PHIL6334/ECON6614-Lecture Notes 2: Estimation

Aris Spanos [Spring 2019]

1 Introduction

1.1 The frequentist approach to statistical inference

The frequentist approach is different from the Bayesian in terms of the following features:

[a] The **interpretation of probability** is *frequentist*: the relative frequencies associated with the long-run metaphor (in a hypothetical set up) reflect the corresponding probabilities; the formal link comes in the form of the Strong Law of Large Numbers (SLLN).

[b] The chance regularities exhibited by data $\mathbf{x}_0$:=$(x_1, x_2, \ldots, x_n)$ constitute the **only relevant statistical information** for selecting the probabilistic assumptions comprising the statistical model. The probabilistic information that aims to account for all the chance regularities in data $\mathbf{x}_0$ is specified in the form of a statistical model:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\},\ \mathbf{x}\in\mathbb{R}^n_X,$$

where $\boldsymbol{\Theta}$ denotes the parameter space, $\mathbb{R}^n_X$ the sample space, and $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}\in\mathbb{R}^n_X$ the (joint) distribution of the sample. Equivalently, the data $\mathbf{x}_0$ are viewed as a 'typical realization' of a stochastic mechanism described by a statistical model,

[c] The **primary aim** of the frequentist approach is to *learn from data* about the "true" statistical data GM:

$$\mathcal{M}_{\boldsymbol{\theta}^*}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}^*)\},\ \mathbf{x}\in\mathbb{R}^n_X.$$

The expression "$\boldsymbol{\theta}^*$ denotes the true value of $\boldsymbol{\theta}$" is a shorthand for saying that "data $\mathbf{x}_0$ constitute a typical realization of the sample $\mathbf{X}$ with distribution $f(\mathbf{x};\boldsymbol{\theta}^*)$".

1.2 Basic frequentist concepts and distinctions

Fisher (1922) recast the modern frequentist approach to statistics by introducing numerous new concepts and ideas in an attempt to addressed several confusions permeating Karl Pearson's approach to statistics. To render the preliminary discussion the basic concepts and important distinctions in frequentist inference less abstract, let us focus on a particular example.

Consider the **simple Bernoulli model** specified by:

$$\mathcal{M}_{\theta}(\mathbf{x}):\ X_k\backsim\mathsf{BerIID}(\theta,\theta(1-\theta)),\ x_k=0,1,\ \theta\in[0,1],\ k=1,2,\ldots n,... \qquad (1)$$

where '$\mathsf{BerIID}(\theta,\theta(1-\theta))$' stands for Bernoulli, Independent and Identically Distributed (IID), with mean θ and variance $\theta(1-\theta)$, k is an index that denotes the order of the sample.

The basic elements of the simple Bernoulli model are:

(i) the parameter space: $\Theta{:=}[1,0]$, i.e. $0 \leq \theta \leq 1$,

(ii) the sample space: $\mathbb{R}^n_X{:=}\{0,1\}^n{=}\{0,1\}{\times}\{0,1\}{\times}\cdots{\times}\{0,1\}$.

Sample vs. sample realization. A crucial distinction is that between a *sample* $\mathbf{X}{:=}(X_1,\ X_2,...,X_n)$, a set of random variables, and a *sample realization* $\mathbf{x}_0{:=}(x_1,x_2,...x_n)$ that represents just one point in $\mathbb{R}^n_X$.

For the Bernoulli model a *sample realization* $\mathbf{x}_0$, say $n{=}30$, would look like:

$$\mathbf{x}_0{:=}(0,0,1,0,1,1,0,0,1,0,0,0,1,1,0,1,0,1,0,0,1,0,1,0,0,0,1,1,0,0).$$

The distinction between a sample $\mathbf{X}$ and one realization $\mathbf{x}_0$ (out of many, often infinite possible ones) takes the form:

$$\begin{array}{rrccccccccc} \textsf{Sample:} & \mathbf{X}{:=} & (X_1, & X_2, & X_3, & X_4, & X_5, & X_6, & \cdots & X_{30}) \\ & & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \cdots & \downarrow \\ \textsf{Sample realization:} & \mathbf{x}_0{:=} & (0 & 0 & 1 & 0 & 1 & 1 & \cdots & 0) \end{array}$$

Distribution of the sample vs. the likelihood function. As mentioned above, the distribution of the sample $f(\mathbf{x};\boldsymbol{\theta}){=}\theta^Y(1{-}\theta)^{n-Y},\ \mathbf{x}{\in}\{0,1\}^n$, where $Y{=}\sum_{i=1}^n X_i$ is determined by the assumptions of a statistical model; BerIID. This in turn determines the *likelihood function* via:

$$L(\boldsymbol{\theta};\mathbf{x}_0){=}c(\mathbf{x}_0){\cdot}f(\mathbf{x}_0;\boldsymbol{\theta}){=}c(\mathbf{x}_0){\cdot}\theta^y(1{-}\theta)^{n-y},\ \forall\boldsymbol{\theta}{\in}\Theta,$$

where $f(\mathbf{x}_0;\boldsymbol{\theta})$ is evaluated at the observed data point $\mathbf{x}_0$. We often choose:

$$c(\mathbf{x}_0){=}(\tfrac{1}{L(\widehat{\boldsymbol{\theta}};\mathbf{x}_0)}) \rightarrow \widehat{L}(\boldsymbol{\theta};\mathbf{x}_0){=}\tfrac{L(\boldsymbol{\theta};\mathbf{x}_0)}{L(\widehat{\boldsymbol{\theta}};\mathbf{x}_0)}.$$

Example. For the *simple Bernoulli model*, assume that in a sample of size $n{=}20$, the observed Y is $y{=}\sum_{k=1}^n x_k{=}17$; see fig. 1 for $f(y;\theta),\ y{=}1,2,..,n$.

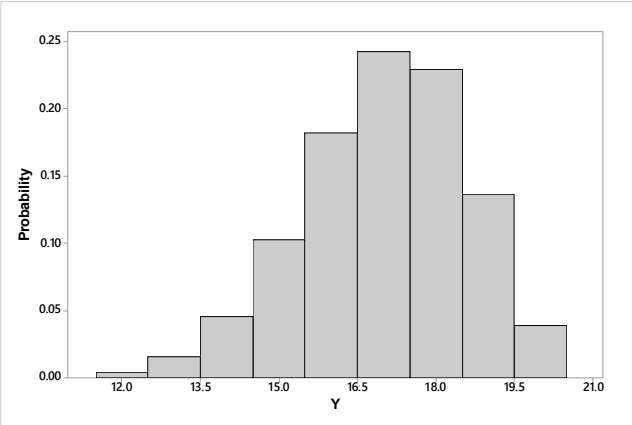


Fig. 1: $Y{\backsim}\mathsf{Bin}(n\theta,\ n\theta(1{-}\theta))$, $n{=}20,\ \theta{=}.85$

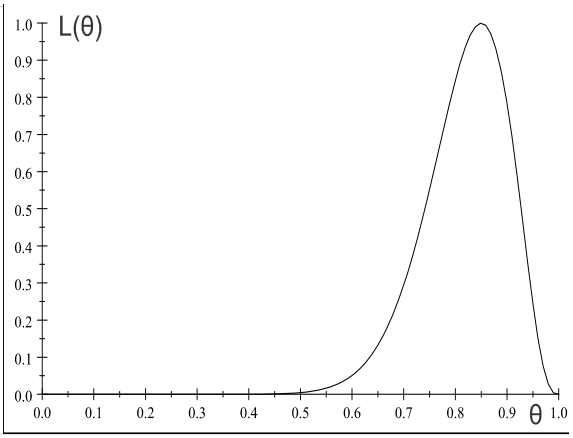


Fig. 2: $L(\theta;\mathbf{x}_0),\ \theta{\in}[0,1],\ Y{=}.85$

2 Point estimation

▶ How would one estimate an unknown parameter θ?

Point estimation. In the context of a statistical model $\mathcal{M}_\theta(\mathbf{x})=\{f(\mathbf{x};\theta),\ \theta\in\Theta\}$, $\mathbf{x}\in\mathbb{R}^n_X$, the data information comes in the form of a particular realization $\mathbf{x}_0$ of the sample $\mathbf{X}:=(X_1,X_2,...,X_n)$.

The primary objective of point estimation is to construct a procedure (mapping) that pin-points the true value θ^* of θ in Θ 'as accurately as possible'. That is, use the data to learn about the 'true' generating mechanism:

$$\mathcal{M}^*(\mathbf{x})=\{f(\mathbf{x};\theta^*)\},\ \mathbf{x}\in\mathbb{R}^n_X.$$

The estimation method comes in the form of a mapping between the sample ($\mathcal{X}:=\mathbb{R}^n_X$) and the parameter ($\Theta$) spaces:

$$h(.):\ \mathcal{X}\to\Theta.$$

This mapping, referred to as a point *estimator* of θ, is denoted by (figure 3):

$$\text{Estimator:}\quad \widehat{\theta}(\mathbf{X})=h(X_1,X_2,...,X_n).$$

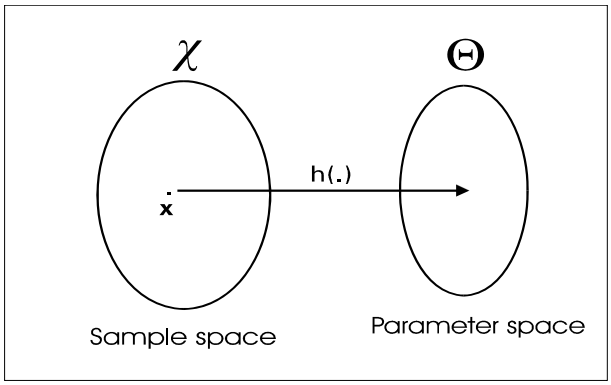


Fig. 3: Defining an estimator

The particular value taken by this estimator, based on the sample realization $\mathbf{X}=\mathbf{x}_0$, is referred to as a point *estimate*:

$$\text{Estimate:}\quad \widehat{\theta}(\mathbf{x}_0)=h(\mathbf{x}_0).$$

NOTE that to avoid cumbersome notation the same symbol $\widehat{\theta}$ is often used to denote both the estimator. When $\widehat{\theta}$ is used without the right hand side, the meaning should be obvious from the context.

Crucial distinction: $\widehat{\theta}(\mathbf{X})$-estimator (mathematical world), $\widehat{\theta}(\mathbf{x}_0)$-estimate (real world), and θ-unknown constant (mathematical world); Fisher (1922).

Unnecessary distinctions and gratuitous jargon? In the early 1920s, Fisher recast frequentist statistics and in that process he introduced several distinctions accompanied by new terminology to avoid confusion. The statisticians of that period reacted negatively to Fisher's changes and considered the distinctions and new terminology unnecessary. Indeed, Fisher's new terminology was widely interpreted as symptomatic of ostentatious display of pedantry. One such distinction was between the unknown parameter θ, its estimator $\widehat{\theta}(\mathbf{X})$ (Fisher called a *statistic* that also

includes test statistics). *Arne Fisher* [American mathematician/statistician] complained to R.A. Fisher in a letter about his "introduction in statistical method of some outlandish and barbarous technical terms. They stand out like quills upon the porcupine, ready to impale the sceptical critic. Where, for instance, did you get that atrocity, *a statistic*?" Fisher calm response was:
"I use special words for the best way of expressing special meanings. Thiele and Pearson were quite content to use the same words for what they were estimating [θ] and for their estimates of it [$\widehat{\theta}(\mathbf{X})$]. Hence the chaos in which they left the problem of estimation. Those of us who wish to distinguish the two ideas prefer to use different words, hence 'parameter' and 'statistic'. No one who does not feel this need is under any obligation to use them. Also, to Hell with pedantry." (Bennett, 1990, pp. 311-313).

Example. It is known that in the Bernoulli model $\theta{=}E(X){=}\mathbb{P}(X{=}1)$. This suggests that an obvious choice of a mapping as an estimator of θ is the sample mean: $\widehat{\theta}{=}\frac{1}{n}\sum_{i=1}^{n}X_i$.

Let us take the idea of an estimator a step further. In light of the fact that the estimator $\widehat{\theta}(\mathbf{X}){=}\frac{1}{n}\sum_{i=1}^{n}X_i$, is a function of the sample $\mathbf{X}$, $\widehat{\theta}(\mathbf{X})$ is a random variable itself, and thus the estimate $\widehat{\theta}(\mathbf{x}_0)$ is just one of the many (often infinitely many) values $\widehat{\theta}(\mathbf{X})$ could have taken. In the case of the above Bernoulli example for each sample realization, say $\mathbf{x}_{(i)}$, $i{=}1,2,...$ there is a different estimate, say:

$$\widehat{\theta}_{(1)}{=}.40,\ \ \widehat{\theta}_{(2)}{=}.43,\ \ \widehat{\theta}_{(3)}{=}.45,\ \ \widehat{\theta}_{(4)}{=}.51,\ \ \widehat{\theta}_{(5)}{=}.35,$$

but all these are values of the same estimator $\widehat{\theta}(\mathbf{X})$. All possible values of $\widehat{\theta}(\mathbf{X})$, together with their corresponding probabilities are described by its sampling distribution:

$$f(\widehat{\theta}(\mathbf{x});\theta),\forall\mathbf{x}{\in}\mathbb{R}_X^n,$$

whose functional form is determined by $f(\mathbf{x};\theta)$ as in (**??**).

Theorem 1. If $\mathbf{X}{:=}(X_1,X_2,...,X_n)$ is a random (IID) sample from the Bernoulli distribution, then the sampling distribution of $Y{=}\sum_{k=1}^{n}X_k$ is Binomial:

$$Y{=}\textstyle\sum_{k=1}^{n}X_k{\backsim}\mathsf{Bin}\left(n\theta,\ n\theta(1{-}\theta);n)\right), \tag{2}$$

Example. For the simple Bernoulli model the sampling distribution of $\widehat{\theta}(\mathbf{X}){=}\frac{1}{n}\sum_{i=1}^{n}X_i$ is a scaled Binomial distribution:

$$\widehat{\theta}(\mathbf{X}){\backsim}\ \mathrm{Bin}\left(\theta,\ \tfrac{\theta(1-\theta)}{n};n\right),\ \forall\mathbf{x}{\in}\{0,1\}^n.$$

The result $E(\widehat{\theta}(\mathbf{X}))=\theta$ follows from theorem 1 and property E2 of the mean in table 1.

Table 1: **Mean - Properties**
E1. $E(c){=}c$,
E2. $E(aX_1+bX_2){=}aE(X_1)+bE(X_2)$.

The result $Var(\widehat{\theta}(\mathbf{X}))=\frac{\theta(1-\theta)}{n}$ follows from theorem 1 and property V2 of the variance in table 2.

Table 2: **Variance - Properties**

V1.	$Var(c)=0$,
V2.	$Var(aX_1+bX_2)=a^2Var(X_1)+b^2Var(X_2)$,

The idea is that $f(\widehat{\theta}(\mathbf{x});\theta)$, $\forall \mathbf{x}\in\{0,1\}^n$ is closely distributed around θ^*, the true θ, as possible. The various concepts associated with the optimality of an estimator will be discussed in chapter 11. Ideally, $f(\widehat{\theta}(\mathbf{x});\theta)$ assigns probability one to θ^*, i.e. it reduces to a degenerate distribution of the form $\mathbb{P}(\widehat{\theta}(\mathbf{X})=\theta^*)=1$.

The notation $\widehat{\theta}(\mathbf{X})$ is used to denote an estimator in order to bring out the fact that it is a function of the sample $\mathbf{X}$, and for different values it generates the sampling distribution $f(\widehat{\theta}(\mathbf{x});\theta)$, for $\mathbf{x}\in\mathfrak{X}$. *Post-data* $\widehat{\theta}(\mathbf{X})$ yields an **estimate** $\widehat{\theta}(\mathbf{x}_0)$, which constitutes a particular value of $\widehat{\theta}(\mathbf{X})$ corresponding to data $\mathbf{x}_0$.

In light of the definition in (**??**), which of the following mappings constitute potential estimators of θ?

Table 3: Estimators of θ?

[a]	$\widehat{\theta}_1(\mathbf{X})=X_n$,
[b]	$\widehat{\theta}_2(\mathbf{X})=X_1-X_n$,
[c]	$\widehat{\theta}_3(\mathbf{X})=\frac{(X_1+X_n)}{2}$,
[d]	$\widehat{\theta}_k(\mathbf{X})=\frac{1}{n}\sum_{i=1}^{n}X_i$, for some $k>3$,
[e]	$\widehat{\theta}_{k+1}(\mathbf{X})=\frac{1}{n+1}\sum_{i=1}^{n}X_i$.

(3)

Do the mappings [a]-[e] in table 3 constitute estimators of θ? All five functions [a]-[e] have $\mathfrak{X}$ as their domain (they are functions of the sample $(X_1,X_2,...,X_n)$, but is the range of each mapping a subset of Θ:=$[0,1]$? Mapping [a], [c]-[e] constitute possible estimators of θ because their ranges are subsets of $[0,1]$. Mapping [b], however, does not constitute an estimator of θ because it can take the value -1 [ensure you understand why!] which lies outside $[0,1]$, the parameter space of θ.

One can easily think of many more functions from $\mathfrak{X}$ to Θ that will qualify as possible estimators of θ. Given the plethora of such possible estimators, how does one decide which one is the most appropriate?

To answer that question let us think about the possibility of an **ideal estimator**, $\theta^*(.):\mathfrak{X}\rightarrow\theta^*$, i.e., $\theta^*(\mathbf{x})=\theta^*$ for all values $\mathbf{x}\in\mathfrak{X}$. That is, $\theta^*(\mathbf{X})$ pinpoints the true value θ^* of θ, whatever the data. A moment's reflection reveals that no such estimator could exist because $\mathbf{X}$ is a random vector with its own distribution $f(\mathbf{x};\theta)$, for all $\mathbf{x}\in\mathfrak{X}$. Moreover, in view of the randomness of $\mathbf{X}$, any mapping of the form (**??**) will be a random variable with its own sampling distribution, $f(\widehat{\theta}(\mathbf{x});\theta)$, which is directly derivable from $f(\mathbf{x};\theta)$. Let us take stock of these distributions.

In the Bernoulli case, all the estimators [a], [c]-[e] are linear functions of $(X_1, X_2, ..., X_n)$ and thus, by (2), their distribution is Binomial. In particular,

Table 4: Estimators and their sampling distributions

[a]	$\widehat{\theta}_1(\mathbf{X})=X_n \backsim \mathsf{Ber}(\theta,\ \theta(1-\theta))$,
[c]	$\widehat{\theta}_3(\mathbf{X})=(X_1+X_n)/2 \backsim \mathsf{Bin}\left(\theta,\ [\frac{\theta(1-\theta)}{2}]\right)$,
[d]	$\widehat{\theta}_n(\mathbf{X})=\frac{1}{n}\sum_{i=1}^{n} X_i \backsim \mathsf{Bin}\left(\theta,\ [\frac{\theta(1-\theta)}{n}]\right)$,
[e]	$\widehat{\theta}_{n+1}(\mathbf{X})=\frac{1}{n+1}\sum_{i=1}^{n} X_i \backsim \mathsf{Bin}\left(\frac{n}{n+1}\theta,\ [\frac{n\theta(1-\theta)}{(n+1)^2}]\right)$.

It is important to emphasize at the outset that the sampling distributions [a]-[e] are evaluated under $\theta=\theta^*$ where θ^* is the true value of θ.

It is clear that none of the sampling distributions of the estimators in table 4 coincides with the *ideal estimator*, $\theta^*(\mathbf{X})$, whose sampling distribution would be of the form:

[i]	$\mathbb{P}(\theta^*(\mathbf{X})=\theta^*)=1.$
[i]a	$E(\theta^*(\mathbf{X}))=\theta^*,$
[i]b	$Var(\theta^*(\mathbf{X}))=0,$

(4)

where [i]a and [i]b denote the first two moments of the ideal estimator. In contrast to the (infeasible) ideal estimator in (4), for all the estimators in table 4 $Var(\theta^*(\mathbf{X}))>0$, but their sampling distributions provide the basis for evaluating the variability represented by a non-zero variance.

3 Properties of optimal point estimators

3.1 Finite sample properties

As mentioned above, the notion of an *optimal estimator* can be motivated by how well the sampling distribution of an estimator $\widehat{\theta}_n(\mathbf{X})$, say $f_n(\widehat{\theta}(\mathbf{X});\theta^*)$, approximates that of the ideal estimator in (4). In particular, the three features of the ideal estimator [i]a-[i]b motivate the following optimal properties of feasible estimators.

Condition [i]a motivates the property known as:

[I] **Unbiasedness**: An estimator $\widehat{\theta}(\mathbf{X})$ is said to be an *unbiased* for θ if:

$$E(\widehat{\theta}(\mathbf{X}))=\theta^*. \tag{5}$$

That is, the mean of the sampling distribution of $\widehat{\theta}(\mathbf{X})$ coincides with the true value of the unknown parameter θ.

Example. In the case of the simple Bernoulli model, we can see from table 4 that the estimators $\widehat{\theta}_1(\mathbf{X})$, $\widehat{\theta}_3(\mathbf{X})$ and $\widehat{\theta}_n(\mathbf{X})$ are unbiased since in all three cases (5) is satisfied. In contrast, estimator $\widehat{\theta}_{n+1}(\mathbf{X})$ is not unbiased because $E\left(\widehat{\theta}_{n+1}(\mathbf{X})\right)=\frac{n}{n+1}\theta \neq \theta$.

Condition [i]b motivates the property known as:

[II] **Relative efficiency**: : An estimator $\widehat{\theta}_1(\mathbf{X})$ is said to be *relatively more efficient than* $\widehat{\theta}_2(\mathbf{X})$ if:

$$Var(\widehat{\theta}_1(\mathbf{X})) > Var(\widehat{\theta}_2(\mathbf{X})),\ \text{for } n > 2.$$

One, however, needs to be careful with such comparisons because they can be very misleading when both estimators are bad (non-optimal)

Example. The estimators $\widehat{\theta}_1(\mathbf{X})$, $\widehat{\theta}_3(\mathbf{X})$ in table 4 are unbiased and $\widehat{\theta}_3(\mathbf{X})$ is relatively more efficient than $\widehat{\theta}_1(\mathbf{X})$ since:

$$Var(\widehat{\theta}_1(\mathbf{X}))=\theta(1-\theta) > Var(\widehat{\theta}_3(\mathbf{X}))=\tfrac{\theta(1-\theta)}{2},\ \text{for } n > 2.$$

does not mean that $\widehat{\theta}_3(\mathbf{X})$ is a good estimator, since there is another unbiased estimator whose variance can be considerably smaller than that of $\widehat{\theta}_3(\mathbf{X})$:

$$Var(\widehat{\theta}_3(\mathbf{X}))=\tfrac{\theta(1-\theta)}{2} > Var(\widehat{\theta}_n(\mathbf{X}))=\tfrac{\theta(1-\theta)}{n},\ \text{for any } n > 2.$$

Hence, relative efficiency is not something to write home about! A much more important property is known as full efficiency.

[III] **Full Efficiency**: An unbiased estimator $\widehat{\theta}_n(\mathbf{X})$ is said to be a *fully efficient* estimator of θ if its variance is as small as it can be, where the latter is expressed by the technical condition:

$$Var(\widehat{\theta}_n(\mathbf{X}))=CR(\theta):=\left[E\left(-\tfrac{d^2\ln f(\mathbf{x};\theta)}{d\theta^2}\right)\right]^{-1},$$

where $E\left(-\frac{d^2\ln f(\mathbf{x};\theta)}{d\theta^2}\right)$ is the Fisher information, and '$CR(\theta)$' stands for the **Cramer-Rao lower bound**.

Example (THE DERIVATIONS ARE NOT IMPORTANT!).
In the case of the simple Bernoulli model:

$$\ln f(\mathbf{x};\theta)=Y\ln\theta+(n-Y)\ln(1-\theta),\quad \text{where } Y=\textstyle\sum_{k=1}^{n}X_k,\ E(Y)=n\theta,$$

$$\tfrac{d\ln f(\mathbf{x};\theta)}{d\theta}=\ (Y)(\tfrac{1}{\theta})-(n-Y)(\tfrac{1}{1-\theta}),\ \tfrac{d^2\ln f(\mathbf{x};\theta)}{d\theta^2}=-Y(\tfrac{1}{\theta^2})-(n-Y)(\tfrac{1}{1-\theta})^2,$$

$$E\left(-\tfrac{d^2\ln f(\mathbf{x};\theta)}{d\theta^2}\right)\ =(\tfrac{1}{\theta^2})E(Y)+[n-E(Y)](\tfrac{1}{1-\theta})^2\ =\tfrac{n}{\theta(1-\theta)},$$

and thus the Cramer-Rao lower bound is: $CR(\theta):=\frac{\theta(1-\theta)}{n}$.
Looking at the estimators of θ in (4) it is clear that only one unbiased estimator achieves that bound, $\widehat{\theta}_n(\mathbf{X})$. Hence, $\widehat{\theta}_n(\mathbf{X})$ is the only estimator of θ which is both *unbiased* and *fully efficient*.

[IV] **Sufficiency**. An estimator $\widehat{\boldsymbol{\theta}}_n(\mathbf{X})$ is said to be sufficient if it is a function of a sufficient statistic $\mathbf{S}(\mathbf{X})$.

Sufficient statistic. In the context of a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, a statistic $\mathbf{S}(\mathbf{X})$ is said to be *sufficient* for $\boldsymbol{\theta}$ if:

$$f(\mathbf{x};\boldsymbol{\theta})=f(\mathbf{x}|\mathbf{s})\cdot f(\mathbf{s};\boldsymbol{\theta}),\ \forall\mathbf{x}\in\mathbb{R}^n_X. \tag{6}$$

That is, $f(\mathbf{x};\boldsymbol{\theta})$ can be reduced to a product of: (i) a conditional distribution of $\mathbf{X}$ given $\mathbf{S}$=$\mathbf{s}$, that is free of $\boldsymbol{\theta}$, and (ii) the marginal distribution of $\mathbf{S}$ which *does* depend on $\boldsymbol{\theta}$.

Intuitively, sufficiency has to do with whether the information contained in a sample $\mathbf{X}$:=$(X_1, X_2, ..., X_n)$ can be compressed into a lower dimensional statistic $\mathbf{S}(\mathbf{X})$ without sacrificing any information relevant for inferences about $\boldsymbol{\theta}$. Given that both the sample $\mathbf{X}$ itself is a sufficient statistic, sufficiency is interesting in cases where $\mathbf{S}$ is of much lower dimensionality than $\mathbf{X}$, and ideally, $\mathbf{S}$ has the same dimensionality as $\boldsymbol{\theta}$. Hence, the form of sufficiency of interest is the smallest dimensionality possible.

Example. It can be shown that in the case of a simple Bernoulli model, the statistic $s(\mathbf{X})=\sum_{k=1}^{n} X_k$ is sufficient for θ; see Spanos (2019), ch. 11. This implies that the only estimators that are sufficient are $\widehat{\theta}_n(\mathbf{X})=\frac{1}{n}\sum_{i=1}^{n} X_i$ and $\widehat{\theta}_{n+1}(\mathbf{X})=\frac{1}{n+1}\sum_{i=1}^{n} X_i$, since they are function of the sufficient statistic $s(\mathbf{X})=\sum_{k=1}^{n} X_k$:

$$\mathbf{X}:=\overbrace{(X_1, X_2, ..., X_n)}^{\text{n-dimensional}} \quad \rightsquigarrow \quad s(\mathbf{X})=\overbrace{\sum_{k=1}^{n} X_k}^{\text{1-dimensional}}$$

3.2 Asymptotic properties

The finite sample properties of an estimator $\widehat{\theta}_n(\mathbf{X})$ are specified in terms of the sampling distribution $f_n(\widehat{\theta};\theta^*)$ associated with a particular sample size n. The asymptotic properties are defined in terms of the asymptotic sampling distribution $f_\infty(\widehat{\theta};\theta^*)$ aiming to approximate $f_n(\widehat{\theta};\theta^*)$ at the limit as $n\to\infty$.

What renders the two unbiased estimators $\widehat{\theta}_1(\mathbf{X})$, $\widehat{\theta}_3(\mathbf{X})$ in table 2 practically useless? An asymptotic property motivated by condition [i] of the ideal estimator, known as *consistency*. Intuitively, an estimator $\widehat{\theta}_n(\mathbf{X})$ is consistent when its precision (how close to θ^* is) improves as the sample size n increases. Condition [i] of the ideal estimator motivates the property known as:

[V] **Consistency**: an estimator $\widehat{\theta}_n(\mathbf{X})$ is *consistent* if:

$$\text{Strong:} \quad \mathbb{P}(\lim_{n\to\infty}\widehat{\theta}_n(\mathbf{X})=\theta^*)=1,$$

$$\text{Weak:} \quad \lim_{n\to\infty}\mathbb{P}(|\widehat{\theta}_n(\mathbf{X})-\theta^*|\leq\varepsilon)=1.$$

That is, an estimator $\widehat{\theta}_n(\mathbf{X})$ is consistent if it approximates (probabilistically) the sampling distribution of the ideal estimator *asymptotically*; as $n\to\infty$. The difference between strong and weak consistency stems from the form of probabilistic convergence they involve, with the former being stronger than the latter. Both of these properties constitute an extension of the Strong and Weak Law of Large Numbers (LLN) which hold for the sample mean $\overline{X}_n=\frac{1}{n}\sum_{i=1}^{n} X_i$ of a process $\{X_k,\ k=1,2,...,n,...\}$, under certain probabilistic assumptions, the most restrictive being that the process is IID; see Spanos (2019), ch. 8.

In practice, it is no-trivial to prove that a particular estimator is consistent or not by verifying directly the conditions in (**??**). However, there is often a short-cut for verifying consistency in the case of unbiased estimators using the *sufficient* condition:

$$Var\left(\widehat{\theta}_n(\mathbf{X})\right) \underset{n\to\infty}{\longrightarrow} 0. \tag{7}$$

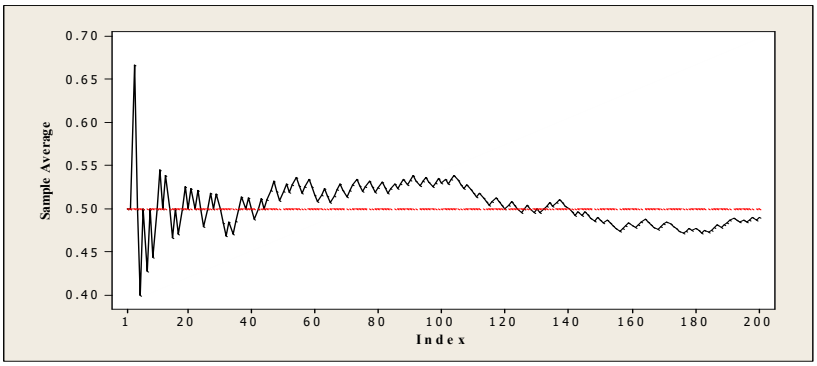


Fig. 4: $\overline{x}_n$ for a BerIIDrealization with n=200

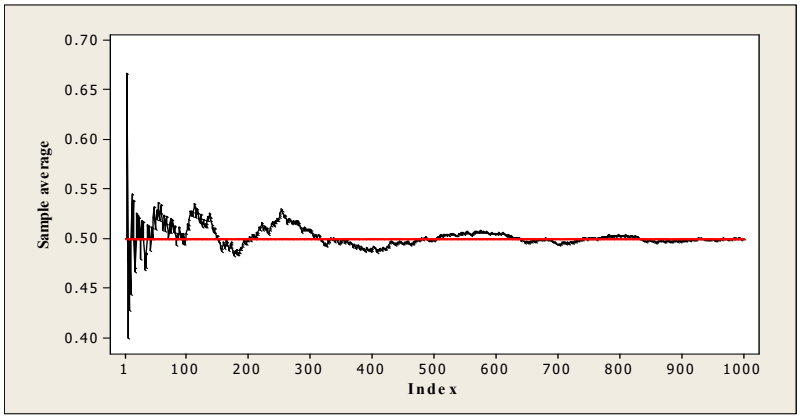


Fig. 5: $\overline{x}_n$ for a BerIIDrealization with n=1000

Example. In the case of the simple Bernoulli model, one can verify that the estimators $\widehat{\theta}_1(\mathbf{X})$ and $\widehat{\theta}_3(\mathbf{X})$ are inconsistent because:

$$Var\left(\widehat{\theta}_1(\mathbf{X})\right) \underset{n\to\infty}{\longrightarrow} \theta(1-\theta)\neq 0, \quad Var\left(\widehat{\theta}_1(\mathbf{X})\right) \underset{n\to\infty}{\longrightarrow} \frac{\theta(1-\theta)}{2}\neq 0,$$

i.e. their variances do not decrease to zero as the sample size n goes to infinity.

Let us take stock of the above properties and how they can used by the practitioner in deciding which estimator is optimal. The property which defines *minimal reliability* for an estimator is that of *consistency*. Intuitively, consistency indicates that as the sample size increases [as $n \to \infty$], the estimator $\widehat{\theta}_n(\mathbf{X})$ approaches θ^*, the true value of θ, in some probabilistic sense; convergence almost surely or convergence in probability. Hence, if an estimator $\widehat{\theta}_n(\mathbf{X})$ is *not* consistent, it is automatically excluded from the subset of potentially optimal estimators, irrespective of any other

properties this estimator might enjoy. In particular, an unbiased estimator which is inconsistent is practically useless. On the other hand, just because an estimator $\widehat{\theta}_n(\mathbf{X})$ is consistent does not imply that it's a 'good' estimator; it only implies that it's minimally acceptable.

■ It is important to emphasize that the properties of unbiasedness and fully efficiency hold for any sample size $n > 1$, and thus we call them *finite sample properties*, but consistency is an *asymptotic property* because it holds as $n \to \infty$.

Example. In the case of the simple Bernoulli model, if the choice between estimators is confined (artificially) among the estimators $\widehat{\theta}_1(\mathbf{X})$, $\widehat{\theta}_3(\mathbf{X})$ and $\widehat{\theta}_{n+1}(\mathbf{X})$, the latter estimator should be chosen, despite being biased, because it's a consistent estimator of θ. On the other hand, among the estimators given in table 4, $\widehat{\theta}_k(\mathbf{X})$ is clearly the best (most optimal) because it satisfies all three properties. In particular, $\widehat{\theta}_n(\mathbf{X})$, not only satisfies the minimal property of consistency, but it also has the smallest variance possible, which means that it comes closer to the ideal estimator than any of the others, for any sample size $n > 2$. The sampling distribution of $\widehat{\theta}_n(\mathbf{X})$, when evaluated under $\theta{=}\theta^*$, takes the form:

$$[\mathrm{d}] \quad \widehat{\theta}_n(\mathbf{X}){=}\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i \overset{\theta=\theta^*}{\backsim} \mathsf{Bin}\left(\theta^*,\ \tfrac{\theta^*(1-\theta^*)}{n}\right), \tag{8}$$

whatever the value θ^* happens to be.

In addition to the properties of estimators mentioned above, asymptotic Normality is often used when the finite sampling distribution $f_n(\widehat{\theta};\theta^*)$ cannot be derived explicitly. In such a case one relies on the asymptotic distribution $f_\infty(\widehat{\theta};\theta^*)$ that aims to approximate $f_n(\widehat{\theta};\theta^*)$ at the limit as $n \to \infty$.

[VI] **Asymptotic Normality**: an estimator $\widehat{\theta}_n(\mathbf{X})$ is said to be asymptotically Normal if:

$$\sqrt{n}(\widehat{\theta}_n(\mathbf{X}) - \theta) \underset{a}{\backsim} \mathsf{N}(0, V_\infty(\theta)),\ V_\infty(\theta){\neq}0, \tag{9}$$

where '$\underset{a}{\backsim}$' stands for 'can be asymptotically approximated by', and $V_\infty(\theta)$ denotes the asymptotic variance.

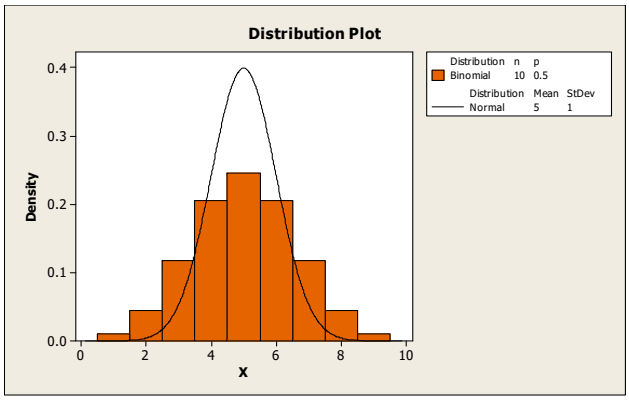
This property is an extension of a well-known result in probability theory: **the Central Limit Theorem** (CLT). The CLT asserts that, under certain probabilistic assumptions on the process $\{X_k,\ k{=}1, 2, ..., n, ...\}$, , the most restrictive being that the process is IID, the sampling distribution of $\overline{X}_n{=}\frac{1}{n}\sum_{i=1}^{n} X_i$ for a 'large enough' n can be approximated by the Normal distribution (Spanos, 1999, ch. 8):

$$\frac{\left(\overline{X}_n - E(X)\right)}{\sqrt{Var(\overline{X}_n)}} \underset{a}{\backsim} \mathsf{N}(0,1). \tag{10}$$

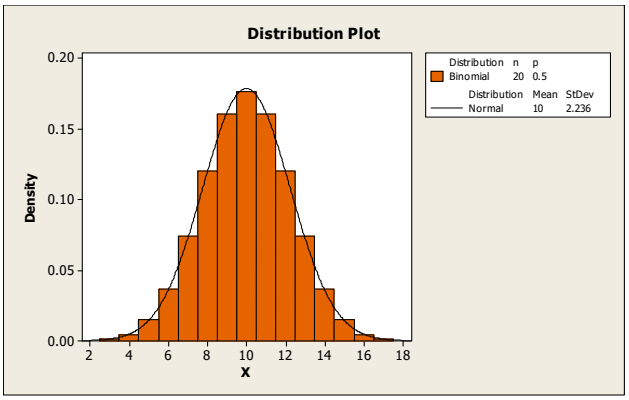
Note that the important difference between (9) and (10) is that $\widehat{\theta}_n(\mathbf{X})$ in the former does not have to coincide with $\overline{X}_n$; it can be any well-behaved function $h(\mathbf{X})$ of the sample $\mathbf{X}$.

Example. In the case of the simple Bernoulli model the sampling distribution of $\widehat{\theta}_n(\mathbf{X})$, which we know is Binomial (see (8)), it can also be approximated using (10).

In the graph below we compare the Normal approximation to the Binomial for n=10 and n=20 in the case where θ=.5, and the improvement is clearly noticeable.



Normal approx. of Bin.:
$f(y;\theta{=}.5,n{=}10)$



Normal approx. of Bin.:
$f(y;\theta{=}.5,n{=}20)$

Summary of the results for the simple Bernoulli model, where UN stands for Unbiasedness, FE for Full Efficiency, S for Sufficiency and SC for strong consistency in table 5.

Table 5: Estimators and their properties	UN	FE	S	SC
[a] $\widehat{\theta}_1(\mathbf{X}){=}X_n \backsim \mathsf{Ber}(\theta,\ \theta(1{-}\theta))$,	✓	×	×	×
[c] $\widehat{\theta}_3(\mathbf{X}){=}(X_1+X_n)/2 \backsim \mathsf{Bin}\left(\theta,\ [\frac{\theta(1-\theta)}{2}]\right)$,	✓	×	×	×
[d] $\widehat{\theta}_n(\mathbf{X}){=}\frac{1}{n}\sum_{i=1}^n X_i \backsim \mathsf{Bin}\left(\theta,\ [\frac{\theta(1-\theta)}{n}]\right)$,	✓	✓	✓	✓
[d] $\widehat{\theta}_{n+1}(\mathbf{X}){=}\frac{1}{n+1}\sum_{i=1}^n X_i \backsim \mathsf{Bin}\left(\frac{n}{n+1}\theta,\ [\frac{n\theta(1-\theta)}{(n+1)^2}]\right)$.	×	✓	✓	✓

Table 5 indicates that the best (optimal) estimator of θ:

$$\widehat{\theta}_n(\mathbf{X}){=}\tfrac{1}{n}\textstyle\sum_{i=1}^n X_i \backsim \mathsf{Bin}\left(\theta,\ [\frac{\theta(1-\theta)}{n}]\right).$$

The bottom line is that any estimator that has all three properties, Unbiased, Fully Efficient and Strongly Consistent is an 'optimal' estimator.

3.3 The simple Normal model: estimation

The simple Normal model is specified by:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})\text{: } X_k \backsim \mathsf{NIID}(\mu,\sigma^2),\ x_k{\in}\mathbb{R},\ \mu{\in}\mathbb{R},\ \sigma^2>0,\ k{\in}\mathbb{N}{:=}(1,2,...,n,...).$$

In what follows we assume (for simplicity) that σ^2 is known.

Theorem 2. If $\mathbf{X}$:=$(X_1, X_2, ..., X_n)$ is a random (IID) sample from the Normal distribution then the sampling distribution of $\sum_{k=1}^{n} X_k$ is:

$$\sum_{k=1}^{n} X_k \backsim \mathsf{N}(n\mu,\ n\sigma^2). \tag{11}$$

Among the estimators in table 6, the sample mean:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \backsim \mathsf{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

constitutes *the* optimal point estimator of μ because it is:

[U] *Unbiased* ($E(\overline{X}_n)=\mu^*$),

[FE] *Fully Efficient* ($Var(\overline{X}_n)=CR(\mu)=\frac{\sigma^2}{n}$),

[S] *Sufficient* ($\sum_{i=1}^{n} X_i$ is a sufficient statistic), and

[SC] *Strongly Consistent* ($\mathbb{P}(\lim_{n\to\infty} \overline{X}_n=\mu^*)=1$).

Table 6: Estimators and their properties		UN	FE	S	SC
[a]	$\widehat{\mu}_1(\mathbf{X})=X_n \backsim \mathsf{N}(\mu, \sigma^2)$	✓	×	×	×
[b]	$\widehat{\mu}_2(\mathbf{X})=X_1 - X_n \backsim \mathsf{N}(0, 2\sigma^2)$	×	×	×	×
[c]	$\widehat{\mu}_3(\mathbf{X})=(X_1 + X_n)/2 \backsim \mathsf{N}(\mu, \frac{1}{2}\sigma^2)$	✓	×	×	×
[d]	$\widehat{\mu}_n(\mathbf{X})=\frac{1}{n}\sum_{i=1}^{n} X_i \backsim \mathsf{N}\left(\mu, \frac{\sigma^2}{n}\right)$	✓	✓	✓	✓
[e]	$\widehat{\mu}_{n+1}(\mathbf{X})=\frac{1}{n+1}\sum_{i=1}^{n} X_i \backsim \mathsf{N}\left(\frac{n}{n+1}\mu, \frac{n\sigma^2}{(n+1)^2}\right)$	×	×	✓	✓

where UN stands for Unbiasedness, FE for Full Efficiency, S for Sufficiency and SC for strong consistency in table 6.

Table 6 lists several possible estimators of μ, together with their properties. It turns out that in the case of the simple Normal model with σ^2- known, $\sum_{i=1}^{n} X_i$ is a sufficient statistic for μ. The above results indicate most clearly that:

$$\widehat{\mu}_n(\mathbf{X})=\frac{1}{n}\sum_{i=1}^{n} X_i \backsim \mathsf{N}(\mu, \tfrac{\sigma^2}{n}),$$

is the best (optimal) estimator of μ. Ensure that you understand the results in table 6.

4 Summary and conclusions

The primary objective in frequentist estimation is to learn about θ^* the true value of the unknown parameter θ of interest using its sampling distribution $f_n(\widehat{\theta};\theta^*)$ associated with a particular sample size n. The finite sample properties are defined directly in terms of $f_n(\widehat{\theta};\theta^*)$ and the asymptotic properties are defined in terms of the asymptotic sampling distribution $f_\infty(\widehat{\theta};\theta^*)$ aiming to approximate $f_n(\widehat{\theta};\theta^*)$ at the limit as $n \to \infty$.

The question that needs to be considered at this stage is: **what combination of the above mentioned properties specifies an 'optimal' estimator?**

A necessary but minimal property for an estimator is **consistency** (preferably strong). By itself, however, consistency does not secure learning from data for a given n; it's a promissory note for potential learning. Hence, for actual learning one needs to supplement consistency with certain finite sample properties like unbiasedness and efficiency to ensure that learning *can* take place with the particular data $\mathbf{x}_0$:=$(x_1, x_2, ..., x_n)$ of sample size n.

Among finite sample properties **full efficiency** is clearly the most important because it secures the highest degree of learning for a given n since it offers the best possible precision.

Relative efficiency, although desirable, needs to be investigated further to find out how large is the class of estimators being compared before passing judgement. Being the best econometrician in my family does not make me a good econometrician!!

Unbiasedness, although desirable, is not considered indispensable by itself. Indeed, as shown above, an *unbiased* but *inconsistent* estimator is practically *useless*, and a *consistent* but *biased* estimator is always preferable.

Sufficiency, is desirable because a sufficient estimator uses all the information in the sample $\mathbf{X}$:=$(X_1, X_2, ..., X_n)$ relevant for inference purposes.

Hence, a **consistent**, **unbiased, sufficient** and **fully efficient** estimator sets the **gold standard** in estimation.

In conclusion, it is important to emphasize that **point estimation** is often considered *inadequate* for the purposes of scientific inquiry because an 'optimal' point estimator $\widehat{\theta}_n(\mathbf{X})$, by itself, does not provide any measure of the reliability and precision associated with the estimate $\widehat{\theta}_n(\mathbf{x}_0)$.

► A practitioner would be **wrong** to presume that point estimation gives rise to an inferential claim of the form $\widehat{\theta}_n(\mathbf{x}_0) \simeq \theta^*$. IT DOES NOT, NO SUCH INFERENCE IS WARRANTED!