

# **Estimation of Abundance Based on Line Transect Data with and without the Shoulder Condition**

**By**

**Ayat Mohammad Mostafa Al Momani**

**Supervisor**

**Dr. Omar M. Eidous**

**Department of Statistics**

**September, 2011**


**Estimation of Abundance Based on Line Transect Data with and without the  
Shoulder Condition**

By

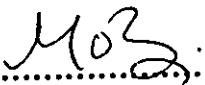
**Ayat Mohammad Mostafa Al Momani**  
B.Sc. Mathematics, Yarmouk University, 2011

**A thesis submitted in partial fulfillment of the requirement for the degree of  
Master of Science in the Department of Statistics, Faculty of  
Science, Yarmouk University, Irbid, Jordan**

Approved by:

**Dr. Omar M. Eidous**..........**Chairman**  
Associate Professor of Statistics, Yarmouk University.

**Professor Mohammad Fraiwan Al-Saleh**..........**Member**  
Professor of Statistics, Yarmouk University.

**Dr. Moustafa Abu-Shawesh**..........**Member**  
Assistant Professor of Statistics, Al-Hashemiah University.

September, 2011

© Arabic Digital Library - Yarmouk University

## الاهناء

الى من خلقتني فأحسن خلقي الله عز وجل  
الى منارة العلم و سيد الخلق الامام المصطفى  
الى من حصد الاشواك من دربي ليمهد لي طريق العلم  
الى من احمل اسمه بكل افتخار والذي العزيز  
الى من علمتني العطاء بدون انتظار  
الى من كان دعائها سر نجاحي امي الحبيبه  
الى رفيق دربي صاحب القلب الطيب زوجي نذير  
الى القلب الطاهر الرقيق و النفس البريئة ابنتي ملاك  
الى سندي و قوتي وملاذي بعد الله اخواني  
الى من تنوقت معهم أجل اللحضات اخواني  
الى الوجه المنعم بالبراءة ابنت اختي شذى  
الى الأرواح الطاهرة اجدادي و جداتي  
الى من بهم أكبر و بوجودهم أعتز أهلي و عشيرتي  
الى من تحلو بالاخاء و تميزوا بالوفاء و العطاء صديقاتي  
الى كل من ساندني و تمنى لي الخير  
اهدي رسالتي

## شكر و تقدير

لا بد لنا ونحن نخطو خطواتنا الأخيرة في الحياة الجامعية من وقفه نعود بها الى أعوام قضيناها في رحاب الجامعة  
مع أساتذتنا الكرام  
الذين قدموا لنا الكثير باذلين بذلك جهودا كبيرة في بناء جيل الغد لتبعث الأمة من جديد  
وقبل ان نمضي تقدم أسمى آيات الشكر والامتنان و التقدير الى الذين حملوا أقدس رسالة في الحياة  
الى الذين مهدوا لنا طريق المعرفة  
الى جميع اساتذتي الأفاضل  
في قسم الرياضيات و قسم الاحصاء  
في هذا الصرح العلمي  
جامعة اليرموك

وأخص بالتقدير والشكر  
الدكتور عمراعدوس  
الذي تفضل بالاشراف على هذه الرسالة فجزاه الله عنا كل خير وله منا كل التقدير والاحترام

# LIST OF FIGURES

<b>Figure 1.1:</b> Diagram represents the line transect method.	1
<b>Figure 1.2:</b> The perpendicular distance $x$ in line transect method	2
<b>Figure 2.1:</b> Represent the detection function $g(x)$ of the Exponential power (EP) model for $\beta = 1.0, 1.5, 2.0,$ and $2.5$	27
<b>Figure 2.2:</b> Represent the detection function $g(x)$ of the Hazard-rate (HR) model for $\beta = 1.5, 2.0, 2.5,$ and $3.0$	28
<b>Figure 2.3:</b> Represent the detection function $g(x)$ of the Beta (BE) model for $\beta = 1.5, 2.0, 2.5,$ and $3.0$	29

## ABSTRACT

**Al-Momani, Ayat Mohammad. Estimation of Abundance Based on Line Transect Data with and without the Shoulder Condition. Master's Degree Thesis, Department of Statistics, Yarmouk University, 2011 (Supervisor: Dr. Omar M. Eidous).**

In this thesis, some nonparametric estimators of the population abundance  $D$  using line transect sampling are presented and compared. These estimators are divided into two categories; the estimators that are developed under the shoulder condition assumption and the estimators that are developed without assuming the shoulder condition. For each case, a new estimator is proposed and investigated mathematically and numerically.

Firstly, we compared between the performances of these estimators for two cases; when the data are simulated from densities that satisfy the shoulder condition and from densities that do not satisfy the shoulder condition. The simulation technique is adopted through the broad range of models to identify the most promising estimator in the case that the shoulder condition is true and in the case that it does not. A comparison of these estimators was undertaken in order to determine whether or not results of different estimators could be combined in analysis of line transect data.

Secondly and based on the simulation results, we proposed different new estimators that combined between the most promising two estimators when the shoulder condition is valid and when it is violated. their performances are compared with the estimator that is called semi-parametric estimator in the literature. The comparison study among the different estimators shows that the performances of the proposed estimators are satisfactory as general estimators for both cases.

**Keywords:** Line transect method; Shoulder condition; Estimation of abundance; Kernel method; Smoothing parameter; Boundary effect.

## المخلص

المومني، ايات محمد. تقدير كثافة المجتمع في حالة البيانات المأخوذة بطريقة الخط العرضي بوجود و بعدم وجود شرط الكتف. رسالة الماجستير في العلوم، قسم الإحصاء، جامعة اليرموك، 2011. (المشرف: الدكتور عمر عدوس).

في هذه الرسالة، عرضنا بعض المقدرات غير المعلمية لكثافة المجتمع في حالة العينة المأخوذة بطريقة الخط العرضي وقمنا بالمقارنة بينها. قسمت هذه المقدرات إلى فئتين؛ مقدرات مطوره عند ادعاء تحقق شرط الكتف، ومقدرات مطورة عند فرض عدم تحقق شرط الكتف، ثم قدمنا في كل حاله مقدر جديد و تحريناه رياضيا و عدديا.

أولا قمنا بمقارنة كفاءة هذه المقدرات باستخدام بيانات مأخوذة من اقترانات تحقق الشرط ومن اقترانات لا تحقق الشرط، و اخترنا تقنية المحاكاة لتعيين أفضل مقدر في كل حالة.

ثانيا و بالاعتماد على نتائج المحاكاة، قدمنا بعض المقدرات المختلفة و هي عبارة عن مقدرات مركبة من افضل مقدر في حالة تحقق شرط الكتف و افضل مقدر في حالة عدم تحققه و قارنا نتائجهما مع نتائج مقدر يسمى بشبه معلمي. من بين المقدرات التي درست في هذه الرسالة أعطت النتائج أن أداء المقدرات المقترحة جيدة كمقدرات عامة في الحالتين عند تحقق شرط الكتف و عند عدم تحققه.

الكلمات المفتاحية: طريقة الخط العرضي، شرط الكتف، تقدير كثافة المجتمع، نظام النواة، معلمة التنعيم، تأثير الحدود.

# CHAPTER ONE

## INTRODUCTION AND LITERATURE REVIEW

### 1.1 Introduction

Line transect method is a popular and convenient technique used to estimate the density (abundance) of a biological population  $D$ , since it is direct, cost efficient and can be carried out on foot, or from a variety of land, air, or watercraft. Assume that the population size is  $N$  and the sampled area is  $A$  then the population density is  $D = N / A$ .

In line transect method, an area of known boundaries and size is divided into non-overlapping strips, each with known length (Figure 1.1). Then an observer moves on the middle of line of the strip and records the perpendicular distances  $x$  from the centerline to a detected object within the strip as illustrated in Figure 1.2. The total length of lines  $l_1, l_2, \dots, l_k$  is denoted by  $L$ .

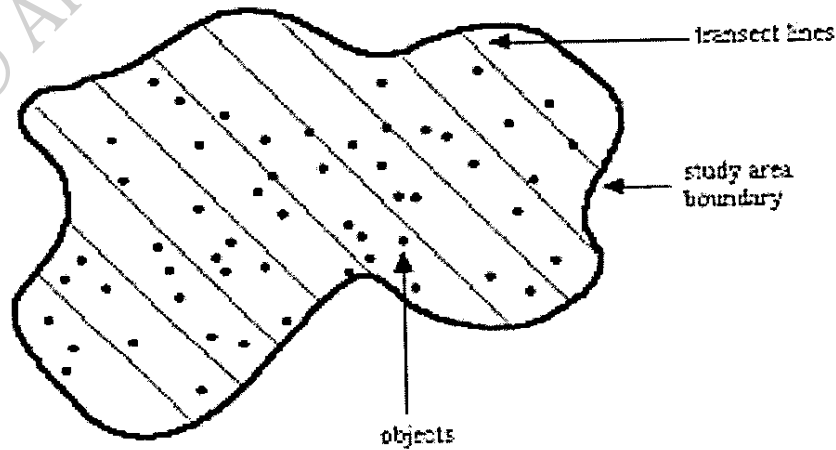
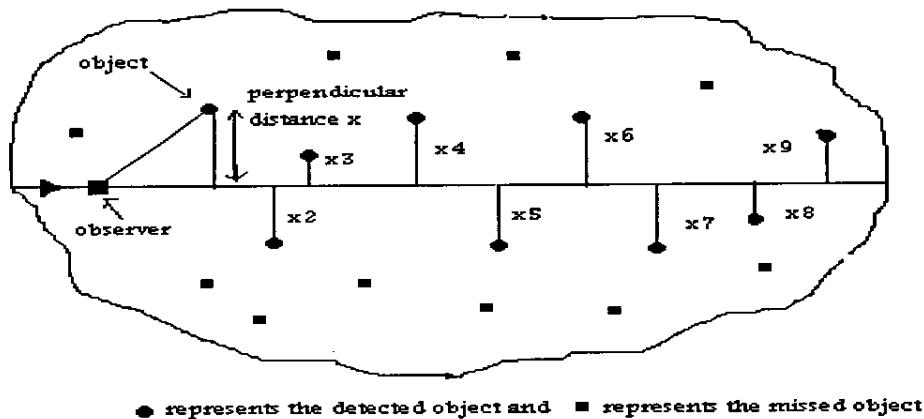


Figure 1.1. Diagram represents the line transect method.





**Figure1.2.** The perpendicular distance  $x$  in line transect method

The fundamental property of line transect sampling method is that not all objects will be detected, some objects will be missed. Moreover, objects near the transect centerline have a greater probability to be detected than objects far from the line.

The detection function  $g(x)$  represents the probability of detecting an object given that its perpendicular distance is  $x$ . The assumptions on  $g(x)$  are (Burnham et al., 1980)

- 1-  $g(x)$  must be monotonically decreasing.
- 2- Objects directly on the transect line will never be missed (i.e.,  $g(0) = 1$ ).

Suppose that the observer detected  $n$  objects with perpendicular distances  $X_1, X_2, \dots, X_n$ . These perpendicular distances form a random sample of size  $n$  that follows a specific pdf  $f(x)$ .

Burnham and Anderson (1976) introduced the basic relationship between  $g(x)$  and  $f(x)$ , which is given by

$$f(x) = \frac{g(x)}{\int_0^w g(x) dx}, \quad 0 \leq x \leq w \quad (1.1)$$

where  $w$  is a truncated distance. They gave the fundamental relationship between

$f(0) = \left[ \int_0^w g(x) dx \right]^{-1}$  and the population abundance,  $D$ , which can be expressed as

$$D = \frac{E(n)f(0)}{2L}, \quad (1.2)$$

where  $n$  is the number of detected objects,  $E(n)$  is the expected value of  $n$ , and  $L$  is the length of the transect lines.

The estimation of  $D$  can be accomplished via the estimation of  $f(0)$  by (Burnham et al., 1980)

$$\hat{D} = \frac{n \hat{f}(0)}{2L}, \quad (1.3)$$

where  $\hat{f}(0)$  is the estimator of  $f(x)$  evaluated on the transect line (i.e., at  $x = 0$ ). As Equation (1.3) demonstrates, the crucial problem in line transect sampling is to estimate  $f(0)$  by  $\hat{f}(0)$ . This leads us to obtain the estimation of density  $D$  by  $\hat{D}$ . Moreover, the estimation of  $D$  is equivalent to estimate the number of objects  $N$  in a specific known area  $A$ . Therefore, the estimation of  $N$  can be accomplished by using  $\hat{N} = A \hat{D}$ .

The estimator  $\hat{f}(0)$  can be obtained by using a parametric approach or a nonparametric approach. The first one assumed that the form of the probability density function  $f(x; \theta)$  is known with unknown parameter  $\theta$  ( $\theta$  may be a vector). A good statistical method – such as the maximum likelihood method - can be used now to estimate  $\theta$  and then  $f(0; \theta)$ . While the parametric method performs well

when the form of  $f(x; \theta)$  is chosen correctly, its performance is not satisfactory otherwise (See for example Buckland et. al., 2001). As an alternative method to the parametric approach, recent works has focused on employing the nonparametric approach to estimate the parameter  $f(0)$  and consequently the parameter  $D$  or  $N$ . A popular method is the kernel method which becomes an important tool in wildlife sampling (See for example, Chen, 1996, Mack and Quang, 1998 and Eidous, 2005a).

## 1.2 The Classical Kernel Density Estimation

The form of the classical kernel estimator  $\hat{f}(x)$  of  $f(x)$  based on a random sample  $X_1, X_2, \dots, X_n$  is given by (Silverman, 1986)

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right), \quad -\infty < x < \infty \quad (1.4)$$

where  $h$  is called a smoothing (or bandwidth) parameter, and  $K$  is the kernel function assumed to be symmetric and satisfies,

$$\int_{-\infty}^{\infty} K(t) dt = 1, \quad \int_{-\infty}^{\infty} tK(t) dt = 0, \quad \int_{-\infty}^{\infty} t^2 K(t) dt = C \neq 0 < \infty. \quad (1.5)$$

Under the assumptions that  $h \rightarrow 0$  and  $nh \rightarrow \infty$  when  $n \rightarrow \infty$ , the bias and variance of  $\hat{f}(x)$  are (Silverman, 1986)

$$\begin{aligned} \text{Bias}\left(\hat{f}(x)\right) &= E \hat{f}(x) - f(x) \\ &= h^2 f''(x) \int_{-\infty}^{\infty} t^2 K(t) dt + o(h^2), \end{aligned} \quad (1.6)$$

and

$$\text{Var}\left(\hat{f}(x)\right) = \frac{f(x)}{nh} \int_{-\infty}^{\infty} K^2(t) dt + o\left(\frac{1}{nh}\right). \quad (1.7)$$

We note that  $\hat{f}(x)$  is a consistent estimator of  $f(x)$ , since Bias  $\rightarrow 0$  and variance  $\rightarrow 0$  when  $h \rightarrow 0$  and  $nh \rightarrow \infty$ . The convergence rate of bias is  $O(h^2)$  and the convergence rate of variance is  $O\left(\frac{1}{nh}\right)$ . The kernel estimator (1.4) can be utilized to estimate the parameter  $f(0)$  when the data are collected via line transect technique. However, some corrections are needed before we can use Estimator (1.4) because the range of perpendicular distance is defined on the positive real line. This issue is explained in Chapter two. By referring to Estimator (1.4), there are two quantities under the user control. The first quantity is the kernel function  $K$  and the other one is the smoothing parameter  $h$ . The using of Estimator (1.4) requires to identify these two quantities, which are discussed in the following two subsections.

### 1.2.1 Kernel Functions

There are infinitely many kernel functions that satisfy the condition (1.5). Some of these functions are given by Silverman (1986). In this subsection we present the name and the formula of each function as given below.

Epanechnikov kernel :

$$K_e(t) = \begin{cases} \frac{3}{4\sqrt{5}} \left(1 - \frac{1}{5}t^2\right), & |t| \leq \sqrt{5}, \\ 0, & o.w \end{cases}$$

Biweight kernel :

$$K_B(t) = \begin{cases} \frac{15}{16} (1-t^2)^2, & |t| < 1, \\ 0, & o.w \end{cases}$$

Triangular kernel :

$$K_T(t) = \begin{cases} 1 - |t|, & |t| < 1 \\ 0, & \text{o.w.} \end{cases}$$

Gaussian kernel :

$$K_G(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2), \quad -\infty < t < \infty,$$

Rectangular kernel :

$$K_R(t) = \begin{cases} \frac{1}{2}, & |t| < 1 \\ 0, & \text{o.w.} \end{cases}$$

Hodges and Lehman (1956) showed that, for fixed  $h$ , the Epanechnikov kernel  $K_e(t)$  minimize the mean integral square error (*MISE*) of  $\hat{f}_k(x)$ . Table (1.1) below is taken from Silverman (1986), which gives the efficiency of each kernel function with respect to Epanechnikov kernel.

**Table 1.1.** The Efficiency (EFF) for several kernel functions (Silverman, 1986).

Kernel	EFF
$K_e(t)$	1
$K_B(t)$	0.9939
$K_T(t)$	0.9859
$K_G(t)$	0.9512
$K_R(t)$	0.9295

The main message of Table (1.1) is that there is very little loss of efficiency when adopting the different kernels in the estimator (1.4) and they all contribute very

similar amount on the basis of *MISE*. Throughout this thesis, the Gaussian kernel function is always used unless otherwise is stated.

### 1.2.2 Smoothing Parameter $h$

The smoothing (or bandwidth) parameter  $h$  controls the smoothness of the fitted density curve. It is well known that the kernel estimator (1.4) is very sensitive to the choice of  $h$ . Large  $h$  produces a smoother estimator with a large bias and small variance, while small  $h$  gives a rougher estimator with small bias and large variance (see Wand and Jones, 1995).

There are many different methods in the literature to choose the value of  $h$ ; most of them are developed based on the minimization of the asymptotic mean integral square error (MISE) or the asymptotic mean square error (MSE) with respect to  $h$ . The formulas of MISE and MSE are given as follows (Silverman, 1986)

$$MISE(\hat{f}(x)) = \int_{-\infty}^{\infty} [E\hat{f}(x) - f(x)]^2 dx + \int_{-\infty}^{\infty} \text{var} \hat{f}(x) dx \quad (1.8)$$

and

$$MSE(\hat{f}(x)) = [\text{bias}(\hat{f}(x))]^2 + \text{var}(\hat{f}(x)). \quad (1.9)$$

The asymptotic bias and asymptotic variance of  $\hat{f}_k(x)$  are

$$\text{Bias}(\hat{f}(x)) = \frac{1}{2} h^2 f''(x) C$$

and

$$\text{Var}(\hat{f}(x)) = \frac{1}{nh} f(x) \int_{-\infty}^{\infty} K^2(t) dt$$

where  $C$  is as defined in (1.5). Therefore equations (1.8) and (1.9) become

$$MISE(\hat{f}(x)) = \frac{1}{4} h^4 C^2 \int_{-\infty}^{\infty} (f''(x))^2 dx + \frac{1}{nh} \int_{-\infty}^{\infty} K^2(t) dt \quad (1.10)$$

and

$$MSE(\hat{f}(x)) = \frac{1}{4} h^4 f''^2(x) C^2 + \frac{1}{nh} f(x) \int_{-\infty}^{\infty} K^2(t) dt \quad (1.11)$$

respectively. By differentiate equations (1.10) and (1.11) with respect to  $h$  and equating the resulting equations to zero, we obtain the value of  $h$ . The optimal formulas of  $h$  that minimize the asymptotic MISE and MSE are respectively,

$$h_{MISE} = C^{-\frac{2}{5}} \left\{ \int_{-\infty}^{\infty} K^2(t) dt \right\}^{\frac{1}{5}} \left\{ \int_{-\infty}^{\infty} f''^2(x) dx \right\}^{-\frac{1}{5}} n^{-\frac{1}{5}} \quad (1.12)$$

and

$$h_{MSE} = C^{-\frac{2}{5}} \left\{ f(x) \int_{-\infty}^{\infty} K^2(t) dt \right\}^{\frac{1}{5}} \left\{ f''^2(x) \right\}^{-\frac{1}{5}} n^{-\frac{1}{5}}. \quad (1.13)$$

The two formulas (1.12) and (1.13) are somewhat disappointing since they show that the optimal  $h$  itself depends on unknown functions  $f(x)$  and  $f''(x)$ . However, formulas (1.12) and (1.13) show an important result:  $h \rightarrow 0$  and  $nh \rightarrow \infty$  as  $n \rightarrow \infty$ . Practically, formulas (1.12) and (1.13) can be adopted by estimate  $f(x)$  and  $f''(x)$  non parametrically and by using the kernel estimator (1.4) and then compute  $h$  iteratively. On the other hand,  $h$  can be obtained parametrically by assuming a reasonable form for  $f(x)$ . A common choice for  $f(x)$  in line transect scheme is the half normal distribution. This issue is discussed in more details in Chapter (2).

### 1.3 Literature Review

Hayne (1949) provided the first estimator that has a rigorous justification in statistical theory. While his method rests on only the use of sighting distances  $r_i$ , the critical assumption made can be tested using the sighting angle distances  $\theta_i$ . But this method is poor if  $\bar{\theta}$  is not approximately  $32.7^\circ$ .

After Hayne's estimator, almost no significant theoretical advances appeared until Gates et al. (1968). They assumed that  $f(x)$  has a negative exponential form, i.e.  $f(x) = a \exp(-ax)$ , where  $a$  is an unknown parameter to be estimated. But since the assumed detection function was very restrictive and might be inappropriate, the resulting estimator of density could be severely biased. Hemingway (1971) suggested the half normal model to estimate  $f(0)$ . This model performs better than the negative exponential model since it satisfies the usual condition that is known in line transect sampling as the shoulder condition assumption. Pollock (1978) suggested to use the exponential power model, which incorporates the negative exponential and the half normal models as special cases. Hayes and Buckland (1983) proposed the hazard rate model to fit line transect data. This model was further studied by Buckland (1985), who also suggested the Hermite polynomial model for grouped line transect data. Karunamuni and Quinn (1995) developed a Bayesian model for estimating the density of a closed animal population from data obtained by the line transect method. A Bayesian estimator is constructed with respect to gamma prior density. Most of the above models used the maximum likelihood method to estimate the corresponding parameter(s) and they perform well when the model is selected to be appropriate for the line transect data at hand. Because the above approaches assume that the form of



the probability density function  $f(x)$  is known, they are known as parametric methods.

However, parametric methods are accurate if the model is properly selected, although they can show poor performance otherwise (Buckland et al., 2001). As an alternative method to the parametric approach, most works have focused on employing the nonparametric method to estimate the parameter  $f(0)$  and consequently the parameter  $D$  or  $N$ . Popular nonparametric methods are the Fourier series method (Burnham *et al.*, 1980) and the kernel method (Silverman, 1986). Burnham *et al.* (1980) published a major monograph on line transect sampling theory and application. Their work provided a review of previous methods, gave guidelines for field use, and identified a small class of estimators that seemed to be useful. Theoretical and numerical studies led them to recommend the use of estimators based on the Fourier series (Crain et al. 1978).

In recent years, researchers have turned their attention to nonparametric kernel method, which becomes an important tool in wildlife sampling. Some initial efforts in applying the kernel method to line transect data made by Buckland (1992), Chen (1996) and Mack and Quang (1998). Mack (1998) used the kernel method to perform a test concerning the validity of the shoulder condition assumption. Mack (2002) considered some methods of bias correction when the kernel method is used in constructing confidence intervals for wildlife abundance based on transect data. Gerard and Schucany (2002) investigated a method for combining estimators from individual transects when each transect has sufficient data to support estimation with the kernel method. It is based on a minimization of the asymptotic mean square error of a linear combination of the individual population density estimator. Eidous (2005b) proposed some methods to improve the performance of the kernel estimator using line

transect data. As a nonparametric method Eidous (2005a) introduced the histogram estimator and investigated its performances using line transect data. Another adaptation of the histogram estimator is suggested and investigated by Eidous (2011). He developed this estimator under the assumption that the shoulder condition assumption is not valid. Barabesi (2000) proposed a semi-parametric technique based on local parametric density estimation. Finally, Eidous and Alshakhateh (2011) investigated the properties of a semi-parametric estimator for  $f(0)$  that combines between the kernel estimator and a specific parametric estimator. The parametric estimator is chosen to be half-normal or negative exponential based on testing the shoulder condition assumption.

#### 1.4 Thesis Objectives

Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  represents the perpendicular distances follow the probability density function  $f(x)$ , where  $f(x)$  and the detection function  $g(x)$  are related as given in Formula (1.1). The crucial problem in line transect sampling is to estimate  $f(0)$ , which leads us to estimate the population abundance  $D$  and the total number of objects  $N$ . As such, this thesis aims to present some existing nonparametric estimators of  $f(0)$  for both cases; when the shoulder condition is assumed to be valid and when it is invalid and to compare between them. For each case, we aim to propose a new estimator for  $f(0)$  and to compare its performance with those existing estimators aiming to identify the most promising estimator(s). In addition, we aim to suggest new estimators that combine the best estimators in each case. These later suggested estimators –as we expected- can be applied when the shoulder condition is valid and when it is violated because we need to perform a test to check the validity of the shoulder condition before we decide

which estimator should be used. Finally, we interest to apply and to study the performances of the different estimators on real data set.

### 1.5 Thesis Outlines

This thesis is divided into five chapters. Chapter One presents an introduction about line transect method and classical kernel density estimation.

The rest of this thesis is structured as follows. In Chapter Two, the shoulder condition is discussed and some known estimators of  $f(0)$  together with a new proposed estimator when the model of the data is assumed to satisfy the shoulder condition are introduced. Also, their performances are studied and compared via simulation technique to identify the most promising estimator(s).

Chapter Three deals with some existing estimators that are developed when the data model does not satisfy the shoulder condition. Another new estimator is proposed in this chapter and its asymptotic properties are also studied. In addition, the performances of the different estimators are studied and compared aiming to identify the best one.

In Chapter Four, we proposed new estimators, each one combines between two estimators; the best one when the shoulder condition is satisfied, and the best one when the shoulder condition is not satisfied. A comparison between these new estimators and the semi-parametric estimator is also performed in this chapter.

Finally, Chapter Five gives numerical example of real data at which the different estimators of this thesis are applied. Some concluding remarks and comments are also given.

## CHAPTER TWO

### SOME ESTIMATORS OF $f(0)$ WITH THE SHOULDER CONDITION

#### 2.1 Introduction

In this chapter, three well known nonparametric estimators for  $f(0)$  that are developed in the literature under the assumption that the shoulder condition is valid are stated. A new proposed estimator is also given under this condition. A comparison study via simulation technique is performed aiming to study the performances of the proposed estimator compared with other estimators aiming to identify the best estimator of them.

#### 2.2 The Shoulder Condition and Classical Kernel Estimator of $f(0)$

To estimate the population abundance  $D$ , we need to estimate  $f(0)$  as equations (1.2) and (1.3) stated. The perpendicular distances that are obtained by applying line transect sampling experiment are non-negative. Therefore, some corrections on the classical kernel estimator (Eq. 1.4) are needed to estimate the parameter  $f(0)$  (i.e.  $f(x)$  at the end point of its support). This issue can be illustrated as follows :

Let  $X_1, X_2, \dots, X_n$  be a random sample of perpendicular distances from  $f(x)$ ,  $x \geq 0$ .

using directly the classical kernel estimator (1.4) to estimate  $f(0)$  we get

$$f^*(0) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{0-X_i}{h}\right). \quad (2.1)$$

The expected value of  $f^*(0)$  is

$$\begin{aligned} E(f^*(0)) &= \frac{1}{h} EK\left(\frac{X}{h}\right) \\ &= \int_0^{\infty} K(u)f(hu)du . \end{aligned}$$

Expanding  $f(hu)$  around zero by using Taylor's series, then we obtain

$$\begin{aligned} E(f^*(0)) &= \int_0^{\infty} K(u) \left[ f(0) + huf'(0) + \frac{h^2u^2}{2} f''(0) + \dots \right] du \\ &= f(0) \int_0^{\infty} K(u)du + hf'(0) \int_0^{\infty} uK(u)du + \frac{h^2 f''(0)}{2} \int_0^{\infty} u^2 K(u)du + \dots \quad (2.2) \end{aligned}$$

Since  $\int_0^{\infty} K(u)du = \frac{1}{2}$ , then estimator (2.1) is not even a consistent estimator for  $f(0)$ .

To correct this problem (known in the literature as the boundary effect), it is obvious

that we need to divide the estimator  $f^*(0)$  by  $\int_0^{\infty} K(u)du = \frac{1}{2}$ . Thus, the consistent

estimator of  $f(0)$  is

$$\hat{f}_k(0) = \frac{2}{nh} \sum_{i=1}^n K\left(\frac{X_i}{h}\right) . \quad (2.3)$$

The bias of  $\hat{f}_k(0)$  is

$$\text{Bias}(\hat{f}_k(0)) = 2hf'(0) \int_0^{\infty} uK(u)du + h^2 f''(0) \int_0^{\infty} u^2 K(u)du + \dots . \quad (2.4)$$

The convergence rate of bias is now  $O(h)$ , slower than the usual rate  $O(h^2)$  of the

classical kernel estimator as given by (1.6). To obtain  $O(h^2)$  bias of  $\hat{f}_k(0)$ , we need

to assume that  $f'(0) = 0$  because  $\int_0^{\infty} uK(u)du \neq 0$ . Since  $x \geq 0$ , then we assumed

$f'(0^+) = 0$ . The condition  $f'(0^+) = 0$  is known in line transect literature as the shoulder condition assumption, which means that the probability of detecting an object in a narrow area around the centerline remains certain.

The variance of estimator (2.3) is (Chen, 1996)

$$\text{Var}(\hat{f}_k(0)) = \frac{4f(0)}{nh} \int_0^\infty K(u)^2 du + o\left(\frac{1}{nh}\right), \quad (2.5)$$

which indicates that the convergence rate for variance of  $\hat{f}_k(0)$  is  $O\left(\frac{1}{nh}\right)$ , the same rate as that of Estimator (1.4).

### 2.3 Testing the Shoulder Condition

There are two methods in the literature that were proposed to test the shoulder condition assumption

$$H_0 : f \in F_0 \quad \text{vs.} \quad H_1 : f \in F \setminus F_0, \quad (2.6)$$

where  $F_0$  = the class of all pdf that satisfy  $f'(0^+) = 0$  and  $F$  = the class of all pdfs that are differentiable at 0.

The first method ( parametric method ) was proposed by Zhang (2001). Let  $X_1, X_2, \dots, X_n$  be a random sample of perpendicular distances with common pdf  $f(x)$ . According to Zhang (2001), reject  $H_0$  for large values of

$$Z = \frac{\sqrt{\sum_{i=1}^n X_i^2}}{\sum_{i=1}^n X_i}.$$

Zhang constructed a table of critical values for the sampling distribution of  $Z$  with respect to different small sample sizes by using Monte Carlo simulations. Borgoni

and Quatto (2011) gave an approximate formula for the critical values when  $n$  is large.

The second method was proposed by Mack (1998) to perform test (2.6). According to Mack (1998) derivations,  $H_0$  is reject (i.e., the shoulder condition is not satisfied) if  $|T| > -Z_{\alpha/2}$ , where  $Z_{\alpha/2}$  represents the  $\alpha/2^{\text{th}}$  quantile of the standard normal distribution. The test statistics  $T$  is defined by

$$T = f'(0) \sqrt{\frac{nb^3}{2\hat{f}_k(0)}} \quad , \quad (2.7)$$

where  $f'(0)$  is estimated by

$$\hat{f}'(0) = \frac{[F_n(2b) - 2F_n(b)]}{b^2} \quad ,$$

$$b = \hat{\sigma} n^{-\frac{1}{4}} \quad ,$$

$\hat{\sigma}$  is given in section (2.4.1) and  $F_n(u)$  is the empirical cumulative distribution function defined by

$$F_n(u) = \frac{\#x_i \in [0, u]}{n} \quad .$$

The p-value for the above test is

$$\begin{aligned} p\text{-value} &= 2pr(Z < -|T|) \\ &= 2\Phi(-|T|) \quad . \end{aligned} \quad (2.8)$$

where  $\Phi$  is the standard normal distribution function. The p-value indicates how strong  $H_0$  is supported by the data.

## 2.4 Estimators of $f(0)$ when $f'(0^+) = 0$

Here we presented three well known estimators for  $f(0)$ , which developed under the assumption that  $f'(0^+) = 0$ . In addition, a new estimator is proposed and studied.

### 2.4.1 Classical Kernel Estimator

Chen (1996) was the first one who suggested the classical kernel estimator  $\hat{f}_k(0)$  (Eq. 2.3) to estimate  $f(0)$ . He derived the asymptotic properties of  $\hat{f}_k(0)$  under the assumption that  $f'(0^+) = 0$ . The bias and the variance of  $\hat{f}_k(0)$  are given by equation (2.4) and (2.5), respectively.

The bandwidth parameter  $h$  controls the smoothness of the fitted density curve. The optimal formula of  $h$  can be obtained by minimizing the asymptotic mean square error (AMSE) of  $\hat{f}_k(0)$

$$\begin{aligned} AMSE(\hat{f}_k(0)) &= Bias^2(\hat{f}_k(0)) + Var(\hat{f}_k(0)) \\ &= h^4 f''^2(0) \left( \int_0^\infty u^2 K(u) du \right)^2 + \frac{4f(0)}{nh} \int_0^\infty K^2(u) du. \end{aligned} \quad (2.9)$$

Differentiate  $AMSE(\hat{f}_k(0))$  with respect to  $h$  and then equating to zero, we get

$$h = \left( \frac{f(0) \int_0^\infty K^2(u) du}{f''(0)^2 \left( \int_0^\infty u^2 K(u) du \right)^2} \right)^{\frac{1}{5}} n^{\frac{1}{5}}. \quad (2.10)$$

If the kernel function  $K$  is assumed to be Gaussian kernel (see subsection, 1.2.1) and  $f(x)$  is half normal distribution with scale parameter  $\sigma^2$  then we get



$h = 0.933 \sigma n^{-1/5}$ , where  $\sigma$  can be replaced by  $\hat{\sigma} = \sqrt{\sum_{i=1}^n x_i^2 / n}$  is the maximum

likelihood estimator of  $\sigma$  under half-normal distribution. It is worthwhile to mention here that there are many methods to select the smoothing parameter  $h$ . Most of them are complicated in their computations. Gerard and Schucany (1999) compared between some of these methods and reported that the selection of  $h$  by assuming the half-normal distribution is very acceptable for line transect sampling technique.

#### 2.4.2 Barabesi Estimator

Barabesi (2001) proposed a new estimator for  $f(0)$  by using line transect method based on local parametric estimation technique. In this technique he assumed  $m(x, \theta, \gamma) = \theta g(x, \gamma)$  is a family of key model, where  $g(x, \gamma)$  is a monotone decreasing density function satisfies  $g'(0, \gamma) = 0$ . The formula of the local parametric estimator for  $f(0)$  is given by

$$\hat{f}_h(0) = \hat{\theta}_h g(0, \hat{\gamma}),$$

where  $\hat{\theta}_h$  is the estimator of  $\theta$ . Hjort and Jones (1996) showed that  $\theta$  can be estimated by solving the following local equation

$$\frac{1}{n} \sum_{i=1}^n K_h(x_i) v(x_i, \theta) - \int_0^{\infty} K_h(t) v(t, \theta) m(t, \theta, \gamma) dt = 0, \quad (2.11)$$

where  $K_h(u) = \frac{1}{h} K\left(\frac{u}{h}\right)$ ,  $K(u)$  is a kernel function and  $v(u, \theta)$  is a weight function

Barabesi assumed in his estimator that  $v(u, \theta) = 1$  and the vector of parameters  $\gamma$  is initially estimated by  $\hat{\gamma}$  which based on a likelihood estimator.

under the assumption  $nh \rightarrow \infty$  and  $h \rightarrow 0$  as  $n \rightarrow \infty$  the bias and variance are given by

$$Bias(\hat{f}_h(0)) = \frac{ha_1}{a_0} f'(0) + \frac{h^2 a_2}{2a_0} (f''(0) - g''(0, \theta_0)) + o(h^2) \quad (2.12)$$

and

$$Var(\hat{f}_h(0)) = \frac{f(0)b}{nha_0^2} + o\left(\frac{1}{nh}\right) \quad (2.13)$$

where  $a_1 = \int u^1 K(u) du$  and  $b = \int K(u)^2 du$ .

By substitute  $m(x, \theta, \gamma)$  in equation (2.11) and solve it to get  $\hat{\theta}_h$ , Barabesi get the local parametric estimator for  $f(0)$ , which is

$$\begin{aligned} \hat{f}_h(0) &= \hat{\theta}_h g(0, \hat{\gamma}) \\ &= \frac{\tilde{f}_h(0)g(0, \hat{\gamma})}{\int K_h(t)g(t, \hat{\gamma})dt} \end{aligned} \quad (2.14)$$

Barabesi takes the simple case when  $g(x, \gamma)$  is a half normal and the kernel function to be the Gaussian. So, his estimator (2.14) becomes

$$\hat{f}_B(0) = \hat{f}_k(0) \left[ \frac{h^2}{\hat{\gamma}^2} + 1 \right]^{\frac{1}{2}}, \quad (2.15)$$

where  $\hat{f}_k(0)$  is given by (2.3) and  $\hat{\gamma}^2 = \frac{\sum_{i=1}^n x_i^2}{n}$ . The estimator (2.15) is simply the

usual classical kernel estimator (2.3) corrected by a factor  $\left[ \frac{h^2}{\hat{\gamma}^2} + 1 \right]^{\frac{1}{2}}$ , which

converges to one as  $h \rightarrow 0$ .

### 2.4.3 Histogram Estimator

Eidous (2005b) introduced a nonparametric frequency histogram method using line transect data. His estimator is given by

$$\hat{f}_E(0) = \frac{1}{nh} \sum_{i=1}^n I_{[0, h)}(X_i), \quad (2.16)$$

where  $h$  is called the bin-width of the histogram estimator, and  $I_B(t)$  is an indicator function of a real set  $B$ .

The bias of  $\hat{f}_E(0)$  is

$$\text{Bias}(\hat{f}_E(0)) = \frac{h}{2} f'(0) + \frac{h^2}{6} f''(0) + o(h^2).$$

The bias convergence rate of  $\hat{f}_E(0)$  is  $O(h)$  if  $f'(0^+) \neq 0$ , while it is  $O(h^2)$  when  $f'(0^+) = 0$ . The variance of estimator (2.16) is

$$\text{Var}(\hat{f}_E(0)) = \frac{f(0)}{nh} + o\left(\frac{1}{nh}\right),$$

which indicates that the convergence rate for variance of  $\hat{f}_E(0)$  is  $O\left(\frac{1}{nh}\right)$ , the same

rate as the classical kernel estimator. The AMSE of  $\hat{f}_E(0)$  is,

$$\text{AMSE}(\hat{f}_E(0)) = \frac{h^4}{36} f''^2(0) + \frac{f(0)}{nh}.$$

The value of  $h$  that minimizing the AMSE is given by

$$h = \left( \frac{9f(0)}{f''^2(0)} \right)^{\frac{1}{5}} n^{-\frac{1}{5}}. \quad (2.17)$$

Assume that the underlying probability density function  $f(x)$  is half normal with scale parameter  $\sigma^2$  then from (2.17) we find  $h \cong 1.624\hat{\sigma} n^{-\frac{1}{5}}$ , where  $\hat{\sigma}$  is the maximum likelihood estimator for  $\sigma$  (see Subsection 2.4.1).

#### 2.4.4 The Proposed Estimator when $f'(0^+) = 0$

Let  $X_1, X_2, \dots, X_n$  be a random sample of perpendicular distances of size  $n$ . Under the assumption that  $f'(0^+) = 0$ , we propose the following estimator for  $f(0)$ ,

$$\hat{f}_{p1}(0) = \frac{2}{nh} \sum_{j=1}^4 \sum_{i=1}^n r_j K\left(\frac{X_i}{jh}\right), \quad (2.18)$$

where  $r_1 = 47/50$ ,  $r_2 = 127/200$ ,  $r_3 = -91/150$  and  $r_4 = 61/400$ .

Let  $D_v = \sum_{j=1}^4 j^v r_j$ , then  $D_1 = 1$ ,  $D_2 = 0.46$ ,  $D_3 = -0.6$ . Also let

$$T(r_1, \dots, r_4) = \int_0^{\infty} K^2(u) du \sum_{j=1}^4 j r_j^2 + 2 \sum_{j=1}^3 \sum_{l=j+1}^4 r_j r_l \int_0^{\infty} K(u/j) K(u/l) du$$
 and  $K(t)$  is the

density of  $N(0,1)$ , then  $T(r_1, \dots, r_4) = 0.1847$ . The optimal value of  $h$  can be

obtained by minimizing the *AMSE* of  $\hat{f}_{p1}(0)$ , which gives  $h = 1.206 \hat{\sigma} n^{-\frac{1}{5}}$  when

$f(x)$  is assumed to be half normal with scale parameter  $\sigma^2$ . The illustrations for the

use of the above notations are given in Section (3.4). The asymptotic properties of

Estimator (2.18) are stated in the following lemma.

**Lemma (2.1).** Suppose that  $f(x)$  is defined on  $[0, \infty)$  and has a continuous second

derivative at  $x = 0$ . Under the assumption that  $h \rightarrow 0$  and  $nh \rightarrow \infty$  as  $n \rightarrow \infty$ , the

expected value and the variance of  $\hat{f}_{p1}(0)$  are,

$$\begin{aligned} E(\hat{f}_{p1}(0)) &= f(0)D_1 + 2hf'(0)D_2 \int_0^{\infty} uK(u)du + h^2 f''(0)D_3 \int_0^{\infty} u^2 K(u)du + o(h^2) \\ &\cong f(0) - 0.6h^2 f''(0) \int_0^{\infty} u^2 K(u)du \end{aligned} \quad (2.19)$$

and

$$\begin{aligned} \text{var}(\hat{f}_{P_1}(0)) &= \frac{4f(0)}{nh} T(r_1, \dots, r_4) + o(n^{-1}h^{-1}). \\ &\cong \frac{0.7388f(0)}{nh} \end{aligned} \quad (2.20)$$

Note that, because  $D_1 = 1$  then  $\hat{f}_{P_1}(0)$  is asymptotically ( $h \rightarrow 0$  as  $n \rightarrow \infty$ ) unbiased estimator for  $f(0)$  and since  $f'(0^+) = 0$  then the convergence rate for bias of  $\hat{f}_{P_1}(0)$  is  $O(h^2)$ . Also note that the variance of  $\hat{f}_{P_1}(0)$  converges to zero as  $nh \rightarrow \infty$  when  $n \rightarrow \infty$ . More details and illustrations about the results of this section are given in sections (3.4) and (3.5) of Chapter (3) at which the proof of Lemma (2.1) is stated.

## 2.5 Simulation Design

To compare among the performances of the different estimators, a simulation study was performed. The data are simulated from densities that satisfy  $f'(0^+) = 0$  (e.g. half normal) and from densities that do not satisfy  $f'(0^+) = 0$  (e.g. negative exponential). The later case is considered to investigate the performances of the four estimators,  $\hat{f}_k(0)$ ,  $\hat{f}_B(0)$ ,  $\hat{f}_E(0)$  and  $\hat{f}_{P_1}(0)$  when the shoulder condition is violated, while their mathematical derivations assume the validity of this condition. The smoothing parameter  $h$  for the different estimators is computed by using the formula  $h = A \hat{\sigma} n^{-\frac{1}{5}}$ , where  $A = 0.933$  for  $\hat{f}_k(0)$  and  $\hat{f}_B(0)$ ;  $A = 1.624$  for  $\hat{f}_E(0)$ ; and  $A = 1.206$  for the proposed estimator  $\hat{f}_{P_1}(0)$ .

All the results are based on simulated 1000 samples of sizes  $n = 50, 100, 200$ . The data generated from three different families of models which are commonly used in

line transect studies (see Barabesi. 2001 and Eidous, 2009). The first model is the exponential power (EP) family (Pollock,1978)

$$f(x) = \frac{1}{\Gamma\left(1 + \frac{1}{\beta}\right)} \exp(-x^\beta), \quad x \geq 0, \beta \geq 1, \quad (2.21)$$

with detection function  $g(x) = \exp(-x^\beta)$ . The hazard rate (HR) family (Hayes and Buckland,1983)

$$f(x) = \begin{cases} \frac{1}{\Gamma\left(1 - \frac{1}{\beta}\right)} (1 - \exp(-x^{-\beta})) & \text{if } x > 0 \\ \frac{1}{\Gamma\left(1 - \frac{1}{\beta}\right)} & \text{if } x = 0 \end{cases}, \quad \beta > 1, \quad (2.22)$$

with detection function  $g(x) = (1 - \exp(-x^{-\beta}))$ , and the beta (BE) model (Eberhardt,1968)

$$f(x) = (1 + \beta)(1 - x)^\beta, \quad 0 \leq x < 1 \quad \beta \geq 0, \quad (2.23)$$

with detection function  $g(x) = (1 - x)^\beta$ . In our simulation design, these three families were truncated at some distance  $w$ .

Four models were selected from EP family with parameter values  $\beta = 1.0, 1.5, 2.0, 2.5$  and corresponding truncation points given by  $w = 5.0, 3.0, 2.5, 2.0$ . (see Figure 2.1). Four models were selected from HR family with parameter values  $\beta = 1.5, 2.0, 2.5, 3.0$  and corresponding truncation points given by  $w = 20, 12, 8, 6$ . (see Figure 2.2). Moreover, four models were selected

from BE model with parameter values  $\beta = 1.5, 2.0, 2.5, 3.0$  and  $w = 1$  for all cases. (see Figure 2.3). The considered models cover a wide range of perpendicular distance probability density functions which vary near zero from spike to flat. The shoulder condition do not satisfy for BE model with different values of  $\beta$  and for EP model with  $\beta = 1$ . Also, despite the shoulder condition is satisfied for HR model, this model decreases sharply away from the original point (i.e.  $x = 0$ ) when  $\beta = 1.5$  and  $2.0$ . This case may be occur in practice when the visibility away from the transect line is not distinct due to –may be- fog, tall grass...etc.

To simulate the data from the above three families, the acceptance-rejection technique is adopted (See for example, Burnham et. al., 1980 and Ross, 1990).

For each consider estimator and for each sample size, the relative bias

$$RB = \frac{E(\hat{f}(0)) - f(0)}{f(0)},$$

and the relative mean error

$$RME = \frac{\sqrt{MSE(\hat{f}(0))}}{f(0)},$$

are computed. The results are presented in Table (2.1). For more simplicity of comparison, we present the efficiency (EFF) of the different estimators with respect to classical kernel estimator in Table (2.2),

$$EFF = \frac{MSE(\hat{f}_k(0))}{MSE(\hat{f}(0))}.$$

where  $\hat{f}(0)$  in the above formula stands for  $\hat{f}_B(0)$ ,  $\hat{f}_E(0)$  or  $\hat{f}_{P1}(0)$ . The symbols EFF1, EFF2 and EFF3 in Table (2.2) represent the efficiencies of  $\hat{f}_B(0)$ ,  $\hat{f}_E(0)$  and  $\hat{f}_{P1}(0)$  with respect to  $\hat{f}_k(0)$  respectively.

## 2.6 Comparison and Results

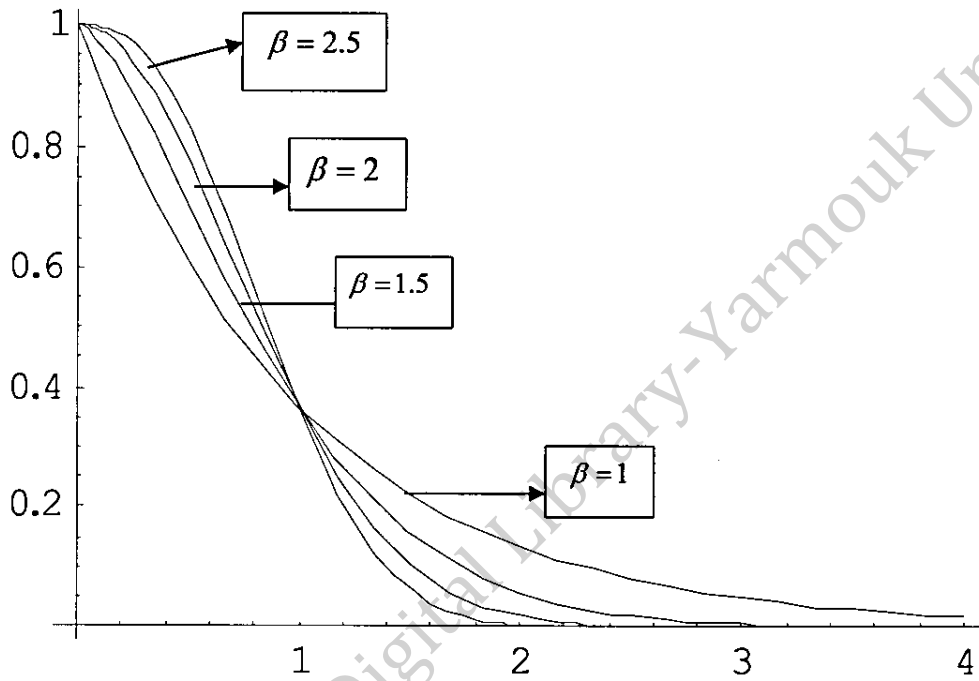
Several conclusions can be drawn based on the simulation results. Based on Table (2.1) it is clear that the *RME*s of the different estimators decrease as the sample size  $n$  increases. This coincides with the asymptotic properties of the different estimators, which assumes that  $nh \rightarrow \infty$  and  $h \rightarrow 0$  when  $n \rightarrow \infty$ . This indicates that the biases and the variances of the different estimators tend to zero as the sample size tends to be large. Tables (2.1) and (2.2) show that the two estimators  $\hat{f}_k(0)$  and  $\hat{f}_E(0)$  perform similar to each other with some preferences for  $\hat{f}_k(0)$  over  $\hat{f}_E(0)$ . This result is also obtained by Eidous (2005c) who suggested some ways to correct the bias of estimator  $\hat{f}_E(0)$ . These two estimators perform well when the shoulder condition of the simulated model is large (e.g. EP with  $\beta = 2.0, 2.5$  and HR with  $\beta = 2.5, 3.0$ ) but not decreasing rapidly away  $x=0$ . However, the simulation results are generally demonstrate the two estimators  $\hat{f}_B(0)$  and  $\hat{f}_{P1}(0)$  to be more promising.

The results of Table (2.1) indicate that Barabesi estimator  $\hat{f}_B(0)$  and the proposed estimator  $\hat{f}_{P1}(0)$  have smaller  $|RB|$ s than the classical kernel estimator  $\hat{f}_k(0)$  and the histogram estimator  $\hat{f}_E(0)$  for most considered cases.

Tables (2.1) and (2.2) show that the performances of  $\hat{f}_B(0)$  and  $\hat{f}_{P1}(0)$  are similar – in some sense - to each other with priority for  $\hat{f}_{P1}(0)$  over  $\hat{f}_B(0)$  when the shoulder

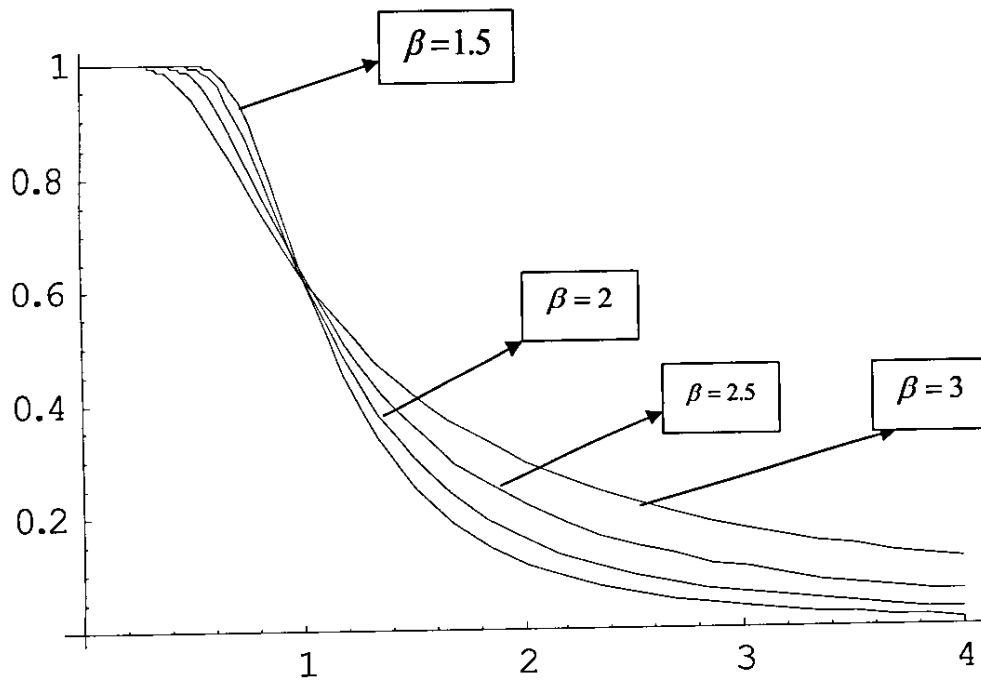


condition is not true (e.g. EP model with  $\beta = 1.0$  and BE model with different values of  $\beta$ ) and when the shoulder condition is moderate (e.g. EP model with  $\beta = 1.5$ ). Moreover, the performance of  $\hat{f}_{P_1}(0)$  is better than that of  $\hat{f}_B(0)$  for HR model with  $\beta = 1.5, 2.0$ . These two cases indicate that the shoulder condition is large but it decreases rapidly away from  $x = 0$ . The biases of the different estimators were large for these two cases compared to the biases of the other cases. Regarding the values of the EFFs in Table (2.2), it is clear that the performances of  $\hat{f}_B(0)$  and  $\hat{f}_{P_1}(0)$  are better than  $\hat{f}_k(0)$  and  $\hat{f}_E(0)$  in most cases that are considered. Based on these results, we may recommend and consider the two estimators  $\hat{f}_B(0)$  and  $\hat{f}_{P_1}(0)$  as the most promising estimators. Therefore, we will consider them again in Chapter (4) to form and to study new proposed estimators.

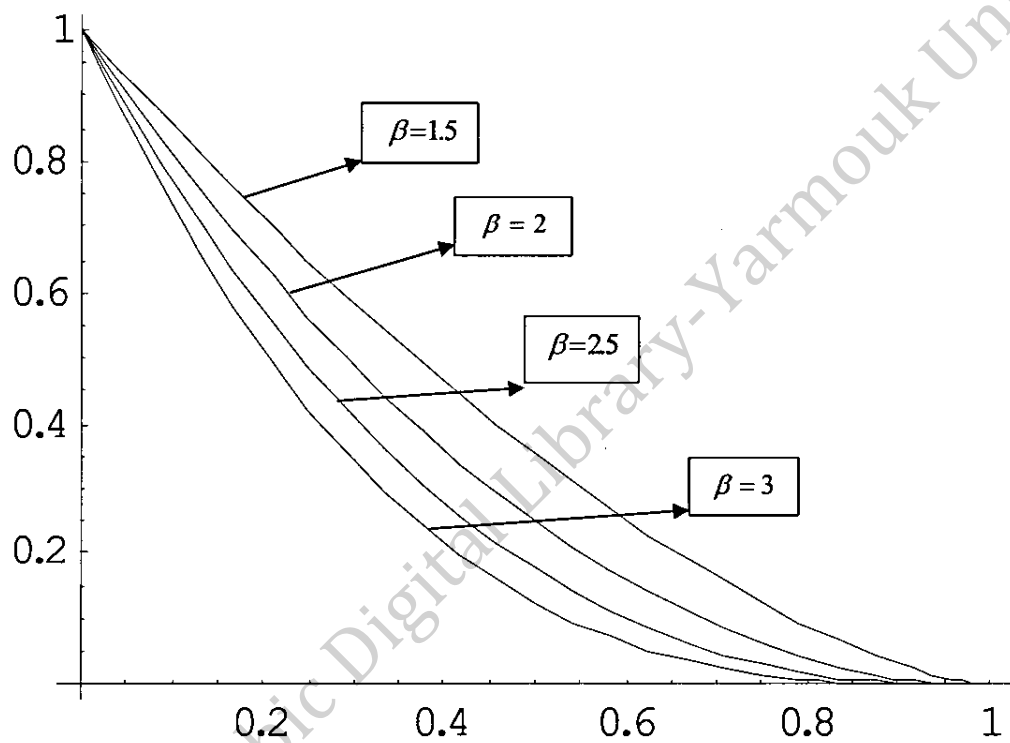


**Figure 2.1.** Represent the detection function  $g(x)$  of the Exponential power (EP) model for  $\beta = 1.0, 1.5, 2.0$  and  $2.5$

© Arabic Digital Library - Yarmouk University



**Figure 2.2.** Represent the detection function  $g(x)$  of the Hazard-rate (HR) model for  $\beta = 1.5, 2.0, 2.5$  and  $3.0$



**Figure 2.3.** Represent the detection function  $g(x)$  of the Beta (BE) model for  $\beta = 1.5, 2.0, 2.5$  and  $3.0$

**Table 2.1.** The Relative Bias (RB) and the Relative Mean Error (RME) for  $\hat{f}_k(0)$ ,  $\hat{f}_B(0)$ ,  $\hat{f}_E(0)$  and  $\hat{f}_{PI}(0)$

			$\hat{f}_k(0)$		$\hat{f}_B(0)$		$\hat{f}_E(0)$		$\hat{f}_{PI}(0)$	
$\beta$	$w$	$n$	RB	RME	RB	RME	RB	RME	RB	RME
Exponential power model										
1.0	5.0	50	-0.323	0.338	-0.264	0.284	-0.358	0.370	-0.309	0.325
		100	-0.290	0.299	-0.242	0.255	-0.322	0.330	-0.270	0.280
		200	-0.262	0.269	-0.225	0.233	-0.289	0.295	-0.234	0.241
1.5	3.0	50	-0.147	0.193	-0.072	0.154	-0.161	0.203	-0.099	0.157
		100	-0.130	0.160	-0.072	0.123	-0.144	0.172	-0.072	0.113
		200	-0.109	0.132	-0.063	0.101	-0.119	0.140	-0.048	0.086
2.0	2.5	50	-0.076	0.154	0.005	0.146	-0.085	0.155	-0.004	0.126
		100	-0.060	0.119	0.003	0.109	-0.063	0.118	0.013	0.097
		200	-0.049	0.094	0.001	0.085	-0.050	0.095	0.036	0.084
2.5	2.0	50	-0.036	0.151	0.049	0.167	0.037	0.154	0.037	0.134
		100	-0.032	0.117	0.033	0.125	-0.030	0.115	0.055	0.116
		200	-0.025	0.088	0.024	0.092	-0.022	0.087	0.074	0.105
Hazard rate model										
1.5	20.0	50	-0.363	0.382	-0.308	0.333	-0.445	0.459	-0.406	0.421
		100	-0.329	0.339	-0.284	0.297	-0.410	0.417	-0.355	0.365
		200	-0.275	0.283	-0.238	0.249	-0.353	0.359	-0.296	0.303
2.0	12.0	50	-0.225	0.257	-0.157	0.208	-0.280	0.308	-0.228	0.264
		100	-0.179	0.202	-0.124	0.159	-0.225	0.246	-0.181	0.205
		200	-0.135	0.150	-0.091	0.114	-0.167	0.181	-0.130	0.147
2.5	8.0	50	-0.102	0.159	-0.023	0.134	-0.108	0.170	-0.074	0.147
		100	-0.067	0.113	-0.005	0.097	-0.066	0.114	-0.031	0.100
		200	-0.040	0.077	0.009	0.072	-0.028	0.075	0.016	0.067
3.0	6.0	50	-0.046	0.129	0.037	0.137	-0.032	0.124	0.012	0.128
		100	-0.023	0.094	0.042	0.106	-0.005	0.090	0.044	0.100
		200	-0.007	0.076	0.043	0.091	-0.006	0.074	0.078	0.106
Beta model										
1.5	1.0	50	-0.166	0.207	-0.093	0.164	-0.182	0.218	-0.126	0.175
		100	-0.149	0.178	-0.092	0.138	-0.164	0.190	-0.095	0.135
		200	-0.127	0.148	-0.083	0.115	-0.139	0.157	-0.061	0.092
2.0	1.0	50	-0.186	0.222	-0.115	0.175	-0.207	0.238	-0.151	0.184
		100	-0.172	0.196	-0.117	0.154	-0.187	0.208	-0.123	0.150
		200	-0.153	0.168	-0.110	0.131	-0.167	0.180	-0.095	0.116
2.5	1.0	50	-0.213	0.243	-0.144	0.192	-0.232	0.257	-0.174	0.205
		100	-0.190	0.209	-0.135	0.165	-0.208	0.225	-0.145	0.167
		200	-0.176	0.189	-0.133	0.152	-0.190	0.202	-0.112	0.130
3.0	1.0	50	-0.223	0.249	-0.155	0.197	-0.243	0.266	-0.192	0.218
		100	-0.200	0.219	-0.146	0.175	-0.218	0.235	-0.162	0.182
		200	-0.180	0.192	-0.138	0.154	-0.198	0.208	-0.129	0.143

Table 2.2. The Efficiency (EFF) for  $\hat{f}_B(0)$ ,  $\hat{f}_E(0)$  and  $\hat{f}_{PI}(0)$

$\beta$	$w$	$n$	EFF1	EFF2	EFF3
Exponential power model					
1.0	5.0	50	1.190	0.909	2.082
		100	1.169	0.917	2.456
		200	1.154	0.907	2.471
1.5	3.0	50	1.242	0.951	0.677
		100	1.300	0.948	0.567
		200	1.328	0.942	0.479
2.0	2.5	50	1.061	0.991	0.419
		100	1.080	0.986	0.346
		200	1.101	0.999	0.263
2.5	2	50	0.904	1.014	0.371
		100	0.933	1.014	0.297
		200	0.930	1.010	0.215
Hazard rate model					
1.5	20.0	50	1.142	0.837	0.937
		100	1.140	0.815	0.930
		200	1.140	0.790	0.926
2.0	12.0	50	1.236	0.833	0.950
		100	1.277	0.822	0.971
		200	1.309	0.820	1.029
2.5	8.0	50	1.132	0.942	1.060
		100	1.140	0.968	1.142
		200	1.045	1.074	1.110
3.0	6.0	50	0.908	0.990	1.036
		100	0.873	1.064	0.957
		200	0.829	1.058	0.723
Beta model					
1.5	1.0	50	1.283	0.929	1.256
		100	1.270	0.951	1.361
		200	1.287	0.934	1.583
2.0	1.0	50	1.276	0.942	1.218
		100	1.277	0.931	1.312
		200	1.274	0.916	1.448
2.5	1.0	50	1.267	0.918	1.178
		100	1.275	0.931	1.253
		200	1.249	0.923	1.371
3.0	1.0	50	1.255	0.929	1.155
		100	1.253	0.929	1.218
		200	1.236	0.925	1.318

## CHAPTER THREE

### SOME ESTIMATORS OF $f(0)$ WITHOUT THE SHOULDER CONDITION

#### 3.1 Introduction

This chapter covers some existing nonparametric estimators for  $f(0)$  when the shoulder condition is not valid (i.e.  $f'(0^+) \neq 0$ ). A new estimator for  $f(0)$  when  $f'(0^+) \neq 0$  is proposed. The asymptotic statistical properties of the proposed estimator are derived. A numerical comparison study among the different estimators based on simulation technique is conducted aiming to identify the most significant one.

#### 3.2 Mack Estimator

Mack et al. (1999) introduced the boundary kernel estimator for  $f(0)$  under the assumption that the shoulder condition is not satisfied (i.e.  $f'(0^+) \neq 0$ ). Their estimator is given by

$$\hat{f}_{ME}(0) = \frac{1}{nh} \sum_{i=1}^n K^* \left( \frac{X_i}{h} \right), \quad (3.1)$$

where  $h$  is the bandwidth (or the smoothing) parameter of the estimator and  $K^*$  is a kernel function satisfying

$$\int_0^{\infty} K^*(u) du = 1, \quad \int_0^{\infty} u K^*(u) du = 0 \quad \text{and} \quad \int_0^{\infty} u^2 K^*(u) du = d \neq 0.$$

what the assumption about  $K^*$  is a little different from those about  $K$ . Here all integrals are defined on  $(0, \infty)$ . The bias of  $\hat{f}_{ME}(0)$  is,

$$Bias(\hat{f}_{ME}(0)) = \frac{h^2 f''(0)}{2} \int_0^{\infty} u^2 K^*(u) du + o(h^2),$$

which is of order  $O(h^2)$  without assuming that  $f'(0^+) = 0$ . the variance of  $\hat{f}_{ME}(0)$  is

$$Var(\hat{f}_{ME}(0)) = \frac{f(0)}{nh} \int_0^{\infty} K^{*2}(u) du + o\left(\frac{1}{nh}\right).$$

Under the assumption that  $h \rightarrow 0$  and  $nh \rightarrow 0$  when  $n \rightarrow \infty$ , the convergence rate for bias and variance of Estimator (3.1) are  $O(h^2)$  and  $O\left(\frac{1}{nh}\right)$  respectively. Which are the same rate as the classical kernel estimator  $\hat{f}_k(0)$ .

The asymptotic mean square error (AMSE) of (3.1) is

$$AMSE(\hat{f}_{ME}(0)) = \frac{h^4}{4} f''^2(0) \left( \int_0^{\infty} u^2 K^*(u) du \right)^2 + \frac{f(0)}{nh} \int_0^{\infty} K^{*2}(u) du, \quad (3.2)$$

and the value of the bandwidth  $h$  that minimize (3.2) is

$$h = \left( \frac{f(0) \int_0^{\infty} K^{*2}(u) du}{f''(0)^2 \left( \int_0^{\infty} u^2 K^*(u) du \right)^2} \right)^{\frac{1}{5}} n^{-\frac{1}{5}}.$$

The boundary kernel function that minimize the AMSE of  $\hat{f}_{ME}(0)$  is (Mack et al., 1999)

$$K^*(u) = 6(1 - 3u + 2u^2)I_{(0,1)}(u), \quad (3.3)$$

where  $I_B(t)$  is an indicator function of a real set  $B$ . Mack et. al. (1999) assumed that the underlying probability density function  $f(x)$  is to be negative exponential with



scale parameter  $\theta$ , then by taking the kernel function as given in (3.3), we found that

$$h = 3.4375 \theta n^{-\frac{1}{5}}, \theta \text{ can be replaced by its MLE } \hat{\theta} = \bar{x}.$$

### 3.3 Eidous Estimator

Eidous (2011) proposed a new estimator for  $f(0)$  without requiring the assumption

$f'(0^+) = 0$ . He named his estimator "additive histogram estimator". The additive

histogram estimator is given by

$$\hat{f}_{PE}(0) = \frac{1}{nh} \sum_{j=1}^4 \sum_{i=1}^n k_j I_j(X_i). \quad (3.4)$$

where the constant  $k_i$ 's are  $k_1 = \frac{107}{60}$ ,  $k_2 = 0.4$ ,  $k_3 = \frac{-59}{60}$ ,  $k_4 = \frac{41}{120}$ ,  $h$  is the bin

width, and  $I_j(x)$  is the indicator function defined by

$$I_j(x) = \begin{cases} 1, & 0 < x < jh \\ 0, & \text{o.w} \end{cases}.$$

Under the assumption that  $h \rightarrow 0$  and  $nh \rightarrow \infty$  as  $n \rightarrow \infty$  and without assuming that

$f'(0^+) = 0$ , the bias and variance of  $\hat{f}_{PE}(0)$  are

$$Bias(\hat{f}_{PE}(0)) = 0.05h^2 f''(0) + o(h^2)$$

and

$$Var(\hat{f}_{PE}(0)) = \frac{2.96361}{nh} f(0) + o\left(\frac{1}{nh}\right).$$

The AMSE of  $\hat{f}_{PE}(0)$  under the same assumption is

$$AMSE(\hat{f}_{PE}(0)) = \frac{2.96361}{nh} f(0) + 0.0025h^4 f''^2(0). \quad (3.5)$$

The value of  $h$  that minimize (3.5) is given by

$$h = 3.12151 \left( \frac{f(0)}{f''(0)} \right)^{\frac{1}{5}} n^{-\frac{1}{5}}. \quad (3.6)$$

By assuming that  $f(x)$  is a negative exponential with scale parameter  $\theta$  then

Formula (3.6) becomes  $h = 3.1215 \hat{\theta} n^{-\frac{1}{5}}$ , where  $\hat{\theta} = \bar{X}$ .

### 3.4 The Proposed Estimator when $f'(0^+) \neq 0$

Given that the shoulder condition is not true (i.e.  $f'(0^+) \neq 0$ ) then we propose to use the following estimator for  $f(0)$ ,

$$\hat{f}_{P_2}(0) = \frac{2}{nh} \sum_{j=1}^4 \sum_{i=1}^n r_j K\left(\frac{X_i}{jh}\right), \quad (3.7)$$

where  $r_1 = 43/30$ ,  $r_2 = 7/10$ ,  $r_3 = -31/30$  and  $r_4 = 19/60$ . Now if  $D_v = \sum_{j=1}^4 j^v r_j$ ,

then  $D_1 = 1$ ,  $D_2 = 0$  and  $D_3 = -0.6$ . Also  $T(r_1, \dots, r_4) = 0.2742$  (see Subsection 2.4.4) when the kernel function  $K$  is chosen to be Gaussian function (i.e.

the density of  $N(0,1)$ ). Assume that  $f(x) = \frac{1}{\theta} e^{-x/\theta}$ ,  $x \geq 0$  and  $K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$  then

the optimal formula estimate of the smoothing parameter  $h$  is  $h = 1.248 \hat{\theta} n^{-\frac{1}{5}}$ , where  $\hat{\theta} = \bar{X}$ .

**Lemma (3.1).** Suppose that  $f(x)$  is defined on  $[0, \infty)$  and has a continuous second positive derivative at  $x=0$ . Under the assumption that  $h \rightarrow 0$  and  $nh \rightarrow \infty$  as  $n \rightarrow \infty$ , the expected value and variance of  $\hat{f}_{P_2}(0)$  are,

$$E(\hat{f}_{P_2}(0)) = f(0) - 0.6 h^2 f''(0) \int_0^{\infty} u^2 K(u) du + o(h^2), \quad (3.8)$$

and

$$\text{var}(\hat{f}_{P_2}(0)) = \frac{1.0896 f(0)}{nh} + o\left(\frac{1}{nh}\right). \quad (3.9)$$

**Proof of Lemma (2.1) and Lemma (3.1).**

Let  $\hat{f}_{P_i}(0)$ ,  $i = 1, 2$  be the proposed estimators, where  $\hat{f}_{P_1}(0)$  is the Estimator (2.18)

and  $\hat{f}_{P_2}(0)$  be the Estimator (3.7). The expected value of  $K(X/jh)$  is

$$\begin{aligned} EK(X/jh) &= \int_0^{\infty} K(x/jh) f(x) dx \\ &= jh \int_0^{\infty} K(u) (f(0) + jhuf'(0) + (jhu)^2 f''(0)/2 + (jhu)^3 f'''(0)/6 + \dots) du \\ &= jhf(0)/2 + (jh)^2 f'(0)R_1 + (jh)^3 f''(0)R_2/2 + (jh)^4 f'''(0)R_3/6 + \dots \end{aligned}$$

where  $R_z = \int_0^{\infty} u^z K(u) du$ . Therefore, the expected value of  $\hat{f}_{P_i}(0)$ ,  $i = 1, 2$  is

$$\begin{aligned} E(\hat{f}_{P_i}(0)) &= \frac{2}{h} \sum_{j=1}^4 r_j E[K(X/jh)] \\ &= \frac{2}{h} \sum_{j=1}^4 r_j [jhf(0)/2 + (jh)^2 f'(0)R_1 + (jh)^3 f''(0)R_2/2 + (jh)^4 f'''(0)R_3/6 + \dots] \\ &= f(0)D_1 + 2hf'(0)R_1D_2 + h^2 f''(0)R_2D_3 + o(h^2). \end{aligned}$$

Now, for estimator  $\hat{f}_{P_1}(0)$  we indicated that  $D_1 = 1$ ,  $D_2 = 0.46$ ,  $D_3 = -0.6$  and we assumed that  $f'(0^+) = 0$ . This yields,

$$E(\hat{f}_{P_1}(0)) = f(0) - 0.6 h^2 f''(0) R_2 + o(h^2).$$

Also, for estimator  $\hat{f}_{P_2}(0)$  we obtained  $D_1 = 1$ ,  $D_2 = 0$ ,  $D_3 = -0.6$  and without assuming that  $f'(0^+) = 0$ , we obtain,

$$E(\hat{f}_{P_2}(0)) = f(0) - 0.6 h^2 f''(0) R_2 + o(h^2).$$

Note that the convergence rate for bias of two estimators is  $O(h^2)$ .

We turn to the variance of  $\hat{f}_{P_i}(0)$ ,  $i = 1, 2$ . Suppose that  $h \rightarrow 0$  and  $nh \rightarrow \infty$  as  $n \rightarrow \infty$  then the variance of  $\hat{f}_{P_i}(0)$  is

$$\begin{aligned} \text{Var}(\hat{f}_{P_i}(0)) &= \frac{4}{nh^2} \text{var} \left[ \sum_{j=1}^4 r_j K(X/jh) \right] \\ &= \frac{4}{nh^2} \left( E \left[ \sum_{j=1}^4 r_j K(X/jh) \right]^2 - \left[ \sum_{j=1}^4 r_j EK(X/jh) \right]^2 \right). \end{aligned} \quad (3.10)$$

By substituting the expression of  $EK(X/jh)$  in the second term of (3.10), then we obtain

$$\begin{aligned} \text{Var}(\hat{f}_{P_i}(0)) &= \frac{4}{nh^2} E \left[ \sum_{j=1}^4 r_j K(X/jh) \right]^2 + o(n^{-1}h^{-1}) \\ &= \frac{4}{nh^2} E \left[ \sum_{j=1}^4 \sum_{l=1}^4 r_j r_l K(X/jh) K(X/lh) \right] + o(n^{-1}h^{-1}) \end{aligned}$$

$$= \frac{4}{nh^2} E \left[ \sum_{j=1}^4 r_j^2 K^2(X/jh) + 2 \sum_{j<l}^4 r_j r_l K(X/jh)K(X/lh) \right] + o(n^{-1}h^{-1})$$

$$= \frac{4}{nh^2} \left[ \sum_{j=1}^4 r_j^2 E(K^2(X/jh)) + 2 \sum_{j=1}^3 \sum_{l=j+1}^4 r_j r_l E(K(X/jh)K(X/lh)) \right] + o(n^{-1}h^{-1}) \quad (3.11)$$

Now,

$$\begin{aligned} E(K^2(X/jh)) &= \int_0^{\infty} K^2(x/jh) f(x) dx \\ &= jh \int_0^{\infty} K^2(u) (f(0) + juh f'(0) + (jhu)^2 f''(0)/2 + \dots) du \\ &= jhf(0) \int_0^{\infty} K^2(u) du + o(h). \end{aligned} \quad (3.12)$$

Also,

$$\begin{aligned} E(K(X/jh)K(X/lh)) &= \int_0^{\infty} K(x/jh)K(x/lh) f(x) dx \\ &= h \int_0^{\infty} K(u/j)K(u/l) (f(0) + huf'(0) + (hu)^2 f''(0)/2 + \dots) du \\ &= hf(0) \int_0^{\infty} K(u/j)K(u/l) du + o(h). \end{aligned} \quad (3.13)$$

Substituting (3.12) and (3.13) into (3.11), we obtain

$$Var(\hat{f}_{pi}(0)) = \frac{4hf(0)}{nh^2} \left[ \int_0^{\infty} K^2(u) du \sum_{j=1}^4 jr_j^2 + 2 \sum_{j=1}^3 \sum_{l=j+1}^4 r_j r_l \int_0^{\infty} K\left(\frac{u}{j}\right)K\left(\frac{u}{l}\right) du \right] + o(n^{-1}h^{-1})$$

$$= \frac{4f(0)}{nh} T(r_1, \dots, r_4) + o(n^{-1}h^{-1}).$$

Now, for the estimator  $\hat{f}_{P1}(0)$  (Eq. 2.18),  $T(r_1, \dots, r_4) = 0.1847$ . This gives,

$$\text{Var}(\hat{f}_{P1}(0)) = \frac{0.7388f(0)}{nh} + o\left(\frac{1}{nh}\right).$$

Also, for the estimator  $\hat{f}_{P2}(0)$  (Eq. 3.7),  $T(r_1, \dots, r_4) = 0.2742$ . Therefore,

$$\text{Var}(\hat{f}_{P2}(0)) = \frac{1.0896f(0)}{nh} + o(n^{-1}h^{-1}).$$

This completes the proof. Note that the convergence rate for variances of  $\hat{f}_{P1}(0)$  and

$\hat{f}_{P2}(0)$  is  $O\left(\frac{1}{nh}\right)$ .

### 3.5 Asymptotic Mean Square Error and Smoothing Parameter $h$

The optimal smoothing parameter  $h$  for the proposed estimators  $\hat{f}_{P1}(0)$  and  $\hat{f}_{P2}(0)$  can be computed by minimizing the asymptotic mean square error of each estimator with respect to  $h$ . The form of the asymptotic mean square error of both estimators  $\hat{f}_{Pi}(0)$ ,  $i = 1, 2$  is

$$AMSE(\hat{f}_{Pi}(0)) = 0.36h^4 (f''^2(0)) \left[ \int_0^\infty u^2 K(u) du \right]^2 + \frac{4T(r_1, \dots, r_4)f(0)}{nh}. \quad (3.14)$$

By differentiate equation (3.14) with respect to  $h$  and equating the resulting equation to zero, we get the formula of  $h$ , which is given by

$$h = \left( \frac{4T(r_1, \dots, r_4)f(0)}{1.44 f''^2(0) \left[ \int_0^\infty u^2 K(u) du \right]^2} \right)^{\frac{1}{5}} n^{-\frac{1}{5}}. \quad (3.15)$$

Assume that the kernel function is Gaussian function, then  $\int_0^\infty u^2 K(u) du = \frac{1}{2}$ .

Therefore, the smoothing parameter of  $\hat{f}_{P1}(0)$  is

$$h = 1.1546 \left( \frac{f(0)}{f''^2(0)} \right)^{\frac{1}{5}} n^{-\frac{1}{5}},$$

and if  $f(x)$  is taken to be half normal with the form  $f(x) = 2 \exp(-x^2 / 2\sigma^2) / \sigma\sqrt{2\pi}$ , then  $f(0) = 2 / \sigma\sqrt{2\pi}$  and  $f''(0) = -2\sqrt{2} / \sigma^3\sqrt{\pi}$ . Therefore,

$$h = 1.206 \hat{\sigma} n^{-\frac{1}{5}}.$$

For estimator  $\hat{f}_{P2}(0)$ , the smoothing parameter is

$$h = 1.2479 \left( \frac{f(0)}{f''^2(0)} \right)^{\frac{1}{5}} n^{-\frac{1}{5}}.$$

Note that  $T(r_1, \dots, r_4) = 0.2742$  for estimator  $\hat{f}_{P2}(0)$ . If  $f(x)$  is taken to be negative exponential with the form  $f(x) = \exp(-x/\theta) / \theta$ , then  $f(0) = 1/\theta$  and  $f''(0) = 1/\theta^3$ .

This gives,

$$h = 1.248 \hat{\theta} n^{-\frac{1}{5}}.$$

### 3.6 Simulation study and Results

A simulation study is performed to investigate the performances of the different estimators of this chapter and to compare among them. Three estimators for  $f(0)$  are considered,  $\hat{f}_{ME}(0)$  (Eq. 3.1),  $\hat{f}_{PE}(0)$  (Eq. 3.4) and  $\hat{f}_{P2}(0)$  (Eq. 3.7). The smoothing parameter  $h$  for the three estimators takes the form  $h = B \hat{\theta} n^{-\frac{1}{5}}$ , where  $B = 3.438$  for estimator  $\hat{f}_{ME}(0)$ ;  $B = 3.122$  for estimator  $\hat{f}_{PE}(0)$ ; and  $B = 1.248$  for estimator  $\hat{f}_{P2}(0)$ . Note that the kernel function (3.3) is used for Mack's estimator  $\hat{f}_{ME}(0)$ , while the Gaussian kernel is used for the proposed estimator  $\hat{f}_{P2}(0)$ .

The data are simulated from the 12 models that given in Section (2.5) with the same values of  $n$ ,  $\beta$  and  $w$ . The  $RB$ s and  $RME$ s for each estimator are given in Table (3.1) and the EFFs of each estimator with respect to the classical kernel estimator,  $\hat{f}_k(0)$  (Eq. 2.3) are demonstrated in Table (3.2). Note that EFF1, EFF2 and EFF3 represent the efficiencies of  $\hat{f}_{ME}(0)$ ,  $\hat{f}_{PE}(0)$  and  $\hat{f}_{P2}(0)$  with respect to  $\hat{f}_k(0)$  respectively.

The results of Table (3.1) show that the bias of  $\hat{f}_{ME}(0)$  is small in most considered cases compared to that of  $\hat{f}_{PE}(0)$  and  $\hat{f}_{P2}(0)$ . However, the corresponding  $RME$  of  $\hat{f}_{ME}(0)$  is larger than that of  $\hat{f}_{PE}(0)$  and  $\hat{f}_{P2}(0)$ . This indicates that the estimator  $\hat{f}_{ME}(0)$  is more volatile than  $\hat{f}_{PE}(0)$  and  $\hat{f}_{P2}(0)$ . In other words, the performances of  $\hat{f}_{PE}(0)$  and  $\hat{f}_{P2}(0)$  are more stable than that of  $\hat{f}_{ME}(0)$ . The comparison among the efficiencies of the three estimators in Table (3.2) leads us to discard  $\hat{f}_{ME}(0)$  as a competitor estimator even for the models that do not satisfy the shoulder condition assumption. The performances of the other two estimators seem to be more



promising. The two estimators perform well for models that do not satisfy the shoulder condition (e.g. EP model with  $\beta = 1$  and BE with different values of  $\beta$ ) and for the HR model with  $\beta = 1.5$ . As we pointed out in Chapter 2, despite that the HR model with  $\beta = 1.5$  has a shoulder at the origin, it decreases sharply away from the origin when  $\beta = 1.5$ . This model (HR model with  $\beta = 1.5$ ) shares most models that do not have the shoulder condition with this property. However, the performances of  $\hat{f}_{PE}(0)$  and  $\hat{f}_{P_2}(0)$  are not acceptable for the other models that satisfy the shoulder condition (e.g. EP model with  $\beta = 2, 2.5$  and HR model with  $\beta = 2.5, 3$ ). But here we need to remember that the asymptotic properties of these two estimators are derived under the assumption that  $f'(0^+) \neq 0$ . Generally speaking, the two estimators  $\hat{f}_{PE}(0)$  and  $\hat{f}_{P_2}(0)$  perform well when the shoulder condition of the data is not true and they are recommended for line transect sampling if the data seem to be spike at the origin. The shoulder or the spike at the origin can be checked by using the traditional histogram method and by taking 4 to 10 intervals (see Buckland et al. 2001). Finally and because of their good performances, the two estimators  $\hat{f}_{PE}(0)$  and  $\hat{f}_{P_2}(0)$  will be used to build new proposed estimators. Then will be discussed in the next chapter.

**Table 3.1.** The Relative Bias (RB) and the Relative Mean Error (RME) for  $\hat{f}_{ME}(0)$ ,  $\hat{f}_{PE}(0)$  and  $\hat{f}_{P2}(0)$

			$\hat{f}_{ME}(0)$		$\hat{f}_{PE}(0)$		$\hat{f}_{P2}(0)$	
$\beta$	$w$	$n$	RB	RME	RB	RME	RB	RME
Exponential power model								
1.0	5.0	50	-0.070	0.208	-0.204	0.236	-0.116	0.189
		100	-0.045	0.160	-0.163	0.188	-0.100	0.150
		200	-0.040	0.125	-0.130	0.149	-0.085	0.121
1.5	3.0	50	0.046	0.252	0.062	0.160	0.057	0.181
		100	0.027	0.187	0.084	0.137	0.074	0.155
		200	0.027	0.151	0.100	0.131	0.057	0.116
2.0	2.5	50	0.064	0.283	0.200	0.246	0.124	0.218
		100	0.049	0.208	0.207	0.233	0.117	0.182
		200	0.029	0.159	0.204	0.222	0.099	0.145
2.5	2.0	50	0.063	0.294	0.298	0.335	0.150	0.239
		100	0.038	0.224	0.286	0.309	0.123	0.187
		200	0.024	0.169	0.253	0.268	0.098	0.149
Hazard rate model								
1.5	20.0	50	0.157	0.279	-0.275	0.318	-0.053	0.206
		100	0.183	0.246	-0.228	0.256	-0.012	0.134
		200	0.223	0.258	-0.153	0.180	0.061	0.120
2.0	12.0	50	0.166	0.275	-0.066	0.184	0.090	0.198
		100	0.180	0.243	0.009	0.136	0.135	0.184
		200	0.176	0.216	0.087	0.131	0.162	0.180
2.5	8.0	50	0.138	0.267	0.142	0.230	0.176	0.246
		100	0.127	0.224	0.211	0.247	0.207	0.244
		200	0.101	0.171	0.272	0.289	0.208	0.230
3.0	6.0	50	0.108	0.272	0.261	0.308	0.216	0.281
		100	0.065	0.197	0.330	0.353	0.215	0.255
		200	0.048	0.167	0.354	0.365	0.199	0.222
Beta model								
1.5	1.0	50	0.002	0.249	0.049	0.140	0.018	0.175
		100	-0.002	0.191	0.055	0.115	0.009	0.133
		200	0.003	0.152	0.044	0.089	0.006	0.100
2.0	1.0	50	0.004	0.244	0.006	0.129	-0.009	0.167
		100	-0.016	0.186	0.014	0.100	-0.010	0.123
		200	-0.007	0.139	0.004	0.077	-0.003	0.103
2.5	1.0	50	-0.025	0.240	-0.030	0.126	-0.028	0.167
		100	-0.018	0.179	-0.015	0.097	-0.033	0.128
		200	-0.026	0.142	-0.016	0.081	-0.011	0.106
3.0	1.0	50	-0.026	0.223	-0.047	0.133	-0.033	0.165
		100	-0.019	0.178	-0.033	0.105	-0.036	0.137
		200	-0.017	0.133	-0.025	0.082	-0.023	0.101

**Table 3.2.** The Efficiency (EFF) for  $\hat{f}_{ME}(0)$ ,  $\hat{f}_{PE}(0)$  and  $\hat{f}_{P2}(0)$

$\beta$	$w$	$n$	EFF1	EFF2	EFF3
Exponential power model					
1.0	5.0	50	1.514	1.428	1.788
		100	1.769	1.607	1.993
		200	2.034	1.834	2.223
1.5	3.0	50	0.745	1.165	1.066
		100	0.790	1.210	1.032
		200	0.867	1.024	1.138
2.0	2.5	50	0.567	0.618	0.706
		100	0.579	0.529	0.654
		200	0.578	0.419	0.648
2.5	2	50	0.500	0.431	0.632
		100	0.522	0.373	0.626
		200	0.520	0.328	0.591
Hazard rate model					
1.5	20.0	50	1.472	1.191	1.854
		100	1.412	1.334	2.530
		200	1.115	1.568	2.358
2.0	12.0	50	0.919	1.360	1.298
		100	0.827	1.533	1.098
		200	0.701	1.215	0.833
2.5	8.0	50	0.553	0.684	0.646
		100	0.495	0.449	0.463
		200	0.432	0.265	0.335
3.0	6.0	50	0.481	0.383	0.459
		100	0.464	0.267	0.369
		200	0.446	0.201	0.342
Beta model					
1.5	1.0	50	0.854	1.532	1.183
		100	0.896	1.573	1.338
		200	1.040	1.685	1.480
2.0	1.0	50	0.986	1.908	1.329
		100	1.047	1.944	1.594
		200	1.189	2.116	1.631
2.5	1.0	50	0.987	1.788	1.455
		100	1.212	2.121	1.633
		200	1.276	2.201	1.783
3.0	1.0	50	1.037	1.774	1.509
		100	1.226	2.103	1.600
		200	1.388	2.292	1.901

## CHAPTER FOUR

### NEW ESTIMATORS FOR $f(0)$ WITH AND WITHOUT THE SHOULDER CONDITION

#### 4.1 Introduction

In the previous two chapters, the comparison (via simulation technique) of the different estimators of  $f(0)$  suggested that the most promising estimators when the model satisfies the shoulder condition were Barabesi's estimator,  $\hat{f}_B(0)$  (Eq. 2.15) and the first proposed estimator  $\hat{f}_{P1}(0)$  (Eq. 2.18). Also, the most promising estimator when the model does not satisfy the shoulder condition were Eidous's estimator,  $\hat{f}_{PE}(0)$  (Eq. 3.4) and the second proposed estimator  $\hat{f}_{P2}(0)$  (Eq. 3.7). In this chapter we suggest new estimators that combine between two good estimators. The combination is based on one estimator that performs well when the shoulder condition is valid and the other one is selected based on the criteria that it performs well when the shoulder condition is violated. The new proposed estimators will be compared with the semi-parametric estimator that was suggested and studied by Eidous and Al-Shakhatreh (2011).

#### 4.2 Semi-Parametric Estimator

Eidous and Alshakhatreh (2011) introduced a new estimator for  $f(0)$ . Their estimator can be considered as a generalization of the semi-parametric estimator of Eidous (2009), which combines the kernel estimator with parametric detection function. The parametric detection function is selected based on testing the shoulder

condition assumption. They assumed that the parametric detection function is of the form

$$g(x, a_1, a_2) = \exp(-(a_1 x^2 + a_2 x)),$$

which implies that  $g(0, a_1, a_2) = 1$  and  $g'(0, a_1, a_2) = -a_2$ . Eidous and Alshakhatreh (2011) proposed the estimator  $\tilde{f}_{a_1, a_2}(0)$  of  $f(0)$  given by

$$\tilde{f}_{a_1, a_2}(0) = \tilde{\alpha}_h(0) = \frac{2}{nh} \sum_{i=1}^n K\left(\frac{X_i}{h}\right) [g(X_i, a_1, a_2)]^{-k}. \quad (4.1)$$

To utilize Estimator (4.1) the shoulder condition needs to be tested, if the shoulder condition is true (i.e.  $f'(0^+) = 0$ ) then  $g(x, a_1, a_2)$  is selected to be  $g(x, a_1, 0)$ . In this case, the detection function is the half-normal. Otherwise, if  $f'(0^+) = 0$  is rejected then  $g(x, a_1, a_2)$  is taken to be  $g(x, 0, a_2)$ . That is, the detection function is the negative exponential. For simplicity, we used the notation  $\tilde{f}_{a_1, 0}(0)$  to represent the estimator (4.1) when  $a_1$  is fixed and  $a_2 = 0$ . Also, we used the notation  $\tilde{f}_{0, a_2}(0)$  to denote the estimator when  $a_2$  is fixed and  $a_1 = 0$ .

Eidous and Alshakhatreh used the likelihood ratio test of Zhang (2001) to decide about the final form of their estimator (4.1) as the following:

Under  $H_0 : f'(0^+) = 0$  (i.e.  $a_2 = 0$ ), the expected value of the estimator (4.1) is

$$E(\tilde{f}_{a_1, 0}(0)) = \frac{f(0)}{\sqrt{1 + 2(1-k)a_1 h^2}}. \quad (4.2)$$

The form of an unbiased estimator for  $f(0)$  (under  $H_0$ ) is

$$\tilde{f}_0(0) = \tilde{f}_{a_1, 0}(0) \sqrt{1 + 2(1-k)a_1 h^2}. \quad (4.3)$$

There is no sense to use (4.3) as an estimator of  $f(0)$  since it contains unknown parameter  $a_1$ . To overcome this deficiency, they replaced the parameter  $a_1$  by its maximum likelihood estimator (MLE);  $\hat{a}_1 = n/2 \sum_{i=1}^n X_i^2$ . So the estimator of  $f(0)$  is

$$\tilde{f}_0^*(0) = \tilde{f}_{\hat{a}_1,0}(0) \sqrt{1 + 2(1-k)\hat{a}_1 h^2}, \quad (4.4)$$

where  $\tilde{f}_{\hat{a}_1,0}$  represent the estimator (4.1) when  $a_1$  is being estimated from the data by using the maximum likelihood method and  $a_2$  is zero.

Eidous and Al-shakatreh (2011) studied the asymptotic properties of  $\tilde{f}_0^*(0)$  as  $n \rightarrow \infty$ . Similarly, under  $H_1 : f'(0^+) \neq 0$  (i.e.  $a_1 \neq 0$ ), the expected value of the estimator (4.1) is

$$E(\tilde{f}_{0,a_2}(0)) = \frac{2 - 2\Phi(a_2 h(1-k))}{\exp\left(\frac{-a_2^2 h^2 (1-k)^2}{2}\right)} f(0), \quad (4.5)$$

where  $f(0) = \mu = a_2$  and  $\Phi(x)$  is the distribution function of the stander normal distribution. From (4.5) an unbiased estimator of  $f(0)$  ( under  $H_1$ ) is

$$\tilde{f}_1(0) = \frac{\exp\left(\frac{-a_2^2 h^2 (1-k)^2}{2}\right)}{2 - 2\Phi(a_2 h(1-k))} \tilde{f}_{0,a_2}(0). \quad (4.6)$$

Again estimator (4.6) contains unknown parameter  $a_2$ . Replacing  $a_2$  by its MLE;

$\hat{a}_2 = \frac{1}{\bar{X}}$ , where  $\bar{x}$  is the sample mean. Therefore, estimator (4.6) becomes,

$$\tilde{f}_1^*(0) = \frac{\exp\left(\frac{-\hat{a}_2^2 h^2 (1-k)^2}{2}\right)}{2 - 2\Phi(\hat{a}_2 h(1-k))} \tilde{f}_{0,\hat{a}_2}(0),$$

where  $\tilde{f}_{0,a_2}$  represent the estimator (4.1) when  $a_2$  is being estimated from the data by using the maximum likelihood method and  $a_1$  is zero. To use  $\tilde{f}_0^*(0)$  and  $\tilde{f}_1^*(0)$  in practice, we need to determine the value of  $k$ . Eidous (2009) suggested to compute the value of  $k$  by using  $k = f''(0)/g''(0, a_1, a_2)f(0)$ . However,  $k$  still depends on unknown quantities  $f''(0)$ ,  $g''(0, a_1, a_2)$  and  $f(0)$ . Therefore and for simplicity, we fixed the value of  $k$  in this study to be 0 and 1.

To choose the value of the bandwidth  $h$  for this estimator we used the  $m$ -nearest-neighbor method which is given by  $h = x_{(m)}$ , where  $x_{(m)}$  represents the  $m^{\text{th}}$  order statistic in the observed sample. A common choice of  $m$  is given by  $m = \lfloor n^\varepsilon \rfloor$ , where  $\lfloor \cdot \rfloor$  denotes the greatest integer function,  $n$  is the sample size and  $0 < \varepsilon < 1$ . In this setting, we used  $\varepsilon = 4/5$  (see for example, Mack and Rosenblatt, 1979, and Barabesi, 2001).

### 4.3 The Proposed Estimators

Let  $X_1, X_2, \dots, X_n$  be a random sample of perpendicular distances of size  $n$  with unknown detection function and probability density function  $f(x)$  where  $x \geq 0$ . Then we proposed the following estimator for  $f(0)$ ,

$$\hat{f}_{P_3}(0) = \begin{cases} \hat{f}_B(0) & \text{if } f \in F_0 \\ \hat{f}_{PE}(0) & \text{if } f \notin F_0 \end{cases}, \quad (4.7)$$

where  $F_0$  is given in (2.6),  $\hat{f}_B(0)$  is the Barabesi estimator and  $\hat{f}_{PE}(0)$  is the Eidous estimator (see chapters 2 and 3). The estimator  $\hat{f}_{P_3}(0)$  can be re-written as follows :

$$\hat{f}_{P3}(0) = \begin{cases} \hat{f}_k(0) \left[ \frac{h^2}{\hat{\gamma}^2} + 1 \right] & \text{if } f \in F_0 \\ \frac{1}{nh} \sum_{j=1}^4 \sum_{i=1}^n k_j I_j(X_i) & \text{if } f \notin F_0 \end{cases}.$$

To compute estimator  $\hat{f}_{P3}(0)$ , we need first to perform the test

$$H_0 : f \in F_0 \quad \text{vs.} \quad H_1 : f \in F \setminus F_0. \quad (4.8)$$

This can be accomplished by using Mack (1998)'s technique, which is stated in Section (2.3). Another proposed estimator for  $f(0)$  is

$$\hat{f}_{P4}(0) = \alpha \hat{f}_B(0) + (1 - \alpha) \hat{f}_{PE}(0), \quad (4.9)$$

where the parameter  $\alpha \in [0,1]$  represents the weight of  $\hat{f}_B(0)$  in the final estimator  $\hat{f}_{P4}(0)$ . In this study, we suggest to choose  $\alpha$  in Estimator (4.9) by using the  $p$ -value of the test of (4.8), since large  $p$ -value supports the hypothesis  $H_0 : f \in F_0$  then  $\hat{f}_B(0)$  is more appropriate (has larger weight) than  $\hat{f}_{PE}(0)$  to estimate  $f(0)$ .

Based on the results of the previous two chapters, we also proposed the following two estimators, that take the same forms of estimators (4.7) and (4.9),

$$\hat{f}_{P5}(0) = \begin{cases} \hat{f}_{P1}(0) & \text{if } f \in F_0 \\ \hat{f}_{P2}(0) & \text{if } f \notin F_0 \end{cases}, \quad (4.10)$$

and

$$\hat{f}_{P6}(0) = \alpha \hat{f}_{P1}(0) + (1 - \alpha) \hat{f}_{P2}(0), \quad (4.11)$$



where  $\hat{f}_{P1}(0)$  is the estimator given by (2.18) and  $\hat{f}_{P2}(0)$  is given by (3.7). The properties of these proposed estimators are studied via simulation in the next Section.

#### 4.4 Comparison and Results

In this section, a simulation study is conducted to study the performances of the proposed estimators,  $\hat{f}_{P3}(0)$ ,  $\hat{f}_{P4}(0)$ ,  $\hat{f}_{P5}(0)$  and  $\hat{f}_{P6}(0)$ . Also, Eidous and Alshakatreh (2011) estimator,  $\tilde{f}_{a1,a2}(0)$  is considered and its results are given. The efficiencies of these different estimators with respect to the classical kernel estimator are compared.

Again we used the three families, which are given in Section (2.5) by equations (2.21), (2.22) and (2.23) to generate the data. The quantities  $n$ ,  $\beta$  and  $w$  were selected in the same way as in Section (2.5) and 1000 samples each of size  $n = 50, 100, 200$  are simulated from these families. The relative bias ( $RB$ ), the relative mean square error ( $RME$ ) and the efficiency ( $EFF$ ) of each estimator are computed. The efficiencies in Table (4.3) are calculated with respect to the classical kernel estimator,  $\hat{f}_k(0)$  (Eq. 2.3). The abbreviations EFF1, EFF2, EFF3, EFF4, EFF5 and EFF6 in Table (4.3) are used to represent the efficiencies of  $\tilde{f}_{a1,a2}(0)$  (with  $k = 0$ ),  $\tilde{f}_{a1,a2}(0)$  (with  $k = 1$ ),  $\hat{f}_{P3}(0)$ ,  $\hat{f}_{P4}(0)$ ,  $\hat{f}_{P5}(0)$  and  $\hat{f}_{P6}(0)$  respectively.

Before we can apply the different estimators we need to perform a test about the validity of the shoulder condition. The test in (4.8) is achieved at level of significance 0.05. Note that the two estimators  $\hat{f}_{P4}(0)$  and  $\hat{f}_{P6}(0)$  do not need any specific value for the level of significance, since the computing of their values depend on the  $p$ -value.

Depending on the simulation results given in Tables (4.1) and (4.2), we observed that the RMEs of different estimators decrease as the sample size increases. This is a good sign for the consistency of them. The estimator  $\tilde{f}_{a_1, a_2}(0)$  with  $k=1$  gives a good performance compared to  $\tilde{f}_{a_1, a_2}(0)$  with  $k=0$ .

The examining of the proposed estimators results show that the performances of  $\hat{f}_{P_3}(0)$  and  $\hat{f}_{P_5}(0)$  are – in general – satisfactory. The other two estimators  $\hat{f}_{P_4}(0)$  and  $\hat{f}_{P_6}(0)$  perform well for models that do not have the shoulder condition. When comparing the proposed estimators, among themselves, it is difficult to determine the best estimator. However, it appears from the results that  $\hat{f}_{P_5}(0)$  can be considered a worthwhile. The estimator  $\hat{f}_{P_5}(0)$  utilizes the two estimators  $\hat{f}_{P_1}(0)$  and  $\hat{f}_{P_2}(0)$  based on testing the shoulder condition assumption. It becomes  $\hat{f}_{P_1}(0)$  when the shoulder condition is accepted and it is  $\hat{f}_{P_2}(0)$  when the shoulder condition is rejected. A deep insight into the RMEs of  $\hat{f}_{P_1}(0)$  in Table (2.1) and that of  $\hat{f}_{P_2}(0)$  in Table (3.1) shows that the RMEs of  $\hat{f}_{P_5}(0)$  compromise between the RMEs of  $\hat{f}_{P_1}(0)$  and  $\hat{f}_{P_2}(0)$ . The same thing can be said about estimator  $\hat{f}_{P_3}(0)$ , which used  $\hat{f}_B(0)$  and  $\hat{f}_{PE}(0)$ .

The other two proposed estimators  $\hat{f}_{P_4}(0)$  and  $\hat{f}_{P_6}(0)$  perform better than  $\hat{f}_{P_3}(0)$  and  $\hat{f}_{P_5}(0)$  for the models that do not have the shoulder at the origin and the converse is true when models satisfy the shoulder at the origin. This may be due to the fact that even when the shoulder condition is true (accept  $H_0$ ), the  $p$ -value may be small (less than 0.5). This indicates that the weight of  $\hat{f}_B(0)$  is less than the weight of

$\hat{f}_{PE}(0)$  in estimator  $\hat{f}_{P4}(0)$ . Also, for estimator  $\hat{f}_{P6}(0)$  the weight of  $\hat{f}_{P1}(0)$  may be less than the weight of  $\hat{f}_{P2}(0)$  even when the shoulder condition is valid. This may illustrate the small values of the efficiencies that correspond to  $\hat{f}_{P4}(0)$  and  $\hat{f}_{P6}(0)$  for the models EP with  $\beta = 2.0, 2.5$  and HR with  $\beta = 2.5, 3.0$ .

**Table 4.1.** The Relative Bias (RB) and the Relative Mean Error (RME) for  $\tilde{f}_{a_1, a_2}(0)$  with  $k = 0$ ,  $k = 1$ ,  $\hat{f}_{P_3}(0)$  and  $\hat{f}_{P_4}(0)$

			$\tilde{f}_{a_1, a_2}(0)$ ( $k = 0$ )		$\tilde{f}_{a_1, a_2}(0)$ ( $k = 1$ )		$\hat{f}_{P_3}(0)$		$\hat{f}_{P_4}(0)$	
$\beta$	$w$	$n$	RB	RME	RB	RME	RB	RME	RB	RME
Exponential power model										
1.0	5.0	50	-0.289	0.320	-0.262	0.291	-0.261	0.283	-0.229	0.256
		100	-0.273	0.292	-0.252	0.269	-0.225	0.243	-0.186	0.208
		200	-0.251	0.262	-0.229	0.240	-0.194	0.211	-0.153	0.170
1.5	3.0	50	-0.175	0.224	-0.080	0.151	-0.068	0.154	-0.006	0.143
		100	-0.149	0.182	-0.079	0.124	-0.063	0.126	0.010	0.113
		200	-0.135	0.155	-0.076	0.106	-0.052	0.105	0.023	0.099
2.0	2.5	50	-0.109	0.174	-0.025	0.127	0.013	0.151	0.097	0.181
		100	-0.096	0.139	-0.010	0.093	0.010	0.118	0.100	0.155
		200	-0.079	0.108	-0.007	0.072	0.010	0.097	0.101	0.143
2.5	2.0	50	-0.085	0.156	-0.090	0.155	0.059	0.175	0.172	0.237
		100	-0.068	0.115	-0.071	0.116	0.044	0.138	0.160	0.208
		200	-0.049	0.088	-0.052	0.083	0.037	0.109	0.140	0.174
Hazard rate model										
1.5	20	50	-0.120	0.226	-0.095	0.226	-0.282	0.322	-0.277	0.319
		100	-0.060	0.162	-0.064	0.161	-0.229	0.256	-0.228	0.256
		200	-0.006	0.114	-0.016	0.115	-0.153	0.180	-0.153	0.180
2.0	12	50	-0.098	0.187	-0.067	0.172	-0.113	0.193	-0.083	0.183
		100	-0.060	0.133	-0.052	0.122	-0.047	0.136	-0.007	0.128
		200	-0.033	0.092	-0.030	0.091	0.015	0.117	0.065	0.117
2.5	8.0	50	-0.068	0.156	-0.001	0.145	-0.006	0.146	0.071	0.177
		100	-0.039	0.114	0.003	0.105	0.011	0.114	0.112	0.169
		200	-0.030	0.086	0.015	0.075	0.027	0.102	0.144	0.153
3.0	6.0	50	-0.051	0.140	0.057	0.151	0.050	0.149	0.150	0.217
		100	-0.030	0.104	0.051	0.115	0.054	0.126	0.181	0.205
		200	-0.019	0.073	0.048	0.084	0.056	0.115	0.197	0.194
Beta model										
1.5	1.0	50	-0.206	0.245	-0.092	0.145	-0.089	0.163	-0.024	0.137
		100	-0.180	0.206	-0.099	0.134	-0.087	0.138	-0.018	0.109
		200	-0.170	0.186	-0.106	0.125	-0.074	0.114	-0.018	0.087
2.0	1.0	50	-0.224	0.260	-0.129	0.181	-0.114	0.175	-0.058	0.143
		100	-0.209	0.230	-0.132	0.161	-0.112	0.153	-0.052	0.119
		200	-0.185	0.200	-0.129	0.146	-0.101	0.129	-0.047	0.093
2.5	1.0	50	-0.239	0.271	-0.152	0.190	-0.141	0.191	-0.088	0.154
		100	-0.220	0.240	-0.150	0.177	-0.128	0.163	-0.073	0.124
		200	-0.192	0.208	-0.148	0.164	-0.124	0.149	-0.069	0.109
3.0	1.0	50	-0.242	0.276	-0.163	0.207	-0.151	0.195	-0.101	0.162
		100	-0.230	0.249	-0.172	0.193	-0.139	0.173	-0.084	0.132
		200	-0.200	0.216	-0.161	0.176	-0.124	0.149	-0.070	0.108

**Table 4.2.** The Relative Biases (RB) and the relative Mean Error (RME) for  $\hat{f}_{P_5}(0)$  and  $\hat{f}_{P_6}(0)$

			$\hat{f}_{P_5}(0)$		$\hat{f}_{P_6}(0)$	
$\beta$	$w$	$n$	RB	RME	RB	RME
Exponential power model						
1.0	5.0	50	-0.247	0.294	-0.168	0.232
		100	-0.193	0.245	-0.129	0.186
		200	-0.171	0.207	-0.105	0.143
1.5	3.0	50	-0.089	0.148	-0.013	0.136
		100	-0.060	0.132	0.011	0.142
		200	-0.035	0.097	0.021	0.099
2.0	2.5	50	-0.001	0.126	0.065	0.163
		100	-0.005	0.097	0.035	0.120
		200	0.028	0.095	0.065	0.125
2.5	2.0	50	0.039	0.137	0.094	0.190
		100	0.067	0.129	0.106	0.165
		200	0.049	0.090	0.070	0.117
Hazard rate model						
1.5	20.0	50	-0.095	0.216	-0.072	0.192
		100	-0.008	0.144	-0.004	0.140
		200	0.065	0.120	0.065	0.120
2.0	12.0	50	-0.008	0.220	0.075	0.200
		100	0.030	0.183	0.104	0.172
		200	0.089	0.193	0.153	0.192
2.5	8.0	50	-0.023	0.186	0.103	0.213
		100	0.012	0.142	0.124	0.189
		200	0.042	0.108	0.128	0.168
3.0	6.0	50	0.037	0.152	0.131	0.219
		100	0.069	0.131	0.141	0.192
		200	0.096	0.122	0.143	0.170
Beta model						
1.5	1.0	50	-0.104	0.178	-0.041	0.176
		100	-0.090	0.138	-0.045	0.126
		200	-0.075	0.117	-0.033	0.107
2.0	1.0	50	-0.126	0.187	-0.061	0.174
		100	-0.113	0.151	-0.056	0.130
		200	-0.094	0.130	-0.046	0.110
2.5	1.0	50	-0.157	0.211	-0.088	0.182
		100	-0.131	0.173	-0.070	0.144
		200	-0.105	0.146	-0.054	0.119
3.0	1.0	50	-0.159	0.210	-0.083	0.176
		100	-0.138	0.179	-0.075	0.142
		200	-0.115	0.153	-0.062	0.118

**Table 4.3.** The Efficiency (EFF) for  $\tilde{f}_{a_1, a_2}(\cdot)(0)$  with  $k = 0, k = 1, \hat{f}_{P_3}(0), \hat{f}_{P_4}(0), \hat{f}_{P_5}(0)$  and  $\hat{f}_{P_6}(0)$  with respect to  $\hat{f}_k(0)$

$\beta$	$w$	$n$	EFF1	EFF2	EFF3	EFF4	EFF5	EFF6
Exponential power model								
1.0	5.0	50	1.020	1.162	1.202	1.318	1.155	1.467
		100	1.014	1.134	1.223	1.438	1.214	1.599
		200	1.036	1.112	1.285	1.602	1.308	1.896
1.5	3.0	50	0.882	1.282	1.260	1.377	1.250	1.368
		100	0.878	1.267	1.270	1.433	1.210	1.122
		200	0.837	1.235	1.232	1.366	1.273	1.250
2.0	2.5	50	0.867	1.238	1.032	0.859	1.142	0.878
		100	0.853	1.282	1.016	0.783	1.309	1.058
		200	0.893	1.394	1.024	0.713	1.073	0.814
2.5	2	50	1.005	0.970	0.881	0.624	1.094	0.787
		100	1.057	1.013	0.838	0.565	0.845	0.658
		200	1.037	1.051	0.845	0.538	0.957	0.741
Hazard rate model								
1.5	20.0	50	1.699	1.714	1.178	1.189	1.808	2.030
		100	2.161	2.115	1.343	1.347	2.349	2.422
		200	2.468	2.358	1.582	1.584	2.332	2.332
2.0	12.0	50	1.358	1.423	1.333	1.400	1.137	1.258
		100	1.467	1.580	1.517	1.603	1.108	1.178
		200	1.607	1.708	1.277	1.237	0.765	0.769
2.5	8.0	50	1.025	1.116	1.090	0.907	0.884	0.769
		100	0.973	1.145	1.012	0.677	0.794	0.596
		200	0.942	1.028	0.830	0.431	0.722	0.645
3.0	6.0	50	0.919	0.919	0.858	0.572	0.899	0.626
		100	0.932	0.854	0.739	0.402	0.787	0.535
		200	0.998	0.862	0.618	0.324	0.583	0.420
Beta model								
1.5	1.0	50	0.854	1.452	1.286	1.569	1.189	1.204
		100	0.837	1.327	1.297	1.660	1.302	1.432
		200	0.817	1.247	1.301	1.740	1.318	1.440
2.0	1.0	50	0.867	1.311	1.280	1.660	1.192	1.280
		100	0.854	1.231	1.293	1.691	1.283	1.492
		200	0.841	1.160	1.295	1.771	1.310	1.554
2.5	1.0	50	0.883	1.259	1.275	1.563	1.164	1.346
		100	0.871	1.191	1.293	1.711	1.224	1.470
		200	0.865	1.129	1.277	1.753	1.266	1.556
3.0	1.0	50	0.898	1.224	1.259	1.513	1.167	1.393
		100	0.889	1.158	1.275	1.654	1.216	1.524
		200	0.882	1.103	1.271	1.753	1.257	1.627

## CHAPTER FIVE

### REAL DATA ANALYSIS AND CONCLUSIONS

#### 5.1. Introduction

In this chapter we applied the different estimators of this thesis on a set of real data, called "wooden stakes data". Final conclusions and comments are also given in this chapter.

#### 5.2 Wooden Stakes Data

The wooden stakes data are given in Burnham et. al. (1980, p:61 ) and re-stated here in Table (5.1). They deal 150 wooden stakes randomly in determined size area with long equal to 1000 meters, then they used a line transect method to estimate the abundance of these stakes in this area, the number of detected stakes was 68, and the perpendicular distance from detected stakes to the transect line was recorded to form the wooden stakes data which are given in Table (5.1), the form of the distribution function for these data is unknown, but the true value of  $f(0)$  is 0.110294 . The true value of  $D$  was 0.00375 stakes/ $m^2$ , which can be calculated by using the fundamental relation  $D = nf(0)/2L$  . Also, the true value of the number of stakes was  $N = 150$ , which gives the area of study to be  $A = 40000 m^2$  by using the relationship  $D = N / A$  .

The different estimators of this thesis are used to estimate  $f(0)$ ,  $D$  and  $N$  of the stakes data. The approximate standard error of each estimator of  $f(0)$  is also computed by using the bootstrap method. The bootstrap method is a technique used to

generate several random samples based on the original observed data. If we have a sample of size  $n$ , we generate a new random sample with the same size by drawing observations with replacement, then we compute the estimators based on the bootstrap sample and repeat this steps  $B$  times as the following:

Let  $X_1, X_2, \dots, X_n$  be the original random sample with probability density function  $f(x)$  and let  $\hat{f}(0)$  be an estimator of  $f(0)$ , then a bootstrap sample is one of size  $n$ , drawn with replacement from the original random sample. Denote the  $i$ th bootstrap sample by  $X_1^{(i)}, X_2^{(i)}, \dots, X_n^{(i)}$  and let  $B$  be the number of bootstrap samples ( $B$  is taken to be 1000 throughout this study). Let  $\hat{f}_i^*(0)$  be the estimator of  $f(0)$  based on  $X_1^{(i)}, X_2^{(i)}, \dots, X_n^{(i)}$ ,  $i = 1, 2, \dots, B$ , then the approximate standard error of  $\hat{f}(0)$  is

$$SE(\hat{f}(0)) = \sqrt{\frac{1}{B-1} \left[ \sum_{i=1}^B \hat{f}_i^{*2}(0) - B \left( \frac{\sum_{i=1}^B \hat{f}_i^*(0)}{B} \right)^2 \right]}$$

By examining these data, we observe that the perpendicular distance  $x_{68} = 31.31$  seems to be an outlier, which can be eliminated before we are going to analysis the data (see Burnham et. al. 1980). However, Table (5.2) gives the point estimate of  $f(0)$ ,  $D$  and  $N$  together with the approximate standard error of the estimator of  $f(0)$  when  $x_{68}$  is included ( $n = 68$ ) and Table (5.3) gives the same result when  $x_{68}$  is excluded ( $n = 67$ ).



**Table 5.1.** The perpendicular distances of wooden stakes data (Burnham et. al. 1980)

$i$	$x_i$	$i$	$x_i$	$i$	$x_i$	$i$	$x_i$
1	2.02	18	1.61	35	3.79	52	8.49
2	0.45	19	4.08	36	15.24	53	6.08
3	10.4	20	6.5	37	3.47	54	0.4
4	3.61	21	8.27	38	3.05	55	9.33
5	0.92	22	4.85	39	7.93	56	0.53
6	1.0	23	1.47	40	18.15	57	1.23
7	3.4	24	18.6	41	10.5	58	1.67
8	2.9	25	0.41	42	4.41	59	4.53
9	8.16	26	0.4	43	1.27	60	3.12
10	6.47	27	0.2	44	13.72	61	3.05
11	5.66	28	11.59	45	6.25	62	6.6
12	2.95	29	3.17	46	3.59	63	4.4
13	3.96	30	7.1	47	9.04	64	4.97
14	0.09	31	10.71	48	7.68	65	3.17
15	11.82	32	3.86	49	4.89	66	7.67
16	14.23	33	6.05	50	9.1	67	18.16
17	2.44	34	6.42	51	3.25	68	31.31

**Table 5.2.** The point estimates of  $f(0)$ ,  $D$ ,  $N$  and the standard error (SE) of the estimators of  $f(0)$  for wooden stakes data ( $n = 68$ ).

<i>estimator</i>	$\hat{f}(0)$	$\hat{D}$	$\hat{N}$	$SE(\hat{f}(0))$
$\hat{f}_k(0)$	0.10125	$3.443 \times 10^{-3}$	137.72	0.012
$\hat{f}_B(0)$	0.10909	$3.709 \times 10^{-3}$	148.36	0.013
$\hat{f}_E(0)$	0.10027	$3.409 \times 10^{-3}$	136.36	0.014
$\hat{f}_{ME}(0)$	0.11814	$4.017 \times 10^{-3}$	160.68	0.025
$\hat{f}_{PE}(0)$	0.13240	$4.502 \times 10^{-3}$	180.08	0.013
$\tilde{f}_{a_1, a_2}(0)$	0.10796	$3.671 \times 10^{-3}$	146.84	0.021
$\hat{f}_{P1}(0)$	0.10834	$3.684 \times 10^{-3}$	147.36	0.012
$\hat{f}_{P2}(0)$	0.12571	$4.274 \times 10^{-3}$	170.96	0.017
$\hat{f}_{P3}(0)$	0.10909	$3.709 \times 10^{-3}$	148.36	0.015
$\hat{f}_{P4}(0)$	0.12625	$4.293 \times 10^{-3}$	171.72	0.014
$\hat{f}_{P5}(0)$	0.10834	$3.684 \times 10^{-3}$	147.36	0.012
$\hat{f}_{P6}(0)$	0.12720	$4.261 \times 10^{-3}$	170.44	0.015

**Table 5.3.** The point estimates of  $f(0)$ ,  $D$ ,  $N$  and the standard error (SE) of the estimators of  $f(0)$  for wooden stakes data ( $n = 67$ ).

<i>estimator</i>	$\hat{f}(0)$	$\hat{D}$	$\hat{N}$	$SE(\hat{f}(0))$
$\hat{f}_k(0)$	0.10432	$3.495 \times 10^{-3}$	139.80	0.013
$\hat{f}_B(0)$	0.11245	$3.767 \times 10^{-3}$	150.68	0.014
$\hat{f}_E(0)$	0.11076	$3.710 \times 10^{-3}$	148.40	0.015
$\hat{f}_{ME}(0)$	0.11921	$3.994 \times 10^{-3}$	159.76	0.029
$\hat{f}_{PE}(0)$	0.13521	$4.530 \times 10^{-3}$	181.20	0.015
$\tilde{f}_{a_1, a_2}(0)$	0.11285	$3.771 \times 10^{-3}$	150.84	0.026
$\hat{f}_{P1}(0)$	0.11254	$3.770 \times 10^{-3}$	150.80	0.012
$\hat{f}_{P2}(0)$	0.12112	$4.118 \times 10^{-3}$	164.72	0.017
$\hat{f}_{P3}(0)$	0.11245	$3.767 \times 10^{-3}$	150.68	0.014
$\hat{f}_{P4}(0)$	0.12803	$4.289 \times 10^{-3}$	171.56	0.014
$\hat{f}_{P5}(0)$	0.11254	$3.770 \times 10^{-3}$	150.80	0.012
$\hat{f}_{P6}(0)$	0.12258	$4.106 \times 10^{-3}$	164.24	0.014

Testing the shoulder condition of these data (level of significance = 0.05) indicates that the shoulder condition is valid (i.e.  $f'(0^+) = 0$  is accepted). In this case, we want to point out that  $\hat{f}_{P3}(0) = \hat{f}_B(0)$  and  $\hat{f}_{P5}(0) = \hat{f}_{P1}(0)$ . Based on Tables (5.2) and (5.3) and by excluding the two estimators  $\hat{f}_{ME}(0)$  and  $\hat{f}_{a1.a2}(0)$ , we observe that the standard error of the other estimators are close to each other when  $n = 68$  and when  $n = 67$ . As we pointed out in Chapter (3), despite that Mack's estimator  $\hat{f}_{ME}(0)$  has small bias –in general- it suffers from the problem of un-stability in its performance. This problem is also clear for this real data since it has the largest standard error among the other estimators. By considering the standard error of different estimators and how these estimators are close to the true values  $f(0)$ ,  $D$  and  $N$ , the results show that the two estimators  $\hat{f}_B(0)$  and  $\hat{f}_{P1}(0)$  perform the best among the other estimators. They exhibit very similar results for both cases  $n = 68$  and  $n = 67$ . We also note that the estimators  $\hat{f}_{ME}(0)$ ,  $\hat{f}_{PE}(0)$  and  $\hat{f}_{P2}(0)$  -which are developed under the constraint  $f'(0^+) \neq 0$ - give overestimate for the true values.

### 5.3 Concluding Remarks and Comments

In view of findings on the simulated and real example in this thesis, we conclude the following:

- The proposed estimator  $\hat{f}_{P1}(0)$  performs well as a general estimator. Despite that this estimator is developed under the constraints  $f'(0^+) = 0$ , it performs well even for models with  $f'(0^+) \neq 0$ . Also, the results of Barabesi's estimator  $\hat{f}_B(0)$  are acceptable in general.

- If the set of data seem to be spike at the origin, the proposed estimator  $\hat{f}_{P_2}(0)$  is a very competitor for the other existing estimators and can be recommended in this case.
- The idea of combining between some estimators based on testing the shoulder condition assumption seems to be success in some cases but not in all cases. Among these combining estimators, the estimators  $\hat{f}_{P_3}(0)$  and  $\hat{f}_{P_5}(0)$  perform well in general. However, the semi-parametric estimator of Eidous and Alshakatreh beats them in some cases.
- It is not easy job to recommend a specific estimator –from those considered in this thesis- as a best estimator for all cases. However, we can close our comments and conclusions by saying that:
  - The classical kernel estimator  $\hat{f}_k(0)$  is recommended when the model of data has a large shoulder at the origin provided that it does not decrease sharply away the origin (e.g. EP model with  $\beta = 2.5$  and HR model with  $\beta = 3.0$ ).
  - The Barabesi's estimator  $\hat{f}_B(0)$  and the proposed estimator  $\hat{f}_{P_1}(0)$  are recommended for data models with moderate shoulder condition at the origin (e.g. EP model with  $\beta = 1.5$  and HR model with  $\beta = 2.0$ )
  - Eidous's estimator  $\hat{f}_{PE}(0)$  and the proposed estimator  $\hat{f}_{P_2}(0)$  are recommended for data models that do not have a shoulder condition at the origin (e.g. EP model with  $\beta = 1.0$  and BE model with different values of  $\beta$ ). Or even for data models that have a shoulder but decreases markedly away the origin (e.g. HR model with  $\beta = 1.5$ ). In these two cases, the proposed estimators  $\hat{f}_{P_3}(0)$ ,  $\hat{f}_{P_4}(0)$ ,  $\hat{f}_{P_5}(0)$  and  $\hat{f}_{P_6}(0)$  are also perform well.

Finally, the real data example shows that  $\hat{f}_{p_1}(0)$ ,  $\hat{f}_B(0)$ ,  $\hat{f}_E(0)$ ,  $\hat{f}_{p_3}(0)$  and  $\hat{f}_{p_5}(0)$  are all perform well for these real data.

© Arabic Digital Library-Yarmouk University

## REFERENCES

**Barabesi, L.** (2000). Local likelihood density estimation in line transect sampling. *Environmetrics*, **11**, 413-422.

**Barabesi, L.** (2001). Local parametric density estimation method in line transect sampling. *Metron*, **LIX**, 21-37.

**Borgoni, R. and Quatto, P.** (2011). Uniformly most powerful unbiased test for shoulder condition in point transect sampling. Submitted to *Statistical papers*.

**Buckland, S. T.** (1985). Perpendicular distance models for line transect sampling. *Biometrics*, **41**, 177-196.

**Buckland, S. T.** (1992). Fitting density functions using polynomials. *Applied Statistics*, **41**, 63-76

**Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., and Thomas, L.** (2001). Introduction to distance sampling. Oxford University Press, Oxford.

**Burnham, K. P. and Anderson, D. R.**(1976). Mathematical models for nonparametric influences from line transect data. *Biometrics*, **32**, 325-336.

**Burnham, K. P., Anderson, D. R., and Laake, J. L. (1980).** Estimation of density from line transect sampling of biological populations. *Wildlife Monograph*, **72**, 1-202.

**Chen, S. X. (1996).** A Kernel estimate for the density of a biological population by using line transect sampling. *Applied Statistics*, **45**, 135-150.

**Crain, B. R., Burnham, K. P., Anderson, D. R., and Laake, J. L. (1978).** A Fourier series estimator of population-density for line transect sampling. Utah St. Univ. Press, Logon, Utah 25 pp.

**Eberhardt, L. L. (1968).** A preliminary appraisal of line transect. *Journal of Wildlife Management*, **32**, 28-88.

**Eidous, O. M. (2005a).** On improving kernel estimators using line transect sampling. *Commun. in Statist.-Theory and Methods*, **34**, 931-941.

**Eidous, O. M. (2005b).** Frequency histogram model for line transect data with and with and without the shoulder condition. *Journal of the Korean Statistical Society*, **34**, 49-60

**Eidous, O. M. (2005c).** Bias correction for histogram estimator using line transect sampling. *Environmetrics*, **16**, 61-69.



**Eidous, O. M. (2009).** Kernel method starting with half-normal detection function for line transect density estimation. *Commun. in Statist.-Theory and Methods*, **38**, 2366-2378.

**Eidous, O. M. (2011).** Additive Histogram Frequency estimator for wildlife abundance using line transect data without the shoulder condition. To appear in *Metron*.

**Eidous, O. M. and Alshakhatreh, M. K. (2011).** Asymptotic unbiased kernel estimator for line transect sampling. To appear in *commun. in Statist.-Theory and Methods*.

**Gates, C. E. Marshall, W. H. and Olson, D. P. (1968).** Line transect method of estimating grouse population densities. *Biometrics*, **24**, 135-145.

**Gerard, D., and Schucany, W. R., (1999).** Local bandwidth selection for kernel estimation of population densities with line transect sampling. *Biometrics*, **55**, 769-773.

**Gerard, P. D., and Schucany, W. R., (2002).** Combining population density estimates in line transect sampling using the kernel method. *Journal of Agricultural, Biological, and Environmental Statistics*, **7(2)**, 233-242.

**Hayes, R. J., and Buckland, S. T. (1983).** Radial distance models of line transect method. *Biometrics*, **39**, 29-42.

**Hayne, D. W.** (1949). An examination of the strip census method for estimating animal populations. *Journal of wildlife management*, **13**, 145-157.

**Hemingway, P.** (1971). Field trials of the line transect method of sampling large populations of herbivores. PP 405-411. *The scientific management of animal and plant communities for conservation*. Blackwell Sci. Publ. Oxford.

**Hodges and Lehman** (1956). The efficiency of some nonparametric competitors of the t-test. *Annals of Mathematics and Statistics*, **27**, 324-335.

**Hjort, N. L. and Jones, M. C.** (1996) Locally parametric nonparametric density estimation. *Annals of Statistics*, **24**, 1619-1647.

**Karunamuni, R. J., and Quinn, T. J.** (1995). Bayesian estimation of animal abundance for line transect sampling. *Biometrics*, **51**, 1325-1337.

**Mack, Y. P.** (1998). Testing for the shoulder condition in transect sampling. *Commun. in Statist.-Theory and Methods*, **27(2)**, 423-432.

**Mack, Y. P.** (2002). Bias-correction confidence intervals for wildlife abundance estimations. *Commun. in Statist.-Theory and Methods*, **31**, 1107-1122.

**Mack, Y. P. and Quang, P. X.** (1998). Kernel methods in line and point transect sampling. *Biometrics*, **54**, 606-619.

**Mack, Y. P., Quang, P. X., and Zhang, S.** (1999). Kernel estimation in transect sampling without the shoulder condition. *Commun. in Statist.-Theory and Methods*, **28**, 2277-2296.

**Mack, Y. P., and Rosenblatt, M.** (1979). Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, **9**, 1-15.

**Pollock, K. H.** (1978). A family of density estimators for line transect sampling. *Biometrics*, **34**, 475-478.

**Ross, S. M.** (1990). A course in simulation. New York: Macmillan Publishing Company.

**Silverman, B. W.** (1986). Density estimation for statistics and data analysis. London :Chapman and Hall.

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

**Zhang, S.** (2001). Generalized likelihood ratio test for the shoulder condition in line transect sampling. *Commun. in Statist.-Theory and Methods*, **30**, 2343-2354.