Summer Seminar: Philosophy of Statistics
Lecture Notes 6: Estimation 1: Properties of Estimators

Aris Spanos [SUMMER 2019]

1 Introduction

Chapter 10 introduced the basic ideas about estimation, focusing on the primary objective of an estimator $\widehat{\theta}_n(\mathbf{X})$, which is to pin-point θ^*, the 'true' value of θ in Θ. How best (optimally) to achieve that objective is the subject matter of the this chapter. Before we begin the discussion of what properties determine an optimal estimator, it is important to emphasize that estimation pressupposes a statistical model:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\subset\mathbb{R}^m\},\ \mathbf{x}\in\mathbb{R}^n_X,\ m<n, \tag{1}$$

whose probabilistic assumptions are assumed to be *valid* at the outset. In practice, one needs to secure the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ beforehand. This is necessary because all the properties of the estimators are derived by invoking the model assumptions. When any of the model assumptions are invalid for one's data, the optimal properties are usually invalidated.

2 What is an estimator?

The theory of optimal estimation is largely due to R.A. Fisher in a series of papers that began with his first paper (1912), when he was an undergraduate at Cambridge, and culminated with several influential papers in the mid 1930s. The concept of the likelihood function and several properties of estimators to be discussed in what follows, such as consistency, asymptotic efficiency, sufficiency, were introduced in Fisher's classic (1922a), the paper that founded modern statistics. In a follow up paper on estimation, Fisher (1925b) introduced the concept we call today Fisher's information, as well as the notions of efficiency in small samples and ancillarity; see Hald (1998) and Gorroochurn (2016) for further discussion.

As argued in chapter 10, commencing with a prespecified statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, as in (1), the estimation of an unknown parameter θ, amounts to defining a function of the form: $h(.)$: $\mathbb{R}^n_X \rightarrow \Theta$, where $\mathbb{R}^n_X$ denotes the *sample space* and Θ the *parameter space*. The data $\mathbf{x}_0$ are viewed as a typical realization of the sample $\mathbf{X}:=(X_1, X_2, ..., X_n)$ whose probabilistic structure is specified by $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. The function $\widehat{\theta}(\mathbf{X})=h(\mathbf{X})$ denotes the (point) estimator, and its observed value $\widehat{\theta}(\mathbf{x}_0)=h(\mathbf{x}_0)$ the estimate.

Example 11.1. Consider the *simple Bernoulli model*, specified in table 11.1.

Table 11.1: The simple Bernoulli model

Statistical GM:		$X_t{=}\theta+u_t,\ t{\in}\mathbb{N}{:=}(1,2,...,n,...)$
[1]	Bernoulli:	$X_t \backsim \mathsf{Ber}(.,.),\ x_t{=}\{0,1\},$
[2]	Constant mean:	$E(X_t){=}\theta,\ 0{\leq}\theta{\leq}1,$ for all $t{\in}\mathbb{N},$
[3]	Constant variance:	$Var(X_t){=}\theta(1{-}\theta),$ for all $t{\in}\mathbb{N},$
[4]	Independence:	$\{X_t,\ t{\in}\mathbb{N}\}$ - independent process.

Note that the specification is given with all the probabilistic assumptions explicitly stated in order to demonstrate how the different assumptions are invoked when the properties of various estimators are derived. For the simple Bernoulli model:

$$\text{sample space: } \mathbb{R}^n_X{=}\{0,1\}^n, \quad \text{parameter space: } \Theta{=}[0,1],$$

The following functions constitute likely estimators of θ:

$$\begin{array}{l}\text{(i) } \widehat{\theta}_1{=}X_1, \text{ (ii) } \widehat{\theta}_2{=}\tfrac{1}{2}(X_1{+}X_2), \text{ (iii) } \widehat{\theta}_3{=}\tfrac{1}{3}(X_1{+}X_2{+}X_n),\\ \text{(iv) } \widehat{\theta}_n{=}\tfrac{1}{n}\sum_{i=1}^n X_i, \text{ (v) } \widehat{\theta}_{n+1}{=}\tfrac{1}{n+1}\sum_{i=1}^n X_i, \text{ (vi) } \widehat{\theta}_{n+2}{=}\tfrac{1}{n+2}\sum_{i=1}^n X_i.\end{array} \tag{2}$$

To verify that all these functions constitute legitimate estimators of θ one needs to check that each of the functions (i)-(vi) are mappings of the form: $h(.)$: $\{0,1\}^n \longrightarrow [1,0]$; ensure you understand that they are such functions.

Example 11.2. For the simple Bernoulli model (table 11.1), consider the functions:

$$\text{(vii) } \widehat{\theta}_4{=}(X_1{-}X_n), \text{ (viii) } \widehat{\theta}_5{=}\left(\tfrac{1}{n}\right)\sum_{i=1}^n X_i^{\beta}, \text{ (ix) } \widehat{\theta}_6{=}.8.$$

None of these functions define legitimate estimators of θ since:

$$\begin{array}{rl}\widehat{\theta}_4{=}(X_1{-}X_n): & \text{the domain is } \{0,1\}^n \text{ but the range is } [-1,1] \neq [0,1],\\ \widehat{\theta}_5{=}\frac{1}{n}\sum_{i=1}^n X_i^{\delta}: & \text{the domain is } \textit{not } \{0,1\}^n \text{ since it includes an unknown } \delta,\\ \widehat{\theta}_6{=}.8: & \text{the domain is } \textit{not } \{0,1\}^n \text{ since } .8 \neq 0,\ .8 \neq 1.\end{array}$$

Example 11.3. Consider the *simple* (one parameter) *Normal model* (table 11.2).

Table 11.2: The simple (one parameter) **Normal model**

Statistical GM:		$X_t{=}\mu+u_t,\ t{\in}\mathbb{N}{:=}(1,2,...,n,...)$
[1]	Normal:	$X_t \backsim \mathsf{N}(.,.),\ x_t{\in}\mathbb{R},$
[2]	Constant mean:	$E(X_t){=}\mu,\ \mu{\in}\mathbb{R},\ \forall t{\in}\mathbb{N},$
[3]	Constant variance:	$Var(X_t){=}\sigma^2,\ \sigma^2$-known, $\forall t{\in}\mathbb{N},$
[4]	Independence:	$\{X_t,\ t{\in}\mathbb{N}\}$-independent process.

For the simple (one parameter) Normal model:

$$\text{sample space: } \mathbb{R}_X^n{=}\mathbb{R}^n, \quad \text{parameter space: } \Theta{=}\mathbb{R},$$

The following functions constitute legitimate estimators of μ:

$$\begin{array}{l} \text{(i) } \widehat{\mu}_1{=}X_1, \text{ (ii) } \widehat{\mu}_2{=}\frac{1}{2}(X_1{+}X_2), \text{ (iii) } \widehat{\mu}_3{=}\frac{1}{2}(X_1{-}X_n), \\ \text{(iv) } \widehat{\mu}_n{=}\frac{1}{n}\sum_{i=1}^n X_i, \text{ (v) } \widehat{\mu}_{n+1}{=}\frac{1}{n+1}\sum_{i=1}^n X_i, \text{ (vi) } \widehat{\mu}_{n+2}{=}\frac{1}{n+2}\sum_{i=1}^n X_i. \end{array} \tag{3}$$

Given that the parameter μ takes values over the whole of the real line ($\mathbb{R}$), it will be impossible to define a function of the sample $(X_1, X_2, ..., X_n)$ which is not an estimator of μ.

This example brings out the fact that it is very easy to define numerous possible estimators. Hence, the question that naturally arises is:

▶ **how does one choose among such a plethora of estimators?**

The answer depends on what is the primary objective of an estimator, which is to pin-point the true value θ^* of θ in Θ as 'closely as possible'; see chapter 10. The problem of choosing a good estimator resembles a situation where an archer is standing at the foot of a hill with the target on the other side of the hill beyond the archer's vision. What the archer has to do is devise a strategy (rule) relating to factors within his/her control, such as the shooting angle and the pulling power, with a view to ensure that the arrow will land as close to the center of the target as possible. The modeler has to choose a rule (an estimator) in a way that ensures close proximity to the unknown true value θ^*. A modeler would like the have the **ideal estimator** $\widehat{\theta}^*(\mathbf{X})$, a function of the form:

$$\widehat{\theta}^*(.) \colon \mathbb{R}_X^n \longrightarrow \theta^*.$$

That is, for all different values $\mathbf{x}$ of the sample $\mathbf{X}$, $\widehat{\theta}^*(\mathbf{X})$ pinpoints θ^*. It should come as no surprise that no such an ideal estimator exists for a finite n; the archer hits θ^* with every arrow! It might, however, exist asymptotically (as $n \to \infty$) when:

$$\mathbb{P}(\lim_{n\to\infty} \widehat{\theta}_n(\mathbf{X}){=}\theta^*){=}1.$$

This is the *strong consistency* property mentioned in chapter 9. In the same chapter, it was emphasized that all the limit theorems (WLLN, SLLN, CLT) assert what happens at the limit $n{=}\infty$, which is never the case in practice, and tell us nothing about how good the approximation is for a given n. This is why asymptotic properties should be viewed with caution as necessary but never sufficient in defining an optimal estimator. What is needed for that are properties that hold for any $n > 1$ known as finite sample properties. These properties are defined in terms of the finite sampling distribution of the estimator in question.

3 Sampling distributions of estimators

The concept of a finite sampling distribution of an estimator or a test was introduced by William Sealy Gosset (1876–1937) in 'Student' (1909) when he put forward the fine sample distribution known today as the Student's t associated with a widely used test of significance. That paper inspired R.A. Fisher (1890–1962) to formally derive several sampling distributions associated with the simple Normal and Linear Regression models over a period of two decades that formed the foundation for frequentist inference.

Since the estimator $\widehat{\theta}(\mathbf{X})$ is a function $Y_n{=}h(\mathbf{X})$ of the sample $\mathbf{X}$, in principle one can always use $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x}{\in}\mathbb{R}_X^n$, defined by the model assumptions, to derive its sampling distribution via:

$$F(y;\boldsymbol{\theta}){=}\mathbb{P}(Y_n{\leq}y;\boldsymbol{\theta}){=}\underbrace{\int\int\cdots\int}_{\{\mathbf{x}:\ h(\mathbf{x})\leq y;\ \mathbf{x}\in\mathbb{R}_X^n\}} f(\mathbf{x};\boldsymbol{\theta}^*)d\mathbf{x},\ \forall y{\in}\mathbb{R}. \tag{4}$$

It is important to emphasize that the sampling distribution of in (4) is derived under $\boldsymbol{\theta}{=}\boldsymbol{\theta}^*$, which renders the reasoning underlying estimation *factual*; the evaluation is under $\boldsymbol{\theta}^*$, whatever that value happens to be.

In light of the multiple integrals in (4), one can appreciate the difficulties of deriving such sampling distributions in the case of complicated functions of the sample. In cases of IID samples, however, $f(\mathbf{x};\boldsymbol{\theta})$ reduces to a product of marginal distributions and the the multiple integrals are easier to handle. In what follows, the results will be given in the form of lemmas to simply the discussion.

Example 11.4. For the simple Bernoulli model (table 11.1), let us derive the sampling distributions of the estimators (i)-(vi) in (2). In light of the fact that all these estimators are linear functions of the sample, we need to derive the sampling distribution of such linear functions, which are special cases of the following lemma.

Lemma 11.1. If $X_1, X_2, ..., X_n$ are Independent Identically Distributed (IID) Bernoulli distributed random variables with parameter θ:

$$X_k \backsim \mathsf{Ber}(\theta, \theta(1{-}\theta)),\ k{=}1,2,...,n,$$

then $Y_n{=}\sum_{k=1}^n X_k$ is Binomially distributed:

$$\sum\nolimits_{k=1}^n X_k{\backsim}\mathsf{Bin}(n\theta, n\theta(1{-}\theta); n).$$

Note that (i) the Bernoulli is a special case of the Binomial with $n{=}1$, and the above evaluation of the sampling distribution of Y_n is under true θ^*, the true value of θ in Θ, but to simplify the notation we leave out the star.

Armed with lemma 1, we can proceed to derive the sampling distributions of estimators (i)-(vi) which, from assumption [1] follows that they are all Binomially distributed, but we need to derive their particular means and variances. To illustrate how these moments are derived using the properties of $E(.)$ and $Var(.)$(chapter 3), let us focus on estimator (iv) $\widehat{\theta}_n{=}\frac{1}{n}\sum_{i=1}^n X_i$. Deriving its mean:

$$E(\widehat{\theta}_n){=}\tfrac{1}{n}E\left(\textstyle\sum_{i=1}^n X_i\right)=\tfrac{1}{n}\textstyle\sum_{i=1}^n E(X_i)\overset{[2]}{=}\tfrac{1}{n}\textstyle\sum_{i=1}^n \theta{=}(\tfrac{1}{n})(n\theta){=}\theta, \tag{5}$$

where the first and second equality stems from the linearity property (**E2**) of $E(.)$, and the third from assumption [2] (table 11.1). Deriving the variance:

$$Var(\widehat{\theta}_n){=}\left(\tfrac{1}{n}\right)^2 Var\left(\textstyle\sum_{i=1}^n X_i\right)\overset{[4]}{=}\left(\tfrac{1}{n}\right)^2\textstyle\sum_{i=1}^n Var(X_i)\overset{[3]}{=}\left(\tfrac{1}{n}\right)^2(n\theta(1{-}\theta))=\tfrac{1}{n}\theta(1{-}\theta), \tag{6}$$

where the first equality stems from the **V2** property of the variance, the second from assumption [4] and the third from assumption [3]; NOTE that if [4] were invalid, the $Var\left(\sum_{i=1}^n X_i\right)$ would have been of the form:

$$Var(\textstyle\sum_{i=1}^n X_i){=}\textstyle\sum_{i=1}^n Var(X_i){+}2\textstyle\sum_{i>j} Cov(X_iX_j),\ i,j{=}1,2,..,n.$$

Using similar derivations for the other estimators in (i)-(iii) that invoke assumptions [2]-[4] (table 11.1), one can show that their sampling distributions are:

$$\begin{array}{ll}\text{(i) }\widehat{\theta}_1\backsim\mathsf{Bin}(\theta,\theta(1{-}\theta);1) & \text{(iv) }\widehat{\theta}_n\backsim\mathsf{Bin}\left(\theta,\frac{\theta(1-\theta)}{n};n\right)\\ \text{(ii) }\widehat{\theta}_2\backsim\mathsf{Bin}\left(\theta,\frac{1}{2}\theta(1{-}\theta);2\right) & \text{(v) }\widehat{\theta}_{n+1}\backsim\mathsf{Bin}\left((\frac{n}{n+1})\theta,\frac{n\theta(1-\theta)}{(n+1)^2};n\right)\\ \text{(iii) }\widehat{\theta}_3\backsim\mathsf{Bin}\left(\theta,\frac{1}{3}\theta(1{-}\theta);3\right) & \text{(vi) }\widehat{\theta}_{n+2}\backsim\mathsf{Bin}\left((\frac{n}{n+2})\theta,\frac{n\theta(1-\theta)}{(n+2)^2};n\right)\end{array} \tag{7}$$

Example 11.5. For the simple Normal model (table 11.2), let us derive the sampling distributions of the estimators (i)-(vi) in (3). Since all these estimators are linear functions of the sample, we need to derive the sampling distribution of such linear functions, which are special cases of the following lemma.

Lemma 11.2. If $X_1, X_2, ..., X_n$ are NIID random variables with parameters $\boldsymbol{\theta}{:=}(\mu,\sigma^2)$:

$$X_k \backsim \mathsf{NIID}(\mu,\sigma^2),\ (\mu,\sigma^2){\in}\mathbb{R}\times\mathbb{R}_+,\ k{=}1,2,...,n,$$

then the function $Y_n{=}\sum_{k=1}^n X_k$ is also Normally distributed: $\sum_{i=1}^n X_k \backsim \mathsf{N}\left(n\mu, n\sigma^2\right)$.

Using lemma 11.2 together with assumption [1] (table 11.2), we can deduce that all the estimators (i)-(iv) in (3), will be Normally distributed. Moreover, using the same derivations as in (5)-(6), we can deduce that their sampling distributions are:

$$\begin{array}{lll}\text{(i) }\widehat{\mu}_1\backsim\mathsf{N}(\mu,\sigma^2) & \text{(iii) }\widehat{\mu}_3\backsim\mathsf{N}(0,\frac{\sigma^2}{2}) & \text{(v) }\widehat{\mu}_{n+1}\backsim\mathsf{N}\left((\frac{n}{n+1})\mu,\frac{n\sigma^2}{(n+1)^2}\right)\\ \text{(ii) }\widehat{\mu}_2\backsim\mathsf{N}(\mu,\frac{\sigma^2}{2}) & \text{(iv) }\widehat{\mu}_n\backsim\mathsf{N}(\mu,\frac{\sigma^2}{n}) & \text{(vi) }\widehat{\mu}_{n+2}\backsim\mathsf{N}\left((\frac{n}{n+2})\mu,\frac{n\sigma^2}{(n+2)^2}\right)\end{array} \tag{8}$$

Note that the above statements concerning the sampling distributions of different estimators should be interpreted as:

$$\widehat{\mu}_k \overset{\mu=\mu^*}{\backsim} \mathsf{N}(.,.),$$

to bring out the fact that these distributions are evaluated under μ^*, the 'true' value of θ in Θ. Similarly, the sampling distributions in (7) are evaluated under θ^*.

4 Finite sample properties of estimators

In an attempt to bring out the finite sample nature of the properties in this section, the generic estimator will be denoted by $\widehat{\theta}_n(\mathbf{X})$, to emphasize that these properties hold for any finite any $1 < n < \infty$. In light of the fact that a frequentist (point) estimator $\widehat{\theta}_n(\mathbf{X})$ aims to pin-point θ^*, and its optimality is evaluated by how effectively it achieves that, it is extremely important to emphasize at the outset that all finite sample frequentist properties for estimators are defined in relation to θ^*. In order to avoid cumbersome notation, however, in most derivations the superscript will be omitted when it is clear from the context. It is also important to bring out at the outset the fact that there is no single property that defines what an optimal estimator is, and thus optimality is defined using a combination of properties.

4.1 Unbiasedness

An estimator $\widehat{\theta}_n(\mathbf{X})$ is said to be an *unbiased estimator* of θ if the mean of its sampling distribution is equal to θ^*, the true value of θ in Θ:

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

It is worth re-iterating that $E(.)$ is with respect to the sampling distribution of $\widehat{\theta}_n(\mathbf{X})$ evaluated under $\theta=\theta^*$. Otherwise $\widehat{\theta}_n(\mathbf{X})$ is said to be *biased* with the bias defined by:

$$\text{Bias:}\quad \mathcal{B}(\widehat{\theta}_n(\mathbf{X});\theta^*)=E(\widehat{\theta}_n(\mathbf{X}))-\theta^*.$$

Note that both unbiasedness and the bias are defined at a point $\theta=\theta^*$, and not for all possible values of θ in Θ. The above definitions should be contrasted with $E(\widehat{\theta}_n(\mathbf{X}))=\theta,\ \forall\theta\in\Theta$, and $\mathcal{B}(\widehat{\theta}_n(\mathbf{X});\theta)=E(\widehat{\theta}_n(\mathbf{X}))-\theta,\ \forall\theta\in\Theta$, encountered in traditional textbooks. The quantifier $\forall\theta\in\Theta$ makes no sense from the frequentist perspective, but does make sense when viewed from the Bayesian/Decision-theoretic perspective; see Spanos (2017b).

Example 11.6. The sampling distributions in (7) in table 11.1 for the simple Bernoulli model, indicate that the estimators $\widehat{\theta}_1$, $\widehat{\theta}_2$, $\widehat{\theta}_3$ and $\widehat{\theta}_n$ are unbiased estimators of θ, but $\widehat{\theta}_{n+1}$ and $\widehat{\theta}_{n+2}$ are biased since:

$$\text{(v)}\quad \mathcal{B}(\widehat{\theta}_{n+1})=-\left(\tfrac{1}{n+1}\right)\theta,\quad \text{(vi)}\quad \mathcal{B}(\widehat{\theta}_{n+2})=-\left(\tfrac{1}{n+2}\right)\theta.$$

Example 11.7. For the simple Normal model (table 11.2), the sampling distributions in (8) of the estimators (i)-(vi) in (3) suggest that the estimators $\widehat{\mu}_1, \widehat{\mu}_2$ and $\widehat{\mu}_n$ are unbiased estimators of μ, but $\widehat{\mu}_3$, $\widehat{\mu}_{n+1}$ and $\widehat{\mu}_{n+2}$ are biased estimators of θ, and their bias is:

$$\text{(iii) } \mathcal{B}(\widehat{\mu}_3) = -\mu, \quad \text{(v) } \mathcal{B}(\widehat{\mu}_{n+1}) = -(\tfrac{1}{n+1})\mu, \quad \text{(vi) } \mathcal{B}(\widehat{\mu}_{n+2}) = -(\tfrac{1}{n+2})\mu.$$

The obvious question that arises from the examples 11.6-7 is:

▶ **Are the biased estimators inferior to the unbiased ones**?

The answer is not as obvious as it appears at first sight. This is because there is no single property that defines what an 'optimal' estimator is, and unbiasedness is not the most desirable property for an optimal estimator. As Fisher (1956) put it in a letter to C.R. Rao (one of his most famous students): "... lack of bias, which ... is not invariant for functional transformation of parameters has never had the least interest for me." (Bennett, 1990, p. 196). What is important to emphasize at the outset is that there are minimal properties, such as consistency, which are necessary for the optimality of an estimator, but never sufficient. Unbiasedness is not one of those minimal properties. Other properties relating to second moment of the sampling distribution of $\widehat{\theta}_n(\mathbf{X})$ are more important.

The notion of unbiasedness is intuitively appealing but is not without its problems.

1. Unbiased estimators do not always exist.

Example 11.8. Consider the *simple Exponential model* in table 11.3 with a density function:

$$f(x;\theta) = \theta\exp\{-\theta x\}, \ \theta > 0, \ x > 0.$$

Table 11.3: The simple Exponential model

Statistical GM:		$X_t = \frac{1}{\theta} + u_t, \ t \in \mathbb{N} := (1, 2, ..., n, ...)$
[1]	Exponential:	$X_t \backsim \mathsf{Exp}(.,.), \ x_t \in \mathbb{R}_+,$
[2]	Constant mean:	$E(X_t) = \frac{1}{\theta}, \ \theta \in \mathbb{R}, \ \forall t \in \mathbb{N},$
[3]	Constant variance:	$Var(X_t) = (1/\theta)^2, \ \forall t \in \mathbb{N},$
[4]	Independence:	$\{X_t, \ t \in \mathbb{N}\}$-independent process.

It can be shown that no unbiased estimator of θ exists; see Schervish (1995), p. 297. This example is directly related to the above comment by Fisher (1956).

2. Unbiasedness is not invariant to transformations of $\boldsymbol{\theta}$. Assuming that there exists an unbiased estimator $\widehat{\theta}_n(\mathbf{X})$ of θ, i.e., $E(\widehat{\theta}_n(\mathbf{X})) = \theta^*$, for $\varphi = g(\theta)$, where $g(.)$: $\Theta \to \Phi$, and $\widehat{\varphi}_n = g(\widehat{\theta}_n(\mathbf{X}))$, then in general:

$$E(\widehat{\varphi}_n) \neq \varphi.$$

Example 11.9. For the simple Exponential model (table 11.3), we have seen in example 11.8 that no unbiased estimator of θ exists, but we can show that when

θ is reparameterized into $\phi=(1/\theta)$, one can show that $\widehat{\phi}=\frac{1}{n}\sum_{i=1}^{n}X_i$ is an unbiased estimator of ϕ since assumption [2] (table 11.3) becomes $E(X_t)=\phi$ and thus $E(\widehat{\phi})\overset{[2]}{=}\frac{1}{n}\sum_{i=1}^{n}E(X_i)=\phi$. This is the reason why the Exponential density is often parameterized in terms of ϕ: $f(x;\phi)=(1/\phi)\exp\{-(x/\phi)\},\ \phi>0,\ x>0.$

4.2 Efficiency: relative vs. full efficiency

Given that a frequentist (point) estimator $\widehat{\theta}_n(\mathbf{X})$ aims to pin-point θ^*, and its optimality is evaluated by how effectively it achieves that, it is natural to extend the search for optimal properties to the *second moment* of the sampling distribution of $\widehat{\theta}_n(\mathbf{X})$. The second moment provides information about the dispersion of the sampling distribution around θ^*, defining the property referred to as *efficiency.* In practice, one needs to distinguish clearly between two different forms of efficiency, *relative* and *full efficiency*, that are sometimes confused with serious consequences. This is because by itself relative efficiency is of limited value, but full efficiency is priceless.

4.2.1 Relative efficiency

For any two unbiased estimators $\widehat{\theta}_1(\mathbf{X})$ and $\widehat{\theta}_2(\mathbf{X})$ of θ, $\widehat{\theta}_1(\mathbf{X})$ is said to be *relatively more efficient* than $\widehat{\theta}_2(\mathbf{X})$ if:

$$Var(\widehat{\theta}_1(\mathbf{X}))\leq Var(\widehat{\theta}_2(\mathbf{X}))$$

Example 11.10. For the simple Bernoulli model (table 11.1), the sampling distributions in (7) of the estimators (i)-(vi) in (2) indicate that $\widehat{\theta}_1,\widehat{\theta}_2$, and $\widehat{\theta}_3$ are unbiased estimators of θ, but in terms of relative efficiency we have a clear ordering based on their variances:

$$Var(\widehat{\theta}_3)=\tfrac{\theta(1-\theta)}{3}<Var(\widehat{\theta}_2)=\tfrac{\theta(1-\theta)}{2}<Var(\widehat{\theta}_1)=\tfrac{\theta(1-\theta)}{1},\ \text{for } n\geq 3$$

Example 11.11. For the simple Normal model (table 11.2), the sampling distributions in (8) of the estimators (i)-(vi) in (3) indicate that $\widehat{\mu}_1,\widehat{\mu}_2$, and $\widehat{\mu}_n$ are unbiased estimators of θ, but in terms of relative efficiency we have a clear ordering based on their variances:

$$Var(\widehat{\mu}_n)=\tfrac{\sigma^2}{n}<Var(\widehat{\mu}_2)=\tfrac{\sigma^2}{2}<Var(\widehat{\mu}_1)=\tfrac{\sigma^2}{1},\ \text{for } n\geq 3$$

The problem with using relative efficiency to compare different estimators is that the comparison is local in the sense that it all depends on the pool of estimators in hand. This is equivalent to my assertion that 'I am the best econometrician in my family'. That is true, but that does not make me a good econometrician because the pool of comparison is much too narrow. In the case of example 11.10, $\widehat{\theta}_3$ is the best on relative efficiency grounds, but it is a terrible estimator because it is *inconsistent* and the comparison is narrowed to a group of inconsistent estimators. Hence, the question that immediately comes to mind is:

▶ **is there a global notion of efficiency**? The surprising answer is *yes*!

4.2.2 Full efficiency: the Cramer-Rao lower bound

The challenge of devising an absolute lower bound was met successfully in the 1940s by two pioneers of modern frequentist statistics, Harald Cramér (1893–1985) and C.R. Rao (1920– $\cdots$), in Cramer (1946b) and Rao (1945). Using different approaches they both reached the conclusion that the global lower bound for unbiased estimators is related to *the Fisher information;* a concept introduced by Fisher (1922a).

Fisher's information of the sample. Like all results with any generality, the Fisher information of the sample and the associated Cramer-Rao lower bound applies to 'regular' statistical models, where regularity is defined in terms of certain restrictions pertaining the their distribution of the sample $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}{\in}\mathbb{R}_X^n$.

A statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is said to be regular if its $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}{\in}\mathbb{R}_X^n$ satisfies the following regularity conditions:

Table 11.4: Regularity for $\mathcal{M}_{\theta}(\mathbf{x}){=}\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}{\in}\Theta\},\ \mathbf{x}{\in}\mathbb{R}_X^n$

(**R1**)	The parameter space Θ is an open subset of $\mathbb{R}^m,\ m<n$,
(**R2**)	the *support* of **X**, $\mathbb{R}_X^n{:=}\{\mathbf{x}{:}\ f(\mathbf{x};\boldsymbol{\theta}){>}0\}$, is the same $\forall\boldsymbol{\theta}{\in}\Theta$,
(**R3**)	$(\frac{\partial\ln f(\mathbf{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}})<\infty$, exists and is finite $\forall\boldsymbol{\theta}{\in}\Theta,\ \forall\mathbf{x}{\in}\mathbb{R}_X^n$,
(**R4**)	for any Borel function $h(\mathbf{X})$ differentiation and integration can be interchanged: $\frac{d}{d\boldsymbol{\theta}}\left(\int\cdots\int h(\mathbf{x})\cdot f(\mathbf{x};\boldsymbol{\theta})d\mathbf{x}\right)=\ \int\cdots\int h(\mathbf{x})\left[\frac{\partial}{\partial\boldsymbol{\theta}}f(\mathbf{x};\boldsymbol{\theta})\right]d\mathbf{x}<\infty.$

R1 excludes boundary points to ensure that derivatives (from both sides of a point) exist, and **R1** ensures that the support of $f(\mathbf{x};\boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$. For such regular probability models we can proceed to define the Fisher information of the sample which is designed to provide a measure of the information rendered by the sample for a parameter $\theta{\in}\Theta$.

Focusing on the case of a single parameter θ, the **Fisher information** of the sample $\mathbf{X}{:=}(X_1,X_2,...,X_n)$ is defined by:

$$\mathcal{I}_n(\theta){:=}E\left\{\left(\tfrac{d\ln f(\mathbf{x};\theta)}{d\theta}\right)^2\right\} \tag{9}$$

There are several things to NOTE about this concept.

(a) Under the *regularity conditions* (**R1**)-(**R4**) it can be shown that:

$$\mathcal{I}_n(\theta){:=}E\left\{\left(\tfrac{d\ln f(\mathbf{x};\theta)}{d\theta}\right)^2\right\}{=}E\left(-\tfrac{d^2\ln f(\mathbf{x};\theta)}{d\theta^2}\right) \tag{10}$$

This often provides a more convenient way to derive Fisher's information and thus the Cramer-Rao lower bound.

(b) The form of $\mathcal{I}_n(\theta)$ depends crucially on the sampling model. For example, in the case of an independent sample:

$$E\left(\tfrac{d\ln f(\mathbf{x};\theta)}{d\theta}\right)=\textstyle\sum_{i=1}^n E\left(\tfrac{d\ln f(x_i;\theta)}{d\theta}\right)$$

and in the IID sample case Fisher's information takes the even simpler form:

$$\mathcal{I}_n(\theta)=n\mathcal{I}(\theta):=nE\left\{\left(\tfrac{d\ln f(x;\theta)}{d\theta}\right)^2\right\},$$

in an obvious notation, where $f(x;\theta)$ denotes the marginal density function and $\mathcal{I}(\theta)$ represents the Fisher information of *a single observation.*

Example 11.12. For the simple Normal model (table 11.2):

$$f(x;\theta)=\tfrac{1}{\sqrt{2\pi}}\exp\{-\tfrac{(x-\theta)^2}{2\sigma^2}\},\ \tfrac{d}{d\theta}\ln f(x;\theta)=\tfrac{(x-\theta)}{\sigma^2}\implies\mathcal{I}(\theta)=1\text{ and }\mathcal{I}_n(\theta)=n.$$

This example suggests that as the information increases, the Fisher information of the sample $\mathcal{I}_n(\theta)$ increases and thus more information about θ is gained.

Cramer-Rao lower bound. Assuming that the Fisher information of the sample exists and $\mathcal{I}_n(\theta)>0$, $\forall\theta\in\Theta$, the variance of any unbiased estimator of a parameter θ, say $\widehat{\theta}_n(\mathbf{X})$, cannot be smaller than the inverse of $\mathcal{I}_n(\theta)$:

$$Var(\widehat{\theta})\geq\textsf{C-R}(\theta)=\mathcal{I}_n^{-1}(\theta).\tag{11}$$

In the case where one is interested in some differentiable function of θ, say $q(\theta)$, and $\widehat{q}(\theta)$ is an estimator of $q(\theta)$, the Cramer-Rao lower bound takes the form:

$$Var(\widehat{q}(\theta))\geq\textsf{C-R}(q(\theta))=\left(\tfrac{d}{d\theta}E(q(\theta))\right)^2\mathcal{I}_n^{-1}(\theta).\tag{12}$$

General C-R lower bound. Using (12) we can extend the Cramer-Rao lower bound to the case of any estimator $\widehat{\theta}_n(\mathbf{X})$ of θ, including biased ones:

$$Var(\widehat{\theta}_n(\mathbf{X}))\geq\textsf{GC-R}(\theta)=\left(\tfrac{dE(\widehat{\theta}_n(\mathbf{X}))}{d\theta}\right)^2\left\{E\left(\tfrac{d\ln f(\mathbf{x};\theta)}{d\theta}\right)^2\right\}^{-1}.\tag{13}$$

The following example illustrates the role of condition **R2** in deriving the C-R lower bound.

Example 11.13. Consider the *simple Uniform model:*

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})\text{: }X_t\backsim\textsf{UIID}(0,\theta),\ 0<x_k<\theta,\ \theta>0,\ k\in\mathbb{N},\tag{14}$$

whose density function takes the form:

$$f(x;\theta)=(\tfrac{1}{\theta}),\ \theta>0,\ 0<x<\theta;\ E(X_k)=(\tfrac{\theta}{2}),Var(X_k)=(\tfrac{\theta^2}{12}),\ \forall t\in\mathbb{N}.$$

In this case, the range of X depends on θ, and the regularity condition **R2** is not satisfied. If we were to derive the C-R lower bound, by ignoring this regularity problem, we will get very misleading results:

$$\tfrac{d\ln f(\mathbf{x};\theta)}{d\theta}=\tfrac{d}{d\theta}\left(-n\ln\theta\right)=-\tfrac{n}{\theta}\implies\mathcal{I}_n(\theta)=E\left(\tfrac{d\ln f(\mathbf{x};\theta)}{d\theta}\right)^2=\left(\tfrac{n}{\theta}\right)^2.$$

This is because there are other unbiased estimators of θ, say $\widehat{\theta}_{[n]}(\mathbf{X}){=}\left(\frac{n+1}{n}\right)\max\left(\mathbf{X}\right),$ where $\max\left(\mathbf{X}\right)$ is the maximum order statistic, whose variance is smaller than $\mathcal{I}_n^{-1}(\theta)$ (Cassela and Berger, 2002) since:

$$Var\left(\widehat{\theta}_{[n]}(\mathbf{X})\right){=}\tfrac{\theta^2}{n(n+2)}<\mathcal{I}_n^{-1}(\theta){=}\tfrac{\theta^2}{n^2}.$$

The equality in (10) offers a simple way to confirm that a statistical model is non-regular. For this statistical model in table 11.5:

$$\tfrac{d^2\ln f(\mathbf{x};\theta)}{d\theta^2}{=}\tfrac{n}{\theta^2}\implies E(-\tfrac{d^2\ln f(\mathbf{x};\theta)}{d\theta^2}){=}-\tfrac{n}{\theta^2}\neq\tfrac{\theta^2}{n^2}{=}E\left(\tfrac{d\ln f(\mathbf{x};\theta)}{d\theta}\right)^2.$$

Indeed, $E(-\frac{d^2\ln f(\mathbf{x};\theta)}{d\theta^2}){=}-\frac{n}{\theta^2}{<}0$ makes no statistical sense. Hence, using the Cramer-Rao lower bound in the case of non-regular statistical models will give rise to misleading results.

Table 11.5: Non-Regular Probability models: R2 invalid

Name	Density	Parameter space	Support
Uniform $\mathsf{U}(0,\theta)$:	$f(x;\theta){=}\frac{1}{\theta}$,	$\theta{\in}\mathbb{R}_+$ $(\theta>0)$,	$0<x<\theta$
Uniform $\mathsf{U}(-\frac{\theta}{2},\frac{\theta}{2})$:	$f(x;\theta){=}\frac{1}{\theta}$,	$\theta{\in}\mathbb{R}_+$ $(\theta>0)$,	$-\frac{\theta}{2}{<}x{<}\frac{\theta}{2}$
Uniform $\mathsf{U}(\theta{-}\frac{1}{2},\theta{+}\frac{1}{2})$:	$f(x;\theta){=}1$,	$\theta{\in}\mathbb{R}_+$ $(\theta>0)$,	$\theta{-}\frac{1}{2}{<}x{<}\theta{+}\frac{1}{2}$
Uniform (discr.) $\mathsf{Ud}(N)$:	$f(x;\theta){=}\frac{1}{N}$,	$N\geq 1$,	$x{=}1,2,..,N$
Triangular: $\mathsf{Triang}(\theta)$:	$f(x;\theta){=}1{-}\left\|x{-}\theta\right\|$,	$0\leq\theta\leq 1$,	$0<x<\theta$,
Location $\mathsf{Loc}(\theta)$:	$f(x;\theta){=}e^{-(x-\theta)}$,	$\theta{\in}\mathbb{R}_+$ $(\theta>0)$,	$x>\theta$,
Location/scale $\mathsf{LS}(\theta,\sigma)$:	$f(x;\theta){=}\frac{1}{\sigma}e^{-(\frac{x-\theta}{\sigma})}$,	$\theta>0,\ \sigma>0$,	$x>\theta$.

Full efficiency. An estimator $\widehat{\theta}_n(\mathbf{X})$ of θ is said to be a *fully efficient* if its variance is equal to the *Cramer-Rao lower bound*:

$$\begin{aligned}&\textsf{Unbiased:}\quad Var(\widehat{\theta}_n(\mathbf{X})){=}\textsf{C-R}(\theta){=}\mathcal{I}_n^{-1}(\theta){=}\left\{E\left(\tfrac{d\ln f(\mathbf{x};\theta)}{d\theta}\right)^2\right\}^{-1},\\&\textsf{Biased:}\quad Var(\widehat{\theta}_n(\mathbf{X})){=}\textsf{GC-R}(\theta){=}\left(\tfrac{dE(\widehat{\theta}_n(\mathbf{X}))}{d\theta}\right)^2\left\{E\left(\tfrac{d\ln f(\mathbf{x};\theta)}{d\theta}\right)^2\right\}^{-1}.\end{aligned}\tag{15}$$

Necessary and sufficient condition. An estimator $\widehat{\theta}_n(\mathbf{X})$ satisfies (15) iff:

$$(\widehat{\theta}_n(\mathbf{X})-\theta){=}h(\theta)\left[\tfrac{d\ln f(\mathbf{x};\theta)}{d\theta}\right],\tag{16}$$

for some function $h(\theta),\ \forall\theta{\in}\Theta$; see Casella and Berger (2002).

Example 11.14. For the simple Normal model (table 11.2), $f(\mathbf{x};\mu)$ and $\ln f(\mathbf{x};\mu)$ are:

$$f(\mathbf{x};\mu)=(\tfrac{1}{\sqrt{2\pi\sigma^2}})^n\exp\{-\tfrac{1}{2}\sum_{t=1}^{n}\tfrac{(x_t-\mu)^2}{\sigma^2}\},\quad \ln f(\mathbf{x};\mu)=-\tfrac{n}{2}\ln(2\pi\sigma^2)-\tfrac{1}{2}\sum_{t=1}^{n}\tfrac{(x_t-\mu)^2}{\sigma^2}.$$

Given that the first and second derivatives are:

$$\tfrac{d\ln f(\mathbf{x};\mu)}{d\mu}=\sum_{i=1}^{n}\tfrac{(x_i-\mu)}{\sigma^2},\ \ \tfrac{d^2\ln f(\mathbf{x};\mu)}{d\mu^2}=\text{-}\tfrac{n}{\sigma^2}\implies \mathcal{I}_n(\mu)=\tfrac{n}{\sigma^2},$$

and thus, the Cramer-Rao lower bound for μ is $\mathsf{C\text{-}R}(\mu)=\frac{\sigma^2}{n}$. This confirms that $\widehat{\mu}_n=\frac{1}{n}\sum_{i=1}^{n}X_i$ of μ is both unbiased and fully efficient. Notice that (16) holds in this case since:

$$(\widehat{\mu}_n-\mu)=\tfrac{1}{n}\left[\tfrac{d\ln f(\mathbf{x};\mu)}{d\mu}\right]=\tfrac{1}{n}\sum_{i=1}^{n}\tfrac{(x_i-\mu)}{\sigma^2}.$$

Example 11.15. For the simple Bernoulli model (table 11.1), $f(\mathbf{x};\theta)$ is:

$$f(\mathbf{x};\theta)=\theta^{\sum_{i=1}^{n}x_i}(1-\theta)^{\sum_{i=1}^{n}(1-x_i)}\Rightarrow\ \ \ln f(\mathbf{x};\theta)=(\textstyle\sum_{i=1}^{n}X_i)\ln\theta+(\sum_{i=1}^{n}[1-X_i])\ln(1-\theta).$$

On the basis of the first and second derivatives:

$$\tfrac{d\ln f(\mathbf{x};\theta)}{d\theta}=(\textstyle\sum_{i=1}^{n}X_i)(\tfrac{1}{\theta})-([\sum_{i=1}^{n}(1-X_i)](\tfrac{1}{1-\theta}),$$

$$\tfrac{d^2\ln f(\mathbf{x};\theta)}{d\theta^2}=\text{-}(\textstyle\sum_{i=1}^{n}X_i)(\tfrac{1}{\theta^2})-[\sum_{i=1}^{n}(1-X_i)](\tfrac{1}{1-\theta})^2,$$

$$\mathcal{I}_n(\theta)=E\left(-\tfrac{d^2\ln f(\mathbf{x};\theta)}{d\theta^2}\right)=\tfrac{n}{\theta(1-\theta)}\implies \mathsf{C\text{-}R}(\theta)=\mathcal{I}_n^{-1}(\theta)=\tfrac{\theta(1-\theta)}{n}.$$

This confirms that $\widehat{\theta}_n=\frac{1}{n}\sum_{i=1}^{n}X_i$ with a sampling distribution $\widehat{\theta}_n\backsim\mathsf{Bin}(\theta,\frac{\theta(1-\theta)}{n};n)$ is both unbiased and fully efficient.

4.3 Minimum MSE estimators and admissibility

The above measures of efficiency enable us to choose between unbiased estimators on one hand and biased estimators on the other using (11) and (13), respectively. These lower bounds, however, offer no guidance on the question of choosing between a biased and an unbiased estimator. Such a comparison is of interest in practice because fully efficient and unbiased estimators do not always exist and unbiased estimators are not always good estimators. There are circumstances where a biased estimator is preferable to an unbiased one. The question then is: ▶ How do we compare biased and unbiased estimators?

It makes sense to penalize a biased estimator in order the derive a dispersion measure that makes a biased and an unbiased estimator comparable. The most widely used such measure is the *Mean Square Error*:

$$\mathsf{MSE}(\widehat{\theta}_n(\mathbf{X});\theta^*)=E\{(\widehat{\theta}_n(\mathbf{X})-\theta^*)^2\}.\tag{17}$$

It is very important to emphasize that the MSE for frequentist estimators, like the bias, is defined at the point $\theta{=}\theta^*$. It can be shown to represent a penalized variance:

$$\mathsf{MSE}(\widehat{\theta}_n(\mathbf{X});\theta^*){=}Var(\widehat{\theta}_n(\mathbf{X})){+}[\mathcal{B}(\widehat{\theta}_n(\mathbf{X});\theta^*)]^2, \tag{18}$$

with the penally equal to the square of the $\mathcal{B}(\widehat{\theta}_n;\theta^*){=}E(\widehat{\theta}_n){-}\theta^*$. This can be derived by adding and subtracting $E(\widehat{\theta}_n){=}\theta_m$ in the definition of the MSE as follows:

$$\begin{aligned}\mathsf{MSE}(\widehat{\theta}_n;\theta^*) &= E\{[(\widehat{\theta}_n-\theta_m)+(\theta_m-\theta^*)]^2\}= \\ &= E[\widehat{\theta}_n{-}\theta_m]^2{+}2[\theta_m{-}\theta^*]E[\widehat{\theta}_n{-}\theta_m]{+}[\theta_m{-}\theta^*]^2= \\ &= E[\widehat{\theta}_n{-}\theta_m]^2+[\theta_m{-}\theta^*]^2{=}Var(\widehat{\theta}_n)+[\mathcal{B}(\widehat{\theta}_n;\theta^*)]^2.\end{aligned} \tag{19}$$

Minimum MSE estimator. An estimator $\widehat{\theta}_n(\mathbf{X})$ is said to be a *minimum MSE estimator* of θ if:

$$\mathsf{MSE}(\widehat{\theta}_n(\mathbf{X});\theta^*)\leq\mathsf{MSE}(\widetilde{\theta}_n(\mathbf{X});\theta^*),$$

for any other estimator $\widetilde{\theta}_n(\mathbf{X})$.

Example 11.21. For the simple Bernoulli model (table 11.1), the estimators $(\widehat{\theta}_{n+1},\widehat{\theta}_{n+2})$ are preferable in terms of their MSE than $\widehat{\theta}_1$, $\widehat{\theta}_2$ and $\widehat{\theta}_3$ since for $n>3$:

$$\begin{aligned}\mathsf{MSE}(\widehat{\theta}_{n+1}) &= (\tfrac{n}{(n+1)^2})\theta(1{-}\theta){+}(\tfrac{-\theta}{(n+1)})^2{=}\tfrac{n\theta(1-\theta)+\theta^2}{(n+1)^2}\leq\mathsf{MSE}(\widehat{\theta}_k){=}\tfrac{\theta(1-\theta)}{k},\ \ k{=}1,2,3, \\ \mathsf{MSE}(\widehat{\theta}_{n+2}) &= (\tfrac{n}{(n+2)^2})\theta(1{-}\theta){+}(\tfrac{-\theta}{(n+2)})^2{=}\tfrac{n\theta(1-\theta)+\theta^2}{(n+2)^2}\leq\mathsf{MSE}(\widehat{\theta}_k){=}\tfrac{\theta(1-\theta)}{k},\ \ k{=}1,2,3.\end{aligned} \tag{20}$$

Cautionary note. The frequentist definition of the MSE in (18) should be contrasted with a non-frequentist definition in some textbooks defined for all θ in Θ $[\forall\theta{\in}\Theta]$:

$$MSE(\widehat{\theta}){=}E(\widehat{\theta}-\theta)^2,\ \ \forall\theta{\in}\Theta. \tag{21}$$

In light of the fact that both the bias $\mathcal{B}(\widehat{\theta}_n;\theta^*){=}E(\widehat{\theta}_n){-}\theta^*$ and $Var(\widehat{\theta}_n){=}E(\widehat{\theta}_n{-}\theta_m)^2$ involve only two particular values of θ in Θ, i.e. $\theta_m{=}E(\widehat{\theta})$ and θ^*, defining $\mathcal{B}(\widehat{\theta}_n;\theta)$ and $Var(\widehat{\theta}_n)$ for all $\theta{\in}\Theta$, makes no frequentist sense, but it does make sense in both Bayesian and Decision-theoretic statistics. The problem is that the non-frequentist definition of the MSE in (21) is employed to define another property of estimators known as *admissibility.*

Admissibility. An estimator $\widetilde{\theta}_n(\mathbf{X})$ is *inadmissible* if there exists another estimator $\widehat{\theta}_n(\mathbf{X})$ such that:

$$MSE(\widehat{\theta}_n,\theta)\leq MSE(\widetilde{\theta}_n,\theta),\ \forall\theta{\in}\Theta, \tag{22}$$

and the strict inequality ($<$) holds for at least one value of θ. Otherwise, $\widetilde{\theta}(\mathbf{X})$ is said to be *admissible* with respect to the loss function $L_2(\widehat{\theta}_n,\theta){=}(\widehat{\theta}_n{-}\theta)^2$, $\forall\theta{\in}\Theta$.

Example 11.21 (continued). Comparing the estimators in terms of their MSEs:

$$MSE(\widehat{\theta}_{n+1}) > MSE(\widehat{\theta}_{n+2}) \text{ for any } n > 1,$$

indicating that $\widehat{\theta}_{n+1}$ is inadmissible, but does that make $\widehat{\theta}_{n+2}$ a good estimator? Not necessarily because this is only a relative comparison since admissibility is a form of *relative efficiency;* with respect to a particular loss function. In the above case the loss function is quadratic.

Admissibility can be highly misleading as a finite sample property because it can be shown that estimators with excellent optimal frequentist properties are *not* always optimal in terms of admissibility. This is because the nature of the quantifier $\forall\theta\in\Theta$ associated with admissibility in (22) has no bearing on the effectiveness of frequentist estimation that focuses on the capacity of an estimator to pinpoint $\theta{=}\theta^*$. Hence, admissibility is *not* an interesting property for a frequentist estimator because it concerns values of θ other than the true value θ^*. Indeed, a moment's reflection suggests that there is something wrong-headed about the use of the quantifier $\forall\theta\in\Theta$ in (21) because it gives rise to dubious results when viewed from the frequentist perspective. The factual nature of frequentist reasoning in estimation also brings out the inappropriateness of the notion of admissibility as a *minimal property* (necessary but not sufficient). To bring out this problem consider the following example.

Example 11.22. In the context of the simple Normal model in (table 11.2), let us compare two estimators of θ in terms of admisibility:

(i) the unbiased and fully efficient: $\overline{X}_n{=}\frac{1}{n}\sum_{k=1}^{n} X_k$,

(ii) the 'crystalball' estimator: $\mu_{cb}{=}7405926, \ \forall\mathbf{x}\in\mathbb{R}_X^n$.

When compared on admissibility grounds, these two estimators are both *admissible* since neither dominates the other on MSE grounds:

$$MSE(\overline{X}_n, \mu) \gtrless MSE(\mu_{cb}, \mu), \ \forall\mu\in\mathbb{R},$$

and thus equally acceptable. Common sense, however, suggests that if a particular criterion of optimality cannot distinguish between $\overline{X}_n$ [the best estimator] and θ_{cb}, a number picked from the air *that ignores the data altogether*, it is not much of a minimal property. A moment's reflection suggests that its inappropriateness stems from the reliance of admissibility on the quantifier '$\forall\theta\in\Theta$'. The admissibility of θ_{cb} stems from the fact that for certain values of θ close to θ_{cb}, say $\theta\in(\theta_{cb}\pm\frac{\lambda}{\sqrt{n}})$, for $0 < \lambda < 1$, on MSE grounds θ_{cb} is 'better' than $\overline{X}_n$:

$$MSE(\overline{X}_n; \theta){=}\tfrac{1}{n} > MSE(\theta_{cb}; \theta) \leq \tfrac{\lambda^2}{n} \text{ for } \theta\in(\theta_{cb}\pm\tfrac{\lambda}{\sqrt{n}}).$$

This example indicates that admissibility is totally ineffective as a *minimal* property because it does not filter out exceptionally "bad" estimators such as θ_{cb}! Instead, it did exclude potentially good estimators like the *sample median*; see Cox and Hinkley (1974). This highlights the extreme relativism of admissibility to the particular loss function, i.e. $L_2(\widehat{\theta}(\mathbf{X});\theta)$, since, as mentioned above, the sample median would have been the optimal estimator in the case of the absolute loss function $L_1(\widehat{\theta}(\mathbf{X});\theta)=|\widehat{\theta}(\mathbf{X})-\theta|$.

4.3.1 Full Efficiency vs. MSE and biased estimators

C-R lower bound comparison. Let us return to the comparison of estimators in terms of the C-R lower bound that allows for biased estimators.

Example 11.23. In the context of the simple Bernoulli model in (table 11.1), let us compare the estimators $(\widehat{\theta}_{n+1},\widehat{\theta}_{n+2})$ in terms of their respective Cramer-Rao lower bounds. Given that:

$$E(\widehat{\theta}_{n+1})=(\tfrac{n}{n+1})\theta,\ \tfrac{dE(\widehat{\theta}_{n+1})}{d\theta}=(\tfrac{n}{n+1}),\quad E(\widehat{\theta}_{n+2})=(\tfrac{n}{n+2})\theta,\ \tfrac{dE(\widehat{\theta}_{n+2})}{d\theta}=(\tfrac{n}{n+2}).$$

their respective the Cramer-Rao lower bounds based on (13) are:

$$\mathsf{C\text{-}R}(\widehat{\theta}_{n+1})=(\tfrac{n}{n+1})^2(\tfrac{\theta(1-\theta)}{n})=\tfrac{n\theta(1-\theta)}{(n+1)^2}<\mathsf{MSE}(\widehat{\theta}_{n+1}),$$
$$\mathsf{C\text{-}R}(\widehat{\theta}_{n+2})=(\tfrac{n}{n+2})^2(\tfrac{\theta(1-\theta)}{n})=\tfrac{n\theta(1-\theta)}{(n+2)^2}<\mathsf{MSE}(\widehat{\theta}_{n+1}).$$

Returning to the MSE comparison between the unbiased estimators $\widehat{\theta}_1$, $\widehat{\theta}_2$ and $\widehat{\theta}_3$ and the biased estimators $\widehat{\theta}_{n+1}$ and $\widehat{\theta}_{n+2}$ in (20), there is another sense in which $(\widehat{\theta}_{n+1},\widehat{\theta}_{n+2})$ are much better estimators than $(\widehat{\theta}_1,\widehat{\theta}_2,\widehat{\theta}_3)$. A closer look at the variances of $(\widehat{\theta}_1,\widehat{\theta}_2,\widehat{\theta}_3)$ are not just bigger than those of $(\widehat{\theta}_{n+1},\widehat{\theta}_{n+2})$, but they do not decrease as n increases. In contrast, the MSE of both $(\widehat{\theta}_{n+1},\widehat{\theta}_{n+2})$ goes to zero as $n\rightarrow\infty$.

This brings up the asymptotic properties of estimators. As argued in chapter 9, the asymptotic properties for estimators amount to extending the limit theorems WLLN, SLLN, CLT that hold for the function $\sum_{k=1}^{n}X_k$, to more general functions $\widehat{\theta}_n{:=}h(X_1,X_2,...,X_n)$.

5 Asymptotic properties of estimators

Since the target of an ideal estimator θ^* with a degenerate sampling distribution $\mathbb{P}(\theta_n^*(\mathbf{X}){=}\theta){=}1$ is impossible to achieve with a sample size $n<\infty$, it makes sense to seek estimators that achieve this ideal as $n\rightarrow\infty$. In particular, when an estimator $\widehat{\theta}_n(\mathbf{X}){=}h(X_1,X_2,...,X_n)$ converges to θ^* almost surely:

$$\mathbb{P}(\lim_{n\rightarrow\infty}\widehat{\theta}_n(\mathbf{X}){=}\theta){=}1,$$

which is an extension of the SLLN; see chapter 9. In view of the fact that the probabilistic convergence results associated with the WLLN, SLLN and the CLT can only inform one about what happens at the limit $n{=}\infty$, why should the asymptotic properties of weak and strong consistency, and asymptotic Normality, extending these limit theorem results, be relevant for any inference based on a given $n < \infty$?

Common sense suggests that an estimator $\widehat{\theta}_n(\mathbf{X})$ whose aptitude to pin-point θ^* is enhanced as n increases is preferable to one whose capacity is the same whether $n{=}3$ or $n{=}3^{10}$. We call such properties *asymptotic* because, in contrast to the above finite sample properties which relate to the finite sampling distribution $f(\widehat{\theta}_n; \mathbf{x})$, they relate to the sequence of sampling distributions $\{f(\widehat{\theta}_n; \mathbf{x})\}_{n=1}^{\infty}$.

5.1 Consistency (weak)

An estimator $\widehat{\theta}_n$ is said to be a (weakly) *consistent* estimator of θ, if for any $\varepsilon > 0$:

$$\lim_{n\rightarrow\infty} \mathbb{P}(\left|\widehat{\theta}_n - \theta^*\right| < \varepsilon){=}1, \text{ and denoted by: } \widehat{\theta}_n \overset{\mathbb{P}}{\rightarrow} \theta^*. \tag{23}$$

This reads "the limit of the probability of the event that $\widehat{\theta}_n$ differs from θ^*, the true θ, by less than some positive constant $\varepsilon > 0$, goes to one as n goes to infinity"; see chapter 9.

REMARKS: (i) To avoid messy notation in what follows, the star $(^*)$ in θ^* will be dropped unless it is important to highlight its role.

(ii) $\widehat{\theta}_n$ in this definition stands for a generic estimator; the sub-script n is used to emphasize the role of the sample size.

(iii) In general, verifying (23) is non-trivial, but in one case where $\widehat{\theta}_n$ has a bounded variance, it can be easily verified using *Chebyshev's inequality* (see Appendix 9.A):

$$\mathbb{P}(\left|\widehat{\theta}_n - \theta\right| \leq \varepsilon) \geq 1 - \tfrac{E(\widehat{\theta}_n - \theta)^2}{\varepsilon^2}.$$

This is because $E(\widehat{\theta}_n - \theta)^2 {=} \mathsf{MSE}(\widehat{\theta}_n)$, and thus when $\mathsf{MSE}(\widehat{\theta}_n) \underset{n\rightarrow\infty}{\rightarrow} 0$, then $\frac{E(\widehat{\theta}_n - \theta)^2}{\varepsilon^2} \underset{n\rightarrow\infty}{\rightarrow} 0$ and (23) holds. From (18) we know that:

$$MSE(\widehat{\theta}_n) \rightarrow 0 \ \text{ if } Var(\widehat{\theta}_n) \rightarrow 0 \text{ and } \mathcal{B}(\widehat{\theta}_n; \theta) \rightarrow 0.$$

This suggests two easily verifiable conditions for $\widehat{\theta}_n$ to be a *consistent* estimator of θ when the required moments of its sampling distribution exist:

$$\text{(a) } \lim_{n\rightarrow\infty} \ E(\widehat{\theta}_n){=}\theta, \quad \text{(b) } \lim_{n\rightarrow\infty} \ Var(\widehat{\theta}_n){=}0.$$

This suggests that in the case where $Var(\widehat{\theta}_n) < \infty$, we can verify the consistency of $\widehat{\theta}_n$ by checking the above (sufficient) conditions. The notion of consistency based on (a)-(b) is sometimes called *mean-square consistency.*

Example 11.24. In the context of the simple Bernoulli model in (table 11.1), the estimators $\widehat{\theta}_1$, $\widehat{\theta}_2$ and $\widehat{\theta}_3$ satisfy (a) but not (b) because their variances do not decrease as $n\rightarrow\infty$. automatically, and thus they are all *inconsistent.* In contrast, the estimators $(\widehat{\theta}_{n+1},\widehat{\theta}_{n+2})$, are consistent because they satisfy both (a) and (b) (verify!). The estimators $\widehat{\theta}_1$, $\widehat{\theta}_2$ and $\widehat{\theta}_3$, being inconsistent, can be eliminated from the list of good estimators of θ and the choice is now between $\widehat{\theta}_n$, $\widehat{\theta}_{n+1}$ and $\widehat{\theta}_{n+2}$. Given that $\widehat{\theta}_n$ is both unbiased and fully efficient but $(\widehat{\theta}_{n+1},\widehat{\theta}_{n+2})$ are biased, there is a slight preference for $\widehat{\theta}_n$, at least on intuitive grounds, unless the gain from the MSE is sizeable enough.

Consistency as a minimal property. It is important to emphasize the fact that consistency is a *minimal* (necessary but not sufficient) property. That is, when an estimator is inconsistent it is not worth serious consideration, but since consistency holds at the limit ($n{=}\infty$), it should be accompanied by certain other desirable finite sample properties to define a 'good' estimator for inference purposes. There are numerous examples of just consistent estimators that are practically useless.

Example 11.25. In the case of the simple Normal model (table 11.2), $\overline{X}_n{=}\frac{1}{n}\sum_{k=1}^{n}X_k$ is a consistent estimator of μ, but so are the following two estimators of μ:

$$\overline{X}_n^{\ddagger}{=}\left\{\begin{array}{ll}0 & \text{for } n\leq 10^{24}\\ \overline{X}_n & \text{for } n>10^{24}\end{array}\right.,\ \overline{X}_n^{\sharp}{=}\left(\tfrac{n-10^5}{n-10^7}\right)\overline{X}_n.$$

Although $\overline{X}_n^{\ddagger}$ and $\overline{X}_n^{\sharp}$ are consistent estimators of μ, they are practically useless! Hence, in practice, one needs to supplement consistency with *finite sample* properties; see chapter 12. Let us illustrate that.

Example 11.26. For the simple Normal model (table 11.2), a summary of the results on unbiasedness and consistency is given in table 11.6. The general conclusion is that unbiased but inconsistent estimators are not good estimators.

Table 11.6: Unbiased and Consistent estimators

Estimator:	Unbiased	Consistent
(i) $\widehat{\mu}_1{=}X_1$:	$E(\widehat{\mu}_1){=}\mu,\ \ \surd$	$Var(\widehat{\mu}_1){=}\sigma^2,\ \ \times$
(ii) $\widehat{\mu}_3{=}\frac{1}{2}(X_1{-}X_n)$:	$E(\widehat{\mu}_2){=}\mu,\ \ \surd$	$Var(\widehat{\mu}_2){=}\frac{\sigma^2}{2},\ \ \times$
(iii) $\widehat{\mu}_3{=}\frac{1}{2}(X_1{-}X_n)$:	$E(\widehat{\mu}_3){=}0,\ \ \times$	$Var(\widehat{\mu}_3){=}\frac{\sigma^2}{2},\ \ \times$
(iv) $\widehat{\mu}_n{=}\frac{1}{n}\sum_{i=1}^{n}X_i$:	$E(\widehat{\mu}_n){=}\mu,\ \ \surd$	$Var(\widehat{\mu}_n){=}\frac{\sigma^2}{n},\ \ \surd$
(v) $\widehat{\mu}_{n+1}{=}(\frac{1}{n+1})\sum_{i=1}^{n}X_i$:	$E(\widehat{\mu}_{n+1}){=}\left(\frac{n}{n+1}\right)\mu,\ \ \times$	$Var(\widehat{\mu}_{n+1}){=}\frac{n\sigma^2}{(n+1)^2},\ \ \surd$
(vi) $\widehat{\mu}_{n+2}{=}(\frac{1}{n+2})\sum_{i=1}^{n}X_i$:	$E(\widehat{\mu}_{n+2}){=}\left(\frac{n}{n+2}\right)\mu,\ \ \times$	$Var(\widehat{\mu}_{n+1}){=}\frac{n\sigma^2}{(n+2)^2},\ \ \surd$

From the above comparison we can conclude that $\widehat{\mu}_n{=}\frac{1}{n}\sum_{i=1}^{n}X_i$ is the best estimator of μ because it is unbiased, consistent and fully efficient.

It is important to note that in the case of the above examples (and in most cases in practice), we utilize only their first two moments when deciding the optimality of the various estimators. That is, higher moments or other features of the sampling distribution are not explicitly utilized. For statistical inference purposes in general, however, we usually require the sample distribution itself, not just its first two moments.

5.2 Consistency (strong)

An estimator $\widehat{\theta}_n$ is said to be a (strongly) *consistent* estimator of θ if:

$$\mathbb{P}(\lim_{n\to\infty}\widehat{\theta}_n{=}\theta^*){=}1, \text{ and denoted by: } \widehat{\theta}_n \overset{a.s.}{\to} \theta^*.$$

As argued in chapter 9, *almost sure* (a.s.) implies *convergence* in probability. In chapter 9 it is shown that almost sure convergence is stronger than convergence in probability and not surprisingly, the former implies the latter.

Example 11.27. In light of the fact that the simple Bernoulli (table 11.1) and Normal (11.2) models assume that their respective underlying processes $\{X_k,\ k{\in}\mathbb{N}\}$ are IID and the estimators (i)-(vi) in (2) and (3) are functions of $\frac{1}{n}\sum_{i=1}^{n}X_i$, we can invoke directly the SLLN in chapter 9 to show which estimators are strongly consistent (table 11.7).

Table 11.7: Strong consistency of estimators

Bernoulli:	$\widehat{\theta}_1\overset{a.s.}{\nrightarrow}\theta,$	$\widehat{\theta}_2\overset{a.s.}{\nrightarrow}\theta,$	$\widehat{\theta}_3\overset{a.s.}{\nrightarrow}\theta,$	$\widehat{\theta}_n \overset{a.s.}{\to} \theta,$	$\widehat{\theta}_{n+1} \overset{a.s.}{\to} \theta,$	$\widehat{\theta}_{n+2} \overset{a.s.}{\to} \theta,$
Normal:	$\widehat{\mu}_1\overset{a.s.}{\nrightarrow}\mu,$	$\widehat{\mu}_2\overset{a.s.}{\nrightarrow}\mu,$	$\widehat{\mu}_3\overset{a.s.}{\nrightarrow}\mu,$	$\widehat{\mu}_n \overset{a.s.}{\to} \mu,$	$\widehat{\mu}_{n+1} \overset{a.s.}{\to} \mu,$	$\widehat{\mu}_{n+2} \overset{a.s.}{\to} \mu.$

5.3 Asymptotic Normality

The next asymptotic property, known as asymptotic Normality, is an extension of the *Central Limit Theorem* (CLT), discussed in chapter 9.

An estimator $\widehat{\theta}_n$ of θ is said to be *asymptotically Normal* if there exists a normalizing sequence $\{c_n\}_{n=1}^{\infty}$ such that:

$$c_n(\widehat{\theta}_n-\theta) \underset{n\to\infty}{\backsim} \mathsf{N}(0,V_\infty(\theta)), \text{ for } V_\infty(\theta){\neq}0.$$

Remarks: (i) " $\underset{n\to\infty}{\backsim}$ " reads asymptotically distributed. (ii) $V_\infty(\theta)$ denotes the asymptotic variance of $\widehat{\theta}_n$. (iii) The sequence $\{c_n\}_{n=1}^{\infty}$ is a function of n, and in the case of an IID sample $c_n{=}\sqrt{n}$.

Example 11.28. Given that the simple Bernoulli (table 11.1) and Normal (11.2) models assume that their respective underlying processes $\{X_k,\ k{\in}\mathbb{N}\}$ are IID and all the estimators (i)-(vi) in (2) and (3) are functions of $\frac{1}{n}\sum_{i=1}^{n}X_i$, we can invoke directly the CLT in chapter 9 to show that the *consistent* estimators in table 11.7 are also asymptotically Normal (table 11.8).

Table 11.8: Asymptotic Normality of estimators

Bernoulli:	$\sqrt{n}(\widehat{\theta}_n-\theta) \underset{n\rightarrow\infty}{\backsim} \mathsf{N}\left(0,\theta(1-\theta)\right),$	$\sqrt{n}(\widehat{\theta}_{n+k}-\theta) \underset{n\rightarrow\infty}{\backsim} \mathsf{N}\left(0,\theta(1-\theta)\right),\ k{=}1,2$
Normal:	$\sqrt{n}(\widehat{\mu}_n-\mu) \underset{n\rightarrow\infty}{\backsim} \mathsf{N}\left(0,\sigma^2\right),$	$\sqrt{n}(\widehat{\mu}_{n+k}-\mu) \underset{n\rightarrow\infty}{\backsim} \mathsf{N}\left(0,\sigma^2\right),\ k{=}1,2.$

The above example brings up a very important class of estimators known as *Consistent and Asymptotically Normal* (CAN) *estimators* of θ. The question that naturally arises is whether within the class of CAN estimators one can use the asymptotic variance $V_\infty(\theta)$ in order to choose between them. Let us discuss this question next.

5.4 Asymptotic efficiency

Is there a way to choose between CAN estimators by comparing their asymptotic variances $V_\infty(\theta)$? The answer is that there is an asymptotic Cramer-Rao ($\mathsf{C\text{-}R}_\infty(\theta)$) lower bound with respect to which $V_\infty(\theta)$ can be compared.

Assuming that the regularity conditions in table 11.4 are valid, the $\mathsf{C\text{-}R}_\infty(\theta)$ takes the form:

$$\mathsf{C\text{-}R}_\infty(\theta){=}[\mathcal{I}_\infty(\theta)]^{-1},\ \text{where } \mathcal{I}_\infty(\theta){=}\lim_{n\rightarrow\infty}\left((\tfrac{1}{c_n^2})\mathcal{I}_n(\theta)\right),$$

and $\mathcal{I}_\infty(\theta)$ is referred to as the *Asymptotic Fisher information.*

Example 11.29. In the context of the simple Bernoulli model in (table 11.1):

$$E\left(-\tfrac{d^2\ln f(\mathbf{x};\theta)}{d\theta^2}\right){=}\tfrac{n}{\theta(1-\theta)} \implies \mathcal{I}_\infty(\theta){=}\lim_{n\rightarrow\infty}\left((\tfrac{1}{n})\tfrac{n}{\theta(1-\theta)}\right){=}\tfrac{1}{\theta(1-\theta)},$$

and thus the asymptotic lower bound is: $\mathsf{C\text{-}R}_\infty(\theta){=}\theta(1{-}\theta)$.

Example 11.30. In the context of the simple Normal model in (table 11.2):

$$E\left(-\tfrac{d^2\ln f(\mathbf{x};\theta)}{d\theta^2}\right){=}\tfrac{n}{\sigma^2} \implies \mathcal{I}_\infty(\mu){=}\lim_{n\rightarrow\infty}\left((\tfrac{1}{n})\tfrac{n}{\sigma^2}\right){=}\tfrac{1}{\sigma^2} \implies CR_\infty(\theta){=}\sigma^2.$$

A CAN estimator $\widehat{\theta}_n$ of θ is said to be *asymptotically efficient* if:

$$c_n(\widehat{\theta}_n-\theta) \underset{n\rightarrow\infty}{\backsim} \mathsf{N}\left(0,[\mathcal{I}_\infty(\theta)]^{-1}\right),\ \text{when } \mathcal{I}_\infty(\theta){>}0.$$

That is, the asymptotic variance equals *the asymptotic Cramer-Rao lower bound*:

$$V_\infty(\theta){=}\mathsf{C\text{-}R}_\infty(\theta){=}[\mathcal{I}_\infty(\theta)]^{-1}$$

Example 11.31. For the simple Bernoulli model in (table 11.1), the results in table 11.8 suggest that all three CAN estimators $(\widehat{\theta}_n,\widehat{\theta}_{n+1},\widehat{\theta}_{n+1})$ are asymptotically efficient. Although $(\widehat{\theta}_n,\widehat{\theta}_{n+1},\widehat{\theta}_{n+1})$ are asymptotically equivalent, the fact that $\widehat{\theta}_n$ is fully efficient for any $n<\infty$, renders it the best among the three. Note that the same comments apply to the estimators $(\widehat{\mu}_n,\widehat{\mu}_{n+1},\widehat{\mu}_{n+1})$ in the context of the simple Normal model (table 11.2).

6 The simple Normal model: estimation

Table 11.9: The simple Normal model

Statistical GM:		$X_t{=}\mu + u_t,\ t{\in}\mathbb{N}{:=}(1,2,...,n,...)$
[1]	Normal:	$X_t \backsim \mathsf{N}(.,.),\ x_t{\in}\mathbb{R},$
[2]	Constant mean:	$E(X_t){=}\mu,\ \mu{\in}\mathbb{R},\ \forall t{\in}\mathbb{N},$
[3]	Constant variance:	$Var(X_t){=}\sigma^2,\ \forall t{\in}\mathbb{N},$
[4]	Independence:	$\{X_t,\ t{\in}\mathbb{N}\}$-independent process.

In the previous section we use two very simple examples in an attempt to keep the technical difficulties at a minimum and concentrate on the ideas and concepts. In this section we use a marginally more complicated model, that happens to be one of the most widely discussed model in statistics. This simple model provides the cornerstone of several widely used statistical models that we discuss in the sequel.

The sampling distribution of $\widehat{\mu}_n{=}\frac{1}{n}\sum_{k=1}^{n} X_k$.

Consider the *simple Normal model* in table 11.9.

As argued above, the best estimator of μ, in the case of a simple one-parameter Normal model is $\widehat{\mu}_n{=}\frac{1}{n}\sum_{k=1}^{n} X_k{:=}\overline{X}_n$. The obvious argument is that since $\mu{=}E(X_t)$ it makes intuitive sense that the sample mean $\widehat{\mu}$ should be a good estimator. We have shown above that the sampling distribution of $\widehat{\mu}_n$ is:

$$\widehat{\mu}_n \backsim \mathsf{N}\left(\mu, \tfrac{\sigma^2}{n}\right), \tag{24}$$

and shown that $\widehat{\mu}_n$ is unbiased, fully efficient and strongly consistent. The derivation of these properties, however, was based on assuming that σ^2 is known. We would like to know if any of these properties change when σ^2 is an unknown parameter to be estimated along side μ.

Although matching distribution and sample moments is not always a valid argument, for simplicity let us employ the same intuitive argument to estimate $\sigma^2{=}Var(X_t)$ using the *sample variance*:

$$\widehat{\sigma}_n^2{=}\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(X_i - \widehat{\mu}_n)^2. \tag{25}$$

Since unbiasedness and consistency are not affected by assuming that σ^2 is also unknown, the only thing that might change is the Cramer-Rao lower bound. To derive that we need the distribution of the sample:

$$\begin{aligned} f(\mathbf{x};\mu,\sigma^2) &={(\tfrac{1}{\sigma\sqrt{2\pi}})}^n \exp\left\{-\tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^{n}(x_i-\mu)^2\right\}, \\ \ln f(\mathbf{x};\mu,\sigma^2) &= -\tfrac{n}{2}\left[\ln(2\pi)\right] - \tfrac{n}{2}\ln(\sigma^2) - \tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^{n}(x_i-\mu)^2. \end{aligned}$$

Taking first and second derivatives of $\ln f(\mathbf{x};\mu,\sigma^2)$ yields:

$$\begin{array}{ll}
\frac{\partial \ln f(\mathbf{x};\mu,\sigma^2)}{\partial\mu}{=}\frac{1}{\sigma^2}\sum_{i=1}^n(x_i{-}\mu), & \frac{\partial \ln f(\mathbf{x};\mu,\sigma^2)}{\partial\sigma^2}{=}-\frac{n}{2\sigma^2}{+}\frac{1}{2(\sigma^2)^2}\sum_{i=1}^n(x_i{-}\mu)^2, \\
\frac{\partial^2 \ln f(\mathbf{x};\mu,\sigma^2)}{\partial\mu^2}{=}\frac{1}{\sigma^2}\sum_{i=1}^n(-1), & \frac{\partial^2 \ln f(\mathbf{x};\mu,\sigma^2)}{\partial(\sigma^2)^2}{=}\frac{n}{2\sigma^4}-\frac{1}{\sigma^6}\sum_{i=1}^n(x_i-\mu)^2, \\
\frac{\partial^2 \ln f(\mathbf{x};\mu,\sigma^2)}{\partial\sigma^2\partial\mu}{=}-\frac{1}{\sigma^4}\sum_{i=1}^n(x_i{-}\mu). &
\end{array}$$

Since $E(-\frac{\partial^2 \ln f(\mathbf{x};\mu,\sigma^2)}{\partial\sigma^2\partial\mu}){=}0$, the *Fisher information matrix* takes the form:

$$\mathcal{I}_n(\mu,\sigma^2){:=}\left(\begin{array}{cc} E\left(-\frac{\partial^2 \ln f(\mathbf{x};\mu,\sigma^2)}{\partial\mu^2}\right) & E\left(-\frac{\partial^2 \ln f(\mathbf{x};\mu,\sigma^2)}{\partial\sigma^2\partial\mu}\right) \\ E\left(-\frac{\partial^2 \ln f(\mathbf{x};\mu,\sigma^2)}{\partial\sigma^2\partial\mu}\right) & E\left(-\frac{\partial^2 \ln f(\mathbf{x};\mu,\sigma^2)}{\partial(\sigma^2)^2}\right)\end{array}\right)=\left(\begin{array}{cc}\frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4}\end{array}\right)$$

and thus the C-R lower bound for any unbiased estimators of (μ,σ^2) is:

$$\mathsf{C\text{-}R}(\mu,\sigma^2){=}\left[\mathcal{I}_n(\mu,\sigma^2)\right]^{-1} \implies \mathsf{C\text{-}R}(\mu){=}\frac{\sigma^2}{n} \text{ and } \mathsf{C\text{-}R}(\sigma^2){=}\frac{2\sigma^4}{n}. \tag{26}$$

This shows that $\widehat{\mu}_n$ achieves its C-R bound, and retains its asymptotic properties of consistency and asymptotic Normality.

The question is whether the estimator $\widehat{\sigma}_n^2$ has similar properties, and for that we need to derive its sampling distribution.

The sampling distribution of $\widehat{\sigma}_n^2{=}\frac{1}{n}\sum_{k=1}^n(X_k{-}\widehat{\mu}_n)^2$

Since $\widehat{\sigma}_n^2$ is a quadratic function of Normally distributed random variables, we will use the following lemma to derive its sampling distribution.

Lemma 11.3. Assume that the random variables $(Z_1,Z_2,...,Z_n)$ are distributed as:

$$\mathsf{NIID}(0,1),\ k{=}1,2,...,n,$$

then the function $V_n{=}\sum_{k=1}^n Z_k^2$ is *Chi-square* distributed with n degrees of freedom:

$$V_n{=}\sum_{k=1}^n Z_k^2 \backsim \chi^2(n).$$

Since this lemma pertains to $\mathsf{N}(0,1)$ random variables, we need to standardize and then use it to claim that:

$$Z_k{=}(\tfrac{X_k-\mu}{\sigma}){\backsim}\mathsf{NIID}(0,1) \implies \sum_{k=1}^n Z_k^2{=}\sum_{k=1}^n(\tfrac{X_k-\mu}{\sigma})^2 \backsim \chi^2(n).$$

Given that $\widehat{\sigma}_n^2{=}\frac{1}{n}\sum_{i=1}^n(X_i-\widehat{\mu}_n)^2$ differs from $\sum_{k=1}^n(\frac{X_k-\mu}{\sigma})^2$, we need an identify to relate the two in the form of (Spanos, 1986, p. 240):

$$\sum_{k=1}^n\left(\tfrac{X_k-\mu}{\sigma}\right)^2{=}\sum_{k=1}^n\left(\tfrac{X_k-\widehat{\mu}_n}{\sigma}\right)^2+n\left(\tfrac{\widehat{\mu}_n-\mu}{\sigma}\right)^2, \tag{27}$$

Standardizing $\widehat{\mu}_n$ in (24) can deduce that:

$$\Big(\frac{\widehat{\mu}_n-\mu}{\sqrt{\frac{\sigma^2}{n}}}\Big)=\Big(\frac{\sqrt{n}(\widehat{\mu}_n-\mu)}{\sigma}\Big)\backsim\mathsf{N}\,(0,1)\overset{\text{lemma 11.3}}{\Longrightarrow} n\Big(\frac{\widehat{\mu}_n-\mu}{\sigma}\Big)^2\backsim\chi^2(1).$$

In addition, it can show that $V_1=n(\frac{\widehat{\mu}_n-\mu}{\sigma})^2$ and $V_2=\sum_{k=1}^n(\frac{X_k-\widehat{\mu}_n}{\sigma})^2$ are independent (Casella and Berger, 2002), we need to relate the left and the right hand side of (27). For that we need the following lemma.

Lemma 11.4. Let $V_1\backsim\chi^2(m_1)$, $V_2\backsim\chi^2(m_2)$ and V_1 and V_1 be independent, then the sum $V=V_1+V_2$ is chi-square distributed with $m=m_1+m_2$ degrees of freedom:

$$(V_1+V_2)\backsim\chi^2(m_1+m_2).$$

In light of this lemma, since $\sum_{k=1}^n(\frac{X_k-\mu}{\sigma})^2$ is distributed as $\chi^2(n)$ and the right hand side is composed of two independent random variables and one has a $\chi^2(1)$ distributed, it follows from lemma 11.4 that:

$$\Big(\frac{n\widehat{\sigma}_n^2}{\sigma^2}\Big)=\sum_{k=1}^n\left(\frac{X_k-\widehat{\mu}}{\sigma}\right)^2\backsim\chi^2(n-1). \tag{28}$$

Using the fact that for $V\backsim\chi^2(m)$, $E(V)=m$, we deduce:

$$E\Big(\frac{n\widehat{\sigma}_n^2}{\sigma^2}\Big)=(n-1)\Rightarrow\quad E(\widehat{\sigma}_n^2)=\frac{(n-1)}{n}\sigma^2\neq\sigma^2, \tag{29}$$

and thus $\widehat{\sigma}_n^2$ is a biased estimator of σ^2.

The sampling distribution of $s_n^2=\frac{1}{n-1}\sum_{i=1}^n(X_i-\widehat{\mu}_n)^2$

The result in (29) also suggests that:

$$s_n^2=\Big(\frac{n}{n-1}\Big)\widehat{\sigma}_n^2=\frac{1}{n-1}\sum_{k=1}^n(X_k-\widehat{\mu}_n)^2,\ E(s_n^2)=\sigma^2,$$

that is, a rescaled form of $\widehat{\sigma}_n^2$, $s_n^2=\frac{1}{n-1}\sum_{i=1}^n(X_i-\widehat{\mu}_n)^2$ is an unbiased estimator of σ^2 and:

$$\frac{(n-1)s_n^2}{\sigma^2}=\sum_{k=1}^n\frac{(X_k-\widehat{\mu}_n)^2}{\sigma^2}\backsim\chi^2(n-1), \tag{30}$$

The question which arises is whether s_n^2, in addition to being unbiased, has any further advantages over $\widehat{\sigma}_n^2$. To derive the variance of s_n^2 we use the result that for $V\backsim\chi^2(m)$, $Var(V)=2m$ (see appendix 3.A), to deduce that:

$$Var\Big(\frac{(n-1)s_n^2}{\sigma^2}\Big)=2(n-1)\implies Var(s_n^2)=\frac{2\sigma^4}{n-1}>\mathsf{C\text{-}R}(\sigma^2)=\frac{2\sigma^4}{n}.$$

Hence, s_n^2 does not achieve the Cramer-Rao lower bound. Similarly, using (28) we can deduce that $\widehat{\sigma}_n^2$ does not achieve the lower bound given by $\mathsf{GC\text{-}R}(\sigma^2)$ since:

$$Var\Big(\frac{n\widehat{\sigma}_n^2}{\sigma^2}\Big)=2(n-1)\implies Var(\widehat{\sigma}_n^2)=\frac{2(n-1)\sigma^2}{n^2}>\mathsf{GC\text{-}R}(\sigma^2)=\left(\frac{2(n-1)^2}{n^3}\right)\sigma^4,$$

$$\mathsf{GC\text{-}R}(\sigma^2){=}\left(\tfrac{dE(\widehat{\sigma}_n^2)}{d\sigma^2}\right)^2\left\{E\left(\tfrac{d\ln f(\mathbf{x};\mu,\sigma^2)}{d\sigma^2}\right)^2\right\}^{-1}=\left(\tfrac{n-1}{n}\right)^2\left(\tfrac{2\sigma^4}{n}\right)=\left(\tfrac{2(n-1)^2}{n^3}\right)\sigma^4.$$

Sufficiency. In the case of the simple Normal model (table 11.9), there exists a minimal sufficient statistic
$\mathbf{S}(\mathbf{X}){=}\left(\sum_{k=1}^n X_k,\ \sum_{k=1}^n X_k^2\right)$, and all three estimators $\left(\widehat{\mu}_n, s_n^2, \widehat{\sigma}_n^2\right)$ are minimal sufficient since they are all functions of $\mathbf{S}(\mathbf{X})$.

Asymptotic properties. In terms of their asymptotic properties both estimators $\widehat{\sigma}_n^2$ and s_n^2 of σ^2 enjoy all the optimal asymptotic properties: consistency, asymptotic Normality and asymptotic efficiency since:

$$\sqrt{n}\left(\widehat{\sigma}_n^2{-}\sigma^2\right)\underset{n\to\infty}{\backsim}\mathsf{N}\left(0,2\sigma^4\right),\quad \sqrt{n}\left(s_n^2{-}\sigma^2\right)\underset{n\to\infty}{\backsim}\mathsf{N}\left(0,2\sigma^4\right),$$

in view of the fact that the asymptotic Fisher's information matrix is:

$$\mathcal{I}_\infty(\mu,\sigma^2)=\lim_{n\to\infty}\left(\tfrac{1}{n}\mathcal{I}_n(\mu,\sigma^2)\right)=\begin{pmatrix}\frac{1}{\sigma^2} & 0\\ 0 & \frac{1}{2\sigma^4}\end{pmatrix}.$$

The sampling distribution of $\left(\frac{\sqrt{n}(\widehat{\mu}_n-\mu)}{s_n}\right)$

For the simple Normal model (table 11.9), the two best estimators of (μ,σ^2) in the form of $\widehat{\mu}_n$ and s_n^2 come together when one is interested in the sampling distribution of the ratio $\frac{\sqrt{n}(\widehat{\mu}_n-\mu)}{s_n}$, $\widehat{\mu}_n{=}\overline{X}_n$. This is the basis of the well known Student's t statistic that began modern statistics at Guinness brewery in Dublin, Ireland. An employee of this brewery by the name William Gosset, published a paper in 1908, under the pseudonym "Student", that inspired Fisher to focus attention on finite sampling distributions as a basis for frequentist inference.

At that time it was known that in the case where $\mathbf{X}{:=}(X_1,X_2,...,X_n)$ is a random sample (IID) from a distribution with unknown mean (μ) and variance (σ^2), the estimator $\overline{X}_n{=}\frac{1}{n}\sum_{k=1}^n X_k$ is asymptotically $(n\to\infty)$ Normally distributed:

$$\overline{X}_n\underset{n\to\infty}{\backsim}\mathsf{N}\left(\mu,\tfrac{\sigma^2}{n}\right)\Rightarrow\tfrac{\sqrt{n}(\overline{X}_n-\mu)}{\sigma}\underset{n\to\infty}{\backsim}\mathsf{N}\left(0,1\right). \tag{31}$$

It was also known that when σ^2 is replaced by $\widehat{\sigma}_n^2{=}\frac{1}{n}\sum_{i=1}^n(X_i{-}\overline{X}_n)^2$, the asymptotic distribution in (31) remains the same:

$$\tfrac{\sqrt{n}(\overline{X}_n-\mu)}{\widehat{\sigma}_n}\underset{n\to\infty}{\backsim}\mathsf{N}\left(0,1\right).$$

Motivated by his experimental work at the Guinness Brewery, Gosset was interested in drawing inferences with a small $n<15$, and the asymptotic approximation in (31) was not accurate enough for that: "in such cases [small n] it is sometimes necessary to judge of the certainty of the results from a very small sample, which itself affords the only indication of the variability." (Student, 1908, p. 2) Gosset realized that when $\mathbf{X}$ is a

NIID sample, the result in (31) is exact and not asymptotic, i.e. it holds for any $n > 1$:

$$\overline{X}_n \backsim \mathsf{N}\left(\mu, \tfrac{\sigma^2}{n}\right) \Rightarrow \tfrac{\sqrt{n}(\overline{X}_n-\mu)}{\sigma} \backsim \mathsf{N}(0,1). \tag{32}$$

This was the first finite sampling distribution. His asked the question about the finite sampling distribution of:

$$\tfrac{\sqrt{n}(\overline{X}_n-\mu)}{s_n} \overset{?}{\backsim} \mathsf{D}(0,1),$$

and used simulated data in conjunction with Karl Pearson's approach to statistics (chapter 10) to conclude (guess) that for any $n{>}1$:

$$\tfrac{\sqrt{n}(\overline{X}_n-\mu)}{s_n} \backsim \mathsf{St}(n{-}1), \tag{33}$$

where $\mathsf{St}(n{-}1)$ denotes a Student's t distribution with $(n{-}1)$ degrees of freedom. This result was formally derived by Fisher (1915) whose derivation of (33) is encapsulated by the following lemma.

Lemma 11.5. Assuming that Z and V are two independent random variables with distributions:

$$(Z \backsim \mathsf{N}(0,1),\ V \backsim \chi^2(m)) \implies \tfrac{Z}{\sqrt{\left(\frac{V}{m}\right)}} \backsim \mathsf{St}(m).$$

That is, the ratio $(\frac{Z}{(V/m)})$ is *Student's t* distributed with m degrees of freedom.

Example 11.34. In the context of the simple Normal model in (table 11.9), the finite sampling distributions of $\overline{X}_n{=}\frac{1}{n}\sum_{k=1}^{n} X_k$ and $s_n^2{=}\frac{1}{n-1}\sum_{i=1}^{n}(X_i{-}\overline{X}_n)^2$ in (32) and (30) are:

$$Z{=}\tfrac{\sqrt{n}(\overline{X}_n-\mu)}{\sigma} \backsim \mathsf{N}(0,1),\ V{=}\tfrac{(n-1)s_n^2}{\sigma^2} \backsim \chi^2(n{-}1).$$

In addition, $\overline{X}_n$ and s_n^2 are independent. Using lemma 11.5 we can deduce that:

$$\left(\frac{\frac{\sqrt{n}(\overline{X}_n-\mu)}{\sigma}}{\sqrt{\frac{(n-1)s_n^2}{(n-1)\sigma^2}}}\right){=}\tfrac{\sqrt{n}(\overline{X}_n-\mu)}{s_n} \backsim \mathsf{St}(n{-}1).$$

After deriving the finite sampling distributions of the Student's t ratio and that of the sample correlation coefficient in 1915, Fisher went on to derive almost all the finite sampling distributions to the simple Normal and Linear Regression models.

7 Confidence Intervals (interval estimation)

It is often insufficiently appreciated that although point estimation provides the basis for all others forms of statistical inference, interval estimation, hypothesis testing, prediction and simulation, it does not, by itself, output an inferential claim as often erroneously presumed. What is the inferential claim associated with an 'optimal' point estimator $\widehat{\theta}_n(\mathbf{X})$ of θ in Θ, in the context of a prespecified statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$? $\widehat{\theta}_n(\mathbf{X})$ does *not* justify the obvious inferential claim:

$$\widehat{\theta}_n(\mathbf{x}_0) \text{ is 'close in some sense' to } \theta^* \text{ (the true value of } \theta\text{)},$$

since $\widehat{\theta}_n(\mathbf{x}_0)$ represents only a single value in the infinite parameter space Θ. The optimal properties calibrate the capacity of the estimator $\widehat{\theta}_n(\mathbf{X})$ to pinpoint θ^*, but the relevant probabilities do not extend to the estimate $\widehat{\theta}_n(\mathbf{x}_0)$. This is the reason why the reported point estimates are usually accompanied by the estimated Standard Error (SE) of $\widehat{\theta}_n(\mathbf{X})$ to convey, in some broad sense, how close that point estimate is likely to be to θ^*. In the case of a simple Bernoulli model (table 11.1), a point estimate $\widehat{\theta}_n(\mathbf{x}_0){=}.51$ with $SE(\widehat{\theta}_n){=}\sqrt{Var(\widehat{\theta}_n(\mathbf{X}))}{=}.01$, in some intuitive sense appears to be closer to θ^* than if $SE(\widehat{\theta}_n){=}\sqrt{Var(\widehat{\theta}_n(\mathbf{X}))}{=}.1$. An attempt to formalize this intuition gave rise to confidence interval estimation. A two-sided $(1{-}\alpha)$ *Confidence Interval* (CI) for θ takes the form:

$$\mathbb{P}\left(\widehat{\theta}_L(\mathbf{X}) \leq \theta \leq \widehat{\theta}_U(\mathbf{X});\ \theta{=}\theta^*\right) = 1{-}\alpha, \tag{34}$$

where $\widehat{\theta}_L(\mathbf{X})$ and $\widehat{\theta}_U(\mathbf{X})$ denote the lower and upper bound statistics that define the different CIs for θ. The notation $\theta{=}\theta^*$ indicates that the evaluation of the probability $(1{-}\alpha)$ is under $\theta{=}\theta^*$; the true value of θ, whatever that happens to be. The $(1{-}\alpha)$ refers to the probability that the random bound(s) $\left(\widehat{\theta}_L(\mathbf{X}),\ \widehat{\theta}_U(\mathbf{X})\right)$ 'cover' (overlay) the true value θ^*. Equivalently, the 'coverage error' probability α denotes the probability that the random interval based on the bounds $[\widehat{\theta}_L(\mathbf{X}),\ \widehat{\theta}_U(\mathbf{X})]$ do *not* cover θ^*. NOTE for *discrete* probabilistic models $f(\mathbf{x};\theta)$ one uses $\geq 1{-}\alpha$ because the equality might not be available for a particular α.

A *one-sided* $(1{-}\alpha)$ CI for θ takes the form:

$$\mathbb{P}\left(\theta \geq \widehat{\theta}_{L_1}(\mathbf{X});\ \theta{=}\theta^*\right) = 1{-}\alpha,\quad \mathbb{P}\left(\theta \leq \widehat{\theta}_{U_1}(\mathbf{X});\ \theta{=}\theta^*\right) = 1{-}\alpha, \tag{35}$$

where $\widehat{\theta}_{L_1}(\mathbf{X})$ and $\widehat{\theta}_{U_1}(\mathbf{X})$ denote the relevant lower and upper statistics.

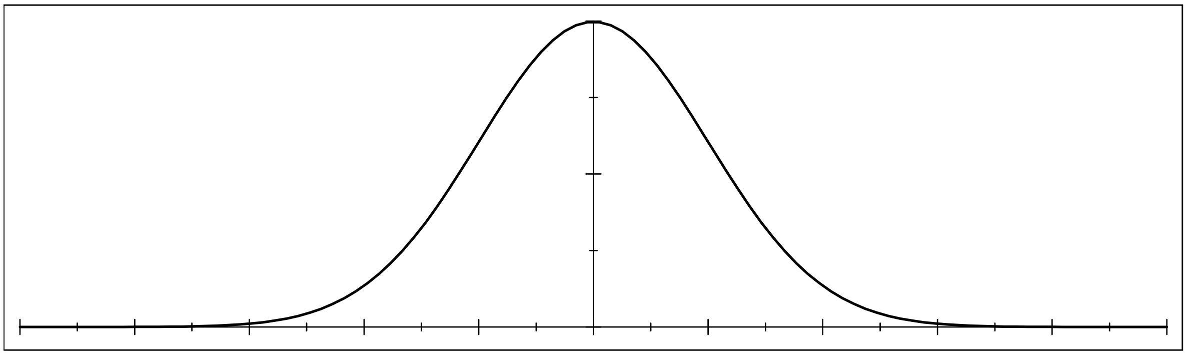
7.1 Long-run 'interpretation' of CIs

An intuitive way to understand the coverage probability is to invoke the metaphor of repeatability (in principle), and draw a sequence of different sample $(\mathbf{X})$ realizations of size n, say $\mathbf{x}_1,\ \mathbf{x}_2, ...,\ \mathbf{x}_N$, and evaluate the *observed CIs*: $[\widehat{\theta}_L(\mathbf{x}_i),\ \widehat{\theta}_U(\mathbf{x}_i)],\ i{=}1, ..., N$. It is crucial to note that no probability can be attached to an *observed CI*:

$$\left(\widehat{\theta}_L(\mathbf{x}_0) \leq \theta \leq \widehat{\theta}_U(\mathbf{x}_0)\right),$$

because the latter is *not* stochastic and it either includes or excludes θ^*. The coverage probability $(1{-}\alpha)$ will ensure that a proportion of at least $(1{-}\alpha)\,\%$ of these observed CIs cover (overlay) the true θ, whatever its value θ^* is.

Example. The .95 Confidence Interval (CI) in (**??**) can be visualized using the long-run metaphor with N=26 in the graph below. The relative frequency of coverage is $\frac{24}{26}$=.923; 24 out of 26 observed CIs overlay the true μ^*. For a better approximation of the coverage probability .95 one would need $N > 100$.



μ^*

1. $\vdash - - - - \overline{x}_n - - - - \dashv$
2. $\vdash - - - - - -\overline{x}_n - - - - - - \dashv$
3. $\vdash - - - - - \overline{x}_n - - - - - \dashv$
4. $\vdash - - - - - \overline{x}_n - - - - - \dashv$
5. $\vdash - - - - - \overline{x}_n - - - - - \dashv$
6. $\vdash - - - - - \overline{x}_n - - - - - \dashv$
7. $\vdash - - - - - \overline{x}_n - - - - - \dashv$
8. $\vdash - - - - - \overline{x}_n - - - - - \dashv$
9. $\vdash - - - - - \overline{x}_n - - - - - \dashv$
10. $\vdash - - - - - \overline{x}_n - - - - - \dashv$
11. $\vdash - - - - - \overline{x}_n - - - - - \dashv$
12. $\vdash - - - - - \overline{x}_n - - - - - \dashv$
13. $\blacktriangleright$ $\vdash - - - -\overline{x}_n- - - - \dashv$
14. $\vdash - - - - - \overline{x}_n - - - - - \dashv$
15. $\vdash - - - - - \overline{x}_n - - - - - \dashv$
16. $\vdash - - - - - \overline{x}_n - - - - - \dashv$
17. $\vdash - - - - - -\overline{x}_n - - - - - - \dashv$
18. $\vdash - - - - - \overline{x}_n - - - - - \dashv$
19. $\vdash - - - - - \overline{x}_n - - - - - \dashv$
20. $\vdash - - - - - \overline{x}_n - - - - - \dashv$
21. $\vdash - - - - - \overline{x}_n - - - - - \dashv$ $\blacktriangleleft$
22. $\vdash - - - - - \overline{x}_n - - - - - \dashv$
23. $\vdash - - - - - \overline{x}_n - - - - - \dashv$
24. $\vdash - - - - - \overline{x}_n - - - - - \dashv$
25. $\vdash - - - - - \overline{x}_n - - - - - \dashv$
26. $\vdash - - - - - \overline{x}_n - - - - - \dashv$

7.2 Constructing a Confidence Interval (CI)

▶ **Where do CIs come from?** The simplest way to construct a confidence interval is to use what Fisher (1935b) called a *pivot* (or a pivotal quantity) $q(\mathbf{X},\theta)$ when it exists. A pivot is defined to be a function of both the sample $\mathbf{X}$ and the unknown parameter θ:

$$q(.,.)\text{: } (\mathbb{R}^n_X\times\Theta)\rightarrow\mathbb{R},$$

whose distribution $f(q(\mathbf{x},\theta))$ is *known*; it does not involve unknown parameters.

Example 11.35. For the simple (one-parameter) Normal model (table 11.2). In this case, a *pivotal quantity* exists and takes the form:

$$q(\mathbf{X},\mu){=}\sqrt{n}(\tfrac{\overline{X}_n-\mu}{\sigma})\backsim\mathsf{N}(0,1)$$

To bring out the fact that this result is an example of factual reasoning, it should be written more correctly as:

$$q(\mathbf{X},\mu){=}\sqrt{n}(\tfrac{\overline{X}_n-\mu^*}{\sigma})\overset{\mu=\mu^*}{\backsim}\mathsf{N}(0,1),\tag{36}$$

where the evaluation of the pivot is under $\mu{=}\mu^*$, whatever μ^* happens to be. Indeed, if one were to choose a particular value for μ, say $\mu{=}\mu_0{\neq}\mu^*$, then the above result would *not* hold since:

$$q(\mathbf{X},\mu){=}\sqrt{n}(\tfrac{\overline{X}_n-\mu_0}{\sigma})\overset{\mu=\mu^*}{\backsim}\mathsf{N}(\delta_0,1),\;\;\delta_0{=}\sqrt{n}(\tfrac{\mu^*-\mu_0}{\sigma}).$$

One can use the pivot in (36) to construct a two-sided Confidence Interval (CI) in two steps.

Step 1. Select α and use the distribution $\mathsf{N}(0,1)$ in (36) to define the constants a and b such that:

$$\mathbb{P}\left(a\leq\sqrt{n}(\tfrac{\overline{X}_n-\mu}{\sigma})\leq b;\;\;\mu{=}\mu^*\right){=}1-\alpha.$$

In this case, for $\alpha{=}.05$, $-a{=}b{=}c_{\frac{\alpha}{2}}{=}1.96$.

Step 2. 'Solve' the pivot to isolate μ between the two inequalities, to derive the two-sided CI:

$$\mathbb{P}\left(\overline{X}_n{-}\tfrac{\sigma}{\sqrt{n}}c_{\frac{\alpha}{2}}\leq\mu\leq\overline{X}_n{+}\tfrac{\sigma}{\sqrt{n}}c_{\frac{\alpha}{2}};\;\;\mu{=}\mu^*\right){=}1{-}\alpha,$$

$$\mathbb{P}\left(\overline{X}_n-\tfrac{1.96}{\sqrt{n}}\leq\mu\leq\overline{X}_n+\tfrac{1.96}{\sqrt{n}};\;\;\mu{=}\mu^*\right){=}.95,$$

or equivalently in terms of the *coverage error*:

$$\mathbb{P}\left(\overline{X}_n-\tfrac{\sigma}{\sqrt{n}}c_{\frac{\alpha}{2}}\leq\mu\leq\overline{X}_n+\tfrac{\sigma}{\sqrt{n}}c_{\frac{\alpha}{2}};\;\;\mu{\neq}\mu^*\right){=}\alpha,$$

$$\mathbb{P}\left(\overline{X}_n-\tfrac{1.96\sigma}{\sqrt{n}}\leq\mu\leq\overline{X}_n+\tfrac{1.96\sigma}{\sqrt{n}};\;\;\mu{\neq}\mu^*\right){=}.05.$$

By the same token, the probability $(1{-}\alpha)$ cannot be assigned to the observed CI:

$$(\overline{x}_n-\tfrac{1}{\sqrt{n}}c_{\frac{\alpha}{2}}\leq\mu\leq\overline{x}_n+\tfrac{1}{\sqrt{n}}c_{\frac{\alpha}{2}}),$$

where $\overline{x}_n$ denotes the observed value of $\overline{X}_n$.

7.3 Optimality of Confidence Intervals

How does one assess the optimality (effectiveness) of Confidence Intervals (CIs)?

Intuitively, for a given confidence level $(1-\alpha)$, the shorter the CI the more informative it is for learning about the true value θ^* of the unknown parameter θ. An ideal CI will be one that reduces the interval to a single point $\theta{=}\theta^*$, but as in the case of an ideal estimator, no such interval exists for a finite sample size n.

▶ **But how can one measure the length of a CI?**

In the case of the $(1-\alpha)$ CI for μ:

$$\mathbb{P}\left(\overline{X}_n - \tfrac{1}{\sqrt{n}}c_{\frac{\alpha}{2}} \leq \mu \leq \overline{X}_n + \tfrac{1}{\sqrt{n}}c_{\frac{\alpha}{2}};\ \mu{=}\mu^*\right){=}1{-}\alpha,$$

one can measure its *length* by subtracting the lower from the upper bound:

$$\widehat{\theta}_U(\mathbf{X}) - \widehat{\theta}_L(\mathbf{X}){=}\left(\overline{X}_n{+}\tfrac{1}{\sqrt{n}}c_{\frac{\alpha}{2}}\right) - \left(\overline{X}_n{-}\tfrac{1}{\sqrt{n}}c_{\frac{\alpha}{2}}\right){=}\tfrac{2}{\sqrt{n}}c_{\frac{\alpha}{2}}.$$

It turns out that this CI is the shortest possible because of the choice of the quantiles of the Normal distribution to be: $-a{=}b{=}c_{\frac{\alpha}{2}}$. Any choice that does not satisfy this equality gives rise to CIs of bigger length.

In general, evaluating the length of a CI is not as easy because the difference $\widehat{\theta}_U(\mathbf{X}) - \widehat{\theta}_L(\mathbf{X})$ often gives rise to a statistic, which is a random variable, and one has to evaluate the *expected length*: $E\left[\widehat{\theta}_U(\mathbf{X}) - \widehat{\theta}_L(\mathbf{X})\right]$.

It turns out, however, that there is a *duality* between optimal tests and optimal CIs, that renders the optimality of both procedures easier to understand because for every optimal test there is an analogous optimal CI; see chapter 13.

8 Bayesian estimation

A key argument used by Bayesians to taut their in favorite approach to statistics is its simplicity in the sense that all forms of inference revolve around a single function, the posterior distribution: $\pi(\boldsymbol{\theta}|\mathbf{x}_0)\propto\pi(\boldsymbol{\theta})\cdot f(\mathbf{x}_0|\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}\in\Theta$. This, however, is only half the story. The other half is how the posterior distribution is utilized to yield 'optimal' inferences. The issue of optimality, however, is intrinsically related to what the primary objective of Bayesian inference is.

An outsider looking at Bayesian approach would surmise that its primary objective is to yield 'the probabilistic ranking' (ordering) of all values of $\boldsymbol{\theta}$ in Θ. The modeling begins with an a priori probabilistic ranking based on $\pi(\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}\in\Theta$, which is revised after observing $\mathbf{x}_0$ to derive $\pi(\boldsymbol{\theta}|\mathbf{x}_0),\ \forall\boldsymbol{\theta}\in\Theta$; hence the key role of the quantifier $\forall\boldsymbol{\theta}\in\Theta$. Indeed, O'Hagan's (1994) argues that the revised probabilistic ranking *is* the inference: "The most usual inference question is this: After seeing the data x_0, what do we now know about the parameter θ? The only answer to this question is to present the entire posterior distribution." (p. 6). He goes on to argue: "Classical inference theory is very concerned with constructing good inference rules. The primary concern of Bayesian inference, ..., is entirely different. The objective is to extract information concerning θ from the posterior distribution, and to present it helpfully via effective summaries." (p. 14). Where do these effective summaries come from? O'Hagan argues that the criteria for 'optimal' Bayesian inferences are only *parasitical* on the Bayesian approach and enter the picture via the decision theoretic perspective: "... a study of decision theory ... helps identify suitable summaries to give Bayesian answers to stylized inference questions which classical theory addresses." (p. 14).

Decision-theoretic framing of inference; initially proposed by Wald (1939; 1950). The decision-theoretic set-up has three basic components.

(i) A prespecified statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

(ii) A decision space D containing all mappings $d(.)$: $\mathbb{R}^n_X \rightarrow A$, where A denotes the set of all actions available to the statistician.

(iii) A loss function $L(.,.)$: $[D \times \Theta] \rightarrow R$, representing the numerical loss if the statistician takes action $a{\in}A$ when the state of nature is $\theta{\in}\Theta$; see Ferguson (1967), Berger (1985), Wasserman (2004).

The basic idea is that, when the decision-maker selects action a, he/she does not know the 'true' state of nature, represented by $\boldsymbol{\theta}^*$. However, contingent on each action $a{\in}A$, the decision-maker 'knows' the losses (gains, utilities) resulting from different choices $(d,\boldsymbol{\theta})\in[D \times \Theta]$. The decision-maker observes data $\mathbf{x}_0$, which provides some information about $\boldsymbol{\theta}^*$, and then maps each $\mathbf{x}{\in}\mathbb{R}^n_X$ to a certain action $a{\in}A$ guided solely by $L(d,\boldsymbol{\theta})$. This intuitive argument needs to be tempered somewhat since $\boldsymbol{\theta}^*$ is unknown, and thus the loss function will penalize $\boldsymbol{\theta}^*$ as every other value of $\boldsymbol{\theta}$ in Θ.

8.1 Optimal Bayesian rules

The decision theoretic setup provides optimal Bayesian rules the risk function $R(\theta,\widehat{\theta}){=}E_{\mathbf{X}}\left[L(\theta,\widehat{\theta}(\mathbf{X}))\right]$ to define the Bayes risk:

$$\textbf{Bayes risk}:\quad R_B(\widehat{\theta}){=}\int_{\boldsymbol{\theta}\in\Theta} R(\theta,\widehat{\theta})\pi(\theta)d\theta,$$

whose minimization with respect to all such rules $\widetilde{\theta}(\mathbf{x})$ yields:

$$\textbf{Bayes rule}:\quad \inf_{\widetilde{\theta}(\mathbf{x})} R_B(\widehat{\theta}){=}\inf_{\widetilde{\theta}(\mathbf{x})}\int_{\theta\in\Theta} R(\theta,\widehat{\theta})\pi(\theta)d\theta.$$

In light of the fact that $R_B(\widehat{\theta})$ can be expressed in the form (Bansal, 2007):

$$R_B(\widehat{\theta})\ \ ={\int_{\mathbf{x}\in\mathbb{R}^n_X}}\int_{\theta\in\Theta} L(\widehat{\theta}(\mathbf{X}),\theta)\pi(\boldsymbol{\theta}|\mathbf{x})d\theta d\mathbf{x}. \tag{37}$$

where $\pi(\boldsymbol{\theta}|\mathbf{x}){\propto}f(\mathbf{x}|\theta)\pi(\theta)$. In light of (37), a Bayesian rule is 'optimal' relative to a particular loss function $L(\widehat{\theta}(\mathbf{X}),\theta)$, when it minimizes $R_B(\widehat{\theta})$. This makes it clear that what constitutes an 'optimal' Bayesian rule is primarily determined by $L(\widehat{\theta}(\mathbf{X}),\theta)$ (Schervish, 1995):

(i) when $L_2(\widehat{\theta},\theta){=}(\widehat{\theta}-\theta)^2$ the Bayes estimate $\widehat{\theta}$ is the *mean* of $\pi(\theta|\mathbf{x}_0)$,

(ii) when $L_1(\widetilde{\theta},\theta){=}|\widetilde{\theta}-\theta|$ the Bayes estimate $\widetilde{\theta}$ is the *median* of $\pi(\theta|\mathbf{x}_0)$,

(iii) when $L_{0-1}(\overline{\theta},\theta){=}\delta(\overline{\theta},\theta){=}\begin{cases}0 & \text{for } \left|\overline{\theta}{-}\theta\right| < \varepsilon\\ 1 & \text{for } \left|\overline{\theta}{-}\theta\right| \geq \varepsilon\end{cases}$ for $\varepsilon{>}0$, the Bayes estimate $\overline{\theta}$ is the *mode* of $\pi(\theta|\mathbf{x}_0)$.

In practice, the most widely used loss function is the square:

$$L_2(\widehat{\theta}(\mathbf{X});\theta){=}(\widehat{\theta}(\mathbf{X})-\theta)^2,\ \forall\theta{\in}\Theta,$$

whose risk function is the decision-theoretic *Mean Square Error (MSE)*:

$$R(\theta,\widehat{\theta}){=}E(\widehat{\theta}(\mathbf{X}){-}\theta)^2{=}MSE(\widehat{\theta}(\mathbf{X});\theta),\ \forall\theta{\in}\Theta. \tag{38}$$

It is important to note that (38) is the source of confusion between that and the frequentist definition in (17).

Example 11.36. As shown in example 10.10, for the *simple Bernoulli model* (table 11.1), with $\pi(\theta){\backsim}\mathsf{Beta}(\alpha,\beta)$, the posterior distribution is $\pi(\theta|\mathbf{x}_0){\backsim}(\alpha^*,\beta^*)$, where: $\alpha^*{=}n\overline{x}{+}\alpha$, $\beta^*{=}n(1{-}\overline{x}){+}\beta$. Note that for $Z\backsim\mathsf{Beta}(\alpha,\beta)$, $\text{mode}(Z){=}\frac{\alpha-1}{\alpha+\beta-2}$, $E(Z){=}\frac{\alpha}{\alpha+\beta}$ and $Var(Z){=}\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ (Appendix 3.A).

(a) When the relevant loss function is $L_{0-1}(\overline{\theta},\theta)$, the optimal Bayesian rule is the *mode* of $\pi(\theta|\mathbf{x}_0){\backsim}(\alpha^*,\beta^*)$, which takes the form:

$$\widetilde{\theta}_B{=}\tfrac{\alpha^*-1}{\alpha^*+\beta^*-2}{=}\tfrac{(n\overline{X}+\alpha-1)}{(n+\alpha+\beta-2)}. \tag{39}$$

(b) When the relevant loss function is $L_2(\widehat{\theta},\theta){=}(\widehat{\theta}-\theta)^2$, the optimal Bayesian rule is the *mode* of $\pi(\theta|\mathbf{x}_0){\backsim}\mathsf{Beta}(\alpha^*,\beta^*)$, which takes the form:

$$\widehat{\theta}_B{=}\tfrac{\alpha^*}{\alpha^*+\beta^*}{=}\tfrac{(n\overline{X}+\alpha)}{(n+\alpha+\beta)}. \tag{40}$$

For a Jeffreys prior $\pi(\theta){\backsim}\mathsf{Beta}(.5,.5)$, $n\overline{x}{=}4$, $n{=}20$, $\alpha^*{=}n\overline{x}{+}\alpha{=}4.5$, $\beta^*{=}n(1\text{-}\overline{x}){+}\beta{=}16.5$:

$$\widetilde{\theta}_B{=}\tfrac{3.5}{21-2}{=}.184,\quad \widehat{\theta}_B{=}\tfrac{4.5}{4.5+16.5}{=}.214. \tag{41}$$

Lehmann (1984) warned statisticians about the perils of arbitrary loss functions: "It is argued that the choice of a loss function, while less crucial than that of the model, exerts an important influence on the nature of the solution of a statistical decision problem, and that an arbitrary choice such as squared error may be baldly misleading as to the relative desirability of the competing procedures." (p. 425). Tukey (1960) went even further arguing that the decision-theoretic framing distorts frequentist testing by replacing error probabilities with losses and costs: "Wald's decision theory ... has given up fixed probability of errors of the first kind, and has focused on gains, losses or regrets." (p. 433). He went on to echo Fisher's (1955) view by contrasting decisions vs. inferences: "Conclusions are established with careful regard to evidence, but without regard to consequences of specific actions in specific circumstances." (p. 425).

Hacking (1965) brought out the key difference between an 'inference pertaining to evidence' for or against a hypothesis, and a 'decision to do something' as a result of an inference: "... to conclude that an hypothesis is best supported is, apparently, to decide that the hypothesis in question is best supported. Hence it is a decision like any other. But this inference is fallacious. Deciding that something *is* the case differs from deciding to *do* something. ... Hence deciding to do something falls squarely in the province of decision theory, but deciding that something is the case does not." (p. 31).

8.2 Bayesian Credible Intervals

A Bayesian $(1-\alpha)$ credible interval for θ is constructed by finding the area between the $\frac{\alpha}{2}$ and $(1-\frac{\alpha}{2})$ percentiles of the posterior distribution, say a and b, respectively:

$$\Pi(a \leq \theta < b)=1-\alpha, \quad \int_a^1 \pi(\theta|\mathbf{x}_0)d\theta=(1-\tfrac{\alpha}{2}), \ \int_b^1 \pi(\theta|\mathbf{x}_0)d\theta=\tfrac{\alpha}{2}, \tag{42}$$

where $\Pi(.)$ denotes probabilistic assignments based on the *posterior* distribution $\pi(\theta|\mathbf{x}_0)$, $\forall\theta\in\Theta$. In practice one can define an infinity of $(1-\alpha)$ credible intervals using the same posterior $\pi(\theta|\mathbf{x}_0)$. To avoid this indeterminancy one needs to impose additional restrictions like the interval with the *shortest length* or one with *equal tails*; see Robert (2007).

Example 11.37. In the case of the simple Bernoulli model (table 11.1), the end points of an equal-tail credible interval can be evaluated by transforming the Beta distribution into the F distribution via:

$$Z \backsim \mathsf{Beta}(\alpha^*, \beta^*) \Rightarrow \quad \tfrac{\beta^* Z}{\alpha^*(1-Z)} \backsim \mathsf{F}(2\alpha^*, 2\beta^*). \tag{43}$$

Denoting the $\frac{\alpha}{2}$ and $(1-\frac{\alpha}{2})$ percentiles of the $\mathsf{F}(2\alpha^*, 2\beta^*)$ distribution, by $\mathsf{f}(\frac{\alpha}{2})$ and $\mathsf{f}(1-\frac{\alpha}{2})$, respectively, the Bayesian $(1-\alpha)$ credible interval for θ is:

$$\left(1+\tfrac{\beta^*}{\alpha^*\mathsf{f}(1-\frac{\alpha}{2})}\right)^{-1} \leq \theta \leq \left(1+\tfrac{\beta^*}{\alpha^*\mathsf{f}(\frac{\alpha}{2})}\right)^{-1}. \tag{44}$$

Example 11.38. For the simple Bernoulli model (table 11.1), with Jeffreys prior $\pi_J(\theta) \backsim \mathsf{Beta}(.5, .5)$, $n\overline{x}$=2, n=20, α=.05:

$$\alpha^*=n\overline{x}+\alpha=2.5, \ \beta^*=n(1-\overline{x})+\beta=18.5, \ \mathsf{f}(1-\tfrac{\alpha}{2})=.163, \ \mathsf{f}(\tfrac{\alpha}{2})=2.93,$$

$$\left(1+\tfrac{18.5}{2.5(.163)}\right)^{-1} \leq \theta \leq \left(1+\tfrac{18.5}{2.5(2.93)}\right)^{-1} \Leftrightarrow \ (.0216 \leq \theta \leq .284)\,. \tag{45}$$

It important to contrast the frequentist Confidence Interval (CI) with the Bayesian Credible Interval to bring out their key differences.

The first important difference is that the basis of a CI is a pivot $q(\mathbf{X}, \theta)$, whose sampling distribution is evaluated under $\theta{=}\theta^*$, with the probability firmly attached to $\mathbf{x}{\in}R^n_X$. In contrast, a $(1{-}\alpha)$ Credible Interval represents the highest posterior density interval. That is, it is simply the shortest interval whose area (integral) under the posterior density function $\pi(\theta|\mathbf{x}_0)$ has value $(1{-}\alpha)$, where the probabilities are firmly attached to $\theta{\in}\Theta$. Hence, any comparison of tail areas amounts to likening apples to oranges.

The second key difference is that the CI:

$$\left(\widehat{\theta}_L(\mathbf{X}) \leq \theta \leq \widehat{\theta}_U(\mathbf{X});\ \theta{=}\theta^*\right),$$

is **random** and its primary purpose is to cover θ^* with probability $(1{-}\alpha)$. There is nothing random about a $(1{-}\alpha)$ Credible Interval, and nothing connects that interval to θ^*. Bayesians would like to think that it does, since it covers a large part of the posterior, but there is nothing in the above derivation or its underlying reasoning that ensures that. Indeed, the very idea of treating θ as a random variable runs afoul any notion of true value θ^* that could have generated data $\mathbf{x}_0$. In the case of a CI, the evaluation of the sampling distribution of $q(\mathbf{X}, \theta)$ under $\theta{=}\theta^*$ secures exactly that.

9 Summary and conclusions

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

▶ **What combination of properties define an optimal estimator**?

A *minimal property* (necessary but not sufficient) for an estimator is *consistency* (weak or strong). As an extension of the Law of Large Numbers (LLN), a consistent estimator of θ indicates potential (as $n \to \infty$) learning from data about the unknown parameter θ. By itself, however, it does not secure learning for a particular n. This suggests that going from potential learning to actual learning one needs to supplement consistency with certain finite sample properties to ensure learning for the particular data $\mathbf{x}_0$ of sample size n. Among finite sample properties *full efficiency* is clearly the most important because it secures the highest degree of precision in learning for a given n. *Relative efficiency*, although desirable, needs to be investigated further to find out how large is the class of estimators being compared before passing judgement. *Unbiasedness*, although wanted, 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 for a large enough n. Sufficiency is clearly a desirable property because it ensures that no information relevant for inference involving θ is forfeited.

In summary, a *consistent, unbiased, fully efficient and sufficient* estimator sets the gold standard in estimation. When no consistent estimator can achieve this standard, one should be careful about trading a loss in efficiency to secure unbiasedness. Analogously, *minimum MSE*, when properly defined at $\theta{=}\theta^*$, is not a particularly essential property by itself.

What about the case where an estimator is *consistent, asymptotically Normal* (CAN) and possibly *asymptotically efficient.* Although in practice statisticians and econometricians consider CAN as being close to the gold standard, the fact of the matter is that *relying exclusively on asymptotic properties* **is a bad strategy in general**, as illustrated in chapter 9. The reason is that the asymptotic sampling distribution $f_\infty(\widehat{\theta}_n;\theta^*)$, because it asserts what happens *in the limit*, might provide a terrible approximation for $f_n(\widehat{\theta}_n;\theta^*)$, the relevant distribution for inferences with data $\mathbf{x}_0$. Even worse, one has no way to make an informed appraisal of how bad this approximation might be for a given n. Hence, relying solely on asymptotic properties like CAN is *not* a good strategy for learning from data, because the reliability of inference is at best non-ascertainable and at worst highly misleading. Recall Le Cam (1986):

"... limit theorems 'as n tends to infinity' are logically devoid of content about what happens at any particular n.

Point estimation **does not**, by itself, output an *inferential claim* of the form:

$$\widehat{\theta}_n(\mathbf{x}_0) \simeq \theta^*, \quad \times$$

irrespective of how optimal the estimator $\widehat{\theta}_n(\mathbf{X})$ is. The reason is that $\mathbf{x}_0$ is one point in the sample space $\mathbb{R}^n_X$, which often allows for an infinite number of values for $\mathbf{x}$. It needs to be supplemented by a certain measure of the precision associated with the estimate $\widehat{\theta}_n(\mathbf{x}_0)$. This is the reason why $\widehat{\theta}_n(\mathbf{x}_0)$ is often accompanied by its standard error $[SE(\widehat{\theta}_n(\mathbf{X}))]$. *Interval estimation* rectifies this omission of point estimation by providing the relevant inferential claim based on a pre-specified *coverage error probability* for the true value θ^*.

Additional references: Arnold (1990), Azzalini (1996), Davison (2003), Spanos (2008).

Important concepts

Ideal estimator, sampling distribution of an estimator, finite sample properties of estimators, unbiasedness, relative efficiency, full efficiency, Cramer-Rao lower bound, Fisher's information, irregular statistical models, sufficiency, Mean Square Error (MSE), Bias of an estimator, admissibility of an estimator, asymptotic properties of estimators, weak consistency, strong consistency, consistency as a minimall property, asymptotic Normality, asymptotic efficiency, mode and median unbiased estimator, inferential claims and confidence intervals, long-run metaphor, pivotal function, coverage probability, minimum length confidence intervals, optimal Bayes rules, Loss and Risk functions, factorization theorem, minimal sufficiency, completeness, exponential family of distributions.

Crucial distinctions

Ideal vs. feasible estimator, finite sample vs. asymptotic properties, relative efficiency vs. full efficiency, frequentist definition of the MSE vs. the Bayesian definition, point vs. interval estimation, statistical modeling vs. statistical inference, Confidence Intervals vs. Credible Intervals.

Essential ideas

- Properties of an estimator aim to gauge its generic capacity to pinpoint θ^* for all values $\mathbf{X}=\mathbf{x}$, $\mathbf{x}\in\mathbb{R}_X^n$.

- The gold standard for an optimal estimator $\widehat{\theta}_n(\mathbf{X})$ comes in the form of a combination of properties. $\widehat{\theta}_n(\mathbf{X})$ needs to satisfy consistency as a minimal property, combined with full efficiency and sufficiency when the latter exists. Reparameterization invariance is a more desirable property than unbiasedness.

- An optimal point estimator $\widehat{\theta}_n(\mathbf{X})$, although fundamental for all forms of statistical inference (confidence intervals, hypothesis testing, prediction), does not carry with it an inferential claim, such as $\widehat{\theta}_n(\mathbf{x}_0)$ is close enough to θ^*.

- A consistent and asymptotically Normal (CAN) estimator $\widehat{\theta}_n(\mathbf{X})$ does not guarantee the reliability of any inference procedure based on it.

- The Bayesian definition of the Mean Square Error (MSE), based on the quantifier $\forall\theta\in\Theta$, and the related property of admissibility, are at odds with the underlying reasoning and primary aim of frequentist estimation. Consistency, and not admissibility, is the relevant minimal property for frequentist estimators.

- In statistical models for which the minimal sufficient and maximal ancillary statistics co-exist, one can separate the modeling from the inference facet, by using the ancillary statistic for the former and the sufficient statistic for the latter. The exponential family of distributions includes several such statistical models that are widely used in empirical modeling, including the simple Normal and the Linear Regression models.