DISTRIBUCIÓN LÍMITE DE CUANTILES MUESTRALES

MARÍA ELENA BERLANGA ORTIZ

TESIS

Presentada como Requisito Parcial para
Obtener el Grado de:

MAESTRÍA EN
ESTADÍSTICA APLICADA

UNIVERSIDAD AUTÓNOMA AGRARIA
"ANTONIO NARRO"

Buenavista, Saltillo, Coahuila, México
Mayo de 2012

Universidad Autónoma Agraria Antonio Narro
Subdirección de Postgrado

DISTRIBUCIÓN LÍMITE DE CUANTILES MUESTRALES

TESIS

Por:

MARÍA ELENA BERLANGA ORTIZ

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

MAESTRA EN
ESTADÍSTICA APLICADA

Comité Particular

Asesor principal: ______________________
Dr. Rolando Cavazos Cadena

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

Asesor: ______________________
M. C. Luis Rodríguez Gutiérrez

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

COMPENDIO

DISTRIBUCIÓN LÍMITE DE CUANTILES MUESTRALES

Por

MARÍA ELENA BERLANGA ORTIZ

MAESTRÍA EN

ESTADÍSTICA APLICADA

UNIVERSIDAD AUTÓNOMA AGRARIA
ANTONIO NARRO

BUENAVISTA, SALTILLO, COAHUILA, Mayo de 2012

Dr. Rolando Cavazos Cadena –Asesor–

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

El tema de este trabajo es el problema de *estimación puntual* en modelos estadísticos paramétricos. Los principales objetivos que se persiguen son los siguientes: (i) Analizar dos métodos de construcción de estimadores, a saber, la técnica de verosimilitud máxima y el método de momentos, (ii) Proporcionar ilustraciones detalladas y completas sobre ideas básicas en la teoría, como insesgamiento, consistencia y normalidad asintótica, y (iii) Presentar una deducción detallada de la distribución límite de cuantiles muestrales. La principal contribución de este trabajo se ubica en el último de estos objetivos, proporcionando una demostración que combina el teorema central de límite con una relación entre la distribución de un estadístico de orden y variables aleatorias binomiles, e indicando claramente cada una de las etapas técnicas del desarrollo.

ABSTRACT

LIMIT DISTRIBUTION OF SAMPLE
QUANTILES

BY

MARÍA ELENA BERLANGA ORTIZ

MASTER IN

APPLIED STATISTICS

UNIVERSIDAD AUTÓNOMA AGRARIA
ANTONIO NARRO

BUENAVISTA, SALTILLO, COAHUILA, May, 2012

Dr. Rolando Cavazos Cadena –Advisor–

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

This work concerns the problem of *point estimation* in parametric statistical models, and the three main objectives of this exposition are as follows: (i) To analyze two methods of constructing estimators, namely, the maximum likelihood technique and the method of moments, (ii) To provide detailed illustrations of basic notion in the theory, as unbiasedness, consistency and asymptotic normality, and (iii) To give a detailed derivation of the limit distribution of sample quantiles. The main contribution of this note is to present a complete derivation of this last result which, combines the central limit theorem with the relation between the distribution of an order statistic and binomial random variables.

Contents

Chapter 1

A Panoramic View

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

1.1. Introduction

This work is concerned with the problem of *parametric point estimation*, which is pervasive and plays a central role in the theory and applications of statistics. Indeed, point estimation lays in the core of the statistical methodology, and a major step in every analysis is the determination of estimates (*i.e.*, approximations) to some unknown quantities in terms of the observed data, and every treatise on theoretical or applied statistics dedicates a good amount of space to describe methods of constructing estimators and to analyze its properties; see, for instance, 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 estimators of the quantiles of the underlying population.

These topics are briefly described below.

1.2. The Estimation Problem

In general, the purpose of a statistical analysis is to use the observed data *to gain knowledge* about some unknown aspect of the process generating the observations. The observable data $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ 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) $\mathbb{R}^n$:

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

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 $\mathbb{R}^n$

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

The parameter θ^* satisfying this relation for every (Borel) subset of $\mathbb{R}^n$ is *the true parameter value.* Notice that the model prescribes the existence of $\theta^* \in \Theta$ such that the above equality always holds, but does not specify which is the parameter 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 guesse' (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. Objectives and Main Contribution

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

(i) To present a formal description of two important methods to construct estimators, namely, the maximum likelihood technique, and the method of moments;

(ii) To use selected examples to illustrate the construction of estimators in models involving distributions frequently used in applications,

(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 rigorous derivation of the asymptotic distribution of the sequence of sample quantiles of a given order.

The analysis performed below to achieve this last objective represents *the main contribution of this work*. In fact, the theorem on the limit distribution of a sequence of sample quantiles is usually presented without proof in intermediate level texts, and in the present exposition a serious effort has been made to derive such a result in a clear and concise manner, highlighting the essential statistical and analytical tools that are used to establish the theorem, and indicating clearly the basic steps of the argument.

1.4. The Project Behind This Work

This work stems 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 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 in these directions the basic references in the project are Apostol (1980), Fulks (1980), Khuri (2002) and Royden (2003), which concern mathematical analysis, whereas the algebraic aspects are covered in Graybill (2001) and Harville (2008).

The examples presented in the following chapters reflect the work developed in the project, and it is a pleasure to thank to Alfonso Soto Almaguer, by his generosity and clever discussions.

1.5. The Organization

The remainder of this work has been organized as follows:

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

Next, in Chapter 3 the method of maximum likelihood estimation is introduced, which is based on the intuitive idea that, after observing the data, the estimate of the unknown parameter 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, finally, in Chapter 5 the estimation of sample quentiles is studied.

As already mentioned, all of the notions introduced in this work are illustrated by carefully analyzed examples.

Chapter 2

Statistical Point Estimation

In this chapter some basic concepts in the theory of point estimation are introduced. The exposition begins with a brief description of the idea of parametric statistical model, and then the estimation problem of an unknown parametric function is discussed. The presentation continues with the notions of unbiased estimator and consistency of a sequence of estimators, and the related concept of asymptotically unbiased sequence is also analyzed. All of the concepts introduced in this chapter are illustrated using fully solved examples. A detailed presentation of the material in this chapter can be found, for instance, in Dudewicz and Mishra (1998), Mood *et al.* (1988) or Wackerly *et al.* (2009) at a level similar to the one in this work; a more advanced perspective is given in Lehmann and Casella (1998), Borobkov (1999) or Shao (2010).

2.1. Introduction

Given an observable vector $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$, a parametric statistical model for $\mathbf{X}$ prescribes a family $\{P_\theta\}_{\theta\in\Theta}$ of probability distributions for $\mathbf{X}$. The set of indices Θ is referred to as the *parameter space* and is a subset of an Euclidean space $\mathbb{R}^k$. The essence of a statistical model is that the distribution of $\mathbf{X}$ is supposed to be P_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

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

Alternatively, if $\mathbf{X}$ has a density or a probability function $f_{\mathbf{X}}(\mathbf{x};\theta)$, the model can be specified as

$$\mathbf{X} \sim f_{\mathbf{X}}(\mathbf{x};\theta), \quad \theta \in \Theta.$$

On the other hand, 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. Unbiasedness and Consistency

In this section the ideas of bias of an estimator and consistency of a sequence of estimators are introduced. In general, after obtaining the data, the calculated value of the estimator will not coincide with the unknown quantity

$g(\theta)$. The unbiasdness property requires that, if the estimator is computed repeatedly, then the average of the calculated values converge to $g(\theta)$. On the other hand, the notion of consistency concerns the method used to generate the estimations: The idea is that, as the sample size increases, the estimators converge (in probability) to the unknown quantity being estimated.

Definition 2.2.1. An estimator T_n of $g(\theta)$ based on $X_1, X_2, \ldots, X_n$ is *unbiased* if

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

for every $\theta \in \Theta$; notice that 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.

In general, the value attained by an estimator $T_n = T_n(X_1, X_2, \ldots, X_n)$ of $g(\theta)$, does not coincide with the quantity $g(\theta)$. However, if the estimator T_n is unbiased, and the experiment producing the sample $\mathbf{X}$ is repeated, obtaining the estimators $T_{n\,1}, T_{n\,2}, T_{n\,3}, \ldots$ at each trial, it follows from the law of large numbers that the average

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

converges to $g(\theta)$ as the number k of repetitions increases (Loève, 1984, Dudewicz and Mishra, 1998). Thus, on the average, the estimator T_n 'points to the correct quantity' $g(\theta)$.

Remark 2.2.1. It must be noted that not all of the parametric 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 $T_n = T_n(X_1, X_2, \ldots, X_n)$ 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 unbiasedness property may be too restrictive, and it is is possible to have that an unbiased estimator does not exists in some cases of interest. □

Definition 2.2.2. 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; this comment will be illustrated in the examples below.

Definition 2.2.3. A sequence $\{T_n\}_{n=1,2,\ldots}$ of estimators of $g(\theta)$ is *asymptotically unbiased* if

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

Notice that the above property is equivalent to the requirement that, for each parameter $\theta \in \Theta$, $E_\theta[T_n] \to g(\theta)$ as $n \to \infty$. The following notion also concerns the behavior of the whole sequence of estimators $\{T_n\}$ or, equivalently, the method used to generate the estimators.

Definition 2.2.4. 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

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

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

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

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

In this case, if the variables $X_1, X_2, \ldots, X_n, \ldots$ 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, consider a function $G(x_1, x_2, \ldots, x_r)$ which is continuous at each point $(g_1(\theta), \ldots, g_r(\theta))$, where $\theta \in \Theta$. Within this framework, the new 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 $\varepsilon > 0$,

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

so that

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

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

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

2.3. A Binomial Example and Factorial Moments

In this section the construction of unbiased estimators will be illustrated in an example involving the binomial distribution. In this case, and in similar contexts where the specification of the underlying probability function involves the factorial of an integer, the computation of expectation can be eased by using the moment generating function of factorial moments, an idea that will be introduced after the following example.

Exercise 2.3.1. Let the variables $X_1, X_2, \ldots, X_n$ be independent and identically distributed Bernoulli random variables with probability of success p, and set $T_n = X_1 + X_2 + \cdots + X_n$, whereas $\overline{X}_n = T_n/n$ is the sample mean of the sample. Show that

(a) $T_n(T_n - 1)/c_n$ with $c_n = n(n-1)$ is an unbiased estimator of p^2.

(b) $T_n(T_n - 1)(T_n - 2)/d_n$ with $d_n = n(n-1)(n-2)$ is an unbiased estimator of p^3.

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

(d) Find an unbiased estimator of $p - q$ where, as usual, $q = 1 - p$.

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

$$\binom{n}{t} = \frac{n}{t}\binom{n-1}{t-1}, \quad t \geq 1, \tag{2.3.1}$$

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

$$\binom{n}{t} = \frac{n}{t}\binom{n-1}{t-1} = \frac{n}{t}\cdot\frac{n-1}{t-1}\binom{n-2}{t-2}, \quad t \geq 2. \tag{2.3.2}$$

Therefore,

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

where (2.3.2) was used to set the last equality. Changing the variable t in the last summation to $r = t-2$, it follows that

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

where the last equality used that $\sum_{r=0}^{n-2}\binom{n-2}{r}p^r q^{n-2-r}$ is the sum of all non-null probabilities in a $\mathit{Binomial}(n-2,p)$ distribution, so that the summation equals to 1. Consequently, for $n \geq 2$, $E_p[T_n(T_n-1)/c_n] = p^2$, where $c_n = n(n-1)$, and then, since the parameter $p \in [0,1]$ is arbitrary, $T_n(T_n-1)/c_n$ is an unbiased estimator of p^2.

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

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

Applying (2.3.1) three times, it follows that

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

and these two last displays together yield that

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

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

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

$$\frac{T_n(T_n-1)}{c_n} = \frac{T_n(T_n-1)}{n(n-1)} = \frac{\overline{X}_n(\overline{X}_n-1/n)}{1(1-1/n)} \xrightarrow{\mathrm{P}_p} \frac{p\cdot p}{1\cdot 1} = p^2$$

and, similarly,

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

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

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

Remark 2.3.1. For $a \in \mathbb{R}$ and a positive integer k, set

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

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

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

whenever the expectation exists. With this notation, the core of the solution to Exercise 2.3.1 was the computation of $E_p(T_n)_2]$ and $E_p[(T_n)_3]$, the second and third factorial moments of T_n. In some cases, the evaluation of a factorial moment of W can be simplified by using the following *factorial moments generating function*:

$$\text{FactM}_W(t) = E[t^W], \quad t > 0. \tag{2.3.3}$$

If this function is finite in a neighborhood of 1, then its derivatives of all orders exist about 1; this fact follows from the dominated convergence theorem (Apostol, 1980, Rudin, 1984). Moreover, the derivatives of $\text{FactM}_W(t)$ are given by

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

where $k \geq 1$. Evaluating at $t = 1$, it follows that

$$\begin{aligned}
&\frac{d^k}{dt^k}\text{FactM}_W(t)\bigg|_{t=1} \\
&\qquad = E[(W)_k] = E[W(W-1)(W-2)\cdots(W-k+1)],
\end{aligned} \tag{2.3.4}$$

and then the factorial moments of W can be determined by taking the derivatives of $\text{FactM}_W(t)$ and evaluating them at $t = 1$. For the random variable T_n in the previous exercise, $T_n \sim \textit{Binomial}(n, p)$, and the factorial moments generating function is easily determined:

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

and then

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

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

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

and

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

thus, the generating function of factorial moments provides an alternative way to compute the expectations in Example 2.3.1. □

2.4. Additional Examples

Before concluding this chapter, the ideas of unbiasdness and consistency are illustrated in some additional examples involving continuous and discrete distributions.

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 $T_n \xrightarrow{P_\theta} \theta$ and $T_n' \xrightarrow{P_\theta} \theta$ with the continuity theorem, it follows that $T_n T_n' \xrightarrow{P_\theta} \theta^2$ and $T_n(T_n' - 1) \xrightarrow{P_\theta} \theta(\theta - 1)$, so that the estimators in parts (a) and (b) are consistent. □

Exercise 2.4.2. Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with distribution $\mathcal{N}(\mu, \mu)$ for some $\mu > 0$. Find a consistent unbiased estimator of μ^2. [*Hint:* $E[\overline{X}_n] = \mu$ and $E[S_n^2] = \mu$; consider $T_n = \overline{X}_n S_n^2$.]

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

Exercise 2.4.3. 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 $b_n(\theta) = \mu(\theta) - \theta = 1/2 - 7\theta/6$

(b) Notice that $b_n(\theta) = 1/2 - 7\theta/6 \neq 0$ for all $\theta \in [0,1]$ does not depend on n, so that $\lim_{n\to\infty} b_n(\theta) = 1/2 - 7\theta/6$, 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). □

Exercise 2.4.4. Let the random variables $X_1, X_2, \ldots, X_n$ be independent and identically distributed Poisson random variables with parameter $\lambda > 0$. Show that $T_n = \overline{X}_n^{\,2} - \overline{X}_n$ is a biased estimator of λ^2, find its bias $b_n(\lambda)$ and hence, find an unbiased estimator of λ^2. Does $\lim_{n\to\infty} b_n(\lambda) = 0$ for all λ?

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

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

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

Chapter 3

Maximum Likelihood Estimation

In this chapter a fundamental procedure to obtain an estimator of a parametric function will be presented. The method is based on an intuitive principle that can be roughly described as follows: After observing the value attained by the random 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 from $\hat{\theta}$. The objective of this chapter is to describe formally this method, and illustrate the computation of the corresponding estimators in some specific examples.

3.1. Introduction

The starting point of the exposition on the maximum likelihood procedure is to define a measure of the *likelihood* of an observation $\mathbf{X} = \mathbf{x}$ under the different parameter values. To achieve this goal, consider a statistical model

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

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

$f_{\mathbf{X}}(\mathbf{x};\theta) = P_\theta[\mathbf{X} = \mathbf{x}]$ be the probability function of $\mathbf{X}$ under the condition that 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 proportionality 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 the above approximation to the probability of the observation $\mathbf{X} = \mathbf{x}$.

3.2. Maximum Likelihood and the Invariance Property

In this section the procedure of maximum likelihood to estimate the true parameter value θ will be formally introduced, and the *invariance* property, allowing to estimate an arbitrary parametric function, will be established.

Definition 3.2.1. 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, $\hat\theta$ satisfies

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

see (3.1.1).

This method to construct estimators of θ plays a central role in statistics, and there are, at least, three 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) The asymptotic variance of maximum likelihood estimators is minimal (Dudewicz and Mishra, 1988, Lehmann and Casella, 1998, Shao, 2010).

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

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

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 $\tilde{g}$:

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

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. The Log-Likelihood Function

Before presenting some examples on maximum likelihood estimators, it is convenient to note that, when $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is a sample 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 role to analyze this optimization problem (Fulks, 1980, Khuri, 2002). In particular, if 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 $\hat{\theta}$. These remarks are illustrated in a collection of fully analyzed examples presented in the following section.

3.4. The Maximum Likelihood Technique in Specific Examples

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.

Exercise 3.4.1. Let $X_1, X_2, \ldots, X_n$ be a random sample from the uniform density in $(0, \theta]$, that is, $f(x; \theta) = (1/\theta) I_{(0,\theta]}(x)$, where $\theta \in \Theta = (0, \infty)$. Find the maximum likelihood estimator of θ, say T_n, and show that $\{T_n\}$ is a consistent sequence of estimators.

Solution. The likelihood function is given by

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

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

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

where, to establish the second equality it was used that, when θ is the parameter value, the inequality $\hat{\theta}_n \le \theta$ always holds, and the last step is due to the fact that $P_\theta[X_i \le \theta - \varepsilon] = 1 - \varepsilon/\theta$ for all i. It follows that $P_\theta[|\hat{\theta}_n - \theta| > \varepsilon] \to 0$ as $n \to \infty$, that is, $\hat{\theta}_n \xrightarrow{\mathrm{P}_\theta} \theta$, 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 m from a $\mathcal{N}(\mu, \sigma^2)$ distribution, and let $Y_1, Y_2, \ldots, Y_n$ be a random sample of size n from a $\mathcal{N}(\nu, \sigma^2)$ distribution, where the two samples are independent. Find the maximum likelihood estimator of the overlapping coefficient

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

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

Solution. The likelihood function is given by

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

and its logarithm is given by

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

where C stands for a term that does not involve the parameter θ. The critical points of $\mathcal{L}(\cdot; \mathbf{X}, \mathbf{Y})$ satisfy

$$\partial_\mu \mathcal{L}(\theta; \mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{n} \frac{(X_i-\mu)}{\sigma^2} = 0$$

$$\partial_\nu \mathcal{L}(\theta; \mathbf{X}, \mathbf{Y}) = \sum_{j=1}^{m} \frac{(Y_j-\nu)}{\sigma^2} = 0$$

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

Direct calculations yield that the unique solution (μ_*, ν_*, σ_*) of this system is given by

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

where $\tilde{S}^2_{n\,X} = \sum_{i=1}^{n}(X_i - \overline{X}_n)^2/n$ and $\tilde{S}^2_{m\,Y} = \sum_{j=1}^{m}(Y_j - \overline{Y}_m)^2/m$ are the maximum likelihood estimators of σ^2 based on $\mathbf{X}$ and $\mathbf{Y}$, respectively. Since

$\mathcal{L}(\theta; \mathbf{X}, \mathbf{Y}) \to -\infty$ when $|\mu| + |\nu| \to \infty$ or $\sigma \to 0$, it follows that the above point $(\mu_*, \nu_*, \sigma_*^2)$ is the maximizer of $\mathcal{L}(\cdot; \mathbf{X}, \mathbf{Y})$, that is,

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

When $\min\{n, m\} \to \infty$, the law of large numbers implies that

$$\overline{X}_n \xrightarrow{\mathrm{P}_\theta} \mu, \quad \overline{Y}_m \xrightarrow{\mathrm{P}_\theta} \nu, \quad \tilde{S}^2_n \xrightarrow{\mathrm{P}_\theta} \sigma^2, \quad \text{and} \quad \tilde{S}^2_m \xrightarrow{\mathrm{P}_\theta} \sigma^2$$

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

$$\hat{\sigma}^2_{n\,m} \xrightarrow{\mathrm{P}_\theta} \sigma^2.$$

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

$$\begin{aligned}\hat{\Delta}_{n\,m} &= \Delta(\hat{\theta}_{n\,m}) \\ &= 2\Phi\left(-\frac{|\overline{X}_n - \overline{Y}_n|}{\hat{\sigma}_{n\,m}}\right) \xrightarrow{\mathrm{P}_\theta} 2\Phi\left(-\frac{|\mu - \nu|}{\sigma}\right) = \Delta(\theta) = \Delta,\end{aligned}$$

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

The following example shows that the calculation of a maximum likelihood estimator may be a difficult task, even in familiar and apparently simple cases.

Exercise 3.4.3. 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 its logarithm is given by

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

where, as before, the term C does not involve θ. Assuming that this function has a maximizer in the parameter space $\Theta = \mathbb{R}\times(0,\infty)\times(0,\infty)$, such a point must satisfy the following likelihood system:

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

The first equation yields that

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

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

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

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

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

a cubic equation in μ □

.

In the following example, the estimation of paramenter of a gamma density will be studied when the other parameter is known; in contrast to the case in which both parameters are unknown, an explicit formula for the estimator will be obtained.

Exercise 3.4.4. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the *Beta* (α, β) density

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

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

(a) Find the maximum likelihood estimator of α when $\beta = 1$;

(b) Find the maximum likelihood estimator of α when $\beta = 2$;

(c) Find the maximum likelihood estimator of $\alpha/(1+\alpha)$ in each of the preceding cases (a) and (b).

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

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

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

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

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

whose logarithm is

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

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

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

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

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

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

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

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

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

whose logarithm is given by

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

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

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

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

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

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

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

hence,

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

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

In the following example, a model with discrete parameter space will be studied; naturally, differentiation will not be directly used to determine the maximum likelihood estimator.

Exercise 3.4.5. Let $X_1, X_1, \ldots, X_n$ be a random sample of size n from the (discrete) *Uniform* $(\{1, 2, \ldots, \theta\})$ distribution on the set $\{1, 2, \ldots, \theta\}$, whose probability function is given by

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

Find the maximum likelihood estimator of θ, and its mean. Is this estimator unbiased?

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

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

an expression that can be written as

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

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

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

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

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

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

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

and then

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

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

In the following example a sample of a *Bernoulli* (p) distribution will be studied; an interesting aspect will be the comparison of variances between biased an unbiased estimators.

Exercise 3.4.6. Let X_i, $i = 1, 2, \ldots, n$ be a random sample of size n from the *Bernoulli* (p) distribution, where $p \in [0, 1]$, and set $T_n = X_1 + X_2 + \cdots + X_n$.

(a) Find the maximum likelihood estimator M_n of $pq = p(1-p)$

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

(c) Show that the maximum likelihood estimator of pq is biased, but is asymptotically unbiased.

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

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

(b) In Exercise 2.3.1 it was shown that $T_n(T_n-1)/[n(n-1)]$ is an unbiased estimator of p^2. Since $\overline{X}_n = T_n/n$ is an unbiased estimator of p, it follows that

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

so that U_n is an unbiased estimator of pq.

(c) Notice that

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

Hence,

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

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

(d) As already noted in the previous part, $M_n = [(n-1)/n]U_n$, so that $\mathrm{Var}_p\,[M_n] = \mathrm{Var}_p\,[[(n-1)/n]\,U_n] = [(n-1)/n]^2\mathrm{Var}_p\,[U_n]$. Therefore,

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

showing that the variance of the unbiased estimator U_n is larger than the variance of the maximum likelihood estimator M_n. □

The estimation problem for the logistic location-scale family is studied in the following example. As it will be shown, the maximum likelihood estimators must be determined computationally.

Exercise 3.4.7. Let $X_1, X_2, \ldots, X_n$ be a random variable of size n from the logistic density

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

where α is an unknown real number and β is known. In this context, find the maximum likelihood estimator of α.

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

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

and its logarithm is

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

Thus,

$$\begin{aligned} \partial_\alpha \mathcal{L}(\alpha;\mathbf{X}) &= -n + 2\sum_{i=1}^{n} \frac{e^{-(\alpha+\beta X_i)}}{1+e^{-(\alpha+\beta X_i)}} \\ &= -n + 2\sum_{i=1}^{n} \frac{e^{-(\alpha+\beta X_i)}+1-1}{1+e^{-(\alpha+\beta X_i)}} \\ &= -n + 2\sum_{i=1}^{n}\left[1 - \frac{1}{1+e^{-(\alpha+\beta X_i)}}\right] \\ &= n - 2\sum_{i=1}^{n} \frac{1}{1+e^{-(\alpha+\beta X_i)}}. \end{aligned} \tag{3.4.1}$$

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

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

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

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

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

$$\partial_\alpha \mathcal{L}(\alpha; \mathbf{X})\big|_{\alpha=\alpha^*} = 0. \tag{3.4.2}$$

Notice now that

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

so that $\mathcal{L}(\alpha; \mathbf{X})$ is a concave function of α, and then the point α^* satisfying (3.4.2) is unique, and is the unique maximizer of $\mathcal{L}(\alpha; \mathbf{X})$, that is, $\hat{\alpha}_n(\mathbf{X}) = \alpha^*$. Notice that (3.4.1) and (3.4.2) together yield that $\hat{\alpha}_n$ is the unique solution of the likelihood equation

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

which, as already mentioned, must be solve numerically. □

The following example is simple and closely related with the Bernoulli case previously analyzed.

Exercise 3.4.8. Suppose that

$$\begin{aligned} X_1 &\sim \mathit{Binomial}(n_1, p) \\ X_2 &\sim \mathit{Binomial}(n_2, p) \\ &\vdots \\ X_k &\sim \mathit{Binomial}(n_k, p) \end{aligned}$$

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

Solution. Given $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ such that X_i is an integer between 0 and n_1, the corresponding likelihood function is

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

whose logarithm is given by

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

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

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

□

The following example concerns the estimation of the right tail of a normal distribution.

Exercise 3.4.9. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the $\mathcal{N}\left(\mu, \sigma^2\right)$ distribution, where the vector $(\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$ is unknown. Set

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

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

Solution. The basic properties of the normal distribution yield that

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

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

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

the invariance theorem yields that

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

since $g(\cdot, \cdot)$ is a continuous function in the parameter space Θ and the sequences $\{\overline{X}_n\}$ and $\{S_n^2\}$ estimate consistently to the parameters μ and σ^2, respectively, it follows from the continuity theorem that $\{\hat{g}_n\}$ is a consistent sequence. □

3.5. A Non-smooth Example

In this section the determination of the maximum likelihood estimator for a location family based on the Laplace distribution will be studied. The interesting part of the analysis is that it is not difficult to prove that the likelihood function achieves its maximum at an interior point of the parameter space, but the optimizer can not be determined by direct differentiation and requires a rather careful analysis.

Exercise 3.5.1. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the (Laplace) double exponential density with center $\theta \in \mathbb{R} \equiv \Theta$, which is given by

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

Find the maximum likelihood estimator of θ.

Solution. The likelihood function is

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

and it logarithm is given by

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

where $C = -n\log(2)$. The main difficulty in this problem is that the absolute value function is not differentiable at every point. Indeed, the mapping $\theta \mapsto |x - \theta|$ is not differentiable at $\theta = x$, whereas

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

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

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

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

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

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

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

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

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

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

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

$$
\begin{aligned}
\theta < X_{(1)} &\Rightarrow \partial_\theta \mathcal{L}(\theta;\mathbf{X}) = n\\
X_{(1)} < \theta < X_{(2)} &\Rightarrow \partial_\theta \mathcal{L}(\theta;\mathbf{X}) = n-2\\
X_{(2)} < \theta < X_{(3)} &\Rightarrow \partial_\theta \mathcal{L}(\theta;\mathbf{X}) = n-4\\
X_{(3)} < \theta < X_{(4)} &\Rightarrow \partial_\theta \mathcal{L}(\theta;\mathbf{X}) = n-6\\
&\vdots\\
X_{(k)} < \theta < X_{(k+1)} &\Rightarrow \partial_\theta \mathcal{L}(\theta;\mathbf{X}) = n-2k\\
X_{(k+1)} < \theta < X_{(k+2)} &\Rightarrow \partial_\theta \mathcal{L}(\theta;\mathbf{X}) = n-2(k+1)\\
&\vdots\\
X_{(n-3)} < \theta < X_{(n-2)} &\Rightarrow \partial_\theta \mathcal{L}(\theta;\mathbf{X}) = 6-n\\
X_{(n-2)} < \theta < X_{(n-1)} &\Rightarrow \partial_\theta \mathcal{L}(\theta;\mathbf{X}) = 4-n\\
X_{(n-1)} < \theta < X_{(n)} &\Rightarrow \partial_\theta \mathcal{L}(\theta;\mathbf{X}) = 2-n\\
X_{(n)} < \theta &\Rightarrow \partial_\theta \mathcal{L}(\theta;\mathbf{X}) = -n
\end{aligned}
\tag{3.5.1}
$$

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

$$n \geq 2k^* \quad \text{and} \quad n - 2(k^*+1) < 0. \tag{3.5.2}$$

With this notation, (3.5.1) shows that

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

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

(b) If $\theta \in (X_{(k^*+1)}, X_{(k^*+2)}) \cup (X_{(k^*+2)}, X_{(k^*+3)}) \cup \cdots \cup (X_{(n-1)}, X_{(n)}) \cup (X_{(n)}, \infty)$, then the partial derivative $\partial_\theta \mathcal{L}(\theta;\mathbf{X})$ is negative; in this case, by continuity of $\mathcal{L}(\theta;\mathbf{X})$, the mapping $\theta \mapsto \mathcal{L}(\theta;\mathbf{X})$ is decreasing in $\theta \in [X_{(k^*+1)}, \infty)$. Thus,

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

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

$$\hat{\theta}_n = X_{(k^*+1)}. \tag{3.5.3}$$

If the sample size n is odd, say $n = 2r + 1$, then k^* in (3.5.2) equals r, and $X_{(k^*+1)} = X_{(r+1)}$ is the sample median,

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

On the other hand, if the sample size n is even, $n = 2r$, then k^* in (3.5.2) equals r, and $\partial_\theta \mathcal{L}(\theta; \mathbf{X})$ is zero in the interval $(X_{(r)}, X_{(r+1)})$, so that $\mathcal{L}(\theta; \mathbf{X})$ is constant on the interval $\theta \in [X_{(r)}, X_{(r+1)}]$, and then every point in that interval is a maximizer of $\mathcal{L}(\theta; \mathbf{X})$. Notice that when $n = 2r$ is an even integer, every point in $[X_{(r)}, X_{(r+1)}]$ is a median of the data, and the above expression for $\hat{\theta}_n$ remains valid. Summarizing: the maximum likelihood estimator of θ is any sample median, and if the sample size n is even, $\hat{\theta}_n$ is not unique. $\square$

Chapter 4

Method of Moments

This chapter introduces an additional procedure to construct estimators of parametric functions, namely, *the method of moments*. Essentially, this procedure can be 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*. The method is easily implemented when the interesting parametric quantity is determined in terms of population moments, and renders consistent and asymptotically normal estimators which, generally, are biased but asymptotically unbiased.

4.1. Moment Estimators

In this section the method of moments to build estimators is formally described. 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. Next, let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample of size n of the population P_θ, so that

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

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 is formally stated 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 of the Method of Moments

The objective of this section is to prove that, under mild conditions, the consistency of a sequence $\{\hat{g}_n\}$ of moment estimators is a generic property.

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).

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

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

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

$$|x_i - \mu'_i(\theta)| \le \delta, \quad i = 1, 2, \ldots, r$$
$$\Rightarrow |G(x_1, x_2, \ldots, x_r) - G(\mu'_1(\theta), \mu'_2(\theta), \ldots, \mu'_r(\theta))| \le \varepsilon.$$

This implication is equivalent to

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

Consequently,

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

that is,

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

which can be written as

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

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

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

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

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

establishing (4.2.1). □

Before proceeding to present some examples on the method of moments, it is convenient to summarize the precedent discussion:

Given a sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ 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 of $\mu'_1(\theta), \mu'_2(\theta), \ldots, \mu'_k(\theta)$ 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.

4.3. Examples

The above ideas about the method of moments are illustrated in a series of examples.

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

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

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

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

Solution. (a) First, the parametric quantity $g(\theta) = \theta$ must be expressed in terms of the moments of the parent distribution. In the present case, if $X \sim$ *Uniform* $(0, \theta)$, then $\mu_1'(\theta) = E_\theta[X] = \theta/2$, so that $\theta = 2\mu_1'$. Consequently, the moments estimator of θ is $\hat{\theta}_n = 2{m_1'}_n(\mathbf{X}) = 2\overline{X}_n$. Notice that θ is a linear function of $\mu_1'(\theta)$, and then $\hat{\theta}_n$ is unbiased.

(b) The moments estimator of $g(\theta) = \theta^2$ based on the sample of size n is $\hat{g}_n = g(\hat{\theta}_n) = \hat{\theta}_n^2 = (2\overline{X}_n)^2 = 4\overline{X}_n^2$; since $\hat{\theta}_n$ is not constant, Jensen's inequality yields that $E_\theta[\hat{g}_n] = E_\theta[(\hat{\theta}_n)^2] > E_\theta[\hat{\theta}_n]^2 = \theta^2$, and then $\hat{g}_n$ is a biased estimator; for a discussion of Jensen's inequality, see, for instance, Rudin (1984), or Khuri (2002). To determine an unbiased estimator of $g(\theta) = \theta^2$, notice that

$$
\begin{aligned}
E_\theta[\hat{\theta}_n^2] &= \mathrm{Var}_\theta\left[\hat{\theta}_n\right] + E_\theta[\hat{\theta}_n]^2 \\
&= \mathrm{Var}_\theta\left[2\overline{\mathbf{X}}_n\right] + \theta^2 \\
&= 4\mathrm{Var}_\theta\left[\overline{\mathbf{X}}_n\right] + \theta^2 \\
&= 4\frac{\theta^2}{12n} + \theta^2 = \left(1 + \frac{1}{3n}\right)\theta^2.
\end{aligned}
$$

Consequently, $U_n = 3n/(1+3n)\hat{\theta}_n^2 = (3n/(1+3n)\hat{g}_n = 12n/(1+3n)\overline{X}_n^2$ is an unbiased estimator of θ^2.

(c) Notice that in parts (a) and (b), θ and $g(\theta)$ are continuous functions of the population moments, and then the sequences $\{\hat{\theta}_n\}$ and $\{\hat{g}_n\}$ are consistent for θ and $g(\theta)$, respectively. Also, $U_n = (3n/(1+3n)\hat{g}_n \xrightarrow{\text{P}_\theta} 1\cdot g(\theta) = g(\theta)$, and then $\{U_n\}$ is a consistent sequence for the parametric function $g(\theta)$. □

Exercise 4.3.2. Let $X_1, X_2, \ldots, X_n$ be independent random variables, each with the density $f(x;\theta) = 1/(2\theta)I_{[-\theta,\theta]}(x)$. Find the moments estimator of θ and show directly that is biased.

Solution. The uniform distribution on the interval $[-\theta,\theta]$ has mean $\mu_1' = 0$, so that θ can not be expressed as a function of μ_1'. Therefore, the second moment must be calculated. If $X \sim Uniform(-\theta,\theta)$,

$$\mu_2'(\theta) = E_\theta[X^2] = \text{Var}_\theta\,[X_2] = \frac{(2\theta)^2}{12} = \frac{\theta^2}{3}.$$

Since θ is a positive number, it follows that $\theta = (3\mu_2'(\theta))^{1/2}$, and then the moments estimator of θ is given by

$$\hat{\theta}_n = (3m_{2\,n}')^{1/2}.$$

This estimator is biased. Indeed, the second sample moment is not constant with probability 1 and, using that the function $H(x) = x^{1/2}$ is strictly concave, it follows that

$$E_\theta[\hat{\theta}_n] = E_\theta[(3m_{2\,n}')^{1/2}] = E_\theta[H(3m_{2\,n}')] > H(E_\theta[(3m_{2\,n}']) = H(\theta^2) = \theta,$$

so that $\hat{\theta}_n$ is biased with positive bias function. □

Exercise 4.3.3. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the *Geometric* (p) distribution, so that the common probability function of the variables is

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

Use the method of moments to find an estimator of p. Show that the method of moments used to estimate $1/p$ produces the estimator $\overline{X}_n$.

Solution. If $X \sim \mathit{Geometric}\,(p)$, then

$$\begin{aligned}\mu_1'(\theta) = E_\theta[X] &= \sum_{x=1}^{\infty} x(1-p)^{x-1}p \\ &= p\sum_{x=1}^{\infty} x(1-p)^{x-1} \\ &= p\frac{d}{dp}\left[\sum_{x=1}^{\infty}(1-p)^x\right] \\ &= p\frac{d}{dp}\left[\frac{1}{1-(1-p)}\right] = \frac{p}{p^2} = \frac{1}{p}.\end{aligned}$$

It follows that $p = 1/\mu_1'$, and then the method of moments produces the following estimator of p:

$$\hat{p}_n = \frac{1}{m'_{1\,n}} = \frac{1}{\overline{X}_n}$$

As for the estimation of $g(p) = 1/p$, the previous calculations show that $g(p) = \mu_1'(p)$, and then

$$\hat{g}_n = m'_{1\,n} = \overline{X}_n$$

is the estimator of $g(p)$ produced by the method of moments. □

Exercise 4.3.4. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the discrete uniform distribution on the set $\{1, 2, \ldots, \theta\}$ where θ is an unknown positive integer. Use the method of moments to find an estimator of θ.

Solution. To express the parameter θ in terms of the population moments, just notice that if $X \sim \mathit{Uniform}\,(\{1, 2, \ldots, \theta\})$ then $\mu_1'(\theta) = E_\theta[X] = (1 + \theta)/2$, so that $\theta = 2\mu_1'(\theta) - 1$. Hence, the method of moments produces the estimator $\hat{\theta}_n = 2m'_{1\,n} - 1 = 2\overline{X}_n - 1$; since θ is a linear function of $\mu_1'(\theta)$, it follows that the estimators $\hat{\theta}_n$ are unbiased and the sequence $\{\hat{\theta}_n\}$ is consistent. □

Exercise 4.3.5. Let $X_1, X_2, \ldots, X_n$ be a random sample of the $\mathit{Gamma}\,(\alpha, \lambda)$ distribution, where $\theta = (\alpha, \lambda) \in \Theta = (0, \infty) \times (0, \infty)$. Use the method of moments to obtain estimators of α and λ.

Solution. The starting point is to evaluate the moments of order one and two of the $Gamma(\alpha, \lambda)$ distribution. It is known that if $X \sim Gamma(\alpha, \lambda)$, then

$$\mu_1'(\theta) = E_\theta[X] = \frac{\alpha}{\lambda}, \quad \text{and} \quad \mu_2'(\theta) = E_\theta[X^2] = \frac{\alpha(\alpha+1)}{\lambda^2}. \tag{4.3.1}$$

To express α and λ in terms of $\mu_1'(\theta)$ and $\mu_2'(\theta)$, notice that

$$\frac{\mu_2'(\theta)}{\mu_1'(\theta)^2} = \frac{\alpha(\alpha+1)/\lambda^2}{\alpha^2/\lambda^2} = \frac{\alpha+1}{\alpha} = 1 + \frac{1}{\alpha}.$$

Hence,

$$\frac{\mu_2'(\theta) - \mu_1'(\theta)^2}{\mu_1'(\theta)^2} = \frac{1}{\alpha},$$

which is equivalent to

$$\alpha = \frac{\mu_1'(\theta)^2}{\mu_2'(\theta) - \mu_1'(\theta)^2}.$$

Combining this expression with the first equality in (4.3.1), it follows that

$$\lambda = \frac{\alpha}{\mu_1'(\theta)} = \frac{\mu_1'(\theta)}{\mu_2'(\theta) - \mu_1'(\theta)^2}.$$

Then the method of moments estimation prescribes the estimators

$$\hat{\alpha}_n = \frac{(m_{1\,n}')^2}{m_{2\,n}' - (m_1')^2},$$

and

$$\hat{\lambda}_n = \frac{m_{1\,n}'}{m_{2\,n}' - (m_{1\,n}')^2}.$$

Since $m_{1\,n}' = \overline{X}_n$ and $m_{2\,n}' - (m_{1\,n}')^2 = \sum_{i=1}^n X_i^2/n - \overline{X}_n^2 = \sum_{i=1}^n (X_i - \overline{X}_n)^2/n = \tilde{S}_n^2$ the above estimators can be expressed in more familiar terms:

$$\hat{\alpha}_n = \frac{\overline{X}_n^2}{\tilde{S}_n^2}, \quad \text{and} \quad \hat{\lambda}_n = \frac{\overline{X}_n}{\tilde{S}_n^2}.$$

Since α and λ are continuous functions of $\mu_1'(\theta)$ and $\mu_2'(\theta)$, the sequences $\{\hat{\alpha}_n\}$ and $\{\hat{\lambda}_n\}$ are consistent. □

Remark 4.3.1. An interesting aspect of the precedent problem is that method of moments allowed to obtain explicit formulas for the estimators of α and

λ. In contrast, the maximum likelihood estimators of α and λ must be determined numerically for each data set. □

Exercise 4.3.6. 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 $X \sim \textit{Beta}\,(\alpha, \beta)$, 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

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

concluding the argument. □

Remark 4.3.2. Observe that $\hat{\alpha}_n$ and $\hat{\beta}_n$ contain the factor

$$\left(\frac{m_{1\,n}'(1-m_{1\,n}')}{m_{2\,n}'} - 1 \right) = \left(\frac{\overline{X}_n(1-\overline{X}_n)}{\sum_{i=1}^n X_i^2/n} - 1 \right). \tag{4.3.2}$$

As it will be shown below, that this factor may be negative for some samples, a fact that illustrates a disadvantage of the method of moments, namely, the estimates generated by the method, do not necessarily belong to the parameter space. Consider the sample

$$\mathbf{X} = \mathbf{x} = (\varepsilon, \varepsilon, \ldots, \varepsilon, 1-\varepsilon) \tag{4.3.3}$$

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

$$\lim_{n\to\infty} \overline{X}_n = \frac{1}{n} \quad \text{and} \quad \lim_{n\to\infty} \sum_{i=1}^n X_i^2/n = \frac{1}{n}. \tag{4.3.4}$$

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 (4.3.3), 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* (4.3.4), 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 (4.3.2) 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. □

Exercise 4.3.7. Let $X_1, X_2, \ldots, X_n$ be a sample of the $\mathcal{N}\left(0, \sigma^2\right)$ distribution, where $\sigma \in (0, \infty)$. Find the moments estimator of σ^2 and analyze the consistency of the sequence $\{\hat{\sigma}^2\}$.

Solution. If $X \sim \mathcal{N}\left(0, \sigma^2\right)$, then $\mu_1'(\sigma) = E_\sigma[X] = 0$, so that σ^2 can not be expressed in terms of $\mu_1'(\sigma)$ and it is necessary to compute more moments of X. Next, observe that $\mu_2'(\sigma) = E_\sigma[X^2] = \mathrm{Var}_\sigma\,[X] = \sigma^2$, and it follows that the interesting parametric function—σ^2 in the present problem—equals the second population moment. Thus, the method of moments prescribes the estimator

$$\hat{\sigma}^2 = m_{2\,n}' = \frac{1}{n}\sum_{i=1}^{n} X_i^2;$$

since σ^2 is a linear function of $\mu_2'(\sigma)$, the estimator $\hat{\sigma}^2$ is unbiased for σ^2. □

Chapter 5

Quantile Estimation

In this chapter the problem of estimating a quantile of a distribution function F is considered. Roughly, given $\alpha \in (0,1)$, a quantile q_α of order α for F is the point that has the following property: the total probability accumulated form $-\infty$ to the point q_α is exactly α. Hereafter the discussion is restricted to continuous distributions, and in this context a quantile q_α exists for each $\alpha \in (0,1)$. The estimator of q_α based on a sample of size n is the data value that occupies among the observations a similar position to the one of q_α in the underlying population. Using a relation between the distribution of an order statistic and a binomial random variable, the limit distribution of the estimator of q_α will be determined.

5.1. Population Quantiles

In this section a population quantile q_α of a given order $\alpha \in (0,1)$ is formally defined, and under mild conditions its uniqueness is established.

Definition 5.1.1. Let $F(x)$ be a continuous distribution function F with density $f(x)$, so that

$$F(x) = \int_{-\infty}^{x} f(z)\,dz, \quad x \in \mathbb{R}.$$

For a given number $\alpha \in (0,1)$, a quantile $q_\alpha \equiv q_\alpha(F)$ for the distribution function F is any solution of the equation

$$F(q_\alpha) = \alpha. \tag{5.1.1}$$

A quantile of order .5 satisfies $F(q_{0.5}) = 0.5$ and is called a median of the distribution. Quantiles of order 0.25 and 0.75 satisfy $F(q_{0.25}) = 0.25$ and $F(q_{0.75}) = 0.75$, respectively, and are usually referred to as *quartiles* of F; the difference $q_{0.75} - q_{0.25}$ is the interquartile range, and is frequently used as a descriptive measure of the dispersion of the distribution function F (Montgomery, 2011).

As for the existence of a quantile q_α, recall that $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$. Since F is assumed to be continuous, from the intermediate value property it follows that, for each $\alpha \in (0,1)$, there exists a number q_α satisfying (5.1.1); see, for instance, Khuri (2002). However, in general a quantile of a given order is not unique, since F may be constant in some interval. In the following lemma sufficient conditions are given to ensure the uniqueness of a quantile.

Lemma 5.1.1. Let F be a continuous distribution function with density f and, given $\alpha \in (0,1)$, let q_α be a quantile of order α for F. In this case, if

$$f(x) \text{ is continuous at } x = q_\alpha \text{ and } f(q_\alpha) \neq 0, \tag{5.1.2}$$

then q_α is the *unique* quantile of order α for F.

Proof. Notice that (5.1.2) implies that there exists $\delta > 0$ such that

$$f(x) > 0 \text{ if } x \in (q_\alpha - \delta, q_\alpha + \delta). \tag{5.1.3}$$

It will be shown that this property implies that

$$z \neq q_\alpha \Rightarrow F(z) \neq \alpha,$$

so that q_α is the unique quantile of order α. To achieve this goal, consider the following two exhaustive cases:

(i) $z < q_\alpha$. In this case select a point w such that

$$z < w < q_\alpha \quad \text{and} \quad w \in (q_\alpha - \delta, q_\alpha), \tag{5.1.4}$$

and notice that

$$\begin{aligned} \alpha = F(q_\alpha) &= \int_{-\infty}^{q_\alpha} f(x)\,dx \\ &= \int_{-\infty}^{w} f(x)\,dx + \int_{w}^{q_\alpha} f(x)\,dx; \end{aligned}$$

on the other hand, (5.1.3) and the inclusion in (5.1.4) together yield that $f(x)$ is positive at each point x in the interval (w, q_α), and then $\int_w^{q_\alpha} f(x)\,dx > 0$; combining this inequality with the previous display, it follows that

$$\alpha > F(w);$$

since the inequality $z < w$ implies that $F(w) \geq F(z)$, it follows that $\alpha > F(z)$.

(ii) $z > q_\alpha$. In this context, paralleling the above argument it will be proved that $F(z) > \alpha$. First, select a point w such that

$$z > w > q_\alpha \quad \text{and} \quad w \in (q_\alpha, q_\alpha + \delta); \tag{5.1.5}$$

with this notation,

$$\begin{aligned} F(w) &= \int_{-\infty}^{w} f(x)\,dx \\ &= \int_{-\infty}^{q_\alpha} f(x)\,dx + \int_{q_\alpha}^{w} f(x)\,dx \\ &= \alpha + \int_{q_\alpha}^{w} f(x)\,dx; \end{aligned}$$

since (5.1.3) and the inclusion in (5.1.5) imply that $f(x) > 0$ for $x \in (q_\alpha, w)$, it follows that $\int_{q_\alpha}^{w} f(x)\,dx > 0$, so the above displayed relation leads to

$$F(w) > \alpha;$$

since the inequality $z > w$ implies that $F(z) \geq F(w)$, it follows that $F(z) > \alpha$. □

5.2. Sample Quantiles and Consistency

The estimation of a quantile q_α will be consider in this section. Notice that (5.1.1) means that a probability α is accumulated to the left or at q_α according to F. Thus, given a sample $X_1, X_2, \ldots, X_n$ of that distribution function, it is natural to estimate the quantile q_α by the sample point $\hat{q}_{\alpha\, n}$ which has an analogous property in terms of the sample; such a point will be introduced below. Recall that the vector $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ order statistics of the sample consists of the same data values as the sample, but arranged in a non-decreasing order:

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)};$$

moreover, since the parent distribution function is continuous, the above inequalities are strict with probability one.

Definition 5.2.1. Let $\alpha \in (0,1)$ be arbitrary, and let $(X_1, X_2, \ldots, X_n)$ be a random sample of a distribution function F. Suppose that $1/n < \alpha$ and let $k(\alpha, n)$ be the the largest integer that does not exceed $n\alpha$, that is, $k(\alpha, n)$ is the positive integer that satisfies

$$\frac{k(\alpha, n)}{n} \leq \alpha, \quad \text{and} \quad \alpha < \frac{k(\alpha, n) + 1}{n}. \tag{5.2.1}$$

With this notation, the estimator $\hat{q}_{\alpha\, n}$ of the quantile q_α is defined by

$$\hat{q}_{\alpha,n} = X_{(k(\alpha,n))}. \tag{5.2.2}$$

Notice that the proportion of data that lay at or to the left of $\hat{q}_n = X_{(k(\alpha,n))}$ is $k(\alpha, n)/n$, and this proportion differs from α at most by $1/n$. Thus, $\hat{q}_{\alpha,n}$ occupies among the observations a position that is similar to the one occupied by q_α in the sampled population. The following result shows that $\{\hat{q}_{\alpha,n}\}$ is a consistent sequence of estimators of q_α.

Theorem 5.2.1. Assume that $X_1, X_2, X_3, \ldots$ is a sequence of independent random variables with a common distribution function $F(x)$ with density $f(x)$. Given $\alpha \in (0,1)$, let q_α be a quantile of order α of F, and suppose that the condition (5.1.2) holds, so that the quantile q_α is uniquely determined. In this context, the sequence $\{\hat{q}_{\alpha,n}\}$ in Definition 5.2.1 estimates q_α consistently, that is,

$$\hat{q}_{\alpha,n} \xrightarrow{\text{P}} q_\alpha.$$

The argument used below to establish this theorem relies on the following lemma, which establishes a connection between the binomial distribution and the quantiles of a sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. First, for each number $x \in \mathbb{R}$ define $N_n(\mathbf{X}; x)$ as the number of observations that are less than or equal to x, that is,

$$N_n(\mathbf{X}; x) = \sum_{i=1}^{n} I[X_i \leq x]; \tag{5.2.3}$$

since the independent variables X_i have the continuous distribution function $F(x)$, it follows that the indicators $I[X_i \leq x]$ are independent with common Bernoulli distribution with success parameter $p = F(x)$, a fact that allows to state that

$$N_n(\mathbf{X}; x) \sim \textit{Binomial}(n, F(x)), \tag{5.2.4}$$

and then, for every $\delta > 0$,

$$P\left[\left|\frac{N_n(\mathbf{X}; x)}{n} - F(x)\right| > \delta\right] \leq \frac{F(x)(1 - F(x))}{n\delta^2} \to 0 \quad \text{as } n \to \infty, \tag{5.2.5}$$

by Chebichev's inequality. Observe now that the inequality $X_{(k)} > x$ is equivalent to the statement that the number of observations that are less than or equal to x is less that k:

$$X_{(k)} > x \iff N_n(\mathbf{X}; x) < k \tag{5.2.6}$$

which is also equivalent to

$$X_{(k)} \leq x \iff N_n(\mathbf{X}; x) \geq k. \tag{5.2.7}$$

Lemma 5.2.1. Under the framework of Theorem 5.2.1, the random variables $N_n(\mathbf{X}; x)$ in (5.2.3) satisfy the following properties (i) and (ii), where $\alpha \in (0, 1)$ and $k(\alpha, n)$ is the integer specified in (5.2.1):

(i) $P[N_n(\mathbf{X}; q_\alpha + \varepsilon) < k(\alpha, n)] \to 0$ as $n \to \infty$.

(ii) $P[N_n(\mathbf{X}; q_\alpha - \varepsilon) \geq k(\alpha, n)] \to 0$ as $n \to \infty$.

Proof. Let ε be an arbitrary positive number.

(i) To begin with, notice that

$$\begin{aligned} N_n(\mathbf{X};&q_\alpha + \varepsilon) < k(\alpha, n) \\ &\iff N_n(\mathbf{X}; q_\alpha + \varepsilon) - nF(q_\alpha + \varepsilon) < k(\alpha, n) - nF(q_\alpha + \varepsilon) \\ &\iff \frac{N_n(\mathbf{X}; q_\alpha + \varepsilon)}{n} - F(q_\alpha + \varepsilon) < \frac{k(\alpha, n)}{n} - F(q_\alpha + \varepsilon). \end{aligned} \tag{5.2.8}$$

Observe now that the relation (5.2.1) specifying the integer $k(\alpha, n)$ yields that $k(\alpha, n)/n \to \alpha$ as $n \to \infty$; hence, since (5.1.2) implies that $F(q_\alpha + \varepsilon) > \alpha$, it follows that

$$\lim_{n\to\infty} \frac{k(\alpha, n)}{n} - F(q_\alpha + \varepsilon) = \alpha - F(q_\alpha + \varepsilon) < 0.$$

Thus, defining the positive number δ by $\delta := (F(q_\alpha + \varepsilon) - \alpha)/2$, it follows that there exists an integer m such that

$$\frac{k(\alpha, n)}{n} - F(q_\alpha + \varepsilon) < -\delta \quad \text{if } n > m.$$

Therefore, for $n > m$,

$$\begin{aligned} \frac{N_n(\mathbf{X}; q_\alpha + \varepsilon)}{n} - F(q_\alpha + \varepsilon) &< \frac{k(\alpha, n)}{n} - F(q_\alpha + \varepsilon) \\ &\Rightarrow \frac{N_n(\mathbf{X}; q_\alpha + \varepsilon)}{n} - F(q_\alpha + \varepsilon) < -\delta \\ &\Rightarrow \left| \frac{N_n(\mathbf{X}; q_\alpha + \varepsilon)}{n} - F(q_\alpha + \varepsilon) \right| > \delta, \end{aligned}$$

a fact that together with (5.2.8) allows to conclude that

$$\begin{aligned} &\text{if } n > m \\ &N_n(\mathbf{X}; q_\alpha + \varepsilon) < k(\alpha, n) \Rightarrow \left| \frac{N_n(\mathbf{X}; q_\alpha + \varepsilon)}{n} - F(q_\alpha + \varepsilon) \right| > \delta. \end{aligned}$$

Consequently,n if $n > m$,

$$P\left[N_n(\mathbf{X}; q_\alpha + \varepsilon) < k(\alpha, n)\right] \leq P\left[\left| \frac{N_n(\mathbf{X}; q_\alpha + \varepsilon)}{n} - F(q_\alpha + \varepsilon) \right| > \delta\right];$$

taking the limit as n goes to ∞ in both sides of this relation, (5.2.5) implies that

$$\lim_{n\to\infty} P\left[N_n(\mathbf{X}; q_\alpha + \varepsilon) < k(\alpha, n)\right] = 0,$$

which is the desired conclusion.

(ii) The argument is similar to the one in part (i). Notice that

$$\begin{aligned}
&N_n(\mathbf{X}; q_\alpha - \varepsilon) \geq k(\alpha, n) \\
&\qquad \iff N_n(\mathbf{X}; q_\alpha - \varepsilon) - nF(q_\alpha - \varepsilon) \geq k(\alpha, n) - nF(q_\alpha - \varepsilon) \\
&\qquad \iff \frac{N_n(\mathbf{X}; q_\alpha - \varepsilon)}{n} - F(q_\alpha - \varepsilon) \geq \frac{k(\alpha, n)}{n} - F(q_\alpha - \varepsilon).
\end{aligned} \tag{5.2.9}$$

Combining the convergence $k(\alpha, n)/n \to \alpha$ as $n \to \infty$ with the inequality $F(q_\alpha - \varepsilon) < \alpha$ (which is a consequence of (5.1.2)), it follows that

$$\lim_{n\to\infty} \frac{k(\alpha, n)}{n} - F(q_\alpha - \varepsilon) = \alpha - F(q_\alpha - \varepsilon) > 0,$$

so that the number δ specified by $\delta := (\alpha - F(q_\alpha - \varepsilon))/2$ is positive and there exists an integer m such that $k(\alpha, n)/n - F(q_\alpha - \varepsilon) > \delta$ for $n > m$. Hence, for $n > m$,

$$\begin{aligned}
\frac{N_n(\mathbf{X}; q_\alpha - \varepsilon)}{n} - F(q_\alpha - \varepsilon) &\geq \frac{k(\alpha, n)}{n} - F(q_\alpha - \varepsilon) \\
&\Rightarrow \frac{N_n(\mathbf{X}; q_\alpha - \varepsilon)}{n} - F(q_\alpha - \varepsilon) > \delta \\
&\Rightarrow \left|\frac{N_n(\mathbf{X}; q_\alpha - \varepsilon)}{n} - F(q_\alpha - \varepsilon)\right| > \delta,
\end{aligned}$$

a fact that together with (5.2.9) implies that, if $n > m$

$$N_n(\mathbf{X}; q_\alpha - \varepsilon) \geq k(\alpha, n) \Rightarrow \left|\frac{N_n(\mathbf{X}; q_\alpha - \varepsilon)}{n} - F(q_\alpha - \varepsilon)\right| > \delta.$$

It follows that, for $n > m$,

$$P\left[N_n(\mathbf{X}; q_\alpha - \varepsilon) \geq k(\alpha, n)\right] \leq P\left[\left|\frac{N_n(\mathbf{X}; q_\alpha - \varepsilon)}{n} - F(q_\alpha - \varepsilon)\right| > \delta\right],$$

and taking the limit as n goes to ∞, (5.2.5) yields that

$$\lim_{n\to\infty} P\left[N_n(\mathbf{X}; q_\alpha - \varepsilon) \geq k(\alpha, n)\right] = 0,$$

completing the proof. □

Proof of Theorem 5.2.1. Given $\alpha \in (0, 1)$ it is necessary to show that, for each $\varepsilon > 0$, the convergence $P[|\hat{q}_{\alpha,n} - q_\alpha| > \varepsilon] \to 0$ holds as $n \to \infty$. With this in mind, notice that

$$[|\hat{q}_{\alpha,n} - q_\alpha| > \varepsilon] = [\hat{q}_{\alpha,n} > q_\alpha + \varepsilon] \cup [\hat{q}_{\alpha,n} < q_\alpha - \varepsilon]. \tag{5.2.10}$$

It will be proved that the probability of each event in the right-hand side converges to 0 as n goes to ∞. To begin with, let $\varepsilon > 0$ be arbitrary and recall that $\hat{q}_{\alpha,n} = X_{(k(\alpha,n))}$, so that

$$[\hat{q}_{\alpha,n} > q_\alpha + \varepsilon] = [X_{(k(\alpha,n))} > q_\alpha + \varepsilon] = [N_n(\mathbf{X}, q_\alpha + \varepsilon) < k(\alpha, n)],$$

where (5.2.6) with $q_\alpha + \varepsilon$ instead of x was used to obtain the set the second equality. From this point, an application of Lemma 5.2.1(i) yields that

$$P[\hat{q}_{\alpha,n} > q_\alpha + \varepsilon] = P[N_n(\mathbf{X}, q_\alpha + \varepsilon) < k(\alpha, n)] \to 0 \quad \text{as } n \to \infty. \quad (5.2.11)$$

On the other hand,

$$[\hat{q}_{\alpha,n} \leq q_\alpha - \varepsilon] = [X_{(k(\alpha,n))} \leq q_\alpha - \varepsilon] = [N_n(\mathbf{X}, q_\alpha - \varepsilon) \geq k(\alpha, n)]$$

where the last equality stems form (5.2.7); applying the second part of Lemma 5.2.1, it follows that

$$\lim_{n\to\infty} P[\hat{q}_{\alpha,n} \leq q_\alpha - \varepsilon] = 0. \quad (5.2.12)$$

Taking the limit as $n \to \infty$ in both sides of (5.2.10), the convergences (5.2.11) and (5.2.12) together imply that $P[|\hat{q}_{\alpha,n} - q_\alpha| > \varepsilon] \to 0$ as $n \to \infty$. □

5.3. Asymptotic Distribution of Sample Quantiles

The objective of this section is to state the result on the asymptotic distribution of the quantile estimators $\hat{q}_{\alpha,n}$. To begin with, notice that the convergence

$$\hat{q}_{\alpha,n} \xrightarrow{\text{P}} q_\alpha$$

established in Theorem 5.2.1 means that $\hat{q}_{\alpha,n}$ is 'near' to q_α when n is 'large'. In the remainder of the chapter the distribution of the sample quantile $\hat{q}_{\alpha,n}$ about q_α will be studied, and it will be shown that the distribution of $\sqrt{n}(\hat{q}_{\alpha,n} - q_\alpha)$ is approximately normal. The basic ideas and notation that will be used to state such a result are briefly discussed in the following remark.

Remark 5.3.1. (i) Given a distribution function, a sequence of random objects $\{W_n\}$ converges in distribution to F if the following property holds:

$$\lim_{n\to\infty} P[W_n \le x] = F(x) \text{ occurs at every point } x \text{ at which } F \text{ is continuous.} \tag{5.3.1}$$

When this statement is true, it is indicated by writing

$$W_n \xrightarrow{\text{d}} F \quad \text{ or, if } Z \sim F, \quad W_n \xrightarrow{\text{d}} Z.$$

(ii) The most basic properties of the notion of convergence in distribution establish that, if a sequence converging in distribution to F is perturbed for other sequence whose influence vanishes as as n goes to ∞, then the new sequence also converges in distribution to F. In formal terms, the following results (a) and (b) hold: Let $\{W_n\}$ and $\{V_n\}$ two sequences of random variables. In this case,

(a) If $V_n \xrightarrow{\text{P}} 0$ and $W_n \xrightarrow{\text{d}} F$, then $V_n + W_n \xrightarrow{\text{d}} F$.

(b) If $V_n \xrightarrow{\text{P}} 1$ and $W_n \xrightarrow{\text{d}} F$, then $V_n W_n \xrightarrow{\text{d}} F$.

(iii) An extension of the two properties in part (ii) is as follows: Let W be a random variable with distribution function F. In this case,

(a) If $V_n \xrightarrow{\text{P}} c \in \mathbb{R}$ and $W_n \xrightarrow{\text{d}} W$, then $V_n + W_n \xrightarrow{\text{d}} c + W$.

(b) If $V_n \xrightarrow{\text{P}} c \in \mathbb{R}$ and $W_n \xrightarrow{\text{d}} W$, then $V_n W_n \xrightarrow{\text{d}} cW$.

(iv) The most important instance of the notion of convergence of distribution is the central limit theorem: If $X_1, X_2, X_3, \ldots$ are independent and identically distributed random variables, with mean μ and variance $\sigma^2 < \infty$, then the sample mean $\overline{X}_n$ satisfies that

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \xrightarrow{\text{d}} Z \sim \mathcal{N}(0,1); \tag{5.3.2}$$

since the distribution function $\Phi(\cdot)$ of the standard normal distribution is continuous at every point, the above convergence is equivalent to

$$P\left[\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \le x\right] = P[Z \le x] = \Phi(x), \quad x \in \mathbb{R}.$$

Also notice that, *via* the properties in part (ii), (5.3.2) is equivalent to

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}} = \sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\text{d}} \sigma Z \sim \mathcal{N}\left(0, \sigma^2\right). \tag{5.3.3}$$

□

The main objective of this section is to establish the following theorem on the asymptotic distribution of a sample quantile $\hat{q}_{\alpha,n}$.

Theorem 5.3.1. Let $\alpha \in (0,1)$ be arbitrary, and let $X_1, X_2, X_3, \ldots$ be a sequence of independent and identically distributed random variables with common distribution function F, which is continuous and has density f satisfying the condition (5.1.2). In this case the sequence $\{\hat{q}_{\alpha,n}\}$ of sample quantiles satisfies that

$$\sqrt{n}[\hat{q}_{\alpha,n} - q_\alpha] \xrightarrow{\text{d}} \frac{\sqrt{\alpha(1-\alpha)}}{f(q_\alpha)} Z, \tag{5.3.4}$$

where $Z \sim \mathcal{N}(0,1)$.

Usually, this result is presented without proof in intermediate level texts, and a major objective of this exposition is to show that Theorem 5.3.4 can be derived putting together the central limit theorem with two simple facts: (i) The connection between the quantile estimators and the binomial distribution implied by (5.2.3)–(5.2.7), and (ii) The basic properties of the idea of convergence in distribution mentioned in Remark 5.3.1. Before proceeding with the technical details, it is convenient to state (5.3.4) in an alternative form.

Remark 5.3.2. Notice that the conclusion (5.3.4) is equivalent to

$$f(q_\alpha)\sqrt{\frac{n}{\alpha(1-\alpha)}}[\hat{q}_{\alpha,n} - q_\alpha] \xrightarrow{\text{d}} Z, \tag{5.3.5}$$

by Remark 5.3.1(iii). Thus, in a more explicit way, the conclusion of Theorem 5.3.1 can be expressed as

$$\begin{aligned}\lim_{n\to\infty} P &\left[a < f(q_\alpha)\sqrt{\frac{n}{\alpha(1-\alpha)}}[\hat{q}_{\alpha,n} - q_\alpha] < b\right] \\ &= \Phi(b) - \Phi(a) \qquad (5.3.6)\\ &= \int_a^b \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz,\end{aligned}$$

where the numbers a and b satisfy $a < b$ but are arbitrary otherwise. The above statement may be extremely useful to establish inferences about the unknown distribution of the data, particularly when the methods of maximum likelihood or moments estimation are not easily implemented; an example of such a situation will be analyzed after completing the proof of Theorem 5.3.1. □

5.4. Preliminary Results

The proof of the asymptotic formula (5.3.4) is rather technical and, by convenience, it has been divided into four steps stated in the following lemmas, where the context is as in the statement of Theorem 5.3.1.

Lemma 5.4.1. For each $x \in \mathbb{R}$, the inequalities

$$\sqrt{n}[\hat{q}_{\alpha,n} - q_\alpha] \leq x \quad \text{and} \quad N_n(\mathbf{X}; q_\alpha + x/\sqrt{n}) \geq k(\alpha, n)$$

are equivalent; see (5.2.1) and (5.2.3) for the specifications of $k(\alpha, n)$ and $N_n(\mathbf{X}; q_\alpha + x/\sqrt{n})$.

Proof. Just recall that $\hat{q}_{\alpha,n} = X_{(k(\alpha,n))}$ and observe that

$$\begin{aligned}
\sqrt{n}[\hat{q}_{\alpha,n} - q_\alpha] \leq x &\iff \hat{q}_{\alpha,n} \leq q_\alpha + x/\sqrt{n} \\
&\iff X_{(k(\alpha,n))} \leq q_\alpha + x/\sqrt{n} \\
&\iff N_n(\mathbf{X}; q_\alpha + x/\sqrt{n}) \geq k(\alpha, n)
\end{aligned}$$

where the relation (5.2.7) with $q_\alpha + x/\sqrt{n}$ instead of x was used to set the last equivalence. □

Lemma 5.4.2. Given $x \in \mathbb{R}$, define the random variable V_n as follows:

$$V_n = N_n(\mathbf{X}; q_\alpha + x/\sqrt{n}) - N_n(\mathbf{X}; q_\alpha) - n[F(q_\alpha + x/\sqrt{n}) - F(q_\alpha)]. \quad (5.4.1)$$

With this notation,

$$\frac{1}{\sqrt{n}} V_n \xrightarrow{\text{P}} 0. \qquad (5.4.2)$$

Proof. First, suppose that $x > 0$ and notice that (5.2.3) leads to

$$\begin{aligned}
N_n(\mathbf{X}; q_\alpha + x/\sqrt{n}) - N_n(\mathbf{X}; q_\alpha) &= \sum_{i=1}^n I[X_i \le q_\alpha + x/\sqrt{n}] - \sum_{i=1}^n I[X_i \le q_\alpha] \\
&= \sum_{i=1}^n (I[X_i \le q_\alpha + x/\sqrt{n}] - I[X_i \le q_\alpha]) \\
&= \sum_{i=1}^n I[q_\alpha < X_i \le q_\alpha + x/\sqrt{n}]
\end{aligned}$$

Observing that each indicator $I[q_\alpha < X_i \le q_\alpha + x/\sqrt{n}]$ has Bernoulli distribution with probability of success equal to $P[q_\alpha < X_i \le q_\alpha + x/\sqrt{n}] = F(q_\alpha + x/\sqrt{n}) - F(q_\alpha)$, and recalling that the variables X_i are independent, it follows that

$$N_n(\mathbf{X}; q_\alpha + x/\sqrt{n}) - N_n(\mathbf{X}; q_\alpha) \sim \mathit{Binomial}\big(n, F(q_\alpha + x/\sqrt{n}\big) - F(q_\alpha)).$$

Consequently, the well-known formulae for the expectation and variance of a binomial distribution imply that the variable V_n in (5.4.1) satisfies that

$$\begin{aligned}
E[V_n] &= 0, \quad \text{and} \\
\text{Var}\,[V_n] &= n(F(q_\alpha + x/\sqrt{n}) - F(q_\alpha))(1 - [F(q_\alpha + x/\sqrt{n}) - F(q_\alpha)]).
\end{aligned}$$

From this point, an application of Chebychev's inequality yields that, for each $\varepsilon > 0$,

$$\begin{aligned}
P\left[\left|\frac{V_n}{\sqrt{n}}\right| > \varepsilon\right] &\le \frac{\text{Var}\,[V_n]}{n\varepsilon^2} \\
&= \frac{(F(q_\alpha + x/\sqrt{n}) - F(q_\alpha))(1 - [F(q_\alpha + x/\sqrt{n}) - F(q_\alpha)])}{\varepsilon^2};
\end{aligned}$$

since $F(q_\alpha + x/\sqrt{n}) - F(q_\alpha) \to 0$ as $n \to \infty$, by the continuity of F, the above display immediately leads to

$$\lim_{n\to\infty} P\left[\left|\frac{V_n}{\sqrt{n}}\right| > \varepsilon\right] = 0$$

establishing (5.4.2). The case $x < 0$ can be handled along similar lines. □

Lemma 5.4.3. Let $\alpha \in (0, 1)$ and $x \in \mathbb{R}$ be arbitrary. In this case,

$$\frac{k(\alpha, n) - nF(q_\alpha + x/\sqrt{n})}{\sqrt{n}} \to -f(q_\alpha)x \quad \text{as } n \to \infty.$$

Proof. Define Δ_n by

$$\Delta_n := \frac{F(q_\alpha + x/\sqrt{n}) - F(q_\alpha)}{x/\sqrt{n}} \tag{5.4.3}$$

and notice that, since F has density f which is continuous at q_α, the function F is differentiable at the quantile q_α and $F'(q_\alpha) = f(q_\alpha)$; hence,

$$f(q_\alpha) = F'(q_\alpha) = \lim_{n\to\infty} \frac{F(q_\alpha + x/\sqrt{n}) - F(q_\alpha)}{x/\sqrt{n}} = \lim_{n\to\infty} \Delta_n. \tag{5.4.4}$$

Now, use (5.4.3) to obtain that $F(q_\alpha + x/\sqrt{n}) = F(q_\alpha) + \Delta_n x/\sqrt{n} = \alpha + \Delta_n x/\sqrt{n}$, where the second equality is due to the relation $F(q_\alpha) = \alpha$. Therefore,

$$nF(q_\alpha + x/\sqrt{n}) = n\alpha + \Delta_n x\sqrt{n}. \tag{5.4.5}$$

On the other hand, the specification of $k(\alpha, n)$ in (5.2.1) yields that $\alpha - 1/n \leq k(\alpha, n) \leq \alpha$, so that there exists

$$\beta_n \in [0, 1/n) \tag{5.4.6}$$

such that $k(\alpha, n)/n = \alpha - \beta_n$, that is,

$$k(\alpha, n) = n\alpha - n\beta_n,$$

an equality that together with (5.4.5) implies that

$$\begin{aligned} \frac{k(\alpha, n) - nF(q_\alpha + x/\sqrt{n})}{\sqrt{n}} &= \frac{n\alpha - n\beta_n - (n\alpha + \Delta_n x\sqrt{n})}{\sqrt{n}} \\ &= -\frac{n\beta_n}{\sqrt{n}} - \Delta_n x; \end{aligned}$$

since the inclusion (5.4.6) yields that $0 \leq n\beta_n/\sqrt{n} \leq 1/\sqrt{n} \to 0$ as $n \to \infty$, taking the limit in the above display it follows that

$$\begin{aligned} \lim_{n\to\infty} \frac{k(\alpha, n) - nF(q_\alpha + x/\sqrt{n})}{\sqrt{n}} &= \lim_{n\to\infty} \left[-\frac{n\beta_n}{\sqrt{n}} - \Delta_n x \right] \\ &= 0 - \lim_{n\to\infty} \Delta_n x, \end{aligned}$$

and the conclusion follows *via* (5.4.4). □

In contrast with the previous lemmas, the next one, which is the last step before the proof of Theorem 5.3.1, is concerned with a general property of the notion of convergence in distribution.

Lemma 5.4.4. Let $\{W_n\}$ be a sequence of random variables converging in distribution to a continuous distribution function $G(x)$, and assume that $G(\cdot)$ is continuous in $\mathbb{R}$. In this case, if $\{y_n\}$ is a convergent sequence of real numbers, say

$$y_n \to y \quad \text{as } n \to \infty,$$

then

$$\lim_{n\to\infty} P[W_n \geq y_n] = 1 - G(y)$$

Remark 5.4.1. When G is a continuous distribution function in $\mathbb{R}$ and $W_n \xrightarrow{\text{d}} G$, the definition of convergence in distribution directly yields that

$$P[W_n \geq y] \to 1 - G(y) \qquad \text{as } n \text{ goes to } \infty;$$

the main conclusion of the above lemma is that such a convergence remains valid when the constant y in the left-hand side is replaced by a convergent sequence $\{y_n\}$. □

Proof. Let a and b two real numbers such that $y \in (a, b)$ and, using that the sequence $\{y_n\}$ converges to y, notice that the inclusion $y_n \in (a, b)$ holds if n is large enough, say for $n > m$. In this case it follows that

$$[W_n > b] \subset [W_n \geq y_n] \subset [W_n > a],$$

and then

$$P[W_n > b] \leq P[W_n \geq y_n] \leq P[W_n > a], \quad n > m.$$

Next, observe that $P[W_n > a] = 1 - P[W_n \leq a] \to 1 - G(a)$ as $n \to \infty$, by the assumption that $W_n \xrightarrow{\text{d}} G$; similarly, $P[W_n > b] = 1 - P[W_n \leq b] \to 1 - G(b)$. Thus, the above display leads to

$$\begin{aligned} 1 - G(b) &= \lim_{n\to\infty} P[W_n > b] \\ &\leq \lim_{n\to\infty} P[W_n \geq y_n] \\ &\leq \lim_{n\to\infty} P[W_n > a] = 1 - G(a). \end{aligned}$$

After taking the limits as $b \searrow y$ and $a \nearrow y$, *via* the continuity of G it follows that

$$1 - G(y) = \lim_{b \searrow y} [1 - G(b)] \leq \lim_{n\to\infty} P[W_n \geq y_n] \leq \lim_{a \nearrow y} [1 - G(a)] = 1 - G(y),$$

completing the proof. □

5.5. Proof of the Asymptotic Normality

The preliminary results established in the previous section will be now used to establish that the limit distribution of a sample quantile is normal.

Proof of Theorem 5.3.1. To begin with, let $x \in \mathbb{R}$ be arbitrary and notice that

$$\begin{aligned}
&\left[\sqrt{n}[\hat{q}_{\alpha,n} - q_\alpha] \leq x\right] \\
&\quad = \left[N_n(\mathbf{X}; q_\alpha + x/\sqrt{n}) \geq k(\alpha, n)\right] \\
&\quad = \left[N_n(\mathbf{X}; q_\alpha + x/\sqrt{n}) - nF(q_\alpha + x/\sqrt{n}) \geq k(\alpha, n) - nF(q_\alpha + x/\sqrt{n})\right]
\end{aligned}$$

where the first equality is due to Lemma 5.4.1, and the second one follows from a substraction in both sides of the inequality. $N_n(\mathbf{X}; q_\alpha + x/\sqrt{n}) \geq k(\alpha, n)$. Now, let V_n be the random variable defined in (5.4.1), and observe that

$$N_n(\mathbf{X}; q_\alpha + x/\sqrt{n}) - nF(q_\alpha + x/\sqrt{n}) = V_n + N_n(\mathbf{X}; q_\alpha)) - nF(q_\alpha)$$

a relation the together with the previous display allows to write

$$\begin{aligned}
&\left[\sqrt{n}[\hat{q}_{\alpha,n} - q_\alpha] \leq x\right] \\
&\quad = \left[V_n + N_n(\mathbf{X}; q_\alpha) - nF(q_\alpha) \geq k(\alpha, n) - nF(q_\alpha + x/\sqrt{n})\right] \\
&\quad = \left[\frac{V_n}{\sqrt{n}} + \frac{N_n(\mathbf{X}; q_\alpha) - nF(q_\alpha)}{\sqrt{n}} \geq \frac{k(\alpha, n) - nF(q_\alpha + x/\sqrt{n})}{\sqrt{n}}\right]
\end{aligned} \tag{5.5.1}$$

Now observe that $N_n(\mathbf{X}; q_\alpha) \sim \textit{Binomial}(n, F(q_\alpha)) = \textit{Binomial}(n, \alpha)$; see (5.1.1) and (5.2.4). Using the central limit theorem, this distributional property yields that

$$\frac{N_n(\mathbf{X}; q_\alpha) - nF(q_\alpha)}{\sqrt{n}} \xrightarrow{\text{d}} \mathcal{N}(0, \alpha(1 - \alpha)).$$

On the other hand, as it was proved in Lemma 5.4.2, $V_n/\sqrt{n} \xrightarrow{\text{P}} 0$, and it follows that

$$W_n := \frac{V_n}{\sqrt{n}} + \frac{N_n(\mathbf{X}; q_\alpha) - nF(q_\alpha)}{\sqrt{n}} \xrightarrow{\text{d}} W \sim \mathcal{N}\left(0, \alpha(1-\alpha)\right); \tag{5.5.2}$$

notice that the distribution function of W is given by

$$G(x) = P[W \leq x] = P\left[\frac{W}{\sqrt{\alpha(1-\alpha)}} \leq \frac{x}{\sqrt{\alpha(1-\alpha)}}\right] = \Phi \left. \quad \frac{x}{\sqrt{\alpha(1-\alpha)}}\right).$$

where $\Phi(\cdot)$ is the distribution function of the standard normal distribution. Also, notice that, with the notation in (5.5.2), the relation (5.5.1) can be equivalently written as

$$\begin{aligned}\left[\sqrt{n}[\hat{q}_{\alpha,n} - q_\alpha] \leq x\right] &= \left[W_n \geq \frac{k(\alpha, n) - nF(q_\alpha + x/\sqrt{n})}{\sqrt{n}}\right] \\ &= [W_n \geq y_n]\end{aligned}$$

where

$$y_n := \frac{k(\alpha, n) - nF(q_\alpha + x/\sqrt{n})}{\sqrt{n}} \to -f(q_\alpha)x,$$

and the convergence follow from Lemma 5.4.3. From this point, an application of Lemma 5.4.4 with $-f(q_\alpha)x$ instead of y allows to conclude that

$$\begin{aligned}\lim_{n\to\infty} P\left[\sqrt{n}[\hat{q}_{\alpha,n} - q_\alpha] \leq x\right] &= \lim_{n\to\infty} P\left[W_n \geq y_n\right] \\ &= P[W > y] = 1 - G(y) = 1 - G(-f(q_\alpha)x),\end{aligned}$$

and the specification of the distribution function $G(\cdot)$ yields that

$$\begin{aligned}\lim_{n\to\infty} P\left[\sqrt{n}[\hat{q}_{\alpha,n} - q_\alpha] \leq x\right] &= 1 - \Phi \left. \quad \frac{-f(q_\alpha)x}{\sqrt{\alpha(1-\alpha)}}\right) \\ &= \Phi \left. \quad \frac{f(q_\alpha)x}{\sqrt{\alpha(1-\alpha)}}\right)\end{aligned}$$

where the second equality is a consequence of the symmetry about 0 of the standard normal distribution. Finally, observe that if $Z \sim \mathcal{N}(0, 1)$, then the distribution function of $\sqrt{\alpha(1-\alpha)}/f(q_\alpha)Z$ is given by

$$P\left[\frac{\sqrt{\alpha(1-\alpha)}}{f(q_\alpha)} Z \leq x\right] = P\left[Z \leq \frac{f(q_\alpha)x}{\sqrt{\alpha(1-\alpha)}}\right] = \Phi \left. \quad \frac{f(q_\alpha)x}{\sqrt{\alpha(1-\alpha)}}\right),$$

and these two last displayed relations together yield that

$$\sqrt{n}[\hat{q}_{\alpha,n} - q_\alpha] \xrightarrow{\mathrm{d}} \frac{\sqrt{\alpha(1-\alpha)}}{f(q_\alpha)} Z \sim \mathcal{N}\left(0, \frac{\alpha(1-\alpha)}{f(q_\alpha)^2}\right),$$

completing the proof. □

Example 5.5.1. Theorem 5.3.1 is quite general and may be particularly useful to establish inferences, specially when the maximum likelihood or the moments methods are not applicable. For instance, consider a sample $X_1, X_2, \ldots, X_n$ of the Cauchy density with center $\theta \in \mathbb{R}$, which is given by

$$f(x;\theta) = \frac{1}{\pi}\,\frac{1}{1+(x-\theta)^2}.$$

For this density, the maximum likelihood estimator of θ does not admit an explicit formula and must be determined numerically for each sample data. On the other hand, this density does not have moments of any order ≥ 1, so the method of moments can not be applied to estimate θ. However, notice that

$$f(\theta+h;\theta) = f(\theta-h;\theta), \quad h \in \mathbb{R},$$

a symmetry property that immediately yields the equality

$$\int_{-\infty}^{\theta} f(x;\theta)\,dx = \frac{1}{2};$$

hence, θ is the median of the distribution, that is,

$$\theta = q_{0.5}(\theta) \equiv q_{0.5}.$$

Thus, the sample median $\hat{q}_{0.5\,n}$ is a consistent estimator of θ, and

$$\sqrt{n}\;[\hat{q}_{0.5\,n} - \theta] \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \frac{\alpha(1-\alpha)}{f(\theta;\theta)^2}\right) = \mathcal{N}\left(0, \frac{\pi^2}{4}\right)$$

This convergence implies that, when n is 'large',

$$.95 \approx P\left[-2 \leq \sqrt{n}\,\frac{\hat{q}_{0.5\,n} - \theta}{\pi/2} \leq 2\right] = P\left[\hat{q}_{0.5\,n} - \frac{\pi}{\sqrt{n}} \leq \theta \leq \hat{q}_{0.5\,n} + \frac{\pi}{\sqrt{n}}\right],$$

and it follows that $[\hat{q}_{0.5\,n} - \pi/\sqrt{n},\;\; \hat{q}_{0.5\,n} + \pi/\sqrt{n}]$ is a confidence interval for θ, whose confidence level is 'approximately' 0.95. □

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.