Econ 240A (1st Half)
Section 4: Fall 2018
Friday, September 21

Fengshi Niu*

Contents

[*] Thanks to Mikkel Sølvsten and Matias D. Cattaneo for letting me use their old notes as an inspiration.

1 Introduction to Statistics

In the next (final) three Sections we will cover the basics of Statistics In particular, we will concentrate on some of the main results in mathematical statistics, namely point estimation, confidence interval estimation and hypothesis testing, which constitute the basis for the analysis discussed in the remaining of the ECON-240A-B sequence.

We begin our discussion by defining one of the key concepts in statistics.

Definition 1.1. *Let $\mathbf{X} = (X_1, X_2, ..., X_N)'$ be an N-dimensional random vector. The random variables $X_1, X_2, ..., X_N$ constitute a **random sample** if they are (mutually) independent and identically distributed (usually denoted by iid).*

Recall from previous sections, that one of the basic properties of independent random variables is that its joint distribution can be recovered by using the marginal distributions of each of the random variables. In particular, we have

$$F_{\mathbf{X}}(\mathbf{x}) = \prod\nolimits_{n=1}^{N} F_{X_n}(x_n) = \prod\nolimits_{n=1}^{N} F(x_n),$$

where in the first and second equalities we used the fact that the random variables are independent and identical distributed, respectively. Moreover, if all the random variables are discrete (continuous) with pmf (pdf) given by $f_{X_n}(x_n)$, then we also have

$$f_{\mathbf{X}}(\mathbf{x}) = \prod\nolimits_{n=1}^{N} f_{X_n}(x_n) = \prod\nolimits_{n=1}^{N} f(x_n).$$

It is important to observe that these results are in fact conclusions of a more general case. This general case is presented in the following exercise.

Exercise 1.1. *Let $\mathbf{X} = (X_1, X_2, ..., X_N)'$ be an N-dimensional random vector and let $\mathbf{X}_n = (X_1, X_2, ..., X_n)$. Show the following results:*

$$\begin{aligned} F_{\mathbf{X}}(\mathbf{x}) &= \prod\nolimits_{n=1}^{N} F_{X_n|\mathbf{X}_{n-1}}(x_n|\,\mathbf{x}_{n-1}), \quad \textit{where } F_{X_1|\mathbf{X}_0}(x_1|\,\mathbf{x}_0) \equiv F_{X_1}(x_1), \quad \textit{and} \\ f_{\mathbf{X}}(\mathbf{x}) &= \prod\nolimits_{n=1}^{N} f_{X_n|\mathbf{X}_{n-1}}(x_n|\,\mathbf{x}_{n-1}), \quad \textit{where } f_{X_1|\mathbf{X}_0}(x_1|\,\mathbf{x}_0) \equiv f_{X_1}(x_1). \end{aligned}$$

Why the random sample case is a particular case of this general setup?

Now we present the general definition of the object that we will be studying in this and the following meetings.

Definition 1.2. *Let $\mathbf{X} = (X_1, X_2, ..., X_N)'$ be a random sample and let $T : \mathbb{R}^N \to \mathbb{R}^K$, $K \in \mathbb{N}$, be a function. The random vector $T(\mathbf{X}) = T(X_1, X_2, ..., X_N)$ is called a **statistic** and its distribution is called the **sampling distribution**.*

Observe that an statistic is nothing more than a function that takes the random variables and construct a new random vector. Typical examples are:

$$\begin{aligned}
\text{Sample Mean} &: \bar{X} = \frac{1}{N}\sum\nolimits_{n=1}^{N} X_n. \\
\text{Sample Variance} &: S^2 = \frac{1}{N-1}\sum\nolimits_{n=1}^{N}\left(X_n - \bar{X}\right)^2. \\
\text{Sample Standard Deviation} &: S = \sqrt{S^2}. \\
\text{Order Statistic} &: X_{(n)} = n\text{-}th \text{ smallest } \{X_n\}_{n=1}^{N}.
\end{aligned}$$

Important results for some of these basic statistics are listed in the following theorem.

Theorem 1.1. *Let* $\mathbf{X} = (X_1, X_2, ..., X_N)'$ *be a random sample from a distribution with mean* μ *and variance* σ^2*, then*

1. $\mathbb{E}\left[\bar{X}\right] = \mu$.
2. $\mathbb{V}ar\left[\bar{X}\right] = \frac{\sigma^2}{N}$.
3. $\mathbb{E}\left[S^2\right] = \sigma^2$.

Proof. Observe that in this proof we only use the fact that $X_1, X_2, ..., X_N$ are *iid*. To see (1), note that

$$\mathbb{E}\left[\bar{X}\right] = \mathbb{E}\left[\frac{1}{N}\sum\nolimits_{n=1}^{N} X_n\right] = \frac{1}{N}\sum\nolimits_{n=1}^{N}\mathbb{E}\left[X_n\right] = \mu.$$

To see (2), note that

$$\mathbb{V}ar\left[\bar{X}\right] = \mathbb{V}ar\left[\frac{1}{N}\sum\nolimits_{n=1}^{N} X_n\right] = \frac{1}{N^2}\sum\nolimits_{n=1}^{N}\mathbb{V}ar\left[X_n\right] = \frac{\sigma^2}{N}.$$

To see (3), note that

$$\begin{aligned}
\mathbb{E}\left[S^2\right] &= \mathbb{E}\left[\frac{1}{N-1}\sum\nolimits_{n=1}^{N}\left(X_n - \bar{X}\right)^2\right] \\
&= \mathbb{E}\left[\frac{1}{N-1}\left(\sum\nolimits_{n=1}^{N} X_n^2 - N\cdot\bar{X}^2\right)\right] \\
&= \frac{1}{N-1}\left(\sum\nolimits_{n=1}^{N}\mathbb{E}\left[X_n^2\right] - N\cdot\mathbb{E}\left[\bar{X}^2\right]\right) \\
&= \frac{1}{N-1}\left(N\cdot\left(\sigma^2+\mu^2\right) - N\cdot\left(\frac{\sigma^2}{N}+\mu^2\right)\right) \\
&= \frac{1}{N-1}\left(N\cdot\sigma^2 - \sigma^2\right) \\
&= \sigma^2,
\end{aligned}$$

which concludes the proof. □

In statistics, average of random variables are usually employed mainly because of one important result: asymptotic theory. In this class we will not cover the results derived from this theory. However, for completeness we present one of the basic result: central limit theorem. Basically, the idea is that for N large enough (technically, when $N \to \infty$), under some regularity conditions, we have that

$$\bar{X} \overset{a}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right),$$

and thus in large samples we can approximate the distribution of a sample average by a normal distribution.

This result motivates the following theorem (at least for large samples).

Theorem 1.2. *If $\{X_n\}_{n=1}^N$ are iid $\mathcal{N}(\mu, \sigma^2)$, then*

1. $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$.
2. $\frac{N-1}{\sigma^2} S^2 \sim \chi^2_{N-1} = Gamma\left(\frac{N-1}{2}, 2\right)$.
3. *$\bar{X}$ and S^2 are statistically independent.*

Before we move to the discussion of Statistical Inference, we present a couple of results for order statistics that are important to note. Recall that the order statistics are the sample values placed in ascending order. That is,

$$X_{(1)} \equiv \min\{X_n\}_{n=1}^N \leq X_{(2)} \leq ... \leq X_{(N)} \equiv \max\{X_n\}_{n=1}^N.$$

In the following exercise we list two important results for one of the order statistics.

Exercise 1.2. *Let $\mathbf{X} = (X_1, X_2, ..., X_N)'$ be a random sample with cdf $F(x)$. Show that:*

1. *The cdf of $X_{(N)}$ is given by*
$$F_{X_{(N)}}(x) = [F(x)]^N.$$
2. *If $\mathbf{X} = (X_1, X_2, ..., X_N)'$ are distributed as iid $\mathcal{U}[0, \theta]$, then*
$$\mathbb{E}\left[X_{(N)}\right] = \frac{N}{N+1} \cdot \theta, \quad \text{and} \quad \mathbb{V}ar\left[X_{(N)}\right] = \frac{N}{(N+2)(N+1)^2} \cdot \theta^2.$$

1.1 Statistical Inference

Statistical Inference is mainly concern with learning something from the data. As we have discussed before, in order to do so we need to impose additional assumptions. In particular, in this course we will assumed that the observed vector $\mathbf{x} = (x_1, x_2, ..., x_N)'$ of data is in fact the realization of some random vector $\mathbf{X} = (X_1, X_2, ..., X_N)'$ whose distribution is known up to a finite number of parameters. Thus, we will be only doing parametric estimation and inference.

Under this hypothetical world, we can (in principle) use the observed data to learn something about the parameters of the statistical model (usually known as Data Generating Process, DGP). Consequently, we can see that Probability and Statistics are in fact two sides of the same coin: in probability we have a known distribution but unknown outcomes, while in statistics we have a known outcome but unknown distribution.

In the next definition we formalize most of our discussion and, in particular, we see how probability theory is in fact an important tool for statistics.

Definition 1.3. *A* ***statistical experiment*** *(or* ***statistical model****) is a triple* $(\Omega, \mathcal{F}, \mathcal{P})$*, where* Ω *is a sample space,* $\mathcal{F}$ *is a* σ*-algebra of events, and* $\mathcal{P}$ *is a collection of probability functions.*

In this class we will concentrate in finite dimensional problems; that is, when the parameter $\theta \in \mathbf{\Theta} \subseteq \mathbb{R}^K$ and $K < \infty$. In words, this means that we know everything up to a finite number of parameters. This kind of models is known as **parametric models**. In particular, we observe from this definition that the triple $(\Omega, \mathcal{B}, P)$ is a probability space for every $\mathbb{P}_\theta \in \mathcal{P}$. Notice also that in previous sections we indexed the elements in $\mathcal{P}$ by $\theta \in \mathbf{\Theta}$. Now we can see that this index is doing more than just "signaling" the element in $\mathcal{P}$: it is in fact the value of the parameter that characterize a particular probability distribution. Consequently, the main objective is to choose the "right" or "best" $\theta \in \mathbf{\Theta}$, based on the data (information) available.

The leading special case for this class of probability models will be: $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}(\mathbb{R})$, and $\mathbf{X} = (X_1, X_2, ..., X_N)'$ a random sample from a parametric model. Thus, we will have

$$\mathcal{F} = \left\{ F(\cdot; \theta) : \theta \in \mathbf{\Theta} \subseteq \mathbb{R}^K \right\} \subseteq \mathcal{P},$$

in words, this is a collection of marginal cdfs completely characterized up to a finite set of values (parameters) $\theta \in \Theta \subseteq \mathbb{R}^K$.

In the rest of this course we will be dealing with this assumptions and we will try to work on three things: (a) point estimation, (b) hypothesis testing, and (c) confidence interval estimation. In the rest of this Section Notes we concentrate on point estimation.

2 Point Estimation

We begin by presenting the formal definition of an estimator. Observe that an estimator is nothing more than a statistic.

Definition 2.1. *Let* $\mathbf{X} = (X_1, X_2, ..., X_N)'$ *be a random sample from a distribution with cdf* $F(\cdot; \theta)$*, where* $\theta \in \mathbf{\Theta} \subseteq \mathbb{R}^K$ *is unknown. A* ***point estimator*** *is any function* $W(\mathbf{X}) = W(X_1, X_2, ..., X_N)$*, where* $W : \mathbb{R}^N \to \mathbb{R}^K$*,* $K \in \mathbb{N}$*. The realized value of an estimator,* $W(\mathbf{x}) = W(x_1, x_2, ..., x_N)$*, is called an* ***estimate****.*

It is interesting to point out that, in principle, the definition does not impose $W(\mathbf{X}) \subseteq \mathbf{\Theta}$. Nevertheless, it is natural to expect that a good estimator would be "close" to the true $\theta \in \mathbf{\Theta}$, usually denoted by θ_0. In this notes we discuss two methods for constructing estimators: methods of moments and maximum likelihood. Both methods are very useful and will appear over and over again in applied work.

However, before we discuss these two particular classes of estimators, it is instructive to discuss a much bigger class that encompasses not only these two classes, but also most of the estimators commonly used in statistics and econometrics. This general class is usually known as **M-estimators** or **extremum estimators** since they are outcome of the minimization of a **criterion function** usually based on both the data and the parameter of interest. This criterion function is in general specially chosen because of its properties (either in finite samples or asymptotically).

As we will show below, M-estimation is a general procedure which encompasses most of the usually employed estimators. Suppose we have a (random) sample of size N denoted by $\mathbf{X} = (X_1, X_2, ..., X_N)'$ and we are interested in the unknown population parameter $\theta \in \mathbf{\Theta} \subseteq \mathbb{R}^K$, where we know that $X_n \overset{iid}{\sim} F(\cdot;\theta)$ for all n. Recall from previous discussions that most population parameters of interest can be in fact written as the solution of the following problem:

$$\theta_0 = \arg\min_{\theta \in \mathbf{\Theta}} \mathbb{E}\left[\rho\left(X_n;\theta\right)\right],$$

for an appropriate choice of the function $\rho(\cdot;\cdot)$. This insight motivates the the basic idea in this general framework: an estimator can be defined as the solution of the following (sample analogue) problem

$$\hat{\theta} = \arg\min_{\theta \in \mathbf{\Theta}} \frac{1}{N} \sum\nolimits_{n=1}^{N} \rho\left(x_n;\theta\right),$$

where now the (vector-valued) function $\rho(\cdot;\cdot)$ defines the estimation procedure and is usually known as the (empirical) criterion function. This function is usually chosen because of its asymptotic properties. Although we do not discuss here such properties, recall from undergraduate statistics that the Law of Large Numbers suggests that for large sample sizes

$$\frac{1}{N} \sum\nolimits_{n=1}^{N} \rho\left(x_n;\theta\right) \approx \mathbb{E}\left[\rho\left(X_n;\theta\right)\right],$$

and hence under some regularity conditions we may expect that

$$\hat{\theta} = \arg\min_{\theta \in \mathbf{\Theta}} \frac{1}{N} \sum\nolimits_{n=1}^{N} \rho\left(x_n;\theta\right) \approx \theta_0 = \arg\min_{\theta \in \mathbf{\Theta}} \mathbb{E}\left[\rho\left(X_n;\theta\right)\right].$$

Recall that we have already solved the limiting problem for some particular cases, such as mean or median. It is common practice in statistics to replace population objects by their sample analogous. This idea has received different names in different places such as plug-in principle or sample analogy principle. In this class we will not evaluate the properties of different procedures (which usually is done by means of asymptotics) but rather we will concentrate on two particular well-known classes of estimators.

Before discuss these estimators, we present a closely related estimation procedure known as Z-estimation. Note that the FOC of the M-estimation problem (when it exists) is given by

$$\frac{1}{N}\sum\nolimits_{n=1}^{N}\frac{\partial}{\partial\theta}\rho\left(x_n;\theta\right)\Big|_{\hat{\theta}}\equiv\frac{1}{N}\sum\nolimits_{n=1}^{N}\psi\left(x_n;\hat{\theta}\right)=\mathbf{0}.$$

Hence an alternative estimation procedure, closely related to M-estimation, could be defined in terms of the function $\psi\left(\cdot;\cdot\right)$ and its associated zero given the data. This procedure is called **Z-estimation**.

To close this introduction, now we present the two leading examples that we will discuss in this class. For simplicity, we discuss the scalar parameter case, but observe that these procedures generalize in the obvious way to finite vector-valued parameters.

Example 2.1. *(*Method of Moments Estimation*) Let $\mathbf{X}=\left(X_1,X_2,...,X_N\right)'$ be a random sample from a distribution with cdf $F\left(\cdot;\theta\right)$, where $\theta\in\Theta\subseteq\mathbb{R}$ is unknown. The method of* ***moments estimation*** *is an M-estimation and a Z-estimation procedure where*

$$\rho\left(x_n;\theta\right)=\left[x_n-\mu\left(\theta\right)\right]^2,\qquad\textit{and}\qquad\psi\left(x_n;\theta\right)=-2\cdot\left[x_n-\mu\left(\theta\right)\right]\cdot\frac{d}{d\theta}\mu\left(\theta\right).$$

Example 2.2. *(*Maximum Likelihood Estimation*) Let $\mathbf{X}=\left(X_1,X_2,...,X_N\right)'$ be a random sample from a distribution with cdf $F\left(\cdot;\theta\right)$, where $\theta\in\Theta\subseteq\mathbb{R}$ is unknown, and has density $f\left(\cdot;\theta\right)$. The method of* ***maximum likelihood*** *is an M-estimation and a Z-estimation procedure where*

$$\rho\left(x_n;\theta\right)=-\log f\left(x_n;\theta\right),\qquad\textit{and}\qquad\psi\left(\mathbf{x}_n;\theta\right)=-\frac{d}{d\theta}\log f\left(x_n;\theta\right).$$

2.1 Method of Moments Estimation

We begin by providing a formal definition of the method for an scalar parameter.

Definition 2.2. *Let $\mathbf{X}=\left(X_1,X_2,...,X_N\right)'$ be a random sample from a distribution with cdf $F\left(\cdot;\theta\right)$ with finite first moments, where $\theta\in\Theta\subseteq\mathbb{R}$ is unknown. A* ***method of moments estimator****, denoted by $\hat{\theta}_{MM}$, solves the* ***sample analogue of the moment equation****; that is*

$$\mu\left(\hat{\theta}_{MM}\right)\equiv\mathbb{E}_{\hat{\theta}_{MM}}\left[X_n\right]=\frac{1}{N}\sum\nolimits_{n=1}^{N}X_n\equiv\bar{X}\equiv\widehat{\mathbb{E}}_{\theta}\left[X_n\right].$$

Observe that in principle, existence and uniqueness of $\hat{\theta}_{MM}$ is implicitly assumed in the definition and thus in some situations we may encounter either nonexistence or multiple solutions to the moment equation. As it is usual the case, estimators will be denoted by a hat over the corresponding parameter. The notation $\widehat{\mathbb{E}}_{\theta}\left[\cdot\right]$ reinforces the idea of sample analogue of the population expectation, a natural characterization of the Method of Moments estimator. Using "sample analogues" as estimators for population parameters is a well-established method in statistics and econometrics.

The general procedure to compute the method of moments estimator is as follows:

1. Find the population moments.
2. Solve the sample analogue of the moment equation. (*Remember to check for special cases!*)

Now we present a very simple example.

Example 2.3. *(*Binomial Distribution*) Let* $\mathbf{X} = (X_1, X_2, ..., X_N)'$ *be a random sample from a Binomial distribution, that is* $X_n \sim$ *iid Binomial* (m, p)*. Assuming that* m *is known, the method of moments estimator of* p*, denoted* $\hat{p}_{MM}$*, is given by*

$$\mathbb{E}_{\hat{\theta}_{MM}}[X_n] = m \cdot \hat{p}_{MM} = \frac{1}{N}\sum\nolimits_{n=1}^{N} X_n \equiv \widehat{\mathbb{E}}_{\theta}[X_n] \Longrightarrow \hat{p}_{MM} = \frac{1}{m} \cdot \bar{X}.$$

In the next example we present the result for the continuous uniform distribution and then we introduce a similar result for the discrete uniform distribution.

Example 2.4. *(*Uniform Distribution*) Let* $\mathbf{X} = (X_1, X_2, ..., X_N)'$ *be a random sample from a Uniform distribution, that is* $X_n \sim$ *iid* $\mathcal{U}[0, \theta]$*,* $\theta \in \Theta = \mathbb{R}_{++}$*. The method of moments estimator of* θ*, denoted* $\hat{\theta}_{MM}$*,is given by*

$$\mathbb{E}_{\hat{\theta}_{MM}}[X_n] = \frac{\hat{\theta}_{MM}}{2} = \frac{1}{N}\sum\nolimits_{n=1}^{N} X_n \equiv \widehat{\mathbb{E}}_{\theta}[X_n] \Longrightarrow \hat{\theta}_{MM} = 2 \cdot \bar{X}.$$

It is important to note that this estimator has at least one bad property: it may be the case that $\hat{\theta}_{MM} \notin \Theta$*.*

In the next example we present the moment estimator for the discrete uniform distribution. This example highlights two additional important things: (*i*) it shows that it could be the case that $\hat{\theta}_{MM} \notin \Theta$ (just as it was in the continuous case), and (*ii*) it allows us to practice one more time a change of variables argument.

Example 2.5. *(*Discrete Uniform Distribution*) Let* $\mathbf{X} = (X_1, X_2, ..., X_N)'$ *be a random sample from a discrete Uniform distribution, with support* $\{c, ..., d\}$*,* $c, d \in \mathbb{N}$ *and* $c < d$*. It is very easy to see that*

$$f_{X_n}(x) = \frac{1}{d-c+1} \cdot \mathbb{I}\{x \in \{c, ..., d\}\},$$

and thus we have

$$\begin{aligned}
\mathbb{E}[X_n] &= \sum\nolimits_{x=c}^{d} x \frac{1}{d-c+1} \\
&= \frac{1}{d-c+1}\sum\nolimits_{x=c}^{d}(x+c-c) \\
&= c + \frac{1}{d-c+1}\sum\nolimits_{x=c}^{d}(x-c) \\
\text{(change vars. } [w = x-c]\text{)} \qquad &= c + \frac{1}{d-c+1}\sum\nolimits_{w=0}^{d-c} w \\
&= c + \frac{1}{d-c+1}\frac{(d-c)\cdot(d-c+1)}{2} \\
&= \frac{c+d}{2},
\end{aligned}$$

where in the last line we used the fact that

$$\sum\nolimits_{n=0}^{N} n = \frac{N \cdot (N+1)}{2}.$$

Now, for simplicity we assume that $c = 0$, and thus we have the method of moments estimator of d, denoted $\hat{d}_{MM}$, given by

$$\mathbb{E}_{\hat{\theta}_{MM}}\left[X_n\right] = \frac{\hat{d}_{MM}}{2} = \frac{1}{N}\sum\nolimits_{n=1}^{N} X_n \equiv \widehat{\mathbb{E}}_{\theta}\left[X_n\right] \Longrightarrow \hat{d}_{MM} = 2 \cdot \bar{X}.$$

Observe that this is, in fact, the estimator for the continuous case as well.

Observe that, as we mentioned before, the method of moments relies on the key assumption that moments exist. This assumption may be in some cases too restrictive. However, even when moments do not exist for the random variable, we can use the Method of Moments estimator with the following generalization:

Definition 2.3. *Let $\mathbf{X} = (X_1, X_2, ..., X_N)'$ be a random sample from a distribution with cdf $F(\cdot;\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$ is unknown. Let $g : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}_\theta\left[\left|g\left(X_n\right)\right|\right] < \infty$. Then, a **method of moments estimator**, denoted by $\hat{\theta}_{MM}$, solves the **sample analogue of the moment equation**; that is*

$$\mu\left(\hat{\theta}_{MM}\right) \equiv \mathbb{E}_{\hat{\theta}_{MM}}\left[g\left(X_n\right)\right] = \frac{1}{N}\sum\nolimits_{n=1}^{N} g\left(X_n\right) \equiv \widehat{\mathbb{E}}_{\theta}\left[g\left(X_n\right)\right].$$

In the next exercise we discuss an example of this generalization.

Exercise 2.1. *Let $\mathbf{X} = (X_1, X_2, ..., X_N)'$ be a random sample from a continuous distribution with pdf given by*

$$f_{X_n}\left(x_n;\theta\right) = \theta \cdot x_n^{-2} \cdot \mathbb{I}\left\{x_n > \theta\right\},$$

where $\theta \in \Theta = \mathbb{R}_{++}$.

1. *Identify the family and particular parametrization, and determine whether it belongs to the exponential family.*

2. *Find the method of moments estimator, denoted $\hat{\theta}_{MM}$.*

Another way of generalizing method of moments estimation is by allowing multiple parameters. This generalization is straightforward and is given in the following definition.

Definition 2.4. *Let $\mathbf{X} = (X_1, X_2, ..., X_N)'$ be a random sample from a distribution with cdf $F(\cdot;\theta)$, where $\theta \in \boldsymbol{\Theta} \subseteq \mathbb{R}^K$ is unknown. Then, a **method of moments estimator**, denoted by $\hat{\theta}_{MM}$, solves the **sample analogue of the moment equations**; that is*

$$\mu_k'\left(\hat{\theta}_{MM}\right) \equiv \mathbb{E}_{\hat{\theta}_{MM}}\left[X_n^k\right] = \frac{1}{N}\sum\nolimits_{n=1}^{N} X_n^k \equiv \widehat{\mathbb{E}}_{\theta}\left[X_n^k\right],$$

for $k = 1, 2, ..., K$.

In the next two exercises we practice this method of estimation for the Binomial distribution and for (our close friend now...) the Gamma distribution.

Exercise 2.2. *Let $\mathbf{X} = (X_1, X_2, ..., X_N)'$ be a random sample from a Binomial distribution with parameters $\theta = (m, p) \in \mathbf{\Theta} = \mathbb{N} \times [0, 1]$. Find the method of moment estimator $\hat{\theta}_{MM}$.*

Exercise 2.3. *Let $\mathbf{X} = (X_1, X_2, ..., X_N)'$ be a random sample from a Gamma distribution with parameters $\theta = (\alpha, \beta) \in \mathbf{\Theta} = \mathbb{R}^2_{++}$. Find the method of moment estimator $\hat{\theta}_{MM}$.*

2.2 Maximum Likelihood Estimation

To begin our discussion of the method of Maximum Likelihood estimation, we define the concept of likelihood and log-likelihood. This is done in the following definition.

Definition 2.5. *(*LIKELIHOOD AND LOG-LIKELIHOOD*) Let $\mathbf{X} = (X_1, X_2, ..., X_N)'$ be a random sample from a distribution with pmf/pdf $f(\cdot; \theta)$, where $\theta \in \mathbf{\Theta} \subseteq \mathbb{R}^K$ is unknown, for some $K \in \mathbb{N}$. The* ***likelihood function*** *is defined as $L(\cdot; \mathbf{x}) : \mathbf{\Theta} \to \mathbb{R}_+$, where*

$$L(\theta; \mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}; \theta) = \prod_{n=1}^{N} f_{X_n}(x_n; \theta).$$

The ***log-likelihood function*** *is defined as $l(\cdot | \mathbf{x}) : \mathbf{\Theta} \to \mathbb{R}$, where*

$$l(\theta; \mathbf{x}) = \log L(\theta; \mathbf{x}) = \log f_{\mathbf{X}}(\mathbf{x}; \theta) = \log \left\{ \prod_{n=1}^{N} f_{X_n}(x_n; \theta) \right\} = \sum_{n=1}^{N} \log f_{X_n}(x_n; \theta)$$

Recall that in this class we will deal only with parametric specifications and thus we will always know $L(\theta | \mathbf{x})$ up to a finite number of parameters collected in $\theta \in \mathbf{\Theta} \subseteq \mathbb{R}^K$, for some $K \in \mathbb{N}$. As it can be seen, both $L(\theta; \mathbf{x})$ and $l(\theta; \mathbf{x})$ are functions of the parameters given the realization of the random sample $\mathbf{X} = \mathbf{x}$. In turn, as we mentioned in the introduction of this section, $f_{\mathbf{X}}(\mathbf{x}; \theta)$ is the joint distribution of the random sample (which is a function of the data), given θ.

In the next definition we present the Maximum Likelihood Estimator.

Definition 2.6. *(*MAXIMUM LIKELIHOOD ESTIMATION*) The* ***maximum likelihood estimate*** *is given by*

$$\hat{\theta}_{ML} = \hat{\theta}_{ML}(\mathbf{x}) = \arg\max_{\theta \in \mathbf{\Theta}} L(\theta; \mathbf{x}) = \arg\max_{\theta \in \mathbf{\Theta}} l(\theta; \mathbf{x})$$

The statistic $\hat{\theta}_{ML}(\mathbf{X})$ is called the ***maximum likelihood estimator****, which is a random variable since it is a function of the random sample.*

The MLE is probably the best estimator available for most purposes since under fairly general conditions, it behaves "very good" in large samples. It can be shown that the MLE is also a sample analog estimator, however, the details on this are beyond the scope of this class. This will be discussed in some detail in ECON-240B.

The general procedure to compute MLE is as follows:

1. Find the likelihood function $L(\theta; \mathbf{x})$.

2. Maximize the likelihood function with respect to $\theta \in \mathbf{\Theta}$. (*Remember to check second order conditions and boundary conditions!*)

In particular, when $L(\theta; \mathbf{x})$ is differentiable (in θ) and the solution to the MLE problem is interior (in $\mathbf{\Theta}$), then $\hat{\theta}_{ML}$ solves the **likelihood equations** given by

$$\left.\frac{\partial}{\partial \theta} L(\theta; \mathbf{x})\right|_{\theta=\hat{\theta}_{ML}} = 0$$

Alternatively, when $L(\theta; \mathbf{x})$ is not differentiable (in θ) and/or the solution to the MLE problem is not interior (in $\mathbf{\Theta}$), we need to solve it by inspection or by using other (usually ad-hoc) methods.

Below we present some examples of cases when MLE is more difficult to obtain. But before presenting difficult cases, we discuss an easy one in the next example.

Example 2.6. *(*Binomial Distribution*) Let $\mathbf{X} = (X_1, X_2, ..., X_N)'$ be a random sample from a Binomial distribution, that is $X_n \sim$ iid Binomial(m, p). Suppose that m is known and let $\theta = p \in \mathbf{\Theta} = [0, 1]$. The likelihood function is*

$$\begin{aligned}
L(\theta; \mathbf{x}) &= f_{\mathbf{X}}(\mathbf{x}; \theta) \\
&= \prod_{n=1}^{N} f_{X_n}(x_n | \theta) \\
&= \prod_{n=1}^{N} \binom{m}{x_n} \cdot \theta^{x_n} \cdot (1-\theta)^{m-x_n} \cdot \mathbb{I}\{x_n \in \{0, 1, ..., m\}\} \\
&= \theta^{\sum_{n=1}^{N} x_n} \cdot (1-\theta)^{mN - \sum_{n=1}^{N} x_n} \cdot \prod_{n=1}^{N} \left\{ \binom{m}{x_n} \cdot \mathbb{I}\{x_n \in \{0, 1, ..., m\}\} \right\},
\end{aligned}$$

and the log-likelihood function is

$$\begin{aligned}
l(\theta; \mathbf{x}) &= \log L(\theta; \mathbf{x}) \\
&= \left(\sum_{n=1}^{N} x_n\right) \cdot \log \theta + \left[m \cdot N - \left(\sum_{n=1}^{N} x_n\right)\right] \cdot \log(1-\theta) + \log \left\{ \prod_{n=1}^{N} \binom{m}{x_n} \cdot \mathbb{I}\{x_n \in \{0, 1, ..., m\}\} \right\}
\end{aligned}$$

for $\theta \in (0, 1)$.

First, if $\sum_{n=1}^{N} x_n \notin \{0, m \cdot N\}$, then observe that since $l(\theta; \mathbf{x})$ is continuously differentiable function of θ and this density does not have parameter-dependence support, we just compute FOC and SOC, which are

$$\begin{aligned}
\frac{\partial}{\partial \theta} l(\theta; \mathbf{x}) &= \left(\sum_{n=1}^{N} x_n\right) \cdot \frac{1}{\theta} - \left[m \cdot N - \left(\sum_{n=1}^{N} x_n\right)\right] \cdot \frac{1}{1-\theta} = 0 \Longrightarrow \hat{\theta}_{ML}(\mathbf{x}) = \frac{1}{m}\bar{x}, \\
\frac{\partial^2}{\partial \theta^2} l(\theta; \mathbf{x}) &= -\left(\sum_{n=1}^{N} x_n\right) \cdot \frac{1}{\theta^2} - \left[m \cdot N - \left(\sum_{n=1}^{N} x_n\right)\right] \cdot \frac{1}{(1-\theta)^2} < 0,
\end{aligned}$$

and therefore we have

$$\hat{\theta}_{ML} = \frac{1}{m} \cdot \bar{X}.$$

Second, if $\sum_{n=1}^{N} x_n \in \{0, m \cdot N\}$, we cannot use the log-likelihood (why?), but rather we need to use the likelihood function. Observe that we have

$$\begin{aligned} L(\theta; \mathbf{x}) &= (1-\theta)^{m \cdot N} \cdot \prod\nolimits_{n=1}^{N} \left\{ \binom{m}{x_n} \cdot \mathbb{I}\{x_n \in \{0,1,...,m\}\} \right\}, \quad \textit{if } \sum\nolimits_{n=1}^{N} x_n = 0, \\ L(\theta; \mathbf{x}) &= \theta^{\sum_{n=1}^{N} x_n} \cdot \prod\nolimits_{n=1}^{N} \left\{ \binom{m}{x_n} \cdot \mathbb{I}\{x_n \in \{0,1,...,m\}\} \right\}, \quad \textit{if } \sum\nolimits_{n=1}^{N} x_n = m \cdot N, \end{aligned}$$

and observe that in both cases the solutions to the optimization problem agree with the MLE previously derived.

In the next example we present the MLE for the continuous uniform distribution. Observe that this is the first example where the distribution is not differentiable and thus other methods have to be employed in order to compute the MLE.

Example 2.7. *(*Uniform Distribution*) Let $\mathbf{X} = (X_1, X_2, ..., X_N)'$ be a random sample from a Uniform distribution, that is $X_n \sim$ iid $\mathcal{U}[0,\theta]$, $\theta \in \Theta = \mathbb{R}_{++}$. The likelihood function is*

$$\begin{aligned} L(\theta; \mathbf{x}) &= f_{\mathbf{X}}(\mathbf{x}; \theta) \\ &= \prod\nolimits_{n=1}^{N} f_{X_n}(x_n | \theta) \\ &= \prod\nolimits_{n=1}^{N} \frac{1}{\theta} \cdot \mathbb{I}\{0 \le x_n \le \theta\} \\ &= \frac{1}{\theta^N} \cdot \mathbb{I}\left\{0 \le x_{(1)}\right\} \cdot \mathbb{I}\left\{x_{(N)} \le \theta\right\}, \end{aligned}$$

and the log-likelihood function is

$$l(\theta; \mathbf{x}) = \log L(\theta; \mathbf{x}) = -N \cdot \log \theta + \log \left\{ \mathbb{I}\left\{0 \le x_{(1)}\right\} \cdot \mathbb{I}\left\{x_{(N)} \le \theta\right\} \right\}.$$

Observe that these functions are clearly not differentiable in θ, since the support of the distribution depends on θ. However, we can see that

$$L(\theta; \mathbf{x}) = \begin{cases} \frac{1}{\theta^N} & \textit{if } x_{(N)} \le \theta \\ 0 & \textit{if } x_{(N)} > \theta \end{cases},$$

and thus it follows that $\hat{\theta}_{ML}(\mathbf{x}) = x_{(N)}$ and thus the ML estimator is given by

$$\hat{\theta}_{ML}(\mathbf{X}) = \hat{\theta}_{ML} = X_{(N)}.$$

In the next exercise we present another example of MLE under parameter-dependence support.

Exercise 2.4. *Let* $\mathbf{X} = (X_1, X_2, ..., X_N)'$ *be a random sample from a continuous distribution with pdf given by*

$$f_{X_n}(x_n;\theta) = \theta \cdot x_n^{-2} \cdot \mathbb{I}\{x_n > \theta\},$$

where $\theta \in \Theta = \mathbb{R}_{++}$*. Find the maximum likelihood estimator, denoted* $\hat{\theta}_{ML}$*. Does it exists? Under what conditions does the MLE exist?*

Finally to close the MLE discussion we present the well known example of the Normal distribution. This example illustrates the method of maximum likelihood for the case of two parameters ($K = 2$) and also shows how to proceed when it is necessary to check the boundary conditions.

Example 2.8. *(*Normal Distribution*) Let* $\mathbf{X} = (X_1, X_2, ..., X_N)'$ *be a random sample from a Normal distribution, that is* $X_n \sim iid\ \mathcal{N}[\mu, \sigma^2]$*,* $\theta = (\mu, \sigma^2) \in \mathbf{\Theta} = \mathbb{R} \times \mathbb{R}_{++}$*. The likelihood function is*

$$\begin{aligned} L(\theta;\mathbf{x}) &= f_{\mathbf{X}}(\mathbf{x};\theta) \\ &= \prod\nolimits_{n=1}^{N} f_{X_n}(x_n;\theta) \\ &= \prod\nolimits_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{1}{2}\frac{(x_n-\mu)^2}{\sigma^2}\right\} \\ &= \left(2\pi\sigma^2\right)^{-\frac{N}{2}} \cdot \exp\left\{-\frac{1}{2\sigma^2}\sum\nolimits_{n=1}^{N}(x_n-\mu)^2\right\}, \end{aligned}$$

and the log-likelihood function is

$$\begin{aligned} l(\theta;\mathbf{x}) &= \log L(\theta;\mathbf{x}) \\ &= -\frac{N}{2}\cdot\log(2\pi) - \frac{N}{2}\cdot\log\sigma^2 - \frac{1}{2\sigma^2}\sum\nolimits_{n=1}^{N}(x_n-\mu)^2. \end{aligned}$$

Observe that since $l(\theta;\mathbf{x})$ *is continuously differentiable function of* θ *and this density does not have parameter-dependence support, we just compute FOC and SOC. So we have for FOC*

$$\begin{aligned} \frac{\partial}{\partial\mu} l(\theta;\mathbf{x}) &= -\frac{1}{\sigma^2}\sum\nolimits_{n=1}^{N}(x_n-\mu)^2 = 0 \Longrightarrow \hat{\mu}_{ML} = \bar{X}, \quad \textit{and} \\ \frac{\partial}{\partial\sigma^2} l(\theta;\mathbf{x}) &= -\frac{N}{2}\cdot\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum\nolimits_{n=1}^{N}(x_n-\mu)^2 \Longrightarrow \hat{\sigma}^2_{ML} = \frac{1}{N}\sum\nolimits_{n=1}^{N}(x_n-\hat{\mu}_{ML})^2. \end{aligned}$$

Now we check SOC (a bit of algebra will give...)

$$\left.\frac{\partial^2}{\partial\theta\partial\theta'} l(\theta;\mathbf{x})\right|_{\theta=\hat{\theta}_{ML}} = \begin{bmatrix} -\hat{\sigma}_{ML}^{-2}\cdot N & 0 \\ 0 & -2\hat{\sigma}_{ML}^{-4}\cdot N \end{bmatrix},$$

which is negative definite. Observe that since we are maximizing over an open set, we need to check for boundary conditions as well. That is, in this case we see

$$\lim_{\mu\to+\infty} L(\theta;\mathbf{x}) = \lim_{\mu\to-\infty} L(\theta;\mathbf{x}) = 0, \quad \textit{and} \quad \lim_{\sigma^2\to 0} L(\theta;\mathbf{x}) = \lim_{\sigma^2\to+\infty} L(\theta;\mathbf{x}) = 0,$$

and since $L\left(\hat{\theta}_{ML};\mathbf{x}\right) = L\left(\hat{\mu}_{ML}, \hat{\sigma}^2_{ML};\mathbf{x}\right) > 0$*, we conclude that in fact* $\hat{\theta}_{ML}$ *is the desired optimum.*