DISTRIBUCIÓN ASINTÓTICA DE ESTIMADORES DE MOMENTOS

ALBERTO AGUILAR RODRÍGUEZ

TESIS

Presentada como Requisito Parcial para
Obtener el Grado de:

MAESTRO EN
ESTADÍSTICA APLICADA

UNIVERSIDAD AUTÓNOMA AGRARIA
"ANTONIO NARRO"
PROGRAMA DE GRADUADOS
Buenavista, Saltillo, Coahuila, México
Abril de 2012

Universidad Autónoma Agraria Antonio Narro
Subdirección de Postgrado

DISTRIBUCIÓN ASINTÓTICA DE ESTIMADORES DE MOMENTOS

TESIS

Por:

ALBERTO AGUILAR RODRÍGUEZ

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: ____________________________
M. C. Luis Rodríguez Gutiérrez

Asesor: ____________________________
M. C. Félix de Jesús Sánchez Pérez

____
Dr. Fernando Ruiz Zárate
Subdirector de Postgrado
Buenavista, Saltillo, Coahuila, Abril de 2012

COMPENDIO

DISTRIBUCIÓN ASINTÓTICA DE ESTIMADORES DE MOMENTOS

Por

ALBERTO AGUILAR RODRÍGUEZ

MAESTRÍA EN

ESTADÍSTICA APLICADA

UNIVERSIDAD AUTÓNOMA AGRARIA
ANTONIO NARRO

BUENAVISTA, SALTILLO, COAHUILA, Abril de 2012

Dr. Rolando Cavazos Cadena –Asesor–

Palabras clave: Teorema central de límite, Convergencia en distribución, Método delta, Consistencia y normalidad asintótica, Función diferenciable.

Este trabajo trata sobre ideas fundamentales en la teoría de *estimación puntual* para modelos estadísticos paramétricos. El principal objetivo de la exposición es ilustrar conceptos básicos, como insesgamiento, consistencia y normalidad asintótica, por medio de una serie de ejemplos cuidadosamente analizados. Se inicia con el estudio de dos métodos de estimación, a saber, la técnica de verosimilitud máxima, y el método de momentos, y se concluye con una deducción detallada de la distribución límite de estos últimos estimadores, la cual combina el teorema central de límite con la propiedad de invarianza de la propiedad de convergencia hacia una distribución normal bajo transformaciones diferenciables.

ABSTRACT

ASYMPTOTIC DISTRIBUTION OF MOMENTS ESTIMATORS

BY

ALBERTO AGUILAR RODRÍGUEZ

MASTER IN

APPLIED STATISTICS

UNIVERSIDAD AUTÓNOMA AGRARIA
ANTONIO NARRO

BUENAVISTA, SALTILLO, COAHUILA, April, 2012

Dr. Rolando Cavazos Cadena –Advisor–

Key Words: Central limit theorem, Invariance property of the convergence in distribution to a normal variate, Delta method, Consistency and asymptotic normality, Differentiable function.

This work is concerned with basic aspects of the theory of *point estimation* in the context of parametric statistical models. The main objective of the exposition is to illustrate fundamental notions, like unbiasedness, consistency and asymptotic normality, presenting a series of fully analyzed examples. To achieve this goal, two methods of constructing estimators, namely, the maximum likelihood technique and the method of moments, are carefully presented and, combining the central limit theorem with the invariance property of the asymptotic normality under the application of smooth functions, a detailed derivation of the limit distribution of moments estimators is given.

Contents

empty

Chapter 1

Perspective of This Work

In this chapter presents a brief outline of this work, establishing the main objectives and describing the organization of this subsequent material. The main contributions are clearly stated.

1.1. Introduction

This work deals with the problem of *parametric point estimation.* Undoubtedly, the area of point estimation lays in the core of the statistical methodology, and a major step in every theoretical or applied analysis is the determination of estimates (*i.e.*, approximations) to some unknown quantities in terms of the observed data; moreover, every treatise on statistics dedicates a good amount of space to describe methods of constructing estimators and to analyze its properties (Dudewicz and Mishra , 1988, Wackerly *et al.* 2009, Lehmann and Casella, 1999, or Graybill, 2000).

The topics analyzed 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;

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

(iii) The analysis of the asymptotic behavior of momenst estimators.

Each one of these topics are briefly described below.

1.2. Parametric 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 $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ 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) $\mathrm{I\!R}^n$:

$$P_{\mathbf{X}} \in \mathcal{F}. \tag{1.2.1}$$

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 θ whose components are real numbers; in such a case the set of possible values of θ, which is referred to as the parameter space, will be denoted be Θ and $\mathcal{F}$ can be written as

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

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 $\mathrm{I\!R}^n$

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

The parameter θ^* satisfying this relation for every (Borel) subset of $\mathrm{I\!R}^n$ 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 θ^*; it is only supposed that θ^* belongs to the parameter space Θ, and the main objective of the analyst is to determine θ^* using the value attained by the vector $\mathbf{X}$, say $\mathbf{X} = \mathbf{x}$. Indeed, the lack of exact knowledge of θ^* 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, θ^* 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 θ^*. 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 θ^*. 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. Contribution and Main Goals

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,

(iii) To show the usefulness of elementary analytical tools in the analysis of basic notions in the theory of point estimation, as unbiasedness, consistency, asymptotic normality and convergence in distribution.

On the other hand, this work is also concerned with the more specific problem of *estimating a quantile of a continuos distribution function*, and the main purpose in this direction is the following:

(iv) To provide a derivation of the asymptotic distribution of the sequence of moments estimators.

The analysis performed below to achieve these objectives, as well as the numerous and detailed examples on the theory, represent *the main contribution of this work*. Indeed, due to the technical difficulties that a rigorous analysis of a statistical problem requires, frequently the delicate and more demanding parts of the arguments are usually described, but not proved; in the present exposition a serious effort has been made to derive and explain the results in a clear and concise manner, highlighting the essential statistical and analytical tools that are used to establish the conclusions, and indicating clearly the basic steps of the arguments.

1.4. The Origin of This Work

This work arose form the activities developed in 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 founder students were Mary Carmen Ruiz Moreno and Alfonso Soto Almaguer. The basic aims of the project are:

(i) To be a framework were 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 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, were the results of their activities are discussed with the members of the program.

Presently, 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 ne 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 that are used below to illustrate the basic statistical notions studied in this work are a direct product of the activities of the different participants in the project, and enjoying the discussions and different perspectives of analysis of a problem has been an experience in a lifetime. In particular, it is a real pleasure to thank to my classmate, Alfonso Soto Almaguer, by his generous help and clever suggestions.

1.5. The Organization

The remainder of this work has been organized as follows:

In Chapter 2 the 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 exposition 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 that data, the estimate of the unknown parameter θ is the value $\hat{\theta}$ in the parameter space that assigns highest probability to the observations. Then, Chapter 4 is concerned with the method of moments and the presentation concludes in Chapter 5 analyzing the limit distribution of a sequence of moments estimators.

Chapter 2

Consistency and Unbiasednenss

This chapter is concerned with the basic notions in the theory of point estimation. After a brief description of a parametric statistical model, the problem of estimating a function of the unknown parameter is considered. The idea of estimator is introduced and the main objective is to illustrate fundamental properties, as unbiasedness, consistency and asymptotic unbiasedness. These goals are achieved by analyzing in detail several examples involving familiar distributions.

2.1. Introduction

Let $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$ be an observable random vector. A parametric statistical model for $\mathbf{X}$ prescribes a family $\{P_\theta\}_{\theta\in\Theta}$ of probability distributions for $\mathbf{X}$, where the set of indices Θ is referred to as the *parameter space* and is a subset of $\mathbb{R}^k$ for some integer $k \geq 1$. Thus, a statistical model can be thought of as the hypothesis that of the distribution of $\mathbf{X}$ coincides with P_θ for some parameter $\theta \in \Theta$, but the 'true' parameter value—the one which corresponds to the distribution of the observation vector $\mathbf{X}$—is unknown. The statistical model is briefly described by writing

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

Frequently the components $X_1, X_2, \ldots, X_n$ of the random vector $\mathbf{X}$ are independent and identically distributed with common density or probability function $f(x;\theta)$, and in this case the model will be written as

$$X_i \sim f(x;\theta), \quad \theta \in \Theta,$$

where it is understood that the involved variables are independent with the common distribution determined by $f(x;\theta)$.

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 the observation vector $\mathbf{X}$ are combined in some way to obtain a function

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

and, after observing $\mathbf{X} = \mathbf{x} = (x_1, x_2, \ldots, x_n)$, the function T_n is evaluated at $\mathbf{x}$ to obtain $T_n(\mathbf{x}) = T_n(x_1, x_2, \ldots, x_n)$, a value that 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 $\mathbf{X} = \mathbf{x}$. Notice that this idea of estimator is quite general; indeed, an estimator is an arbitrary function of the available data whose values are used as an approximation of the unknown value of the parametric quantity $g(\theta)$; thus, some criteria are needed to distinguish among diverse estimators and to select one with desirable properties.

2.2. The Estimation Problem

In this section the problem of point estimation is discussed, and two basic properties of estimators are discussed, namely, unbiasedness and consistency. A parametric statistical model for a random (and observable) vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ postulates that the distribution of $\mathbf{X}$ is a member of a a family $\{P_\theta\}_{\theta\in\Theta}$ of probability distributions on the Borel subsets of $\mathbb{R}^n$. The set of indices Θ is referred to as the *parameter space* and is a subset of an Euclidean space $\mathbb{R}^k$. Thus, a statistical model stipulates that the distribution of $\mathbf{X}$ coincides with P_θ for some parameter $\theta \in \Theta$, but the 'true' parameter value—the one which corresponds to the distribution of the observation vector $\mathbf{X}$—is unknown. Such a 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 the observation vector $\mathbf{X}$ are combined in some way to obtain a function

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

whose values are used as 'approximations' of the unknown quantity $g(\theta)$. Thus, after performing the underlying experiment and observing $\mathbf{X} = \mathbf{x} = (x_1, x_2, \ldots, x_n)$, the value $T_n(\mathbf{x}) = T_n(x_1, x_2, \ldots, x_n)$ is used as the analyst's guess for the $g(\theta)$. 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

$$E_\theta[T_n] = g(\theta), \quad \theta \in \Theta;$$

the subindex θ in the expectation operator is used to indicate that the expected value is computed under the condition that θ is the true parameter value. Generally, the value attained by an estimator is not equal to $g(\theta)$ but, if the estimator T_n is unbiased and the experiment producing the sample $\mathbf{X}$ is repeated again and again, the estimates $T_{n\,1}, T_{n\,2}, T_{n\,3}, \ldots$ obtained at each repetition satisfy that, with probability 1,

$$\frac{T_{n\,1} + T_{n\,2} + T_{n\,3} + \cdots + T_{n\,k}}{k}$$

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

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

when the x_is are zero or one, it follows that

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

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 that n, as $g(\theta) = \theta^{n+1}$. Thus, the unbiasdness property may be too restrictive, and it is is possible to have that an unbiased estimator does not exists in some cases of interest.

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

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

so that T_n is unbiased if $b_{T_n}(\theta) = 0$ for every $\theta \in \Theta$. To compute the bias of an estimator T_n it is necessary to compute 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 $\{T_n\}_{n=1,2,\ldots}$ of estimators of $g(\theta)$ is *asymptotically unbiased* if

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

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

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

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

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

$$T_n \xrightarrow{\mathrm{P}_\theta} g(\theta).$$

2.3. Instruments to Study the Consistency Property

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 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 $X_1, X_2, \ldots, X_n, \ldots$ are independent and identically distributed, setting

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

the law of large numbers yields that $T_n \xrightarrow{\mathrm{P}_\theta} g(\theta)$, that is the sequence $\{T_n\}$ of estimators of $g(\theta)$ is consistent.

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

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

$$T_{i\,n} \xrightarrow{\mathrm{P}_\theta} g_i(\theta), \quad i = 1, 2, \ldots, r.$$

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

$$G(T_{1\,n}, T_{2\,n}, \ldots, T_{r\,n}) \xrightarrow{\mathrm{P}_\theta} G(g_1(\theta), g_2(\theta), \ldots, g_r(\theta)).$$

(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

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

The notation $T_n \xrightarrow{L^p} g(\theta)$ will be used to indicate that this 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 $\theta \in \Theta$, $E_\theta[(T_n - g(\theta))^2] \to 0$ as $n \to \infty$. 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 $\varepsilon > 0$,

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

so that

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

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 for some $p > 0$, than to prove consistency directly. When considering consistency in the mean of order 2, it is useful to keep in mind the the mean square error $E_\theta[(T_n - g(\theta))^2]$, the variance and the bias function of T_n are related by

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

2.4. Unbiasedness and Consistency in Simple Cases

In this section the ideas recently introduced will be illustrated for statistical models involving some common distributions and standard statistical concepts, like independence. The idea is to generate some insight on the necessary computations to evaluate the bias of an estimator, and to establish the consistency or asymptotic unbiasedness of a sequence of estimators.

Exercise 2.4.1. Let T_n and T_n' be two independent unbiased and consistent estimators of θ.

(a) Find and unbiased estimator of θ^2;

(b) Find and 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 θ,

$$E_\theta[T_n T_n] = E_\theta[T_n] E_\theta[T_n'] = \theta \cdot \theta = \theta^2$$

and then $T_n T_n'$ is an unbiased estimator of θ^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 θ, combining the convergences $T_n \xrightarrow{\mathrm{P}_\theta} \theta$ and $T_n' \xrightarrow{\mathrm{P}_\theta} \theta$ with the continuity theorem, it follows that $T_nT_n' \xrightarrow{\mathrm{P}_\theta} \theta^2$ and $T_n(T_n'-1) \xrightarrow{\mathrm{P}_\theta} \theta(\theta-1)$, so that the estimators in parts (a) and (b) are consistent. □

Exercise 2.4.2. Let $X_1, X_2, \ldots, X_n$ 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 $\overline{X}_n$ is a biased estimator of θ and find its bias $b_n(\theta)$,

(b) Does $\lim_{n\to\infty} b_n(\theta) = 0$ for all θ?

(c) Is $\overline{X}_n$ consistent in mean square?

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

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

(a) Since $E_\theta[\overline{X}_n] = \mu(\theta) \neq \theta$, the sample mean $\overline{X}_n$ is a biased estimator of θ, 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 $\lim_{n\to\infty} b_n(\theta) = 1 - 7\theta/6$, and then $b_n(\theta)$ does not converge to zero at any parameter value; in particular, considering $\overline{X}_n$ as an estimator of θ, the sequence $\{\overline{X}_n\}$ is not asymptotically unbiased.

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

2.5. The Usefulness of a Symmetry Property

The objective of this section is to show that the computations to analyze the properties of an estimator can be eased by the application of symmetry properties of the underlying probability measures postulated by the model.

Exercise 2.5.1. Let $X_1, X_2, \ldots, X_n$ 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 θ belongs to $\mathbb{R}$. If $Y_1 \leq Y_2 \leq \cdots \leq Y_n$ are the order statistics, show that $T_n = (Y_1 + Y_n)/2$ is an unbiased estimator of θ. □

Solution. The key fact to keep in mind is that the underlying density is symmetric about θ, 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

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

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

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

Notice now that

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

whereas

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

Combining the three last displays, it follows that

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

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 θ. □

2.6. Additional Examples

The remainder of the chapter is dedicated to provide further illustrations of the basic conceptos introduced in Section 2.2.

Exercise 2.6.1. (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 θ based on a sample of size 1.

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

(c) Let $Y_1 \leq Y_2 \leq \cdots \leq Y_n$ 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 θ based on a sample of size 1.

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

(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, for $\theta \leq y \leq 1$,

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

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

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

To obtain an explicit formula, chage 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

$$E_\theta[Y_1 - \theta] = \frac{n(1-\theta)\Gamma(3/2)\Gamma(n)}{\Gamma(n+3/2)}. \tag{2.6.1}$$

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 (2.6.1) 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. □

Exercise 2.6.2. Let $X_1, X_2, \ldots, X_n$ be independent random variables each one with distribution *Gamma* (α, λ), which has density

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

where α and λ are positive. Suppose that α is known, and define

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

(a) Show that T_n is an unbiased estimator of β 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 β^2.

Solution. To begin with, recall that the first and second moments of the *Gamma* (α, λ) distribution are given by

$$E_\lambda[X_1] = \frac{\alpha}{\lambda} = \alpha\beta, \quad \text{and} \quad E_\lambda[X_1^2] = \frac{\alpha(\alpha+1)}{\lambda^2} = \alpha(\alpha+1)\beta^2, \qquad (2.6.2)$$

relations that yield

$$\text{Var}_\lambda[X_1] = \frac{\alpha}{\lambda^2} = \alpha\beta^2. \qquad (2.6.3)$$

(a) The first equation in (2.6.2) yields that $E_\lambda[\overline{X}_n] = \alpha\beta$, and then $E_\lambda[T_n] = E_\lambda[\overline{X}_n/\alpha] = \beta$, that is, T_n is an unbiased estimator of β. On the other hand, from (2.6.3) it follows that $\text{Var}_\lambda[\overline{X}_n] = \text{Var}_\lambda[X_1]/n = \alpha\beta^2/n$, and then

$$\begin{aligned} E_\lambda[(T_n - \beta)^2] &= \text{Var}_\lambda[T_n] \\ &= \text{Var}_\lambda[\overline{X}_n/\alpha] = \frac{1}{\alpha^2}\text{Var}_\lambda[\overline{X}_n] = \frac{\beta^2}{n\alpha} \to 0 \quad \text{as} \quad n \to \infty, \end{aligned}$$

so that T_n is consistent in mean square as estimator of β.

(b) The second equality in (2.6.2) and the law of large numbers together yield that

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

and

$$\frac{X_1^2 + X_2^2 + \cdots + X_n^2}{n} \xrightarrow{\mathrm{P}_\lambda} \alpha(\alpha+1)\beta^2.$$

Hence,

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

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

Exercise 2.6.3. Let $X_1, X_2, \ldots, X_n$ be independent random variables with *Exponential*(λ) 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 λ is $1/\overline{X}_n$. Show that this estimator is biased, and compute the bias $b_{1/\overline{X}_n}(\lambda)$.

(b) Based on part (a), find an unbiased estimator of λ.

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

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

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,

$$E_\lambda[1/\overline{X}_n] = E_\lambda[n/Y] = \frac{n\lambda}{n-1} = \lambda + \frac{\lambda}{n-1}; \tag{2.6.4}$$

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

(b) Equality (2.6.4) yields that $E_\lambda[(n-1)/(n\overline{X}_n)] = \lambda$, so that

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

is an unbiased estimator of λ. □

Exercise 2.6.4. Let $X_1, X_2, \ldots, X_n$ be a random sample from the triangular density

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

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

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

$$f((a+b)/2 + w; a, b) = \frac{(b-a)/2 - |w|}{c} I_{-(b-a)2,\ (b-a)/2)}(w), \qquad (2.6.5)$$

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

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

and using (2.6.5), 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 $\{\overline{X}_n\}$ and $\{S_n^2\}$ as estimators of μ and σ^2, it is important to keep in mind that they are always consistent, as it is shown in Dudewciz y Mishra (1998). □

Chapter 3

Maximum Likelihood Estimation

The idea of estimator as presented before is rather arbitrary, in the sense that any function T of the observations is considered as an estimator of a parametric quantity $g(\theta)$, as soon as the analyst is willing to think that the values attained by T can be used as approximations for $g(\theta)$. In this chapter a technique to generate estimators that can be reasonably thought of as 'approximations' for $g(\theta)$ is presented, namely the *method of maximum likelihood.* The technique is based on an intuitive principle that can be roughly described as follows: After observing the value attained by the random vector $\mathbf{X}$, say $\mathbf{X} = \mathbf{x}$, the estimate of the unknown parameter θ 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 form $\hat{\theta}$. The objective of the chapter is to present a formal description of these idea, and illustrate its application.

3.1. Introduction

In this section a measure of the likelihood of an observation $\mathbf{X}$ under the different parameter values is introduced, and then it is used to generate estimators of parametric quantities. Consider a statistical model

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

and, as a starting point, 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 θ 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_θ, and then is a measure of the 'likelihood' of the observation $\mathbf{x}$ if θ is the true parameter. Thus, the *likelihood function* corresponding to the data $\mathbf{X} = \mathbf{x}$ is defined by

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

When $\mathbf{X}$ is continuous it has a density $f_X(\mathbf{x};\theta)$ depending on θ, 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 θ 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

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

it follows that the likelihood function is (approximately) proportional to the probability of observing $\mathbf{X} = \mathbf{x}$; moreover, the proportionallty constant does not depend on θ, and then when the maximizer of the function $L(\cdot;\mathbf{X})$ is determined, such a point also maximizes (approximately) the probability of the observation $\mathbf{X} = \mathbf{x}$.

3.2. Maximum Likelihood Estimators and Invariance Principle

The *maximum likelihood estimator of* θ, 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 θ, that is, $\hat\theta(\mathbf{X})$ satisfies

$$L(\hat\theta;\mathbf{X}) \geq L(\theta;\mathbf{X}), \quad \theta \in \Theta. \tag{3.2.1}$$

This maximum likelihood method to construct estimators of θ 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.

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

$$L_{\tilde{g}}(\mathbf{X}) := \max_{\theta:\ g(\theta)=\tilde{g}} L(\theta;\mathbf{X}), \tag{3.2.2}$$

so that $L_{\tilde{g}}(\mathbf{X})$ is the largest likelihood that can be achieved among the parameters θ 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 $\tilde{g}$:

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

The maximizing value can be determined easily. Set

$$\hat{g} = g(\hat{\theta}) \tag{3.2.3}$$

and notice that (3.2.1) and (3.2.2) 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 (3.2.3). 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 θ. This result is called *the invariance principle (or property)* of the maximum likelihood estimation procedure.

3.3. Logarithmic Transformation

Before going to the analysis of specific examples, it is useful to note that, when the observation vector $\mathbf{X}$ is a sample $(X_1, X_2, \ldots, X_n)$ 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 plays a central to analyze this optimization problem, In particular, of the likelihood function is 'smooth' as a function of θ and the maximizer belongs to the interior of the parameter space, the following *likelihood equation* is satisfied:

$$D_\theta \mathcal{L}(\theta;\mathbf{X}) = 0, \tag{3.3.1}$$

where D_θ is the gradient operator, whose components are the partial derivatives with respect to each element of the parameter θ; thus, when θ is a vector, (3.3.1) 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 (3.3.1) is no longer necessarily satisfied by the optimizer $\hat{\theta}$.

3.4. Elementary Applications

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. The first example concerns a normal model with unitary coefficient of variation.

Exercise 3.4.1. Let $X_1, X_2, \ldots, X_n$ be a random sample from the $\mathcal{N}\left(\theta, \theta^2\right)$ distribution, where $\theta \in (0, \infty)$. Find the maximum likelihood estimator of θ. 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 it logarithm is given by

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

Hence

$$\partial_\theta \mathcal{L}(\theta; \mathbf{X}) = -\frac{n}{\theta} + \sum_{i=1}^{n} \frac{X_i(X_i - \theta)}{\theta^3}$$

From this expression, direct calculations show that the equation $\partial_\theta \mathcal{L}(\theta; \mathbf{X}) = 0$ is equivalent to $\theta^2 + m_1\theta - m_2 = 0$, where m_i is the ith sample moment about 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 $\partial\mathcal{L}(\theta; \mathbf{X}) \to -\infty$ as $\theta \to 0$ or $\theta \to \infty$, it follows that θ^* maximizes the likelihood, 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

$$\sum_{i=1}^{n} X_i^2/n \xrightarrow{\mathrm{P}_\theta} E_\theta[X_1^2] = \mathrm{Var}_\theta\,[X_1] + (E_\theta[X_1])^2 = \theta^2 + \theta^2 = 2\theta^2$$

and

$$\sum_{i=1}^{n} X_i/n \xrightarrow{\mathrm{P}_\theta} E_\theta[X_1] = \theta.$$

Combining these convergences with the continuity theorem it follows that

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

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

Exercise 3.4.2. Let $X_1, X_2, \ldots, X_n$ 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

$$\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \approx \log(\alpha) - \frac{1}{2\alpha} \tag{3.4.1}$$

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 it 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

$$\begin{aligned}\partial_\alpha\mathcal{L}(\theta;\mathbf{X}) &= n\log(\lambda) - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^n \log(X_i) = 0\\ \partial_\lambda\mathcal{L}(\theta;\mathbf{X}) &= n\frac{\alpha}{\lambda} - \sum_{i=1}^n X_i = 0\end{aligned} \tag{3.4.2}$$

The second equation yields that

$$\frac{\alpha}{\lambda} = \overline{X}_n \tag{3.4.3}$$

Combining the first equation in (3.4.2) with (3.4.1) it follows that

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

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* (3.4.3) leads to

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

and then

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

This expression and (3.4.3) yield that

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

□

Exercise 3.4.3. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the *Poisson* (λ) distribution. Find the maximum likelihood estimator of $p(0) + p(1)$.

Solution. The interesting function must be expressed in terms of the para meter λ. 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 λ by λ_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

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

Observe that $\mathcal{L}(\lambda; \mathbf{X}) \to -\infty$ as $\lambda \to 0$ or $\lambda \to \infty$, 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 $\lambda^* = \overline{X}_n$. Thus, $\hat{\lambda}_n = \overline{X}_n$, and then

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

Notice now the the strong law of large numbers yields that $\hat{\lambda}_n \xrightarrow{\mathrm{P}_\lambda} \lambda$; since that function $g(\lambda)$ is continuous, an application of the continuity theorem yields that $\hat{g}_n = g(\hat{\lambda}_n) \xrightarrow{\mathrm{P}_\lambda} g(\lambda)$, that is, the sequence $\{\hat{g}_n\}$ is consistent. □

Exercise 3.4.4. Let $X_1, X_2, \ldots, X_n$ be a random sample of size m from a $\mathcal{N}\left(\mu, \sigma_1^2\right)$ distribution and, independently, let $Y_1, Y_2, \ldots, Y_n$ be a random sample of size n from the $\mathcal{N}\left(\mu, \sigma_2^2\right)$ 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 it logarithm is given by

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

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:

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

$$\partial_{\sigma_1} \mathcal{L}(\theta; \mathbf{X}, \mathbf{Y}) = -\frac{m}{\sigma_1} + \sum_{i=1}^{m} \frac{(X_i - \mu)^2}{\sigma_1^3}$$

$$\partial_{\sigma_2} \mathcal{L}(\theta; \mathbf{X}, \mathbf{Y}) = -\frac{n}{\sigma_2} + \sum_{j=1}^{n} \frac{(Y_j - \mu)^2}{\sigma_2^3}$$

The first equation yields that

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

that is,

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

whereas the last two likelihood equations are equivalent to

$$\sigma_1^2 = \frac{1}{m}\sum_{i=1}^{m}(X_i - \mu)^2 = \tilde{S}_{X\,m}^2 + (\overline{X}_m - \mu)^2$$
$$\sigma_2^2 = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \mu)^2 = \tilde{S}_{Y\,n}^2 + (\overline{Y}_n - \mu)^2$$

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

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

a cubic equation in μ □

.

Exercise 3.4.5. Let $X_1, X_2, \ldots, X_n$ 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 θ. Is this estimator unbiased? Consistent?

Solution. The likelihood function is given by

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

Observing that

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

it follows that

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

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

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

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

$$P_\theta[\hat{\theta}_n < \theta] = 1; \tag{3.4.4}$$

this structural property implies that $E_\theta[\hat{\theta}_n] < \theta$, and then $\hat{\theta}_n$ is a biased estimator of θ, 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,

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

Since (3.4.4) implies that $P_\theta[\hat{\theta}_n \geq \theta + \varepsilon] = 0$, it follows that

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

that is, $\hat{\theta}_n \xrightarrow{\mathrm{P}_\theta} \theta$, 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 $\alpha(\theta,\varepsilon)^n\to$, this implies that $\limsup_{n\to\infty}|b_{\hat\theta_n}(\theta)|\leq\varepsilon$; hence, since $\varepsilon>0$ is arbitrary, $\lim_{n\to\infty}b_{\hat\theta_n}(\theta)=0$, that is, $\{\hat\theta_n\}$ is asymptotically unbiased. □

Remark 3.4.1. The above analysis of the unbiasedness property for θ_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,

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

The expectation of $\hat\theta_n$ can be now computed explicitly as follows:

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

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

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

the bounded convergence theorem implies that

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

which implies that $\{\hat\theta_n\}$ is asymptotically unbiased. □

3.5. Further Examples

In some cases, the determination of a maximum likelihood estimator does not have simple expressions, and examples of this and other difficulties of the method are illustrated in this section.

Exercise 3.5.1. 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 θ.

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

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

□

Exercise 3.5.2. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the *Poisson* (λ) 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 λ, then $\hat{g}_n = g(\hat{\lambda}_n)$. To find $\hat{\lambda}_n$, notice that, given a sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ 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

$$\mathcal{L}(\lambda; \mathbf{X}) = \log(\lambda) \sum_{i=1}^{n} X_i - n\lambda + \log\left(\prod_{i=1}^{n} \frac{1}{X_i!}\right), \quad \lambda \in [0, \infty). \tag{3.5.1}$$

(i) Suppose that $X_i > 0$ for some i. In this case, the basic properties of the logarithmic function yield that $\mathcal{L}(\lambda; \mathbf{X}) \to -\infty$ as $\lambda \to 0$ or as $\lambda \to \infty$. 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 $\lambda = \overline{X}_n = \sum_{i=1}^{n} X_i/n$. Hence,

$$\hat{\lambda}_n = \overline{X}_n. \tag{3.5.2}$$

(ii) Suppose now that $X_i = 0$ for all i. In this context, (3.5.1) shows that the likelihood function reduces to $\mathcal{L}(\lambda; \mathbf{X}) = -n\lambda$, and then its maximizer is $\hat{\lambda}_n = 0 = \overline{X}_n$. Thus, *in any circumstance*, the maximum likelihood estimator of λ is the sample mean, and for $g(\lambda) = e^{-\lambda}$,

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

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

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

(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 λ) which is a function of $\overline{X}_n$. Let $G(\overline{X}_n)$ be such that

$$E_\lambda[G(\overline{X}_n)] = e^{-\lambda} \quad \text{for every } \lambda \in [0, \infty). \tag{3.5.3}$$

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

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

and then

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

where the classical expansion $e^a = \sum_{k=0}^{\infty} a^k/k!$ was used in the last step. Therefore (3.5.3) 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

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

that is,

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

Consequently,

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

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

Exercise 3.5.3. Let $\mathbf{X} = ((X_{1\,1}, X_{2\,1}), (X_{1\,2}, X_{2\,2}), \ldots, (X_{1\,n}, X_{2\,n}))$ be a random sample of size n from the bivariate normal distribution with means μ_1, μ_2, variances σ_1^2 and σ_2^2 and correlation coefficient ρ. Suppose that $\rho \in (-1, 1)$ is unknown and find the maximum likelihood estimator of ρ if

(a) μ_1, μ_2, and σ_1^2 and σ_2^2 are also *unknown*.

(b) μ_1, μ_2, and σ_1^2 and σ_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

$$L(\rho; \mathbf{X}) = e^{-Q/[2(1-\rho^2)]} \prod_{i=1}^{n} \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}} \tag{3.5.4}$$

where the quadratic form Q is given by

$$Q = \sum_{i=1}^{n} \left[\left(\frac{X_{1\,i} - \mu_1}{\sigma_1}\right)^2 + \left(\frac{X_{2\,i} - \mu_2)}{\sigma_2}\right)^2 - 2\rho \left(\frac{X_{1\,i} - \mu_1)}{\sigma_1}\right) \left(\frac{X_{2\,i} - \mu_2}{\sigma_2}\right) \right]. \tag{3.5.5}$$

The logarithm of the likelihood function is

$$\mathcal{L}(\theta; \mathbf{X}) = -\frac{Q}{2(1-\rho^2)} - n\log(\sigma_1) - n\log(\sigma_2) - \frac{n}{2}\log(1-\rho^2) - n\log(2\pi), \tag{3.5.6}$$

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

(i) First, given σ_1, σ_2 and ρ, the maximizer of $\mathcal{L}(\cdot; \mathbf{X})$ with respect to μ_1 and μ_2 will be determined.. Notice that (3.5.5) and (3.5.6) together yield that

$\mathcal{L}(\cdot;\mathbf{X})$ is a concave quadratic form in (μ_1,μ_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

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

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

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

Thus,

$$\mathcal{L}(\theta;\mathbf{X}) \le \mathcal{L}(\overline{X}_{1\,n}, \overline{X}_{2\,n}, \sigma_1, \sigma_2, \rho), \quad \theta = (\mu_1,\mu_2,\sigma_1,\sigma_2,\rho) \in \Theta. \tag{3.5.7}$$

(ii) Next, the function $\mathcal{L}(\overline{X}_{1\,n}, \overline{X}_{2\,n}, \sigma_1, \sigma_2, \rho)$ will be maximized with respect to σ_1, σ_2 and ρ. To achieve this goal, notice that

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

where the quadratic form $\tilde{Q}$ is given by

$$\begin{aligned} \tilde{Q} = \sum_{i=1}^n \Bigg[&\left(\frac{X_{1\,i}-\overline{X}_{1\,n}}{\sigma_1}\right)^2 + \left(\frac{X_{2\,i}-\overline{X}_{2\,n})}{\sigma_2}\right)^2 \\ &-2\rho\left(\frac{X_{1\,i}-\overline{X}_{1\,n})}{\sigma_1}\right)\left(\frac{X_{2\,i}-\overline{X}_{2\,n}}{\sigma_2}\right)\Bigg] \end{aligned} \tag{3.5.8}$$

From this expressions, it follows that, as σ_1 or σ_2 goes to 0 or ∞ or $\rho \to \pm 1$, the function $\mathcal{L}(\overline{X}_{1\,n}, \overline{X}_{2\,n}, \sigma_1, \sigma_2, \rho)$ converges to $-\infty$, and then the mapping $\mathcal{L}(\overline{X}_{1\,n}, \overline{X}_{2\,n}, \cdot, \cdot, \cdot)$ attains its maximum at some point

$$(\sigma_1, \sigma_2, \rho) \in (0,\infty) \times (0,\infty) \times (0,1),$$

which satisfies

$$
\begin{aligned}
&\partial_{\sigma_1}\mathcal{L}(\overline{X}_{1\,n},\overline{X}_{2\,n},\sigma_1,\sigma_2,\rho)\\
&\quad= \frac{-1}{2(1-\rho^2)}\left[\frac{-2}{\sigma_1}\sum_{i=1}^{n}\left(\frac{X_{1\,i}-\overline{X}_{1\,n}}{\sigma_1}\right)^2\right.\\
&\qquad\qquad\left.+\frac{2\rho}{\sigma_1}\sum_{i=1}^{n}\left(\frac{X_{1\,i}-\overline{X}_{1\,n}}{\sigma_1}\right)\left(\frac{X_{2\,i}-\overline{X}_{2\,n}}{\sigma_1}\right)\right]-\frac{n}{\sigma_1}=0\\
&\partial_{\sigma_2}\mathcal{L}(\overline{X}_{1\,n},\overline{X}_{2\,n},\sigma_1,\sigma_2,\rho)\\
&\quad= \frac{-1}{2(1-\rho^2)}\left[\frac{-2}{\sigma_2}\sum_{i=1}^{n}\left(\frac{X_{2\,i}-\overline{X}_{2\,n}}{\sigma_2}\right)^2\right.\\
&\qquad\qquad\left.+\frac{2\rho}{\sigma_2}\sum_{i=1}^{n}\left(\frac{X_{1\,i}-\overline{X}_{1\,n}}{\sigma_1}\right)\left(\frac{X_{2\,i}-\overline{X}_{2\,n}}{\sigma_2}\right)\right]-\frac{n}{\sigma_2}=0\\
&\partial_{\rho}\mathcal{L}(\overline{X}_{1\,n},\overline{X}_{2\,n},\sigma_1,\sigma_2,\rho)\\
&\qquad= \frac{-\rho}{(1-\rho^2)^2}\tilde{Q}-\frac{1}{2(1-\rho^2)}\partial_\rho\tilde{Q}+\frac{n\rho}{1-\rho^2}=0
\end{aligned}
\tag{3.5.9}
$$

The first equation immediately yields that

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

and then, multiplying both sides by σ_1/n,

$$\frac{S_1^2}{\sigma_1^2}-\rho\frac{S_{1\,2}}{\sigma_1\sigma_2}=1-\rho^2 \tag{3.5.10}$$

where

$$
\begin{aligned}
S_1^2 &= \sum_{i=1}^{n}\left(X_{1\,i}-\overline{X}_{1\,n}\right)^2/n\\
S_2^2 &= \sum_{i=1}^{n}\left(X_{2\,i}-\overline{X}_{2\,n}\right)^2/n\\
S_{1\,2} &= \sum_{i=1}^{n}\left(X_{1\,i}-\overline{X}_{1\,n}\right)\left(X_{2\,i}-\overline{X}_{2\,n}\right)/n
\end{aligned}
\tag{3.5.11}
$$

Similarly, from the second equation in (3.5.10) it follows that

$$\frac{S_2^2}{\sigma_2^2}-\rho\frac{S_{1\,2}}{\sigma_1\sigma_2}=1-\rho^2 \tag{3.5.12}$$

Combining the specification of $\tilde{Q}$ in (3.5.8) with (3.5.11) it follows that

$$\tilde{Q} = n\left[\frac{S_1^2}{\sigma_1^2} - \rho\frac{S_{1\,2}}{\sigma_1\sigma_1} + \frac{S_2^2}{\sigma_1^2} - \rho\frac{S_{1\,2}}{\sigma_1\sigma_1}\right],$$

and then, at the solution of the system (3.5.9), equalities (3.5.10) and (3.5.12) yield that

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

combining this relation with the third equation in (3.5.9), 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 $\partial_\rho\tilde{Q} = -2nS_{1\,2}/(\sigma_1\sigma_2)$ (see (3.5.8) and (3.5.11)), it follows that

$$\rho = \frac{S_{1\,2}}{\sigma_1\sigma_2}. \tag{3.5.13}$$

Together with (3.5.10) 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, (3.5.12) and (3.5.13) together yield that

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

and then (3.5.13) becomes

$$\rho = \frac{S_{1\,2}}{S_1S_2}$$

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

$$\begin{aligned}&\mathcal{L}(\overline{X}_{1\,n}, \overline{X}_{2\,n}, \sigma_1, \sigma_2, \rho)\\ &\quad \le \mathcal{L}(\overline{X}_{1\,n}, \overline{X}_{2\,n}, S_1, S_2, S_{1\,2}/[S_1S_2]), \quad \sigma_1, \sigma_2 > 0, \quad \rho \in (-1, 1),\end{aligned}$$

and combining this inequality with (3.5.7), it follows that

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

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

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

in particular, the maximum likelihood estimator of the population correlation coefficient ρ is the sample correlation coefficient $S_{1\,2}/[S_1S_2]$.

(b) When μ_1, μ_2 and σ_1 and σ_2 are known, the likelihood function is given by (3.5.4), where Q is specified by (3.5.5), that is,

$$L(\rho;\mathbf{X}) = e^{-Q/[2(1-\rho^2)]} \prod_{i=1}^{n} \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}} \tag{3.5.14}$$

and the corresponding logarithm is

$$\mathcal{L}(\rho;\mathbf{X}) = -\frac{Q}{2(1-\rho^2)} - n\log(\sigma_1) - n\log(\sigma_2) - \frac{n}{2}\log(1-\rho^2) - n\log(2\pi), \tag{3.5.15}$$

where, writing

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

Q is given by

$$Q = n[\tilde{S}_1^2 + \tilde{S}_2^2 - 2\rho\tilde{S}_{1\,2}]$$

The value of ρ 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

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

that is,

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

equality that is equivalent to

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

this cubic equation should be solved numerically. □

Remark 3.5.1. (i) At first, sight, part (b) seemed to be easier than part (a), since in part (b) only ρ 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. □

Exercise 3.5.4. Let $X_1, X_2, \ldots, X_m$ be a random sample of size m from the $\mathcal{N}\left(\mu, \sigma^2\right)$ distribution and, independently, let $Y_1, Y_2, \ldots, Y_n$ be a random sample of size n from $\mathcal{N}\left(\mu, \lambda\sigma^2\right)$ where $\lambda > 0$ is unknown.

(a) If μ and σ are known, find the maximum likelihood estimator of λ.

(b) If μ and σ and λ are unknown, find the maximum likelihood estimator of $\theta = (\mu, \sigma, \lambda)$.

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

$$Y_1, \ldots, Y_n \text{ are i.i.d. } \mathcal{N}\left(\mu, \lambda\sigma^2\right) \text{ random variables, } \lambda \in (0, \infty). \qquad (3.5.16)$$

Since λ is an arbitrary real number, setting

$$\sigma_1^2 = \lambda\sigma^2 \tag{3.5.17}$$

the statistical model (3.5.16) is equivalent to

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

For this model, the maximum likelihood estimator of σ_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 λ 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

(i) $Y_1, \ldots, Y_n$ are i.i.d. $\mathcal{N}\left(\mu, \lambda\sigma^2\right)$ random variables,

(ii) $X_1, \ldots, X_m$ are i.i.d. $\mathcal{N}\left(\mu, \sigma^2\right)$ random variables,

(iii) The vectors $(X_1, \ldots, X_m)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ are independent, and

(iv) $\mu \in \mathbb{R}, \quad \sigma \in (0, \infty), \quad \lambda \in (0, \infty)$.

Defining $\sigma_1 > 0$ by

$$\sigma_1^2 = \lambda\sigma^2, \tag{3.5.18}$$

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) $Y_1, \ldots, Y_n$ are i.i.d. $\mathcal{N}\left(\mu, \sigma_1^2\right)$ random variables,

(ii) $X_1, \ldots, X_m$ are i.i.d. $\mathcal{N}\left(\mu, \sigma^2\right)$ random variables,

(iii) The vectors $(X_1, \ldots, X_m)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ are independent, and

(iv) $\mu \in \mathbb{R}, \quad \sigma \in (0, \infty), \quad \sigma_1 \in (0, \infty)$.

This model was studied in Exercise 3.4.4, 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 the 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, (3.5.18) and the invariance property together yield that

$$\hat{\lambda} = \frac{\sigma_1^2}{\sigma^2}$$

is the maximum likelihood estimator of λ. □

Exercise 3.5.5. Let $(X_1, X_2, \ldots, X_k)$ be a random vector with multinomial distribution with parameter $p = (p_1, p_2, \ldots, p_k)$ and n trials, where n is known and the probabilities p_1 are unknown numbers is $[0, 1]$ satisfying $\sum_{i=1}^{k} p_i = 1$. Find the maximum likelihood estimator $\hat{p} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k)$.

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

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

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,

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

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

$$L(\hat{p}; \mathbf{X}) \geq \text{Ł}(p; \mathbf{X}), \quad p \in \mathcal{P}. \tag{3.5.19}$$

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

$$D = \{i \in \{1, 2, \ldots, n\} \mid X_i \neq 0\}, \tag{3.5.20}$$

so that

$$L(p;\mathbf{X}) = C\prod_{j\in D} p_j^{X_j}, \tag{3.5.21}$$

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

(a) $L(\hat{p};\mathbf{X}) > 0$. Indeed, the k-dimensional vector $u = (1/k, 1/k, \ldots, 1/k) \in \mathcal{P}$ satisfies $L(u;\mathbf{X}) = C/k^n > 0$, and then (3.5.19) implies that Ł$(\hat{p};\mathbf{X}) > 0$.

(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 (3.5.21) 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 (3.5.21) 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 (3.5.21), and then $L(\hat{p};\mathbf{X}) = 0$, 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 (3.5.21) 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 form 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 (3.5.19) 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 $i = 1, 2, \ldots, k$,

$$\hat{p}_i = \frac{X_i}{n}. \tag{3.5.22}$$

To establish this assertion, first notice that if $i \notin D$, then $X_i = 0$, by (3.5.20), 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 (3.5.22) 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 (3.5.20)) and $\sum_{r=1}^{k} X_r = n$. Consequently, (3.5.22) 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 (3.5.22) also occurs when $i \in D$. In short, for every $i = 1, 2, \ldots, k$, the maximum likelihood estimator of p_i is $\hat{p}_i = X_i/n$. □

Exercise 3.5.6. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ 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 θ and verify that $\{\hat{\theta}_n\}$ is consistent.

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

(c) Find the estimate $\hat{g}_5$ corresponding to $\mathbf{x} = (2.9, 1.48, 5.62, 4.0, 1.22)$, so that $n = 5$.

Solution. (a) Given $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ 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

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

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 $\varepsilon > 0$,

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

Since $P_\theta[\hat{\theta}_n < \theta] = 0$, it follows that $P_\theta[|\hat{\theta}_n - \theta| > \varepsilon] = P_\theta[\hat{\theta}_n > \theta + \varepsilon] \to 0$ as $n \to \infty$, 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

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

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

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

□

Exercise 3.5.7. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample from a *Geometric* (p) distribution, where $p \in [0, 1]$, so that the common probability function of the X_is is

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

(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, we find how many people we need 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 $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ 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_is 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 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,

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

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

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

Chapter 4

Method of Moments

This chapter presents other procedure to build estimators of parametric functions, namely, *the method of moments.* In contrast with the maximum likelihood technique, when applicable the method of moments always produces explicit formulas for the estimators, and can be roughly described as follows: A population moment is estimated by the corresponding sample moment, and a parametric function that is a function of the population moments, is estimated by the same function *evaluated at the sample moments.* Under mild conditions, the generated estimators are consitent, and, as it will be proved later, are asymptotically normal.

4.1. Estimation Using Sample Moments

This section introduces the method of moments to produce estimators of parametric functions. Consider a random variable X whose distribution depends on an unknown parameter θ,

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

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

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

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

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

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

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

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

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

so that the sequence $\{m'_{k\,n}\}_{n=1,2,3,\ldots}$ 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 $X_1, X_2, \ldots, X_n$ as in (4.1.2), then

(i) The kth 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 $\mu'_1(\theta), \mu'_2(\theta), \ldots, \mu'_r(\theta)$, say

$$g(\theta) = G(\mu'_1(\theta), \mu'_2(\theta), \ldots, \mu'_r(\theta)), \tag{4.1.5}$$

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

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

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 of the corresponding sample moments.

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.

4.2. Consistency

The method of moments produces estimators that, under very mild conditions, are consistent. Such a result is precisely stated in the following theorem.

Theorem 4.2.1. Suppose that the function $G(z_1, z_2, \ldots, z_r)$ is continuous at each point $(\mu'_1(\theta), \mu'_2(\theta), \ldots, \mu'_r(\theta))$, $\theta \in \Theta$. 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).

A detailed proof of this result can be seen in TESIS1. Before proceeding to present some examples on the method of moments, it is convenient to summarize the precedent discussion: Given a sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ of a population P_θ, 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 $G(\mu'_1(\theta), \mu'_2(\theta), \ldots, \mu'_k(\theta))$ is constructed evaluating the same function at the sample moments $m'_{1\,n}, m'_{2,n}, \ldots, m'_{k\,n}$.

(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.*

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. On the other hand, it should be mentioned that, when the parametric function $g(\theta)$ being estimated depends in a ‘smooth’ manner of the population moments, the sequence $\{\hat{g}_n\}$

of moments estimators have a normal limit distribution as the sample size n increases. This property will be established in the following chapter and represents the most important result presented in this work.

4.3. The Values of Moments Estimators

In this section a first illustration of method of moments will be presented, and the analysis will be used to show that in general, the method generates estimators whose values do not necessarily belong to the parameter space.

Exercise 4.3.1. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the *Beta*(α, β) distribution, where $\theta = (\alpha, \beta) \in (0, \infty) \times (0, \infty)$. Determine the moment estimators of α and β.

Solution. If $X \sim$ *Beta*(α, β), the first two moments of X are

$$\mu_1' = E_\theta[X] = \frac{\alpha}{\alpha+\beta}, \qquad \mu_2' = \frac{\alpha\beta}{(\alpha+\beta)^2(1+\alpha+\beta)}.$$

Now, the parameters α and β will be expressed in terms of μ_1' and μ_2'. Notice that

$$\mu_2' = \frac{\mu_1'(1-\mu_1')}{1+\alpha+\beta}, \quad \text{and then} \quad \alpha+\beta = \frac{\mu_1'(1-\mu_1')}{\mu_2'} - 1.$$

Since $\alpha = \mu_1'(\alpha+\beta)$, it follows that

$$\alpha = \mu_1' \left(\frac{\mu_1'(1-\mu_1')}{\mu_2'} - 1 \right)$$

On the other hand, notice that $1-\mu_1' = 1-E_\theta[X] = 1-\alpha/(\alpha+\beta) = \beta/(\alpha+\beta)$, so that

$$\beta = (1-\mu_1')(\alpha+\beta) = (1-\mu_1') \left(\frac{\mu_1'(1-\mu_1')}{\mu_2'} - 1 \right)$$

From these two last displays, it follows that the moments estimators of α and β based on a sample of size n are given by

$$\begin{aligned} \hat{\alpha}_n &= m_{1\,n}' \left(\frac{m_{1\,n}'(1-m_{1\,n}')}{m_{2\,n}'} - 1 \right) \\ \hat{\beta}_n &= (1-m_{1\,n}') \left(\frac{m_{1\,n}'(1-m_{1\,n}')}{m_{2\,n}'} - 1 \right), \end{aligned}$$

concluding the argument. □

Remark 4.3.1. Observe that $\hat{\alpha}_n$ and $\hat{\beta}_n$ contain the factor

$$\left(\frac{m'_{1\,n}(1-m'_{1\,n})}{m'_{2\,n}}-1\right)=\left(\frac{\overline{X}_n(1-\overline{X}_n)}{\sum_{i=1}^n X_i^2/n}-1\right). \tag{4.3.1}$$

and it will be shown that this factor may be negative for some samples. Consider the sample

$$\mathbf{X}=\mathbf{x}=(\varepsilon,\varepsilon,\ldots,\varepsilon,1-\varepsilon) \tag{4.3.2}$$

of size n and notice that

$$\overline{X}_n=[(n-1)\varepsilon+1-\varepsilon]/n \quad \text{and} \quad \sum_{i=1}^n X_i^2/n=[(n-1)\varepsilon^2+(1-\varepsilon)^2]/n.$$

so that

$$\lim_{n\to\infty}\overline{X}_n=\frac{1}{n} \quad \text{and} \quad \lim_{n\to\infty}\sum_{i=1}^n X_i^2/n=\frac{1}{n}. \tag{4.3.3}$$

On the other hand,

$$\begin{aligned}\left(\frac{\overline{X}_n(1-\overline{X}_n)}{\sum_{i=1}^n X_i^2/n}-1\right)\geq 0 &\iff \overline{X}_n(1-\overline{X}_n)\geq\sum_{i=1}^n X_i^2/n\\ &\iff \overline{X}_n(1-\overline{X}_n)\geq\sum_{i=1}^n X_i^2/n\end{aligned}$$

Suppose now that, for the sample (4.3.2), the factor () is nonnegative, so that the last inequality in the previous display holds; taking the limit as ε goes to 0, it follows that

$$\lim_{\varepsilon\searrow 0}\overline{X}_n(1-\overline{X}_n)\geq\lim_{\varepsilon\searrow 0}\sum_{i=1}^n X_i^2/n,$$

a relation that, *via* (4.3.3), is equivalent to $(1/n)[1-1/n]\geq 1/n$, which in turn yields that $1-1/n\geq 1$, which is a contradiction. It follows that

$$\lim_{\varepsilon\searrow 0}\overline{X}_n(1-\overline{X}_n)<\lim_{\varepsilon\searrow 0}\sum_{i=1}^n X_i^2/n,$$

and then $\overline{X}_n(1-\overline{X}_n)<\sum_{i=1}^n X_i^2/n$ when $\varepsilon>0$ is small enough, a fact the implies that, with positive probability, the factor in (4.3.1) is negative, and then the estimators $\hat{\alpha}_n$ and $\hat{\beta}_n$ are negative with positive probability. This discussion shows explicitly that the estimators generated by the method of moments do not necessarily belong to the parameter space. □

4.4. Additional Examples

In this section the method of moments will be applied to estimate parametric functions in familiar models. The first one analyzes a normal model with unitary coefficient of variation.

Exercise 4.4.1. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a $\mathcal{N}\left(\theta, \theta^2\right)$ distribution for some $\theta \in \Theta = (0, \infty)$. Find an estimator of θ^2 using the method of moments.

Solution. Let $X \sim \mathcal{N}\left(\theta, \theta^2\right)$ and notice that the first population moment is $\mu_1'(\theta) = E_\theta[X] = \theta$, Thus, a method of moments estimator of θ is given by $\hat{\theta}_n = {m'_1}_n = \overline{X}_n$. 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 θ; see Exercise 3.4.1. □

Exercise 4.4.2. Let $X_1, X_2, \ldots, X_n$ 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 α and λ, 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 $Y = (X - \alpha)/\lambda \sim$ *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

$$\mu_1'(\theta) = E_\theta\left[X\right] = \alpha + \lambda \tag{4.4.1}$$

whereas the second part implies that

$$E_\theta\left[(X-\alpha)^2\right] = 2\lambda^2,$$

so that

$$E_\theta\left[X^2 - 2X\alpha + \alpha^2\right] = 2\lambda^2,$$

a relation that leads to

$$\begin{aligned}\mu_2'(\theta) = E_\theta[X^2] &= 2\lambda^2 - \alpha^2 + 2E_\theta[X]\alpha \\ &= 2\lambda^2 - \alpha^2 + 2(\lambda+\alpha)\alpha \\ &= 2\lambda^2 + 2\alpha\lambda + \alpha^2 \qquad (4.4.2) \\ &= 2\lambda(\lambda+\alpha) + \alpha^2 \\ &= 2\lambda\mu_1'(\theta) + \alpha^2\end{aligned}$$

Using that $\lambda = \mu_1'(\theta) - \alpha$, by (4.4.1), it follows that

$$\begin{aligned}\mu_2'(\theta) &= 2(\mu_1'(\theta)-\alpha)\mu_1'(\theta) + \alpha^2 \\ &= 2\mu_1'(\theta)^2 - 2\mu_1'(\theta)\alpha + \alpha^2 = \mu_1(\theta)^2 + (\mu_1(\theta)-\alpha)^2.\end{aligned}$$

Consequently,

$$\lambda^2 = (\mu_1(\theta)-\alpha)^2 = \mu_2'(\theta) - \mu_1'(\theta)^2,$$

where the first equality is due to (4.4.1), and it is useful to observe that the relation $\mu_2'(\theta) - \mu_1'(\theta)^2 \geq 0$ holds, by Jensen's inequality. 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:

$$\begin{aligned}\hat\lambda_n &= \sqrt{m_{2\,n}' - (m_{1\,n}')^2} \\ \hat\alpha_n &= m_{1\,n}' - \sqrt{m_{2\,n}' - (m_{1\,n}')^2}.\end{aligned}$$

Since λ and α are continuous functions of μ_1' and μ_2', it follows that these estimators are consistent, and since they are not linear functions of μ_1' and μ_2', they are not unbiased. Before concluding, it is interesting to observe that $m_2' - (m_1')^2 = \sum_{i=1}^n X_i^2/n - \overline{X}_n^2 = \sum_{i=1}^n (X_i - \overline{X}_n)^2/n$ 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 $\hat{\alpha}_n = \overline{X}_n - \tilde{S}_n$. □

Exercise 4.4.3. Let $f_1(x)$ and $f_2(x)$ be two densities with means μ_1 and μ_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 θ 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 μ_1 and μ_2, the expectations of the densities f_1 and f_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 $X_1, X_2, \ldots, X_n$ of the density $f(x; \theta)$ is available, the method of moments prescribes the estimator

$$\hat{\theta}_n = \frac{m'_{1\,n} - \mu_2}{\mu_1 - \mu_2} = \frac{\overline{X}_n - \mu_2}{\mu_1 - \mu_2},$$

which is unbiased. □

Chapter 5

Limit Behavior of Moments Estimators

The objective of this chapter is to study the limit distribution of the sequence $\{\hat{g}_n\}$ of moments estimators of a parametric function $g(\theta)$. The main result in this direction establishes that, after multiplying the difference $[\hat{g}_n - g(\theta)]$ by the square root of the sample size n, the resulting random quantity converges in distribution to a normal variable. This result is obtained using the central limit theorem together with an invariance result on convergence to normality, which can be described as follows: if a sequence of random vectors $\{W_n\}$ $\sqrt{n}[W_n - \mu]$ converges in distribution to normality, then a transformed sequence $\{G(W_n)\}$ has a similar behavior whenever the function G is continuously differentiable.

5.1. Asymptotic Normality

In this section the basic notion on the limit behavior of a sequence of random variables (or vectors) is introduced. The formal statement of this idea involves the concept of *convergence in distribution* which is carefully analyzed, for instance, in Deudewicz y Mishra (1988), Mood *et al.* (1984), or Wackerly *et al.* (2009). Briefly, if $F(\cdot)$ is a distribution function in $\mathbb{R}^k$, a sequence

$\{W_n\}$ of k-dimensional random vectors converges in distribution to F if

$$\lim_{n\to\infty} P[W_n \leq \mathbf{x}] = F(\mathbf{x})$$

at each point point $\mathbf{x}$ at which F is continuous, a property that is indicated by writing

$$W_n \xrightarrow{\mathrm{d}} F.$$

A common instance of convergence in distribution corresponds to the case in which F is the distribution function of a normal probability measure with mean μ and covariance matrix M, and in the notation

$$W_n \xrightarrow{\mathrm{d}} \mathcal{N}(\mu, M)$$

is used.

Definition 5.1.1. Consider a parametric function $g: \Theta \to \mathbb{R}^d$ defined on the parameter space Θ and taking values in $\mathbb{R}^d$ and, for each positive integer n, let $\hat{g}_n$ be an estimator of $g(\theta)$ based on the first n observations $X_1, X_2, \ldots, X_n$. In this case, the sequence $\{\hat{g}_n\}$ of estimators is *consistent and asymptotically normal* if, and only,

$$\sqrt{n}\ [\hat{g}_n - g(\theta)] \xrightarrow{\mathrm{d}} \mathcal{N}\,(0, J(\theta)^2);$$

where $J(\theta)^2$ is square nonnegative matrix of order $d \times d$. In this case, $J(\theta)^2$ is referred to as the asymptotic variance of $\sqrt{n}\ [\hat{g}_n - g(\theta)]$.

Remark 5.1.1. The matrix $J(\theta)^2$ in the above definition is also referred to as the (asymptotic) *information matrix of the sequence* $\{\hat{g}_n\}$ since, when the sample size is 'large', $J(\theta)^2$ determines the length of the confidence intervals for linear combinations of the components of $g(\theta)$ that can be obtained from the estimator $\hat{g}_n$. □

5.2. Invariance Principle

As already mentioned, the main objective of this chapter is to show that the moments estimators of a parametric quantity $g(\theta)$ are *consistent and* *asymptotically normal.* The main tool to establish this result are the multivariate central limit theorem, and an invariance property of the convergence to normality, which can be roughly stated as follows: If a sequence of random vectors $\{W_n\}$ converges to a (multivariate) normal distribution, and if g is a smooth function, then under mild conditions the transformed sequence $\{g(W_n)\}$ also converges to a normal distribution. These two fundamental results are formally stated below:

Theorem 5.2.1. [Multivariate Central Limit Theorem.] Consider a random vector $X = (X^{(1)}, X^{(2)}, \ldots, X^{(k)})'$ with mean μ and variance matrix M, that is,

$$\mu = (\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(k)})' = (E[X^{(1)}], E[X^{(2)}], \ldots, E[X^{(k)}])'$$
$$M = [m_{i\,j}] = \text{Cov}\left(X^{(i)}, X^{(j)}\right).$$

Suppose that $X_1, X_2, X_3, \ldots,$ is a sequence of independent and identically distributed random vectors with the same distribution as X. In this case

$$\sqrt{n}\ [\overline{X}_n - \mu] \xrightarrow{\text{d}} \mathcal{N}_k(0, M).$$

The following result shows that convergence to normality is not altered under the application of differentiable transformations.

Theorem 5.2.2. Suppose that $\{W_n\}$ is a sequence of k-dimensional random vectors such that

$$\sqrt{n}\ [W_n - \mu] \xrightarrow{\text{d}} \mathcal{N}_k(0, M)$$

for some nonnegative matrix M of order $k \times k$ and $\mu \in \mathbb{R}^k$. In this case, let g be a function defined on an open set of $\mathbb{R}^k$ containing the vector μ, suppose that g takes value in $\mathbb{R}^d$ and that g is differentiable at μ. In this case

$$\sqrt{n}\ [g(W_n) - g(\mu)] \xrightarrow{\text{d}} \mathcal{N}_d(0, Dg(\mu) M Dg(\mu)'),$$

where $Dg(\mu)$ is the (matrix) derivative of g at μ and has order $d \times k$.

A discussion and proof of these fundamental results can be found, for instance, in Dudewicz y Mishra (1988), Wackerly *et al.* (2007) or Lehmann and Casella (1999). The proof of the central limit theorem involves the idea of characteristic function and the so called 'continuty theorem', that relates the notions of convergence in distribution and uniform convergence of characteristic functions. The invariance principle is derived, essentially, using the idea of derivative of a function as a linear transformation, and applying Slutsky's theorem, as presented in the aforementioned books.

5.3. The Invariance Property in Specific Cases

Before going any further, the application of Theorem 5.2.2 is illustrated in some particular cases.

Example 5.3.1. Suppose that $X_1, X_2, \ldots$ is a sequence of independent and identically distributed random variables with mean μ and variance $\sigma^2 < \infty$. The central limit theorem yields that

$$\sqrt{n}\,[\overline{X}_n - \mu] \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \sigma^2\right). \tag{5.3.1}$$

Now, the asymptotic distribution of some transformations $\{g(\overline{X}_n)\}$ will be obtained by an application of Theorem 5.2.2.

(i) $g(x) = e^x$. In this case, $g(\overline{X}_n) = e^{\overline{X}_n}$, and observing that $Dg(x) = g'(x) = e^x$, it follows that $Dg(\mu) = e^\mu$. Hence, starting from (5.3.1), an application of Theorem 5.2.2 leads to

$$\sqrt{n}\,[e^{\overline{X}_n} - e^\mu] \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, e^\mu \sigma^2 e^\mu\right) = \mathcal{N}\left(0, e^{2\mu}\sigma^2\right)$$

(ii) $g(x) = \sin(x)$. For this function, $g(\overline{X}_n) = \sin(\overline{X}_n)$, and $Dg(x) = g'(x) = \cos(x)$, so that $Dg(\mu) = \cos(\mu)$. Thus, (5.3.1), and Theorem 5.2.2 together imply that

$$\sqrt{n}\,[\sin(\overline{X}_n) - \sin(\mu)) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \cos(\mu)\,\sigma^2 \cos(\mu)\right) = \mathcal{N}\left(0, \cos(\mu)^2\,\sigma^2\right)$$

(iii) Consider now that transformation $g(x) = (e^x, \sin(x))'$. This function transforms $\mathbb{R} = \mathbb{R}^1$ into $\mathbb{R}^2$, and its derivative Dg is the following matrix of order 2×1:

$$Dg(x) = \begin{bmatrix} \dfrac{d}{dx} e^x \\ \dfrac{d}{dx}\sin(x) \end{bmatrix} = \begin{bmatrix} e^x \\ \cos(x) \end{bmatrix}.$$

Therefore,

$$\sqrt{n}\ [g(\overline{X}_n) - g(\mu)) \xrightarrow{\text{d}} \mathcal{N}_2(0, Dg(\mu)\sigma^2 Dg(\mu)');$$

more explicitly,

$$\begin{aligned}\sqrt{n}\ \left[\begin{pmatrix} e^{\overline{X}_n} \\ \sin(\overline{X}_n) \end{pmatrix} - \begin{pmatrix} e^{\mu} \\ \sin(\mu) \end{pmatrix}\right] &\xrightarrow{\text{d}} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} e^{\mu} \\ \cos(\mu) \end{bmatrix} \sigma^2 [e^{\mu} \quad \cos(\mu)]\right) \\ &= \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} e^{2\mu} & e^{\mu}\cos(\mu) \\ e^{\mu}\cos(\mu) & \cos^2(\mu) \end{bmatrix}\right).\end{aligned}$$

□

The next example concerns the estimation of the variation coefficient for a normal population.

Example 5.3.2. Suppose that $X_1, X_2, X_3, \dots$ are independent and identically distributed random variables with $\mathcal{N}\left(\mu, \sigma^2\right)$ distribution.

(i) The asymptotic distribution of the sample standard deviation can be determined as follows: Recall that the sample variance $S_n^2 = \sum_{i=1}^n (X_i - \overline{X}_n)^2/(n-1)$ has the χ^2_{n-1} distribution, and then the central limit theorem yields that

$$\sqrt{n-1}\ [S_n^2 - \sigma^2] \xrightarrow{\text{d}} \mathcal{N}\left(0, 2\sigma^4\right)$$

a statement that, because to the convergence $\sqrt{n}/\sqrt{n-1} \to 1$ as $n \to \infty$, is equivalent to

$$\sqrt{n}\ [S_n^2 - \sigma^2] \xrightarrow{\text{d}} \mathcal{N}\left(0, 2\sigma^4\right).$$

Consider now the function $g(x) = \sqrt{x}$, so that $Dg(x) = g'(x) = 1/[2\sqrt{x}]$. The above convergence and Theorem 5.2.2 together yield that

$$\begin{aligned}\sqrt{n}\ [S_n - \sigma] &= \sqrt{n}\ [g(S_n^2) - g(\sigma^2)] \\ &\xrightarrow{\text{d}} \mathcal{N}\left(0, g'(\sigma^2)\,(2\sigma^4))g'(\sigma^2)\right) = \mathcal{N}\left(0, \sigma^2/2\right).\end{aligned} \tag{5.3.2}$$

(ii) The *variation coefficient*

$$\text{CV} = \frac{\mu}{\sigma}$$

is naturally estimated by

$$\widehat{\text{CV}}_n = \frac{\overline{X}_n}{S_n},$$

which is the maximum likelihood estimator as well as the moments estimator. The present objective is to determine its asymptotic distribution. To achieve this goal, recall the well-known fact that for the normal model $\overline{X}_n$ and S_n are independent random variables; combining this fact with (5.3.2) and the convergence $\sqrt{n}\ [\overline{X}_n - \mu] \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \sigma^2\right)$, it follows that

$$\sqrt{n}\ \left[\begin{pmatrix}\overline{X}_n\\ S_n\end{pmatrix} - \begin{pmatrix}\mu\\ \sigma\end{pmatrix}\right] \xrightarrow{\mathrm{d}} \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma^2 & 0\\ 0 & \sigma^2/2\end{bmatrix}\right) \tag{5.3.3}$$

Next, consider the function transforming a vector in $\mathbb{R}^2$ with no-null second component into the a real number specified as follows:

$$g\begin{pmatrix}x_1\\ x_2\end{pmatrix} = \frac{x_1}{x_2}.$$

The derivative of g is the matrix of order 1×2 given by

$$Dg\begin{pmatrix}x_1\\ x_2\end{pmatrix} = [\partial_{x_1} g, \partial_{x_2} g] = [1/x_2,\ \ -x_1/x_2^2],$$

and it follows that

$$g\begin{pmatrix}\mu\\ \sigma\end{pmatrix} = \frac{\mu}{\sigma} = \mathrm{CV}, \quad g\begin{pmatrix}\overline{X}_n\\ S_n\end{pmatrix} = \frac{\overline{X}_n}{S_n} = \widehat{\mathrm{CV}}_n, \quad Dg\begin{pmatrix}\mu\\ \sigma\end{pmatrix} = [1/\sigma,\ \ -\mu/\sigma^2],$$

and then

$$Dg\begin{pmatrix}\mu\\ \sigma\end{pmatrix}\begin{bmatrix}\sigma^2 & 0\\ 0 & \sigma^2/2\end{bmatrix} Dg\begin{pmatrix}\mu\\ \sigma\end{pmatrix}' = 1 + \frac{\mu^2}{2\sigma^2} = 1 + \frac{\mathrm{CV}^2}{2}$$

Thus, starting with (5.3.3), and application of Theorem 5.2.2 with the function g specified above yields that

$$\sqrt{n}\ \left[\widehat{\mathrm{CV}}_n - \mathrm{CV}\right] = \sqrt{n}\ \left[g\begin{pmatrix}\overline{X}_n\\ S_n\end{pmatrix} - g\begin{pmatrix}\mu\\ \sigma\end{pmatrix}\right] \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, 1 + \frac{\mathrm{CV}^2}{2}\right).$$

□

5.4. Limit Distribution of Moment Estimators

In this section it will be shown that Theorems 5.2.1 and 5.2.2 together imply that, generally, the moments estimators are asymptotically normal.

Theorem 5.4.1. Suppose that X is a random variable whose distribution P_θ depends on an unknown parameter $\theta \in \Theta$, and that the moment of order $2k$ of P_θ is finite for every $\theta \in \Theta$, *i.e.*,

$$E_\theta[X^{2k}] = \mu'_{2k}(\theta) < \infty, \quad \theta \in \Theta.$$

Let the k-dimensional vector $\mu(\theta)$ and the matrix $M(\theta)$ of order $k \times k$ be the mean and covariance of the vector $X, X^2, \ldots X^k)$, that is,

$$\mu(\theta) = (\mu'_1(\theta), \mu'_2(\theta), \ldots, \mu'_k(\theta)' = (E_\theta[X], E_\theta[X^2], \ldots, E_\theta[X^k])' \qquad (5.4.1)$$

and

$$M(\theta) = [M_{i\,j}(\theta)] = [\mathrm{Cov}_\theta(X^i, X^j)] = [\mu'_{i+j}(\theta) - \mu'_i(\theta)\mu'_j(\theta))] \qquad (5.4.2)$$

Consider now a sequence $X_1, X_2, X_3, \ldots,$ of independent and identically distributed random variables from the same distribution as X, and let $m_{r\,n}$ be the sample moment of order r based on the first n observations:

$$m_{r\,n} = \frac{1}{n}\sum_{i=1}^{n} X_i^r. \qquad (5.4.3)$$

Finally, for each positive integer n define the k-dimensional vector W_n by

$$W_n = (m_{1\,n}, m_{2\,n}, \ldots, m_{k\,n})'. \qquad (5.4.4)$$

In this case,

(i) The sequence $\{W_n\}$ is asymptotically normal with mean $\mu(\theta)$ and covariance matrix $M(\theta)$. More explicitly,

$$\sqrt{n}\,[W_n - \mu(\theta)] \xrightarrow{\mathrm{d}} \mathcal{N}_k\left(0, M(\theta)\right).$$

(ii) Let $g\colon \mathcal{O} \to \mathbb{R}^p$ be a continuously differentiable function defined on an open set $\mathcal{O} \subset \mathbb{R}^k$, and suppose that $\mathcal{O}$ contains the vector $\mu(\theta)$ for each $\theta \in \Theta$. Consider the parametric function

$$g(\mu(\theta)) = g(\mu'_1(\theta), \mu'_2(\theta) \ldots, \mu'_k(\theta)),$$

and let $\hat{g}_n$ be the moment estimator of $g(\mu(\theta))$ based on the first n observations, that is,

$$\hat{g}_m = g(m_{1\,n}, m_{2\,n}, \ldots, m_{k\,n}) = g(W_n).$$

In this case, the sequence $\{\hat{g}_n\}$ is asymptotically normal with mean $g(\mu(\theta))$ and covariance matrix

$$M_g(\theta) = D_\mu g(\mu(\theta))M(\theta)D_\mu g(\mu(\theta))'.$$

that is,

$$\sqrt{n}\ [g(W_n) - g(\mu(\theta))] \xrightarrow{\text{d}} \mathcal{N}_p\left(0, M_g(\theta)\right). \tag{5.4.5}$$

Proof. When $X \sim P_\theta$, the mean and covariance matrix of the random vector $\mathbf{X} = (X, X^2, \ldots, X^k)$ are $\mu(\theta)$ and $M(\theta)$, respectively, as specified in (5.4.1) and (5.4.2), and the vectors $\mathbf{X}_i = (X_i, X_i^2, \ldots, X_i^k)$, $i = 1, 2, 3, \ldots$ are independent with the same distribution as $\mathbf{X}$. Moreover, the vector W_n defined in (5.4.4) and (5.4.3) is the sample mean of $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$:

$$W_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i.$$

Thus, the multivariate central limit theorem leads to

$$\sqrt{n}\ [W_n - \mu(\theta)] \xrightarrow{\text{d}} \mathcal{N}_k(0, M(\theta)),$$

and the relation (5.4.5) follows combining this convergence with Theorem 5.2.2. □

Remark 5.4.1. In words, the above theorem establishes that the asymptotic normality is a 'generic' property of moments estimators. Indeed, the interesting parametric functions g that arise in practice and can be expressed in terms of population moments are generally smooth (continuously differentiable) functions of the moments. In this case, if the moments of sufficiently high order of the underlying population are finite, Theorem 5.4.1 ensures the asymptotic normality of the moments estimators $\hat{g}_n$. □

5.5. Arcsine and Risk Ratio Examples

The application of the above theorem is illustrated in the following examples. The first one concerns a transformation that is frequently used in experimental design to study the problem of comparing proportions, whereas the second one refers to the limit distribution of the risk ratio as studied in categorical data analysis.

Example 5.5.1. Let $X_1, X_2, X_3, \ldots$ be independent random variables from a *Bernoulli* (p) population, where the parameter $p \in (0,1)$ is unknown. The first population moment is p, so that the moments estimator of p is $\hat{p}_n = \overline{X}_n$. Since the population variance is $\sigma^2 = p(1-p)$, the central limit theorem yields that

$$\sqrt{n}\ [\overline{X}_n - p] \xrightarrow{\mathrm{d}} \mathcal{N}(0, p(1-p))$$

Consider now the smooth function

$$g(p) = \arcsin(\sqrt{p}),$$

so that

$$D_p g(p) = g'(p) = \frac{d}{dp}\arcsin(\sqrt{p}) = \frac{1}{\sqrt{1-(\sqrt{p})^2}}\frac{1}{2\sqrt{p}} = \frac{1}{2}\frac{1}{\sqrt{1-p}}.$$

An application of Theorem 5.4.1 yields that

$$\begin{aligned}\sqrt{n}\ [\arcsin(\overline{X}_n) - \arcsin(p)] &= \sqrt{n}\ [g(\overline{X}_n) - g(p)] \\ &\xrightarrow{\mathrm{d}} \mathcal{N}(0, Dg(p)p(1-p)Dg(p)) = \mathcal{N}\left(0, \frac{1}{4}\right)\end{aligned}$$

notice that the (asymptotic) variance of the transformed mean, that is, $\arcsin(\overline{X}_n)$, does not depend on the value of p; this stabilizing transformation is frequently used when comparing proportions, since an essential assumption in the analysis of variance is that the standard deviations of the different populations being compared is the same. □

.

Example 5.5.2. Consider two samples $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ of the *Binomial* (p_1) and *Binomial* (p_2) populations, respectively. In health studies, p_i is interpreted as the probability if acquiring some illness and is 'samall', and the ratio

$$r = \frac{p_1}{p_2}$$

is referred as the *risk ratio*. The moments estimator of r based on the two samples of size n is

$$\hat{r}_n = \frac{\overline{X}_n}{\overline{Y}_n},$$

and obtaining an approximation for the distribution of $\hat{r}_n$ for large samples is an interesting and important problem. Notice that

$$\sqrt{n}\;[\overline{X}_n - p_1] \xrightarrow{\mathrm{d}} \mathcal{N}\,(0, p_1(1-p_1))$$

and $\sqrt{n}\;[\overline{Y}_n - p_2] \xrightarrow{\mathrm{d}} \mathcal{N}\,(0, p_2(1-p_2))$, by the central limit theorem, and that the independence of the samples implies that

$$\sqrt{n}\;\left[\begin{bmatrix}\overline{X}_n\\ \overline{Y}_n\end{bmatrix} - \begin{bmatrix}p_1\\ p_2\end{bmatrix}\right] \xrightarrow{\mathrm{d}} \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}p_1(1-p_1) & 0\\ 0 & p_2(1-p_2)\end{bmatrix}\right). \qquad (5.5.1)$$

Now, consider the function

$$g(p_1, p_2) = \log(p_2/p_1) = \log(p_2) - \log(p_1),$$

and notice that

$$Dg(p_1,p_2) = (\partial_{p_1} g(p_1,p_2),\; \partial_{p_2} g(p_1,p_2)) = \left(-\frac{1}{p_1}, \frac{1}{p_2}\right),$$

as well as

$$Dg(p_1,p_2)\begin{bmatrix}p_1(1-p_1) & 0\\ 0 & p_2(1-p_2)\end{bmatrix} Dg(p_1,p_2)' = (1-p_1)/p_1 + (1-p_2)/p_2.$$

and, starting with (5.5.1), an application of Theorem 5.2.2 yields that

$$\begin{aligned}\sqrt{n}\;[\log(\overline{X}_n/\overline{Y}_n) - \log(p_1/p_2)] &= \sqrt{n}\;\left[g\begin{bmatrix}\overline{X}_n\\ \overline{Y}_n\end{bmatrix} - g\begin{bmatrix}p_1\\ p_2\end{bmatrix}\right]\\ &\xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \frac{1-p_1}{p_1} + \frac{1-p_2}{p_2}\right).\end{aligned}$$

□

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.