

IDEAS BÁSICAS Y EJEMPLOS SOBRE  
ESTIMACIÓN PUNTUAL

*ALFONSO SOTO ALMAGUER*

TESIS

Presentada como Requisito Parcial para  
Obtener el Grado de:

MAESTRO EN  
ESTADÍSTICA APLICADA



UNIVERSIDAD AUTÓNOMA AGRARIA  
“ANTONIO NARRO”

Buenavista, Saltillo Coahuila, México  
Enero de 2014

Universidad Autónoma Agraria Antonio Narro  
Subdirección de Postgrado

IDEAS BÁSICAS Y EJEMPLOS SOBRE  
ESTIMACIÓN PUNTUAL

TESIS

Por:

ALFONSO SOTO ALMAGUER

Elaborada bajo la supervisión del comité particular de asesoría  
y aprobada como requisito parcial para optar al grado de

MAESTRO EN  
ESTADÍSTICA APLICADA

Comité Particular

Asesor principal:


  
Dr. Rolando Cavazos Cadena

Asesor:

  
Dr. Mario Cantú Sifuentes

Asesor:

  
M. C. Luis Rodríguez Gutiérrez

  
Dr. Fernando Ruiz Zárate  
Subdirector de Postgrado

Buenavista, Saltillo, Coahuila, Enero de 2014

# Acknowledgement

Agradezco sinceramente  
al Dr. Rolando Cavazos Cadena por la ayuda  
y orientación brindada durante mis estudios,  
a aquellas personas que de alguna manera me  
impulsaron en mi formación profesional,  
y muy especialmente ... ¡a mi familia!

# Dedication

A Dios.

COMPENDIO

IDEAS BÁSICAS Y EJEMPLOS SOBRE  
ESTIMACIÓN PUNTUAL

Por

ALFONSO SOTO ALMAGUER

MAESTRÍA EN

ESTADÍSTICA APLICADA

UNIVERSIDAD AUTÓNOMA AGRARIA  
ANTONIO NARRO

BUENAVISTA, SALTILLO, COAHUILA, Enero de 2014

Dr. Rolando Cavazos Cadena –Asesor–

**Palabras clave:** Distribución binomial, Teorema central de límite, Convergencia en distribución, Convergencia en probabilidad, Comportamiento asintótico, Métodos de estimación.

Este trabajo trata sobre el problema de *estimación puntual* en modelos estadísticos paramétricos. Los principales objetivos que se persiguen son los siguientes: (i) Analizar dos métodos de construcción de estimadores, a saber, la técnica de verosimilitud máxima y el método de momentos, y (ii) Proporcionar ilustraciones detalladas y completas sobre ideas básicas en la teoría, como insesgamiento, consistencia y normalidad asintótica. La principal contribución de este trabajo se ubica en el último de estos objetivos, ilustrando de manera formal y rigurosa las técnicas de estimación en modelos que surgen en las aplicaciones e involucran distribuciones de uso común.

ABSTRACT

FUNDAMENTAL IDEAS AND EXAMPLES  
ON POINT ESTIMATION

BY

ALFONSO SOTO ALMAGUER

MASTER IN

APPLIED STATISTICS

UNIVERSIDAD AUTÓNOMA AGRARIA  
ANTONIO NARRO

BUENAVISTA, SALTILLO, COAHUILA, January, 2014

Dr. Rolando Cavazos Cadena –Advisor–

**Key Words:** Binomial distribution, Central limit theorem, Convergence in distribution, Convergence in probability, Asymptotic behavior, Estimation methods, Parametric Statistics.

This work concerns the problem of *point estimation* in parametric statistical models, and the three main objectives of this exposition are as follows: (i) To analyze two methods of constructing estimators, namely, the maximum likelihood technique and the method of moments, and (ii) To provide detailed illustrations of basic notion in the theory, as unbiasedness, consistency and asymptotic normality, The main contribution of this note concerns the second objective, presenting complete illustrations of the estimation techniques studied in the thesis in a rigorous and formal manner.

# Contents

<b>1. Presentation</b> .....	<b>1</b>
1.1 Introduction .....	1
1.2 The Estimation Problem .....	2
1.3 Main Goals and Contribution .....	3
1.4 The Origin of This Work .....	3
1.5 The Organization .....	4
<b>2. Point Estimation</b> .....	<b>6</b>
2.1 Statistical Point Estimation: Concepts .....	6
2.2 Examples and Factorial Moments .....	10
2.3 Unbiasedness and Consistency .....	14
2.4 Additional Examples .....	18
<b>3. Maximum Likelihood</b> .....	<b>28</b>
3.1 Maximum Likelihood Estimation .....	28
3.2 The Method in Specific Cases .....	31
3.3 Estimation of the Mean of a Laplace Distribution .....	36
3.4 The Poisson and Normal Distributions .....	39
3.5 Estimating the Parameters of a Beta Distribution .....	45
3.6 Additional Examples .....	47
3.7 Bivariate Normal Distribution .....	53
3.8 Logistic Model .....	59
<b>4. Method of Moments</b> .....	<b>71</b>
4.1 Description of the Method .....	71

4.2 Consistency of the Estimators .....	72
4.3 Applications .....	74
4.4 Further Examples .....	76
<b>References .....</b>	<b>84</b>



# Chapter 1

## Presentation

This chapter presents a general perspective of the material presented in the subsequent development. The main goals, contributions and the motivation behind this work are clearly stated, and the organization and content of the following chapters is briefly described.

### 1.1. Introduction

This work deals with the problem of *parametric point estimation*, which is pervasive and plays a central role in the theory and applications of statistics. Certainly, point estimation lays in the core of the statistical methodology, and a major step in every analysis is the determination of estimates (*i.e.*, approximations) to some unknown quantities in terms of the observed data, and every treatise on theoretical or applied statistics dedicates a good amount of space to the analysis of diverse methods to construct estimators and to study its properties; see, for instance, Dudewicz and Mishra (1988), Wackerly *et al.* (2009), Lehmann and Casella (1999), or Graybill (2000).

The topics presented in the following chapters are mainly concentrated on three aspects of the estimation problem:

- (i) The construction of estimators *via* the maximum likelihood technique and the method of moments, and

(ii) The study of particular models to illustrate the estimation procedures, and to point out the technical difficulties to obtain explicit formulas.

The basic estimation problem is briefly described below.

## 1.2. The Estimation Problem

In general, the purpose of a statistical analysis is to use the observed data *to gain knowledge* about some unknown aspect of the process generating the observations. The observable data  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  is thought of as a random vector whose distribution is not completely known. Rather, theoretical or modeling considerations lead to assume that the distribution of  $\mathbf{X}$ , say  $P_{\mathbf{X}}$ , belongs to a certain family  $\mathcal{F}$  of probability measures defined on (the Borel class of)  $\mathbb{R}^n$ :

$$P_{\mathbf{X}} \in \mathcal{F}. \quad (1.2.1)$$

This is a statistical model, and in any practical instance it is necessary to include a precise definition of the family  $\mathcal{F}$ . In this work, the main interest concentrates on *parametric models*, for which the family  $\mathcal{F}$  can be indexed by a  $k$ -dimensional vector  $\theta$  whose components are real numbers; in such a case the set of possible values of  $\theta$ , which is referred to as the parameter space, will be denoted by  $\Theta$  and  $\mathcal{F}$  can be written as

$$\mathcal{F} = \{P_{\theta} \mid \theta \in \Theta\}.$$

In this context the model (1.2.1) ensures that *there exists some parameter*  $\theta^* \in \Theta$  such that  $P_{\mathbf{X}} = P_{\theta^*}$ , that is, for every (Borel) subset  $A$  of  $\mathbb{R}^n$

$$P[X \in A] = P_{\mathbf{X}}[A] = P_{\theta^*}[A]. \quad (1.2.2)$$

The parameter  $\theta^*$  satisfying this relation for every (Borel) subset of  $\mathbb{R}^n$  is *the true parameter value*. Notice that the model prescribes the existence of  $\theta^* \in \Theta$  such that the above equality always holds, but does not specify which is the parameter  $\theta^*$ ; it is only supposed that  $\theta^*$  belongs to the parameter space  $\Theta$ , and the main objective of the analyst is to determine  $\theta^*$  using the value attained by the vector  $\mathbf{X}$ , say  $\mathbf{X} = \mathbf{x}$ . Indeed, the lack of exact knowledge of  $\theta^*$  represents ‘the aspects that are unknown’ to the analyst about the real process generating the observation vector  $\mathbf{X}$ . On the other hand, in any practical situation,  $\theta^*$  can not be determined exactly after observing the

value of  $\mathbf{X}$ , so that the real goal of the analyst is to make an ‘educated guess’ about the true parameter value using the observed value of  $\mathbf{X}$ ; this means that a function  $T(\mathbf{X})$  must be constructed so that, after observing  $\mathbf{X} = \mathbf{x}$ , the value  $T(\mathbf{x})$  will represent ‘the guess’ (approximation) of the analyst to the true parameter value  $\theta^*$ . More generally, the interest may be to obtain an ‘approximation’ to the value  $g(\theta^*)$  attained by some function  $g(\theta)$  at the true parameter value  $\theta^*$ . The estimation problem consists in constructing a function  $T(\mathbf{X})$  whose values will be used as approximations to  $g(\theta^*)$  such that the estimator  $T(\mathbf{X})$  has good statistical properties. As already mentioned, this work analyzes methods to construct estimators.

### 1.3. Main Goals and Contribution

The main goals of this work can be described as follows:

- (i) To present a formal description of two important methods to construct estimators, namely, the maximum likelihood technique, and the method of moments;
- (ii) To use selected examples to illustrate the construction of estimators in models involving distributions frequently used in applications,

*The main contribution* of the presentation consists in presenting a rigorous and formal analysis of diverse examples involving common distribution arising in applications.

### 1.4. The Origin of This Work

This work was developed as part of the project *Mathematical Statistics: Elements of Theory and Examples*, started on July 2011 by the Graduate Program in Statistics at the Universidad Autónoma Agraria Antonio Narro. The author and Mary Carmen Ruiz Moreno were the initial students in the project, and it is a pleasure to thank Mary Carmen for a lot of interesting and stimulating discussions.

The basic aims of the project are:

- (i) To be a framework where statistical problems can be freely and fruitfully discussed;
- (ii) To promote the *understanding* of basic statistical and analytical tools through the analysis and detailed solution of exercises.

(iii) To develop the *writing skills* of the participants, generating an organized set of neatly solved examples, which can be used by other members of the program, as well as by the statistical communities in other institutions and countries.

(iv) To develop the *communication skills* of the students and faculty through the regular participation in seminars, where the results of their activities are discussed with the members of the program.

The work of the project has been concerned with fundamental statistical theory at an intermediate (non-measure theoretical) level, as in the book *Mathematical Statistics* by Dudewicz and Mishra (1998). When necessary, other more advanced references that have been useful are Lehmann and Casella (1998), Borobkov (1999) and Shao (2002), whereas deeper probabilistic aspects have been studied in the classical text by Loève (1984). On the other hand, statistical analysis requires algebraic and analytical tools, and in these directions the basic references in the project are Apostol (1980), Fulks (1980), Khuri (2002) and Royden (2003), which concern mathematical analysis, whereas the algebraic aspects are covered in Graybill (2001) and Harville (2008).

The examples presented in the following chapters reflect the work developed in the project, and it is a pleasure to thank to Mary Carmen Ruiz Moreno by clever discussions, and to and to the Statistics Program promoting the project, by the opportunity to collaborate in the project.

## 1.5. The Organization

The material presented below has been organized as follows: In Chapter 2 some basic concepts in the theory of point estimation are introduced, presenting a description of the idea of parametric statistical model, and discussing the estimation problem of an unknown parametric function. The presentation continues with the notions of unbiased estimator and consistency of a sequence of estimators, and the related concept of asymptotically unbiased sequence is also analyzed.

Next, in Chapter 3 the method of maximum likelihood estimation is introduced, which is based on the intuitive idea that, after observing the data, the estimate of the unknown parameter  $\theta$  is the value  $\hat{\theta}$  in the parameter space that assigns highest probability to the observation and, finally, Chapter

4 is concerned with the method of moments; as already mentioned, all of the notions introduced in this work are illustrated by carefully analyzed examples.

# Chapter 2

## Point Estimation

This chapter introduces fundamental concepts in the theory of Point Estimation. The notion of estimator is introduced and the ideas of unbiasedness and consistency are discussed and illustrated.

### 2.1. Statistical Point Estimation: Concepts

A parametric statistical model for an observable vector

$$\mathbf{X} = (X_1, X_2, \dots, X_n)$$

prescribes a family  $\{P_\theta\}_{\theta \in \Theta}$  of possible probability distributions for  $\mathbf{X}$ . The set of indices  $\Theta$  is referred to as the *parameter space* and is a subset of an Euclidean space  $\mathbb{R}^k$ . The essence of a statistical model is that the distribution of  $\mathbf{X}$  is supposed to be  $P_\theta$  for some parameter  $\theta \in \Theta$ , but the ‘true’ parameter value—the one which corresponds to the distribution of  $\mathbf{X}$ —is unknown. The statistical model is briefly described by writing

$$\mathbf{X} \sim P_\theta, \quad \theta \in \Theta.$$

The main objective of the analyst is to determine, at least approximately, the value of the true parameter or, more generally, the value of a function

$g(\theta)$  at the true parameter. To achieve this goal the components of  $\mathbf{X}$  are combined in some way to obtain a function

$$T_n \equiv T_n(\mathbf{X}) = T_n(X_1, X_2, \dots, X_n),$$

and after observing  $\mathbf{X} = \mathbf{x} = (x_1, x_2, \dots, x_n)$ , the value

$$T_n(\mathbf{x}) = T_n(x_1, x_2, \dots, x_n)$$

is used as an ‘approximation’ of the unknown quantity  $g(\theta)$ . The random variable  $T_n$  is called an *estimator* of  $g(\theta)$  and  $T_n(\mathbf{x})$  is the *estimate* corresponding to the observation  $\mathbf{X} = \mathbf{x}$ .

An estimator of  $g(\theta)$  is *unbiased* if

$$E_\theta[T_n] = g(\theta)$$

for every  $\theta \in \Theta$ ; notice that the subindex  $\theta$  in the expectation operator is used to indicate that the expected value is computed under the condition that  $\theta$  is the true parameter value. In general, the value attained by an estimator  $T_n = T_n(X_1, X_2, \dots, X_n)$ , does not coincide with the quantity  $g(\theta)$ . However, if the estimator  $T_n$  is unbiased, and the experiment generating  $\mathbf{X}$  is repeated, obtaining the estimators  $T_{n1}, T_{n2}, T_{n3}, \dots$  at each repetition, it follows from the law of large numbers that the average

$$\frac{T_{n1} + T_{n2} + T_{n3} + \dots + T_{nk}}{k}$$

converges to  $g(\theta)$  as the number  $k$  of repetitions increases. Thus, on the average, the estimator  $T_n$  ‘points to the correct quantity’  $g(\theta)$ . It must be noted that not all of the parametric quantities  $g(\theta)$  admit an unbiased estimator. For instance, suppose that  $X_1, X_2, \dots, X_n$  is a sample from the *Bernoulli*( $\theta$ ) distribution, where  $\theta \in \Theta = [0, 1]$ , and assume that  $T_n = T_n(X_1, X_2, \dots, X_n)$  is an unbiased estimator for  $g(\theta)$ . Observing that

$$P_\theta[X_1 = x_1, X_2 = x_2, \dots, X_n = x_n] = \theta^{\sum_i x_i} (1 - \theta)^{n - \sum_i x_i}$$

when  $x_i$ s are zero or one for all  $i$ , it follows that

$$E_\theta[T_n] = \sum_{x_1, \dots, x_k=0,1} T(x_1, x_2, \dots, x_n) \theta^{\sum_i x_i} (1 - \theta)^{n - \sum_i x_i}$$

is a polynomial of degree less than or equal to  $n$ , so that  $E_\theta[T_n] = g(\theta)$  for all  $\theta \in \Theta$  can not be satisfied for functions that are not polynomials, as  $g(\theta) = e^\theta$  or  $g(\theta) = \sin(\theta)$ , or even for polynomial functions with degree larger than  $n$ , as  $g(\theta) = \theta^{n+1}$ . Thus, the unbiasedness property may be too restrictive, and it is possible to have that an unbiased estimator does not exist in some cases of interest.

The *bias function* of an estimator  $T_n$  of  $g(\theta)$  is defined by

$$b_{T_n, g}(\theta) \equiv b_{T_n}(\theta) := E_\theta[T_n] - g(\theta), \quad \theta \in \Theta,$$

so that  $T_n$  is unbiased if  $b_{T_n}(\theta) = 0$  for every  $\theta \in \Theta$ . To evaluate the bias function of an estimator  $T_n$  it is necessary to determine the expected value  $E_\theta[T_n]$ , and usually this task requires to know the density or probability function of  $T_n$ ; however, occasionally symmetry conditions may help to simplify the computation.

A sequence  $\{T_n\}_{n=1,2,\dots}$  of estimators of  $g(\theta)$  is *asymptotically unbiased* if

$$\lim_{n \rightarrow \infty} b_{T_n}(\theta) = 0, \quad \theta \in \Theta,$$

a condition that is equivalent to requiring that, for each parameter  $\theta \in \Theta$ ,  $E_\theta[T_n] \rightarrow g(\theta)$  as  $n \rightarrow \infty$ .

On the other hand, a sequence  $\{T_n\}_{n=1,2,\dots}$  of estimators of  $g(\theta)$  is *consistent* if for each  $\varepsilon > 0$ ,

$$\lim_{n \rightarrow \infty} P_\theta[|T_n - g(\theta)| > \varepsilon] = 0, \quad \theta \in \Theta,$$

that is, the sequence  $\{T_n\}$  always converges in probability to  $g(\theta)$  with respect to the distribution  $P_\theta$ . The above convergence will be alternatively written as

$$T_n \xrightarrow{P_\theta} g(\theta).$$

There are three main tools to show consistency of a sequence of estimators, which are briefly discussed in the following points (i)–(iii):

(i) *The strong law of large numbers*: Assume that the quantity  $g(\theta)$  is the expectation of a random variable  $Y = Y(X_1)$ , that is,

$$g(\theta) = E_\theta[Y(X_1)]$$



In this case, if the variables  $X_1, X_2, \dots, X_n, \dots$  are independent and identically distributed, setting

$$T_n = \frac{Y(X_1) + Y(X_2) + \dots + Y(X_n)}{n},$$

the law of large numbers ensures that  $T_n \xrightarrow{P_\theta} g(\theta)$ , *i.e.*, that the sequence  $\{T_n\}$  of estimators of  $g(\theta)$  is consistent.

(ii) The *continuity theorem*. Roughly, this result establishes that the consistency property is preserved under the application of a continuous function, a conclusion that is formally stated as follows:

Suppose that the parametric functions  $g_1(\theta), g_2(\theta), \dots, g_r(\theta)$  are estimated consistently by the sequences  $\{T_{1n}\}, \{T_{2n}\}, \dots, \{T_{rn}\}$ , that is

$$T_{in} \xrightarrow{P_\theta} g_i(\theta), \quad i = 1, 2, \dots, r.$$

Additionally, let the function  $G(x_1, x_2, \dots, x_r)$  be a function that is continuous at each point  $(g_1(\theta), \dots, g_r(\theta))$ , where  $\theta \in \Theta$ . In this context, the sequence  $\{G(T_{1n}, T_{2n}, \dots, T_{rn})\}$  of estimators of the parametric function  $G(g_1(\theta), g_2(\theta), \dots, g_r(\theta))$  is consistent, *i.e.*,

$$G(T_{1n}, T_{2n}, \dots, T_{rn}) \xrightarrow{P_\theta} G(g_1(\theta), g_2(\theta), \dots, g_r(\theta)).$$

(iii) The idea of *convergence in the mean*. If  $p$  is a positive number, a sequence of random variables  $\{T_n\}$  converges in the mean of order  $p$  to  $g(\theta)$  if

$$\lim_{n \rightarrow \infty} E_\theta[|T_n - g(\theta)|^p] = 0, \quad \theta \in \Theta;$$

the notation  $T_n \xrightarrow{L^p} g(\theta)$  will be used to indicate that the above condition holds. The most common instance in applications arises when  $p = 2$ , so that  $T_n \xrightarrow{L^2} g(\theta)$  is equivalent to the statement that, for each  $\theta \in \Theta$ ,  $E_\theta[(T_n - g(\theta))^2] \rightarrow 0$  as  $n \rightarrow \infty$ . When  $T_n \xrightarrow{L^p} g(\theta)$  the sequence  $\{T_n\}$  of estimators of  $g(\theta)$  is referred to as *consistent in the mean of order  $p$* . Suppose now that  $T_n \xrightarrow{L^p} g(\theta)$ , and notice that Markov's inequality yields that, for each  $\varepsilon > 0$ ,

$$P_\theta[|T_n - g(\theta)| > \varepsilon] \leq \frac{E_\theta[|T_n - g(\theta)|^p]}{\varepsilon^p} \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

so that

$$T_n \xrightarrow{L^p} g(\theta) \Rightarrow T_n \xrightarrow{P} g(\theta);$$

in words, if the sequence  $\{T_n\}$  of estimators of  $g(\theta)$  is consistent in the mean of order  $p$ , then  $\{T_n\}$  is consistent (in probability). This implication is useful, since it is frequently easier to establish consistency in the mean of some order  $p > 0$ , than to prove consistency directly. When considering consistency in the mean of order 2, it is useful to keep in mind that the mean square error  $E_\theta[(T_n - g(\theta))^2]$ , the variance and the bias function of  $T_n$  are related by

$$E_\theta[(T_n - g(\theta))^2] = b_{T_n}(\theta)^2 + \text{Var}_\theta(T_n).$$

## 2.2. Examples and Factorial Moments

The following examples illustrate the ideas recently introduced.

**Exercise 2.2.1.** Let  $X_1, X_2, \dots, X_n$  be independent and identically distributed Bernoulli random variables with common probability of success  $p$ , and denote  $T_n = X_1 + X_2 + \dots + X_n$ , whereas  $\bar{X}_n = T_n/n$  is the sample mean of the sample. Show that

- (a)  $T_n(T_n - 1)/c_n$  with  $c_n = n(n - 1)$  is an unbiased estimator of  $p^2$ .
- (b)  $T_n(T_n - 1)(T_n - 2)/d_n$  with  $d_n = n(n - 1)(n - 2)$  is an unbiased estimator of  $p^3$ .
- (c) Investigate the consistency of the estimators in parts (a) and (b).
- (d) Find an unbiased estimator of  $p - q$  where, as usual,  $q = 1 - p$ .

**Solution.** (a) It must be shown that  $E_p[T_n(T_n - 1)/[n(n - 1)]] = p$  for every  $p \in [0, 1]$ . To compute the expectation, first notice that  $T_n \sim \text{Binomial}(n, p)$ , so that  $E_p[T_n(T_n - 1)] = \sum_{t=0}^n t(t-1) \binom{n}{t} p^t q^{n-t}$ . To evaluate this summation, recall the identity

$$\binom{n}{t} = \frac{n}{t} \binom{n-1}{t-1}, \quad t \geq 1, \quad (2.2.1)$$

to obtain, after two successive applications of this relation, that

$$\binom{n}{t} = \frac{n}{t} \binom{n-1}{t-1} = \frac{n}{t} \cdot \frac{n-1}{t-1} \binom{n-2}{t-2}, \quad t \geq 2. \quad (2.2.2)$$

Therefore,

$$\begin{aligned}
E_p[T_n(T_n - 1)] &= \sum_{t=0}^n t(t-1) \binom{n}{t} p^t q^{n-t} \\
&= \sum_{t=2}^n t(t-1) \frac{n}{t} \cdot \frac{n-1}{t-1} \binom{n-2}{t-2} p^t q^{n-t} \\
&= n(n-1) \sum_{t=2}^n \binom{n-2}{t-2} p^t q^{n-t},
\end{aligned}$$

where (2.2.2) was used to set the last equality. Changing the variable  $t$  in the last summation to  $r = t - 2$ , it follows that

$$\begin{aligned}
E_p[T_n(T_n - 1)] &= n(n-1) \sum_{r=0}^{n-2} \binom{n-2}{r} p^{r+2} q^{n-2-r} \\
&= n(n-1) p^2 \sum_{r=0}^{n-2} \binom{n-2}{r} p^r q^{n-2-r} \\
&= n(n-1) p^2,
\end{aligned}$$

where to set the third equality it was used that  $\sum_{r=0}^{n-2} \binom{n-2}{r} p^r q^{n-2-r}$  is the sum of all no-null probabilities in the *Binomial*( $n-2, p$ ) distribution, so that the summation equals to 1. Consequently, for  $n \geq 2$ ,  $c_n = n(n-1) > 0$  and  $E_p[T_n(T_n - 1)/c_n] = p^2$ ; since the parameter  $p \in [0, 1]$  is arbitrary,  $T_n(T_n - 1)/c_n$  is an unbiased estimator of  $p^2$ .

(b) The argument parallels the one used in part (a). It is necessary to evaluate

$$\begin{aligned}
E_p[T_n(T_n - 1)(T_n - 2)] &= \sum_{t=0}^n t(t-1)(t-2) \binom{n}{t} p^t q^{n-t} \\
&= \sum_{t=3}^n t(t-1)(t-2) \binom{n}{t} p^t q^{n-t}.
\end{aligned}$$

Applying (2.2.1) three times, it follows that

$$\binom{n}{t} = \frac{n}{t} \cdot \frac{n-1}{t-1} \cdot \frac{n-2}{t-2} \binom{n-3}{t-3}, \quad t \geq 3,$$

and these two last displays together yield that

$$\begin{aligned}
& E_p[T_n(T_n - 1)(T_n - 2)] \\
&= \sum_{t=3}^n t(t-1)(t-2) \frac{n}{t} \cdot \frac{n-1}{t-1} \cdot \frac{n-2}{t-2} \binom{n-3}{t-3} p^t q^{n-t} \\
&= n(n-1)(n-2) \sum_{t=3}^n \binom{n-3}{t-3} p^t q^{n-t} \\
&= n(n-1)(n-2) \sum_{r=0}^{n-3} \binom{n-3}{r} p^{r+3} q^{n-3-r} \\
&= n(n-1)(n-2) p^3 \sum_{r=0}^{n-3} \binom{n-3}{r} p^r q^{n-3-r} \\
&= n(n-1)(n-2) p^3
\end{aligned}$$

where the change of variable  $r = t-3$  was used to set the third equality. Thus, for  $n \geq 3$ ,  $d_n = n(n-1)(n-2) \neq 0$  and  $E_p[T_n(T_n - 1)(T_n - 2)/d_n] = p^3$  for every parameter value  $p \in [0, 1]$ , that is,  $T_n(T_n - 1)(T_n - 2)/d_n$  is an unbiased estimator of  $p^3$ .

(c) By the strong law of large numbers,  $T_n/n = \bar{X}_n \xrightarrow{P_p} E_p[X_1] = p$ . Consequently, by the continuity theorem,

$$\frac{T_n(T_n - 1)}{c_n} = \frac{T_n(T_n - 1)}{n(n-1)} = \frac{\bar{X}_n(\bar{X}_n - 1/n)}{1(1 - 1/n)} \xrightarrow{P_p} \frac{p \cdot p}{1 \cdot 1} = p^2$$

and, similarly,

$$\begin{aligned}
\frac{T_n(T_n - 1)(T_n - 2)}{d_n} &= \frac{T_n(T_n - 1)(T_n - 2)}{n(n-1)(n-2)} \\
&= \frac{\bar{X}_n(\bar{X}_n - 1/n)(\bar{X}_n - 2/n)}{1(1 - 1/n)(1 - 2/n)} \xrightarrow{P_p} \frac{p \cdot p \cdot p}{1 \cdot 1 \cdot 1} = p^3.
\end{aligned}$$

Thus, the sequences  $\{T_n(T_n - 1)/c_n\}$  and  $\{T_n(T_n - 1)(T_n - 2)/d_n\}$  are consistent for  $p^2$  and  $p^3$ , respectively.

(d) Notice that  $g(p) = p - q = p - (1 - p) = 2p - 1$ ; since  $E_p[\bar{X}_n] = p$ , it follows that  $E_p[2\bar{X}_n - 1] = 2p - 1 = g(p)$ , that is,  $2\bar{X}_n - 1$  is an unbiased estimator of  $g(p) = 2p - 1$ .  $\square$

**Remark 2.2.1.** For  $a \in \mathbb{R}$  and a positive integer  $k$ , set

$$(a)_k := a(a-1) \cdots (a-k+1).$$

If  $k$  is a positive integer, for each random variable  $W$  the  $k$ th factorial moment is given by

$$E[(W)_k] = E[W(W-1)\cdots(W-k+1)]$$

whenever the expectation exists. With this notation, the core of the solution to Exercise 2.2.1 was the computation of  $E_p[(T_n)_2]$  and  $E_p[(T_n)_3]$ , the second and third factorial moments of  $T_n$ . In some cases, computation of a factorial moment of  $W$  can be simplified by using the following *factorial moments generating function*:

$$\text{FactM}_W(t) = E[t^W], \quad t > 0. \quad (2.2.3)$$

If this function is finite in a neighborhood of 1, then the derivatives of all orders exist about 1, and are given by

$$\begin{aligned} \frac{d}{dt} \text{FactM}_W(t) &= E[Wt^{W-1}] = E[(W)_1 t^{W-1}] \\ \frac{d^2}{dt^2} \text{FactM}_W(t) &= E[W(W-1)t^{W-2}] = E[(W)_2 t^{W-2}] \\ \frac{d^3}{dt^3} \text{FactM}_W(t) &= E[W(W-1)(W-2)t^{W-3}] = E[(W)_3 t^{W-3}] \\ &\vdots \\ \frac{d^k}{dt^k} \text{FactM}_W(t) &= E[W(W-1)(W-2)\cdots(W-k+1)t^{W-k}] \\ &= E[(W)_k t^{W-k}], \quad k \geq 1. \end{aligned}$$

Evaluating at  $t = 1$ , it follows that

$$\left. \frac{d^k}{dt^k} \text{FactM}_W(t) \right|_{t=1} = E[(W)_k] = E[W(W-1)(W-2)\cdots(W-k+1)], \quad (2.2.4)$$

so that the factorial moments of  $W$  can be evaluated from the knowledge of  $\text{FactM}_W(t)$  and its derivatives. For the random variable  $T_n$  in the previous exercise,  $T_n \sim \text{Binomial}(n, p)$ , so that

$$\text{FactM}_{T_n}(t) = \sum_{k=0}^n t^k \binom{n}{k} p^k q^{n-k} = \sum_{k=0}^n \binom{n}{k} (pt)^k q^{n-k} = (q + tp)^n,$$

and then

$$\begin{aligned} \frac{d}{dt} \text{FactM}_{T_n}(t) &= np(q + tp)^{n-1} \\ \frac{d^2}{dt^2} \text{FactM}_{T_n}(t) &= n(n-1)p^2(q + tp)^{n-2} \\ \frac{d^3}{dt^3} \text{FactM}_{T_n}(t) &= n(n-1)(n-2)p^3(q + tp)^{n-3}; \end{aligned}$$

evaluating the second and third derivatives at  $t = 1$ , it follows that

$$E_p[(T_n)_2] = E_p[T_n(T_n - 1)] = \left. \frac{d^2}{dt^2} \text{FactM}_{T_n}(t) \right|_{t=1} = n(n-1)p^2,$$

and

$$E_p[(T_n)_3] = E_p[T_n(T_n - 1)(T_n - 2)] = \left. \frac{d^3}{dt^3} \text{FactM}_{T_n}(t) \right|_{t=1} = n(n-1)(n-2)p^3,$$

providing an alternative way to compute the relevant expectations in Exercise 2.2.1.  $\square$

### 2.3. Unbiasedness and Consistency

In this section the ideas of unbiased estimator and consistency of a sequence of estimators are illustrated in some specific examples.

**Exercise 2.3.1.** Let  $T_n$  and  $T'_n$  be two independent unbiased and consistent estimators of  $\theta$ .

- (a) Find an unbiased estimator of  $\theta^2$ ;
- (b) Find an unbiased estimator of  $\theta(\theta - 1)$ ;
- (c) Are the estimator in parts (a) and (b) consistent?

**Solution.** (a) The independence and unbiasedness properties of  $T_n$  and  $T'_n$  yield that, for each parameter  $\theta$ ,

$$E_\theta[T_n T'_n] = E_\theta[T_n] E_\theta[T'_n] = \theta \cdot \theta = \theta^2$$

and then  $T_n T'_n$  is an unbiased estimator of  $\theta^2$ .

- (b) Using that  $E_\theta[T_n T'_n] = \theta^2$  and  $E_\theta[T_n] = \theta$ , it follows that

$$E_\theta[T_n(T'_n - 1)] = E_\theta[T_n T'_n - T_n] = \theta^2 - \theta = \theta(\theta - 1),$$

that is,  $T_n(T'_n - 1)$  is an unbiased estimator of  $\theta(\theta - 1)$ .

- (c) Since  $T_n$  and  $T'_n$  are consistent estimators of  $\theta$ , combining the convergences  $T_n \xrightarrow{P_\theta} \theta$  and  $T'_n \xrightarrow{P_\theta} \theta$  with the continuity theorem, it follows that

$T_n T'_n \xrightarrow{P_\theta} \theta^2$  and  $T_n(T'_n - 1) \xrightarrow{P_\theta} \theta(\theta - 1)$ , so that the estimators in parts (a) and (b) are consistent.  $\square$

**Exercise 2.3.2.** Let  $X_1, X_2, X_3, \dots$ , be independent and identically distributed random variables with distribution  $\mathcal{N}(\mu, \mu)$  for some  $\mu > 0$ . Find a consistent unbiased estimator of  $\mu^2$ . [Hint:  $E[\bar{X}_n] = \mu$  and  $E[S_n^2] = \mu$ ; consider  $T_n = \bar{X}_n S_n^2$ .]

**Solution.** Recall that in the context of a normal model  $\bar{X}_n$  and  $S_n^2$  are independent; since  $E_\mu[\bar{X}_n] = \mu$  and  $E_\mu[S_n^2] = \mu$  (because in the present model the population variance and mean coincide), it follows that  $E_\mu[\bar{X}_n S_n^2] = E_\mu[\bar{X}_n] E_\mu[S_n^2] = \mu^2$ , and then  $T_n = \bar{X}_n S_n^2$  is an unbiased estimator of  $\mu^2$ . Finally, using that  $X_n \xrightarrow{P_\mu} \mu$  and  $S_n^2 \xrightarrow{P_\mu} \mu$ , it follows that  $T_n \xrightarrow{P_\mu} \mu \cdot \mu = \mu^2$ , by the continuity theorem, and then  $\{T_n\}$  is a consistent sequence of estimators of  $\mu^2$ .  $\square$

**Exercise 2.3.3.** Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from the density  $f(x; \theta) = [(1 - \theta) + \theta/(2\sqrt{x})]I_{[0,1]}(x)$ .

- (a) Show that  $\bar{X}_n$  is a biased estimator of  $\theta$  and find its bias  $b_{\bar{X}_n}(\theta) = b_n(\theta)$ ,
- (b) Does  $\lim_{n \rightarrow \infty} b_n(\theta) = 0$  for all  $\theta$ ?
- (c) Is  $\bar{X}_n$  consistent in mean square?

**Solution.** The mean of the density  $f(x; \theta)$  is

$$\mu(\theta) = \int_{\mathbb{R}} x f(x; \theta) dx = \int_0^1 x[(1 - \theta) + \theta/(2\sqrt{x})] dx = \frac{1 - \theta}{2} + \frac{\theta}{3} = \frac{1}{2} - \frac{\theta}{6}.$$

(a) Since  $E_\theta[\bar{X}_n] = \mu(\theta) \neq \theta$ , the sample mean  $\bar{X}_n$  is a biased estimator of  $\theta$ , and  $b_n(\theta) = \mu(\theta) - \theta = 1 - 7\theta/6$

(b) Notice that  $b_n(\theta) = 1 - 7\theta/6 \neq 0$  for all  $\theta \in [0, 1]$  does not depend on  $n$ , so that  $\lim_{n \rightarrow \infty} b_n(\theta) = 1 - 7\theta/6$ , and then  $b_n(\theta)$  does not converge to zero at any parameter value; in particular, considering  $\bar{X}_n$  as an estimator of  $\theta$ , the sequence  $\{\bar{X}_n\}$  is not asymptotically unbiased.

(c) The sequence  $\{\bar{X}_n\}$  is not consistent in mean square; indeed  $E_\theta[(\bar{X}_n - \theta)^2] \geq b_n^2(\theta)$ , and then  $E_\theta[(\bar{X}_n - \theta)^2]$  does not converges to zero as  $n \rightarrow \infty$ , by part (b).  $\square$

**Exercise 2.3.4.** Let  $X_1, X_2, \dots, X_n$  be independent and identically distributed Poisson random variables with parameter  $\lambda > 0$ . Show that  $T_n = \overline{X_n}^2 - \overline{X_n}$  is a biased estimator of  $\lambda^2$ , find its bias  $b_n(\lambda)$  and hence, find an unbiased estimator of  $\lambda^2$ . Does  $\lim_{n \rightarrow \infty} b_n(\lambda) = 0$  for all  $\lambda$ ?

**Solution.** Recall that for a *Poisson*( $\lambda$ ) distribution the mean  $\mu(\lambda)$  and the variance  $\sigma(\lambda)^2$  are equal to  $\lambda$ . Thus,  $E_\lambda[\overline{X_n}] = \mu(\lambda) = \lambda$  and  $E_\lambda[\overline{X_n}^2] = \text{Var}_\lambda[\overline{X_n}] + (E_\lambda[\overline{X_n}])^2 = \sigma(\lambda)^2/n + \mu(\lambda)^2 = \lambda/n + \lambda^2$ . Thus,

$$E_\lambda[T_n] = E_\lambda[\overline{X_n}^2 - \overline{X_n}] = (\lambda/n + \lambda^2) - \lambda.$$

Thus, as an estimator of  $\lambda^2$ ,  $T_n$  is a biased estimator, and its bias function, which is given by  $b_n(\lambda) = E_\lambda[T_n] - \lambda^2 = \lambda/n - \lambda$  converges to  $\lambda \neq 0$  as  $n$  goes to  $\infty$ . To find an unbiased estimator of  $\lambda^2$ , recall that  $E_\lambda[\overline{X_n}^2] = \lambda^2 + \lambda/n$ , and combine this equality with  $E_\lambda[\overline{X_n}/n] = \lambda/n$  to conclude that  $E_\lambda[\overline{X_n}^2 - \overline{X_n}/n] = \lambda^2$ , showing that  $\overline{X_n}^2 - \overline{X_n}/n$  is an unbiased estimator of  $\lambda$ .  $\square$

**Exercise 2.3.5.** Let  $X_1, X_2, \dots, X_n$  be independent random variables each with the same ‘displaced Laplace density’

$$f(x; \theta) = \frac{1}{2}e^{-|x-\theta|}, \quad x \in \mathbb{R},$$

where the parameter  $\theta$  belongs to  $\mathbb{R}$ . If  $Y_1 \leq Y_2 \leq \dots \leq Y_n$  are the order statistics, show that  $T_n = (Y_1 + Y_n)/2$  is an unbiased estimator of  $\theta$ .  $\square$

**Solution.** The key fact to keep in mind is that the underlying density is symmetric about  $\theta$ , so that  $X_i - \theta$  and  $\theta - X_i$  have the same Laplace density  $f(x) = (1/2)e^{-|x|}$ . Using the independence of the variables  $X_i$ , it follows that

$$(X_1 - \theta, X_2 - \theta, \dots, X_n - \theta) \stackrel{d}{=} (\theta - X_1, \theta - X_2, \dots, \theta - X_n),$$

a relation that, after applying the minimum functions in both sides, leads to

$$\min\{X_i - \theta, i = 1, 2, \dots, n\} \stackrel{d}{=} \min\{\theta - X_i, i = 1, 2, \dots, n\}.$$

Notice now that

$$\min\{X_i - \theta, i = 1, 2, \dots, n\} = \min\{X_i, i = 1, 2, \dots, n\} - \theta = Y_1 - \theta$$



whereas

$$\begin{aligned}\min\{\theta - X_i, i = 1, 2, \dots, n\} &= \theta + \min\{-X_i, i = 1, 2, \dots, n\} \\ &= \theta - \max\{X_i, i = 1, 2, \dots, n\} \\ &= \theta - Y_n.\end{aligned}$$

Combining the three last displays, it follows that

$$Y_1 - \theta \stackrel{d}{=} \theta - Y_n,$$

and then both sides in this relation have the same expectation, that is,  $E[Y_1 - \theta] = E[\theta - Y_n]$ . Therefore,  $E[Y_1 + Y_n] = 2\theta$ , *i.e.*,  $E_\theta[(Y_1 + Y_n)/2] = \theta$ , showing that  $T_n = (Y_1 + Y_n)/2$  is an unbiased estimator of  $\theta$ .  $\square$

**Exercise 2.3.6.** Let  $X_1, X_2, \dots, X_n$  be independent random variables each with density  $f(x; \theta) = (1/\theta)I_{[\theta, 2\theta]}(x)$ .

(a) Show that  $Y_1, Y_n$  are biased estimators of  $\theta$ , and find their respective biases. Do these biases converge to zero as  $n \rightarrow \infty$ ?

(b) Based on part (a), find unbiased estimators of  $\theta$  based on  $Y_1$  alone,  $Y_n$  alone, and a linear combinations of  $Y_1$  and  $Y_n$ .

(c) Two intuitive estimators of  $\theta$  are  $T_n = Y_n - Y_1$  and  $T'_n = (Y_n + Y_1)/3$ . Show that  $T_n$  is biased but  $T'_n$  is unbiased.

**Solution.** (a) The expectations  $E_\theta[Y_1]$  and  $E_\theta[Y_n]$  are required to evaluate the biases of  $Y_1$  and  $Y_n$  as estimators of  $\theta$ . To compute these quantities the densities of  $Y_1$  and  $Y_n$  will be determined. First, notice that the distribution function  $F(x; \theta)$  of the density  $f(x; \theta)$  satisfies

$$F(x; \theta) = (x - \theta)/\theta, \quad x \in [\theta, 2\theta]$$

and, using the formula for the density of  $Y_1$ , it follows that

$$\begin{aligned}f_{Y_1}(y; \theta) &= nf(y; \theta)(1 - F(y; \theta))^{n-1} \\ &= \frac{n}{\theta}[1 - (y - \theta)/\theta]^{n-1} \\ &= \frac{n(2\theta - y)^{n-1}}{\theta^n}, \quad \theta \leq y \leq 2\theta.\end{aligned}$$

Therefore,

$$E_{\theta}[2\theta - Y_1] = \int_{\theta}^{2\theta} (2\theta - y) \frac{n(2\theta - y)^{n-1}}{\theta^n} dy = \frac{n}{n+1}\theta$$

and then

$$E_{\theta}[Y_1] = \theta \frac{n+2}{n+1}, \quad b_{Y_1 n}(\theta) = E_{\theta}[Y_1] - \theta = \frac{\theta}{n+1} \quad (2.3.1)$$

Similarly, using the formula for the density of  $Y_n$

$$\begin{aligned} f_{Y_n}(y; \theta) &= n f(y; \theta) (F(y; \theta))^{n-1} \\ &= \frac{n}{\theta} [(y - \theta)/\theta]^{n-1} \\ &= \frac{n(y - \theta)^{n-1}}{\theta^n}, \quad \theta \leq y \leq 2\theta, \end{aligned}$$

so that

$$E_{\theta}[Y_n - \theta] = \int_{\theta}^{2\theta} (y - \theta) \frac{n(y - \theta)^{n-1}}{\theta^n} dy = \frac{n}{n+1}\theta.$$

Consequently,

$$E_{\theta}[Y_n] = \theta \frac{2n+1}{n+1}, \quad b_{Y_n n}(\theta) = E_{\theta}[Y_n] - \theta = \theta \frac{n}{n+1} \quad (2.3.2)$$

From (2.3.1) and (2.3.2) it follows that  $Y_1$  and  $Y_n$  are biased estimators of  $\theta$ . Since  $\lim_{n \rightarrow \infty} b_{Y_1 n}(\theta) = \lim_{n \rightarrow \infty} [\theta/(n+1)] = 0$ ,  $Y_1$  is asymptotically unbiased; on the other hand,  $\lim_{n \rightarrow \infty} b_{Y_n n}(\theta) = \lim_{n \rightarrow \infty} [n\theta/(n+1)] = \theta$ , so that  $Y_n$  is not asymptotically unbiased.

(b) The first equations in (2.3.1) and (2.3.2) yield that the random variables  $\tilde{Y}_1 = (n+1)\theta/n$  and  $\tilde{Y}_n = [(n+1)/(2n+1)]$  are unbiased estimators of  $\theta$ , which are based on  $Y_1$  and  $Y_n$  alone, respectively. On the other hand,

$$E_{\theta}[Y_1 + Y_n] = \theta \frac{n+2}{n+1} + \theta \frac{2n+1}{n+1} = 3\theta$$

and then the linear combination  $(Y_1 + Y_n)/3$  is an unbiased estimator of  $\theta$ .

(c) The numbered relations in part (a) yield that

$$E_{\theta}[T_n] = E_{\theta}[Y_n - Y_1] = \theta \frac{2n+1}{n+1} - \theta \frac{n+2}{n+1} = \theta \frac{n-1}{n+1} \neq \theta,$$

and

$$E_{\theta}[T'_n] = E_{\theta}[(Y_n + Y_1)/3] = \frac{1}{3} \cdot \theta \frac{2n+1}{n+1} + \frac{1}{3} \cdot \theta \frac{n+2}{n+1} = \theta \frac{3n+3}{3(n+1)} = \theta;$$

thus, as estimators of  $\theta$ ,  $T_n$  is biased, whereas  $T'_n$  is unbiased.  $\square$

## 2.4. Additional Examples

This section contains additional examples concerning basic computations and the basic concepts introduced in this chapter.

**Exercise 2.4.1.** (a) Let  $X$  have density  $f(x; \theta) = [2/(1 - \theta)^2](x - \theta)I_{(\theta, 1)}$ , where  $\theta \in [0, 1)$ . Show that  $E_{\theta}[X - \theta] = 2(1 - \theta)/3$ , and hence find an unbiased estimator of  $\theta$  based on a sample of size 1.

(b) If  $X_1, X_2, \dots, X_n$  is a random sample of size  $n$  from the density in part (a), find a function of  $\bar{X}_n$  that is unbiased for  $\theta$ , and also find the bias of  $\bar{X}_n$ .

(c) Let  $Y_1 \leq Y_2 \leq \dots \leq Y_n$  be the order statistics of the sample in part (b). Find  $E_{\theta}[Y_1]$ .

**Solution.** (a) Notice that

$$E_{\theta}[X - \theta] = \int_{\mathbb{R}} (x - \theta)f(x; \theta) dx = [2/(1 - \theta)^2] \int_{\theta}^1 (x - \theta)^2 dx = \frac{2}{3}(1 - \theta);$$

hence, the mean of the density  $f(x; \theta)$  is

$$\mu(\theta) = E_{\theta}[X] = \frac{2}{3} + \frac{\theta}{3}.$$

and  $E_{\theta}[3X - 2] = \theta$ , that is,  $T = 3X_1 - 2$  is an unbiased estimator of  $\theta$  based on a sample of size 1.

(b) Because the expectation of the sample average equals the population mean, part (a) yields that  $E_{\theta}[\bar{X}_n] = (2 + \theta)/3$ , that is,  $E_{\theta}[3\bar{X}_n - 2] = \theta$ , so that  $\bar{T}_n = 3\bar{X}_n - 2$  is a function of  $\bar{X}_n$  and is an unbiased estimator of  $\theta$ . The bias of  $\bar{X}_n$  as an estimator of  $\theta$  is  $b_{\bar{X}_n}(\theta) = E_{\theta}[\bar{X}_n] - \theta = 2(1 - \theta)/3$ .

(c) To evaluate  $E_{\theta}[Y_1]$  it is necessary to determine the density of  $Y_1$ . Observe that the distribution function of the density  $f(x; \theta)$  satisfies

$$F(x; \theta) = \frac{(x - \theta)^2}{(1 - \theta)^2}, \quad \theta \leq x \leq 1.$$

An application of the formula for the density of  $Y_1$  yields that

$$\begin{aligned} f_{Y_1}(y; \theta) &= nf(y; \theta)[1 - F(y; \theta)]^{n-1} \\ &= n \frac{2(y - \theta)}{(1 - \theta)^2} \left[ 1 - \frac{(y - \theta)^2}{(1 - \theta)^2} \right]^{n-1}, \quad \theta \leq y \leq 1, \end{aligned}$$

an expression the leads to

$$E_\theta[Y_1 - \theta] = \int_\theta^1 (y - \theta) \cdot n \frac{2(y - \theta)}{(1 - \theta)^2} \left[ 1 - \frac{(y - \theta)^2}{(1 - \theta)^2} \right]^{n-1} dy.$$

Changing the variable in the integral to  $z = (y - \theta)/(1 - \theta)$ , and observing that  $dy = (1 - \theta)dz$  and that  $z = 0$  when  $y = \theta$  and  $z = 1$  when  $y = 1$ , it follows that

$$E_\theta[Y_1 - \theta] = 2n(1 - \theta) \int_0^1 z^2 [1 - z^2]^{n-1} dz.$$

To obtain an explicit formula, change the variable in this last integral by setting  $w = z^2$  to obtain, using that  $z = w^{1/2}$  and  $dz = (1/2)w^{-1/2}$ , that

$$E_\theta[Y_1 - \theta] = n(1 - \theta) \int_0^1 w^{3/2-1} (1 - w)^{n-1} dw.$$

Recall now that  $\int_0^1 x^{\alpha-1} (1 - x)^{\beta-1} = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta)$ , and combine this expression with the previous display to obtain

$$E_\theta[Y_1 - \theta] = \frac{n(1 - \theta)\Gamma(3/2)\Gamma(n)}{\Gamma(n + 3/2)}. \quad (2.4.1)$$

The right-hand side can be simplified by observing that

$$\begin{aligned} \Gamma(3/2) &= (1/2)\Gamma(1/2) = \sqrt{\pi}/2 \\ \Gamma(n) &= (n - 1)! \\ \Gamma(n + 3/2) &= (n + 1/2)\Gamma(n + 1/2) \\ &= (n + 1/2)(n - 1/2)\Gamma(n - 1/2) \\ &\vdots \\ &= (n + 1/2)(n - 1/2) \cdots (1/2)\Gamma(1/2) \\ &= \left(\frac{2n + 1}{2}\right) \left(\frac{2n - 1}{2}\right) \cdots \left(\frac{1}{2}\right) \sqrt{\pi} \\ &= \frac{(2n + 1)(2n - 1) \cdots 1}{2^n} \sqrt{\pi} \\ &= \frac{(2n + 1)(2n)(2n - 1)(2n - 2) \cdots 2 \cdot 1}{2^n(2n)(2n - 2) \cdots 2} \sqrt{\pi} \\ &= \frac{(2n + 1)!}{2^{2n}n!} \sqrt{\pi}. \end{aligned}$$

Combining these expressions with (2.4.1) it follows that

$$\begin{aligned} E_\theta[Y_1 - \theta] &= \frac{n(1 - \theta)[\sqrt{\pi}/2](n - 1)!}{[(2n + 1)!/2^{2n}n!]\sqrt{\pi}} \\ &= \frac{1 - \theta}{2(2n + 1)} \cdot \frac{1}{[(2n)!/2^{2n}(n!)^2]} \\ &= \frac{1 - \theta}{2(2n + 1)} \cdot \frac{2^{2n}}{\binom{2n}{n}}. \end{aligned}$$

Thus,

$$E_\theta[Y_1] = \theta + \frac{1 - \theta}{2(2n + 1)} \cdot \frac{2^{2n}}{\binom{2n}{n}},$$

concluding the argument.  $\square$

**Exercise 2.4.2.** Let  $T_{1n}$  and  $T_{2n}$  be independent unbiased estimators of  $\theta$ , with  $\text{Var}_\theta[T_{1n}] = \sigma_{1n}^2$  and  $\text{Var}_\theta[T_{2n}] = \sigma_{2n}^2$ . For each  $\alpha \in \mathbb{R}$  show that  $T_{3n} = \alpha T_{1n} + (1 - \alpha)T_{2n}$  is an unbiased estimator of  $\theta$  and find the value of  $\alpha$  for which  $\text{Var}_\theta[T_{3n}]$  is minimum.

**Solution.** Just notice that  $E_\theta[T_{3n}] = \alpha E_\theta[T_{1n}] + (1 - \alpha)E_\theta[T_{2n}] = \alpha\theta + (1 - \alpha)\theta = \theta$ , so that  $T_{3n}$  is an unbiased estimator of  $\theta$ . On the other hand, using the independence of  $T_{1n}$  and  $T_{2n}$ , the variance of  $T_{3n}$  is given by  $\text{Var}_\theta[T_{3n}] = \text{Var}_\theta[\alpha T_{1n} + (1 - \alpha)T_{2n}] = \alpha^2 \text{Var}_\theta[T_{1n}] + (1 - \alpha)^2 \text{Var}_\theta[T_{2n}] = \alpha^2 \sigma_{1n}^2 + (1 - \alpha)^2 \sigma_{2n}^2$ ; the value of  $\alpha$  that minimizes this variance is the solution of  $2\alpha\sigma_{1n}^2 - 2(1 - \alpha)\sigma_{2n}^2 = 0$ , that is,

$$\alpha^* = \frac{\sigma_{2n}^2}{\sigma_{1n}^2 + \sigma_{2n}^2} \quad \text{and} \quad 1 - \alpha^* = \frac{\sigma_{1n}^2}{\sigma_{1n}^2 + \sigma_{2n}^2}.$$

Notice that when  $\alpha = \alpha^*$ , the statistic that receives the largest weight is the one with smaller variance.  $\square$

**Exercise 2.4.3.** Let  $X_1, X_2, \dots, X_n$  be independent random variables each one with distribution *Gamma*( $\alpha, \lambda$ ), which has density

$$f(x; \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} I_{(0, \infty)}(x),$$

where  $\alpha$  and  $\lambda$  are positive. Suppose that  $\alpha$  is known, and define

$$\beta \equiv \beta(\lambda) = 1/\lambda, \quad \text{and} \quad T_n = \bar{X}_n/\alpha.$$

(a) Show that  $T_n$  is an unbiased estimator of  $\beta$  which is consistent in mean square.

(b) Show that  $(X_1^2 + X_2^2 + \cdots + X_n^2)/[n\alpha(\alpha + 1)]$  is unbiased and consistent as estimator of  $\beta^2$ .

**Solution.** To begin with, recall that the first and second moments of the *Gamma*( $\alpha, \lambda$ ) distribution are given by

$$E_\lambda[X_1] = \frac{\alpha}{\lambda} = \alpha\beta, \quad \text{and} \quad E_\lambda[X_1^2] = \frac{\alpha(\alpha + 1)}{\lambda^2} = \alpha(\alpha + 1)\beta^2, \quad (2.4.2)$$

relations that yield

$$\text{Var}_\lambda[X_1] = \frac{\alpha}{\lambda^2} = \alpha\beta^2. \quad (2.4.3)$$

(a) The first equation in (2.4.2) yields that  $E_\lambda[\bar{X}_n] = \alpha\beta$ , and then  $E_\lambda[T_n] = E_\lambda[\bar{X}_n/\alpha] = \beta$ , that is,  $T_n$  is an unbiased estimator of  $\beta$ . On the other hand, from (2.4.3) it follows that  $\text{Var}_\lambda[\bar{X}_n] = \text{Var}_\lambda[X_1]/n = \alpha\beta^2/n$ , and then

$$E_\lambda[(T_n - \beta)^2] = \text{Var}_\lambda[T_n] = \text{Var}_\lambda[\bar{X}_n/\alpha] = \frac{1}{\alpha^2} \text{Var}_\lambda[\bar{X}_n] = \frac{\beta^2}{n\alpha} \rightarrow 0,$$

so that  $T_n$  is consistent in mean square as estimator of  $\beta$ .

(b) The second equality in (2.4.2) and the law of large numbers together yield that

$$E_\lambda \left[ \frac{X_1^2 + X_2^2 + \cdots + X_n^2}{n} \right] = \alpha(\alpha + 1)\beta^2, \\ \frac{X_1^2 + X_2^2 + \cdots + X_n^2}{n} \xrightarrow{P_\lambda} \alpha(\alpha + 1)\beta^2$$

and then

$$E_\lambda \left[ \frac{X_1^2 + X_2^2 + \cdots + X_n^2}{n\alpha(\alpha + 1)} \right] = \beta^2, \quad \text{and} \quad \frac{X_1^2 + X_2^2 + \cdots + X_n^2}{n\alpha(\alpha + 1)} \xrightarrow{P_\lambda} \beta^2$$

showing that  $(X_1^2 + X_2^2 + \cdots + X_n^2)/[n\alpha(\alpha + 1)]$  is an unbiased estimator of  $\beta^2$ , and that the sequence  $\{(X_1^2 + X_2^2 + \cdots + X_n^2)/[n\alpha(\alpha + 1)]\}$  estimates  $\beta^2$  consistently.  $\square$

**Exercise 2.4.4.** Let  $X_1, X_2, \dots, X_n$  be independent random variables each with density  $\mathcal{N}(\mu, 1)$ . Find an unbiased estimator of  $\mu^2$  that is a function of  $\bar{X}_n$ .

**Solution.** Notice that  $E_\mu[\bar{X}_n^2] = \text{Var}_\mu[\bar{X}_n] + E_\mu[\bar{X}_n]^2 = 1/n + \mu^2$ , so that  $E_\mu[\bar{X}_n^2 - 1/n] = \mu^2$ , that is,  $T_n = \bar{X}_n^2 - 1/n$  is an unbiased estimator of  $\mu^2$ .  $\square$

**Exercise 2.4.5.** Let  $X_1, X_2, \dots, X_n$  be independent random variables with *Exponential*( $\lambda$ ) distribution, which has density

$$f(x; \lambda) = \lambda e^{-\lambda x} I_{(0, \infty)}(x),$$

where  $\lambda > 0$ . Note that  $E_\lambda[X_i] = 1/\lambda$ .

(a) An intuitive estimator for  $\lambda$  is  $1/\bar{X}_n$ . Show that this estimator is biased, and compute the bias  $b_{1/\bar{X}_n}(\lambda)$ .

(b) Based on part (a), find an unbiased estimator of  $\lambda$ .

**Solution.** Let  $n$  be a fixed positive integer and notice that

$$Y := X_1 + \dots + X_n \sim \text{Gamma}(n, \lambda),$$

so that

$$\begin{aligned} E_\lambda[1/Y] &= \int_0^\infty \frac{1}{y} \frac{\lambda^n}{\Gamma(n)} y^{n-1} e^{-\lambda y} dy \\ &= \frac{\lambda^n}{\Gamma(n)} \int_0^\infty y^{n-2} e^{-\lambda y} dy \\ &= \frac{\lambda^n}{\Gamma(n)} \cdot \frac{\Gamma(n-1)}{\lambda^{n-1}} = \frac{\lambda}{n-1} \end{aligned}$$

Hence,

$$E_\lambda[1/\bar{X}_n] = E_\lambda[n/Y] = \frac{n\lambda}{n-1} = \lambda + \frac{\lambda}{n-1}; \quad (2.4.4)$$

this relation shows that  $1/\bar{X}_n$  is a biased estimator of  $\lambda$ , with bias  $b_{1/\bar{X}_n}(\lambda) = \lambda/(n-1)$ .

(b) Equality (2.4.4) yields that  $E_\lambda[(n-1)/(n\bar{X}_n)] = \lambda$ , so that

$$T_n = (n-1)/(X_1 + X_2 + \dots + X_n)$$

is an unbiased estimator of  $\lambda$ .  $\square$

**Exercise 2.4.6.** Let  $X_1, X_2, \dots, X_n$  be independent and identically distributed with mean  $\mu$  and variance  $\sigma^2$ . Show that  $S_n^2$  and  $\bar{X}_n$  are unbiased and consistent estimators of  $\sigma^2$  and  $\mu$ , respectively.

**Solution.** Using that  $E[X_i] = \mu$  for all  $i$ , the linearity of the expectation yield that

$$\begin{aligned} E[\bar{X}_n] &= E\left[\frac{X_1 + X_2 + \dots + X_n}{n}\right] \\ &= \frac{E[X_1] + E[X_2] + \dots + E[X_n]}{n} = \frac{n\mu}{n} = \mu; \end{aligned}$$

thus,  $\bar{X}_n$  is an unbiased estimator of  $\mu$  whereas, by the law of large numbers,  $\bar{X}_n \xrightarrow{P} \mu$ , so that the sequence  $\{\bar{X}_n\}$  estimates consistently the population mean  $\mu$ .

Next, recall that

$$E(X_i^2) = \text{Var}[X_i] + E[X_i]^2 = \sigma^2 + \mu^2;$$

also, it is known that  $\text{Var}[\bar{X}_n] = \sigma^2/n$ , and then

$$E(\bar{X}_n^2) = \text{Var}[\bar{X}_n] + E[\bar{X}_n]^2 = \frac{\sigma^2}{n} + \mu^2;$$

Combining the two last displays with

$$\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \sum_{i=1}^n X_i^2 - n\bar{X}_n^2,$$

it follows that

$$\begin{aligned} E\left[\sum_{i=1}^n (X_i - \bar{X}_n)^2\right] &= \sum_{i=1}^n E[X_i^2] - nE[\bar{X}_n^2] \\ &= \sum_{i=1}^n (\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) \\ &= n\sigma^2 + n\mu^2(\sigma^2 + n\mu^2) \\ &= (n-1)\sigma^2. \end{aligned}$$

Hence,

$$E\left[\frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n-1}\right] = \sigma^2,$$



so that  $S_n^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2 / (n - 1)$  is an unbiased estimator of  $\sigma^2$ . To conclude, the consistency of the sequence  $\{S_n^2\}$  will be shown. To achieve this goal, first observe that

$$\begin{aligned} S_n^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \\ &= \frac{1}{n-1} \left[ \sum_{i=1}^n X_i^2 - n\bar{X}_n^2 \right] \\ &= \left( \frac{n}{n-1} \right) \left[ \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2 \right]; \end{aligned} \tag{2.4.5}$$

next, use the law of large numbers applied to the variables  $\{X_i^2\}$  to obtain that

$$\frac{1}{n} \sum_{i=1}^n X_i^2 \xrightarrow{P} E[X_1^2] = \sigma^2 + \mu^2;$$

now, combine the consistency of the sequence  $\{\bar{X}_n\}$  with the continuity theorem to obtain that

$$\bar{X}_n^2 \xrightarrow{P} \mu^2;$$

Using again the continuity theorem, the two last displays lead to

$$\frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2 \xrightarrow{P} [\sigma^2 + \mu^2] - \mu^2 = \sigma^2,$$

and then since  $n/(n-1) \rightarrow 1$ , (2.4.5) yields that  $S_n^2 \rightarrow \sigma^2$ , establishing the consistency of the sequence  $\{S_n^2\}$ .  $\square$

**Exercise 2.4.7.** Let  $X_1, X_2, \dots, X_n$  be a random sample from the triangular density

$$f(x; a, b) = \begin{cases} \frac{x-a}{c}, & \text{if } a \leq x \leq (a+b)/2, \\ \frac{b-x}{c}, & \text{if } (a+b)/2 \leq x \leq b, \\ 0 & \text{otherwise,} \end{cases}$$

where  $a$  and  $b$  are arbitrary real numbers satisfying  $a < b$ , and  $c = c(a, b) = (b-a)^2/4$ . Show that  $\bar{X}_n$  is an unbiased estimator of  $E(X_1)$  (the parental mean), and that  $\text{Var}[\bar{X}_n] = (b-a)^2/(24n)$ .

**Solution.** The specification of  $f(x; a, b)$  (or a sketch of its graph) makes it evident that, as a function of  $x$ ,  $f(\cdot; a, b)$  is symmetric about  $(a + b)/2$ ; this property can be verified analytically as follows:

If  $w \in [0, (b - a)/2]$ , then  $(a + b)/2 + w \in [(a + b)/2, b]$  and

$$f((a + b)/2 + w; a, b) = \frac{b - [w + (a + b)/2]}{c} = \frac{(b - a)/2 - w}{c}.$$

Similarly, when  $w \in [0, (b - a)/2]$ , the inclusion  $(a + b)/2 - w \in [a, (a + b)/2]$  holds, so that

$$f((a + b)/2 - w; a, b) = \frac{[(a + b)/2 - w] - a}{c} = \frac{(b - a)/2 - w}{c}.$$

These two last displays yield that

$$f((a + b)/2 + w; a, b) = \frac{(b - a)/2 - |w|}{c} I_{-(b-a)/2, (b-a)/2}(w), \quad (2.4.6)$$

showing explicitly that  $f(\cdot; a, b)$  is symmetric about  $(a + b)/2$ . Consequently, the mean of the density is  $(a + b)/2$ , that is

$$\mu \equiv \mu(a, b) = \int_{\mathbb{R}} x f(x; a, b) dx = (a + b)/2,$$

and the variance of the density is

$$\begin{aligned} \sigma^2 \equiv \sigma^2(a, b) &= \int_{\mathbb{R}} (x - (a + b)/2)^2 f(x; a, b) dx \\ &= \int_a^b (x - (a + b)/2)^2 f(x; a, b) dx \\ &= \int_{-(b-a)/2}^{(b-a)/2} w^2 f(w + (a + b)/2; a, b) dw \end{aligned}$$

where the change of variable  $w = x - (a + b)/2$  was used to obtain the third equality. Using (2.4.6), it follows that

$$\begin{aligned} \sigma^2 &= \int_{-(b-a)/2}^{(b-a)/2} w^2 \frac{(b - a)/2 - |w|}{c} dx \\ &= 2 \int_0^{(b-a)/2} w^2 \frac{(b - a)/2 - w}{c} dx \\ &= \frac{2}{c} \left[ \frac{[(b - a)/2]^4}{3} - \frac{[(b - a)/2]^4}{4} \right] \\ &= \frac{[(b - a)/2]^4}{6c} = \frac{(b - a)^4}{96c} = \frac{(b - a)^2}{24} \end{aligned}$$

Concerning the consistency of the sequences  $\{\bar{X}_n\}$  and  $\{S_n^2\}$  as estimators of  $\mu$  and  $\sigma^2$ , recall that they are always consistent, as it was shown in Exercise 2.4.6.  $\square$

**Exercise 2.4.8.** In Exercise 2.4.2, suppose that  $\text{Corr}(T_{1n}, T_{2n}) = \rho$ . Find the value of  $\alpha$  for which  $\text{Var}[T_{3n}]$  is minimized.

**Solution.** Recall that  $T_{3n} = \alpha T_{1n} + (1 - \alpha)T_{2n}$ , so that

$$\begin{aligned} \text{Var}[T_{3n}] &= \text{Var}[\alpha T_{1n} + (1 - \alpha)T_{2n}] \\ &= \alpha^2 \text{Var}[T_{1n}] + (1 - \alpha)^2 \text{Var}[T_{2n}] + 2\alpha(1 - \alpha)\text{Cov}(T_{1n}, T_{2n}) \\ &= \alpha^2 \sigma_1^2 + (1 - \alpha)^2 \sigma_2^2 + 2\alpha(1 - \alpha)\sigma_1 \sigma_2 \rho \end{aligned}$$

Taking the derivative with respect to  $\alpha$ , it follows that the value of  $\alpha$  that minimizes this expression is the solution  $\alpha^*$  of the following equation:

$$2\alpha\sigma_1^2 - 2(1 - \alpha)\sigma_2^2 + 2(1 - 2\alpha)\rho\sigma_1\sigma_2 = 0,$$

and then  $\alpha^* = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$ .  $\square$

# Chapter 3

## Maximum Likelihood

This chapter concerns a fundamental technique to construct estimators, namely, the method of maximum likelihood, which generates estimators with good asymptotic properties, as asymptotic normality.

### 3.1. Maximum Likelihood Estimation

In this section a fundamental procedure to obtain an estimator of a parametric function will be presented. The method is based on an intuitive principle that can be roughly described as follows: After observing the value attained by the random observation, say  $\mathbf{X} = \mathbf{x}$ , the estimate of the unknown parameter  $\theta$  is the value  $\hat{\theta}$  in the parameter space that assigns highest probability to the observed data. In other words, under the condition that  $\hat{\theta}$  is the true parameter value, the occurrence of the observed event  $[\mathbf{X} = \mathbf{x}]$  is more likely than under the condition that the true parameter is different from  $\hat{\theta}$ . To establish this idea on firm grounds, it is necessary to define a measure of the likelihood of an observation  $\mathbf{X} = \mathbf{x}$  under the different parameter values. To achieve this goal, consider a statistical model

$$\mathbf{X} \sim P_{\theta}, \quad \theta \in \Theta,$$

and, to begin with, suppose that  $\mathbf{X}$  is a discrete vector. In this case, let

$f_{\mathbf{X}}(\mathbf{x}; \theta) = P_{\theta}[\mathbf{X} = \mathbf{x}]$  be the probability function of  $\mathbf{X}$  under the condition that  $\theta$  is the true parameter value. As a function of  $\theta \in \Theta$ , the value  $f(\mathbf{x}; \theta)$  indicates the probability of observing  $\mathbf{X} = \mathbf{x}$  if the true distribution of  $\mathbf{X}$  is  $P_{\theta}$ , and then is a measure of the ‘likelihood’ of the observation  $\mathbf{x}$  under the condition that  $\theta \in \Theta$  is the true parameter. Thus, the *likelihood function* corresponding to the data  $\mathbf{X} = \mathbf{x}$  is defined by

$$L(\theta; \mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}; \theta), \quad \theta \in \Theta \quad (3.1.1)$$

When  $\mathbf{X}$  is continuous it has a density  $f_{\mathbf{X}}(\mathbf{x}; \theta)$  depending on  $\theta$ , and the likelihood function associated with the observation  $\mathbf{X} = \mathbf{x}$  is also defined by (3.1.1); notice that in this case,  $f(\mathbf{x}; \theta)$  is not a probability. However, suppose that the measurement instrument used to determine the observation has a certain precision  $h$ , where  $h$  is ‘small’, so that when  $\mathbf{X} = \mathbf{x}$  is reported, the practical meaning is that the vector  $\mathbf{X}$  belongs to a ball  $B(\mathbf{x}; h)$  with center  $\mathbf{x}$  and radius  $h$ ; , when  $\theta$  is the true parameter value, the probability of such an event is

$$\int_{\mathbf{y} \in B(\mathbf{x}; h)} f_{\mathbf{X}}(\mathbf{y}; \theta) \, d\mathbf{y}$$

and, if the density  $f_{\mathbf{X}}(\cdot; \theta)$  is continuous, the above integral is approximately equal to

$$\text{Volume of } B(\mathbf{x}; h) f(\mathbf{x}; \theta) = \text{Volume of } B(\mathbf{x}; h) L(\theta; \mathbf{x});$$

it follows that the likelihood function is (approximately) proportional to the probability of observing  $\mathbf{X} = \mathbf{x}$ ; moreover, the proportionality constant does not depend on  $\theta$ , and then when the maximizer of the function  $L(\cdot; \mathbf{X})$  is determined, such a point also maximizes the approximate probability of the event  $[\mathbf{X} \in B(\mathbf{x}, h)]$ .

The *maximum likelihood estimator* of  $\theta$ , hereafter denoted by  $\hat{\theta} \equiv \hat{\theta}(\mathbf{X})$ , is (any) maximizer of the likelihood function  $L(\theta; \mathbf{X})$  as a function of  $\theta$ , that is,  $\hat{\theta}(\mathbf{X})$  satisfies

$$L(\hat{\theta}; \mathbf{X}) \geq L(\theta; \mathbf{X}), \quad \theta \in \Theta. \quad (3.1.2)$$

This maximum likelihood method to construct estimators of  $\theta$  plays a central role in Statistics, and there are, at least, two reasons for its importance: (i) The method is intuitively appealing, and (ii) The procedure generates estimators that, in general, have nice behavior. For instance, as the sample

size increases, the sequence of maximum likelihood is generally consistent, and the estimators are asymptotically unbiased. Moreover, (iii) As it will be seen later, the asymptotic variance of maximum likelihood estimators is minimal.

On the other hand, frequently what is desired is to estimate the value of a parametric function  $g(\theta)$  at the true parameter value. In this context, it is necessary to decide what value  $\hat{g}$  is ‘more likely’ when  $\mathbf{X} = \mathbf{x}$  has been observed. To determine such a value, consider the likelihood function  $L(\cdot; \mathbf{x})$  of the data and define, for each possible value  $\tilde{g}$  of the function  $g(\theta)$ , the *reduced likelihood* corresponding the value  $\tilde{g}$  of  $g(\theta)$  by

$$L_{\tilde{g}}(\mathbf{X}) := \max_{\theta: g(\theta)=\tilde{g}} L(\theta; \mathbf{X}), \quad (3.1.3)$$

so that  $L_{\tilde{g}}(\mathbf{X})$  is the largest likelihood that can be achieved among the parameters  $\theta$  that produce the value  $\tilde{g}$  for  $g(\theta)$ . The maximum likelihood method prescribes to estimate  $g(\theta)$  by the value  $\hat{g}$  that maximizes  $L_{\tilde{g}}(\mathbf{X})$  as a function of  $\tilde{g}$ :

$$L_{\hat{g}}(\mathbf{X}) \geq L_{\tilde{g}}(\mathbf{X}), \quad \tilde{g} \text{ an arbitray value of } g(\theta).$$

The maximizing value can be determined easily. Set

$$\hat{g} = g(\hat{\theta}) \quad (3.1.4)$$

and notice that (3.1.2) and (3.1.3) imply that, for each possible value  $\tilde{g}$  of  $g(\theta)$ ,

$$L(\hat{\theta}; \mathbf{X}) \geq \max_{\theta: g(\theta)=\tilde{g}} L(\theta; \mathbf{X}) = L_{\tilde{g}}(\mathbf{X})$$

and

$$L(\hat{\theta}; \mathbf{X}) = \max_{\theta: g(\theta)=\hat{g}} L(\theta; \mathbf{X}) = L_{\hat{g}}(\mathbf{X})$$

It follows that  $L_{\hat{g}}(\mathbf{X}) \geq L_{\tilde{g}}(\mathbf{X})$ , and then the reduced likelihood is maximized by  $\hat{g}$  in (3.1.4). In short, the maximum likelihood estimator of a parametric function  $g(\theta)$  is  $\hat{g} = g(\hat{\theta})$ , the value that is obtained by evaluating the function  $g$  at the maximum likelihood estimator of  $\theta$ . This result is called *the invariance principle (or property)* of the maximum likelihood estimation procedure. Before concluding this presentation, it is useful to note that, when

the observation vector  $\mathbf{X}$  is a sample  $(X_1, X_2, \dots, X_n)$  of size  $n$  from a population with probability function or density  $f(x; \theta)$ , the likelihood function is given by

$$L(\theta; \mathbf{X}) = \prod_{i=1}^n f(X_i; \theta);$$

since the logarithmic function is strictly increasing, maximizing this product is equivalent to maximizing its logarithm, which is given by

$$\mathcal{L}(\theta; \mathbf{X}) = \sum_{i=1}^n \log(f(X_i; \theta)).$$

In any case, whether  $L(\cdot; \mathbf{X})$  or  $\mathcal{L}(\theta; \mathbf{X})$  is being maximized, the problem of obtaining its maximizer is an interesting one. As it should be expected, the differentiation technique is of central importance in the analysis of this optimization problem. In particular, if the likelihood function is ‘smooth’ as a function of  $\theta$  and the maximizer belongs to the interior of the parameter space, the following *likelihood equation* is satisfied:

$$D_\theta \mathcal{L}(\theta; \mathbf{X}) = 0, \tag{3.1.5}$$

where  $D_\theta$  is the gradient operator, whose components are the partial derivatives with respect to each element of the parameter  $\theta$ ; thus, when  $\theta$  is a vector, (3.1.5) represent a system of equations satisfied by  $\hat{\theta}$ . On the other hand, when  $\hat{\theta}$  belongs to the boundary of the parameter space, the requirement (3.1.5) is no longer necessarily satisfied by the optimizer  $\hat{\theta}$ . The following examples illustrate the application of the maximum likelihood method for the construction of estimators in models that frequently appear in statistics, and show that the application of the technique leads to interesting problems, even for familiar models as the normal one.

## 3.2. The Method in Specific Cases

This section contains examples illustrating the application of the method of maximum likelihood to construct estimators.

**Exercise 3.2.1.** Let  $X_1, X_2, \dots, X_n$  be a random sample from the uniform density in  $(0, \theta)$ , that is,  $f(x; \theta) = (1/\theta)I_{(0, \theta]}(x)$ , where  $\theta \in \Theta = (0, \infty)$ .

Find the maximum likelihood estimator of  $\theta$ , say  $T_n$ , and show that  $\{T_n\}$  is a consistent sequence of estimators.

**Solution.** The likelihood function is given by

$$L(\theta; \mathbf{X}) = \prod_{i=1}^n (1/\theta) I_{(0,\theta)}(X_i) = \begin{cases} 1/\theta^n, & \text{if } 0 < X_i \leq \theta \text{ for } i = 1, 2, \dots, n \\ 0, & \text{otherwise.} \end{cases}$$

From this expression it follows that  $L(\theta; \mathbf{X})$  is maximized by the smallest number  $\theta$  which satisfies  $X_i \leq \theta$  for every  $i$ , and such a number is  $\hat{\theta}_n = \max\{X_1, \dots, X_n\} = X_{(n)}$ , the largest order statistic of the sample. The sequence  $\{\hat{\theta}_n\}$  is consistent; indeed, given  $\theta \in (0, \infty)$  and  $\varepsilon \in (0, \theta)$ ,

$$\begin{aligned} P_\theta[|\hat{\theta}_n - \theta| > \varepsilon] &= P_\theta[\hat{\theta}_n > \theta + \varepsilon] + P_\theta[\hat{\theta}_n < \theta - \varepsilon] \\ &= P_\theta[\hat{\theta}_n < \theta - \varepsilon] \\ &= P_\theta[X_1 \leq \theta - \varepsilon, X_2 \leq \theta - \varepsilon, \dots, X_n \leq \theta - \varepsilon] \\ &= (1 - \varepsilon/\theta)^n \end{aligned}$$

where, to establish the second equality it was used that, when  $\theta$  is the parameter value, the inequality  $\hat{\theta}_n \leq \theta$  always holds, and the last step is due to the fact that  $P_\theta[X_i \leq \theta - \varepsilon] = 1 - \varepsilon/\theta$  for all  $i$ . It follows that  $P_\theta[|\hat{\theta}_n - \theta| > \varepsilon] \rightarrow 0$  as  $n \rightarrow \infty$ , that is,  $\hat{\theta}_n \xrightarrow{P_\theta} \theta$ , establishing the consistency of  $\{\hat{\theta}_n\}$ .  $\square$

**Exercise 3.2.2.** Let  $X_1, X_2, \dots, X_n$  be a random sample from the  $\mathcal{N}(\theta, \theta^2)$  distribution, where  $\theta \in (0, \infty)$ . Find the maximum likelihood estimator of  $\theta$ . Is the sequence  $\{\hat{\theta}_n\}$  consistent?

**Solution.** The likelihood function is given by

$$L(\theta; \mathbf{X}) = \prod_{i=1}^n (1/\sqrt{2\pi}\theta) e^{-(X_i - \theta)^2/[2\theta^2]}$$

and its logarithm is given by

$$\mathcal{L}(\theta; \mathbf{X}) = -\frac{n}{2} \log(2\pi) - n \log(\theta) - \frac{1}{2} \sum_{i=1}^n \left( \frac{X_i - \theta}{\theta} \right)^2$$



Hence

$$\begin{aligned}\partial_\theta \mathcal{L}(\theta; \mathbf{X}) &= -\frac{n}{\theta} + \sum_{i=1}^n \frac{X_i(X_i - \theta)}{\theta^3} \\ &= -\frac{n}{\theta^3}[\theta^2 + m_1\theta - m_2]\end{aligned}$$

where  $m_i$  is the  $i$ th sample moment about 0,  $i = 1, 2$ . From this expression, direct calculations show that the equation  $\partial_\theta \mathcal{L}(\theta; \mathbf{X}) = 0$  is equivalent to  $\theta^2 + m_1\theta - m_2 = 0$ , The unique positive solution of this likelihood equation is

$$\theta^* = \frac{\sqrt{m_1^2 + 4m_2} - m_1}{2} = \frac{4m_2}{2[\sqrt{m_1^2 + 4m_2} + m_1]}.$$

Since  $\partial_\theta \mathcal{L}(\theta; \mathbf{X}) \rightarrow -\infty$  as  $\theta \rightarrow \infty$ , and  $\partial_\theta \mathcal{L}(\theta; \mathbf{X}) \rightarrow +\infty$  as  $\theta \rightarrow 0$ , it follows that  $\theta^*$  is the unique maximizer of the likelihood function, that is,

$$\begin{aligned}\hat{\theta}_n &= \frac{4m_2}{2[\sqrt{m_1^2 + 4m_2} + m_1]} \\ &= \frac{4 \sum_{i=1}^n X_i^2/n}{2[\sqrt{(\sum_{i=1}^n X_i/n)^2 + 4 \sum_{i=1}^n X_i^2/n} + \sum_{i=1}^n X_i/n]}\end{aligned}$$

To analyze the consistency of  $\{\hat{\theta}_n\}$ , recall that the law of large numbers implies that

$$\begin{aligned}\sum_{i=1}^n X_i^2/n &\xrightarrow{P_\theta} E_\theta[X_1^2] = \text{Var}_\theta[X_1] + (E_\theta[X_1])^2 = \theta^2 + \theta^2 = 2\theta^2 \\ \text{and} \\ \sum_{i=1}^n X_i/n &\xrightarrow{P_\theta} E_\theta[X_1] = \theta.\end{aligned}$$

Combining these convergences with the continuity theorem it follows that

$$\hat{\theta}_n \xrightarrow{P_\theta} \frac{4(2\theta^2)}{2[\sqrt{(\theta)^2 + 4(2\theta^2)} + \theta]} = \frac{8\theta^2}{2[\sqrt{9\theta^2} + \theta]} = \theta$$

establishing the consistency of  $\{\hat{\theta}_n\}$ .  $\square$

**Exercise 3.2.3.** Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $m$  from a  $\mathcal{N}(\mu, \sigma^2)$  distribution, and let  $Y_1, Y_2, \dots, Y_n$  be a random sample of size  $n$

from a  $\mathcal{N}(\nu, \sigma^2)$  distribution, where the two samples are independent. Find the maximum likelihood estimator of the overlapping coefficient

$$\Delta(\theta) \equiv \Delta = 2\Phi\left(-\frac{|\nu - \mu|}{\sigma}\right), \quad \theta = (\mu, \nu, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times (0, \infty) = \Theta.$$

Show that, as  $\min\{n, m\} \rightarrow \infty$ , the sequence of maximum likelihood estimators  $\{\hat{\Delta}_{m,n}\}$  is consistent for  $\Delta$ . Also find the maximum likelihood estimator of  $\theta = (\mu, \nu, \sigma^2)$ .

**Solution.** The likelihood function is given by

$$L(\theta; \mathbf{X}, \mathbf{Y}) = \prod_{i=1}^n (1/\sqrt{2\pi}\sigma) e^{-(X_i - \mu)^2/[2\sigma^2]} \prod_{j=1}^m (1/\sqrt{2\pi}\sigma) e^{-(Y_j - \nu)^2/[2\sigma^2]}$$

and its logarithm is given by

$$\mathcal{L}(\theta; \mathbf{X}, \mathbf{Y}) = C - (n + m) \log(\sigma) - \frac{1}{2} \sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^2} - \frac{1}{2} \sum_{j=1}^m \frac{(Y_j - \nu)^2}{\sigma^2},$$

where the term  $C$  does not involve the parameters. The critical points of  $\mathcal{L}(\cdot; \mathbf{X}, \mathbf{Y})$  satisfy

$$\begin{aligned} \partial_\mu \mathcal{L}(\theta; \mathbf{X}, \mathbf{Y}) &= \sum_{i=1}^n \frac{(X_i - \mu)}{\sigma^2} = 0 \\ \partial_\nu \mathcal{L}(\theta; \mathbf{X}, \mathbf{Y}) &= \sum_{j=1}^m \frac{(Y_j - \nu)}{\sigma^2} = 0 \\ \partial_\sigma \mathcal{L}(\theta; \mathbf{X}, \mathbf{Y}) &= -\frac{n + m}{\sigma} + \sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^3} + \sum_{j=1}^m \frac{(Y_j - \nu)^2}{\sigma^3} = 0 \end{aligned}$$

Direct calculations yield that the unique solution  $(\mu_*, \nu_*, \sigma_*)$  of this system is given by

$$\begin{aligned} \mu_* &= \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \\ \nu_* &= \bar{Y}_m = \frac{1}{m} \sum_{j=1}^m Y_j \\ \sigma_*^2 &= \frac{1}{n + m} \left[ \sum_{i=1}^n (X_i - \bar{X}_n)^2 + \sum_{j=1}^m (Y_j - \bar{Y}_m)^2 \right] \\ &= \frac{n}{n + m} \tilde{S}_n^2 X + \frac{m}{n + m} \tilde{S}_m^2 Y \end{aligned}$$

where  $\tilde{S}_{nX}^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2/n$  and  $\tilde{S}_{mY}^2 = \sum_{j=1}^m (Y_j - \bar{Y}_m)^2/m$  are the maximum likelihood estimators of  $\sigma^2$  based only on  $\mathbf{X}$  and  $\mathbf{Y}$ , respectively. Since  $\mathcal{L}(\theta; \mathbf{X}, \mathbf{Y}) \rightarrow -\infty$  when  $|\mu| + |\nu| \rightarrow \infty$  or  $\sigma \rightarrow 0$ , it follows that the above point  $(\mu_*, \nu_*, \sigma_*^2)$  is the maximizer of  $\mathcal{L}(\cdot; \mathbf{X}, \mathbf{Y})$ , that is,

$$\hat{\theta}_{nm} = (\hat{\mu}_{nm}, \hat{\nu}_{nm}, \hat{\sigma}_{nm}^2) = \left( \bar{X}_n, \bar{Y}_m, \frac{n}{n+m} \tilde{S}_{nX}^2 + \frac{m}{n+m} \tilde{S}_{mY}^2 \right)$$

When  $\min\{n, m\} \rightarrow \infty$ , it was shown in Exercise 2.4.6 that the law of large numbers implies that

$$\bar{X}_n \xrightarrow{P_\theta} \mu, \quad \bar{Y}_m \xrightarrow{P_\theta} \nu, \quad \tilde{S}_n^2 \xrightarrow{P_\theta} \sigma^2, \quad \text{and} \quad \tilde{S}_m^2 \xrightarrow{P_\theta} \sigma^2$$

and then, since  $\hat{\sigma}_{nm}^2$  is a convex combination of  $\tilde{S}_n^2$  and  $\tilde{S}_m^2$ ,

$$\hat{\sigma}_{nm}^2 \xrightarrow{P_\theta} \sigma^2.$$

Hence, the sequence  $\{\hat{\theta}_{nm}\}$  is consistent when  $\min\{m, n\}$  goes to  $\infty$ . Since the overlapping coefficient  $\Delta = \Delta(\theta)$  is a continuous function of  $\theta$ , it follows from the above displays and the continuity theorem, that as  $\min\{n, m\} \rightarrow \infty$ ,

$$\begin{aligned} \hat{\Delta}_{nm} &= \Delta(\hat{\theta}_{nm}) \\ &= 2\Phi\left(-\frac{|\bar{X}_n - \bar{Y}_m|}{\hat{\sigma}_{nm}}\right) \xrightarrow{P_\theta} 2\Phi\left(-\frac{|\mu - \nu|}{\sigma}\right) = \Delta(\theta) = \Delta, \end{aligned}$$

establishing that  $\{\hat{\Delta}_{nm}\}$  is a consistent sequence as  $n$  and  $m$  increase.  $\square$

**Exercise 3.2.4.** Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from the gamma density  $f(x; \alpha, \lambda) = \lambda^\alpha x^{\alpha-1} e^{-\lambda x} / \Gamma(\alpha) I_{(0, \infty)}(x)$ , where  $\theta = (\alpha, \lambda) \in \Theta = (0, \infty) \times (0, \infty)$ . Use the approximation

$$\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \approx \log(\alpha) - \frac{1}{2\alpha} \tag{3.2.1}$$

to find an approximate formula for the maximum likelihood estimator  $\hat{\theta}_n = (\hat{\alpha}_n, \hat{\lambda}_n)$ .

**Solution.** Under the condition  $X_i > 0$  for all  $i$  (which in the present context always holds with probability 1), the likelihood function is

$$L(\theta; \mathbf{X}) = \prod_{i=1}^n (\lambda^\alpha / \Gamma(\alpha)) X_i^{\alpha-1} e^{-\lambda X_i}, \quad \theta = (\alpha, \lambda) \in (0, \infty) \times (0, \infty).$$

and its logarithm is given by

$$\mathcal{L}(\theta; \mathbf{X}) = n\alpha \log(\lambda) - n \log(\Gamma(\alpha)) + (\alpha - 1) \sum_{i=1}^n \log(X_i) - \lambda \sum_{i=1}^n X_i$$

Thus, a critical point of  $\mathcal{L}(\cdot; \mathbf{X})$  satisfies

$$\begin{aligned} \partial_\alpha \mathcal{L}(\theta; \mathbf{X}) &= n \log(\lambda) - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^n \log(X_i) = 0, \\ \partial_\lambda \mathcal{L}(\theta; \mathbf{X}) &= n \frac{\alpha}{\lambda} - \sum_{i=1}^n X_i = 0. \end{aligned} \tag{3.2.2}$$

The second equation yields that

$$\frac{\alpha}{\lambda} = \bar{X}_n. \tag{3.2.3}$$

Combining the first equation in (3.2.2) with (3.2.1) it follows that

$$n \log(\lambda) - n \left[ \log(\alpha) - \frac{1}{2\alpha} \right] + \sum_{i=1}^n \log(X_i) \approx 0,$$

that is,

$$-\log\left(\frac{\alpha}{\lambda}\right) + \frac{1}{2\alpha} + \frac{1}{n} \sum_{i=1}^n \log(X_i) \approx 0,$$

a relation that *via* (3.2.3) leads to

$$-\log(\bar{X}_n) - \frac{1}{2\alpha} + \frac{1}{n} \sum_{i=1}^n \log(X_i) \approx 0,$$

and then

$$\hat{\alpha}_n \approx \frac{1}{2 \left[ \sum_{i=1}^n \log(X_i)/n - \log(\bar{X}_n) \right]}.$$

Combining this expression and (3.2.3) it follows that

$$\hat{\lambda}_n \approx \frac{1}{2\bar{X}_n \left[ \sum_{i=1}^n \log(X_i)/n - \log(\bar{X}_n) \right]},$$

concluding the argument.  $\square$

In the following section an the maximum likelihood method is applied to a ‘non-smooth’ model.

### 3.3. Estimation of the Mean of a Laplace Distribution

Maximizing the likelihood function can be a challenging task when it is not smooth. In this section a statistical model based on the Laplace density is studied to illustrate the application of the method, when the likelihood function does not have a derivative at some points.

**Exercise 3.3.1.** Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from the (Laplace) double exponential density with center  $\theta \in \mathbb{R} \equiv \Theta$ , which is given by

$$f(x; \theta) = \frac{1}{2} e^{-|x-\theta|}.$$

Find the maximum likelihood estimator of  $\theta$ .

**Solution.** The likelihood function is

$$L(\theta; \mathbf{X}) = 2^{-n} \prod_{i=1}^n e^{-|X_i - \theta|}.$$

and its logarithm is given by

$$\mathcal{L}(\theta; \mathbf{X}) = -n \log(2) - \sum_{i=1}^n |X_i - \theta|.$$

The main difficulty in this problem is that the absolute value function is not differentiable at every point. Indeed, the mapping  $\theta \mapsto |x - \theta|$  is not differentiable at  $\theta = x$ , whereas

$$\frac{d}{d\theta} |x - \theta| = -\text{sign}(x - \theta), \quad \theta \neq x.$$

where  $\text{sign}(a) = 1$  if  $a > 0$  and  $\text{sign}(a) = -1$  for  $a < 0$ . Notice now that  $\mathcal{L}(\theta; \mathbf{X})$  is a continuous function of  $\theta$  and observe the following facts:

(i) When  $\theta \leq \min\{X_i, i = 1, 2, \dots, n\} = X_{(1)}$ , the relations  $|X_i - \theta| = X_i - \theta$  hold for every  $i$ , and in this case  $\mathcal{L}(\theta; \mathbf{X}) = -\sum_{i=1}^n [X_i - \theta] = n\theta - \sum_{i=1}^n X_i$ ; consequently,

$$\lim_{\theta \rightarrow -\infty} \mathcal{L}(\theta; \mathbf{X}) = -\infty.$$

(ii) For  $\theta \geq \max\{X_i, i = 1, 2, \dots, n\} = X_{(n)}$ , the equalities  $|X_i - \theta| = \theta - X_i$  are always valid, so that  $\mathcal{L}(\theta; \mathbf{X}) = -\sum_{i=1}^n [X_i - \theta] = -n\theta + \sum_{i=1}^n X_i$ ; thus,

$$\lim_{\theta \rightarrow \infty} \mathcal{L}(\theta; \mathbf{X}) = -\infty.$$

These properties (i) and (ii) together with the continuity of  $\mathcal{L}(\theta; \mathbf{X})$  with respect to  $\theta$  yield that, given  $\mathbf{X}$ ,  $\mathcal{L}(\theta; \mathbf{X})$  attains its maximum at some point  $\hat{\theta}_n \in \mathbb{R}$ . To determine such a point, it is convenient to write

$$\mathcal{L}(\theta; \mathbf{X}) = -n \log(2) - \sum_{i=1}^n |X_{(i)} - \theta|$$

where  $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$  are the order statistics of the sample  $X_1, \dots, X_n$ ; this expression for the log-likelihood function is equivalent to the original one, because the vector of order statistics is just a permutation of the original data. Now, let  $\theta \neq X_{(1)}, X_{(2)}, \dots, X_{(n)}$  and notice that

$$\begin{aligned} \partial_\theta \mathcal{L}(\theta; \mathbf{X}) &= \sum_{i=1}^n \text{sign}(X_{(i)} - \theta) \\ &= \#\{i \mid X_{(i)} > \theta\} - \#\{j \mid X_{(j)} < \theta\} \end{aligned}$$

where  $\#A$  stands for the number of elements of the set  $A$ . Hence,

$$\begin{aligned} \theta < X_{(1)} &\Rightarrow \partial_\theta \mathcal{L}(\theta; \mathbf{X}) = n \\ X_{(1)} < \theta < X_{(2)} &\Rightarrow \partial_\theta \mathcal{L}(\theta; \mathbf{X}) = n - 2 \\ X_{(2)} < \theta < X_{(3)} &\Rightarrow \partial_\theta \mathcal{L}(\theta; \mathbf{X}) = n - 4 \\ X_{(3)} < \theta < X_{(4)} &\Rightarrow \partial_\theta \mathcal{L}(\theta; \mathbf{X}) = n - 6 \\ &\vdots \\ X_{(k)} < \theta < X_{(k+1)} &\Rightarrow \partial_\theta \mathcal{L}(\theta; \mathbf{X}) = n - 2k \\ X_{(k+1)} < \theta < X_{(k+2)} &\Rightarrow \partial_\theta \mathcal{L}(\theta; \mathbf{X}) = n - 2(k+1) \\ &\vdots \\ X_{(n-3)} < \theta < X_{(n-2)} &\Rightarrow \partial_\theta \mathcal{L}(\theta; \mathbf{X}) = 6 - n \\ X_{(n-2)} < \theta < X_{(n-1)} &\Rightarrow \partial_\theta \mathcal{L}(\theta; \mathbf{X}) = 4 - n \\ X_{(n-1)} < \theta < X_{(n)} &\Rightarrow \partial_\theta \mathcal{L}(\theta; \mathbf{X}) = 2 - n \\ X_{(n)} < \theta &\Rightarrow \partial_\theta \mathcal{L}(\theta; \mathbf{X}) = -n \end{aligned} \tag{3.3.1}$$

Suppose that  $n \geq 2$  and let  $k^*$  be the largest positive integer such that  $n \geq 2k^*$ , that is,  $k^*$  satisfies

$$n \geq 2k^* \quad \text{and} \quad n - 2(k^* + 1) < 0. \tag{3.3.2}$$

With this notation, (3.3.1) shows that

(a)  $\partial_\theta \mathcal{L}(\theta; \mathbf{X}) \geq 0$  when  $\theta \in (-\infty, X_{(1)}) \cup (X_{(1)}, X_{(2)}) \cup \cdots \cup (X_{(k^*)}, X_{(k^*+1)})$ , and then the continuity of  $\mathcal{L}(\theta; \mathbf{X})$  implies that  $\mathcal{L}(\theta; \mathbf{X})$  is an increasing function of  $\theta$  in the interval  $(-\infty, X_{(k^*+1)}]$ , so that

$$\mathcal{L}(\theta; \mathbf{X}) \leq \mathcal{L}(X_{(k^*+1)}; \mathbf{X}), \quad \theta \in (-\infty, X_{(k^*+1)}].$$

(b)  $\partial_\theta \mathcal{L}(\theta; \mathbf{X}) < 0$  when

$$\theta \in (X_{(k^*+1)}, X_{(k^*+2)}) \cup \cdots \cup (X_{(n-1)}, X_{(n)}) \cup (X_{(n)}, \infty),$$

and then, by continuity of  $\mathcal{L}(\theta; \mathbf{X})$ , the mapping  $\theta \mapsto \mathcal{L}(\theta; \mathbf{X})$  is decreasing in  $\theta \in [X_{(k^*+1)}, \infty)$ . Thus,

$$\mathcal{L}(\theta; \mathbf{X}) \leq \mathcal{L}(X_{(k^*+1)}; \mathbf{X}), \quad \theta \in [X_{(k^*+1)}, \infty).$$

The two last displays together yield that  $\theta \mapsto \mathcal{L}(\theta; \mathbf{X})$  attains its maximum at

$$\hat{\theta}_n = X_{(k^*+1)}. \tag{3.3.3}$$

If the sample size  $n$  is odd, say  $n = 2r + 1$ , then  $k^*$  in (3.3.2) equals  $r$ , and  $X_{(k^*+1)} = X_{(r+1)}$  is the sample median,

$$\hat{\theta}_n = \text{median}(X_1, \dots, X_n) = \text{median}(\mathbf{X})$$

On the other hand, if the sample size  $n$  is even,  $n = 2r$ , then  $k^*$  in (3.3.2) equals  $r$ , and  $\partial_\theta \mathcal{L}(\theta; \mathbf{X})$  is zero in the interval  $(X_{(r)}, X_{(r+1)})$ , and then  $\mathcal{L}(\theta; \mathbf{X})$  is constant on the interval  $\theta \in [X_{(r)}, X_{(r+1)}]$ , and it follows that every point in that interval is a maximizer of  $\mathcal{L}(\theta; \mathbf{X})$ . Notice that when  $n = 2r$  is an even integer, every point in  $[X_{(r)}, X_{(r+1)}]$  is a median of the data, and the above expression for  $\hat{\theta}_n$  remains valid. Summarizing: the maximum likelihood estimator of  $\theta$  is any sample median, and if the sample size  $n$  is even,  $\hat{\theta}_n$  is not unique.  $\square$

### 3.4. The Poisson and Normal Distributions

In this section additional examples about the application of the maximum likelihood method are presented for models involving common distributions.

**Exercise 3.4.1.** Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from the  $Poisson(\lambda)$  distribution, where  $\lambda \in [0, \infty)$ . Find the maximum likelihood estimator of  $p(0) + p(1)$ .

**Solution.** The interesting function must be expressed in terms of the parameter  $\lambda$ . Notice that

$$p(0) + p(1) = P_\lambda[X = 0] + P_\lambda[X = 1] = e^{-\lambda} + \lambda e^{-\lambda} =: g(\lambda).$$

The maximum likelihood estimator of  $g(\lambda)$  will be constructed using the invariance principle: first,  $\hat{\lambda}_n$  will be determined, and then  $\hat{g}_n$  will be obtained by replacing  $\lambda$  by  $\hat{\lambda}_n$  in the above expression for  $g(\lambda)$ . To develop this plan, notice that the likelihood function is

$$L(\lambda; \mathbf{X}) = \prod_{i=1}^n e^{-\lambda} \frac{\lambda^{X_i}}{X_i!} = e^{-n\lambda} \lambda^{\sum_{i=1}^n X_i} \prod_{i=1}^n \frac{1}{X_i!},$$

whose logarithm is given by

$$\mathcal{L}(\lambda; \mathbf{X}) = -n\lambda + \log(\lambda) \sum_{i=1}^n X_i - \sum_{i=1}^n \log(X_i!),$$

Suppose that  $\sum_i X_i > 0$  and observe that in this case  $\mathcal{L}(\lambda; \mathbf{X}) \rightarrow -\infty$  as  $\lambda \rightarrow 0$  or  $\lambda \rightarrow \infty$ , so that  $\lambda \mapsto \mathcal{L}(\lambda; \mathbf{X})$  attains its maximum at some point  $\hat{\lambda}_n \in (0, \infty)$ , which is be a solution of

$$\partial_\lambda \mathcal{L}(\lambda; \mathbf{X}) = -n + \frac{1}{\lambda} \sum_{i=1}^n X_i = 0,$$

an equation that has the unique solution  $\lambda^* = \bar{X}_n$ . Thus,  $\hat{\lambda}_n = \bar{X}_n$ , and then

$$\hat{g}_n = g(\hat{\lambda}_n) = (1 + \hat{\lambda}_n)e^{-\hat{\lambda}_n} = (1 + \bar{X}_n)e^{-\bar{X}_n}. \quad (3.4.1)$$

Consider now the case  $\sum_i X_i = 0$ . In this context,  $\mathcal{L}(\lambda; \mathbf{X}) = -n\lambda$ , so that  $\mathcal{L}(\lambda; \mathbf{X})$  is decreasing as a function of  $\lambda \in [0, \infty)$ , and then attains its maximum at  $\hat{\lambda}_n = 0 = \bar{X}_n$ , and the invariance principle yields that the expression (3.4.1) for  $\hat{g}_n$  is also valid in this context. Since  $\sum_i X_i \geq 0$  with probability 1, it follows that  $\hat{g}_n = (1 + \bar{X}_n)e^{-\bar{X}_n}$ . Notice now the the strong law of large numbers yields that  $\hat{\lambda}_n \xrightarrow{P_\lambda} \lambda$ ; since that function  $g(\lambda)$



is continuous, an application of the continuity theorem yields that  $\hat{g}_n = g(\hat{\lambda}_n) \xrightarrow{P_\lambda} g(\lambda)$ , that is, the sequence  $\{\hat{g}_n\}$  is consistent.  $\square$

**Exercise 3.4.2.** Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $m$  from a  $\mathcal{N}(\mu, \sigma_1^2)$  distribution and, independently, let  $Y_1, Y_2, \dots, Y_n$  be a random sample of size  $n$  from the  $\mathcal{N}(\mu, \sigma_2^2)$  distribution. Find the maximum likelihood estimators of  $\mu, \sigma_1^2, \sigma_2^2$ , and find the variance of these estimators.

**Solution.** A solution to this problem will not be presented. The analysis below shows that finding the maximum likelihood estimator of  $\theta = (\mu, \sigma_1^2, \sigma_2^2)$  requires to solve a cubic equation; although an explicit formula for the solution of a cubic equation is available, it is not simple. The likelihood function is

$$L(\theta; \mathbf{X}, \mathbf{Y}) = \prod_{i=1}^m (1/\sqrt{2\pi}\sigma_1) e^{-(X_i - \mu)^2/[2\sigma_1^2]} \prod_{j=1}^n (1/\sqrt{2\pi}\sigma_2) e^{-(Y_j - \mu)^2/[2\sigma_2^2]}$$

and its logarithm is given by

$$\mathcal{L}(\theta; \mathbf{X}) = C - m \log(\sigma_1) - n \log(\sigma_2) - \frac{1}{2} \sum_{i=1}^m \frac{(X_i - \mu)^2}{\sigma_1^2} - \frac{1}{2} \sum_{j=1}^n \frac{(Y_j - \mu)^2}{\sigma_2^2},$$

where  $C$  stands for a quantity not involving the parameters. Assuming that this function has a maximizer in the parameter space  $\Theta = \mathbb{R} \times (0, \infty) \times (0, \infty)$ , such a point must satisfy the following likelihood system:

$$\begin{aligned} \partial_\mu \mathcal{L}(\theta; \mathbf{X}, \mathbf{Y}) &= \sum_{i=1}^m \frac{(X_i - \mu)}{\sigma_1^2} + \sum_{j=1}^n \frac{(Y_j - \mu)}{\sigma_2^2} = 0 \\ \partial_{\sigma_1} \mathcal{L}(\theta; \mathbf{X}, \mathbf{Y}) &= -\frac{m}{\sigma_1} + \sum_{i=1}^m \frac{(X_i - \mu)^2}{\sigma_1^3} \\ \partial_{\sigma_2} \mathcal{L}(\theta; \mathbf{X}, \mathbf{Y}) &= -\frac{n}{\sigma_2} + \sum_{j=1}^n \frac{(Y_j - \mu)^2}{\sigma_2^3} \end{aligned}$$

The first equation yields that

$$\frac{m(\bar{X}_m - \mu)}{\sigma_1^2} + \frac{n(\bar{Y}_n - \mu)}{\sigma_2^2} = 0$$

that is,

$$m(\bar{X}_m - \mu)\sigma_2^2 + n(\bar{Y}_n - \mu)\sigma_1^2 = 0$$

whereas the last two likelihood equations are equivalent to

$$\begin{aligned}\sigma_1^2 &= \frac{1}{m} \sum_{i=1}^m (X_i - \mu)^2 = \tilde{S}_{X_m}^2 + (\bar{X}_m - \mu)^2 \\ \sigma_2^2 &= \frac{1}{n} \sum_{j=1}^n (Y_j - \mu)^2 = \tilde{S}_{Y_n}^2 + (\bar{Y}_n - \mu)^2\end{aligned}$$

where  $\tilde{S}_{X_m}^2 = \sum_{i=1}^m (X_i - \mu)^2/m$  and  $\tilde{S}_{Y_n}^2 = \sum_{j=1}^n (Y_j - \mu)^2/n$ . The two last displays together lead to

$$m(\bar{X}_m - \mu)[\tilde{S}_{Y_n}^2 + (\bar{Y}_n - \mu)^2] + n(\bar{Y}_n - \mu)[\tilde{S}_{X_m}^2 + (\bar{X}_m - \mu)^2] = 0,$$

a cubic equation in  $\mu$  □

**Exercise 3.4.3.** Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from the truncated Laplace density

$$f(x; \theta) = \frac{1}{2(1 - e^{-\theta})} e^{-|x|} I_{[-\theta, \theta]}(x)$$

where  $\theta \in \Theta = (0, \infty)$ . Find the maximum likelihood estimator of  $\theta$ . Is this estimator unbiased? Consistent?

**Solution.** The likelihood function is given by

$$\begin{aligned}L(\theta; \mathbf{X}) &= \prod_{i=1}^n \frac{1}{2(1 - e^{-\theta})} e^{-|X_i|} I_{[-\theta, \theta]}(X_i) \\ &= \frac{1}{2^n (1 - e^{-\theta})^n} e^{-\sum_{i=1}^n |X_i|} \prod_{i=1}^n I_{[-\theta, \theta]}(X_i)\end{aligned}$$

Observing that

$$I_{[-\theta, \theta]}(x) = 1 \iff -\theta \leq x \leq \theta \iff |x| \leq \theta$$

it follows that

$$L(\theta; \mathbf{X}) = \begin{cases} \frac{1}{2^n (1 - e^{-\theta})^n} e^{-\sum_{i=1}^n |X_i|}, & \text{if } \theta \geq |X_i|, \quad i = 1, 2, \dots, n \\ 0, & \text{otherwise.} \end{cases}$$

Notice now that  $\theta \mapsto (1/(1 - e^{-\theta})^n$  is a decreasing function, a fact that implies that  $L(\theta; \mathbf{X})$  is maximized at the smallest value at which the function is positive, that is,

$$\hat{\theta}_n = \max\{|X_1|, |X_2|, \dots, |X_n|\}.$$

To analyze the bias of  $\hat{\theta}_n$ , notice that  $P_\theta[|X_i| < \theta] = 1$  for every  $i$ , so that  $P_\theta[|X_i| < \theta, i = 1, 2, \dots, n] = 1$ , *i.e.*, for every  $\theta \in \Theta$

$$P_\theta[\hat{\theta}_n < \theta] = 1; \tag{3.4.2}$$

this structural property implies that  $E_\theta[\hat{\theta}_n] < \theta$ , and then  $\hat{\theta}_n$  is a biased estimator of  $\theta$ , and its bias function  $b_{\hat{\theta}_n}(\theta) = E_\theta[\hat{\theta}_n] - \theta$  is negative. To study the consistency, notice that if  $\varepsilon \in (0, \theta)$ , then

$$\begin{aligned} P_\theta[|X_i| \leq \theta - \varepsilon] &= P_\theta[-(\theta - \varepsilon) \leq X_i \leq \theta - \varepsilon] \\ &= \int_{-(\theta - \varepsilon)}^{\theta - \varepsilon} \frac{1}{2(1 - e^{-\theta})} e^{-|x|} dx =: \alpha(\theta, \varepsilon) < 1. \end{aligned}$$

Hence,

$$\begin{aligned} P_\theta[\hat{\theta}_n \leq (\theta - \varepsilon)] &= P_\theta[|X_i| \leq \theta - \varepsilon, i = 1, 2, \dots, n] \\ &= \prod_{i=1}^n P_\theta[|X_i| \leq \theta - \varepsilon] = \alpha(\theta, \varepsilon)^n \rightarrow 0 \quad \text{as } n \rightarrow \infty. \end{aligned}$$

Since (3.4.2) implies that  $P_\theta[\hat{\theta}_n \geq \theta + \varepsilon] = 0$ , it follows that

$$\begin{aligned} P_\theta[|\hat{\theta}_n - \theta| \geq \varepsilon] &= P_\theta[\hat{\theta}_n \leq \theta - \varepsilon] + P_\theta[\hat{\theta}_n \geq \theta + \varepsilon] \\ &= P_\theta[\hat{\theta}_n \leq \theta - \varepsilon] = \alpha(\theta, \varepsilon)^n \rightarrow 0 \quad \text{as } n \rightarrow \infty, \end{aligned}$$

that is,  $\hat{\theta}_n \xrightarrow{P_\theta} \theta$ , so that the sequence  $\{\hat{\theta}_n\}$  is consistent. A natural question is to see whether the sequence  $\{\hat{\theta}_n\}$  is asymptotically unbiased. To study this problem observe that

$$\begin{aligned} |b_{\hat{\theta}_n}(\theta)| &= |E_\theta[\hat{\theta}_n] - \theta| \\ &\leq E_\theta[|\hat{\theta}_n - \theta|] \\ &= E_\theta[|\hat{\theta}_n - \theta|I[|\hat{\theta}_n - \theta| < \varepsilon]] + E_\theta[|\hat{\theta}_n - \theta|I[|\hat{\theta}_n - \theta| \geq \varepsilon]] \\ &\leq \varepsilon + E_\theta[|\hat{\theta}_n - \theta|I[|\hat{\theta}_n - \theta| \geq \varepsilon]] \end{aligned}$$

Observing that  $P_\theta[|\hat{\theta}_n - \theta| \leq \theta] = 1$ , it follows that

$$|b_{\hat{\theta}_n}(\theta)| \leq \varepsilon + \theta E_\theta[I[|\hat{\theta}_n - \theta| \geq \varepsilon]] \leq \varepsilon + \theta \alpha(\theta, \varepsilon)^n$$

and then, because  $\alpha(\theta, \varepsilon)^n \rightarrow 0$ , this implies that  $\limsup_{n \rightarrow \infty} |b_{\hat{\theta}_n}(\theta)| \leq \varepsilon$ ; hence, since  $\varepsilon > 0$  is arbitrary,  $\lim_{n \rightarrow \infty} b_{\hat{\theta}_n}(\theta) = 0$ , that is,  $\{\hat{\theta}_n\}$  is asymptotically unbiased.  $\square$

**Remark 3.4.1.** The above analysis of the unbiasedness property for  $\theta_n$  was not based on a direct computation of the expectation of  $\hat{\theta}_n$ . If an explicit formula for the bias function is required, such an expectation must be calculated using the density of  $\hat{\theta}_n$ , which is determined as follows: Notice that the distribution function of  $|X_i|$  is

$$\begin{aligned} G(x; \theta) &= P_\theta[|X_i| \leq x] = \int_{-x}^x \frac{1}{2(1 - e^{-\theta})} e^{-|t|} I_{[-\theta, \theta]}(t) dt \\ &= \int_0^x \frac{1}{(1 - e^{-\theta})} e^{-t} dt = \frac{1 - e^{-x}}{1 - e^{-\theta}}, \quad x \in [0, \theta] \end{aligned}$$

an expression that renders the following formula for the density of  $|X_i|$ :

$$g(x; \theta) = \frac{e^{-x}}{1 - e^{-\theta}} I_{[0, \theta]}(x).$$

Using the formula for the density of the maximum of independent and identically distributed random variables,

$$f_{\hat{\theta}_n}(x; \theta) = n g(x; \theta) G(x; \theta)^{n-1} = \frac{n e^{-x}}{1 - e^{-\theta}} \left( \frac{1 - e^{-x}}{1 - e^{-\theta}} \right)^{n-1} I_{[0, \theta]}(x).$$

The expectation of  $\hat{\theta}_n$  can be now computed explicitly as follows:

$$\begin{aligned} E_\theta[\hat{\theta}_n] &= \int_0^\theta x \frac{n e^{-x}}{1 - e^{-\theta}} \left( \frac{1 - e^{-x}}{1 - e^{-\theta}} \right)^{n-1} dx \\ &= x \left( \frac{1 - e^{-x}}{1 - e^{-\theta}} \right)^n \Big|_{x=0}^\theta - \int_0^\theta \left( \frac{1 - e^{-x}}{1 - e^{-\theta}} \right)^n dx \\ &= \theta - \int_0^\theta \left( \frac{1 - e^{-x}}{1 - e^{-\theta}} \right)^n dx. \end{aligned}$$

Therefore, the bias function of  $\hat{\theta}_n$  is

$$b_{\hat{\theta}_n}(\theta) = E_\theta[\hat{\theta}_n] - \theta = - \int_0^\theta \left( \frac{1 - e^{-x}}{1 - e^{-\theta}} \right)^n dx, \quad \theta > 0$$

showing explicitly that the bias is always negative. Also, observing that

$$\lim_{n \rightarrow \infty} \left( \frac{1 - e^{-x}}{1 - e^{-\theta}} \right)^n = 0, \quad x \in [0, \theta),$$

the bounded convergence theorem implies that

$$\lim_{n \rightarrow \infty} \int_0^\theta \left( \frac{1 - e^{-x}}{1 - e^{-\theta}} \right)^n dx = 0,$$

a convergence that yields that  $\{\hat{\theta}_n\}$  is asymptotically unbiased.  $\square$

### 3.5. Estimating the Parameters of a Beta Distribution

**Exercise 3.5.1.** Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from the  $Beta(\alpha, \beta)$  density

$$f(x; \alpha, \beta) := \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} I_{(0,1)}(x)$$

where  $\alpha$  and  $\beta$  are positive numbers. In the following questions  $\beta$  is a known number, but  $\alpha \in (0, \infty)$  is unknown.

- (a) Find the maximum likelihood estimator of  $\alpha$  when  $\beta = 1$ ;
- (b) Find the maximum likelihood estimator of  $\alpha$  when  $\beta = 2$ ;
- (c) Find the maximum likelihood estimator of  $\alpha/(1 + \alpha)$  in each of the preceding cases (a) and (b).

**Solution.** Set  $\mathbf{X} = (X_1, X_2, \dots, X_n)$ .

(a) When  $\beta = 1$  the density of each observation  $X_i$  is

$$f(x; \alpha, 1) := \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha)\Gamma(1)} x^{\alpha-1} (1-x)^{1-1} I_{(0,1)}(x) = \alpha x^{\alpha-1} I_{(0,1)}(x).$$

where it was used that  $\Gamma(1) = 1$  and  $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$  to set the second equality. Thus, the likelihood function associated to a sample  $\mathbf{X} \in (0, 1)^n$  is

$$L(\alpha; \mathbf{X}) = \prod_{i=1}^n \alpha X_i^{\alpha-1} = \alpha^n \left( \prod_{i=1}^n X_i \right)^{\alpha-1}$$

whose logarithm is

$$\mathcal{L}(\alpha; \mathbf{X}) = n \log(\alpha) + (\alpha - 1) \sum_{i=1}^n \log(X_i), \quad \alpha \in (0, \infty)$$

Recalling that  $\log(x) \rightarrow -\infty$  as  $x \searrow 0$  and that  $\log(x) < 0$  when  $x \in (0, 1)$ , it follows that  $\mathcal{L}(\alpha; \mathbf{X}) \rightarrow -\infty$  as  $\alpha \searrow 0$  or  $\alpha \rightarrow \infty$ , and then  $\mathcal{L}(\cdot; \mathbf{X})$  has a maximizer  $\hat{\alpha}_n$  in  $(0, \infty)$ . Such a point is a solution of the likelihood equation

$$\partial_\alpha \mathcal{L}(\alpha; \mathbf{X}) = \frac{n}{\alpha} + \sum_{i=1}^n \log(X_i) = 0,$$

whose unique solution is  $\alpha = -n / (\sum_{i=1}^n \log(X_i))$ . Consequently,

$$\hat{\alpha}_n = -\frac{n}{\sum_{i=1}^n \log(X_i)}.$$

(b) Suppose that  $\beta = 2$ . In this case, the density of each observation  $X_i$  is

$$f(x; \alpha, 2) = \frac{\Gamma(\alpha + 2)}{\Gamma(\alpha)\Gamma(2)} x^{\alpha-1} (1-x)^{2-1} I_{(0,1)}(x) = \alpha(\alpha+1)x^{\alpha-1}(1-x)I_{(0,1)}(x);$$

as for the second equality, recall that  $\Gamma(2) = 1$  and  $\Gamma(\alpha + 2) = (\alpha + 1)\alpha\Gamma(\alpha)$ . It follows that the likelihood function associated to a sample  $\mathbf{X} \in (0, 1)^n$  is

$$L(\alpha; \mathbf{X}) = \prod_{i=1}^n \alpha(\alpha+1)X_i^{\alpha-1}(1-X_i) = [\alpha(\alpha+1)]^n \left( \prod_{i=1}^n X_i \right)^{\alpha-1} \prod_{i=1}^n (1-X_i)$$

whose logarithm is given by

$$\mathcal{L}(\alpha; \mathbf{X}) = n \log(\alpha) + n \log(\alpha + 1) + (\alpha - 1) \sum_{i=1}^n \log(X_i) + \sum_{i=1}^n \log(1 - X_i),$$

where  $\alpha \in (0, \infty)$ . As in the previous part, it is not difficult to see that  $\mathcal{L}(\alpha; \mathbf{X}) \rightarrow -\infty$  as  $\alpha \searrow 0$  or  $\alpha \rightarrow \infty$ , so that  $\mathcal{L}(\cdot; \mathbf{X})$  has a maximizer  $\hat{\alpha}_n$  in  $(0, \infty)$  which satisfies that the likelihood equation

$$\partial_\alpha \mathcal{L}(\alpha; \mathbf{X}) = \frac{n}{\alpha} + \frac{n}{\alpha + 1} + \sum_{i=1}^n \log(X_i) = 0;$$

after some simple algebra, this equation is equivalent to  $(2\alpha+1)+\alpha(\alpha+1)Y = 0$ , where  $Y = \sum_{i=1}^n \log(X_i)/n$ . This quadratic equation in  $\alpha$  can be written as  $\alpha^2 Y + \alpha(2+Y) + 1 = 0$ , which has roots

$$\alpha = \frac{-(2+Y) \pm \sqrt{(2+Y)^2 - 4Y}}{2Y} = \frac{-(2+Y) \pm \sqrt{4+Y^2}}{2Y};$$

Recalling that  $Y < 0$ , the root that is positive is given by

$$\alpha = \frac{-(2+Y) - \sqrt{4+Y^2}}{2Y} = \frac{2}{\sqrt{4+Y^2} + (2-Y)};$$

hence,

$$\hat{\alpha}_n = \frac{2}{\sqrt{4 + (\sum_{i=1}^n \log(X_i))^2} + (2 - \sum_{i=1}^n \log(X_i))}.$$

(c) By the invariance principle, the maximum likelihood estimator of  $g(\alpha) = \alpha/(\alpha + 1)$  is given by  $\hat{g} = \frac{\hat{\alpha}_n}{1 + \hat{\alpha}_n}$ .  $\square$

### 3.6. Additional Examples

**Exercise 3.6.1.** Let  $f_1(x)$  and  $f_2(x)$  be two density functions and consider a random sample  $Z_1, Z_2$  of size two of the mixture

$$f(z; \theta) = \theta f_1(z) + (1 - \theta) f_2(z) = f_2(z) + \theta [f_1(z) - f_2(z)],$$

where  $\theta \in [0, 1]$ . Find the maximum likelihood estimator of  $\theta$ .

**Solution.** The likelihood function of the data  $\mathbf{Z} = (Z_1, Z_2)$  is

$$L(\theta; \mathbf{Z}) = [f_2(Z_1) + \theta d(Z_1)][f_2(Z_2) + \theta d(Z_2)], \quad \theta \in [0, 1],$$

where

$$d(z) := f_1(z) - f_2(z).$$

To find the maximizers of  $L(\cdot; \mathbf{Z})$ , consider the following exhaustive cases:

(i)  $d(Z_1)d(Z_2) > 0$ : In this context, the mapping

$$\theta \mapsto [f_2(Z_1) + \theta d(Z_1)][f_2(Z_2) + \theta d(Z_2)]$$

is convex, and its unique critical point is a minimizer. Thus,  $L(\cdot; \mathbf{Z})$  attains its maximum at  $\theta = 0$  or  $\theta = 1$ . Observing that  $L(0; \mathbf{Z}) = f_2(Z_1)f_2(Z_2)$  and  $L(1; \mathbf{Z}) = f_1(Z_1)f_1(Z_2)$ , it follows that

$$\hat{\theta}_2(\mathbf{Z}) = \begin{cases} 1, & \text{if } f_1(Z_1)f_1(Z_2) > f_2(Z_1)f_2(Z_2) \\ 0, & \text{if } f_1(Z_1)f_1(Z_2) < f_2(Z_1)f_2(Z_2) \\ 0 \text{ or } 1, & \text{if } f_1(Z_1)f_1(Z_2) = f_2(Z_1)f_2(Z_2). \end{cases}$$

(ii)  $d(Z_1)d(Z_2) < 0$ : In this framework, the mapping

$$\theta \mapsto [f_2(Z_1) + \theta d(Z_1)][f_2(Z_2) + \theta d(Z_2)]$$

is concave, and attains its maximum (with respect to all the points  $\theta \in \mathbb{R}$ ) at the unique critical point

$$\theta^*(Z) = -\frac{d(Z_1)f_2(Z_2) + d(Z_2)f_2(Z_1)}{2d(Z_1)d(Z_2)}$$

and the maximizer of  $L(\cdot; \mathbf{Z})$  is given by

$$\hat{\theta}_2(\mathbf{Z}) = \begin{cases} \theta^*(\mathbf{Z}), & \text{if } \theta^*(\mathbf{Z}) \in [0, 1] \\ 0, & \text{if } \theta^*(\mathbf{Z}) < 0 \\ 1, & \text{if } \theta^*(\mathbf{Z}) > 1. \end{cases}$$

(iii)  $d(Z_1) = 0$  and  $d(Z_2) \neq 0$ : In this framework,  $L(\theta; \mathbf{Z})$  is a linear function of  $\theta$  with slope  $f_2(Z_1)d(Z_2)$ , and it follows that

$$\hat{\theta}_2(\mathbf{Z}) = \begin{cases} 1, & \text{if } f_2(Z_1)d(Z_2) > 0 \\ 0, & \text{if } f_2(Z_1)d(Z_2) < 0 \\ \text{any point in } [0, 1], & \text{if } f_2(Z_1) = 0. \end{cases}$$

Similarly,

(iv)  $d(Z_1) \neq 0$  and  $d(Z_2) = 0$ : In these circumstances,  $L(\theta; \mathbf{Z})$  is a linear function of  $\theta$  with slope  $f_2(Z_2)d(Z_1)$ , and

$$\hat{\theta}_2(\mathbf{Z}) = \begin{cases} 1, & \text{if } f_2(Z_2)d(Z_1) > 0 \\ 0, & \text{if } f_2(Z_2)d(Z_1) < 0 \\ \text{any point in } [0, 1], & \text{if } f_2(Z_2) = 0. \end{cases}$$

Finally,

(iv)  $d(Z_1) = 0$  and  $d(Z_2) = 0$ : In this case  $L(\theta; \mathbf{Z})$  is a constant function, so that

$$\hat{\theta}_2(\mathbf{Z}) = \text{any point in } [0, 1],$$

and the solution is complete. □

**Exercise 3.6.2.** Let  $X_1, X_1, \dots, X_n$  be a random sample of size  $n$  from the (discrete) *Uniform*  $(\{1, 2, \dots, \theta\})$  distribution on the set  $\{1, 2, \dots, \theta\}$ , whose probability function is given by

$$f(x; \theta) = \frac{1}{\theta} I_{\{1, 2, \dots, \theta\}}(x).$$



Find the maximum likelihood estimator of  $\theta$ , and its mean. Is this estimator unbiased?

**Solution.** Given  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  with positive integer components, the corresponding likelihood function is given as follows: For a positive integer  $\theta$ ,

$$L(\theta; \mathbf{X}) = \prod_{i=1}^n \frac{1}{\theta} I_{\{1,2,3,\dots,\theta\}}(X_i) = \begin{cases} 1/\theta^n, & \text{if } X_i \leq \theta \text{ for all } i = 1, 2, \dots, n, \\ 0, & \text{otherwise,} \end{cases}$$

an expression that can be written as

$$L(\theta; \mathbf{X}) = \begin{cases} 1/\theta^n, & \text{if } \max\{X_i, i = 1, 2, \dots, n\} \leq \theta, \\ 0, & \text{otherwise.} \end{cases}$$

Since  $\theta \mapsto 1/\theta^n$  is a decreasing mapping on the set of positive integers, it follows that the maximizer of  $L(\cdot; \mathbf{X})$  is the minimal value of  $\theta$  at which  $L(\theta; \mathbf{X})$  is positive, that is,

$$\hat{\theta}_n = X_{(n)} = \max\{X_1, X_2, \dots, X_n\}.$$

To find the expectation of  $\hat{\theta}_n$ , first the distribution function of the estimator will be determined. Given a positive integer  $\theta$ , notice that

$$\begin{aligned} P_\theta[\hat{\theta}_n \leq k] &= P_\theta[X_i \leq k, i = 1, 2, \dots, n] \\ &= \prod_{i=1}^k P_\theta[X_i \leq k] = \prod_{i=1}^k \left(\frac{k}{\theta}\right) = \left(\frac{k}{\theta}\right)^n, \quad k = 1, 2, \dots, \theta, \end{aligned}$$

Thus, the probability function of  $\hat{\theta}_n$  is determined by

$$\begin{aligned} f_{\hat{\theta}_n}(k; \theta) &= P_\theta[\hat{\theta}_n = k] \\ &= P_\theta[\hat{\theta}_n \leq k] - P_\theta[\hat{\theta}_n \leq (k-1)] \\ &= \left(\frac{k}{\theta}\right)^n - \left(\frac{k-1}{\theta}\right)^n, \quad k = 1, 2, \dots, \theta, \end{aligned}$$

and then

$$\begin{aligned}
E_\theta[\hat{\theta}_n] &= \sum_{k=1}^{\theta} k P_\theta[\hat{\theta}_n = k] \\
&= \sum_{k=1}^{\theta} k \left[ \left(\frac{k}{\theta}\right)^n - \left(\frac{k-1}{\theta}\right)^n \right] \\
&= \sum_{k=1}^{\theta} \frac{k^{n+1} - k(k-1)^n}{\theta^n} \\
&= \sum_{k=1}^{\theta} \frac{k^{n+1} - (k-1)^{n+1} - (k-1)^n}{\theta^n} \\
&= \sum_{k=1}^{\theta} \frac{k^{n+1} - (k-1)^{n+1}}{\theta^n} - \sum_{k=1}^{\theta} \frac{(k-1)^n}{\theta^n} \\
&= \frac{\theta^{n+1} - (1-1)^{n+1}}{\theta^n} - \sum_{k=1}^{\theta} \frac{(k-1)^n}{\theta^n} \\
&= \theta - \sum_{k=1}^{\theta-1} \frac{k^n}{\theta^n}.
\end{aligned}$$

and it follows that  $\hat{\theta}_n$  is a biased estimator of  $\theta$ .  $\square$

**Exercise 3.6.3.** Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from the *Poisson*( $\lambda$ ) distribution, where  $\lambda \in [0, \infty)$  is unknown.

- (a) Find the maximum likelihood estimator of  $e^{-\lambda}$ .  
(b) Find an unbiased estimator of  $e^{-\lambda}$ .

**Solution.** (a) The maximum likelihood estimator  $\hat{g}_n$  of  $g(\lambda) = e^{-\lambda}$  will be constructed *via* the invariance principle, that is, if  $\hat{\lambda}_n$  is the maximum likelihood estimator of  $\lambda$ , then  $\hat{g}_n = g(\hat{\lambda}_n)$ . To find  $\hat{\lambda}_n$ , notice that, given a sample  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  whose components are nonnegative integers, the corresponding likelihood function is given by

$$L(\lambda; \mathbf{X}) = \prod_{i=1}^n \frac{\lambda^{X_i}}{X_i!} e^{-\lambda} = \lambda^{\sum_{i=1}^n X_i} e^{-n\lambda} \prod_{i=1}^n \frac{1}{X_i!}, \quad \lambda \in [0, \infty)$$

and its logarithm is

$$\mathcal{L}(\lambda; \mathbf{X}) = \log(\lambda) \sum_{i=1}^n X_i - n\lambda + \log \left( \prod_{i=1}^n \frac{1}{X_i!} \right), \quad \lambda \in [0, \infty). \quad (3.6.1)$$

(i) Suppose that  $X_i > 0$  for some  $i$ . In this case, the basic properties of the logarithmic function yield that  $\mathcal{L}(\lambda; \mathbf{X}) \rightarrow -\infty$  as  $\lambda \rightarrow 0$  or as  $\lambda \rightarrow \infty$ . Therefore,  $\mathcal{L}(\cdot; \mathbf{X})$  attains its maximum at some positive point, which satisfies

$$\partial_\lambda \mathcal{L}(\lambda; \mathbf{X}) = \frac{1}{\lambda} \sum_{i=1}^n X_i - n = 0;$$

this equation has the unique solution  $\lambda = \bar{X}_n = \sum_{i=1}^n X_i/n$ . Hence,

$$\hat{\lambda}_n = \bar{X}_n. \quad (3.6.2)$$

(ii) Suppose now that  $X_i = 0$  for all  $i$ . In this context, (3.6.1) shows that the likelihood function reduces to  $\mathcal{L}(\lambda; \mathbf{X}) = -n\lambda$ , and then its maximizer is  $\hat{\lambda}_n = 0 = \bar{X}_n$ . Thus, *in any circumstance*, the maximum likelihood estimator of  $\lambda$  is the sample mean. Thus, for  $g(\lambda) = e^{-\lambda}$ , the maximum likelihood estimator of  $g(\lambda)$  is

$$\hat{g}_n = e^{-\hat{\lambda}_n} = e^{-\bar{X}_n}.$$

It is interesting to observe that this estimator is *biased*. Indeed, using that the population mean of the *Poisson*( $\lambda$ ) distribution is  $\lambda$ , it follows that  $E_\lambda[\bar{X}_n] = \lambda$ , and then observing that the function  $H(x) = e^{-x}$  is strictly convex, Jensen's inequality implies that

$$e^{-\lambda} = H(\lambda) = H(E_\lambda[\bar{X}_n]) < E_\lambda[H(\bar{X}_n)] = E_\lambda[e^{-\bar{X}_n}].$$

(b) To determine an unbiased estimator of  $e^{-\lambda}$ , notice that

$$e^{-\lambda} = P_\lambda[X_1 = 0] = E_\lambda[I[X_1 = 0]].$$

Thus,  $I[X_1 = 0]$  is an unbiased estimator of  $e^{-\lambda}$ ; since all the  $X_i$  have the same distribution, it follows that, for every  $i$ ,  $I[X_i = 0]$  is also an unbiased estimator, and then so is their average  $T = \sum_{i=1}^n I[X_i = 0]/n$ . However, the idea behind this problem is to determine an unbiased estimator of  $\lambda$  which is a function of  $\bar{X}_n$ . Let  $G(\bar{X}_n)$  be such that

$$E_\lambda[G(\bar{X}_n)] = e^{-\lambda} \quad \text{for every } \lambda \in [0, \infty). \quad (3.6.3)$$

Since  $\bar{X}_n = T_n/n$  where  $T_n = X_1 + X_2 + \cdots + X_n \sim \text{Poisson}(n\lambda)$ , it follows that

$$E_\lambda[G(\bar{X}_n)] = \sum_{k=0}^{\infty} G(k/n) P_\lambda[T_n = k] = \sum_{k=0}^{\infty} G(k/n) \frac{(n\lambda)^k}{k!} e^{-n\lambda},$$

and then

$$\begin{aligned}
E_\lambda[G(\bar{X}_n)] = e^{-\lambda} &\iff \sum_{k=0}^{\infty} G(k/n) \frac{(n\lambda)^k}{k!} e^{-n\lambda} = e^{-\lambda} \\
&\iff \sum_{k=0}^{\infty} \frac{G(k/n)n^k}{k!} \lambda^k = e^{(n-1)\lambda} \\
&\iff \sum_{k=0}^{\infty} \frac{G(k/n)n^k}{k!} \lambda^k = \sum_{k=0}^{\infty} \frac{(n-1)^k}{k!} \lambda^k
\end{aligned}$$

where the classical expansion  $e^a = \sum_{k=0}^{\infty} a^k/k!$  was used in the last step. Therefore (3.6.3) is equivalent to

$$\sum_{k=0}^{\infty} \frac{G(k/n)n^k}{k!} \lambda^k = \sum_{k=0}^{\infty} \frac{(n-1)^k}{k!} \lambda^k, \quad \lambda \in [0, \infty).$$

Now, using the known fact that two power series coincide in an interval if and only if they have the same coefficients, this last display is equivalent to

$$\frac{G(k/n)n^k}{k!} = \frac{(n-1)^k}{k!}, \quad k = 0, 1, 2, 3, \dots,$$

that is,

$$G(k/n) = \frac{(n-1)^k}{n^k} = \left(1 - \frac{1}{n}\right)^k, \quad k = 1, 2, 3, \dots$$

Consequently,

$$G(\bar{X}_n) = G(T_n/n) = \left(1 - \frac{1}{n}\right)^{T_n} = \left(1 - \frac{1}{n}\right)^{n\bar{X}_n}$$

is the unique unbiased estimator of  $e^{-\lambda}$  which is a function of  $\bar{X}_n$ .  $\square$

**Exercise 3.6.4.** Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from the *Bernoulli*( $p$ ) distribution, where  $p \in [0, 1]$ , and set  $T_n = X_1 + X_2 + \dots + X_n$ .

- Find the maximum likelihood estimator  $M_n$  of  $pq = p(1-p)$
- Show that  $U_n = T_n(n - T_n)/[n(n-1)]$  is an unbiased estimator of  $pq = p(1-p)$ .
- Show that the maximum likelihood estimator of  $pq$  is biased, but is asymptotically unbiased.

(d) Show that the unbiased estimator of  $pq$  has *larger* variance than the maximum likelihood estimator.

**Solution.** (a) The maximum likelihood estimator of  $p$  is  $\bar{X}_n$ , so that, by the invariance property,  $\bar{X}_n(1 - \bar{X}_n) = M_n$  is the maximum likelihood estimator of  $p(1 - p) = pq$ .

(b) In Exercise 2.2.1 it was shown that  $T_n(T_n - 1)/[n(n - 1)]$  is an unbiased estimator of  $p^2$ . Since  $\bar{X}_n = T_n/n$  is an unbiased estimator of  $p$ , it follows that

$$\begin{aligned} pq &= p(1 - p) \\ &= p - p^2 = E_p \left[ \frac{T_n}{n} - \frac{T_n(T_n - 1)}{n(n - 1)} \right] = E_p \left[ \frac{T_n(n - T_n)}{n(n - 1)} \right] = E_p[U_n], \end{aligned}$$

that is,  $U_n = T_n(n - T_n)/[n(n - 1)]$  is an unbiased estimator of  $pq$ .

(c) Notice that

$$\begin{aligned} M_n &= \bar{X}_n(1 - \bar{X}_n) \\ &= \frac{T_n}{n} \left( 1 - \frac{T_n}{n} \right) = \frac{T_n(n - T_n)}{n^2} = \frac{n - 1}{n} \frac{T_n(T_n - 1)}{n(n - 1)} = \frac{n - 1}{n} U_n. \end{aligned}$$

Hence,

$$E_p[M_n] = E_p \left[ \frac{n - 1}{n} U_n \right] = \frac{n - 1}{n} E_p[U_n] = \frac{n - 1}{n} pq = pq - \frac{pq}{n}.$$

It follows that  $b_{M_n}(p) = E_p[M_n] - pq = -pq/n$ , so that  $M_n$  is biased; since  $b_{M_n}(p) = pq/n \rightarrow 0$  as  $n \rightarrow \infty$ ,  $M_n$  is asymptotically unbiased.

(d) As already noted in the previous part,  $M_n = [(n - 1)/n]U_n$ , and then  $\text{Var}_p[M_n] = \text{Var}_p \left[ \frac{(n - 1)}{n} U_n \right] = \left[ \frac{(n - 1)}{n} \right]^2 \text{Var}_p[U_n]$ . Therefore,

$$\text{Var}_p[U_n] = \left( \frac{n}{n - 1} \right)^2 \text{Var}_p[M_n] > \text{Var}_p[M_n],$$

showing that the variance of the unbiased estimator  $U_n$  is larger than the variance of the maximum likelihood estimator  $M_n$ .  $\square$

### 3.7. Bivariate Normal Distribution

**Exercise 3.7.1.** Let  $\mathbf{X} = ((X_{11}, X_{21}), (X_{12}, X_{22}), \dots, (X_{1n}, X_{2n}))$  be a random sample of size  $n$  from the bivariate normal distribution with means  $\mu_1, \mu_2$ , variances  $\sigma_1^2$  and  $\sigma_2^2$  and correlation coefficient  $\rho$ . Suppose that  $\rho \in (-1, 1)$  is unknown and find the maximum likelihood estimator of  $\rho$  if

- (a)  $\mu_1, \mu_2$ , and  $\sigma_1^2$  and  $\sigma_2^2$  are also *unknown*.  
 (b)  $\mu_1, \mu_2$ , and  $\sigma_1^2$  and  $\sigma_2^2$  are *known*.

**Solution.** (a) In this case the parameter is  $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) \in \Theta = \mathbb{R} \times \mathbb{R} \times (0, \infty) \times (0, \infty) \times (-1, 1)$ , and the likelihood of the sample  $\mathbf{X}$  is given by

$$L(\rho; \mathbf{X}) = e^{-Q/[2(1-\rho^2)]} \prod_{i=1}^n \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}} \quad (3.7.1)$$

where

$$Q = \sum_{i=1}^n \left[ \left( \frac{X_{1i} - \mu_1}{\sigma_1} \right)^2 + \left( \frac{X_{2i} - \mu_2}{\sigma_2} \right)^2 - 2\rho \left( \frac{X_{1i} - \mu_1}{\sigma_1} \right) \left( \frac{X_{2i} - \mu_2}{\sigma_2} \right) \right]. \quad (3.7.2)$$

The logarithm of the likelihood function is

$$\mathcal{L}(\theta; \mathbf{X}) = -\frac{Q}{2(1-\rho^2)} - n \log(\sigma_1) - n \log(\sigma_2) - \frac{n}{2} \log(1-\rho^2) - n \log(2\pi), \quad (3.7.3)$$

and without loss of generality it will be supposed that the vectors  $(X_{1i}, i = 1, 2, \dots, n)$  and  $(X_{2i}, i = 1, 2, \dots, n)$  are not constant. The maximizer of  $\mathcal{L}(\cdot; \mathbf{X})$  will be determined in two phases:

- (i) First, given  $\sigma_1, \sigma_2$  and  $\rho$ , the maximizer of  $\mathcal{L}(\cdot; \mathbf{X})$  with respect to  $\mu_1$  and  $\mu_2$  will be determined.. Notice that (3.7.2) and (3.7.3) together yield that  $\mathcal{L}(\cdot; \mathbf{X})$  is a concave quadratic form in  $(\mu_1, \mu_2)$ , and then it is maximized at the pair satisfying the following critical equations:

$$\partial_{\mu_1} \mathcal{L}(\theta; \mathbf{X}) = 0, \quad \text{and} \quad \partial_{\mu_2} \mathcal{L}(\theta; \mathbf{X}) = 0,$$

which are equivalent to

$$\begin{aligned} \sum_{i=1}^n \left( \frac{X_{1i} - \mu_1}{\sigma_1} \right) - \rho \sum_{i=1}^n \left( \frac{X_{2i} - \mu_2}{\sigma_2} \right) &= 0, \\ -\rho \sum_{i=1}^n \left( \frac{X_{1i} - \mu_1}{\sigma_1} \right) + \sum_{i=1}^n \left( \frac{X_{2i} - \mu_2}{\sigma_2} \right) &= 0. \end{aligned}$$

Since  $\rho \in (-1, 1)$ , these equations have the unique solution

$$\hat{\mu}_1 = \frac{1}{n} \sum_{i=1}^n X_{1i} =: \bar{X}_{1n}, \quad \text{and} \quad \hat{\mu}_2 = \frac{1}{n} \sum_{i=1}^n X_{2i} =: \bar{X}_{2n}.$$

Thus,

$$\mathcal{L}(\theta; \mathbf{X}) \leq \mathcal{L}(\bar{X}_{1n}, \bar{X}_{2n}, \sigma_1, \sigma_2, \rho), \quad \theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) \in \Theta. \quad (3.7.4)$$

(ii) Next, the function  $\mathcal{L}(\bar{X}_{1n}, \bar{X}_{2n}, \sigma_1, \sigma_2, \rho)$  will be maximized with respect to  $\sigma_1, \sigma_2$  and  $\rho$ . To achieve this goal, notice that

$$\begin{aligned} & \mathcal{L}(\bar{X}_{1n}, \bar{X}_{2n}, \sigma_1, \sigma_2, \rho) \\ &= -\frac{\tilde{Q}}{2(1-\rho^2)} - n \log(\sigma_1) - n \log(\sigma_2) - \frac{n}{2} \log(1-\rho^2) - n \log(2\pi), \end{aligned}$$

where

$$\begin{aligned} \tilde{Q} &= \sum_{i=1}^n \left[ \left( \frac{X_{1i} - \bar{X}_{1n}}{\sigma_1} \right)^2 + \left( \frac{X_{2i} - \bar{X}_{2n}}{\sigma_2} \right)^2 \right] \\ &\quad - 2\rho \left[ \left( \frac{X_{1i} - \bar{X}_{1n}}{\sigma_1} \right) \left( \frac{X_{2i} - \bar{X}_{2n}}{\sigma_2} \right) \right] \end{aligned} \quad (3.7.5)$$

From this expressions, it follows that, as  $\sigma_1$  or  $\sigma_2$  goes to 0 or  $\infty$  or  $\rho \rightarrow \pm 1$ , the function  $\mathcal{L}(\bar{X}_{1n}, \bar{X}_{2n}, \sigma_1, \sigma_2, \rho)$  converges to  $-\infty$ , and then the mapping  $(\sigma_1, \sigma_2, \rho) \mapsto \mathcal{L}(\bar{X}_{1n}, \bar{X}_{2n}, \sigma_1, \sigma_2, \rho; \mathbf{X})$  attains its maximum at some point  $(\sigma_1, \sigma_2, \rho) \in (0, \infty) \times (0, \infty) \times (-1, 1)$ , which satisfies

$$\begin{aligned} & \partial_{\sigma_1} \mathcal{L}(\bar{X}_{1n}, \bar{X}_{2n}, \sigma_1, \sigma_2, \rho) \\ &= \frac{-1}{2(1-\rho^2)} \left[ \frac{-2}{\sigma_1} \sum_{i=1}^n \left( \frac{X_{1i} - \bar{X}_{1n}}{\sigma_1} \right)^2 \right. \\ &\quad \left. + \frac{2\rho}{\sigma_1} \sum_{i=1}^n \left( \frac{X_{1i} - \bar{X}_{1n}}{\sigma_1} \right) \left( \frac{X_{2i} - \bar{X}_{2n}}{\sigma_1} \right) \right] - \frac{n}{\sigma_1} = 0 \\ & \partial_{\sigma_2} \mathcal{L}(\bar{X}_{1n}, \bar{X}_{2n}, \sigma_1, \sigma_2, \rho) \\ &= \frac{-1}{2(1-\rho^2)} \left[ \frac{-2}{\sigma_2} \sum_{i=1}^n \left( \frac{X_{2i} - \bar{X}_{2n}}{\sigma_2} \right)^2 \right. \\ &\quad \left. + \frac{2\rho}{\sigma_2} \sum_{i=1}^n \left( \frac{X_{1i} - \bar{X}_{1n}}{\sigma_1} \right) \left( \frac{X_{2i} - \bar{X}_{2n}}{\sigma_2} \right) \right] - \frac{n}{\sigma_2} = 0 \\ & \partial_{\rho} \mathcal{L}(\bar{X}_{1n}, \bar{X}_{2n}, \sigma_1, \sigma_2, \rho) \\ &= \frac{-\rho}{(1-\rho^2)^2} \tilde{Q} - \frac{1}{2(1-\rho^2)} \partial_{\rho} \tilde{Q} + \frac{n\rho}{1-\rho^2} = 0 \end{aligned} \quad (3.7.6)$$

The first equation immediately yields that

$$\begin{aligned} & \frac{n(1 - \rho^2)}{\sigma_1} \\ &= \left[ \frac{1}{\sigma_1} \sum_{i=1}^n \left( \frac{X_{1i} - \bar{X}_{1n}}{\sigma_1} \right)^2 - \frac{\rho}{\sigma_1} \sum_{i=1}^n \left( \frac{X_{1i} - \bar{X}_{1n}}{\sigma_1} \right) \left( \frac{X_{2i} - \bar{X}_{2n}}{\sigma_2} \right) \right] \end{aligned}$$

and then, multiplying both sides by  $\sigma_1/n$ ,

$$\frac{S_1^2}{\sigma_1^2} - \rho \frac{S_{12}}{\sigma_1 \sigma_2} = 1 - \rho^2 \quad (3.7.7)$$

where

$$\begin{aligned} S_1^2 &= \sum_{i=1}^n (X_{1i} - \bar{X}_{1n})^2 / n \\ S_2^2 &= \sum_{i=1}^n (X_{2i} - \bar{X}_{2n})^2 / n \\ S_{12} &= \sum_{i=1}^n (X_{1i} - \bar{X}_{1n}) (X_{2i} - \bar{X}_{2n}) / n \end{aligned} \quad (3.7.8)$$

Similarly, from the second equation in (3.7.7) it follows that

$$\frac{S_2^2}{\sigma_2^2} - \rho \frac{S_{12}}{\sigma_1 \sigma_2} = 1 - \rho^2 \quad (3.7.9)$$

Combining the specification of  $\tilde{Q}$  in (3.7.5) with (3.7.8) it follows that

$$\tilde{Q} = n \left[ \frac{S_1^2}{\sigma_1^2} - \rho \frac{S_{12}}{\sigma_1 \sigma_1} + \frac{S_2^2}{\sigma_1^2} - \rho \frac{S_{12}}{\sigma_1 \sigma_1} \right],$$

and then, at the solution of the system (3.7.6), equalities (3.7.7) and (3.7.9) yield that

$$\tilde{Q} = 2n(1 - \rho^2);$$

combining this relation with the third equation in (3.7.6), it follows that

$$\frac{-\rho}{(1 - \rho^2)^2} [2n(1 - \rho^2)] - \frac{1}{2(1 - \rho^2)} \partial_\rho \tilde{Q} + \frac{n\rho}{1 - \rho^2} = 0,$$

that is,

$$\frac{-2n\rho}{(1 - \rho^2)} - \frac{1}{2(1 - \rho^2)} \partial_\rho \tilde{Q} + \frac{n\rho}{1 - \rho^2} = 0,$$



equality that immediately yields that

$$\rho = -\frac{1}{2n}\partial_\rho\tilde{Q}.$$

Since  $\partial_\rho\tilde{Q} = -2nS_{12}/(\sigma_1\sigma_2)$  (see (3.7.5) and (3.7.8)), it follows that

$$\rho = \frac{S_{12}}{\sigma_1\sigma_2}. \quad (3.7.10)$$

Together with (3.7.7) this implies that  $S_1^2/\sigma_1^2 - \rho^2 = 1 - \rho^2$ , that is  $S_1^2/\sigma_1^2 = 1$ , so that

$$\sigma_1^2 = S_1^2.$$

Similarly, (3.7.9) and (3.7.10) together yield that

$$\sigma_2^2 = S_2^2,$$

and then (3.7.10) becomes

$$\rho = \frac{S_{12}}{S_1S_2}$$

In short, the mapping  $(\sigma_1, \sigma_2, \rho) \mapsto \mathcal{L}(\bar{X}_{1n}, \bar{X}_{2n}, \sigma_1, \sigma_2, \rho)$  attains its maximum at the point specified in the three previous displays, that is,

$$\mathcal{L}(\bar{X}_{1n}, \bar{X}_{2n}, \sigma_1, \sigma_2, \rho) \leq \mathcal{L}(\bar{X}_{1n}, \bar{X}_{2n}, S_1, S_2, S_{12}/[S_1S_2]), \quad \sigma_1, \sigma_2 > 0,$$

where  $\rho \in (-1, 1)$ , and combining this inequality with (3.7.4), it follows that

$$\mathcal{L}(\theta; \mathbf{X}) \leq \mathcal{L}(\bar{X}_{1n}, \bar{X}_{2n}, S_1, S_2, S_{12}/[S_1S_2]; \mathbf{X}), \quad \theta \in \Theta, .$$

showing that the maximum likelihood estimator  $\hat{\theta} = (\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2, \hat{\rho})$  is given by

$$(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2, \hat{\rho}) = (\bar{X}_{1n}, \bar{X}_{2n}, S_1, S_2, S_{12}/[S_1S_2]);$$

in particular, the maximum likelihood estimator of the population correlation coefficient  $\rho$  is the sample correlation coefficient  $S_{12}/[S_1S_2]$ .

(b) When  $\mu_1, \mu_2$  and  $\sigma_1$  and  $\sigma_2$  are known, the likelihood function is given by (3.7.1), where  $Q$  is specified by (3.7.2), that is,

$$L(\rho; \mathbf{X}) = e^{-Q/[2(1-\rho^2)]} \prod_{i=1}^n \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}} \quad (3.7.11)$$

and the corresponding logarithm is

$$\mathcal{L}(\rho; \mathbf{X}) = -\frac{Q}{2(1-\rho^2)} - n \log(\sigma_1) - n \log(\sigma_2) - \frac{n}{2} \log(1-\rho^2) - n \log(2\pi), \quad (3.7.12)$$

where, writing

$$\begin{aligned} \tilde{S}_1^2 &= \frac{1}{n} \sum_{i=1}^n \left( \frac{X_{1i} - \mu_1}{\sigma_1} \right)^2 \\ \tilde{S}_2^2 &= \frac{1}{n} \sum_{i=1}^n \left( \frac{X_{2i} - \mu_2}{\sigma_2} \right)^2 \\ \tilde{S}_{12} &= \frac{1}{n} \sum_{i=1}^n \left( \frac{X_{1i} - \mu_1}{\sigma_1} \right) \left( \frac{X_{2i} - \mu_2}{\sigma_2} \right) \end{aligned}$$

$Q$  is given by

$$Q = n[\tilde{S}_1^2 + \tilde{S}_2^2 - 2\rho\tilde{S}_{12}]$$

The value of  $\rho$  maximizing  $\mathcal{L}(\rho; \mathbf{X})$  in the interval  $(-1, 1)$  satisfies the likelihood equation

$$\partial_\rho \mathcal{L}(\rho; \mathbf{X}) = -\frac{\rho Q}{(1-\rho^2)^2} - \frac{\partial_\rho Q}{2(1-\rho^2)} + \frac{n\rho}{1-\rho^2} = 0,$$

which is equivalent to

$$-\frac{n\rho[\tilde{S}_1^2 + \tilde{S}_2^2 - 2\rho\tilde{S}_{12}]}{(1-\rho^2)^2} - \frac{-2n\tilde{S}_{12}}{2(1-\rho^2)} + \frac{n\rho}{1-\rho^2} = 0,$$

that is,

$$-\frac{\rho[\tilde{S}_1^2 + \tilde{S}_2^2 - 2\rho\tilde{S}_{12}]}{(1-\rho^2)} + \tilde{S}_{12} + \rho = 0,$$

equality that can be written as

$$(1-\rho^2)[\tilde{S}_{12} + \rho] - \rho[\tilde{S}_1^2 + \tilde{S}_2^2 - 2\rho\tilde{S}_{12}] = 0;$$

this cubic equation should be solved numerically.  $\square$

**Remark 3.7.1.** (i) At first, sight, part (b) seemed to be easier than part (a), since in part (b) only  $\rho$  is unknown. However, the maximum likelihood estimators was explicitly found when all the quantities determining the distribution of the observation data were unknown.

(ii) A very elegant method to determine the maximum likelihood estimators when the observation vectors have a multivariate normal distribution with unknown mean and covariance matrix, can be found, for instance in Chapter 1 of Anderson (2002). The argument relies on the spectral theory of positive matrices and on a factorization result in terms of triangular matrices.  $\square$

### 3.8. Logistic Model

In this section the maximum likelihood method is applied to logistic, binomial and normal models.

**Exercise 3.8.1.** Let  $X_1, X_2, \dots, X_n$  be a random variable of size  $n$  from the logistic density

$$f(x; \alpha) = \beta \frac{e^{-(\alpha + \beta x)}}{(1 + e^{-(\alpha + \beta x)})^2}$$

where  $\alpha$  is an unknown real number and  $\beta$  is known. In this context, find the maximum likelihood estimator of  $\alpha$ .

**Solution.** This is a case where an explicit formula for the maximum likelihood estimator  $\hat{\alpha}_n$  does not exist, and  $\hat{\alpha}_n(\mathbf{X}) = \hat{\alpha}(X_1, X_2, \dots, X_n)$  must be found numerically. In the following argument it will be shown that  $\hat{\alpha}_n$  exists, and is uniquely determined and is the unique critical point of the likelihood function. The likelihood function associated to the sample  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  is given by

$$\begin{aligned} L(\alpha; \mathbf{X}) &= \prod_{i=1}^n \beta \frac{e^{-(\alpha + \beta X_i)}}{(1 + e^{-(\alpha + \beta X_i)})^2} \\ &= \frac{\beta^n e^{-\sum_{i=1}^n (\alpha + \beta X_i)}}{\prod_{i=1}^n (1 + e^{-(\alpha + \beta X_i)})^2} = \beta^n \frac{e^{-n(\alpha + \beta \bar{X}_n)}}{\prod_{i=1}^n (1 + e^{-(\alpha + \beta X_i)})^2}, \end{aligned}$$

and its logarithm is

$$\mathcal{L}(\alpha; \mathbf{X}) = n \log(\beta) - n(\alpha + \beta \bar{X}_n) - 2 \sum_{i=1}^n \log(1 + e^{-(\alpha + \beta X_i)}).$$

Thus,

$$\begin{aligned}
\partial_\alpha \mathcal{L}(\alpha; \mathbf{X}) &= -n + 2 \sum_{i=1}^n \frac{e^{-(\alpha+\beta X_i)}}{1 + e^{-(\alpha+\beta X_i)}} \\
&= -n + 2 \sum_{i=1}^n \frac{e^{-(\alpha+\beta X_i)} + 1 - 1}{1 + e^{-(\alpha+\beta X_i)}} \\
&= -n + 2 \sum_{i=1}^n \left[ 1 - \frac{1}{1 + e^{-(\alpha+\beta X_i)}} \right] \\
&= n - 2 \sum_{i=1}^n \frac{1}{1 + e^{-(\alpha+\beta X_i)}}.
\end{aligned} \tag{3.8.1}$$

Notice that  $\lim_{\alpha \rightarrow \infty} 1 + e^{-(\alpha+\beta X_i)} = 1$ , so that

$$\lim_{\alpha \rightarrow \infty} \partial_\alpha \mathcal{L}(\alpha; \mathbf{X}) = -n < 0,$$

and  $\lim_{\alpha \rightarrow -\infty} 1 + e^{-(\alpha+\beta X_i)} = \infty$ , and then

$$\lim_{\alpha \rightarrow -\infty} \partial_\alpha \mathcal{L}(\alpha; \mathbf{X}) = n > 0.$$

These two last displays together imply that there exists (at least) a point  $\alpha^*(\mathbf{X}) \equiv \alpha^*$  such that

$$\partial_\alpha \mathcal{L}(\alpha; \mathbf{X})|_{\alpha=\alpha^*} = 0. \tag{3.8.2}$$

Notice now that

$$\begin{aligned}
\partial_\alpha^2 \mathcal{L}(\alpha; \mathbf{X}) &= \partial_\alpha \left[ n - 2 \sum_{i=1}^n \frac{1}{1 + e^{-(\alpha+\beta X_i)}} \right] \\
&= -2 \sum_{i=1}^n \frac{e^{-(\alpha+\beta X_i)}}{(1 + e^{-(\alpha+\beta X_i)})^2} < 0,
\end{aligned}$$

so that  $\mathcal{L}(\alpha; \mathbf{X})$  is a concave function of  $\alpha$ , and then the point  $\alpha^*$  satisfying (3.8.2) is unique, and is the unique maximizer of  $\mathcal{L}(\alpha; \mathbf{X})$ , that is,  $\hat{\alpha}_n(\mathbf{X}) = \alpha^*$ . Notice that (3.8.1) and (3.8.2) together yield that  $\hat{\alpha}_n$  is the unique solution of the likelihood equation

$$\sum_{i=1}^n \frac{1}{1 + e^{-(\alpha+\beta X_i)}} = \frac{n}{2}$$

which, as already mentioned, must be solve numerically.  $\square$

**Exercise 3.8.2.** Let  $X_1, X_2, \dots, X_m$  be a random sample of size  $m$  from the  $\mathcal{N}(\mu, \sigma^2)$  distribution and, independently, let  $Y_1, Y_2, \dots, Y_n$  be a random sample of size  $n$  from  $\mathcal{N}(\mu, \lambda\sigma^2)$  where  $\lambda > 0$  is unknown.

- (a) If  $\mu$  and  $\sigma$  are known, find the maximum likelihood estimator of  $\lambda$ .
- (b) If  $\mu$  and  $\sigma$  and  $\lambda$  are unknown, find the maximum likelihood estimator of  $\theta = (\mu, \sigma, \lambda)$ .

**Solution.** (a) Suppose that  $\mu$  and  $\sigma^2$  are known. In this case the distribution of the random vector  $\mathbf{X} = (X_1, X_2, \dots, X_m)$  does not involve  $\lambda$ , and the estimation of this parameter relies only on  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ . The statistical model for this last vector is

$$Y_1, \dots, Y_n \text{ are i.i.d. } \mathcal{N}(\mu, \lambda\sigma^2) \text{ random variables, } \lambda \in (0, \infty). \quad (3.8.3)$$

Since  $\lambda$  is an arbitrary real number, setting

$$\sigma_1^2 = \lambda\sigma^2 \quad (3.8.4)$$

the statistical model (3.8.3) is equivalent to

$$Y_1, \dots, Y_n \text{ are i.i.d. } \mathcal{N}(\mu, \sigma_1^2) \text{ random variables, } \sigma_1 \in (0, \infty).$$

For this model, the maximum likelihood estimator of  $\sigma_1^2$  is given by

$$\hat{\sigma}_1^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2;$$

since  $\lambda = \sigma_1^2/\sigma^2$ , the maximum likelihood estimator of  $\lambda$  is

$$\hat{\lambda} = \frac{\hat{\sigma}_1^2}{\sigma^2} = \frac{1}{n} \sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2.$$

(b) The statistical model for  $(\mathbf{X}, \mathbf{Y})$  is determined by

- (i)  $Y_1, \dots, Y_n$  are i.i.d.  $\mathcal{N}(\mu, \lambda\sigma^2)$  random variables,
- (ii)  $X_1, \dots, X_m$  are i.i.d.  $\mathcal{N}(\mu, \sigma^2)$  random variables,
- (iii) The vectors  $(X_1, \dots, X_m)$  and  $\mathbf{Y} = (Y_1, \dots, Y_n)$  are independent,
- (iv)  $\mu \in \mathbb{R}$ ,  $\sigma \in (0, \infty)$ ,  $\lambda \in (0, \infty)$ .

Defining  $\sigma_1 > 0$  by

$$\sigma_1^2 = \lambda\sigma^2, \quad (3.8.5)$$

the mapping  $(\mu, \sigma, \lambda) \mapsto (\mu, \sigma, \sigma_1)$  is a bijection of the parameter space  $\mathbb{R} \times (0, \infty) \times (0, \infty)$ . Hence, the above statistical model is equivalent to the following:

- (i)  $Y_1, \dots, Y_n$  are i.i.d.  $\mathcal{N}(\mu, \sigma_1^2)$  random variables,
- (ii)  $X_1, \dots, X_m$  are i.i.d.  $\mathcal{N}(\mu, \sigma^2)$  random variables,
- (iii) The vectors  $(X_1, \dots, X_m)$  and  $\mathbf{Y} = (Y_1, \dots, Y_n)$  are independent,
- (iv)  $\mu \in \mathbb{R}$ ,  $\sigma \in (0, \infty)$ ,  $\sigma_1 \in (0, \infty)$ .

This model was studied in Exercise 3.4.2, where it was shown that  $\hat{\mu}$  is determined as the root of a cubic equation, and then  $\hat{\sigma}$  and  $\hat{\sigma}_1$  are determined by

$$\hat{\sigma}^2 = \frac{1}{m} \sum_{i=1}^m (X_i - \hat{\mu})^2 \quad \text{and} \quad \hat{\sigma}_1^2 = \frac{1}{n} \sum_{j=1}^n (Y_j - \hat{\mu})^2;$$

then, (3.8.5) and the invariance property together yield that

$$\hat{\lambda} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}^2}$$

is the maximum likelihood estimator of  $\lambda$ . □

**Exercise 3.8.3.** Suppose that

$$X_1 \sim \text{Binomial}(n_1, p)$$

$$X_2 \sim \text{Binomial}(n_2, p)$$

$$\vdots$$

$$X_k \sim \text{Binomial}(n_k, p)$$

are independent random variables. Find the maximum likelihood estimator of  $p$ .

**Solution.** Given  $\mathbf{X} = (X_1, X_2, \dots, X_k)$  such that  $X_i$  is an integer between 0 and  $n_i$ , the corresponding likelihood function is

$$L(p; \mathbf{X}) = \prod_{i=1}^k \binom{n_i}{X_i} p^{X_i} (1-p)^{n_i - X_i}$$

whose logarithm is given by

$$\begin{aligned}\mathcal{L}(p; \mathbf{X}) &= \sum_{i=1}^k \log \left[ \binom{n_i}{X_i} \right] + \sum_{i=1}^k [X_i \log(p) + (n_i - X_i) \log(1 - p)] \\ &= \sum_{i=1}^k \log \left[ \binom{n_i}{X_i} \right] + \log(p) \sum_{i=1}^k X_i + \log(1 - p) \left[ N - \sum_{i=1}^k X_i \right] \\ &= \sum_{i=1}^k \log \left[ \binom{n_i}{X_i} \right] + \log(p)T + \log(1 - p) [N - T]\end{aligned}$$

where  $N = \sum_{i=1}^k n_i$  and  $T = \sum_{i=1}^k X_i$ . The kernel in this expression (the part involving the parameter  $p$ ), is the same as the kernel of a sample  $Y_1, Y_2, \dots, Y_N$  of size  $N$  from the *Bernoulli*( $p$ ) distribution when the grand total  $Y_1 + Y_2 + \dots + Y_N$  is equal to  $T$ . The computations for this case are well-known and yield that, in the present problem, the maximum likelihood estimator of  $p$  is

$$\hat{p} = \frac{T}{N} = \frac{X_1 + X_2 + \dots + X_k}{n_1 + n_2 + \dots + n_k}.$$

□

**Exercise 3.8.4.** Let  $(X_1, X_2, \dots, X_k)$  be a random vector with multinomial distribution with parameter  $p = (p_1, p_2, \dots, p_k)$  and  $n$  trials, where  $n$  is known and the probabilities  $p_i$  are unknown numbers in  $[0, 1]$  satisfying  $\sum_{i=1}^k p_i = 1$ . Find the maximum likelihood estimator  $\hat{p} = (\hat{p}_1, \hat{p}_2, \dots, \hat{p}_k)$ .

**Solution.** Given  $\mathbf{X} = (X_1, X_2, \dots, X_k)$  with positive components adding up to  $n$ , the corresponding likelihood function is

$$L(p; \mathbf{X}) = \binom{n}{X_1, X_2, \dots, X_k} p_1^{X_1} p_2^{X_2} \dots p_k^{X_k} \equiv C p_1^{X_1} p_2^{X_2} \dots p_k^{X_k},$$

where the convention  $0^0 = 1$  is enforced, and the multinomial coefficient has been denoted by  $C$ , since it does not involve the unknown vector parameter  $p$ . Let  $\mathcal{P}$  be the set of all admissible values of the vector  $p$ , that is,

$$\mathcal{P} = \left\{ p = (p_1, p_2, \dots, p_k) \in \mathbb{R}^k \mid \sum_{i=1}^k p_i = 1, \quad p_i \geq 0, \quad i = 1, 2, \dots, k \right\}.$$

This set is closed and bounded, so that the continuous function  $L(\cdot; \mathbf{X})$  attains its maximum at some point  $\hat{p} = (\hat{p}_1, \hat{p}_2, \dots, \hat{p}_k)$ :

$$L(\hat{p}; \mathbf{X}) \geq L(p; \mathbf{X}), \quad p \in \mathcal{P}. \quad (3.8.6)$$

To determine this point, let  $D$  be the set of all indices  $i$  such that  $X_i$  is no-null, that is,

$$D = \{i \in \{1, 2, \dots, n\} \mid X_i \neq 0\}, \quad (3.8.7)$$

so that

$$L(p; \mathbf{X}) = C \prod_{j \in D} p_j^{X_j}, \quad (3.8.8)$$

Now, observe the following properties (a)–(e):

(a)  $L(\hat{p}; \mathbf{X}) > 0$ . Indeed, the  $k$ -dimensional vector  $u = (1/k, 1/k, \dots, 1/k) \in \mathcal{P}$  satisfies  $L(u; \mathbf{X}) = C/k^n > 0$ , and then (3.8.6) implies that  $L(\hat{p}; \mathbf{X}) > 0$ .

(b) If  $X_i = 0$  then  $\hat{p}_i = 0$ . Proceeding by contradiction suppose that  $X_i = 0$  but  $\hat{p}_i > 0$ . In these circumstances, notice that  $i \notin D$  and that  $\hat{p}_i < 1$ , since otherwise  $\hat{p}_i = 1$ , and then  $\hat{p}_j = 0$  for all  $j \neq i$ ; in particular,  $\hat{p}_j = 0$  for every  $j \in D$ , and then (3.8.8) yields that  $L(\hat{p}; \mathbf{X}) = 0$ , which contradicts the fact (a) stated above. To continue, define the new vector  $\tilde{p} \in \mathcal{P}$  as follows:

$$\tilde{p}_i = 0, \quad \tilde{p}_j = \hat{p}_j / (1 - \hat{p}_i), \quad j \neq i,$$

so that  $\tilde{p}_j = \hat{p}_j / (1 - \hat{p}_i)$  for every  $j \in D$ , and then

$$\begin{aligned} L(\tilde{p}; \mathbf{X}) &= C \prod_{j \in D} \left( \frac{\hat{p}_j}{1 - \hat{p}_i} \right)^{X_j} \\ &= \frac{1}{\prod_{j \in D} (1 - \hat{p}_i)^{X_j}} C \prod_{j \in D} \hat{p}_j^{X_j} \\ &= \frac{1}{\prod_{j \in D} (1 - \hat{p}_i)^{X_j}} L(\hat{p}; \mathbf{X}) \end{aligned}$$

where (3.8.8) with  $\hat{p}$  instead of  $p$  was used in the last step. Since  $\hat{p}_i \in (0, 1)$  and  $X_j > 0$  for  $j \in D$ , the above display yields that  $L(\tilde{p}; \mathbf{X}) > L(\hat{p}; \mathbf{X})$ , which is a contradiction, since  $\hat{p}$  maximizes  $L(\cdot; \mathbf{X})$  on the set  $\mathcal{P}$  and  $\tilde{p} \in \mathcal{P}$ . It follows that  $X_i = 0$  implies that  $\hat{p}_i = 0$ , establishing the desired conclusion.



(c)  $\hat{p}_i = 0$  implies  $X_i = 0$ . Indeed, if  $\hat{p}_i = 0$  but  $X_i \neq 0$ , it follows that  $i \in D$  and then the factor  $\hat{p}_i^{X_i} = 0$  appears in the right hand side of (3.8.8), so that  $L(\hat{p}; \mathbf{X}) = 0$ , in contradiction with fact (a).

The discussion in (a)-(c) can be summarized as follows: The largest value of the likelihood function is positive, and a coordinate  $\hat{p}_i$  of the maximizer  $\hat{p}$  is positive if, and only if, the observation  $X_i$  is positive.

(d) Suppose that  $D$  is a singleton, say  $D = \{j^*\}$ . In this case  $\hat{p}_{j^*} = 1$ .

When  $D = \{j^*\}$ , notice that (3.8.8) yields that  $L(p; \mathbf{X}) = Cp_{j^*}^{X_{j^*}}$ , which is an increasing function of  $p_{j^*}$ , and then attains its maximum when  $p_{j^*} = 1$ , so that  $\hat{p}_{j^*} = 1$ .

(e) Suppose that  $S$  contains two or more indices and let  $j^* \in D$  be fixed. For every  $i \in D$  the equality

$$\frac{X_i}{\hat{p}_i} = \frac{X_{j^*}}{p_{j^*}}$$

occurs.

To verify this assertion, for a real number  $h$  satisfying  $|h| < \min\{\hat{p}_i, \hat{p}_{j^*}\}$ , define the  $k$ -dimensional vector  $p(h)$  by

$$p(h)_j = \begin{cases} \hat{p}_j, & \text{if } j \neq i, j^* \\ \hat{p}_i - h, & \text{if } j = i \\ \hat{p}_{j^*} + h, & \text{if } j = j^* \end{cases}$$

It follows from this specification  $p(h) \in \mathcal{P}$  and  $p(0) = \hat{p}$ . Defining  $g(h) = L(p(h); \mathbf{X})$  for  $|h| < \min\{\hat{p}_i, \hat{p}_{j^*}\}$ , relation (3.8.6) yields that  $0 \neq L(\hat{p}; \mathbf{X}) = g(0) \geq g(h)$ , that is, the function  $g$  attains its maximum at  $h = 0$ , so that  $g'(0) = 0$ . Observing that  $g(h) = \tilde{C}(\hat{p}_i - h)^{X_i}(\hat{p}_{j^*} + h)^{X_{j^*}}$  where  $\tilde{C}$  is a no-null term which does not depend on  $h$ , it follows that  $g'(h) = [X_{j^*}/(p_{j^*} + h) - X_i/(p_i - h)]g(h)$ . Therefore,

$$0 = g'(0) = \left[ \frac{X_{j^*}}{p_{j^*}} - \frac{X_i}{p_i} \right] g(0),$$

and then, since  $g(0) \neq 0$ ,

$$\frac{X_{j^*}}{p_{j^*}} = \frac{X_i}{p_i}.$$

Using the previous facts, it will be shown that, for  $i = 1, 2, \dots, k$ ,

$$\hat{p}_i = \frac{X_i}{n}. \tag{3.8.9}$$

To establish this assertion, first notice that if  $i \notin D$ , then  $X_i = 0$ , by (3.8.7), and this implies that  $\hat{p}_i = 0$ , by part (b), so that the above equality always holds when  $i \notin D$ . To conclude it will be shown that (3.8.9) occurs when  $i \in D$ . To achieve this goal, consider the following two exhaustive cases.

(i)  $D$  is a singleton, say  $D = \{j^*\}$ . In this context  $\hat{p}_{j^*} = 1$ , by part (c), and  $X_{j^*} = n$ , since the  $X_i = 0$  for  $i \neq j^*$  (by (3.8.7)) and  $\sum_{r=1}^k X_r = n$ . Consequently, (3.8.9) also holds when  $i = j^*$ .

(ii)  $D$  contains two or more indices. In this case, part (e) yields that the quotient

$$\frac{X_i}{\hat{p}_i} = \lambda$$

is constant when  $i$  varies in  $D$ . Thus,  $X_i = \lambda \hat{p}_i$  and, using that  $X_i = 0 = \hat{p}_i$  when  $i \notin D$ , it follows that

$$n = \sum_{r=1}^n X_r = \sum_{r \in D} X_r = \sum_{r \in D} \lambda \hat{p}_r = \lambda \sum_{r=1}^k \hat{p}_r = \lambda,$$

and then  $\hat{p}_i = X_i/n$  for all  $i \in D$ , showing that (3.8.9) also occurs when  $i \in D$ . In short, for every  $i = 1, 2, \dots, k$ , the maximum likelihood estimator of  $p_i$  is  $\hat{p}_i = X_i/n$ .  $\square$

**Exercise 3.8.5.** Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from the  $\mathcal{N}(\mu, \sigma^2)$  distribution, where the vector  $(\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$  is unknown. Set

$$g(\mu, \sigma^2) = P_{(\mu, \sigma^2)}[X > c],$$

where  $c$  is a known constant. Determine the maximum likelihood estimator  $\hat{g}_n$  of this parametric function and show that the  $\{\hat{g}_n\}$  is a consistent sequence.

**Solution.** The basic properties of the normal distribution yield that

$$g(\mu, \sigma^2) = 1 - \Phi\left(\frac{c - \mu}{\sigma}\right),$$

where, as usual,  $\Phi(\cdot)$  is the cumulative distribution function of the standard normal distribution. Recalling that the maximum likelihood estimator of  $(\mu, \sigma^2)$  is

$$(\hat{\mu}_n, \hat{\sigma}_n^2) = (\bar{X}_n, S_n^2),$$

the invariance theorem yields that

$$\hat{g}_n = 1 - \Phi\left(\frac{c - \bar{X}_n}{S_n}\right);$$

since  $g(\cdot, \cdot)$  is a continuous function in the parameter space  $\Theta$  and the sequences  $\{\bar{X}_n\}$  and  $\{S_n^2\}$  estimate consistently to the parameters  $\mu$  and  $\sigma^2$ , respectively, the continuity theorem yields that  $\{\hat{g}_n\}$  is a consistent sequence for  $g(\mu, \sigma^2)$ .  $\square$

**Exercise 3.8.6.** Let  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  be a random sample from the density  $f(x; \theta) = (\theta/x^2)I_{[\theta, \infty)}(x)$ , where  $\theta \in \Theta = (0, \infty)$ .

(a) Find the maximum likelihood estimator  $\{\hat{\theta}_n\}$  of  $\theta$  and verify that  $\{\hat{\theta}_n\}$  is consistent.

(b) Find the maximum likelihood estimator of  $g(\theta) = P_\theta[X \leq c]$ , where  $c$  is a known constant, and show the consistency of the sequence  $\{\hat{g}_n\}$ .

(c) If  $n = 5$  find the estimate  $\hat{g}_5$  corresponding to

$$\mathbf{x} = (2.9, 1.48, 5.62, 4.0, 1.22).$$

**Solution.** (a) Given  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  with positive components, the corresponding likelihood function is

$$L(\theta, \mathbf{X}) = \prod_{i=1}^n \frac{\theta}{X_i^2} I_{[\theta, \infty)}(X_i) = \frac{\theta^n}{\prod_{i=1}^n X_i^2} \prod_{i=1}^n I_{[\theta, \infty)}(X_i), \quad \theta \in (0, \infty).$$

Observing that

$$\begin{aligned} \prod_{i=1}^n I_{[\theta, \infty)}(X_i) = 1 &\iff I_{[\theta, \infty)}(X_i) = 1 \text{ for all } i = 1, 2, \dots, n \\ &\iff \theta \leq X_i \text{ for all } i = 1, 2, \dots, n \\ &\iff \theta \leq X_{(1)} = \min\{X_1, X_2, \dots, X_n\}, \\ &\iff I_{(0, X_{(1)}]}(\theta) = 1, \end{aligned}$$

it follows that

$$L(\theta, \mathbf{X}) = \frac{1}{\prod_{i=1}^n X_i^2} \theta^n I_{(0, X_{(1)}]}(\theta).$$

This expression shows that  $L(\cdot; \mathbf{X})$  is strictly increasing in  $(0, X_{(1)})$  and is null outside this interval. Hence,  $\hat{\theta}_n = X_{(1)}$ . Observe now that, for  $\varepsilon > 0$ ,

$$\begin{aligned} P_\theta[\hat{\theta}_n > \theta + \varepsilon] &= P_\theta[X_i > \theta + \varepsilon, i = 1, \dots, n] \\ &= \prod_{i=1}^n P_\theta[X_i > \theta + \varepsilon] \\ &= \prod_{i=1}^n \int_{\theta+\varepsilon}^{\infty} \frac{\theta}{x^2} dx \\ &= \left( \frac{\theta}{\theta + \varepsilon} \right)^n \rightarrow 0 \text{ as } n \rightarrow \infty. \end{aligned}$$

Since  $P_\theta[\hat{\theta}_n < \theta] = 0$ , it follows that  $P_\theta[|\hat{\theta}_n - \theta| > \varepsilon] = P_\theta[\hat{\theta}_n > \theta + \varepsilon] \rightarrow 0$  as  $n \rightarrow \infty$ , that is,  $\{\hat{\theta}_n\}$  is a consistent sequence.

(b) By the invariance principle, the maximum likelihood estimator of  $g(\theta)$  is

$$\hat{g}_n = g(\hat{\theta}_n) = g(X_{(1)}).$$

On the other hand, the function  $g(\theta)$  is explicitly given by

$$g(\theta) = \int_0^c f(x; \theta) dx = \int_0^c \frac{\theta}{x^2} I[\theta, \infty)(x) dx = \begin{cases} 1 - \theta/c, & \text{if } \theta \leq c, \\ 0 & \text{if } c < \theta, \end{cases}$$

and it is clear the  $g(\cdot)$  is continuous in the parameter space. Using that  $\{\hat{\theta}_n\}$  is a consistent sequence, the continuity theorem yields the consistency of  $\{\hat{g}_n\}$ .

The estimate  $\hat{\theta}_5(\mathbf{x})$  corresponding to the given data is

$$\hat{\theta}_5(\mathbf{x}) = \min\{x_1, x_2, x_3, x_4, x_5\} = 1.48,$$

and then

$$\hat{g}_5(\mathbf{x}) = g(1.48) = \begin{cases} 1 - 1.48/c, & \text{if } \theta \leq 1.48, \\ 0 & \text{if } 1.48 < \theta. \end{cases}$$

□

**Exercise 3.8.7.** Let  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  be a random sample of size  $n$  from the displaced exponential density

$$f(x; \delta, \lambda) = (1/\lambda)e^{-(x-\delta)/\lambda} I_{[\delta, \infty)}(x),$$

where  $\theta = (\delta, \lambda) \in \Theta = \mathbb{R} \times (0, \infty)$  is unknown. In this context, the maximum likelihood estimator of  $\theta$  is

$$\hat{\theta}_n = (\hat{\delta}_n, \hat{\lambda}_n) = (X_{(1)}, \bar{X}_n - X_{(1)}),$$

where  $X_{(1)}$  is the minimum of the sample  $\mathbf{X}$ . If  $n = 5$  and the sample takes the value  $\mathbf{x}$  as in Exercise 3.8.6, find the estimate of the parametric function  $g(\theta) = P_\theta[X > c]$ , where  $c$  is a known constant and  $X$  has density  $f(\cdot; \delta, \lambda)$ .

**Solution.** Notice the following two facts: (i) If  $X$  has density  $f(x; \delta, \lambda)$ , then  $Y = (X - \delta)/\lambda$  has the *Exponential*(1) distribution, an assertion that follows from the change of variable formula, and (ii) If  $Y \sim \text{Exponential}(1)$ , then  $P[Y > y] = \min\{1, e^{-y}\}$ . It follows that  $g(\theta) = P[X > c] = P[(X - \delta)/\lambda > (c - \delta)/\lambda] = \min\{1, e^{-(c-\delta)/\lambda}\}$ , and then

$$\hat{g}_n = g(\hat{\theta}_n) = g(\hat{\delta}_n, \lambda_n) = \min\{1, e^{-(c-\hat{\delta}_n)/\hat{\lambda}_n}\}.$$

For the data vector in Exercise 3.8.6,  $n = 5$ , and the observed value of  $\hat{\theta}_5 = (\hat{\delta}_5, \hat{\lambda}_5)$  is  $(1.48, 5.055 - 1.480) = (1.48, 3.575)$ . Therefore, the estimate  $\hat{g}_5(\mathbf{x})$  is given by

$$\begin{aligned} \hat{g}_5(\mathbf{x}) &= g(1.48, 3.575) \\ &= \min\{1, e^{-(c-1.48)/3.575}\} = \begin{cases} 1, & \text{if } c < 1.48, \\ e^{-(c-1.48)/3.575}, & \text{if } c \geq 1.48, \end{cases} \end{aligned}$$

concluding the argument. □

**Exercise 3.8.8.** Let  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  be a random sample from a *Geometric*( $p$ ) distribution, where  $p \in [0, 1]$ , so that the common probability function of the variables  $X_i$  is given by

$$f(x; p) = (1 - p)^{x-1} p I_{\{1, 2, 3, \dots\}}(x).$$

(a) Find the maximum likelihood estimator of  $p$ .

(b) A state has 36 counties. Assume that each county has equal proportions of people who favor a certain gun control proposal. In each of 8 randomly selected counties, it is determined how many people is needed to sample to find the first person who favors the proposal. The results are

3, 8, 9, 6, 5, 3, 2

(e.g., in the first county sampled, the first two persons sampled were opposed, and the third one was in favor). Based on this data, compute the maximum likelihood estimator of  $p$ .

**Solution.** (a) Given a sample  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  whose components are positive integers, the corresponding likelihood function is

$$L(p; \mathbf{X}) \prod_{i=1}^n (1-p)^{X_i-1} p = (1-p)^{T_n-n} p^n, \quad p \in [0, 1],$$

where

$$T_n = \sum_{i=1}^n X_i.$$

The function  $L(\cdot; \mathbf{X})$  is continuous in  $[0, 1]$ , and then it has a maximizer  $\hat{p}_n$ . To determine such a point, notice that  $T_n \geq n$ , since the  $X_i$ s are positive integers, and consider the following two exhaustive cases.

(i)  $T_n = n$ . In this context,  $L(p; \mathbf{X}) = p^n$  is an increasing function in  $[0, 1]$ , so that the likelihood function is maximized at the unique point  $\hat{p}_n = 1$ .

(ii)  $T_n > n$ . In this case  $L(p; \mathbf{X})$  is null at the extreme points  $p = 0$  and  $p = 1$  of its domain, and is positive for  $p \in (0, 1)$ . It follows that  $L(p; \mathbf{X})$  attains its maximum inside the open interval  $(0, 1)$ , and the maximizer must satisfy the likelihood equation

$$\partial_p L(p; \mathbf{X}) = -\frac{T_n - n}{1 - p} L(p; \mathbf{X}) + \frac{n}{p} L(p; \mathbf{X}) = 0$$

where  $L(p; \mathbf{X}) \neq 0$ . Hence,

$$\frac{T_n - n}{1 - p} = \frac{n}{p},$$

which is equivalent to  $p(T_n - n) = n(1 - p)$ , that is,  $pT_n = n$ , and the unique solution is  $p = n/T_n$ . Consequently,

$$\hat{p}_n = \frac{n}{T_n} = \frac{1}{\bar{X}_n},$$

a relation that is also valid when  $T_n = n$ , since in this case  $\hat{p}_n = 1$  and  $\bar{X}_n = 1$ . In short, the maximum likelihood estimator of  $p$  is  $\hat{p}_n = 1/\bar{X}_n$ .

(b) For the data set  $\mathbf{x}$  in the problem,  $\bar{X}_8$  attains the value  $\bar{x}_8 = 40/8 = 5$ , and the corresponding estimate of  $p$  is  $\hat{p}_8 = 1/5 = 0.2$ .  $\square$

# Chapter 4

## Method of Moments

### 4.1. Description of the Method

This section introduces another method to produce estimators of parametric functions. Consider a random variable  $X$  whose distribution depends on an unknown parameter  $\theta$ ,

$$X \sim P_\theta, \quad \theta \in \Theta,$$

where the parameter space  $\Theta$  is a subset of  $\mathbb{R}^m$  for some  $m$ . Now, let  $\mu'_k(\theta)$  be the  $k$ th moment of the distribution  $P_\theta$ , that is,

$$\mu'_k(\theta) = E_\theta[X^k], \tag{4.1.1}$$

which is supposed to be finite. Now, let  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  be a random sample of size  $n$  of the population  $P_\theta$ , so that

$$\begin{aligned} X_1, X_2, \dots, X_n \text{ are independent and identically} \\ \text{distributed with common distribution } P_\theta. \end{aligned} \tag{4.1.2}$$

The  $k$ th sample moment of the data  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  is defined by

$$m'_{kn} = \frac{1}{n} \sum_{i=1}^n X_i^k. \tag{4.1.3}$$

This sample moment is naturally considered as an estimator of  $\mu'_k$ ; indeed, since the powers  $X_1^k, X_2^k, \dots, X_n^k$  are independent with the same distribution as  $X^k$ , the law of large numbers yields that

$$m'_{k n} = \frac{1}{n} \sum_{i=1}^n X_i^k \xrightarrow{P_\theta} E_\theta[X^k] = \mu'_k(\theta) \quad (4.1.4)$$

so that the sequence  $\{m'_{k n}\}_{n=1,2,3,\dots}$  estimates  $\mu'_k(\theta)$  consistently. Moreover,  $E_\theta[m'_{k n}] = \sum_{i=1}^n E_\theta[X_i^k]/n = n\mu'_k(\theta)/N = \mu'_k(\theta)$ , so that  $m'_{k n}$  is an unbiased estimator of  $\mu'_k(\theta)$ .

The *method of moments* can be now stated formally as follows: Given  $X_1, X_2, \dots, X_n$  as in (4.1.2), then

- (i) The  $k$ th population moment  $\mu'_k(\theta)$  is estimated by  $m'_{k n}$ ;
- (ii) If a parametric quantity  $g(\theta)$  can be expressed in terms of the population moments  $\mu'_1(\theta), \mu'_2(\theta), \dots, \mu'_r(\theta)$ , say

$$g(\theta) = G(\mu'_1(\theta), \mu'_2(\theta), \dots, \mu'_r(\theta)), \quad (4.1.5)$$

then the estimator of  $g(\theta)$  based on  $X_1, X_2, \dots, X_n$  is given by

$$\hat{g}_n = G(m'_{1 n}, m'_{2 n}, \dots, m'_{r n}); \quad (4.1.6)$$

in words, if the parametric quantity  $g(\theta)$  is a function of some population moments, then the estimator  $\hat{g}_n$  is *the same* function evaluated at the corresponding sample moments.

## 4.2. Consistency of the Estimators

As it was already noted, the estimator  $m'_{k n}$  of  $\mu'_k(\theta)$  is unbiased. However, the above estimator  $\hat{g}_n$  of the parametric function in (4.1.5) is not, in general, unbiased if the function  $G$  is not linear; this assertion will be exemplified several times below. On the other hand, it will be now proved that the sequence  $\{\hat{g}_n\}$  is generally consistent. The following is *the continuity theorem*, and it was stated without proof in Section 2.1

**Theorem 4.2.1.** Suppose that the function  $G(z_1, z_2, \dots, z_r)$  is continuous at each point  $(\mu'_1(\theta), \mu'_2(\theta), \dots, \mu'_r(\theta))$ ,  $\theta \in \Theta$ . In this case, within the framework determined by (4.1.2), the parametric function  $g(\theta)$  in (4.1.5) is estimated consistently by the sequence  $\{\hat{g}_n\}$  specified in (4.1.6).



**Proof.** It must be shown that, for each  $\theta \in \Theta$  and  $\varepsilon > 0$ ,

$$\lim_{n \rightarrow \infty} P_\theta[|\hat{g}_n - g(\theta)| > \varepsilon] = 0. \quad (4.2.1)$$

To establish the conclusion, let  $\theta \in \Theta$  be arbitrary but fixed. By the continuity of the function  $G$ , given  $\varepsilon > 0$ , there exists  $\delta > 0$  such that

$$\begin{aligned} |x_i - \mu'_i(\theta)| &\leq \delta, \quad i = 1, 2, \dots, r \\ \Rightarrow |G(x_1, x_2, \dots, x_r) - G(\mu'_1(\theta), \mu'_2(\theta), \dots, \mu'_r(\theta))| &\leq \varepsilon. \end{aligned}$$

This implication is equivalent to

$$\begin{aligned} |G(x_1, x_2, \dots, x_r) - G(\mu'_1(\theta), \mu'_2(\theta), \dots, \mu'_r(\theta))| &> \varepsilon \\ \Rightarrow |x_i - \mu'_i(\theta)| &> \delta, \quad \text{for some } i = 1, 2, \dots, r. \end{aligned}$$

Consequently,

$$\begin{aligned} |G(m'_{1n}, m'_{2n}, \dots, m'_{rn}) - G(\mu'_1(\theta), \mu'_2(\theta), \dots, \mu'_r(\theta))| &> \varepsilon \\ \Rightarrow |m'_{in} - \mu'_i(\theta)| &> \delta, \quad \text{for some } i = 1, 2, \dots, r. \end{aligned}$$

that is,

$$\begin{aligned} &[|G(m'_{1n}, m'_{2n}, \dots, m'_{rn}) - G(\mu'_1(\theta), \mu'_2(\theta), \dots, \mu'_r(\theta))| > \varepsilon] \\ &\subset \bigcup_{i=1}^r [|m'_{in} - \mu'_i(\theta)| > \delta], \end{aligned}$$

which can be written as

$$[|\hat{g}_n - g(\theta)| > \varepsilon] \subset \bigcup_{i=1}^r [|m'_{in} - \mu'_i(\theta)| > \delta];$$

see (4.1.5) and (4.1.6). From this point, the monotonicity and subadditivity properties of a probability distribution yield that

$$P_\theta[|\hat{g}_n - g(\theta)| > \varepsilon] \leq \sum_{i=1}^r P_\theta[|m'_{in} - \mu'_i(\theta)| > \delta].$$

Recalling the  $P_\theta[|m'_{in} - \mu'_i(\theta)| > \delta] \rightarrow 0$  as  $n \rightarrow \infty$ , by (4.1.4), taking the limit as  $n$  goes to  $\infty$  in the above display, it follows that

$$\lim_{n \rightarrow \infty} P_\theta[|\hat{g}_n - g(\theta)| > \varepsilon] \leq \sum_{i=1}^r \lim_{n \rightarrow \infty} P_\theta[|m'_{in} - \mu'_i(\theta)| > \delta] = 0,$$

establishing (4.2.1). □

Before proceeding to present some examples on the method of moments, it is convenient to summarize the precedent discussion: Given a sample  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  of a population  $P_\theta$ , where  $\theta \in \Theta$ ,

(i) The method of moments prescribes to estimate a population moment by the corresponding sample moment;

(ii) The estimator of a function of the moments  $\mu'_1(\theta), \mu'_2(\theta), \dots, \mu'_k(\theta)$  is constructed evaluating the same function at the sample moments

$$m'_{1n}, m'_{2n}, \dots, m'_{kn}.$$

(iii) When estimating a continuous function of population moments, the method of moments produces consistent estimators.

(iv) If a linear function of population moments is being estimated, the method of moments generates unbiased estimators; however, the estimators of nonlinear functions of population moments are generally *biased*.

### 4.3. Applications

One of the appealing features of the method of moments is that, as soon as the parametric function of interest can be expressed as a function of the population moments, the construction of the estimator corresponding to a given sample is straightforward. In some cases the method can be applied successfully, particularly in problems for which the maximum likelihood estimate needs to be determined numerically.

The above ideas are illustrated in the following examples.

**Exercise 4.3.1.** Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from the *Uniform*(0,  $\theta$ ) distribution, where  $\theta \in \Theta = (0, \infty)$ .

(a) Find the method of moments estimator of  $\theta$  and show that it is unbiased.

(b) Find the method of moments estimator of  $\theta^2$  and show that it is biased. Also, find an unbiased estimator of  $\theta^2$ .

(c) Show the consistency of the estimators in parts (a) and (b).

**Solution.** (a) First, the parametric quantity  $g(\theta) = \theta$  must be expressed in terms of the moments of the parent distribution. In the present case, if  $X \sim \text{Uniform}(0, \theta)$ , then  $\mu'_1(\theta) = E_\theta[X] = \theta/2$ , so that  $\theta = 2\mu'_1$ . Consequently, the moments estimator of  $\theta$  is  $\hat{\theta}_n = 2m'_{1n}(\mathbf{X}) = 2\bar{X}_n$ . Notice that  $\theta$  is a linear function of  $\mu'_1(\theta)$ , and then  $\hat{\theta}_n$  is unbiased.

(b) The moments estimator of  $g(\theta) = \theta^2$  based on the sample of size  $n$  is  $\hat{g}_n = g(\hat{\theta}_n) = \hat{\theta}_n^2 = (2\bar{X}_n)^2 = 4\bar{X}_n^2$ ; since  $\hat{\theta}_n$  is not constant, Jensen's inequality yields that  $E_\theta[\hat{g}_n] = E_\theta[(\hat{\theta}_n)^2] > E_\theta[\hat{\theta}_n]^2 = \theta^2$ , and then  $\hat{g}_n$  is a biased estimator. To determine an unbiased estimator of  $g(\theta) = \theta^2$ , notice that

$$\begin{aligned} E_\theta[\hat{\theta}_n^2] &= \text{Var}_\theta [\hat{\theta}_n] + E_\theta[\hat{\theta}_n]^2 \\ &= \text{Var}_\theta [2\bar{X}_n] + \theta^2 \\ &= 4\text{Var}_\theta [\bar{X}_n] + \theta^2 \\ &= 4\frac{\theta^2}{12n} + \theta^2 = \left(1 + \frac{1}{3n}\right)\theta^2. \end{aligned}$$

Consequently,  $U_n = 3n/(1 + 3n)\hat{\theta}_n^2 = (3n/(1 + 3n))\hat{g}_n = 12n/(1 + 3n)\bar{X}_n^2$  is an unbiased estimator of  $\theta^2$ .

(c) Notice that in parts (a) and (b),  $\theta$  and  $g(\theta)$  are continuous functions of the population moment  $\mu'_1(\theta)$ , and then the sequences  $\{\hat{\theta}_n\}$  and  $\{\hat{g}_n\}$  are consistent for  $\theta$  and  $g(\theta)$ , respectively. Also,  $U_n = (3n/(1 + 3n))\hat{g}_n \xrightarrow{P_\theta} 1 \cdot g(\theta) = g(\theta)$ , and then  $\{U_n\}$  is a consistent sequence for the parametric function  $g(\theta)$ .  $\square$

**Exercise 4.3.2.** Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from the  $\text{Beta}(\alpha, \beta)$  distribution, where  $\theta = (\alpha, \beta) \in (0, \infty) \times (0, \infty)$ . Determine the moment estimators of  $\alpha$  and  $\beta$ .

**Solution.** If  $X \sim \text{Beta}(\alpha, \beta)$ , the first two moments of  $X$  are

$$\mu'_1 = E_\theta[X] = \frac{\alpha}{\alpha + \beta}, \quad \mu'_2 = \frac{\alpha\beta}{(\alpha + \beta)^2(1 + \alpha + \beta)} + (\mu'_1)^2.$$

Now, the parameters  $\alpha$  and  $\beta$  will be expressed in terms of  $\mu'_1$  and  $\mu'_2$ . Notice that

$$\mu'_2 - (\mu'_1)^2 = \frac{\mu'_1(1 - \mu'_1)}{1 + \alpha + \beta}, \quad \text{and then} \quad \alpha + \beta = \frac{\mu'_1(1 - \mu'_1)}{\mu'_2 - (\mu'_1)^2} - 1.$$

Since  $\alpha = \mu'_1(\alpha + \beta)$ , it follows that

$$\alpha = \mu'_1 \left( \frac{\mu'_1(1 - \mu'_1)}{\mu'_2 - (\mu'_1)^2} - 1 \right)$$

On the other hand, notice that  $1 - \mu'_1 = 1 - E_\theta[X] = 1 - \alpha/(\alpha + \beta) = \beta/(\alpha + \beta)$ , so that

$$\beta = (1 - \mu'_1)(\alpha + \beta) = (1 - \mu'_1) \left( \frac{\mu'_1(1 - \mu'_1)}{\mu'_2 - (\mu'_1)^2} - 1 \right)$$

From these two last displays, it follows that the moments estimators of  $\alpha$  and  $\beta$  based on a sample of size  $n$  are given by

$$\begin{aligned} \hat{\alpha}_n &= m'_{1n} \left( \frac{m'_{1n}(1 - m'_{1n})}{m'_{2n} - (m'_{1n})^2} - 1 \right) = m'_{1n} \left( \frac{m_{1n} - m'_{2n}}{m'_{2n} - (m'_{1n})^2} \right) \\ \hat{\beta}_n &= (1 - m'_{1n}) \left( \frac{m'_{1n}(1 - m'_{1n})}{m'_{2n} - (m'_{1n})^2} - 1 \right) = (1 - m'_{1n}) \left( \frac{m_{1n} - m'_{2n}}{m'_{2n} - (m'_{1n})^2} \right), \end{aligned}$$

concluding the argument.  $\square$

**Exercise 4.3.3.** Let  $X_1, X_2, \dots, X_n$  be independent random variables, each with the density  $f(x; \theta) = 1/(2\theta)I_{[-\theta, \theta]}(x)$ . Find the moments estimator of  $\theta$  and show directly that is biased.

**Solution.** The uniform distribution on the interval  $[-\theta, \theta]$  has mean  $\mu'_1 = 0$ , so that  $\theta$  can not be expressed as a function of  $\mu'_1$ . Therefore, the second moment must be calculated. If  $X \sim Uniform(-\theta, \theta)$ ,

$$\mu'_2(\theta) = E_\theta[X^2] = \text{Var}_\theta[X_2] = \frac{(2\theta)^2}{12} = \frac{\theta^2}{3}.$$

Since  $\theta$  is a positive number, it follows that  $\theta = (3\mu'_2(\theta))^{1/2}$ , and then the moments estimator of  $\theta$  is given by

$$\hat{\theta}_n = (3m'_{2n})^{1/2}.$$

This estimator is biased. Indeed, the second sample moment is not constant with probability 1 and, using that the function  $H(x) = x^{1/2}$  is strictly concave, it follows from Jensen's inequality that

$$E_\theta[\hat{\theta}_n] = E_\theta[(3m'_{2n})^{1/2}] = E_\theta[H(3m'_{2n})] > H(E_\theta[(3m'_{2n})]) = H(\theta^2) = \theta,$$

so that  $\hat{\theta}_n$  is biased with positive bias function.  $\square$

#### 4.4. Further Examples

**Exercise 4.4.1.** Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from a  $\mathcal{N}(\theta, \theta^2)$  distribution for some  $\theta \in \Theta = (0, \infty)$ . Find an estimator of  $\theta^2$  using the method of moments.

**Solution.** Let  $X \sim \mathcal{N}(\theta, \theta^2)$  and notice that the first population moment is  $\mu'_1(\theta) = E_\theta[X] = \theta$ . Thus, a method of moments estimator of  $\theta$  is given by  $\hat{\theta}_n = m'_{1n} = \bar{X}_n$ . This estimator was obtained quite directly, and the simplicity of the present argument should be contrasted with the effort required to determine the maximum likelihood estimator of  $\theta$ ; see Exercise 3.2.2.  $\square$

**Exercise 4.4.2.** Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from the *Geometric*( $p$ ) distribution, so that the common probability function of the variables is

$$f(x; p) = (1 - p)^{x-1} p I_{\{1, 2, 3, \dots\}}(x).$$

Use the method of moments to find an estimator of  $p$ . Show that the method of moments used to estimate  $1/p$  produces the estimator  $\bar{X}_n$ .

**Solution.** If  $X \sim \text{Geometric}(p)$ , then

$$\begin{aligned} \mu'_1(\theta) = E_\theta[X] &= \sum_{x=1}^{\infty} x(1-p)^{x-1} p \\ &= p \sum_{x=1}^{\infty} x(1-p)^{x-1} \\ &= p \frac{d}{dp} \left[ \sum_{x=1}^{\infty} (1-p)^x \right] = p \frac{d}{dp} \left[ \frac{1}{1 - (1-p)} \right] = \frac{p}{p^2} = \frac{1}{p}. \end{aligned}$$

It follows that  $p = 1/\mu'_1$ , and then the method of moments produces the following estimator of  $p$ :

$$\hat{p}_n = \frac{1}{m'_{1n}} = \frac{1}{\bar{X}_n}$$

As for the estimation of  $g(p) = 1/p$ , the previous calculations show that  $g(p) = \mu'_1(p)$ , and then

$$\hat{g}_n = m'_{1n} = \bar{X}_n$$

is an estimator of  $g(p)$  produced by the method of moments.  $\square$

**Exercise 4.4.3.** Let  $X_1, X_2, \dots, X_n$  be a random sample from the ‘displaced’ exponential population with density

$$f(x; \alpha, \lambda) = \frac{1}{\lambda} e^{(x-\alpha)/\lambda} I_{(\alpha, \infty)}(x),$$

where  $\theta = (\alpha, \lambda) \in \mathbb{R} \times (0, \infty) = \Theta$ . Use the method of moments to generate estimators of  $\alpha$  and  $\lambda$ , and investigate their unbiasedness and consistency.

**Solution.** To begin with, the first two population moments of the given population will be determined, The task is simplified by the following observation:

If  $X$  has the density  $f(x; \alpha, \lambda)$ , then  $Y = (X - \alpha)/\lambda \sim \text{Exponential}(1)$ .

It follows that  $E[Y] = 1 = \text{Var}[Y] = E[Y^2] - 1$ , so that

$$E \left[ \frac{X - \alpha}{\lambda} \right] = 1 = E \left[ \left( \frac{X - \alpha}{\lambda} \right)^2 \right] - 1.$$

The first part of this relation yields that

$$\mu'_1(\theta) = E_\theta [X] = \alpha + \lambda \tag{4.4.1}$$

whereas the second part implies that  $E_\theta [(X - \alpha)^2] = 2\lambda^2$ , so that

$$E_\theta [X^2 - 2X\alpha + \alpha^2] = 2\lambda^2,$$

a relation that leads to

$$\begin{aligned} \mu'_2(\theta) = E_\theta[X^2] &= 2\lambda^2 - \alpha^2 + 2E_\theta[X]\alpha \\ &= 2\lambda^2 - \alpha^2 + 2(\lambda + \alpha)\alpha \\ &= 2\lambda^2 + 2\alpha\lambda + \alpha^2 \\ &= 2\lambda(\lambda + \alpha) + \alpha^2 \\ &= 2\lambda\mu'_1(\theta) + \alpha^2 \end{aligned} \tag{4.4.2}$$

where (4.4.1) was used in the last step. Using (4.4.1) again, notice that  $\lambda = \mu'_1(\theta) - \alpha$ , and then

$$\begin{aligned}\mu'_2(\theta) &= 2(\mu'_1(\theta) - \alpha)\mu'_1(\theta) + \alpha^2 \\ &= 2\mu'_1(\theta)^2 - 2\mu'_1(\theta)\alpha + \alpha^2 = \mu'_1(\theta)^2 + (\mu'_1(\theta) - \alpha)^2.\end{aligned}$$

Consequently,

$$\lambda^2 = (\mu_1(\theta) - \alpha)^2 = \mu'_2(\theta) - \mu'_1(\theta)^2;$$

observe that the relation  $\mu'_2(\theta) - \mu'_1(\theta)^2$  is the population variance, so that  $\mu'_2(\theta) - \mu'_1(\theta)^2 \geq 0$ . Hence, recalling the  $\lambda > 0$ ,

$$\lambda = \sqrt{\mu'_2(\theta) - \mu'_1(\theta)^2},$$

and

$$\alpha = \mu'_1(\theta) - \lambda = \mu'_1(\theta) - \sqrt{\mu'_2(\theta) - \mu'_1(\theta)^2}.$$

From these expressions, the method of moments renders the following estimators:

$$\begin{aligned}\hat{\lambda}_n &= \sqrt{m'_{2n} - (m'_{1n})^2} \\ \hat{\alpha}_n &= m'_{1n} - \sqrt{m'_{2n} - (m'_{1n})^2}.\end{aligned}$$

Since  $\lambda$  and  $\alpha$  are continuous functions of  $\mu'_1$  and  $\mu'_2$ , it follows that these estimators are consistent, and since they are not linear functions of  $\mu'_1$  and  $\mu'_2$ , they are not unbiased. Before concluding, it is interesting to observe that  $m'_2 - (m'_1)^2 = \sum_{i=1}^n X_i^2/n - \bar{X}_n^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2/n$  is the sample variance  $\tilde{S}_n^2$  (with denominator  $n$ ), and then  $\hat{\lambda}_n$  is the sample standard deviation  $\tilde{S}_n$ , whereas  $\hat{\alpha}_n = \bar{X}_n - \tilde{S}_n$ .  $\square$

**Exercise 4.4.4.** Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from the discrete uniform distribution on the set  $\{1, 2, \dots, \theta\}$  where  $\theta$  is an unknown positive integer. Use the method of moments to find an estimator of  $\theta$ .

**Solution.** To express the parameter  $\theta$  in terms of the population moments, just notice that if  $X \sim \text{Uniform}(\{1, 2, \dots, \theta\})$  then  $\mu'_1(\theta) = E_\theta[X] = (1 + \theta)/2$ , so that  $\theta = 2\mu'_1(\theta) - 1$ . Hence, the method of moments produces the estimator  $\hat{\theta}_n = 2m'_{1n} - 1 = 2\bar{X}_n - 1$ ; since  $\theta$  is a linear function of  $\mu'_1(\theta)$ , it follows that the estimators  $\hat{\theta}_n$  are unbiased and the sequence  $\{\hat{\theta}_n\}$  is consistent.  $\square$

**Exercise 4.4.5.** Let  $f_1(x)$  and  $f_2(x)$  be two densities with means  $\mu_1$  and  $\mu_2$ , respectively, where  $\mu_1 \neq \mu_2$ . For each  $\theta \in [0, 1] = \Theta$  define the mixture

$$f(x; \theta) = \theta f_1(x) + (1 - \theta) f_2(x).$$

Use the method of moments to find an estimator of  $\theta$  based on a random sample of size  $n$  from  $f(x; \theta)$ .

**Solution.** Observe that if  $X \sim f(x; \theta)$  then

$$\begin{aligned} \mu'_1(\theta) &= E_\theta[X] \\ &= \int_{\mathbb{R}} x[\theta f_1(x) + (1 - \theta) f_2(x)] dx \\ &= \theta \int_{\mathbb{R}} x f_1(x) + (1 - \theta) \int_{\mathbb{R}} x f_2(x) dx \\ &= \theta \mu_1 + (1 - \theta) \mu_2 = \mu_2 + \theta(\mu_1 - \mu_2); \end{aligned}$$

notice that the expectations of the densities  $f_1$  and  $f_2$  ( $\mu_1$  and  $\mu_2$ , respectively) are known numbers. Since  $\mu_1 \neq \mu_2$ , it follows that

$$\theta = \frac{\mu'_1(\theta) - \mu_2}{\mu_1 - \mu_2}$$

and then, when a random sample  $X_1, X_2, \dots, X_n$  of the density  $f(x; \theta)$  is available, the method of moments prescribes the estimator

$$\hat{\theta}_n = \frac{m'_{1n} - \mu_2}{\mu_1 - \mu_2} = \frac{\bar{X}_n - \mu_2}{\mu_1 - \mu_2},$$

which is unbiased. □

**Exercise 4.4.6.** Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from the  $Gamma(\alpha, \lambda)$  distribution, where  $\theta = (\alpha, \lambda) \in \Theta = (0, \infty) \times (0, \infty)$ . Use the method of moments to obtain estimators of  $\alpha$  and  $\lambda$ .

**Solution.** The starting point is to evaluate the moments of order one and two of the  $Gamma(\alpha, \lambda)$  distribution. It is known that if  $X \sim Gamma(\alpha, \lambda)$ , then

$$\mu'_1(\theta) = E_\theta[X] = \frac{\alpha}{\lambda}, \quad \text{and} \quad \mu'_2(\theta) = E_\theta[X^2] = \frac{\alpha(\alpha + 1)}{\lambda^2}. \quad (4.4.3)$$



To express  $\alpha$  and  $\lambda$  in terms of  $\mu'_1(\theta)$  and  $\mu'_2(\theta)$ , notice that

$$\frac{\mu'_2(\theta)}{\mu'_1(\theta)^2} = \frac{\alpha(\alpha + 1)/\lambda^2}{\alpha^2/\lambda^2} = \frac{\alpha + 1}{\alpha} = 1 + \frac{1}{\alpha}.$$

Hence,

$$\frac{\mu'_2(\theta) - \mu'_1(\theta)^2}{\mu'_1(\theta)^2} = \frac{1}{\alpha},$$

which is equivalent to

$$\alpha = \frac{\mu'_1(\theta)^2}{\mu'_2(\theta) - \mu'_1(\theta)^2}.$$

Combining this expression with the first equality in (4.4.3), it follows that

$$\lambda = \frac{\alpha}{\mu'_1(\theta)} = \frac{\mu'_1(\theta)}{\mu'_2(\theta) - \mu'_1(\theta)^2}.$$

Then the method of moments estimation prescribes the estimators

$$\hat{\alpha}_n = \frac{(m'_{1n})^2}{m'_{2n} - (m'_{1n})^2},$$

and

$$\hat{\lambda}_n = \frac{m'_{1n}}{m'_{2n} - (m'_{1n})^2}.$$

Since  $m'_{1n} = \bar{X}_n$  and

$$m'_{2n} - (m'_{1n})^2 = \sum_{i=1}^n X_i^2/n - \bar{X}_n^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2/n = \tilde{S}_n^2$$

the above estimators can be expressed in more familiar terms:

$$\hat{\alpha}_n = \frac{\bar{X}_n^2}{\tilde{S}_n^2}, \quad \text{and} \quad \hat{\lambda}_n = \frac{\bar{X}_n}{\tilde{S}_n^2}.$$

Since  $\alpha$  and  $\lambda$  are continuous functions of  $\mu'_1(\theta)$  and  $\mu'_2(\theta)$ , the sequences  $\{\hat{\alpha}_n\}$  and  $\{\hat{\lambda}_n\}$  are consistent.  $\square$

**Remark 4.4.1.** An interesting aspect of the precedent problem is that method of moments allowed to obtain explicit formulas for the estimators of  $\alpha$  and  $\lambda$ . In contrast, the maximum likelihood estimators of  $\alpha$  and  $\lambda$  must be determined numerically for each data set.  $\square$

**Exercise 4.4.7.** Let  $X_1, X_2, \dots, X_n$  be a sample of the *Bernoulli*( $p$ ) distribution, where  $p \in [0, 1]$ . Find the moments estimator of  $p$ .

**Solution.** If  $X \sim \text{Bernoulli}(p)$ , then  $\mu'_1(p) = E_p[X] = p$ , so that the moments estimator of  $p$  is  $\hat{p}_n = m'_{1n} = \bar{X}_n$ .  $\square$

**Exercise 4.4.8.** Let  $X_1, X_2, \dots, X_n$  be a sample of the density

$$f(x; \theta) = \frac{\theta}{x^2} I_{[\theta, \infty)}(x),$$

where  $\theta \in \Theta = (0, \infty)$ . Find an estimator of  $\theta$  using the method of moments.

**Solution.** As usual, the starting point is to compute moments of the distribution until the parametric quantity to be estimated— $\theta$  in the present case—can be expressed in terms of the available moments. Let  $X$  have the density  $f(x; \theta)$  and notice that

$$\mu'_1(\theta) = E_\theta[X] = \int_{\mathbb{R}} x f(x; \theta) dx = \int_{\theta}^{\infty} x \frac{\theta}{x^2} dx = \theta \int_{\theta}^{\infty} \frac{1}{x} dx = \infty.$$

Since the first population moment is not finite, all the other moments of order  $k \geq 1$  are  $\infty$ . Thus,  $\theta$  can not be expressed in terms of the moments of  $X$  which have order equal or larger than one. However, in the present case an alternative is to consider fractional moments, that is, expected values of fractional powers of  $X$ . For instance, consider  $X^{1/2}$  and notice that

$$\begin{aligned} \mu'_{1/2}(\theta) &= E_\theta[X^{1/2}] \\ &= \int_{\mathbb{R}} x^{1/2} f(x; \theta) dx = \int_{\theta}^{\infty} x^{1/2} \frac{\theta}{x^2} dx = \theta \int_{\theta}^{\infty} \frac{1}{x^{3/2}} dx = 2\theta^{1/2}. \end{aligned}$$

Therefore,

$$\theta = \left( \frac{\mu'_{1/2}(\theta)}{2} \right)^2,$$

an expression that leads to consider the estimator

$$\hat{\theta}_n = \left( \frac{m'_{1/2, n}}{2} \right)^2 = \left( \frac{\sum_{i=1}^n X_i^{1/2}/n}{2} \right)^2 = \left( \frac{\sum_{i=1}^n X_i^{1/2}}{2n} \right)^2,$$

which is biased, since  $\theta$  is not a linear function of  $\mu'_{1/2}(\theta)$ ; however,  $\theta$  is a continuous function of  $\mu_{1/2}(\theta)$ , so that the sequence  $\{\hat{\theta}_n\}$  is consistent. Of course, other fractional moments may be used in this problem.  $\square$

**Exercise 4.4.9.** Let  $X_1, X_2, \dots, X_n$  be a sample of the *Poisson*( $\lambda$ ) distribution, where  $\lambda \in [0, \infty)$ . Find the moments estimator of  $\lambda$ .

**Solution.** If  $X \sim \text{Poisson}(\lambda)$ , then  $\mu'_1(\lambda) = E_\lambda[X] = \lambda$ , so that the moments estimator of  $\lambda$  is  $\hat{\lambda}_n = m'_{1n} = \bar{X}_n$ .  $\square$

**Exercise 4.4.10.** Let  $X_1, X_2, \dots, X_n$  be a sample of the  $\mathcal{N}(0, \sigma^2)$  distribution, where  $\sigma \in (0, \infty)$ . Find the moments estimator of  $\sigma^2$  and analyze the consistency of the sequence  $\{\hat{\sigma}^2\}$ .

**Solution.** If  $X \sim \mathcal{N}(0, \sigma^2)$ , then  $\mu'_1(\sigma) = E_\sigma[X] = 0$ , so that  $\sigma^2$  can not be expressed in terms of  $\mu'_1(\sigma)$  and it is necessary to compute more moments of  $X$ . Next, observe that  $\mu'_2(\sigma) = E_\sigma[X^2] = \text{Var}_\sigma[X] = \sigma^2$ , and it follows that the interesting parametric function— $\sigma^2$  in the present problem—equals the second population moment. Thus, the method of moments prescribes the estimator

$$\hat{\sigma}^2 = m'_{2n} = \frac{1}{n} \sum_{i=1}^n X_i^2;$$

since  $\sigma^2$  is a linear function of  $\mu'_2(\sigma)$ ,  $\hat{\sigma}^2$  is an unbiased estimator of  $\sigma^2$ .  $\square$

# References

- [1]. T. M. Apostol (1980), *Mathematical Analysis*, Addison Wesley, Reading, Massachusetts.
- [2]. A. A. Borovkov (1999), *Mathematical Statistics*, Gordon and Breach, New York
- [3]. E. Dudewicz y S. Mishra (1998). *Mathematical Statistics*, Wiley, New York.
- [4]. W. Fulks (1980), *Cálculo Avanzado*, Limusa, México, D. F.
- [5]. F. A. Graybill (2000), *Theory and Application of the Linear Model*, Duxbury, New York.
- [6]. F. A. Graybill (2001), *Matrices with Applications in Statistics* Duxbury, New York.
- [7]. D. A. Harville (2008), *Matrix Algebra Form a Statistician's Perspective*, Springer-Verlaf, New York.
- [8]. A. I. Khuri (2002), *Advanced Calculus with Applications in Statistics*, Wiley, New York.
- [9]. E. L. Lehmann and G. B. Casella, (1998), *Theory of Point Estimation*, Springer, New York.
- [10]. M. Loève (1984), *Probability Theory, I*, Springer-Verlag, New York.
- [11]. D. C. Montgomery (2011), *Introduction to Statistical Quality Control*, 6th Edition, Wiley, New York.
- [12]. A. M. Mood, D. C. Boes and F. A. Graybill (1984), *Introduction to the Theory of Statistics*, McGraw-Hill, New York.
- [13]. W. Rudin (1984), *Real and Complex Analysis*, McGraw-Hill, New York.
- [14]. H. L. Royden (2003), *Real Analysis*, MacMillan, London.
- [15]. J. Shao (2010), *Mathematical Statistics*, Springer, New York.
- [16]. D. Wackerly, W. Mendenhall y R. L. Scheaffer (2009), *Mathematical Statistics with Applications*, Prentice-Hall, New York.