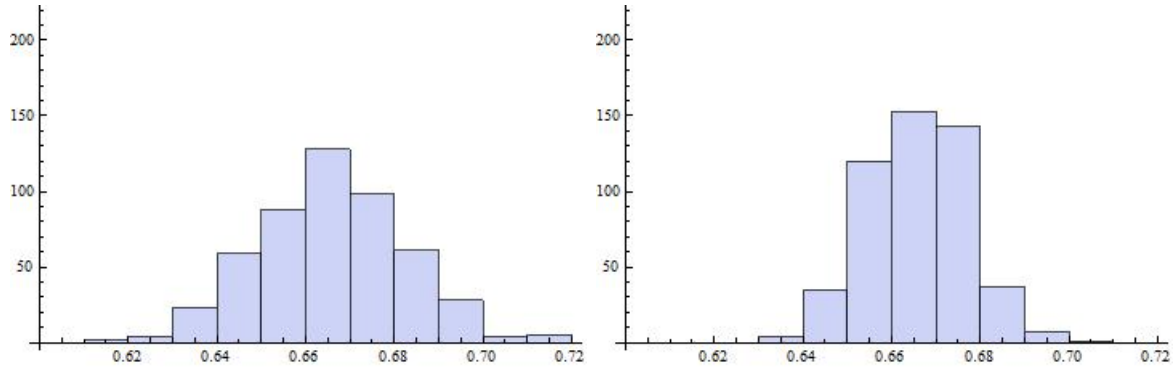


Figure 2.7: Histogram of 500 simulations of the estimator $M_{X,n}$, where the random variable X has a square root Uniform distribution on $[0, 1]$. Left: For $n = 500$ observations, the mean of the $M_{X,n}$ samples is 0.666017 and the sample variance is equal to 2.75786×10^{-4}. Right: For $n = 1000$ observations, the mean of the $M_{X,n}$ samples is 0.665877 and the sample variance is equal to 1.19894×10^{-4}.

Now we will check the result of our Theorem 2.2.1., where $M_{X,n}$ has an $\mathcal{AN}\left(\mu, \frac{\sigma^2_V}{n}\right)$ distribution. When X has a square root Uniform distribution on $[0, 1]$, then $\mu = \frac{2}{3}$ and the variance $\sigma^2 = \frac{1}{18} = 5.56 \times 10^{-2}$, and we can write in general that $M_{X,n}$ has an $\mathcal{AN}\left(\frac{2}{3}, \frac{5}{36n}\right)$ distribution.

For $n = 500$ observations, $M_{X,500}$ approximately has a $\mathcal{N}\left(0.67, 2.78 \times 10^{-4}\right)$ distribution and for $n = 1000$ observations, $M_{X,1000}$ approximately has a $\mathcal{N}\left(0.67, 1.39 \times 10^{-4}\right)$ distribution.

Simulations of the estimator $M_{X,n}$ gave us values, which are close to the real values. The simulation results support the theorem.

Let us check also Theorem 2.2.2., where $S^2_{X,n}$ has an $\mathcal{AN}\left(\sigma^2, \frac{\sigma^2_{1s}}{n}\right)$ distribution. When X has a square root Uniform distribution on $[0, 1]$, then we can write in general that $S^2_{X,n}$ has an $\mathcal{AN}\left(\frac{1}{18}, \frac{23}{810n}\right)$ distribution.

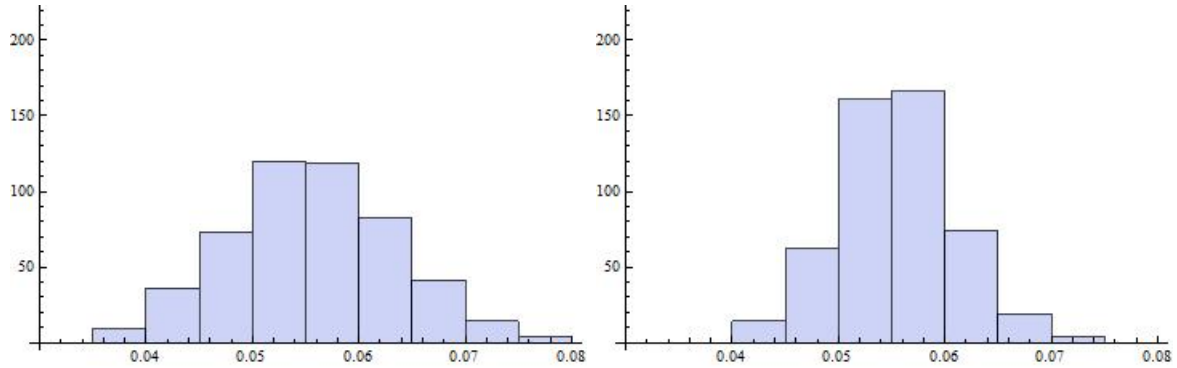


Figure 2.8: Histogram of 500 simulations of the estimator $S^2_{X,n}$, where the random variable X has a square root Uniform distribution on $[0, 1]$. Left: For $n = 500$ observations, the mean of the $S^2_{X,n}$ samples is 5.56057×10^{-2} and the sample variance is equal to 6.21251×10^{-5}. Right: For $n = 1000$ observations, the mean of the $S^2_{X,n}$ samples is 5.54688×10^{-2} and the sample variance is equal to 2.88202×10^{-5}.

For $n = 500$ observations, $S^2_{X,500}$ approximately has a $\mathcal{N}\left(5.56 \times 10^{-2}, 5.68 \times 10^{-5}\right)$ distribution and for $n = 1000$ observations, $S^2_{X,1000}$ approximately has a $\mathcal{N}\left(5.56 \times 10^{-2}, 2.84 \times 10^{-5}\right)$ distribution. Simulations of the estimator $S^2_{X,n}$ gave us values, which confirm our Theorem 2.2.2..

Chapter 3

Interval censoring case 1.

3.1 Estimators of mean and variance with known distribution of the observation times

In this section we will discuss the case when the density q is known. We assume that the densities f and q are concentrated on $[0, 1]$. First we derive the estimator for the mean in interval censoring case 1. Because of the equations (1.3) and (2.1) we can write

$$\mathrm{E}\left(\frac{V_i}{\bar{q}(V_i)}\right) = \int_{-\infty}^{\infty} \frac{v}{\bar{q}(v)} g(v)dv = \int_{-\infty}^{\infty} v(F(v) - F(v-1))dv = \mu_V = \mu + \frac{1}{2} \tag{3.1}$$

Then we again use the sample mean to estimate $\mathrm{E}\left(\frac{V_i}{\bar{q}(V_i)}\right)$ and we get our estimator of the mean μ. We get

$$M_{X,n} = \frac{1}{n}\sum_{i=1}^{n} \frac{V_i}{\bar{q}(V_i)} - \frac{1}{2}. \tag{3.2}$$

To simplify the notation we denote

$$Y_i := \frac{V_i}{\bar{q}(V_i)}. \tag{3.3}$$

To derive the estimator for the variance, recall (2.2). First we derive the second moment of the random variable $\frac{V_i}{\sqrt{\bar{q}(V_i)}}$, which will be useful in the next step of the computation. We have

$$\begin{aligned} \mathrm{E}\left(\frac{V_i^2}{\bar{q}(V_i)}\right) &= \int_{-\infty}^{\infty} \frac{v^2}{\bar{q}(v)} g(v)dv \\ &= \int_{-\infty}^{\infty} v^2(F(v) - F(v-1))dv \\ &= \sigma^2 + \mu^2 + \mu + \frac{1}{3} \end{aligned} \tag{3.4}$$

As an estimator for the second moment of the random variable $\frac{V_i}{\sqrt{\bar{q}(V_i)}}$, we use

$$\frac{1}{n}\sum_{i=1}^{n} \left(\frac{V_i^2}{\bar{q}(V_i)}\right). \tag{3.5}$$

Then we can easily show that the estimator for the variance σ^2 will be

$$\begin{aligned} S^2_{X,n} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{V_i^2}{\bar{q}(V_i)}\right) - M^2_{X,n} - M_{X,n} - \frac{1}{3} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{V_i^2}{\bar{q}(V_i)}\right) - \left(\bar{Y}_n - \frac{1}{2}\right)^2 - \bar{Y}_n + \frac{1}{2} - \frac{1}{3} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{V_i^2}{\bar{q}(V_i)}\right) - \bar{Y}_n^2 - \frac{1}{12} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{V_i^2}{\bar{q}(V_i)}\right) - \left(\frac{1}{n}\sum_{i=1}^{n}\frac{V_i}{\bar{q}(V_i)}\right)^2 - \frac{1}{12}. \end{aligned} \tag{3.6}$$

We can see that if $\bar{q}(.) \equiv 1$, the estimators for the mean and the variance are exactly the same as we derived them in 2nd chapter, see (2.5) and (2.9). We can take it as an indication that we proceed correctly.

3.1.1 Limit distribution of the estimators

Before we start with the theorem about the asymptotic distribution of the estimator $M_{X,n}$, let us first derive one assumption which is crucial for the following theorem in order for the asymptotic variance to be well defined. Note

$$\begin{aligned} &\left|\int_0^2 \frac{v^2}{\bar{q}(v)}(F(v) - F(v-1))dv\right| \\ &\leq \left|\int_0^1 \frac{v^2}{q(v)}F(v)dv\right| + \left|\int_1^2 \frac{v^2}{q(v-1)}(1 - F(v-1))dv\right| \\ &\leq \int_0^1 \frac{1}{q(v)}dv + \int_1^2 \frac{4}{q(v-1)}dv \\ &= 5\int_0^1 \frac{1}{q(v)}dv. \end{aligned} \tag{3.7}$$

Theorem 3.1.1. *Assume that $\int_0^1 \frac{1}{q(v)}dv < \infty$. As n approaches infinity, the random variable $\sqrt{n}(M_{X,n} - \mu)$, where $M_{X,n}$ is defined by (3.2), converges to a normal $\mathcal{N}(0, \sigma^2_{qm})$ distribution,*

$$\sqrt{n}\,(M_{X,n} - \mu) \xrightarrow{D} \mathcal{N}(0, \sigma^2_{qm}), \tag{3.8}$$

with $\sigma^2_{qm} = \int_0^2 \frac{v^2}{\bar{q}(v)}(F(v) - F(v-1))dv - \mu_V^2$ *and* $\mu_V = \int_0^2 v(F(v) - F(v-1))dv$.

Proof. First note that $\int_0^1 \frac{1}{q(v)}dv < \infty$ implies that σ^2_{qm} is well defined. To prove (3.8) we must show that the requirements of the Central Limit Theorem (CLT) are satisfied. First we will derive the mean and the variance of the estimator $M_{X,n}$. We have by (1.3) and (3.1)

$$\mathrm{E}\, M_{X,n} = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{V_i}{\bar{q}(V_i)} - \frac{1}{2}\right) = \mu,$$

and

$$\begin{aligned}\operatorname{Var}(M_{X,n}) &= \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{V_i}{\bar{q}(V_i)}\right) = \frac{1}{n}\operatorname{Var}\left(\frac{V_i}{\bar{q}(V_i)}\right) = \frac{1}{n}\left(\operatorname{E}\left(\frac{V_i}{\bar{q}(V_i)}\right)^2 - \left(\operatorname{E}\left(\frac{V_i}{\bar{q}(V_i)}\right)\right)^2\right)\\ &= \frac{1}{n}\left(\int_0^2 \frac{v^2}{(\bar{q}(v))^2}\bar{q}(v)(F(v)-F(v-1))dv - \left(\int_0^2 v(F(v)-F(v-1))dv\right)^2\right)\\ &= \frac{1}{n}\left(\int_0^2 \frac{v^2}{\bar{q}(v)}(F(v)-F(v-1))dv - \mu_V^2\right) = \frac{1}{n}\sigma_{qm}^2.\end{aligned}$$

Now by applying the CLT to the estimator $M_{X,n}$ we get the required result. □

Before we start with the theorem about the asymptotic distribution of the estimator $S_{X,n}^2$, let us first derive one assumption which is crucial for the following theorem in order for the asymptotic variance to be well defined. We have

$$\begin{aligned}\operatorname{Var}\left(\frac{V_i^2}{\bar{q}(V_i)} - 2\mu_V\frac{V_i}{\bar{q}(V_i)}\right) &= \operatorname{E}\left(\frac{V_i^2}{\bar{q}(V_i)} - 2\mu_V\frac{V_i}{\bar{q}(V_i)}\right)^2 - \left(\operatorname{E}\left(\frac{V_i^2}{\bar{q}(V_i)} - 2\mu_V\frac{V_i}{\bar{q}(V_i)}\right)\right)^2\\ &\le \operatorname{E}\left(\frac{V_i^2}{\bar{q}(V_i)} - 2\mu_V\frac{V_i}{\bar{q}(V_i)}\right)^2\\ &\le \operatorname{E}\left(\frac{V_i^4}{(\bar{q}(V_i))^2}\right) + 4\mu_V^2\operatorname{E}\left(\frac{V_i}{\bar{q}(V_i)}\right)^2\\ &\le 17\int_0^1\frac{1}{q(v)}dv + 20\mu_V^2\int_0^1\frac{1}{q(v)}dv\\ &= (20\mu_V^2+17)\int_0^1\frac{1}{q(v)}dv.\end{aligned} \tag{3.9}$$

Theorem 3.1.2. *Assume that $\int_0^1 \frac{1}{q(v)}dv < \infty$. As n approaches infinity, the random variable $\sqrt{n}(S_{X,n}^2 - \sigma^2)$, where $S_{X,n}^2$ is defined by (3.6), converges to a normal $\mathcal{N}(0,\sigma_{qs}^2)$ distribution,*

$$\sqrt{n}\left(S_{X,n}^2 - \sigma^2\right) \xrightarrow{D} \mathcal{N}(0,\sigma_{qs}^2), \tag{3.10}$$

with $\sigma_{qs}^2 = \operatorname{Var}\left(\frac{V_i^2}{\bar{q}(V_i)} - 2\mu_V\frac{V_i}{\bar{q}(V_i)}\right)$ and $\mu_V = \operatorname{E}\left(\frac{V_i}{\bar{q}(V_i)}\right)$.

Proof. First note that $\int_0^1 \frac{1}{q(v)}dv < \infty$ implies that σ_{qs}^2 is well defined. Write $Y_i = \frac{V_i}{\bar{q}(V_i)}$, $\operatorname{E} Y_i = \mu_y$ and $\operatorname{Var} Y_i = \sigma_y^2$ to simplify the notation. Note that $\mu_y = \mu_V$. Let $\{Y_i\}_{i=\{1,\dots,n\}}$ be a sequence of *i.i.d.* random variables, then we can write

$$\sqrt{n}\bar{Y}_n^2 = \sqrt{n}(\bar{Y}_n - \mu_y)^2 + 2\mu_y\sqrt{n}(\bar{Y}_n - \mu_y) + \sqrt{n}\mu_y^2.$$

Then from this equation it follows that

$$\sqrt{n}(\bar{Y}_n^2 - \mu_y^2) = \sqrt{n}(\bar{Y}_n - \mu_y)(\bar{Y}_n - \mu_y) + 2\mu_y\sqrt{n}(\bar{Y}_n - \mu_y), \tag{3.11}$$

where

$$\sqrt{n}(\bar{Y}_n - \mu_y)(\bar{Y}_n - \mu_y) \xrightarrow{P} 0, \tag{3.12}$$

$$2\mu_y\sqrt{n}(\bar{Y}_n - \mu_y) \xrightarrow{D} \mathcal{N}(0, 4\mu_y^2\sigma_y^2). \tag{3.13}$$

Equations (3.12), (3.13) are true, because of the CLT, the Kolmogorov Theorem (law of large numbers) and Slutsky's Theorem, see Serfling(1980). Then we can write

$$\sqrt{n}(\bar{Y}_n^2 - \mu_y^2) \xrightarrow{D} \mathcal{N}(0, 4\mu_y^2\sigma_y^2), \tag{3.14}$$

because of Slutsky's Theorem. Then from the equation (3.11) is obvious that

$$\bar{Y}_n^2 - \mu_y^2 = 2\mu_y\bar{Y}_n - 2\mu_y^2 + o_p(\frac{1}{\sqrt{n}}). \tag{3.15}$$

To prove (3.10) it suffices to show that

$$\begin{aligned} &\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\left[\frac{V_i^2}{\bar{q}(V_i)} - 2\mu_y\frac{V_i}{\bar{q}(V_i)}\right] - \mathrm{E}\left[\frac{1}{\bar{q}(V_i)}(V_i^2 - 2\mu_y V_i)\right]\right) \\ &= \sqrt{n}\left(S_{X,n}^2 - \sigma^2\right) + o_p(1). \end{aligned} \tag{3.16}$$

Note that

$$\sigma^2 = \mathrm{E}\left(\frac{V_i^2}{\bar{q}(V_i)}\right) - \mu^2 - \mu - \frac{1}{3},$$

see (3.4). Thus we can write

$$\begin{aligned} &\sqrt{n}\left(S_{X,n}^2 - \sigma^2\right) \\ &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{V_i^2}{\bar{q}(V_i)}\right) - \left(\frac{1}{n}\sum_{i=1}^{n}\frac{V_i}{\bar{q}(V_i)}\right)^2 - \frac{1}{12} - \sigma^2\right) \\ &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{V_i^2}{\bar{q}(V_i)}\right) - \left(\frac{1}{n}\sum_{i=1}^{n}\frac{V_i}{\bar{q}(V_i)}\right)^2 - \mathrm{E}\left(\frac{V_i^2}{\bar{q}(V_i)}\right) + \mu_y^2\right) \\ &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{V_i^2}{\bar{q}(V_i)}\right) - \mathrm{E}\left(\frac{V_i^2}{\bar{q}(V_i)}\right) - \left(\left(\frac{1}{n}\sum_{i=1}^{n}\frac{V_i}{\bar{q}(V_i)}\right)^2 - \mu_y^2\right)\right) \end{aligned}$$

Now we can apply (3.15) to the previous expression and we get

$$\begin{aligned} &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{V_i^2}{\bar{q}(V_i)}\right) - \mathrm{E}\left(\frac{V_i^2}{\bar{q}(V_i)}\right) - 2\mu_y\frac{1}{n}\sum_{i=1}^{n}\frac{V_i}{\bar{q}(V_i)} + 2\mu_y^2\right) + o_p(1) \\ &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\left[\frac{V_i^2}{\bar{q}(V_i)} - 2\mu_y\frac{V_i}{\bar{q}(V_i)}\right] - \mathrm{E}\left[\frac{1}{\bar{q}(V_i)}(V_i^2 - 2\mu_y V_i)\right]\right) + o_p(1). \end{aligned}$$

This proves the equality in (3.16). The CLT says, as n approaches infinity,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\left[\frac{V_i^2}{\bar{q}(V_i)}-2\mu_y\frac{V_i}{\bar{q}(V_i)}\right]-\mathrm{E}\left[\frac{1}{\bar{q}(V_i)}(V_i^2-2\mu_y V_i)\right]\right)\xrightarrow{D}\mathcal{N}(0,\sigma_{qs}^2), \tag{3.17}$$

where

$$\sigma_{qs}^2=\mathrm{Var}\left[\frac{V_i^2}{\bar{q}(V_i)}-2\mu_y\frac{V_i}{\bar{q}(V_i)}\right]. \tag{3.18}$$

□

3.2 Simulations of the estimators

In this Section we generate a sequence of random variables $X_i, i=1,\ldots,n$ and a sequence of random variables $T_i, i=1,\ldots,n$, from different distributions. Then we easily compute Δ_i, where $\Delta_i = 1_{[X_i\leq T_i]}$. To construct the random variable V_i, we use formula (1.2) and we will restrict only to Uniform distribution on $[0,1]$, its transformations (i.e. squared Uniform and square root Uniform distribution on $[0,1]$), and Beta distribution with scale parameters α and β. Then we will compute our estimates of the mean (3.2) and the variance (3.6), and compare them to the result of the theorems 3.1.1, and 3.1.2 for different values of n.

Note that, when the density q of the observation times T_i is Uniform on $[0,1]$, then $\bar{q}(.)\equiv 1$, see (1.4), and therefore our limit theorems for the estimators $M_{X,n}$ and $S_{X,n}^2$, see (2.10), and (2.11) are the same as (3.8), and (3.10).

i) Let us first consider the case, when the density q of the observation times T_i is Uniform on $[0,1]$ and the random variable X has an Uniform distribution on $[0,1]$. As we can see on Figures 3.1 and 3.2 simulations gave us similar results as in the previous chapter, see Figures 2.1 and 2.2, which also confirm that, when the density q is Uniform on $[0,1]$, the theorems 3.1.1 and 3.1.2 are the same as the theorems 2.2.1, and 2.2.2. Because of different generations of the random variable V the simulation results are not the same.

ii) The density q of the observation times T_i is squared Uniform on $[0,1]$ and the random variable X has a squared Uniform distribution on $[0,1]$.

Now we will check if the result holds our Theorem 3.1.1, where $M_{X,n}$ has an $\mathcal{AN}\left(\mu,\frac{\sigma_{qm}^2}{n}\right)$ distribution. When X has a squared Uniform distribution on $[0,1]$, then $\mu=\frac{1}{3}=0.33$ and variance $\sigma^2=\frac{4}{45}=8.89\times10^{-2}$, and we can write in general that $M_{X,n}$ has an $\mathcal{AN}\left(0.33,0.48/n\right)$ distribution.

For $n=500$ observations, $M_{X,500}$ approximately has a $\mathcal{N}\left(0.33,9.54\times10^{-4}\right)$ distribution and for $n=1000$ observations, $M_{X,1000}$ approximately has a $\mathcal{N}\left(0.33,4.77\times10^{-4}\right)$ distribution.

Simulations of the estimator $M_{X,n}$ gave us values, which are close to the real values. The simulation results support the theorem.

Let us check also the Theorem 3.1.2, where $S_{X,n}^2$ has an $\mathcal{AN}\left(\sigma^2,\frac{\sigma_{qs}^2}{n}\right)$ distribution. When X has a squared Uniform distribution on $[0,1]$, then we can write in general that $S_{X,n}^2$ has an $\mathcal{AN}\left(8.89\times10^{-2},0.18/n\right)$ distribution.

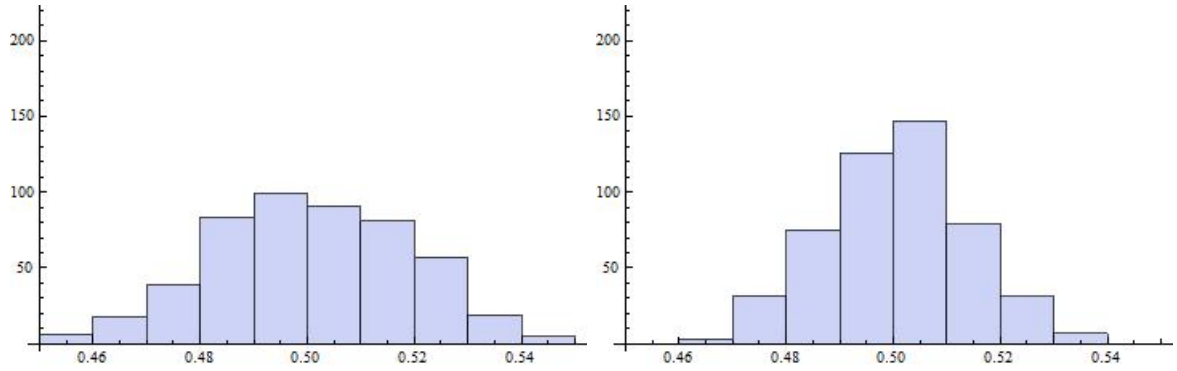


Figure 3.1: Histogram of 500 simulations of the estimator $M_{X,n}$, where the density q is Uniform on $[0,1]$ and the random variable X has an Uniform distribution on $[0,1]$. Left: For $n = 500$ observations, the mean of the $M_{X,n}$ samples is 0.50035 and the sample variance is equal to 3.41719×10^{-4}. Right: For $n = 1000$ observations, the mean of the $M_{X,n}$ samples is 0.500491 and the sample variance is equal to 1.76624×10^{-4}.

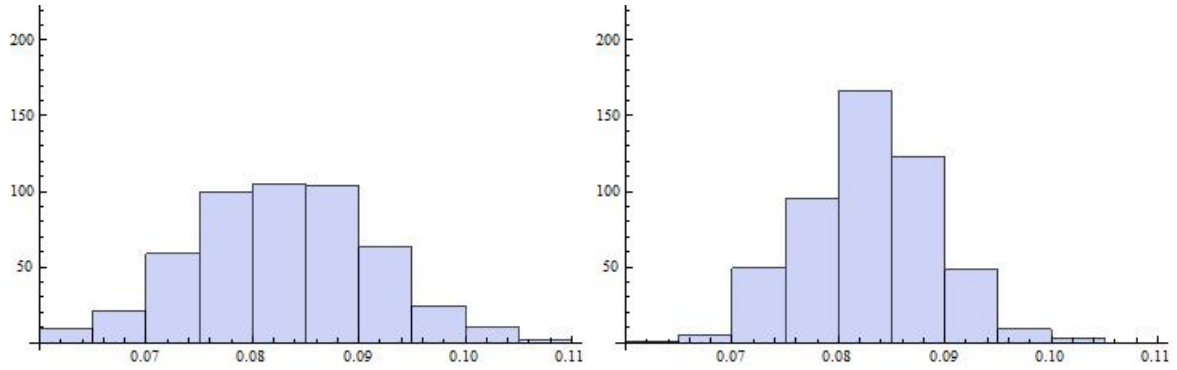


Figure 3.2: Histogram of 500 simulations of the estimator $S^2_{X,n}$, where the density q is Uniform on $[0,1]$ and the random variable X has an Uniform distribution on $[0,1]$. Left: For $n = 500$ observations, the mean of the $S^2_{X,n}$ samples is 8.27407×10^{-2} and the sample variance is equal to 7.36203×10^{-5}. Right: For $n = 1000$ observations, the mean of the $S^2_{X,n}$ samples is 8.30364×10^{-2} and the sample variance is equal to 3.68688×10^{-5}.

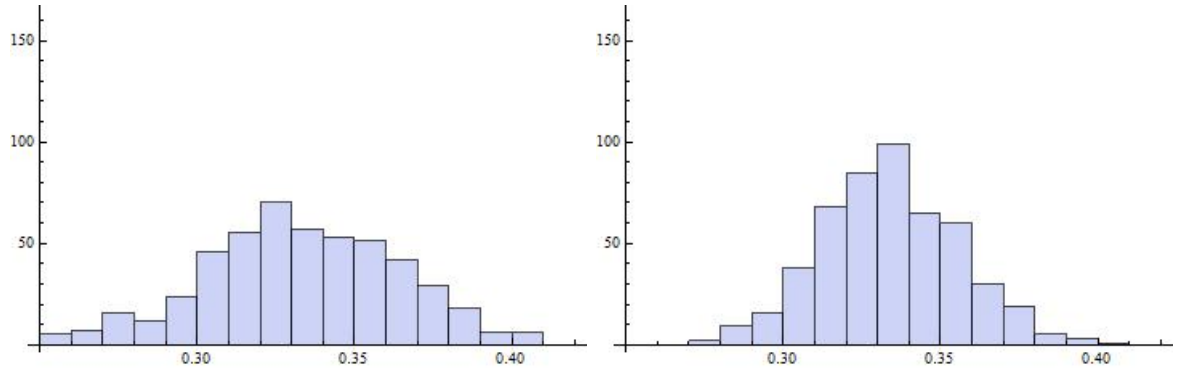


Figure 3.3: Histogram of 500 simulations of the estimator $M_{X,n}$, where the density q is squared Uniform on $[0,1]$ and the random variable X has a squared Uniform distribution on $[0,1]$. Left: For $n = 500$ observations, the mean of the $M_{X,n}$ samples is 0.333618 and the sample variance is equal to 9.88741×10^{-4}. Right: For $n = 1000$ observations, the mean of the $M_{X,n}$ samples is 0.333893 and the sample variance is equal to 4.70584×10^{-4}.

For $n = 500$ observations, $S^2_{X,500}$ approximately has a $\mathcal{N}\left(8.89 \times 10^{-2}, 3.55 \times 10^{-4}\right)$ distribution and for $n = 1000$ observations, $S^2_{X,1000}$ approximately has a $\mathcal{N}\left(8.89 \times 10^{-2}, 1.78 \times 10^{-4}\right)$ distribution.

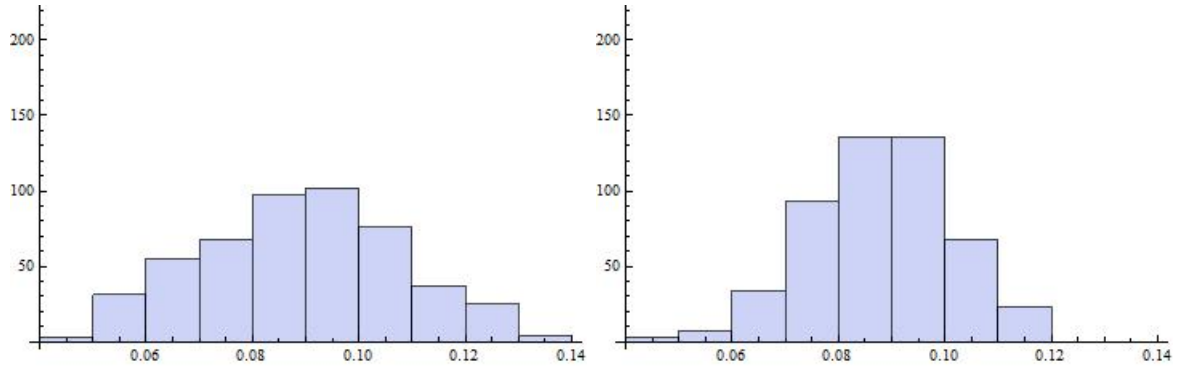


Figure 3.4: Histogram of 500 simulations of the estimator $S^2_{X,n}$, where the density q is squared Uniform on $[0,1]$ and the random variable X has a squared Uniform distribution on $[0,1]$. Left: For $n = 500$ observations, the mean of the $S^2_{X,n}$ samples is 8.89655×10^{-2} and the sample variance is equal to 3.55967×10^{-4}. Right: For $n = 1000$ observations, the mean of the $S^2_{X,n}$ samples is 8.80317×10^{-2} and the sample variance is equal to 1.79164×10^{-4}.

Simulations of the estimator $S^2_{X,n}$ gave us values, which confirm our Theorem 3.1.2.

iii) The density q of the observation times T_i is squared Uniform on $[0,1]$ and the random variable X has a square root Uniform distribution on $[0,1]$.
Now we will check if the result holds our Theorem 3.1.1, where $M_{X,n}$ has an $\mathcal{AN}\left(\mu, \frac{\sigma^2_{qm}}{n}\right)$ distribution. When X has a square root Uniform distribution on $[0,1]$, then $\mu = \frac{2}{3} = 0.67$

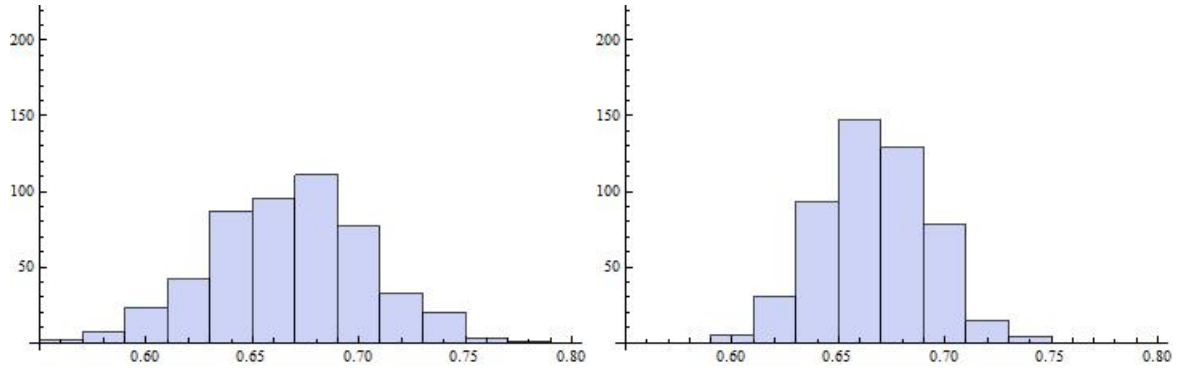


Figure 3.5: Histogram of 500 simulations of the estimator $M_{X,n}$, where the density q is squared Uniform on $[0,1]$ and the random variable X has a square root Uniform distribution on $[0,1]$. Left: For $n = 500$ observations, the mean of the $M_{X,n}$ samples is 0.667043 and the sample variance is equal to 1.33208×10^{-3}. Right: For $n = 1000$ observations, the mean of the $M_{X,n}$ samples is 0.666879 and the sample variance is equal to 6.49092×10^{-4}.

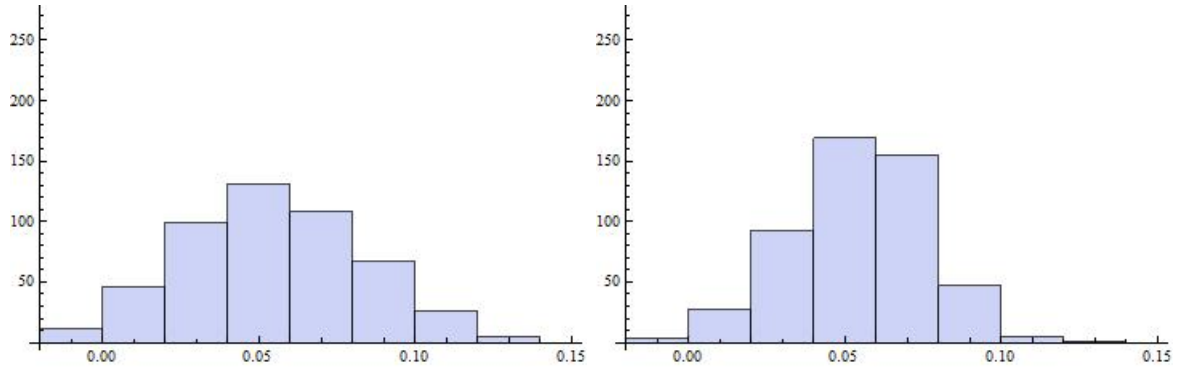


Figure 3.6: Histogram of 500 simulations of the estimator $S^2_{X,n}$, where the density q is squared Uniform on $[0,1]$ and the random variable X has a square root Uniform distribution on $[0,1]$. Left: For $n = 500$ observations, the mean of the $S^2_{X,n}$ samples is 5.30812×10^{-2} and the sample variance is equal to 9.43035×10^{-4}. Right: For $n = 1000$ observations, the mean of the $S^2_{X,n}$ samples is 5.44715×10^{-2} and the sample variance is equal to 4.63484×10^{-4}.

and variance $\sigma^2 = \frac{1}{18} = 5.56 \times 10^{-2}$, and we can write in general, that $M_{X,n}$ has an $\mathcal{AN}(0.67, 0.68/n)$ distribution.

For $n = 500$ observations, $M_{X,500}$ approximately has a $\mathcal{N}(0.67, 1.37 \times 10^{-3})$ distribution and for $n = 1000$ observations, $M_{X,1000}$ approximately has a $\mathcal{N}(0.67, 6.83 \times 10^{-4})$ distribution.

Simulations of the estimator $M_{X,n}$ gave us values, which are close to the real values. The simulation results support the theorem.

Let us check also the Theorem 3.1.2, where $S^2_{X,n}$ has an $\mathcal{AN}\left(\sigma^2, \frac{\sigma^2_{qs}}{n}\right)$ distribution. When X has a square root Uniform distribution on $[0, 1]$, then we can write in general that $S^2_{X,n}$ has an $\mathcal{AN}(5.56 \times 10^{-2}, 0.46/n)$ distribution.

For $n = 500$ observations, $S^2_{X,500}$ approximately has a $\mathcal{N}(5.56 \times 10^{-2}, 9.24 \times 10^{-4})$ distribution and for $n = 1000$ observations, $S^2_{X,1000}$ approximately has a $\mathcal{N}(5.56 \times 10^{-2}, 4.62 \times 10^{-4})$ distribution.
Simulations of the estimator $S^2_{X,n}$ gave us values, which confirm our Theorem 3.1.2.

iv) The density q of the observation times T_i is squared Uniform on $[0, 1]$ and the random variable X has a Beta distribution with scale parameters $\alpha = 5$ and $\beta = 2$.

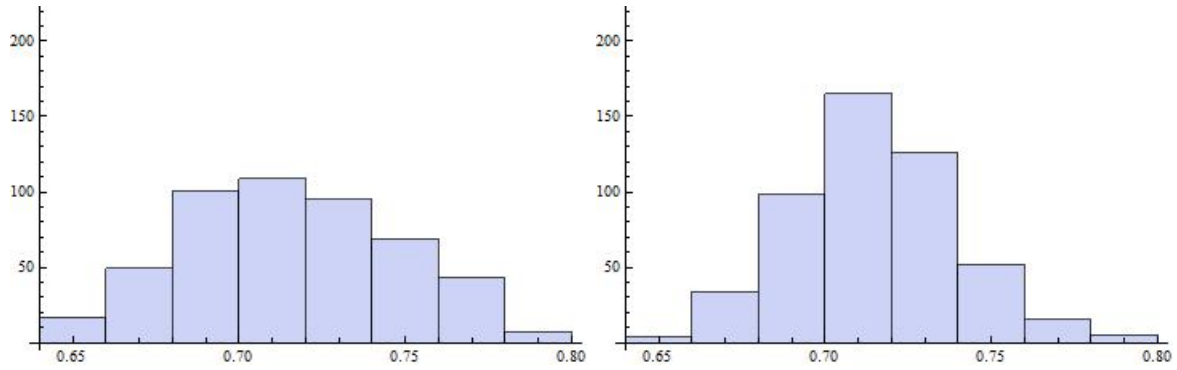


Figure 3.7: Histogram of 500 simulations of the estimator $M_{X,n}$, where the density q is squared Uniform on $[0, 1]$ and the random variable X has a Beta distribution with scale parameters $\alpha = 5$ and $\beta = 2$. Left: For $n = 500$ observations, the mean of the $M_{X,n}$ samples is 0.715547 and the sample variance is equal to 1.18261×10^{-3}. Right: For $n = 1000$ observations, the mean of the $M_{X,n}$ samples is 0.71441 and the sample variance is equal to 6.30535×10^{-4}.

Now we will check if the result holds our Theorem 3.1.1, where $M_{X,n}$ has an $\mathcal{AN}\left(\mu, \frac{\sigma^2_{qm}}{n}\right)$ distribution. When X has a Beta distribution with scale parameters $\alpha = 5$ and $\beta = 2$, then $\mu = \frac{5}{7} = 0.71$ and variance $\sigma^2 = \frac{5}{196} = 2.55 \times 10^{-2}$, and we can write in general, that $M_{X,n}$ has an $\mathcal{AN}(0.71, 0.67/n)$ distribution.

For $n = 500$ observations, $M_{X,500}$ approximately has a $\mathcal{N}(0.71, 1.34 \times 10^{-3})$ distribution and for $n = 1000$ observations, $M_{X,1000}$ approximately has a $\mathcal{N}(0.71, 6.70 \times 10^{-4})$ distribution.

Simulations of the estimator $M_{X,n}$ gave us values, which are close to the real values. The simulation results support the theorem.

Let us check also the Theorem 3.1.2, where $S^2_{X,n}$ has an $\mathcal{AN}\left(\sigma^2, \frac{\sigma^2_{qs}}{n}\right)$ distribution. When X has a Beta distribution with scale parameters $\alpha = 5$ and $\beta = 2$, then we can write in general that $S^2_{X,n}$ has an $\mathcal{AN}\left(2.55 \times 10^{-2}, 0.56/n\right)$ distribution.

For $n = 500$ observations, $S^2_{X,500}$ approximately has a $\mathcal{N}\left(2.55 \times 10^{-2}, 1.12 \times 10^{-3}\right)$ distribution and for $n = 1000$ observations, $S^2_{X,1000}$ approximately has a $\mathcal{N}\left(2.55 \times 10^{-2}, 5.61 \times 10^{-4}\right)$ distribution.

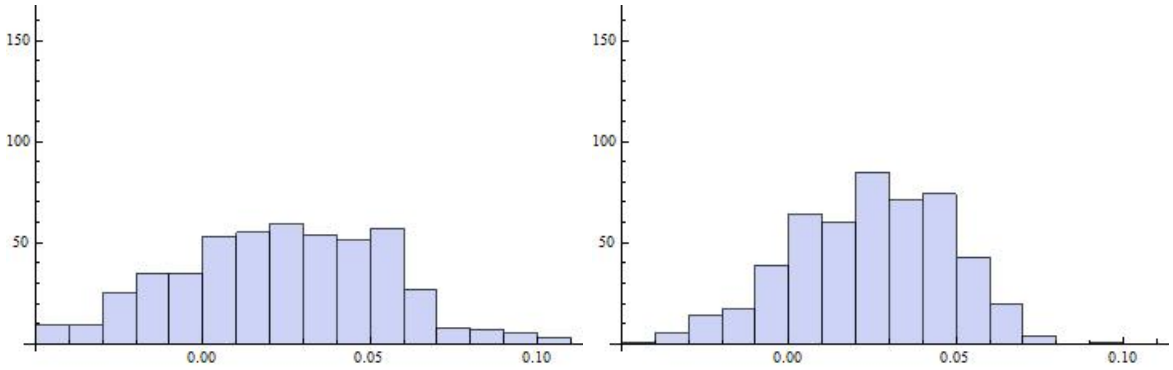


Figure 3.8: Histogram of 500 simulations of the estimator $S^2_{X,n}$, where the density q is squared Uniform on $[0, 1]$ and the random variable X has a Beta distribution with scale parameters $\alpha = 5$ and $\beta = 2$. Left: For $n = 500$ observations, the mean of the $S^2_{X,n}$ samples is 2.3026×10^{-2} and the sample variance is equal to 1.02878×10^{-3}. Right: For $n = 1000$ observations, the mean of the $S^2_{X,n}$ samples is 2.41975×10^{-2} and the sample variance is equal to 5.56178×10^{-4}.

Simulations of the estimator $S^2_{X,n}$ gave us values, which confirm our Theorem 3.1.2.

v) The density q of the observation times T_i is Beta distributed with scale parameters $\alpha = \frac{1}{2}$ and $\beta = \frac{1}{2}$, and the random variable X has a Uniform distribution on $[0, 1]$.

Now we will check if the result holds our Theorem 3.1.1, where $M_{X,n}$ has an $\mathcal{AN}\left(\mu, \frac{\sigma^2_{qm}}{n}\right)$ distribution. When X has a Uniform distribution on $[0, 1]$, then $\mu = \frac{1}{2}$ and variance $\sigma^2 = \frac{1}{12} = 8.33 \times 10^{-2}$, and we can write in general, that $M_{X,n}$ has an $\mathcal{AN}\left(0.5, 0.47/n\right)$ distribution.

For $n = 500$ observations, $M_{X,500}$ approximately has a $\mathcal{N}\left(0.5, 9.30 \times 10^{-4}\right)$ distribution and for $n = 1000$ observations, $M_{X,1000}$ approximately has a $\mathcal{N}\left(0.5, 4.65 \times 10^{-4}\right)$ distribution.

Simulations of the estimator $M_{X,n}$ gave us values, which are close to the real values. The simulation results support the theorem.

Let us check also the Theorem 3.1.2, where $S^2_{X,n}$ has an $\mathcal{AN}\left(\sigma^2, \frac{\sigma^2_{qs}}{n}\right)$ distribution. When X has a Uniform distribution on $[0, 1]$, then we can write in general, that $S^2_{X,n}$ has an $\mathcal{AN}\left(8.33 \times 10^{-2}, 0.16/n\right)$ distribution.

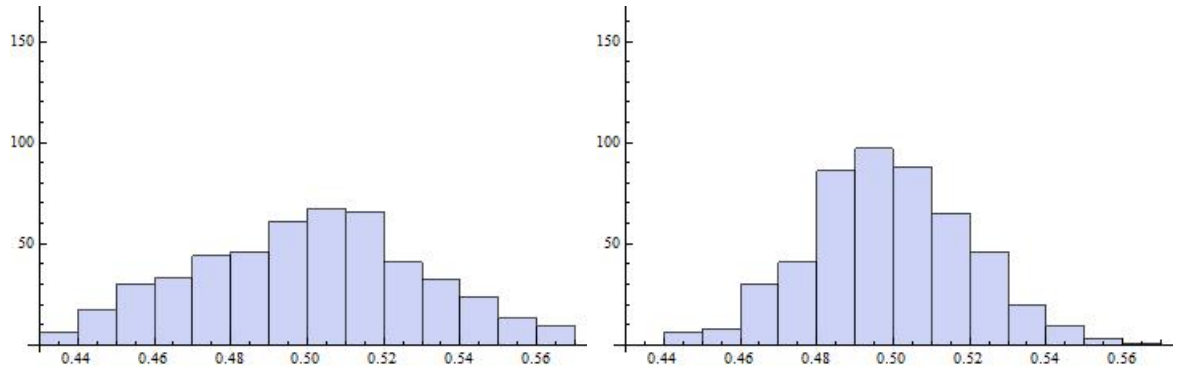


Figure 3.9: Histogram of 500 simulations of the estimator $M_{X,n}$, where the density q is Beta distributed with scale parameters $\alpha = \frac{1}{2}$ and $\beta = \frac{1}{2}$, and the random variable X has a Uniform distribution on $[0, 1]$. Left: For $n = 500$ observations, the mean of the $M_{X,n}$ samples is 0.499842 and the sample variance is equal to 9.81582×10^{-4}. Right: For $n = 1000$ observations, the mean of the $M_{X,n}$ samples is 0.498813 and the sample variance is equal to 4.3118×10^{-4}.

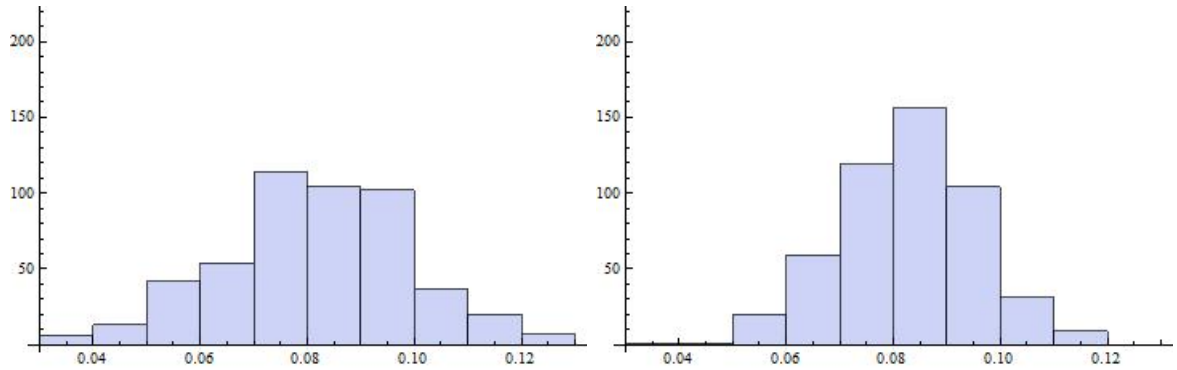


Figure 3.10: Histogram of 500 simulations of the estimator $S^2_{X,n}$, where the density q is Beta distributed with scale parameters $\alpha = \frac{1}{2}$ and $\beta = \frac{1}{2}$, and the random variable X has a Uniform distribution on $[0, 1]$. Left: For $n = 500$ observations, the mean of the $S^2_{X,n}$ samples is 8.14184×10^{-2} and the sample variance is equal to 3.07231×10^{-4}. Right: For $n = 1000$ observations, the mean of the $S^2_{X,n}$ samples is 8.28168×10^{-2} and the sample variance is equal to 1.62075×10^{-4}.

For $n = 500$ observations, $S^2_{X,500}$ approximately has a $\mathcal{N}(8.33 \times 10^{-2}, 3.27 \times 10^{-4})$ distribution and for $n = 1000$ observations, $S^2_{X,1000}$ approximately has a $\mathcal{N}(8.33 \times 10^{-2}, 1.63 \times 10^{-4})$ distribution.
Simulations of the estimator $S^2_{X,n}$ gave us values, which confirm our Theorem 3.1.2.

3.3 Estimators for mean and variance with unknown distribution of observation times

In this section we will discuss the case where the density q is unknown. Consider estimation of the density function q from the observations $T_1, \ldots, T_n$. Several generally applicable methods have been proposed for this problem, but let us review the *direct kernel density estimation*. The kernel density estimator with *kernel function* w and *bandwidth* $h > 0$, is defined by

$$q_{nh}(t) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h} w\left(\frac{t - T_j}{h}\right). \tag{3.19}$$

In the definition of $q_{nh}(.)$, the kernel w and the bandwidth h enter as unspecified parameters. By making the kernel fixed the bandwidth is usually chosen on the basis of the data. The choice of the bandwidth plays a crucial role in the performance of the estimator. Choice of a small bandwidth h leads to an estimator with small bias and large variance what produces noisy estimates whereas too large h leads to highly biased estimator producing flat estimates that do not reveal some interesting characteristics of q. Because of its relevancy, the selection of the bandwidth h is one of the mostly studied topics in kernel density estimation and several approaches have been proposed for choosing h. Wand and Jones (1995) made a good overview of the variety of methods that were proposed and appeared since the late seventies.

In this paper we will briefly review direct plug-in method, one of the most successful among all current methods. The core idea of the direct plug-in method dates back to Woodroofe(1970), later on modified by Nadaraya(1974) and Deheuvels and Hominal(1980). Direct plug-in is a very simple data dependent method for selection of the bandwidth. It is based on asymptotic approximations for the bandwidth h_0 that minimizes the mean integrated square error $MISE(q; n, h) = \mathrm{E}(ISE(q; n, h)) = \mathrm{E}\,\|q_n - q\|_2^2$, where $\|.\|$ denotes the L_2 distance:

$$h_0 = \arg\min_{h>0} MISE(q; n, h). \tag{3.20}$$

Chacón(2007) proved the existance and described asymptotic behaviour of h_0. Under some moment and regularity conditions on w and q, respectively, two asymptotic approximations to the optimal bandwidth h_0 are given by

$$h_1 = C_{1,w}\theta_2^{-1/5} n^{-1/5} \quad \text{and} \quad h_2 = C_{1,w}\theta_2^{-1/5} n^{-1/5} + C_{2,w}\theta_2^{-8/5}\theta_3 n^{-3/5}, \tag{3.21}$$

where θ_r denotes the quadratic functional

$$\theta_r = \int q^{(r)}(t)^2 dt = \left\|q^{(r)}\right\|_2^2, \tag{3.22}$$

with $r = 0, 1, \ldots, q^{(r)}$ the rth derivative of q, and

$$C_{1,w} = \omega_0^{1/5}\omega_2^{-2/5}, \quad C_{2,w} = \frac{1}{20}\omega_0^{3/5}\omega_2^{-11/5}\omega_4, \tag{3.23}$$

with $\omega_0 = \int w(u)^2 du, \omega_j = \int u^j w(u) du$ for $j = 1, 2, \ldots$, see Tenreiro(2010). These asymptotic approximations to h_0 reduce the problem of estimating the optimal bandwidth to that of estimating the quadratic functionals θ_2 and θ_3. This is the idea of the direct plug-in approach to bandwidth selection.

Next we show the mean and the variance of such a kernel density estimate of q. For smooth q, essentially twice continuously differentiable, and symmetric w with integral one, we have

$$\begin{aligned}
\mathrm{E}\, q_{nh}(t) &= \int_{-\infty}^{\infty} \frac{1}{h} w\left(\frac{t-u}{h}\right) q(u) du = q(t) + \frac{1}{2}h^2 q''(t) \int u^2 w(u) du + o(h^2), \\
\operatorname{Var} q_{nh}(t) &= \frac{1}{nh} q(t) \int w^2(u) du + o\left(\frac{1}{nh}\right),
\end{aligned}$$

as $n \to \infty, h \to 0$ and $nh \to \infty$. For proofs and more on direct kernel density estimators see for instance Prakasa Rao (1983), Silverman (1986) and Wand and Jones (1995).

Kernel density estimate of q can be computed in several mathematical and statistical programs like Wolfram Mathematica, R or Matlab, where this function is already predefined or can be easilly added by appropriate packages.

To show the estimators of the mean and the variance first we need to define the function $\bar{q}_{nh}$ by

$$\bar{q}_{nh}(t) = q_{nh}(t) + q_{nh}(t-1). \tag{3.24}$$

The principle of deriving estimators for the mean and the variance is exactly the same as for known distribution of observation times. We just replace the density $q(.)$ by the kernel denstiy estimate $q_{nh}(.)$ in our estimators ($\bar{q}$ by $\bar{q}_{nh}$ respectively). This gives the following estimators

$$M_{X,n} = \frac{1}{n}\sum_{i=1}^{n} \frac{V_i}{\bar{q}_{nh}(V_i)} - \frac{1}{2}, \tag{3.25}$$

$$S^2_{X,n} = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{V_i}{\sqrt{\bar{q}_{nh}(V_i)}}\right)^2 - \left(\frac{1}{n}\sum_{i=1}^{n} \frac{V_i}{\bar{q}_{nh}(V_i)}\right)^2 - \frac{1}{12}. \tag{3.26}$$

Because of lack of time we have not been able to perform simulations of these estimators. The limit theory seems complicated.

Chapter 4

Nonparametric maximum likelihood estimation (NPMLE) in interval censoring case 1.

4.1 One-step procedure for calculation NPMLE

In this section we will review the procedure for calculation of the NPMLE, $\hat{F}_n$, of the distribution function F. Most of the theory presented here is from Groeneboom and Wellner (1992). They showed two different characterizations of the NPMLE, one in terms of the so-called self-consistency equations and the other using concepts from the theory of isotonic regression. Let $(X_1, T_1), \ldots, (X_n, T_n)$ be a sample of random variables in $\mathbb{R}^2_+$, where X_i and T_i are independent (non-negative) random variables with distribution function F and Q, respectively. The setting is as in Chapter 3. The log likelihood for F is given by the function

$$F \mapsto \sum_{i=1}^{n} \{\Delta_i \log F(T_i) + (1 - \Delta_i) \log(1 - F(T_i))\}, \tag{4.1}$$

where F is a right-continous distribution function.

First we study the likelihood equations. The log likelihood, devided by n, can be written in the following way:

$$\psi(F) \stackrel{def}{=} \int_{\mathbb{R}^2_+} \left\{1_{\{x \leq t\}} \log F(t) + 1_{\{x > t\}} \log(1 - F(t))\right\} dP_n(x, t), \tag{4.2}$$

where P_n is the empirical probability measure of the pairs $(X_i, T_i), 1 \leq i \leq n$. The *nonparametric maximum likelihood estimator* $\hat{F}_n$ of F is a (right-continous) distribution function F, maximizing (4.2).

Note that only the values of $\hat{F}_n$ at the observation points matter for the maximization problem. The NPMLE $\hat{F}_n$ is a distribution function, which is piecewise constant, and only has jumps at the observation times T_i. It is possible that the likelihood function will be maximized by a function F such that $F(t) < 1$, at each observation time t. If this happens, we won't specify the location of the remaining mass to the right of the biggest observation time.

A theoretically useful characterization of the NPMLE in terms of the so-called self-consistency equations can be found in Groeneboom and Wellner(1992). We will give a different characterization of the NPMLE, using concepts from the theory of isotonic regression. To describe this method we need to set-up some notation. Suppose $T_{(i)}$ is the i^{th} order statistic of $T_1, \ldots, T_n$, and Δ_i is the corresponding indicator, i.e., if $T_j = T_{(i)}$, then $\Delta_{(i)} = 1_{\{X_j \leq T_j\}}$. The NPMLE corresponds to a vector $\tilde{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$, which maximizes the function

$$\phi(\tilde{x}) = \sum_{i=1}^{n} \left\{ \Delta_{(i)} \log x_i + (1 - \Delta_{(i)}) \log(1 - x_i) \right\}, \tilde{x} \in \mathbb{R}^n, \tag{4.3}$$

under the side condition

$$0 \leq x_1 \leq \cdots \leq x_n \leq 1. \tag{4.4}$$

When we look at the equation (4.3), it is obvious that, when $\Delta_{(i)} = 0, i = 1, \ldots, k$, also $y_1 = \cdots = y_k$ should be equal to 0 in order to maximize (4.3). This makes the second therm in the sum (4.3) as big as possible and puts no additional constraints on the values of y_i, for $i > k$. Similarly, if $\Delta_{(i)} = 1, j \leq i \leq n$, then $y_j = \cdots = y_n$ should be equal to 1 in order to maximize (4.3).

Without loss of generality we may assume that $\Delta_{(i)} = 1$ and $\Delta_{(n)} = 0$ for this maximization problem. According to this assumption we can assume as well without loss of generality that $y_1 > 0$ and $y_n < 1$, if $\tilde{y}$ maximizes (4.3), otherwise we would have $\phi(\tilde{y}) = -\infty$.

The following proposition give us necessary and sufficient conditions for $\tilde{y}$ to be a vector maximizing (4.3), under the constraint (4.4) and the just mentioned restrictions.

Proposition 4.1.1. *Let $\Delta_{(1)} = 1$ and $\Delta_{(n)} = 0$, and let $\tilde{y} = (y_1, \ldots, y_n)$ satisfy (4.4), with x_i replaced by y_i. Then $\tilde{y}$ maximizes (4.3) if and only if*

$$\sum_{j \geq i} \left\{ \frac{\Delta_{(j)}}{y_j} - \frac{1 - \Delta_{(j)}}{1 - y_j} \right\} \leq 0, i = 1, \ldots, n, \tag{4.5}$$

and

$$\sum_{i=1}^{n} \left\{ \frac{\Delta_{(i)}}{y_i} - \frac{1 - \Delta_{(i)}}{1 - y_i} \right\} y_i = 0. \tag{4.6}$$

Moreover, $\tilde{y}$ is uniquely determined by (4.5) and (4.6).

The proof of Proposition 4.1.1 can be found in Groeneboom and Wellner(1992).

Groeneboom and Wellner presented two possible solutions of this maximization problem of (4.3). One is so-called "max-min formula", where $y_m, 1 \leq m \leq n$, is given by

$$y_m = \max_{i \leq m} \min_{k \geq m} \frac{\sum_{i \leq j \leq k} \Delta_{(j)}}{k - i + 1},$$

and the other one can be found graphically by plotting the points $\left(i, \sum_{j \leq i} \Delta_{(j)}\right)$ in the plane, and drawing the *(greatest) convex minorant* of these points on the interval $[0, n]$.

Proposition 4.1.2. *Let function $H^* : [0,n] \to \mathbb{R}$ be the convex minorant of the points $\left(i, \sum_{j\le i}\Delta_{(j)}\right)$ on $[0,n]$, i.e.,*

$$H^*(t) = \sup\left\{H(t) : H(i) \le \sum_{j\le i}\Delta_{(j)}, \text{ for each } i, 0 \le i \le n, H(0) = 0, \text{ and } H \text{ is convex}\right\}, \tag{4.7}$$

for $t \in [0,n]$. Moreover, let y_i be the left derivative of H^* at i. Then $\tilde{y} = (y_1, \ldots, y_n)$ is the unique vector maximizing (4.3) *under the constraint* (4.4).

The proof of Proposition 4.1.2 can be found in Groeneboom and Wellner(1992).

Remark 4.1.1. *Note that in Proposition 4.1.2 (in contrast to Proposition 4.1.1) no restriction is made on $\Delta_{(1)}$ and $\Delta_{(n)}$.*

From the Proposition 4.1.2 it follows that $i \mapsto y_i$ is the isotonic regression of the function $i \mapsto \Delta_{(i)}$ in the class of all isotonic functions $i \mapsto x_i$, with respect to the simple ordering $x_1 \le \cdots \le x_n$ or in other words, the function $i \mapsto y_i$ minimizes

$$\sum_{i=1}^{n}\left\{\Delta_{(i)} - x_i\right\}^2,$$

in the class of isotonic functions $i \mapsto x_i$. For further details on the connection between the derivative of the convex minorant and the solution of the isotonic regression see e.g., Theorem 1.2.1 of Robertson et al. (1988).

So we get that the NPMLE, i.e. $\hat{F}_n$, of F, which maximizes (4.1), is given by

$$\hat{F}_n(T_{(i)}) = y_i,$$

where $\tilde{y}$ is the isotonic regression of the function $i \mapsto \Delta_{(i)}$.

Theorem 4.1.1. *Let t_0 be such that $0 < F(t_0), Q(t_0) < 1$, and let F and Q be differentiable at t_0, with strictly positive derivatives $f(t_0)$ and $q(t_0)$, respectively. Furthermore, let $\hat{F}_n$ be the NPMLE of F. Then we have, as $n \to \infty$,*

$$n^{1/3}\frac{\left\{\hat{F}_n(t_0) - F(t_0)\right\}}{\left\{\frac{1}{2}F(t_0)(1-F(t_0))f(t_0)/q(t_0)\right\}^{1/3}} \xrightarrow{D} 2U,$$

where $\xrightarrow{D}$ denotes convergence in distribution, and where U is the last time where standard two-sided Brownian motion minus the parabola $y(t) = t^2$ reaches its maximum.

The previous theorem describes the asymptotic distribution of $\hat{F}_n(t_0)$, for fixed $t_0 \in \mathbb{R}$, where $n^{1/3}$ is the obtained convergence rate. A straightforward proof of the previous theorem is outlined in Groeneboom and Wellner (1992) Chapter 5. in exercises 1 to 4. Another proof is given in Groeneboom (1987).

Example 4.1.1. *Here we will show some examples of convex minorant and its derivative (NPMLE), for different distributions of random variables X and T.*
Note that in Figure 4.4 the distribution function of the random variable X is $\sqrt{x}$ and the variance of the estimator grows as x goes to 0.

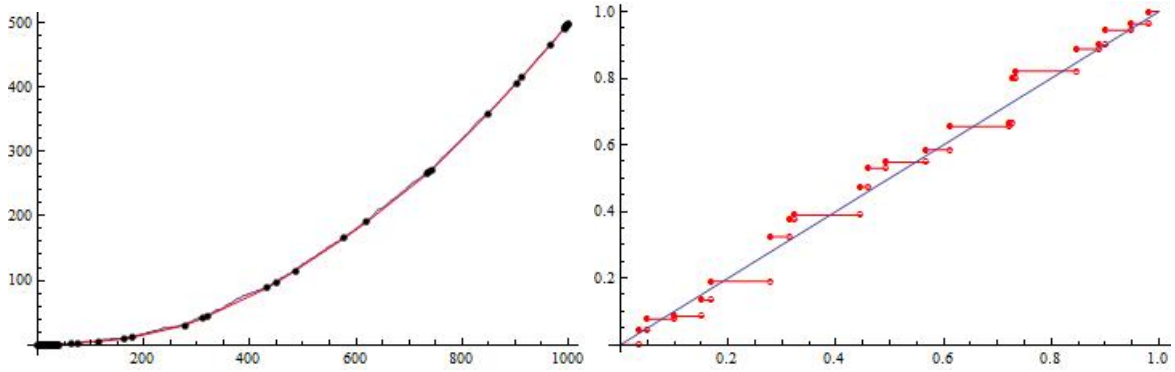


Figure 4.1: The random variables X and T are both from the Uniform distribution on $[0, 1]$, for $n = 1000$ observations. Left: Convex minorant. Right: Derivative of the convex minorant, the NPMLE, i.e. $\hat{F}_n$.

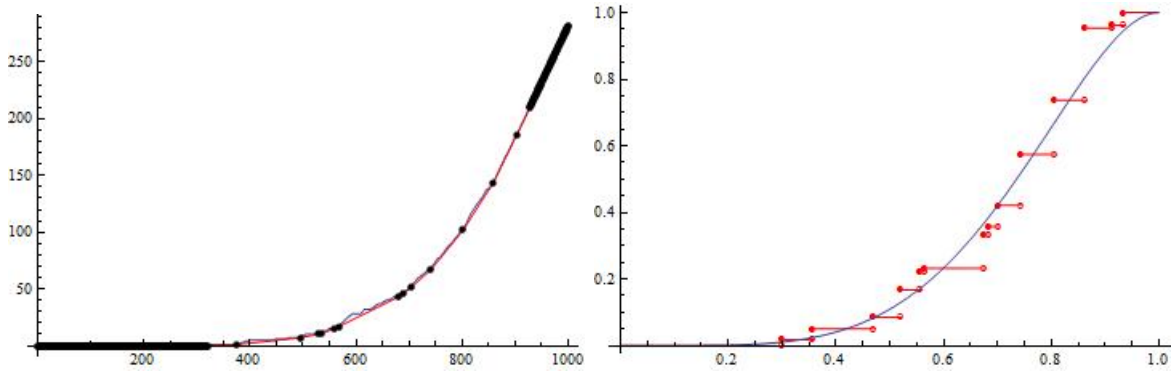


Figure 4.2: The random variable X is from Beta distribution with scale parameters $\alpha = 5$ and $\beta = 2$, and the random variable T is from the Uniform distribution on $[0, 1]$, for $n = 1000$ observations. Left: Convex minorant. Right: Derivative of the convex minorant, the NPMLE, i.e. $\hat{F}_n$.

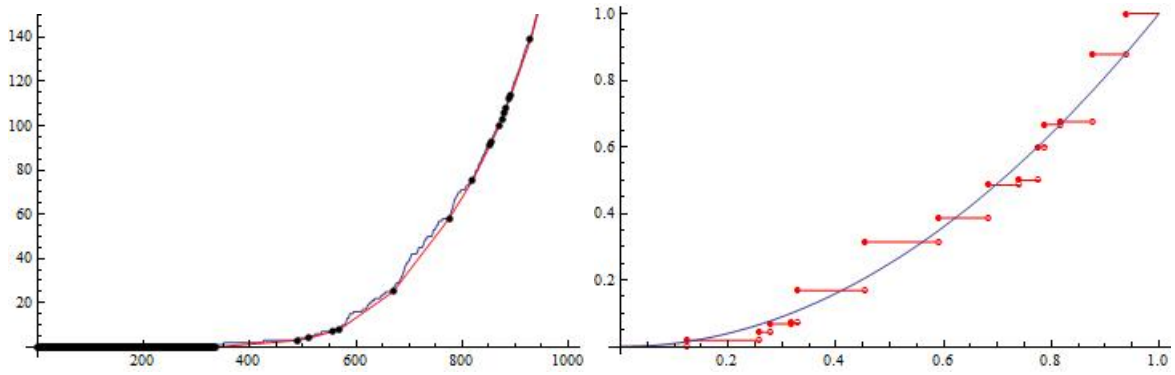


Figure 4.3: The random variable X is from the squared root Uniform distribution on $[0, 1]$, and the random variable T is from the squared Uniform distribution on $[0, 1]$, for $n = 1000$ observations. Left: Convex minorant. Right: Derivative of the convex minorant, the NPMLE, i.e. $\hat{F}_n$.

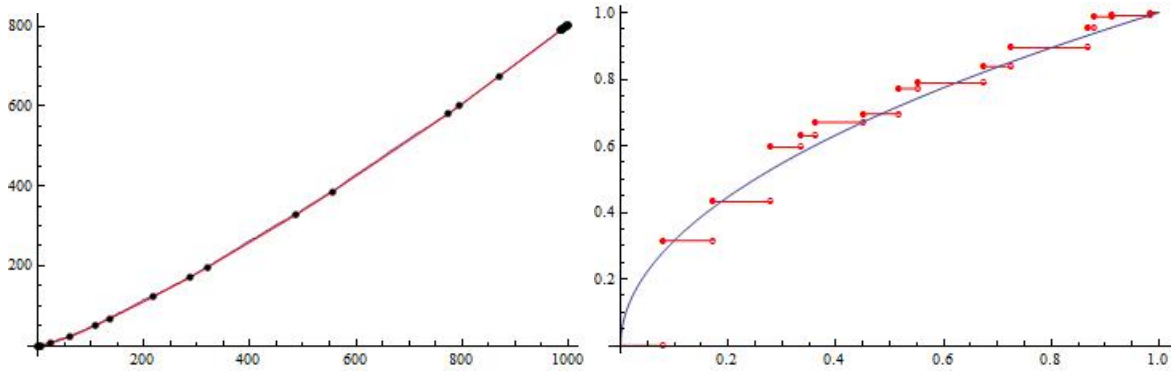


Figure 4.4: The random variable X is from the squared Uniform distribution on $[0, 1]$, and the random variable T is from the squared root Uniform distribution on $[0, 1]$, for $n = 1000$ observations. Left: Convex minorant. Right: Derivative of the convex minorant, the NPMLE, i.e. $\hat{F}_n$.

4.2 Convergence properties of the NPMLE $\mu(\hat{F}_n)$ of the mean $\mu(F)$

From the general theory on differentiable functionals (see e.g., van der Vaart (1991)) we know that, efficient estimators of smooth functionals like the mean

$$\mu_F = \int t dF(t) \tag{4.8}$$

should have $\sqrt{n}$-behavior. We are using the same set-up as in previous sections. Groeneboom and Wellner(1992) assume that the support of P_F is a bounded interval $I = [0, M]$, and that F and Q have densities f and q, respectively, satisfying

$$q(t) \geq \delta > 0, \text{and} \quad f(t) \geq \delta > 0, \quad \text{if} \quad t \in I,$$

for some $\delta > 0$. They assume also that q has a bounded derivative on I. An example of this situation could be the case where F and Q are both the uniform distribution function on $[0, 1]$. They claim that it is certainly possible to prove the following theorem under weaker conditions, but at the cost of an increasing number of technicalities.

Theorem 4.2.1. *Let F and Q satisfy the conditions, listed above, and let $\hat{F}_n$ be the NPMLE of F. Then*

$$\sqrt{n} \int_I (\hat{F}_n(t) - F(t)) dt \xrightarrow{D} Z, \tag{4.9}$$

where Z has a normal distribution with mean zero and variance

$$\sigma_F^2 = \int \frac{F(t)(1 - F(t))}{q(t)} dt. \tag{4.10}$$

When $\mu(\hat{F}_n)$ is the NPMLE of the mean $\mu(F)$, it follows from the previous theorem that

$$\sqrt{n}(\mu(\hat{F}_n) - \mu(F)) \xrightarrow{D} N\left(0, \int \frac{F(t)\,(1 - F(t))}{q(t)} dt\right), \quad \text{as} \quad n \to \infty \tag{4.11}$$

with $q(t)$ the density of the distribution of the observation times. Huang and Wellner (1995) prove a similar result for a wider class of functionals. The proof of Theorem 4.2.1 in Groeneboom (1992) uses the convergence rate of the supremum distance between the NPMLE and the underlying distribution function, which is replaced by a simpler argument based on $\mathcal{L}_2$-distance properties in Huang and Wellner's (1995) proof.

Geskus and Groeneboom (1996) also showed that $\mu(\hat{F}_n)$ is an asymptotically efficient estimator of $\mu(F)$. For a similar estimator $\sigma^2(\hat{F}_n)$ of $\sigma^2(F) = \int \left(t - \int s dF(s)\right)^2 dF(t)$, however, no limit theory exists.

Chapter 5

Uniform Deconvolution vs. NPMLE method in Interval censoring case 1.

In this Chapter we will focus on comparison of the performance of our estimators of Chapter 3. to the NPMLE method in the interval censoring problem for different numbers of observations, and for different distributions of the random variables X and T. We will also show the results of simulations of the estimators for different number of observations and for different distributions. It is most unlikely that different methods will agree exactly, for all kind of distributions. Our method is always worse asymptotically or asymptotically equivalent, because $\mu(\hat{F}_n)$ is an asymptotically efficient estimator. For $\sigma^2(\hat{F}_n)$ this is not clear.

The condition $\int_0^1 \frac{1}{q(v)} < \infty$ in our Theorem 3.1.1 and Theorem 3.1.2 limits us in possible densities for q of the observation times for which we can apply our theorems. Note that when q has an α Uniform density on $[0,1]$, i.e. the distribution of the random variable T is that of U^α, where U is uniformly distributed on $[0,1]$, then the parameter α should satisfy the following condition

$$\alpha > \frac{1}{2}. \tag{5.1}$$

When the density q is Beta(α, β), the parameters α and β should satisfy the following conditions

$$0 < \alpha < 2,\ 0 < \beta < 2. \tag{5.2}$$

Since there is no theory for the limit distribution of the variance estimator based on the NPMLE of F, we will restrict ourselves only to the comparison of the variances of the estimators of μ.

5.1 Comparison of the variances of the estimators of μ

To check how the two methods we used perform for different distributions, we will compare the variances of the estimators of μ for both methods. Theorem 3.1.1 and Theorem 4.2.1 describe the variances of such estimators. So to check if the theorems are exactly the same, we have to check whether the difference between the variances is equal to 0, so whether

$$\int_0^2 \frac{v^2}{\bar{q}(v)}(F(v) - F(v-1))dv - \mu_V^2 - \int_0^1 \frac{F(v)(1-F(v))}{q(v)}dv = 0. \tag{5.3}$$

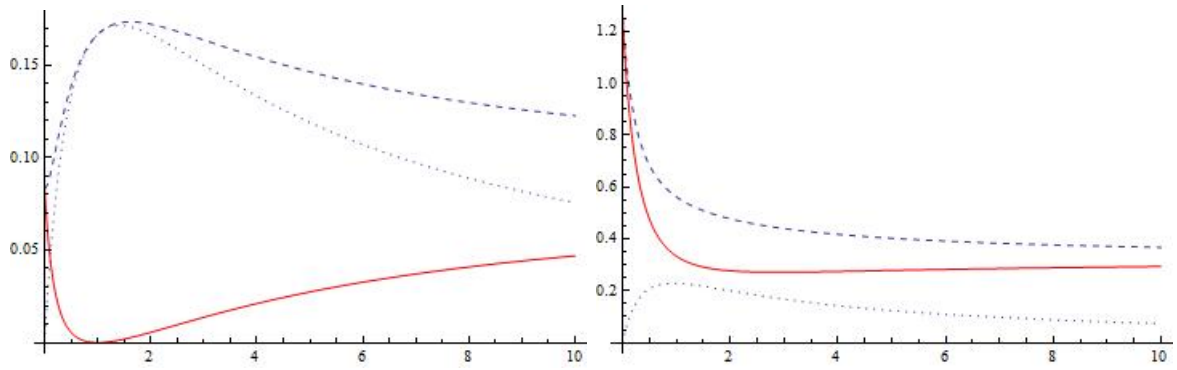


Figure 5.1: Variances of the mean estimate for both methods for $\alpha \in [0, 10]$, when the random variable X has an α Uniform distribution on $[0, 1]$ and T has a β Uniform distribution on $[0, 1]$. The dashed curve is Uniform deconvolution method, the dotted curve is the NPMLE method and the red curve is the difference between these two methods. Left: $\beta = 1$. Right: $\beta = 2$.

First we will consider that F is α Uniform distribution on $[0, 1]$, i.e. U^α, and Q is β Uniform distribution on $[0, 1]$, i.e. U^β, where $\alpha > 0$ and $\beta > \frac{1}{2}$. From equation (5.3) we have

$$\begin{aligned}
&\int_0^1 \left(v^2 v^{\frac{1}{\alpha}} \beta v^{\frac{\beta-1}{\beta}} + (v+1)^2 (1 - v^{\frac{1}{\alpha}}) \beta v^{\frac{\beta-1}{\beta}} \right) dv - \left(\frac{1}{1+\alpha} + \frac{1}{2} \right)^2 - \int_0^1 v^{\frac{1}{\alpha}} (1 - v^{\frac{1}{\alpha}}) \beta v^{\frac{\beta-1}{\beta}} dv \\
&= -\left(\frac{1}{2} + \frac{1}{1+\alpha} \right)^2 - \frac{\alpha\beta^3}{(\beta + \alpha(2\beta-1))\,(2\beta + \alpha(2\beta-1))} \\
&+ \frac{\beta^2 \left(\alpha^2 (1 - 5\beta + 6\beta^2)^2 + 2\beta^2 (2 - 12\beta + 17\beta^2) + \alpha\beta(-5 + 41\beta - 110\beta^2 + 98\beta^3) \right)}{(-1 + 9\beta - 26\beta^2 + 24\beta^3)\,(\beta + \alpha(-1+2\beta))\,(\beta + \alpha(-1+3\beta))}.
\end{aligned} \tag{5.4}$$

Note that for $\alpha = \beta = 1$, i.e. both F and Q are $Un[0, 1]$ the limit variances are equal.

Now we will consider that F is α Uniform distribution on $[0, 1]$ for $\alpha > 0$ and Q is Beta distributed with parameters $a = 1/2$ and $b = 1/2$. From the equation (5.3) we have

$$\begin{aligned}
&\int_0^1 \left(v^2 v^{\frac{1}{\alpha}} \pi \sqrt{v}\sqrt{1-v} + (v+1)^2 \left(1 - v^{\frac{1}{\alpha}}\right) \pi \sqrt{v}\sqrt{1-v} \right) dv - \left(\frac{1}{1+\alpha} + \frac{1}{2} \right)^2 \\
&- \int_0^1 v^{\frac{1}{\alpha}} \left(1 - v^{\frac{1}{\alpha}}\right) \pi \sqrt{v}\sqrt{1-v} dv \\
&= \frac{37}{128}\pi^2 - \left(\frac{1}{2} + \frac{1}{1+\alpha} \right)^2 - \frac{3\pi^{\frac{3}{2}}(1+2\alpha)\Gamma\left(\frac{3}{2} + \frac{1}{\alpha}\right)}{2\alpha\Gamma\left(4 + \frac{1}{\alpha}\right)} - \frac{1}{2}\pi^{\frac{3}{2}} \left(\frac{\Gamma\left(\frac{3}{2} + \frac{1}{\alpha}\right)}{\Gamma\left(3 + \frac{1}{\alpha}\right)} - \frac{\Gamma\left(\frac{3}{2} + \frac{2}{\alpha}\right)}{\Gamma\left(3 + \frac{2}{\alpha}\right)} \right).
\end{aligned} \tag{5.5}$$

Next we consider an F of the α Uniform distribution on $[0, 1]$ for $\alpha > 0$ and Q of the Beta

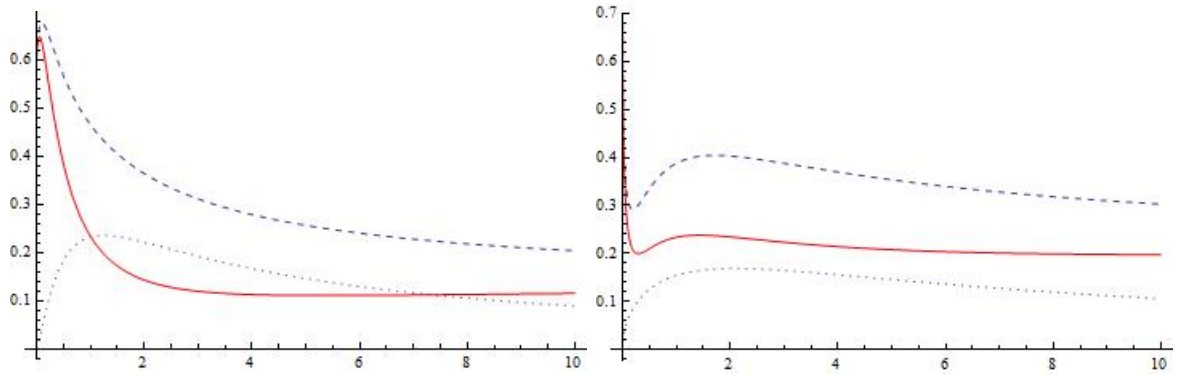


Figure 5.2: Variances of the mean estimate for both methods for $\alpha \in [0, 10]$, when the random variable X has an α Uniform distribution on $[0, 1]$. The dashed curve is Uniform deconvolution method, the dotted curve is the NPMLE method and the red curve is the difference between these two methods. Left: q is Beta distributed with parameters $a = 1/2$ and $b = 1/2$. Right: q is Beta distributed with parameters $a = 3/2$ and $b = 3/2$.

distribution with parameters $a = 3/2$ and $b = 3/2$. From equation (5.3) we have

$$\begin{aligned}&\int_0^1 \left(\frac{v^2 v^{\frac{1}{\alpha}} \pi}{8\sqrt{v}\sqrt{1-v}} + \frac{(v+1)^2 \left(1 - v^{\frac{1}{\alpha}}\right) \pi}{8\sqrt{v}\sqrt{1-v}} \right) dv - \left(\frac{1}{1+\alpha} + \frac{1}{2} \right)^2 - \int_0^1 \frac{v^{\frac{1}{\alpha}} \left(1 - v^{\frac{1}{\alpha}}\right) \pi}{8\sqrt{v}\sqrt{1-v}} dv \\ &= \frac{19}{64}\pi^2 - \left(\frac{1}{2} + \frac{1}{1+\alpha} \right)^2 - \frac{\pi^{\frac{3}{2}}(3+2\alpha)\Gamma\left(\frac{1}{2}+\frac{1}{\alpha}\right)}{8\alpha\Gamma\left(2+\frac{1}{\alpha}\right)} - \frac{1}{8}\pi^{\frac{3}{2}} \left(\frac{\Gamma\left(\frac{1}{2}+\frac{1}{\alpha}\right)}{\Gamma\left(1+\frac{1}{\alpha}\right)} - \frac{\Gamma\left(\frac{1}{2}+\frac{2}{\alpha}\right)}{\Gamma\left(\frac{2+\alpha}{\alpha}\right)} \right).\end{aligned} \tag{5.6}$$

The third example is an F of the α Uniform distribution on $[0, 1]$ for $\alpha > 0$ and Q of the Beta distribution with parameters $a = 5/3$ and $b = 7/6$. From equation (5.3) we have

$$\begin{aligned}&\int_0^1 \left(\frac{v^2 v^{\frac{1}{\alpha}} \Gamma\left(\frac{7}{6}\right)\Gamma\left(\frac{5}{3}\right)}{(1-v)^{\frac{1}{6}} v^{\frac{2}{3}} \Gamma\left(\frac{17}{6}\right)} + \frac{(v+1)^2 \left(1 - v^{\frac{1}{\alpha}}\right) \Gamma\left(\frac{7}{6}\right)\Gamma\left(\frac{5}{3}\right)}{(1-v)^{\frac{1}{6}} v^{\frac{2}{3}} \Gamma\left(\frac{17}{6}\right)} \right) dv - \left(\frac{1}{1+\alpha} + \frac{1}{2} \right)^2 \\ &- \int_0^1 \left(\frac{v^{\frac{1}{\alpha}} \left(1 - v^{\frac{1}{\alpha}}\right) \Gamma\left(\frac{7}{6}\right)\Gamma\left(\frac{5}{3}\right)}{(1-v)^{\frac{1}{6}} v^{\frac{2}{3}} \Gamma\left(\frac{17}{6}\right)} \right) dv \\ &= \frac{2544\sqrt{3}\pi}{5005} - \left(\frac{1}{2} + \frac{1}{1+\alpha} \right)^2 - \frac{3\sqrt{\pi}(18+11\alpha)\Gamma\left(\frac{7}{3}\right)\Gamma\left(\frac{1}{3}+\frac{1}{\alpha}\right)}{55\sqrt[3]{2}\alpha\Gamma\left(\frac{13}{6}+\frac{1}{\alpha}\right)} \\ &- \frac{\pi\Gamma\left(\frac{5}{3}\right)\left(\frac{\Gamma\left(\frac{1}{3}+\frac{1}{\alpha}\right)}{\Gamma\left(\frac{7}{6}+\frac{1}{\alpha}\right)} - \frac{\Gamma\left(\frac{1}{3}+\frac{2}{\alpha}\right)}{\Gamma\left(\frac{7}{6}+\frac{2}{\alpha}\right)} \right)}{3\Gamma\left(\frac{17}{6}\right)}.\end{aligned} \tag{5.7}$$

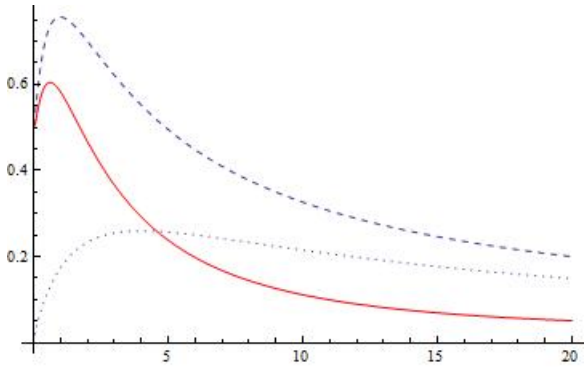


Figure 5.3: Variances of the mean estimate for both methods for $\alpha \in [0, 20]$. The dashed curve is Uniform deconvolution method, the dotted curve is the NPMLE method and the red curve is the difference between these two methods. The density q is Beta distributed with parameters $a = 5/3$ and $b = 7/6$, and F is an α Uniform distribution on $[0, 1]$.

5.2 Comparison of the simulations of the estimators of μ and σ^2

For the special case, when both random variables X and T have a Uniform distribution on $[0, 1]$, the estimators derived from the Uniform deconvolution model for interval censoring case 1. work better for a small number of observations than the NPMLE method, and for larger numbers of observations the resulting values of the estimators in both methods are close to the theoretical values, see Table 5.1 and Figure 5.4-5.5. As we can see in Figure 5.1 (left graph) the variances of the mean estimator should be exactly the same, which is also confirmed in the Table 5.1 of our simulations. Small differences in the simulation results are caused due to different random number generations.

"n" numbers of observations	Estimates	Uniform deconvolution method	NPMLE method
50	$\mathrm{E}\,\hat{\mu}$	0.494729	0.485422
	$n\,\mathrm{Var}\,\hat{\mu}$	0.164099	0.181879
	$\mathrm{E}\,\hat{\sigma}^2$	0.0812273	0.059999
	$n\,\mathrm{Var}\,\hat{\sigma}^2$	0.0389754	0.0262945
100	$\mathrm{E}\,\hat{\mu}$	0.497745	0.494727
	$n\,\mathrm{Var}\,\hat{\mu}$	0.161642	0.169427
	$\mathrm{E}\,\hat{\sigma}^2$	0.0830499	0.0687081
	$n\,\mathrm{Var}\,\hat{\sigma}^2$	0.0393127	0.0280347
500	$\mathrm{E}\,\hat{\mu}$	0.499946	0.498992
	$n\,\mathrm{Var}\,\hat{\mu}$	0.163922	0.182366
	$\mathrm{E}\,\hat{\sigma}^2$	0.0825502	0.0775889
	$n\,\mathrm{Var}\,\hat{\sigma}^2$	0.0368032	0.0329724
1000	$\mathrm{E}\,\hat{\mu}$	0.49936	0.499384
	$n\,\mathrm{Var}\,\hat{\mu}$	0.16819	0.190955
	$\mathrm{E}\,\hat{\sigma}^2$	0.0824645	0.0799377
	$n\,\mathrm{Var}\,\hat{\sigma}^2$	0.0394148	0.0337447

Table 5.1: Mean and variance of 500 simulations of the estimators of the mean μ and variance σ^2 for different numbers of observations. The random variables X and T are both from the Uniform distribution on $[0, 1]$. The theoretical value of $n\,\mathrm{Var}\,\hat{\mu}$ for both methods is equal to 0.17 and the theoretical value of $n\,\mathrm{Var}\,\hat{\sigma}^2$ for the uniform deconvolution method is equal to 0.04.

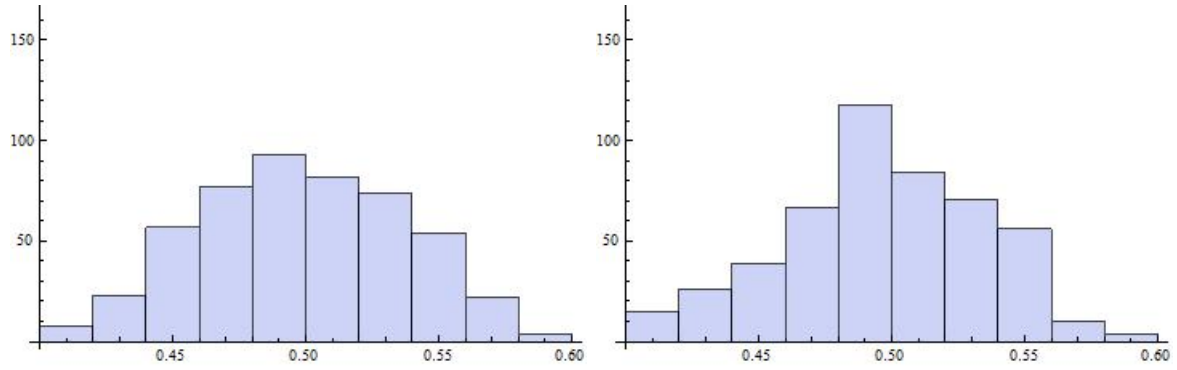


Figure 5.4: Histograms of 500 simulations of the estimator of the mean μ for $n = 100$ observations, where the random variables X and T are from the Uniform distribution on $[0, 1]$. Left: The Uniform deconvolution method in the interval censoring case 1. Right: The NPMLE method in the interval censoring case 1.

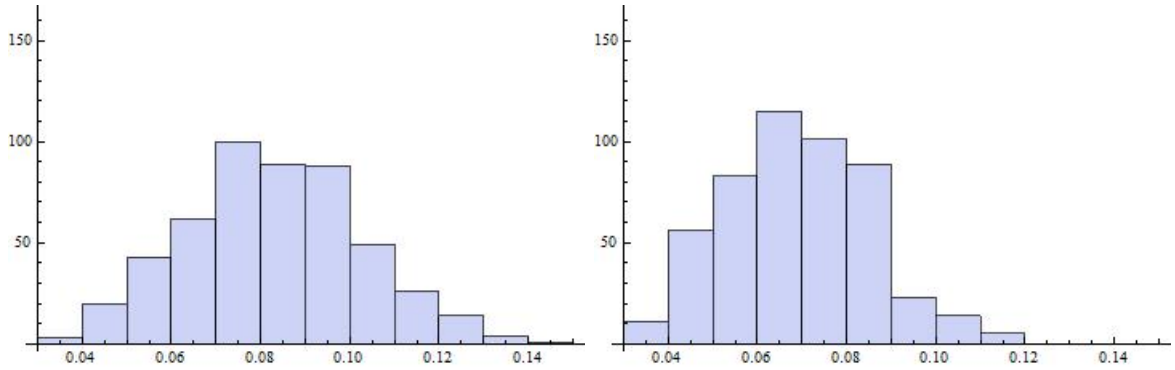


Figure 5.5: Histograms of 500 simulations of the estimator of the variance σ^2 for $n = 100$ observations, where the random variables X and T are from the Uniform distribution on $[0, 1]$. Left: The Uniform deconvolution method in the interval censoring case 1. Right: The NPMLE method in the interval censoring case 1.

"n" numbers of observations	Estimates	Uniform deconvolution method	NPMLE method
50	$\mathrm{E}\,\hat{\mu}$	0.666424	0.607416
	$n\,\mathrm{Var}\,\hat{\mu}$	0.667922	0.473418
	$\mathrm{E}\,\hat{\sigma}^2$	0.0387666	0.0352293
	$n\,\mathrm{Var}\,\hat{\sigma}^2$	0.476514	0.0177889
100	$\mathrm{E}\,\hat{\mu}$	0.665177	0.647004
	$n\,\mathrm{Var}\,\hat{\mu}$	0.697143	0.289355
	$\mathrm{E}\,\hat{\sigma}^2$	0.0483338	0.042259
	$n\,\mathrm{Var}\,\hat{\sigma}^2$	0.444199	0.0219097
500	$\mathrm{E}\,\hat{\mu}$	0.664324	0.66124
	$n\,\mathrm{Var}\,\hat{\mu}$	0.687873	0.217696
	$\mathrm{E}\,\hat{\sigma}^2$	0.0571431	0.0505707
	$n\,\mathrm{Var}\,\hat{\sigma}^2$	0.457352	0.0278321
1000	$\mathrm{E}\,\hat{\mu}$	0.666047	0.662532
	$n\,\mathrm{Var}\,\hat{\mu}$	0.6716	0.233164
	$\mathrm{E}\,\hat{\sigma}^2$	0.0543321	0.0521733
	$n\,\mathrm{Var}\,\hat{\sigma}^2$	0.48904	0.0271602

Table 5.2: Mean and variance of 500 simulations of the estimators of the mean μ and variance σ^2 for different numbers of observations. The random variable X has a square root Uniform distribution on $[0, 1]$ and T is from the squared Uniform distribution on $[0, 1]$. The theoretical value of $n\,\mathrm{Var}\,\hat{\mu}$ for the uniform deconvolution method is equal to 0.68 and for the NPMLE method is equal to 0.21. The theoretical value of $n\,\mathrm{Var}\,\hat{\sigma}^2$ for the uniform deconvolution method is equal to 0.46.

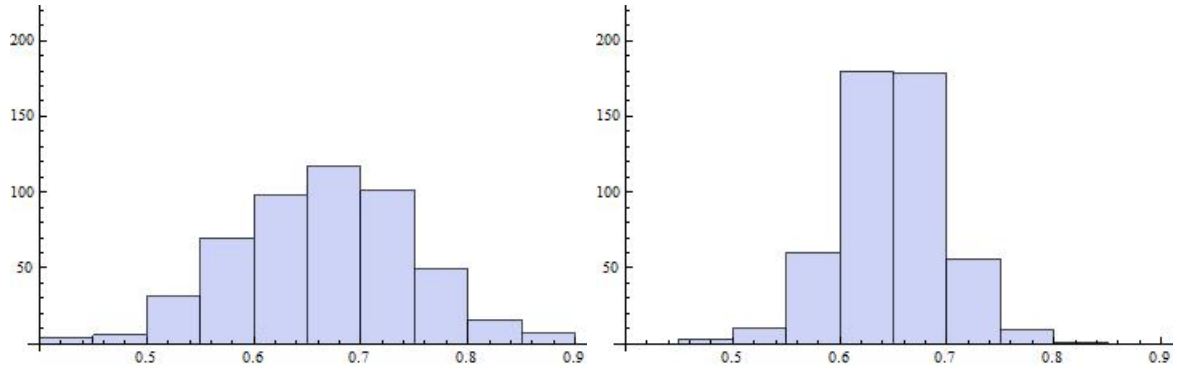


Figure 5.6: Histograms of 500 simulations of the estimator of the mean μ for $n = 100$ observations, where the random variable X has a square root Uniform distribution on $[0, 1]$, and T is from the squared Uniform distribution on $[0, 1]$. Left: The Uniform deconvolution method in the interval censoring case 1. Right: The NPMLE method in the interval censoring case 1.

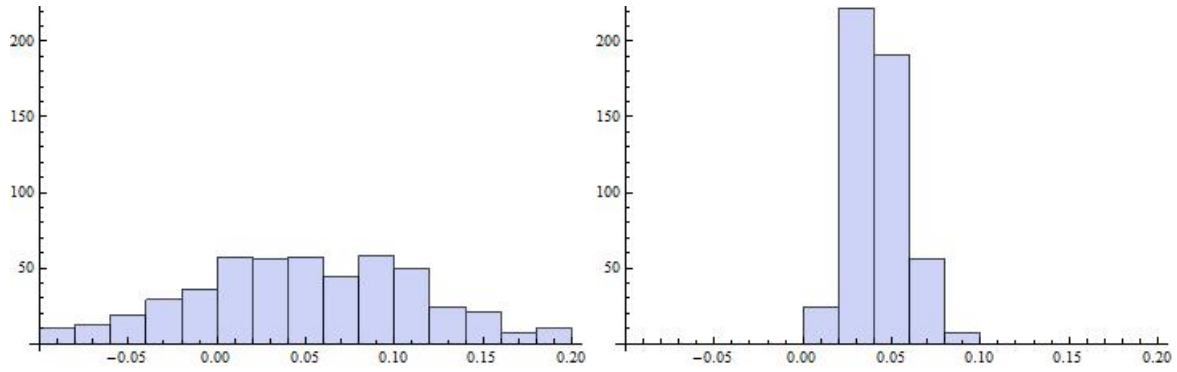


Figure 5.7: Histograms of 500 simulations of the estimator of the variance σ^2 for $n = 100$ observations, where the random variable X has a square root Uniform distribution on $[0, 1]$, and T is from the squared Uniform distribution on $[0, 1]$. Left: The Uniform deconvolution method in the interval censoring case 1. Right: The NPMLE method in the interval censoring case 1.

"n" numbers of observations	Estimates	Uniform deconvolution method	NPMLE method
50	$\mathrm{E}\,\hat{\mu}$	0.327372	0.33308
	$n\,\mathrm{Var}\,\hat{\mu}$	0.487122	0.197324
	$\mathrm{E}\,\hat{\sigma}^2$	0.0797685	0.064273
	$n\,\mathrm{Var}\,\hat{\sigma}^2$	0.229728	0.0307087
100	$\mathrm{E}\,\hat{\mu}$	0.326811	0.337766
	$n\,\mathrm{Var}\,\hat{\mu}$	0.593986	0.216281
	$\mathrm{E}\,\hat{\sigma}^2$	0.0830141	0.0708243
	$n\,\mathrm{Var}\,\hat{\sigma}^2$	0.282814	0.0336281
500	$\mathrm{E}\,\hat{\mu}$	0.331527	0.338828
	$n\,\mathrm{Var}\,\hat{\mu}$	0.596458	0.255957
	$\mathrm{E}\,\hat{\sigma}^2$	0.0879949	0.0806184
	$n\,\mathrm{Var}\,\hat{\sigma}^2$	0.230275	0.0403909
1000	$\mathrm{E}\,\hat{\mu}$	0.335192	0.338021
	$n\,\mathrm{Var}\,\hat{\mu}$	0.54687	0.217025
	$\mathrm{E}\,\hat{\sigma}^2$	0.087535	0.0832436
	$n\,\mathrm{Var}\,\hat{\sigma}^2$	0.214267	0.0416525

Table 5.3: Mean and variance of 500 simulations of the estimators of the mean μ and variance σ^2 for different numbers of observations. The random variable X has a squared Uniform distribution on $[0, 1]$ and T has a Beta distribution with parameters $a = 5/3$ and $b = 7/6$. The theoretical value of $n\,\mathrm{Var}\,\hat{\mu}$ for the uniform deconvolution method is equal to 0.70 and for the NPMLE method is equal to 0.23. The theoretical value of $n\,\mathrm{Var}\,\hat{\sigma}^2$ for the uniform deconvolution method is equal to 0.25.

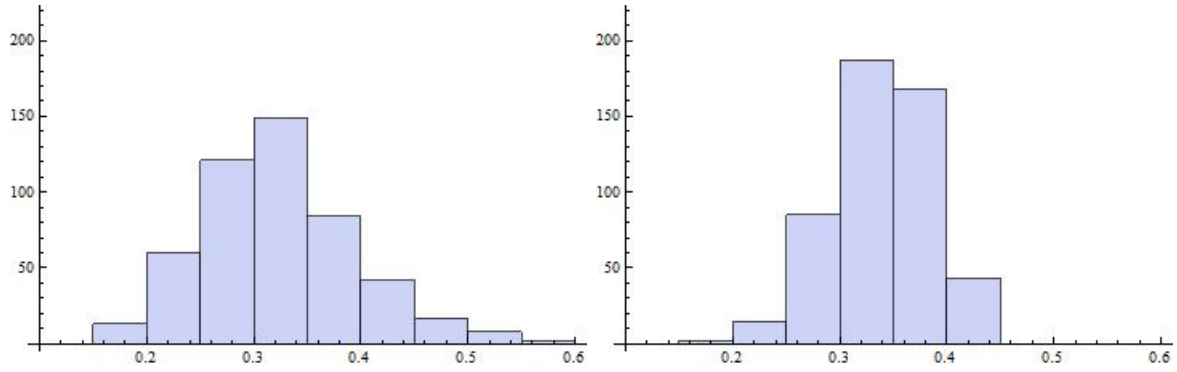


Figure 5.8: Histograms of 500 simulations of the estimator of the mean μ for $n = 100$ observations, where the random variable X has a squared Uniform distribution on $[0, 1]$, and T has a Beta distribution with parameters $a = 5/3$ and $b = 7/6$. Left: The Uniform deconvolution method in the interval censoring case 1. Right: The NPMLE method in the interval censoring case 1.

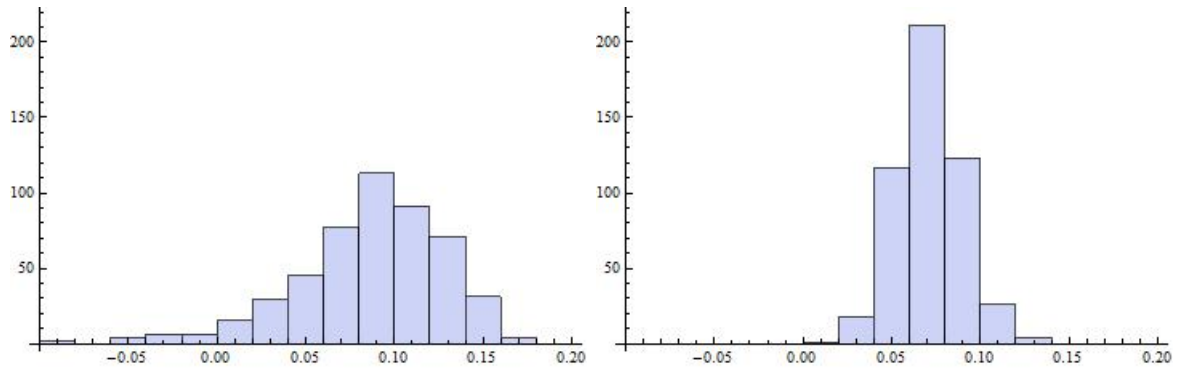


Figure 5.9: Histograms of 500 simulations of the estimator of the variance σ^2 for $n = 100$ observations, where the random variable X has a squared Uniform distribution on $[0, 1]$, and T has a Beta distribution with parameters $a = 5/3$ and $b = 7/6$. Left: The Uniform deconvolution method in the interval censoring case 1. Right: The NPMLE method in the interval censoring case 1.

"n" numbers of observations	Estimates	Uniform deconvolution method	NPMLE method
50	$\mathrm{E}\,\hat{\mu}$	0.0611257	0.069375
	$n\,\mathrm{Var}\,\hat{\mu}$	0.190362	0.128732
	$\mathrm{E}\,\hat{\sigma}^2$	0.0238639	0.0185147
	$n\,\mathrm{Var}\,\hat{\sigma}^2$	0.0432067	0.0115713
100	$\mathrm{E}\,\hat{\mu}$	0.058485	0.0703596
	$n\,\mathrm{Var}\,\hat{\mu}$	0.133171	0.126783
	$\mathrm{E}\,\hat{\sigma}^2$	0.0263796	0.0209462
	$n\,\mathrm{Var}\,\hat{\sigma}^2$	0.0205877	0.0133858
500	$\mathrm{E}\,\hat{\mu}$	0.0625968	0.0685972
	$n\,\mathrm{Var}\,\hat{\mu}$	0.1863	0.13714
	$\mathrm{E}\,\hat{\sigma}^2$	0.0279369	0.0249703
	$n\,\mathrm{Var}\,\hat{\sigma}^2$	0.0211424	0.0151118
1000	$\mathrm{E}\,\hat{\mu}$	0.0608505	0.0673941
	$n\,\mathrm{Var}\,\hat{\mu}$	0.159216	0.147629
	$\mathrm{E}\,\hat{\sigma}^2$	0.0279049	0.0258414
	$n\,\mathrm{Var}\,\hat{\sigma}^2$	0.0218457	0.0188084

Table 5.4: Mean and variance of 500 simulations of the estimators of the mean μ and variance σ^2 for different numbers of observations. The random variable X has a Uniform distribution on $[0, 1]$ to the power 15, i.e. U^{15} and T has a Beta distribution with parameters $a = 5/3$, and $b = 7/6$. The theoretical value of $n\,\mathrm{Var}\,\hat{\mu}$ for the uniform deconvolution method is equal to 0.25 and for the NPMLE method is equal to 0.18. The theoretical value of $n\,\mathrm{Var}\,\hat{\sigma}^2$ for the uniform deconvolution method is equal to 0.02.

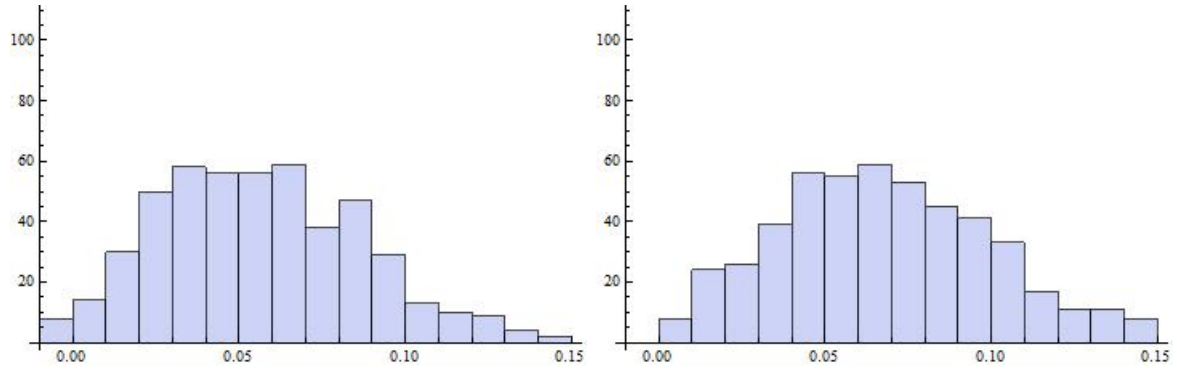


Figure 5.10: Histograms of 500 simulations of the estimator of the mean μ for $n = 100$ observations, where the random variable X has a Uniform distribution to the power 15, i.e. U^{15}, on $[0, 1]$ and T has a Beta distribution with parameters $a = 5/3$, and $b = 7/6$. Left: The Uniform deconvolution method in the interval censoring case 1. Right: The NPMLE method in the interval censoring case 1.

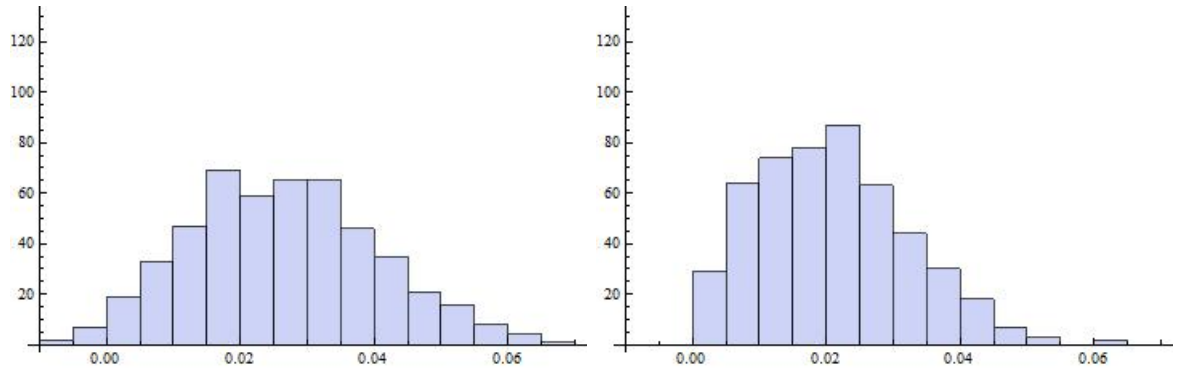


Figure 5.11: Histograms of 500 simulations of the estimator of the variance σ^2 for $n = 100$ observations, where the random variable X has a Uniform distribution to the power 15, i.e. U^{15}, on $[0, 1]$ and T has a Beta distribution with parameters $a = 5/3$, and $b = 7/6$. Left: The Uniform deconvolution method in the interval censoring case 1. Right: The NPMLE method in the interval censoring case 1.

"n" numbers of observations	Estimates	Uniform deconvolution method	NPMLE method
50	$\mathrm{E}\,\hat{\mu}$	0.329078	0.313078
	$n\,\mathrm{Var}\,\hat{\mu}$	0.355461	0.226959
	$\mathrm{E}\,\hat{\sigma}^2$	0.0854564	0.0648559
	$n\,\mathrm{Var}\,\hat{\sigma}^2$	0.110381	0.0369783
100	$\mathrm{E}\,\hat{\mu}$	0.331934	0.324128
	$n\,\mathrm{Var}\,\hat{\mu}$	0.360844	0.224074
	$\mathrm{E}\,\hat{\sigma}^2$	0.086389	0.0747777
	$n\,\mathrm{Var}\,\hat{\sigma}^2$	0.108557	0.039271
500	$\mathrm{E}\,\hat{\mu}$	0.330725	0.329716
	$n\,\mathrm{Var}\,\hat{\mu}$	0.332566	0.214185
	$\mathrm{E}\,\hat{\sigma}^2$	0.0884206	0.0841244
	$n\,\mathrm{Var}\,\hat{\sigma}^2$	0.10038	0.0436377
1000	$\mathrm{E}\,\hat{\mu}$	0.333427	0.332846
	$n\,\mathrm{Var}\,\hat{\mu}$	0.344408	0.220596
	$\mathrm{E}\,\hat{\sigma}^2$	0.0885696	0.0861581
	$n\,\mathrm{Var}\,\hat{\sigma}^2$	0.109132	0.0447196

Table 5.5: Mean and variance of 500 simulations of the estimators of the mean μ and variance σ^2 for different numbers of observations. The random variable X has a squared Uniform distribution on $[0, 1]$ and T has a Beta distribution with parameters $a = 1/2$, and $b = 1/2$. The theoretical value of $n\,\mathrm{Var}\,\hat{\mu}$ for the uniform deconvolution method is equal to 0.36 and for the NPMLE method is equal to 0.22. The theoretical value of $n\,\mathrm{Var}\,\hat{\sigma}^2$ for the uniform deconvolution method is equal to 0.10.

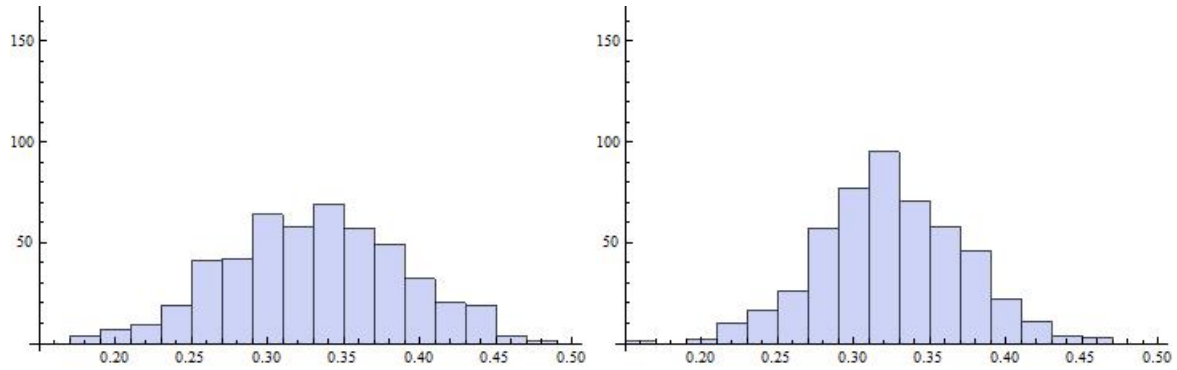


Figure 5.12: Histograms of 500 simulations of the estimator of the mean μ for $n = 100$ observations, where the random variable X has a squared Uniform distribution on $[0, 1]$ and T has a Beta distribution with parameters $a = 1/2$, and $b = 1/2$. Left: The Uniform deconvolution method in the interval censoring case 1. Right: The NPMLE method in the interval censoring case 1.

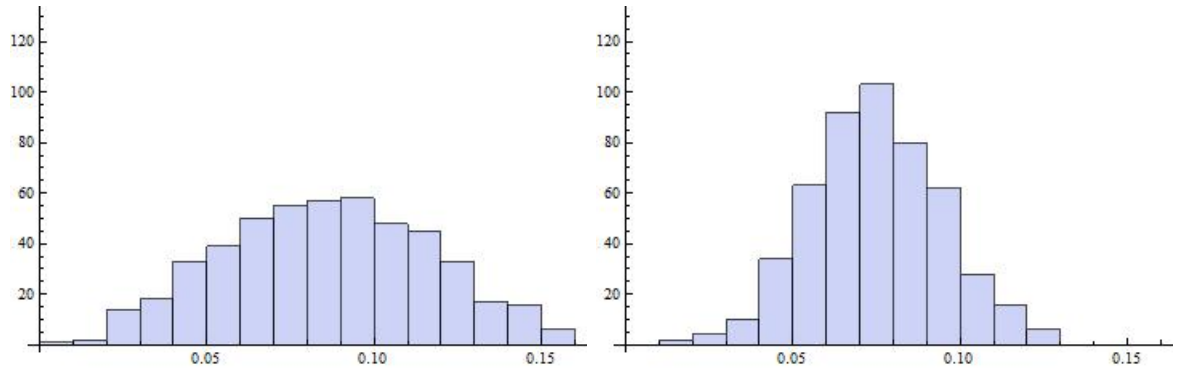


Figure 5.13: Histograms of 500 simulations of the estimator of the variance σ^2 for $n = 100$ observations, where the random variable X has a squared Uniform distribution on $[0, 1]$ and T has a Beta distribution with scale parameters $a = 1/2$, and $b = 1/2$. Left: The Uniform deconvolution method in the interval censoring case 1. Right: The NPMLE method in the interval censoring case 1.

"n" numbers of observations	Estimates	Uniform deconvolution method	NPMLE method
50	$\mathrm{E}\,\hat{\mu}$	0.24892	0.245317
	$n\,\mathrm{Var}\,\hat{\mu}$	0.370779	0.144096
	$\mathrm{E}\,\hat{\sigma}^2$	0.0749734	0.0560885
	$n\,\mathrm{Var}\,\hat{\sigma}^2$	0.151972	0.0216635
100	$\mathrm{E}\,\hat{\mu}$	0.253026	0.251191
	$n\,\mathrm{Var}\,\hat{\mu}$	0.38705	0.147398
	$\mathrm{E}\,\hat{\sigma}^2$	0.0761187	0.0624301
	$n\,\mathrm{Var}\,\hat{\sigma}^2$	0.1108	0.0300449
500	$\mathrm{E}\,\hat{\mu}$	0.249856	0.252665
	$n\,\mathrm{Var}\,\hat{\mu}$	0.44376	0.170586
	$\mathrm{E}\,\hat{\sigma}^2$	0.0788549	0.0729135
	$n\,\mathrm{Var}\,\hat{\sigma}^2$	0.128771	0.0325075
1000	$\mathrm{E}\,\hat{\mu}$	0.248686	0.251408
	$n\,\mathrm{Var}\,\hat{\mu}$	0.403786	0.165569
	$\mathrm{E}\,\hat{\sigma}^2$	0.0807041	0.0747131
	$n\,\mathrm{Var}\,\hat{\sigma}^2$	0.102595	0.0376444

Table 5.6: Mean and variance of 500 simulations of the estimators of the mean μ and variance σ^2 for different numbers of observations. The random variable X has a Uniform distribution to the power 3, i.e. U^3, on $[0, 1]$ and T has a Beta distribution with parameters $a = 3/2$, and $b = 3/2$. The theoretical value of $n\,\mathrm{Var}\,\hat{\mu}$ for the uniform deconvolution method is equal to 0.39 and for the NPMLE method is equal to 0.17. The theoretical value of $n\,\mathrm{Var}\,\hat{\sigma}^2$ for the uniform deconvolution method is equal to 0.10.

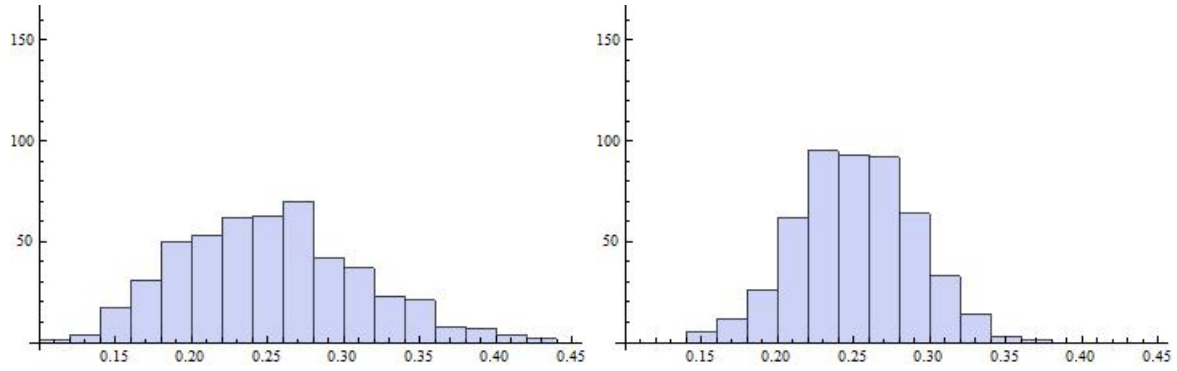


Figure 5.14: Histograms of 500 simulations of the estimator of the mean μ for $n = 100$ observations, where the random variable X has a Uniform distribution to the power 3, i.e. U^3, on $[0, 1]$ and T has a Beta distribution with parameters $a = 3/2$, and $b = 3/2$. Left: The Uniform deconvolution method in the interval censoring case 1. Right: The NPMLE method in the interval censoring case 1.

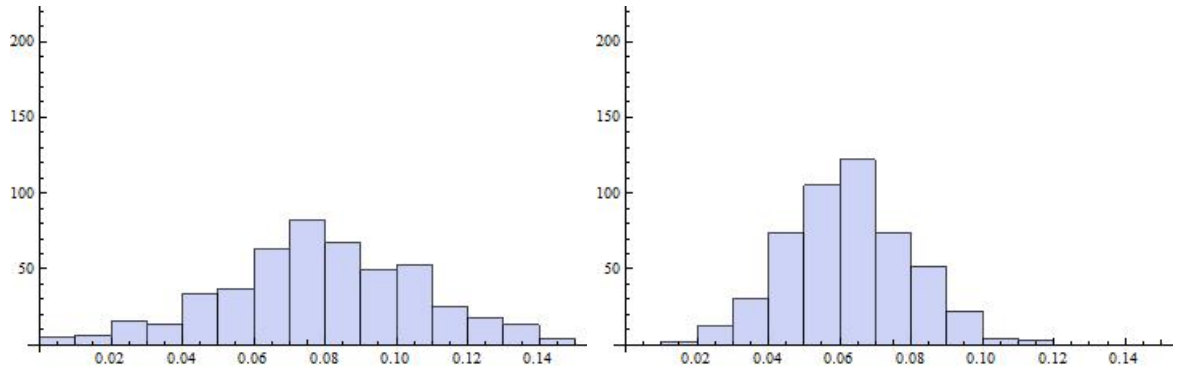


Figure 5.15: Histograms of 500 simulations of the estimator of the variance σ^2 for $n = 100$ observations, where the random variable X has a Uniform distribution to the power 3, i.e. U^3, on $[0, 1]$ and T has a Beta distribution with parameters $a = 3/2$ and $b = 3/2$. Left: The Uniform deconvolution method in the interval censoring case 1. Right: The NPMLE method in the interval censoring case 1.

Chapter 6

Discussion and conclusions

In the preceding chapters we discussed two different approaches for the interval censoring case 1 problem, the uniform deconvolution method and the NPMLE method. In both methods we described the estimators of the mean and the variance and their limit distributions. Then we compared the simulation results of these estimators in order to present their behaviour under some specific distributions of the random variables. By specific distributions we mean distributions of the random variables, which satisfie the necessary condition involved in the theorems about the limit distributions of the estimators of the mean and variance. We expected that the estimators of the mean μ and variance σ^2 in the uniform deconvolution method will perform better than the NPMLE based estimators under some distributions of the random variables X and T.

The uniform deconvolution method is based on the transformation of the data from the uniform deconvolution model into interval censored data. The estimators of the mean and the variance in the uniform deconvolution method are relatively simple, derived from estimators of the moments of the random variables $V = X + U$ in the uniform deconvolution model, and based on the sample mean and sample variance.

The NPMLE method is based on estimating the distribution function F by nonparametric maximum likelihood and then computing the mean and variance of this distribution.

Simulations of the estimators in the uniform deconvolution method show results which are closer to the theoretical values of the mean μ and variance σ^2 than the results of the NPMLE based estimators for small samples. Considering small sample behaviour, the NPMLE of distribution function is inaccurate and then also the mean and variance are far from the theoretical values. With increasing number of observations the NPMLE of distribution function is more precise, i.e. the values of the sample mean and the sample variance are closer to the theoretical values. What makes the uniform deconvolution method worse is the fact that the asymptotic variances of the estimators of the mean μ and variance σ^2 are larger than the asymptotic variances of the estimators in the NPMLE method. This confirms the fact that $\mu(\hat{F}_n)$ is an asymptotically efficient estimator of $\mu(F)$. In the special case when both random variables X and T are uniformly distributed on $[0, 1]$, the estimators of the mean μ for both methods reach the same results. Because no theory about the asymptotic variance behaviour of the estimator of the variance $\sigma^2(F)$ exists, we can only discuss about the efficiency of such an estimator. In all the simulations we showed, the variance of the NPMLE based estimator of the variance σ^2 is always smaller than the variance in the uniform deconvolution method.

This suggests that $\sigma^2(\hat{F}_n)$ could be an asymptotically efficient estimator of $\sigma^2(F)$.

One possible way we could continue and improve the estimators of the mean μ and variance σ^2 in the uniform deconvolution method is to use Le Cam's one-step efficient estimator. The procedure of constructing such an estimator is described in Le Cam(1986).

Bibliography

[1] Chacón J.E., Montanero J., Nogales A.G., Pérez P. (2007), 'On the existance and limit behaviour of the optimal bandwidth for kernel density estimation', *Statistica Sinica*, 17, 289-300.

[2] Deheuvels P., Hominal P. (1980), 'Estimation automatique de la densité', *Revue de Statistique Appliquée*, 28, 25-55.

[3] Geskus R. B., Groeneboom P. (1996), *Asymptotically optimal estimation of smooth functionals for interval censoring, part 1.*, Statist. Neer., volume **50**, Issue 1, pages 69-88.

[4] Groeneboom P. (1987), *Asymptotics for interval censored observations*, Technical Report 87-18, Department of Mathematics, University of Amsterdam.

[5] Groeneboom P., Wellner A. J. (1992), *Information Bounds and Nonparametric Maximum Likelihood Estimation*, Birkhauser Verlag, Basel. Boston. Berlin.

[6] Huang J., Wellner J. A.(1995), *Asymptotic normality of the NPMLE of linear functionals for interval censored data,case 1.*, Statist. Neer. **49**, 153-63.

[7] Le Cam L. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer, New York.

[8] Nadaraya E.A (1974), 'On the integral mean square error of some nonparametric estimates for the density function', *Theory of Probability and Its Applications*, 19, 133-141.

[9] Prakasa Rao B.L.S. (1983), *Nonparametric Functional Estimation*, Academic Press, New York.

[10] Robertson T., Wright F. T., Dykstra R. L. (1988), *Order Restricted Statistical Inference*, Wiley, New York.

[11] Rudin W.(1986), *Real and Complex Analysis, international edition*, McGraw-Hill Book Company, Singapore.

[12] Serfling J. R. (1980), *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, New York. Chichester. Brisbane. Toronto. Singapore.

[13] Silverman B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

[14] Tenreiro C. (2010), 'Fourier series based direct plug-in bandwidth selectors for kernel density estimation', *Journal of Nonparametric Statistics*, 23, 533-545.

[15] van der Vaart, A. W. (1991), *On differentiable functionals*, Ann. Statist. **19**, 178-205.

[16] Wand M. P. and Jones M. C. (1995), *Kernel Smoothing*, Chapman and Hall, London.

[17] Woodroofe M. (1970), *The Annals of Mathematical Statistics*, 'On choosing a delta-sequence' 41, 1665-1671.