Mathematical Statistics

Daniela Rodriguez

Universidad Torcuato Di Tella and CONICET

December 31, 2025

Contents

Chapter 1

Formulation of The Problem of Statistical Estimation.

1.1 Introduction

This course introduces the mathematical principles that form the foundation of statistical theory. The central goal of statistics is to infer properties of an unknown probability distribution from a set of observed data points. One can think of the discipline from two general perspectives:

Applied statistics: This involves the methods for data collection and data analysis used across various fields like the natural sciences, engineering, medicine, and business.

Theoretical statistics: This provides the mathematical framework for understanding the properties and scope of statistical methods.

While there is no single, unifying theory of statistics that can solve every problem posed by a data analyst, a core unifying idea of the field is the concept of statistical models.

Statistical inference is the process of drawing conclusions about a larger, unknown system based on a small set of data. The mathematical foundation for this is provided by probability models.

We can break down this process into three main steps:

1. **Define the Model**: We start with an assumption that the data comes from a specific type of probability model. This model has a known structure but depends on one or more unknown parameters, which we represent with the symbol θ.

2. **Collect the Data**: We then gather a sample of data—for example, n independent observations $(X_1, \ldots, X_n)$. We know this data was generated by our chosen model, but we don't know the exact value of the parameter θ that created it.

3. **Make Inferences**: Our goal is to use this observed data to make educated guesses about the true value of θ and to understand how certain or uncertain our guesses are.

Given this framework, the field of statistics has three primary goals:

- **Estimation**: This is about creating a single "best guess" for the unknown parameter θ. We construct a function of our data, called an estimator $(\hat{\theta})$, that should be as close as possible to the true value of θ.

- **Inference (Uncertainty Quantification)**: This goes beyond a single guess to provide a range of plausible values for θ. We find a confidence interval (C_n) so that we can be highly confident (e.g., 95% confident) that the true value of θ lies within this range. This helps us quantify the uncertainty in our estimate.

- **Hypothesis Testing**: This involves using the data to decide between two competing claims or hypotheses. We set up a null hypothesis (H_0, a default assumption like $\theta = \theta_0$) and an alternative hypothesis (H_1, a competing claim like $\theta \neq \theta_0$). We then use a statistical test to determine which of these two statements is better supported by the evidence from our data.

1.2 Statistical Models

Consider a real-valued random variable X, on a probability space Ω, with distribution defined for all $t \in \mathbb{R}$ by

$$F(t) = P(\omega \in \Omega : X(\omega) \leq t).$$

When X is discrete it is equal to

$$F(t) = \sum_{x \leq t} f(x),$$

and f is called the probability mass function of X (p.m.f.). When X is continuous it is equal to

$$F(t) = \int_{-\infty}^{t} f(x)\, dx,$$

and f is called the probability density function of X (p.d.f.).

We write $X \sim F$ to state that F is the distribution of X. If $\{X_i\}_{i \in I}$ is a collection of independent identically distributed random variables with distribution F , we write $X_i \sim F$ iid. The distribution F will typically depend on one or several parameters that we shall represent as $\theta = (\theta_1, \ldots, \theta_p)^T \in \Theta \subset \mathbb{R}^p$. The space Θ where the parameter θ belongs is called the **parameter space**. To indicate that the distribution F depends on the parameter θ, we will often write F_θ (or $F(x|\theta)$,or $F(x, \theta)$).

Definition 1 (Statistical Model)**.** *A **statistical model** for a sample from X is any family $\{f(\theta, \cdot) : \theta \in \Theta\}$ of p.m.f. or p.d.f. $f(\theta, \cdot)$, or $\{P_\theta : \theta \in \Theta\}$ ($\{F_\theta : \theta \in \Theta\}$) of probability distribution for the law of X (P_θ or F_θ) with parameter space $\Theta \subset \mathbb{R}^p$.*

Simply put, the model F_θ cannot switch between continuous and discrete depending on the value of θ.

Example 1. *Some statistical models and their parameter spaces*

1. *$N(\theta, 1)$; $\theta \in \Theta = \mathbb{R}$.*
2. *$N(\mu, \sigma^2)$; $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$.*
3. *Exp(θ); $\theta \in \Theta = (0, \infty)$.*
4. *$N(\theta, 1)$; $\theta \in \Theta = [-1, 1]$.*

We will assume that the statistical model is well specified, i.e. such that $F = F_{\theta_0}$ for some $\theta_0 \in \Theta$. In words, we assume that the true generating probability law F belongs to the family of distributions postulated by the statistical model.

Definition 2. *For a variable X with distribution F, we say that the model $\{F_\theta : \theta \in \Theta\}$ is **correctly specified** if there exists $\theta_0 \in \Theta$ such that $F_{\theta_0} = F$.*

We will often write θ_0 for the true value of θ to distinguish it from other elements of the parameter space Θ. This particular θ_0 is called the *true parameter*. We will say that the X_i are i.i.d. from the model $\{P_\theta : \theta \in \Theta\}$ in this case.

Example 2. *A very easy example. If we want to know what the percentage of people who like Coke is, we can think of a variable (X) with values **1** if they like it, and **0** if they do not.*

Let p be the probability that they like it: $P(X = 1) = p$.

That is, we have a statistical model $Ber(p)$, and the parameter p is identified. Indeed, trivially a different parameter $p_0 \neq p$ will lead to a model $Ber(p_0)$ that will generate data with a different distribution from that of $Ber(p)$.

For example, if half the people like it and half do not, p will be $1/2$.

Example 3. *As an example, if $X \sim N(2, 1)$ the model in i) is correctly specified but the model in iv) is not.*

Example 4. *If $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ iid , but we only observe $Y_i = \mathbb{I}(X_i \geq 0)$ for $i = 1, \ldots, n$. In this case the parameters μ and σ^2 are not identified. To see this first note that*

$$P(Y_i = 1) = P(X_i \geq 0) = 1 - P(X_i \leq 0) = 1 - \Phi(-\frac{\mu}{\sigma}) = \Phi(\frac{\mu}{\sigma}).$$

Since the ratio $\theta = \mu/\sigma$ completely determines the distribution of the observed random sample $Y_1, \ldots, Y_n$, we can easily see that the pairs $(c\mu, c\sigma)$ and (μ, σ) lead to the same distribution of Y_i for any $c > 0$. In this case only $\theta = \mu/\sigma$ is identified.

When F depends on a parameter θ, we still have

$$F_\theta(t) = P[X \leq t].$$

Since the left-hand side depends on θ, the right-hand side also must depend on θ, even though this is not explicit in our notation. Sometimes we will need to make that clear, in which case we will write P_θ instead of just P in order to remind ourselves of this dependence. Similarly, we will sometimes write E_θ instead of just E for the expectation of X when its distribution is $F_\theta(x)$.

1.3 Exponential Families of Distributions

At a glance, it might not be obvious, but many of the probability models we've studied—both discrete and continuous—share fundamental structural properties. We can therefore introduce a more abstract framework and view these models as specific instances of a larger family: the **exponential family of distributions**. This approach is powerful because any theorems we prove for this general family automatically apply to all its members.

Definition 3 (The Exponential Family of Distributions)**.** *A regular probability distribution is said to be a member of a* ***k-parameter exponential family****, if its density (or probability mass function) can be written in the following form:*

$$f(x) = \exp\left(\sum_{i=1}^{k} \eta_i T_i(x) - A(\eta) + S(x)\right) = \exp\{\eta^T T(x) - A(\eta) + S(x)\}, \quad x \in \mathcal{X}; \ (1.1)$$

where:

1. $\eta = (\eta_1, \ldots, \eta_k)^t$ *is a* k*-dimensional parameter in* $\mathbb{R}^k$*;*
2. $T(x) = (T_1(x), \ldots, T_k(x))^t$ *and* $T_i : \mathcal{X} \to \mathbb{R}$*,* $S(x) : \mathcal{X} \to \mathbb{R}$*, and* $A : \mathbb{R}^k \to \mathbb{R}$ *are real-valued functions;*
3. *The sample space* $\mathcal{X}$ *does not depend on* η*.*

Remark 1. *The parameter* η *is known as the natural parameter.*

Remark 2. *The presence of the exponential function is not the most significant feature of this family. Any density can be written this way. The key characteristic is that the density can be factored into three distinct parts: one that depends solely on the parameter, one that depends only on the data, and a third that connects both in a specific manner as a linear combination of the coordinates of* η *with coefficients that are functions of* x*.*

Remark 3. *The exponential family should not be confused with the exponential distribution. To avoid mix-ups, we always use the word "family" to distinguish the broader concept.*

We will see some example of the exponential family. To do this, we'll need to manipulate their density or frequency functions to match the form in Equation 1.1. Often, the standard parameters used for a distribution don't align with the natural parameters. However, there is typically a smooth, one-to-one transformation between them, so the density can also be written in this form:

$$\exp\left(\sum_{i=1}^{k} c_i(\theta) T_i(x) - d(\theta) + S(x)\right).$$

Both versions are valid, but the natural representation is generally preferred for theoretical work and proving theorems because the parameter appears linearly in the exponent. In contrast, the usual representation is more common in practical applications.

Example 5. *Let* $X \sim Binom(n, p)$*. Recall that this means that* $X \in \{0, 1, 2, \ldots, n\}$ *and* $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$*. Now, we may take the log and then exponentiate to obtain:*

$$\binom{n}{x} p^x (1-p)^{n-x} = \exp\left\{\log\left(\frac{p}{1-p}\right) x + n\log(1-p) + \log\left(\binom{n}{x}\right)\right\}$$

Define:

$$\eta = \log\left(\frac{p}{1-p}\right); \quad T(x) = x; \quad S(x) = \log\left(\binom{n}{x}\right); \quad A(\eta) = -n\log(1-p) = n\log(1+e^{\eta})$$

Thus, if n is held fixed and only p is allowed to vary, the support of f does not depend on η and so we see that the Binomial with fixed n is a 1-parameter exponential family. Here the usual parameter p is a twice differentiable bijection of the natural parameter η:

$$p = \frac{e^\eta}{1+e^\eta} \quad \textit{and} \quad \eta = g(p) = \log\left(\frac{p}{1-p}\right)$$

Here $p \in (0,1)$ but $\eta \in \mathbb{R}$.

Example 6. *Let $X \sim N(\mu, \sigma^2)$. Then we may write:*

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\}$$

Define:

$$\eta_1 = \frac{\mu}{\sigma^2}; \quad \eta_2 = -\frac{1}{2\sigma^2}; \quad T_1(x) = x; \quad T_2(x) = x^2;$$

$$S(x) = -\frac{1}{2}\log(2\pi); \quad A(\eta_1,\eta_2) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-\frac{1}{2\eta_2})$$

and also observe that the support of f is always $\mathbb{R}$, regardless of the parameter values. It follows that the $N(\mu,\sigma^2)$ distribution is a 2-parameter exponential family.

Example 7. *Let $X \sim Unif(\theta_1,\theta_2)$. The support of this distribution is the interval $[\theta_1,\theta_2]$, which clearly depends on the parameters. Because the sample space is not fixed, the uniform distribution does not belong to the exponential family.*

Theorem 1. *Let $\mathbf{X} = (X_1,\ldots,X_q)$ be a random vector whose distribution belongs to a one-parameter exponential family with density given by*

$$p(\mathbf{x},\theta) = A(\theta)e^{c(\theta)T(\mathbf{x})}h(\mathbf{x}) \quad \textit{with } \theta \in \Theta,$$

where Θ is an open set in $\mathbb{R}$ and $c(\theta)$ is infinitely differentiable. Then we have:

(i) $A(\theta)$ is infinitely differentiable.

(ii)

$$E_\theta(T(\mathbf{X})) = -\frac{A'(\theta)}{A(\theta)c'(\theta)}$$

(iii)

$$Var_\theta(T(\mathbf{X})) = \frac{1}{c'(\theta)}\frac{\partial E_\theta(T(\mathbf{X}))}{\partial\theta}$$

Lemma 1. *Let $\mathbf{X} = (X_1,\ldots,X_q)$ be a random vector whose distribution belongs to a discrete or continuous one-parameter exponential family with density given by $p(\mathbf{x},\theta) = A(\theta)e^{c(\theta)T(\mathbf{x})}h(\mathbf{x})$; with $\theta \in \Theta$, where Θ is an open set in $\mathbb{R}$ and $c(\theta)$ is infinitely differentiable. Then, if $m(\mathbf{x})$ is a statistic such that*

$$\int\cdots\int |m(\mathbf{x})|p(\mathbf{x},\theta)dx_1\cdots dx_q < \infty \quad \forall\theta\in\Theta \quad (\textit{continuous case})$$

$$\sum_{x_1} \cdots \sum_{x_q} |m(\mathbf{x})| p(\mathbf{x}, \theta) < \infty \quad \forall \theta \in \Theta \quad (\textit{discrete case})$$

holds, then the derivative with respect to θ can be taken inside the integral/summation sign.

Proof. Suppose X is continuous. The discrete case is completely similar. Since

$$\int \cdots \int A(\theta) e^{c(\theta)T(\mathbf{x})} h(\mathbf{x}) dx_1 \cdots dx_q = 1$$

we have

$$\frac{1}{A(\theta)} = \int \cdots \int e^{c(\theta)T(\mathbf{x})} h(\mathbf{x}) dx_1 \cdots dx_q$$

As the right-hand side of this equality satisfies the conditions of Lemma 1 with $m(\mathbf{x}) = 1$, it follows that the right-hand side is infinitely differentiable. Consequently, $A(\theta)$ is also infinitely differentiable, which proves (i).

Furthermore, we have

$$A(\theta) \int \cdots \int e^{c(\theta)T(\mathbf{x})} h(\mathbf{x}) dx_1 \cdots dx_q = 1 \quad \forall \theta \in \Theta$$

and using Lemma 1, which allows us to differentiate inside the integral sign, we get

$$A'(\theta) \int \cdots \int e^{c(\theta)T(\mathbf{x})} h(\mathbf{x}) dx_1 \cdots dx_q + A(\theta) c'(\theta) \int \cdots \int T(\mathbf{x}) e^{c(\theta)T(\mathbf{x})} h(\mathbf{x}) dx_1 \cdots dx_q = 0$$

Then:

$$\frac{A'(\theta)}{A(\theta)} \int \cdots \int A(\theta) e^{c(\theta)T(\mathbf{x})} h(\mathbf{x}) dx_1 \cdots dx_q + c'(\theta) \int \cdots \int T(\mathbf{x}) A(\theta) e^{c(\theta)T(\mathbf{x})} h(\mathbf{x}) dx_1 \cdots dx_q = 0$$

Recognizing that $p(\mathbf{x}, \theta) = A(\theta) e^{c(\theta)T(\mathbf{x})} h(\mathbf{x})$ and substituting back the expected value $E_\theta(T(\mathbf{X}))$: and thus

$$E_\theta(T(\mathbf{X})) = -\frac{A'(\theta)}{A(\theta) c'(\theta)}$$

which proves (ii).

(iii) The strategy is to differentiate the expected value, $E_\theta(r(\mathbf{X}))$, with respect to θ. We apply the chain rule by differentiating the integral definition of the expected value.

$$\begin{aligned} \frac{\partial E_\theta(r(\mathbf{X}))}{\partial \theta} &= \frac{\partial}{\partial \theta} \left[\int T(\mathbf{x}) p(\mathbf{x}, \theta) d\mathbf{x} \right] = \int T(\mathbf{x}) \frac{\partial p(\mathbf{x}, \theta)}{\partial \theta} d\mathbf{x} \\ &= \int T(\mathbf{x}) \frac{\frac{\partial p(\mathbf{x}, \theta)}{\partial \theta}}{p(\mathbf{x}, \theta)} p(\mathbf{x}, \theta) d\mathbf{x} = \int T(\mathbf{x}) \frac{\partial \log p(\mathbf{x}, \theta)}{\partial \theta} p(\mathbf{x}, \theta) d\mathbf{x} \end{aligned}$$

Next, we use the fact that the derivative of the logarithm of the density is

$$\frac{\partial \log p(\mathbf{x}, \theta)}{\partial \theta} = \frac{A'(\theta)}{A(\theta)} + c'(\theta) T(\mathbf{x})$$

Substituting this back into the expression for the derivative of the expectation:

$$\frac{\partial E_\theta(T(\mathbf{X}))}{\partial \theta} = \int T(\mathbf{x}) \left[\frac{A'(\theta)}{A(\theta)} + c'(\theta) T(\mathbf{x}) \right] p(\mathbf{x}, \theta) d\mathbf{x}$$

We separate the integral terms and factor out $c'(\theta)$ from the second term:

$$\frac{\partial E_\theta(T(\mathbf{X}))}{\partial \theta} = \frac{A'(\theta)}{A(\theta)} \int T(\mathbf{x}) p(\mathbf{x}, \theta) d\mathbf{x} + c'(\theta) \int T^2(\mathbf{x}) p(\mathbf{x}, \theta) d\mathbf{x}$$

$$\frac{\partial E_\theta(T(\mathbf{X}))}{\partial \theta} = \frac{A'(\theta)}{A(\theta)} E_\theta(T(\mathbf{X})) + c'(\theta) E_\theta(T^2(\mathbf{X}))$$

Now we substitute the result from part (ii), $A'(\theta)/A(\theta) = -c'(\theta) E_\theta(T(\mathbf{X}))$:

$$\frac{\partial E_\theta(T(\mathbf{X}))}{\partial \theta} = \left[-c'(\theta) E_\theta(T(\mathbf{X})) \right] E_\theta(T(\mathbf{X})) + c'(\theta) E_\theta(T^2(\mathbf{X}))$$

Factoring out $c'(\theta)$:

$$\frac{\partial E_\theta(T(\mathbf{X}))}{\partial \theta} = c'(\theta) \left[E_\theta(T^2(\mathbf{X})) - (E_\theta(T(\mathbf{X})))^2 \right] = c'(\theta) \cdot \mathrm{Var}_\theta(T(\mathbf{X}))$$

1.3.1 Sampling from a k-Parameter Exponential Family

Consider a probability model described by a k-parameter exponential family. The density or mass function of a single random variable X is expressed in the canonical form:

$$f(x; \boldsymbol{\eta}) = h(x) \exp\left(\sum_{j=1}^{k} \eta_j T_j(x) - A(\boldsymbol{\eta}) \right), \quad x \in \mathcal{X},$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)^\top$ is the vector of **natural parameters**, $T_j(x)$ are the component statistics, and $A(\boldsymbol{\eta})$ is the log-normalizer (or cumulant-generating function).
Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample of size n, where $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) according to $f(x; \boldsymbol{\eta})$.

The joint probability function of the sample $\mathbf{X}$ is the product of the individual densities:

$$f(\mathbf{x}; \boldsymbol{\eta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\eta})$$

Substituting the canonical form and rearranging terms, we obtain the joint distribution in its exponential family form:

$$\begin{aligned} f(\mathbf{x}; \boldsymbol{\eta}) &= \prod_{i=1}^{n} \left[h(x_i) \exp\left(\sum_{j=1}^{k} \eta_j T_j(x_i) - A(\boldsymbol{\eta}) \right) \right] \\ &= \left(\prod_{i=1}^{n} h(x_i) \right) \exp\left(\sum_{j=1}^{k} \eta_j \left(\sum_{i=1}^{n} T_j(x_i) \right) - nA(\boldsymbol{\eta}) \right) \\ &= H(\mathbf{x}) \exp\left(\sum_{j=1}^{k} \eta_j \cdot \mathbf{T}_{n,j}(\mathbf{x}) - nA(\boldsymbol{\eta}) \right) \end{aligned}$$

where $H(\mathbf{x}) = \prod_{i=1}^{n} h(x_i)$, $\mathbf{T}_{n,j}(\mathbf{x}) = \sum_{i=1}^{n} T_j(x_i)$ and the vector of summed component statistics

$$\mathbf{T}_n(\mathbf{X}) = \left(\sum_{i=1}^{n} T_1(X_i), \ldots, \sum_{i=1}^{n} T_k(X_i) \right)^\top.$$

Theorem 2. *Let $X_1, \ldots, X_n$ be a random samplre of size n where $\mathbf{X_i}$ are distributed according to a **one-parameter exponential family** with density given by*

$$p(\mathbf{x}, \theta) = A(\theta) e^{c(\theta) T(\mathbf{x})} h(\mathbf{x}) \quad \text{with } \theta \in \Theta,$$

where Θ is an open set in $\mathbb{R}$ and $c(\theta)$ is infinitely differentiable. Then we can compute the expected value and the variance of the statistic $T_n(\mathbf{X}) = \sum_{i=1}^{n} T(X_i)$ by:

$$E_\theta(T_n(\mathbf{X})) = -n \frac{A'(\theta)}{A(\theta) c'(\theta)}$$

$$Var_\theta(T_n(\mathbf{X})) = n \frac{1}{c'(\theta)} \frac{\partial E_\theta(T(\mathbf{X}))}{\partial \theta}$$

1.4 Review of some useful probability tools

1.4.1 Expected value

Definition 4 (Expected Value (Mean))**.** *The **expected value** of a random variable X, denoted as $E[X]$ or μ, is the weighted average of all possible values that X can take.*

For a discrete random variable X with probability mass function $p(x)$:

$$E[X] = \sum_x x \cdot p(x)$$

For a continuous random variable X with probability density function $f(x)$:

$$E[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

Definition 5 (Variance)**.** *The **variance** of a random variable X, denoted as $Var(X)$ or σ^2, measures the spread or dispersion of its values around the expected value.*

$$Var(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

Definition 6 (Covariance of Random Vectors)**.** *The **covariance** between two random vectors, $\mathbf{X} \in \mathbb{R}^p$ and $\mathbf{Y} \in \mathbb{R}^q$, denoted as $Cov(\mathbf{X}, \mathbf{Y})$, is a $p \times q$ matrix that measures the degree to which their components change together. Its element at row i and column j is the covariance between X_i and Y_j.*

The fundamental definitions are as follows:

$$\textit{Expected Value of a Vector:} \quad E[\mathbf{X}] = E\begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} = \begin{pmatrix} E[X_1] \\ \vdots \\ E[X_p] \end{pmatrix}$$

$$\textit{Covariance Matrix:} \quad Cov(\mathbf{X}, \mathbf{Y}) = E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{Y} - E[\mathbf{Y}])^\top]$$

$$\textit{Variance-Covariance Matrix:} \quad Var(\mathbf{X}) = Cov(\mathbf{X}, \mathbf{X}) = E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^\top]$$

The variance-covariance matrix is symmetric and positive semi-definite.

Remark 4. *The expected value of a random variable has another property, one that we can think of as relating to the interpretation of $E[X]$ as a good guess at a value of X.*

Suppose we measure the distance between a random variable X and a constant b by $(X-b)^2$. It does no good to look for a value of b that minimizes $(X-b)^2$, since the answer would depend on the random values of $X>$

The closer b is to X, the smaller this quantity is. We can now determine the value of b that minimizes $E[(X-b)^2]$ and, hence, will provide us with a good predictor of X.

We could proceed with the minimization of $E[(X-b)^2]$ with respect to b using calculus, but there is a simpler method.

$$\begin{aligned} E[(X-b)^2] &= E[(X-E[X]+E[X]-b)^2] \quad (\textit{add and subtract } E[X],\ \textit{then group terms}) \\ &= E[(X-E[X])^2+(E[X]-b)^2+2(X-E[X])(E[X]-b)] \quad (\textit{expand the square}) \end{aligned}$$

Now, note that

$$E[(X-E[X])(E[X]-b)] = (E[X]-b)E[X-E[X]] = 0,$$

since $(E[X]-b)$ is constant and comes out of the expectation, and $E[X-E[X]] = E[X]-E[E[X]] = E[X]-E[X] = 0$. This means that

$$E[(X-b)^2] = E[(X-E[X])^2] + (E[X]-b)^2. \tag{1.2}$$

We have no control over the first term on the right-hand side of (1.2), and the second term, which is always greater than or equal to 0, can be made equal to 0 by choosing $b = E[X]$. Hence,

$$\min_b E[(X-b)^2] = E[(X-E[X])^2]. \tag{1.3}$$

1.4.2 Convergence

Definition 7. *Let $X_1, X_2, \ldots$ be a sequence of random variables and let X be another random variable. Let F_n denote the cdf of X_n and let F denote the cdf of X.*

1. *X_n converges to X in probability, written $X_n \xrightarrow{P} X$, if, for every $\varepsilon > 0$,*

$$P(|X_n - X| > \varepsilon) \to 0 \quad \textit{as } n \to \infty. \tag{1.4}$$

2. *X_n converges to X in distribution, written $X_n \xrightarrow{D} X$, if*

$$\lim_{n\to\infty} F_n(t) = F(t) \tag{1.5}$$

   at all t for which F is continuous.

3. *X_n converges to X in quadratic mean (also called convergence in L_2), written $X_n \xrightarrow{qm} X$, if*

$$E(X_n - X)^2 \to 0 \tag{1.6}$$

   as $n \to \infty$.

When the limiting random variable is a point mass, we change the notation slightly. If $P(X = c) = 1$ and $X_n \xrightarrow{P} X$ then we write $X_n \xrightarrow{P} c$. Similarly, if $X_n \xrightarrow{D} X$ we write $X_n \xrightarrow{D} c$.
The next theorem gives the relationship between the types of convergence.

Theorem 3 (Relationship between Convergences)**.** *The following relationships hold:*

(a) $X_n \xrightarrow{qm} X$ *implies that* $X_n \xrightarrow{P} X$.

(b) $X_n \xrightarrow{P} X$ *implies that* $X_n \xrightarrow{D} X$.

(c) If $X_n \xrightarrow{D} X$ *and if* $P(X = c) = 1$ *for some real number* c*, then* $X_n \xrightarrow{P} X$.

In general, none of the reverse implications hold except the special case in (c).

Let us now show that the reverse implications do not hold.

Convergence in probability does not imply convergence in quadratic mean.

Let $U \sim \text{Unif}(0,1)$ and let $X_n = \sqrt{n} I_{(0,1/n)}(U)$. Then $P(|X_n| > \varepsilon) = P(\sqrt{n} I_{(0,1/n)}(U) > \varepsilon) = P(0 \leq U < 1/n) = 1/n \to 0$. Hence, $X_n \xrightarrow{P} 0$. But

$$E(X_n^2) = E\left(\left(\sqrt{n} I_{(0,1/n)}(U)\right)^2\right) = \int_0^{1/n} n\, du = n\,[u]_0^{1/n} = 1$$

for all n. Thus, X_n does not converge in quadratic mean.

Convergence in distribution does not imply convergence in probability.

Let $X \sim N(0,1)$. Let $X_n = -X$ for $n = 1, 2, 3, \ldots$; hence $X_n \sim N(0,1)$. X_n has the same distribution function as X for all n so, trivially, $\lim_n F_n(x) = F(x)$ for all x. Therefore, $X_n \xrightarrow{D} X$. But $P(|X_n - X| > \varepsilon) = P(|-X - X| > \varepsilon) = P(|2X| > \varepsilon) = P(|X| > \varepsilon/2) \neq 0$. So X_n does not converge to X in probability.

Warning.

One might conjecture that if $X_n \xrightarrow{P} b$, then $E(X_n) \to b$. This is not true. Let X_n be a random variable defined by $P(X_n = n^2) = 1/n$ and $P(X_n = 0) = 1 - (1/n)$. Now, $P(|X_n| > \varepsilon) = P(X_n = n^2) = 1/n \to 0$. Hence, $X_n \xrightarrow{P} 0$. However, $E(X_n) = [n^2 \times (1/n)] + [0 \times (1 - (1/n))] = n$. Thus, $E(X_n) \to \infty$.

Theorem 4. *Law of Large numbers Let* $X_1, \ldots, X_n$ *be i.i.d. random variables with mean* μ *and such that* $E[|X_i|] < \infty$*. Then*

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{P;a.s.} \mu.$$

Theorem 5. *Central Limit Theorem Let $X_1, \ldots, X_n$ be i.i.d. random variables with mean μ and $\sigma^2 < \infty$. Then*

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{D} \mathcal{N}(0,1).$$

Example 1: Poisson Distribution

Let $X_1, \ldots, X_n$ be independent and identically distributed (iid) random variables, $X_i \sim \text{Poisson}(\lambda)$.

Let $S_n = \sum_{i=1}^{n} X_i$. Since the Poisson distribution is stable under summation, the sum itself is exactly Poisson distributed:

$$S_n \sim \text{Poisson}(n\lambda).$$

However, the **Central Limit Theorem (CLT)** gives us an important asymptotic approximation. It tells us that, for large n, the standardized sample mean ($\bar{X}_n$) converges in distribution to the standard Normal distribution. This result is widely used for inference when $n\lambda$ is large, as the Poisson distribution then becomes well-approximated by the Normal distribution.

The CLT states:

$$\sqrt{n}\frac{(\bar{X}_n - \lambda)}{\sqrt{\lambda}} \xrightarrow{D} \mathcal{N}(0,1) \quad \text{as } n \to \infty.$$

Example 2: Normal (Gaussian) Distribution

Let $X_1, \ldots, X_n$ be independent and identically distributed (iid) random variables, $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

Unlike other distributions, the Normal distribution is reproductive: the sum (or average) of independent Normal variables is **exactly Normal**. Therefore, the Central Limit Theorem (CLT) is not technically needed to describe the distribution of the mean.

The sample mean ($\bar{X}_n$) has an **exact** distribution for any sample size n:

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Consequently, the standardized sample mean is **exactly** the standard Normal distribution for **all** n, making the convergence trivial:

$$\sqrt{n}\frac{(\bar{X}_n - \mu)}{\sigma} \sim \mathcal{N}(0,1).$$

In this case, the distribution **does not converge** to $\mathcal{N}(0,1)$; it is already $\mathcal{N}(0,1)$ regardless of the sample size n.

Some convergence properties are preserved under transformations.

Theorem 6 (Preservation Properties)**.** *Let X_n, X, Y_n, Y be random variables. Let g be a continuous function.*

(a) If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n + Y_n \xrightarrow{P} X + Y$.

(b) If $X_n \xrightarrow{qm} X$ and $Y_n \xrightarrow{qm} Y$, then $X_n + Y_n \xrightarrow{qm} X + Y$.

(c) If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n Y_n \xrightarrow{P} XY$.

(d) If $X_n \xrightarrow{P} X$, *then* $g(X_n) \xrightarrow{P} g(X)$.

(e) If $X_n \xrightarrow{D} X$, *then* $g(X_n) \xrightarrow{D} g(X)$.

Theorem 7. *Slutsky's Theorem Suppose that* $T_n \xrightarrow{D} T$ *and* $S_n \xrightarrow{P} s$. *Then*

(a) $T_n + S_n \xrightarrow{D} T + s$.

(b) $T_n S_n \xrightarrow{D} sT$.

Example 3 Let $X_1, \ldots, X_n \sim$ iidPoisson(λ). The Central Limit Theorem tells us that

$$\sqrt{n}\frac{(\bar{X}_n - \lambda)}{\sqrt{\lambda}} \xrightarrow{D} \mathcal{N}(0,1) \quad \text{as } n \to \infty.$$

We must often use the estimator $\sqrt{\bar{X}_n}$ instead of the unknown true value $\sqrt{\lambda}$. By the Law of Large Numbers, $\bar{X}_n \xrightarrow{P} \lambda$, so by the Continuous Mapping Theorem, $\sqrt{\bar{X}_n} \xrightarrow{P} \sqrt{\lambda}$.

By Slutsky's Theorem, we can substitute the term $\left(\frac{\sqrt{\lambda}}{\sqrt{\bar{X}_n}}\right)$, which converges in probability to $\frac{\sqrt{\lambda}}{\sqrt{\lambda}} = 1$.

The standardized statistic using the sample variance estimator is derived as follows:

$$\begin{aligned}
\sqrt{n}\frac{(\bar{X}_n - \lambda)}{\sqrt{\bar{X}_n}} &= \left(\frac{\sqrt{\lambda}}{\sqrt{\bar{X}_n}}\right) \cdot \left(\sqrt{n}\frac{(\bar{X}_n - \lambda)}{\sqrt{\lambda}}\right) \\
&= \left(\frac{1}{\sqrt{\bar{X}_n}/\sqrt{\lambda}}\right) \cdot Z_n \\
&\xrightarrow{D} \left(\frac{1}{\sqrt{\lambda}/\sqrt{\lambda}}\right) \cdot \mathcal{N}(0,1) \\
&= \mathcal{N}(0,1).
\end{aligned}$$

This result is crucial as it allows us to construct asymptotic confidence intervals and hypothesis tests without knowing the true value of λ.

Theorem 8. *Delta Method Suppose that* $\sqrt{n}(T_n - t) \xrightarrow{D} \mathcal{N}(0, v)$. *If* $g(x)$ *is a function with derivative* $g'(t)$ *at* $x = t$, *then*

$$\sqrt{n}(g(T_n) - g(t)) \xrightarrow{D} g'(t)\mathcal{N}(0, v) = \mathcal{N}(0, [g'(t)]^2 v).$$

Example 4

Let $X_1, \ldots, X_n$ be independent and identically distributed (iid) random variables, $X_i \sim$ Poisson(λ). We aim to find the asymptotic distribution of the square root transformation, $\sqrt{\bar{X}_n}$. We define the function $g(t) = \sqrt{t}$.

- The function is $g(t) = t^{1/2}$.
- The derivative is $g'(t) = \frac{1}{2}t^{-1/2} = \frac{1}{2\sqrt{t}}$.

Applying the Delta Method to $g(\bar{X}_n)$ yields the following asymptotic distribution:

$$\sqrt{n}(g(\bar{X}_n) - g(\lambda)) \xrightarrow{D} \mathcal{N}\left(0, [g'(\lambda)]^2\lambda\right).$$

Substituting the function and derivative into the expression, we get:

$$\begin{aligned}\sqrt{n}(\sqrt{\bar{X}_n} - \sqrt{\lambda}) &\xrightarrow{D} \mathcal{N}\left(0, \frac{1}{4\lambda} \cdot \lambda\right) \\ &= \mathcal{N}\left(0, \frac{1}{4}\right).\end{aligned}$$

The square root transformation successfully stabilizes the variance of the sample mean estimator to a fixed value of $1/4$, which is independent of the true parameter λ. This is highly beneficial for statistical inference.

Theorem 9 (Sampling Distribution of Mean and Variance)**.** *Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. The sample mean $\bar{X}$ and the sample variance S^2 have the following sampling distributions:*

1. $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.
2. *The sample mean $\bar{X}$ is independent of the sample variance S^2.*
3. *The standardized sample variance follows a Chi-squared distribution:*

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

4. *The standardized sample mean using the sample standard deviation (S) follows Student's t-distribution:*

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Here t_{n-1} denotes Student's distribution with $n-1$ degrees of freedom.

The proof of Theorem 9 is omitted, as its technical nature does not contribute new skills required for the remainder of this course. You are only required to know and apply the three results listed above.

Chapter 2

Point Estimation

We are studying a random process that we believe can be described by a specific probability model. This model is part of a family of distributions, each defined by an unknown parameter, θ. Our data consists of n independent and identical observations, $X_1, \ldots, X_n$, which were generated by a specific, true value of the parameter, θ_0. Our goal is to use this data to learn about this unknown θ_0.

The most direct question we can ask is: what is the single best guess for the true parameter θ_0?

This task is called **point estimation**. Because our only information comes from the data, we must use a function of our sample to create this guess. A point estimator is any function of the observed data that provides a single numerical value as a guess for the unknown parameter θ. In other words, it's a rule that takes our sample $(X_1, \ldots, X_n)$ and maps it to a point within the parameter space Θ.

Definition 8 (Point Estimator)**.** *Suppose that the observable random variables of interest are $X_1, \ldots, X_n$. We define a statistic $T_n = T(\mathbf{X})$ to be a function of $\mathbf{X} = (X_1, \ldots, X_n)$ that does not depend on unknown parameters. An **point estimator** of $\theta_0 \in \Theta$ is a statistic whose primary goal is to estimate θ_0. If $\{X_1 = x_1, \ldots, X_n = x_n\}$ are observed, then $T(x_1, \ldots, x_n)$ is called an estimate of θ_0.*

Remark 5. *We commonly use the notation $\hat{\theta}_n$ to represent a point estimator. It's crucial to remember the difference between the true parameter θ, which is a fixed, unknown value, and the estimator $\hat{\theta}$, which is a **random variable**. This is because $\hat{\theta}_n$ is calculated from our random sample, so its value will change every time we collect a new sample.*

Example 8. *Let $X_1, \ldots, X_n \sim$ Bernoulli(p) and let $\hat{p}_n = T(X_1, \ldots X_n) = n^{-1} \sum_i X_i$.*

The definition of an estimator is broad, so how do we choose a good one from all the possibilities? And how can we evaluate an estimator's performance? The key insight is that because an estimator is a random variable, its value will vary from one sample to the next. A high-quality estimator is one whose distribution is tightly clustered around the true parameter θ. This means that most of the time, our estimate will be "close" to the actual value we are trying to find.

2.1 Performance Metrics for Estimators

This section introduces several key criteria that allow us to evaluate and compare different estimators, helping us choose the best one for a given problem.

After defining what an estimator is, the next question is how to determine if it's a good one. To do this, we need a way to measure how "concentrated" its values are around the true parameter θ. While many measures exist, statisticians primarily focus on two key properties of an estimator's distribution: its mean and its variance.

Why these two? First, they are easy to understand.

The **mean** of an estimator tells us its average value, indicating if it's on target. The **variance** tells us how much its values typically spread out from that average. An estimator with a small mean and small variance is generally a good one.

Second, the exact distribution of an estimator is often unknown. In these cases, we rely on approximations. Asymptotic theory often shows that the estimator's distribution becomes normal, and for a normal distribution, the mean and variance are all we need to know about its spread.

Even when the distribution isn't normal, powerful tools like Markov's and Chebyshev's inequalities can be used to set bounds on how far the estimator's value might be from the true parameter, just by knowing its mean and variance.

Definition 9 (Unbiasedness)**.** *Let $\hat{\theta}_n$ be an estimator for a parameter θ_0 of a parametric model $\{F_\theta : \theta \in \Theta\}$. $\hat{\theta}_n$ is called **unbiased** if its expected value is equal to the true parameter it is meant to estimate. Formally, this is expressed as $E[\hat{\theta}_n] = \theta_0$.*

Conceptually, an unbiased estimator is one that gets the correct answer on average. If we were to take many different random samples and calculate our estimate each time, the average of all these estimates would converge to the true value of the parameter. This property is highly desirable as it indicates the estimator does not suffer from a systematic error in one direction.

Definition 10 (Bias)**.** *The **bias** is the difference between the estimator's average value and the true parameter, $bias(\hat{\theta}_n, \theta_0) = E[\hat{\theta}_n] - \theta_0$.*

It represents a systematic deviation from the truth.

Example 9. *Let $X_1, \ldots, X_n \sim Bernoulli(p)$ and let $\hat{p}_n = n^{-1}\sum_i X_i$. Then*

$$E(\hat{p}_n) = n^{-1}\sum_i E(X_i) = p$$

so $\hat{p}_n$ is unbiased.

The **MSE** provides a single, comprehensive measure of an estimator's overall accuracy. It quantifies the average squared difference between the estimator and the true parameter.

Definition 11 (Mean Squared Error (MSE))**.** *Let $\hat{\theta}_n$ be an estimator for a parameter θ_0 of a parametric model $\{F_\theta : \theta \in \Theta\}$. The mean squared error of $\hat{\theta}$ is defined to be*

$$MSE(\hat{\theta}, \theta_0) = E[\|\hat{\theta} - \theta_0\|^2].$$

Example 10. *Let $X_1, \ldots, X_n \sim Bernoulli(p)$ and let $\hat{p}_n = n^{-1}\sum_i X_i$. Then we know that $\hat{p}_n$ is unbiased. Therefore*

$$MSE(\hat{p}_n, p) = V(\hat{p}_n) = \frac{p(1-p)}{n}.$$

A central result in statistical theory is the **Bias-Variance Decomposition**, which shows that the MSE can be broken down into two components: the estimator's bias and its variance.

Theorem 10 (Bias-Variance Decomposition)**.** *Let $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)^T$. The mean squared error of an estimator admits the decomposition*

$$MSE(\hat{\theta}, \theta_0) = \|E[\hat{\theta}] - \theta_0\|^2 + E[\|\hat{\theta} - E[\hat{\theta}]\|^2] = \|bias(\hat{\theta}, \theta_0)\|^2 + \sum_{k=1}^{p} Var[\hat{\theta}_k].$$

Proof. We expand the MSE after adding and subtracting $E[\hat{\theta}]$:

$$\begin{aligned}
E[\|\hat{\theta} - \theta\|^2] &= E[\|\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta\|^2] \\
&= E[(\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta)^T(\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta)] \\
&= \|E[\hat{\theta}] - \theta\|^2 + E[\|\hat{\theta} - E[\hat{\theta}]\|^2] + 2E[(\hat{\theta} - E[\hat{\theta}])^T(E[\hat{\theta}] - \theta)] \\
&= \|E[\hat{\theta}] - \theta\|^2 + E[\|\hat{\theta} - E[\hat{\theta}]\|^2] + 2(E[\hat{\theta}] - E[\hat{\theta}])^T(E[\hat{\theta}] - \theta) \\
&= \|E[\hat{\theta}] - \theta\|^2 + E[\|\hat{\theta} - E[\hat{\theta}]\|^2] + 0 \\
&= \|E[\hat{\theta}] - \theta\|^2 + \sum_{k=1}^{p} E[(\hat{\theta}_k - E[\hat{\theta}_k])^2]
\end{aligned}$$

by linearity of the expectation and since $(E[\hat{\theta}] - \theta)$ is deterministic.

Remark 6. *The bias of the estimator $\hat{\theta}$ at true parameter θ expresses how far off $\hat{\theta}$ is from θ on average. When the bias at some coordinate of θ is positive we have **overestimation**; when it is negative we have **underestimation**; when the bias is zero, we speak of an **unbiased estimator**. Notice that the variances $Var[\hat{\theta}_k]$ can also depend on θ, even though this is not explicitly reflected in the notation.*

For a vector-valued parameter, the MSE is given by

$$MSE(\hat{\theta}_n, \theta_0) = \|bias(\hat{\theta}_n, \theta_0)\|^2 + tr[Cov(\hat{\theta}_n)].$$

*The **covariance** term, $Cov(\hat{\theta}_n)$, measures the random spread of the estimator's values around its own mean. The decomposition reveals that an estimator's total error is a sum of its squared bias and its variance. This forces a crucial trade-off: reducing one can sometimes increase the other, making the MSE a powerful tool for finding the optimal balance.*

The mean squared error is just one method for evaluating an estimator's accuracy, but the concept is much broader. You can define any loss function you want to measure the cost of an estimation error. The estimator's quality is then judged by its average cost, which we call risk. Since the loss function you choose directly defines what you consider a good or bad estimate, picking the right one is crucial. The mean squared error is simply one example of a risk function that uses the squared difference as its penalty.

When comparing two different estimators for the same parameter, we can use their **relative efficiency** to determine which one is superior. This is defined as the ratio of their Mean Squared Errors.

Definition 12 (Relative Efficiency)**.** *Given two estimators, $\hat{\theta}_n$ and $\tilde{\theta}_n$, the relative efficiency is calculated as*

$$\mathcal{E}(\hat{\theta}_n, \tilde{\theta}_n) = \frac{MSE(\hat{\theta}_n, \theta_0)}{MSE(\tilde{\theta}_n, \theta_0)}.$$

An efficiency value less than one, $\mathcal{E} < 1$, indicates that the first estimator, $\hat{\theta}_n$, is more efficient than the second, $\tilde{\theta}_n$, as it achieves a smaller MSE. This provides a clear, quantitative way to rank and select the best estimator from a set of candidates.

Definition 13 (Consistency)**.** *Let $\hat{\theta}_n$ be an estimator for a parameter θ_0 of a parametric model $\{F_\theta : \theta \in \Theta\}$. $\hat{\theta}_n$ is said to be **(weakly) consistent** for the parameter θ_0 if it converges in probability to the true value as the sample size n approaches infinity. This is written as $\hat{\theta}_n \xrightarrow{P} \theta_0$ as $n \to \infty$. That is for every $\epsilon > 0$ we have that $P(||\hat{\theta}_n - \theta_0|| > \epsilon) \to 0$*

This means that as we collect more and more data, the probability that our estimator is far from the true parameter becomes vanishingly small. Consistency is a crucial property for any estimator, as it guarantees that given a large enough dataset, we can get arbitrarily close to the truth. While unbiasedness tells us about the estimator's average performance, **consistency** speaks to its behavior as our sample size grows.

The concentration of an estimator $\hat{\theta}_n$ around the true parameter θ can always be bounded using the mean squared error (provided that the estimator $\hat{\theta}_n$ has finite variance). The concentration of an estimator $\hat{\theta}_n$ around the true parameter θ can always be bounded using the mean squared error (provided that the estimator $\hat{\theta}_n$ has finite variance). This fact relate the concept of consistent estimator with MSE.

Theorem 11. *Let $\hat{\theta}$ be an estimator of $\theta_0 \in \mathbb{R}^p$ such that $Var[\hat{\theta}] < \infty$. Then, for all $\epsilon > 0$,*

$$P[\|\hat{\theta} - \theta_0\| > \epsilon] \leq \frac{MSE(\hat{\theta}, \theta_0)}{\epsilon^2} = \frac{\|E[\hat{\theta}] - \theta_0\|^2 + \sum_{k=1}^{p} Var[\hat{\theta}_k]}{\epsilon^2}.$$

Proof. Let $X = \|\hat{\theta} - \theta\|^2$. Since $\epsilon > 0$, Markov's inequality yields

$$P[\|\hat{\theta} - \theta\| > \epsilon] = P[\|\hat{\theta} - \theta\|^2 > \epsilon^2] \leq \frac{E[\|\hat{\theta} - \theta\|^2]}{\epsilon^2} = \frac{\text{MSE}(\hat{\theta}, \theta)}{\epsilon^2}.$$

Notice that convergence of the MSE to zero implies consistency. The converse is not true in general, though.

There are a few general ways to construct estimators based on an observed random sample. In the following we will discuss some of them..

2.2 Method of Moments

Let's begin by considering a simple case: a model with a single, unknown parameter, θ. The Method of Moments is a straightforward and intuitive way to estimate this parameter. Its core idea stems from a foundational concept in statistics: the Law of Large Numbers. This law guarantees that the average of our observed data, $\frac{1}{n}\sum_{i=1}^{n} X_i$, will get closer and closer to the true, theoretical average of the population, $E[X_1]$, as we collect more data.

The crucial insight is that this theoretical average, $E[X_1]$, is itself a function of our unknown parameter θ. Let's denote this function as $m(\theta)$. The Law of Large Numbers now tells us that our sample average should be a good approximation of this theoretical value:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \approx m(\theta)$$

for a sufficiently large sample size n. This simple relationship gives us a powerful idea for an estimator: we should find a value for our estimator, $\hat{\theta}$, that makes this equation hold true.

This leads directly to the formal definition of the Method of Moments estimator.

Definition 14 (Method of Moments Estimator: Single Parameter Case)**.** *The* ***Method of Moments (MoM)*** *estimator, $\hat{\theta}_n$, is found by solving the following equation:*

$$\frac{1}{n}\sum_{i=1}^{n} X_i = m(\hat{\theta}_n)$$

where $m(\theta) = E_\theta[X_1]$.

In other words, we set the first empirical moment (the sample average) equal to the first theoretical moment and solve for the parameter.

Let's illustrate this technique with simple examples.

Example 11. *Let $X_1, \ldots, X_n$ be an iid random sample with distribution Ber(p). The first moment of a Bernoulli distribution is $E[X_1] = p$, and the empirical moment is $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Therefore, we use the empirical moment to estimate the population moment:*

$$\hat{p}_n^{MoM} = \bar{X}.$$

Example 12. *Let $X_1, \ldots, X_n$ be i.i.d. Exponential random variables with density $f(x) = \lambda e^{-\lambda x}$. The first moment of X is*

$$E[X_1] = \frac{1}{\lambda} = m(\lambda),$$

and the empirical first moment is $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. The resulting MM estimator of λ is

$$\hat{\lambda}_n^{MoM} = m^{-1}(\bar{X}) = \frac{1}{\bar{X}}.$$

Example 13. *Let $X_1, \ldots, X_n$ be i.i.d. with uniform distribution $U(0, \theta)$, the first moment is*
$E(X_1) = \frac{\theta}{2}$. *We equate this theoretical average to our sample average, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$:*

$$\bar{X} = \frac{\hat{\theta}_n}{2}$$

Solving this simple equation for our estimator, $\hat{\theta}_n$, we get a very straightforward result: $\hat{\theta}_n = 2\bar{X}$, which is simply twice the sample average.

Extending to Multiple Parameters

The Method of Moments can be easily extended to problems with multiple parameters. If our model has p unknown parameters, say $\theta_1, \ldots, \theta_p$, the MoM procedure instructs us to match the first p empirical moments of our sample to the first p theoretical moments of the distribution. This process yields a system of p equations with p unknowns. By solving this system, we obtain the Method of Moments estimator for all the parameters.

Definition 15 (Method of Moments Estimator: Multiparameter Case)**.** *For a model with p parameters, the MoM estimator, $\hat{\theta}_n$, is the solution to the system of equations:*

$$\hat{m}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k = m_k(\hat{\theta}_n), \quad \text{for } k = 1, \ldots, p$$

where $m_k(\theta)$ is the k-th theoretical moment, defined as $m_k(\theta) = E_\theta[X^k]$.

The most notable advantage of the Method of Moments is its simplicity. The estimation problem is transformed from a complex search into a simple equation-solving task. The MoM equation is often easier to solve because the data are grouped together on one side, and the parameter is isolated within a function on the other. This allows for a direct solution for the estimator, bypassing a more complicated optimization problem.

Definition 16. *Assuming that the function $\psi(\theta) = (m_1(\theta), \ldots, m_d(\theta))$ is bijective we have that $\theta = \psi^{-1}(m_1(\theta), \ldots, m_d(\theta))$. The method of moments estimator of θ_0 is*

$$\hat{\theta}_n^{MM} = \psi^{-1}(\hat{m}_1, \ldots, \hat{m}_d)$$

provided it exists.

Example 14. *Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d Gamma(α, β) with density function*

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

where $\alpha, \beta > 0$ and $\Gamma(\cdot)$ is the gamma function. In this case we want to estimate the two-dimensional parameter $\theta = (\alpha, \beta)$. The first two moments of this distribution are:

$$E[X_1] = \frac{\alpha}{\beta}; \quad E[X_1^2] = \frac{\alpha(\alpha+1)}{\beta^2};$$

which implies that

$$\alpha = \frac{E[X_1]^2}{E[X_1^2] - E[X_1]^2}; \quad \beta = \frac{E[X_1]}{E[X_1^2] - E[X_1]2};$$

The MOM says that we replace the right-hand sides of these equations by the sample moments and then solve for α and β. In this case, we get

$$\hat{\alpha} = \frac{(\bar{X})^2}{\bar{X^2} - (\bar{X})^2}; \quad \hat{\beta} = \frac{\bar{X}}{\bar{X^2} - (\bar{X})^2}.$$

2.3 Maximum Likelihood Estimation

First, we define the likelihood function. The likelihood function is a central concept in statistics, used to measure how well a statistical model "fits" or explains a set of observed data. It answers the question: "How probable are our observed data, given a specific value for the unknown parameter?"

The easiest way to think about likelihood is in the discrete case. Suposse that you have a random sample of independent observations, $x_1, x_2, \ldots, x_n$, from a population whose probability mass function depends on an unknown parameter, θ. The likelihood

function, denoted as $L(\theta)$, is defined as the joint probability of observing that specific sample:

$$L(\theta; x_1, \ldots, x_n) = P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n).$$

Since the observations are independent, this can be written as the product of their individual probabilities:

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} P_\theta(X_i = x_i).$$

It's a function of the parameter(s) of a model, not the data. This is a crucial distinction. While it's built using the data, the likelihood function's output is a value that changes as you vary the model's parameters. A higher likelihood value for one parameter suggests that the observed data were more probable under that parameter's assumption than under another's.

The goal of **maximum likelihood estimation (MLE)** is to find the value of θ that maximizes this function, as it represents the parameter that makes the observed data most probable.

Definition 17. *Let $X_1, \ldots, X_n$ be an i.i.d. sample of random variables with density or frequency function $f(x; \theta_0)$ and assume that $\Theta \subseteq \mathbb{R}^d$, for some $d \geq 1$. The* ***likelihood function*** *is*

$$L(\theta) = L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta)$$

and the ***maximum likelihood estimator*** *of θ_0 is*

$$\hat{\theta}_n^{MLE} = \arg\max_{\theta \in \Theta} L(\theta; X_1, \ldots, X_n)$$

When the likelihood function has a single, highest point, the parameter value at that point is called the **Maximum Likelihood Estimator (MLE)**. We can find this value using differential calculus. The first step is to find the parameter value $\hat{\theta}$ where the derivative (or gradient, for multiple parameters) of the likelihood function is zero. This gives us a candidate for the MLE:

$$\nabla_\theta L(\theta) = 0$$

However, a derivative of zero doesn't guarantee a maximum; it could be a minimum. To confirm we've found a maximum, we must check the second derivative. For a single parameter, the second derivative must be negative at our candidate value. For multiple parameters, this requires a more complex check on the Hessian matrix to ensure it's negative definite, i.e.,

$$-\nabla^2 L(\theta)\big|_{\theta=\hat{\theta}} > 0$$

Solving for the derivative of the likelihood function can be very difficult because the function is often a product of many terms, as shown by $L(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$. To simplify this, we use a clever and common trick: we maximize the **log-likelihood** $\ell(\theta) = \log L(\theta)$ instead. This works because the logarithm is a monotonic function, meaning it has the same maximum points as the original function. The major advantage is that the logarithm of a product becomes a sum of logarithms, which is much easier to differentiate:

$$\ell(\theta; X_1, \ldots, X_n) = \log\left(\prod_{i=1}^{n} f(X_i; \theta)\right) = \sum_{i=1}^{n} \log f(X_i; \theta).$$

Then **maximum likelihood estimator** of θ_0 is

$$\hat{\theta}_n^{\text{MLE}} = \arg\max_{\theta\in\Theta} L(\theta; X_1, \ldots, X_n) = \arg\max_{\theta\in\Theta} \ell(\theta; X_1, \ldots, X_n).$$

Therefore, the standard procedure is to find the parameter value that makes the derivative of the log-likelihood function equal to zero and then verify with the second derivative that you have indeed found a maximum. An MLE $\hat{\theta}$ will satisfy:

$$\nabla_\theta \ell(\theta)\big|_{\theta=\hat{\theta}} = 0 \quad \text{and} \quad -\nabla^2 \ell(\theta)\big|_{\theta=\hat{\theta}} > 0.$$

Example 15. *Let $X_1, \ldots, X_n \sim Ber(p)$ are iid Bernoulli random variables. The joint density function is*

$$f(x_1, \ldots, x_n; p) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

and the log-likelihood is

$$\ell(p; X_1, \ldots, X_n) = \log(L(p; X_1, \ldots, X_n)) = (\sum_{i=1}^{n} X_i)\log p + (\sum_{i=1}^{n}(1-X_i))\log(1-p).$$

To get the argmax, we take the derivative with respect to p and set it to zero:

$$\frac{\partial \ell(p; X_1, \ldots, X_n)}{\partial p} = \frac{\sum_{i=1}^{n} X_i}{p} - \frac{\sum_{i=1}^{n}(1-X_i)}{1-p} = 0$$

Solving for p, we get the MLE:

$$\hat{p}^{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}.$$

Also,

$$\frac{\partial^2 \ell(p; X_1, \ldots, X_n)}{\partial p^2} = -\frac{\sum_{i=1}^{n} X_i}{p^2} - \frac{\sum_{i=1}^{n}(1-X_i)}{(1-p)^2} = -\left(\frac{n\bar{X}_n}{p^2} + \frac{n(1-\bar{X}_n)}{(1-p)^2}\right) < 0$$

Example 16. *If $X_1, \ldots, X_n \sim \exp(\lambda)$ are i.i.d. Exponential random variables with mean $1/\lambda$, the likelihood function is*

$$L(\lambda; X_1, \ldots, X_n) = \prod_{i=1}^{n} f(X_i; \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} X_i}.$$

The log-likelihood is

$$\ell(\lambda; X_1, \ldots, X_n) = n\log(\lambda) - \lambda \sum_{i=1}^{n} X_i.$$

Taking the derivative with respect to λ and setting it to zero:

$$\frac{\partial \ell(\lambda; X_1, \ldots, X_n)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i = 0.$$

Solving for λ, we get the MLE:

$$\hat{\lambda}^{MLE} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}}.$$

And

$$\frac{\partial^2 \ell(\lambda; X_1, \ldots, X_n)}{\partial \lambda^2} = -\frac{n}{\lambda^2} < 0.$$

Example 17. *Suppose that $X_1, \ldots, X_n \sim \Gamma(\alpha, 1)$ are i.i.d from a Gamma distribution for which the p.d.f is as follows:*

$$f(x;\alpha) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}, \quad \text{for } x > 0.$$

The likelihood function is

$$L(\alpha; X_1, \ldots, X_n) = \frac{1}{\Gamma(\alpha)^n} \left(\prod_{i=1}^{n} X_i \right)^{\alpha-1} e^{-\sum_{i=1}^{n} X_i}.$$

and thus the log-likelihood is

$$\log L(\alpha; X_1, \ldots, X_n) = -n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \log(X_i) - \sum_{i=1}^{n} X_i.$$

The MLE of α will be the value of α that satisfies the equation

$$\frac{\partial}{\partial \alpha} \log L(\alpha) = -n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n} \log(X_i) = 0.$$

i.e.,

$$\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = \frac{1}{n} \sum_{i=1}^{n} \log(X_i).$$

In this case we do not have an analytical solution for the estimator. Instead, we would have to rely on numerical methods (e.g. Newton's method) in order to compute $\hat{\alpha}^{MLE}$.

Example 18. *Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\eta, \sigma^2)$. The likelihood function is the product of the individual probability density functions:*

$$L(\eta, \sigma^2) = \prod_{i=1}^{n} f_{\eta,\sigma^2}(X_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(X_i - \eta)^2}{2\sigma^2} \right\}$$

$$L(\eta, \sigma^2) = \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \eta)^2 \right\}$$

Taking logarithms on both sides, we obtain the log-likelihood function:

$$\ell(\eta, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \eta)^2$$

We calculate the first derivatives with respect to η and σ^2:

$$\frac{\partial \ell(\eta, \sigma^2)}{\partial \eta} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \eta)$$

$$\frac{\partial \ell(\eta, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (X_i - \eta)^2$$

Solving for $\nabla_{(\eta,\sigma^2)}\ell(\eta,\sigma^2)=0$ with respect to (η,σ^2) yields a system of two equations in two unknowns. The unique root of this system can be seen to be $\bar{X}$ and $\frac{1}{n}\sum_{i=1}^n(X_i-\bar{X})^2$. Call this $(\hat{\eta},\hat{\sigma}^2)$. It is our candidate for an MLE, provided that it yields a maximum.

We now calculate the second derivatives:

$$\frac{\partial^2\ell(\eta,\sigma^2)}{\partial\eta^2}=-\frac{n}{\sigma^2}$$

$$\frac{\partial^2\ell(\eta,\sigma^2)}{\partial(\sigma^2)^2}=\frac{n}{2(\sigma^2)^2}-\frac{1}{(\sigma^2)^3}\sum_{i=1}^n(X_i-\eta)^2$$

$$\frac{\partial^2\ell(\eta,\sigma^2)}{\partial\eta\partial\sigma^2}=\frac{\partial^2\ell(\eta,\sigma^2)}{\partial\sigma^2\partial\eta}=-\frac{1}{(\sigma^2)^2}\sum_{i=1}^n(X_i-\eta)$$

Evaluating these second derivatives at $(\hat{\eta},\hat{\sigma}^2)$ yields:

$$\left.\frac{\partial^2\ell}{\partial\eta^2}\right|_{(\hat{\eta},\hat{\sigma}^2)}=-\frac{n}{\hat{\sigma}^2}$$

$$\left.\frac{\partial^2\ell}{\partial(\sigma^2)^2}\right|_{(\hat{\eta},\hat{\sigma}^2)}=\frac{n}{2\hat{\sigma}^4}-\frac{1}{\hat{\sigma}^6}\sum_{i=1}^n(X_i-\bar{X})^2=\frac{n}{2\hat{\sigma}^4}-\frac{n\hat{\sigma}^2}{\hat{\sigma}^6}=\frac{n}{2\hat{\sigma}^4}-\frac{n}{\hat{\sigma}^4}=-\frac{n}{2\hat{\sigma}^4}$$

$$\left.\frac{\partial^2\ell}{\partial\eta\partial\sigma^2}\right|_{(\hat{\eta},\hat{\sigma}^2)}=-\frac{1}{(\hat{\sigma}^2)^2}\sum_{i=1}^n(X_i-\bar{X})=0$$

We conclude that the Hessian matrix $-\nabla^2\ell(\eta,\sigma^2)|_{(\hat{\eta},\hat{\sigma}^2)}$ is diagonal. To show that it is positive definite, it suffices to show that its two diagonal elements are positive, which is true since $\hat{\sigma}^2$ is positive with probability one. Therefore, the unique MLE of (η,σ^2) is:

$$(\hat{\eta},\hat{\sigma}^2)=\left(\bar{X},\frac{1}{n}\sum_{i=1}^n(X_i-\bar{X})^2\right).$$

There are times when we don't want to estimate the parameter θ directly, but instead want to estimate a different value, γ, which is a function of θ. If this function is a one-to-one mapping (a bijection), we don't have to go through the entire estimation process again. This is because the maximum of a function remains the maximum even when you relabel the points.

The invariance property tells us that the MLE of a known function of the unknown parameter can be found by plugging-in the MLE of the unknown parameter.

Proposition 1 (Invariance Property)**.** *Let $\{f(x;\theta):\theta\in\Theta\}$ be a parametric model, where $\Theta\subseteq\mathbb{R}^p$. Suppose that $\hat{\theta}$ is an MLE of θ, on the based on a random sample $X_1,\ldots,X_n$ from $f(x;\theta)$. Let $g:\Theta\to\Gamma\subseteq\mathbb{R}^p$ be a bijection. Then, $\hat{\gamma}=g(\hat{\theta})$ is an MLE of $\gamma=g(\theta)$.*

Proof. The core of the proof is to show that the new estimator, $\hat{\gamma}$, maximizes the likelihood for the new parameterization. Define $h(x;\gamma)=f(x;g^{-1}(\gamma))$, and note that h is a well-defined function, because $g^{-1}:\Gamma\to\Theta$ is well-defined. The function $h(x;\gamma)$ is simply the density/frequency of X_i under the reparametrisation given by $\gamma\in\Gamma$.

We can define a new likelihood function for γ and an MLE of γ, say $\hat{\gamma}$, must satisfy:

$$\prod_{i=1}^{n} h(X_i;\hat{\gamma}) \geq \prod_{i=1}^{n} h(X_i;\gamma), \quad \forall \gamma \in \Gamma.$$

Let $\hat{\theta}$ be an MLE of θ, and let $\hat{\gamma} = g(\hat{\theta})$. Let $\gamma \in \Gamma$ be arbitrary and observe that:

$$\prod_{i=1}^{n} h(X_i;\gamma) = \prod_{i=1}^{n} f(X_i; g^{-1}(\gamma)) \leq \prod_{i=1}^{n} f(X_i;\hat{\theta}) = \prod_{i=1}^{n} f(X_i; g^{-1}(\hat{\gamma})) = \prod_{i=1}^{n} h(X_i;\hat{\gamma})$$

which proves the proposition.

Example 19. *Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\eta, 1)$. Suppose we are interested in estimating the probability that a new observation is less than or equal to a specific value x, which we can write as $P[X \leq x]$.*

$$P[X_1 \leq x] = P[X_1 - \eta \leq x - \eta] = \Phi(x - \eta),$$

where Φ is the standard normal CDF. But the mapping $\eta \mapsto \Phi(x-\eta)$ is a bijection because Φ is monotone. The MLE for the probability will simply be the probability calculated using our MLE for the mean: $\hat{P}[X_1 \leq x] = \Phi(x-\hat{\eta})$, where Φ is the standard normal cumulative distribution function and $\hat{\eta} = \bar{X}$.

There are some cases where finding the maximum likelihood estimator (MLE) through differential calculus is not a viable option. This can happen, for instance, when the set of possible parameter values is discrete, or when the range of the data itself depends on the parameter. When dealing with a single, one-dimensional parameter, the MLE can sometimes be found simply by visual inspection of the likelihood function.

Example 20. *Let $X_1, \ldots, X_n \overset{iid}{\sim} U(0, \theta)$. The likelihood function for n independent and identically distributed random variables from a uniform distribution on $[0, \theta]$ can be written as:*

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\theta} \cdot \mathbf{1}_{\{0 \leq X_i \leq \theta\}} = \theta^{-n} \prod_{i=1}^{n} I_{\{0 \leq X_i \leq \theta\}},$$

where the $I_{\{0 \leq X_i \leq \theta\}}$ term is the indicator function. It equals 1 if the condition inside the brackets is true, and 0 otherwise.

However, this is only valid if all the data points fall within the range of the distribution. This means that every single observation must be less than or equal to θ, and since the observations are also greater than 0, the largest observation, $X_{(n)} = \max\{X_1, \ldots, X_n\}$, must also be less than or equal to θ.

Therefore, the likelihood function is:

$$L(\theta) = \begin{cases} \theta^{-n} & \text{if } \theta \geq X_{(n)} \text{ and } X_{(1)} > 0 \\ 0 & \text{if } \theta < X_{(n)} \end{cases}$$

where $X_{(1)} = \min\{X_1, \ldots, X_n\}$. Assuming the data is positive, the condition simplifies to:

$$L(\theta) = \theta^{-n} \cdot \mathbf{1}_{\{\theta \geq X_{(n)}\}}.$$

By inspecting this function, we can see that if θ is less than the maximum observed value, the likelihood is zero. As θ increases from $X_{(n)}$ to infinity, the likelihood function θ^{-n} continuously decreases. To maximize the likelihood, we must choose the smallest possible value for θ that is still valid. This value is precisely the largest observed data point.

Thus, the maximum likelihood estimator is $\hat{\theta} = X_{(n)}$.

Maximum Likelihood in Exponential Families

Excluding the uniform distribution, every probability model we have examined so far is a member of the exponential family. This naturally leads to the question of whether general properties of the maximum likelihood method can be derived for all models within this family.

The existence and uniqueness of the MLE in our previous examples were not coincidental. This is a characteristic feature of exponential family models. For clarity, we will focus on the single-parameter case.

Proposition 2 (One-Parameter Exponential Family MLE)**.** *Consider an independent and identically distributed sample $X_1, \ldots, X_n$ from a single-parameter exponential family,*

$$f(x, \eta) = \exp\{\eta T(x) - A(\eta) + S(x)\}, \quad x \in \mathcal{X}, \quad \eta \in \mathcal{H}$$

with a parameter space $\mathcal{H} \subseteq \mathbb{R}$ that is an open set and T a non-constant function. If the MLE $\hat{\eta}$ exists, it is guaranteed to be unique. It can be found as the one-of-a-kind solution to the equation

$$A'(\hat{\eta}) = \bar{T}$$

where $\bar{T} = \frac{1}{n}\sum_{i=1}^{n} T(X_i)$.

Proof. To prove this, we first establish the log-likelihood function for our sample. By taking the logarithm of the likelihood, we get a simplified expression:

$$\ell(\eta) = \log L(\eta) = -nA(\eta) + n\bar{T} + \sum_{i=1}^{n} S(X_i)$$

Setting the first derivative to zero, $\ell'(\eta) = -nA'(\eta) + n\bar{T} = 0$, reveals the maximum must satisfy $A'(\hat{\eta}) = \bar{T}$.

The key to the proof is the second derivative of the log-likelihood function. We can show that this second derivative is always negative:

$$\ell''(\eta) = -nA''(\eta) = -n\mathrm{Var}_\eta\left[T(X_1)\right] \leq 0$$

Remark 7. *If the natural parameter η is a bijective function of a different parameter θ, the uniqueness of the MLE is maintained. This is a consequence of the equivariance property of MLEs.*

Example 21. *Bernoulli Distribution*

*Consider an i.i.d. sample $X_1, \ldots, X_n$ from a **Bernoulli**(p) distribution, with probability mass function (PMF) $f(x, p) = p^x(1-p)^{1-x}$, where $x \in \{0, 1\}$.*

We showed that the Bernoull distribution belongst to a 1 parameter Exponential Family, with $\eta = \log\left(\frac{p}{1-p}\right) \implies p = \frac{e^\eta}{1+e^\eta}$, $T(x) = x$ *and* $A(\eta) = -\log(1-p) = \log(1+e^\eta)$

The Proposition guarantees that the MLE $\hat{\eta}$ *is the unique solution to the equation* $\mathbf{A}'(\hat{\eta}) = \bar{\mathbf{T}} = \frac{1}{n}\sum_{i=1}^n X_i$. *We have that* $\mathbf{A}'(\eta) = \frac{e^\eta}{1+e^\eta}$ *and then,*

$$\frac{e^{\hat{\eta}}}{1+e^{\hat{\eta}}} = \bar{X}$$

Therefore, the MLE of p, *is* $\hat{p} = \bar{X}$.

2.4 Asymptotic results

2.4.1 Consistency of Moments Estimators

Theorem 12. *Let* $X_1, \ldots, X_n$ *be a random sample from a distribution belonging to the family* $\mathcal{F} = \{F(x,\theta)$ *with* $\theta \in \Theta \subset \mathbb{R}\}$, *where* $\boldsymbol{\theta_0} \in \boldsymbol{\Theta}$ *is the true parameter that generated the data. Let* $h(x)$ *be a continuous real-valued function. Suppose that the population moment* $E_\theta(h(X_1)) = m(\theta)$ *is a continuous and strictly monotonic function of* θ. *Let the method of moments estimator* $\hat{\theta}_n$ *be defined as the solution to*

$$\frac{1}{n}\sum_{i=1}^n h(X_i) = E_\theta(h(X_1)) = m(\theta).$$

Then, with probability 1, there exists n_0 *such that for all* $n \geq n_0$ *the equation defining* $\hat{\theta}_n$ *has a solution, and* $\hat{\theta}_n$ *is* ***strongly consistent*** *for* $\boldsymbol{\theta_0}$.

Proof. Let $\varepsilon > 0$. We need to show that, with probability 1,

$$\text{there exists } n_0 \text{ such that } |\hat{\theta}_n - \boldsymbol{\theta_0}| < \varepsilon \text{ for } n \geq n_0.$$

Assume, without loss of generality, that $m(\theta)$ is **strictly increasing**. The proof for a strictly decreasing function follows analogously. Since $m(\theta)$ is strictly increasing, we evaluate the bounds around the true parameter θ_0:

$$m(\theta_0 - \varepsilon) < m(\theta_0) < m(\theta_0 + \varepsilon).$$

Let $\delta = \min(m(\theta_0 + \varepsilon) - m(\theta_0), m(\theta_0) - m(\theta_0 - \varepsilon))$. Thus,

$$m(\theta_0 - \varepsilon) \leq m(\theta_0) - \delta < m(\theta_0) < m(\theta_0) + \delta \leq m(\theta_0 + \varepsilon).$$

By the **Strong Law of Large Numbers** (S.L.L.N.), since the true population mean is $E_{\boldsymbol{\theta_0}}(h(X_1)) = m(\boldsymbol{\theta_0})$:

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n h(X_i) = E_{\boldsymbol{\theta_0}}(h(X_1)) = m(\boldsymbol{\theta_0}) \quad \text{w.p. 1 (with probability 1).}$$

Therefore, with probability 1, given $\delta > 0$, there exists n_0 such that for all $n \geq n_0$:

$$\left| \frac{1}{n} \sum_{i=1}^{n} h(X_i) - m(\boldsymbol{\theta_0}) \right| \leq \delta.$$

This inequality implies:

$$m(\boldsymbol{\theta_0}) - \delta \leq \frac{1}{n} \sum_{i=1}^{n} h(X_i) \leq m(\boldsymbol{\theta_0}) + \delta.$$

Combining this with the definition of δ:

$$m(\theta_0 - \varepsilon) \leq \frac{1}{n} \sum_{i=1}^{n} h(X_i) \leq m(\theta_0 + \varepsilon) \quad \text{for } n \geq n_0.$$

By definition, the Method of Moments Estimator $\hat{\theta}_n$ satisfies:

$$\frac{1}{n} \sum_{i=1}^{n} h(X_i) = E_{\hat{\theta}_n}(h(X_1)) = m(\hat{\theta}_n).$$

Substituting $m(\hat{\theta}_n)$ into the inequality:

$$m(\theta_0 - \varepsilon) \leq m(\hat{\theta}_n) \leq m(\theta_0 + \varepsilon) \quad \text{for } n \geq n_0.$$

Since $m(\theta)$ is **continuous** and **strictly increasing** (and thus invertible), we infer that the argument of the function must also be bounded by the same values:

$$\theta_0 - \varepsilon \leq \hat{\theta}_n \leq \theta_0 + \varepsilon \quad \text{for } n \geq n_0,$$

which is equivalent to $|\hat{\theta}_n - \boldsymbol{\theta_0}| < \varepsilon$ for $n \geq n_0$. This proves the strong consistency of $\hat{\theta}_n$ for $\boldsymbol{\theta_0}$.

2.4.2 Consistency of the Maximum Likelihood Estimator (MLE)

We state a theorem establishing the consistency of maximum likelihood estimators for the single-parameter case. We denote the true, unknown parameter that generated the data as $\boldsymbol{\theta_0}$.

Let $X_1, \ldots, X_n$ be a random sample. The MLE, $\hat{\theta}_n$, maximizes the likelihood function:

$$\max_{\theta \in \Theta} \prod_{i=1}^{n} f(x_i, \theta) = \prod_{i=1}^{n} f(x_i, \hat{\theta}_n)$$

It can be shown that under very general conditions, $\hat{\theta}_n$ is **strongly consistent** for $\boldsymbol{\theta_0}$.

Theorem 13 (Strong Consistency of the MLE)**.** *Let $X_1, \ldots, X_n$ be a random sample from a discrete or continuous distribution with density (or PMF) in the family $f(x, \theta)$ with $\theta \in \Theta$, where Θ is an open interval in $\mathbb{R}$ and $\boldsymbol{\theta_0} \in \boldsymbol{\Theta}$ is the true parameter. Assume that $f(x, \theta)$ is differentiable with respect to θ and that the set of support $S = \{x : f(x, \theta) \neq 0\}$ is independent of θ for all $\theta \in \Theta$. Let $\hat{\theta}_n$ be the Maximum Likelihood*

*Estimator of θ, which satisfies the **Score Equation**:*

$$\sum_{i=1}^{n} \frac{\partial \ln f(x_i, \hat{\theta}_n)}{\partial \theta} = 0$$

Finally, assume that the score equation has at most one solution and that $\theta \neq \theta'$ implies that $f(x, \theta) \neq f(x, \theta')$. Then $\lim_{n\to\infty} \hat{\theta}_n = \boldsymbol{\theta_0}$ a.s. (almost surely).

For the sake of simplicity in the proof, the conditions used in the theorem are often stronger than those strictly necessary for the theorem to hold. The theorem also holds in the multiparameter case.

Proof. Below, we provide a heuristic idea of the proof. The goal is to show that the Maximum Likelihood Estimator ($\hat{\theta}_{\text{MV}}$) converges to the true parameter $\boldsymbol{\theta_0}$.

Consider the **average log-likelihood function**

$$\ell_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ln f(X_i; \theta)$$

As $n \to \infty$, by the **Strong Law of Large Numbers**, the sample mean converges almost surely to its expected value, which is the **expected log-likelihood**, $\ell(\theta)$:

$$\ell_n(\theta) \xrightarrow{\text{a.s.}} E_{\boldsymbol{\theta_0}}[\ln f(X; \theta)] = \ell(\theta)$$

Where the limiting function $\ell(\theta)$ is defined as:

$$\ell(\theta) = \int \ln(f(x; \theta)) f(x; \boldsymbol{\theta_0}) dx$$

The essence of consistency is that for large n, the value of θ that maximizes the sample function $\ell_n(\theta)$ ($\hat{\theta}_{\text{MV}}$) should converge to the value that maximizes the limiting function $\ell(\theta)$ ($\boldsymbol{\theta_0}$).

We prove that $\boldsymbol{\theta_0}$ is the value that maximizes $\ell(\theta)$. We use the Kullback-Leibler identity. By Jensen's inequality, the Kullback-Leibler divergence between the true distribution $f(\cdot, \boldsymbol{\theta_0})$ and any other $f(\cdot, \theta)$ is non-negative. This implies:

$$E_{\boldsymbol{\theta_0}} \left[\ln \left(\frac{f(X; \theta)}{f(X; \boldsymbol{\theta_0})} \right) \right] \leq 0$$

Rewriting this in terms of $\ell(\theta)$:

$$\ell(\theta) - \ell(\boldsymbol{\theta_0}) \leq 0 \implies \ell(\theta) \leq \ell(\boldsymbol{\theta_0})$$

This proves that $\boldsymbol{\theta_0}$ is the **unique global maximum** of the limiting function $\ell(\theta)$. Given that:

1. The sample log-likelihood ($\ell_n(\theta)$) converges to the expected log-likelihood ($\ell(\theta)$).
2. The expected function $\ell(\theta)$ is uniquely maximized at the true parameter $\boldsymbol{\theta_0}$.

The formal, rigorous step is proving that if a sequence of functions converges uniformly to a limit function, then the sequence of their maximizing arguments must converge to the maximizer of the limit function.

Based on the convergence of the functions, we **heuristically conclude** that the value $\hat{\theta}_{\text{MV}}$ that maximizes $\ell_n(\theta)$ must converge to $\boldsymbol{\theta_0}$, establishing the strong consistency of the MLE.

2.5 Asymptotic Distribution

The consistency of an estimator $\hat{\theta}_n$ only tells us that it converges to the true parameter $\boldsymbol{\theta_0}$ as $n \to \infty$. The **convergence rate** quantifies how quickly this convergence happens.

Suppose $\mathcal{F} = \{f(x;\theta); \theta \in \Theta\}$ is a model for the distribution of $X_1, \ldots, X_n$, and $\boldsymbol{\theta_0} \in \mathbb{R}^p$ is the true parameter of interest. If for some $\alpha > 0$, the following condition holds for the estimator $\hat{\theta}_n$:

$$n^\alpha(\hat{\theta}_n - \boldsymbol{\theta_0}) \xrightarrow{d} G_{\theta_0} \tag{2.1}$$

where G_{θ_0} is a non-degenerate distribution for all $\theta \in \Theta$, then $\hat{\theta}_n$ is said to be an n^α-consistent estimator of $\boldsymbol{\theta_0}$, and n^α is the **convergence rate** (or **normalization constant**) of $\hat{\theta}_n$.

Example 22. *Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} U(0, \theta_0)$, where $\theta_0 > 0$ is the true parameter. We compare two consistent estimators for θ_0:*

1. $\hat{\theta}_{n,1} = X_{(n)} = \max\{X_1, \ldots, X_n\}$ *(MLE)*
2. $\hat{\theta}_{n,2} = 2\bar{X}_n$ *(MOM)*

1. We want to find the scaling factor n^α such that the limiting distribution of $n^\alpha(\hat{\theta}_{n,1} - \boldsymbol{\theta_0})$ is non-degenerate. Since $X_{(n)}$ converges to $\boldsymbol{\theta_0}$ from below, we define the scaled variable Z_n:

$$Z_n = n(\boldsymbol{\theta_0} - X_{(n)})$$

We compute the cumulative distribution function (CDF) of Z_n for $z > 0$:

$$\begin{aligned} F_{Z_n}(z) &= P(Z_n \leq z) \\ &= P(n(\boldsymbol{\theta_0} - X_{(n)}) \leq z) \\ &= P(\boldsymbol{\theta_0} - X_{(n)} \leq z/n) \\ &= P(X_{(n)} \geq \boldsymbol{\theta_0} - z/n) \end{aligned}$$

Since $X_{(n)} \leq \boldsymbol{\theta_0}$ always holds, the complement event $P(X_{(n)} < \boldsymbol{\theta_0} - z/n)$ requires all X_i to be less than the bound $\boldsymbol{\theta_0} - z/n$.

$$\begin{aligned} F_{Z_n}(z) &= 1 - P(X_{(n)} < \boldsymbol{\theta_0} - z/n) \\ &= 1 - P(X_1 < \boldsymbol{\theta_0} - z/n, \ldots, X_n < \boldsymbol{\theta_0} - z/n) \end{aligned}$$

Due to the independence and identical distribution of X_i:

$$F_{Z_n}(z) = 1 - [P(X_1 < \boldsymbol{\theta_0} - z/n)]^n = 1 - \left[F_{X_1}\left(\boldsymbol{\theta_0} - \frac{z}{n}\right)\right]^n$$

For $X_i \sim U(0, \boldsymbol{\theta_0})$, the CDF is $F_{X_1}(x) = x/\boldsymbol{\theta_0}$ for $x \in (0, \boldsymbol{\theta_0}]$. Substituting the argument:

$$F_{Z_n}(z) = 1 - \left[\frac{\boldsymbol{\theta_0} - z/n}{\boldsymbol{\theta_0}}\right]^n = 1 - \left[1 - \frac{z}{n\boldsymbol{\theta_0}}\right]^n$$

Taking the limit as $n \to \infty$ and using the identity $\lim_{n\to\infty}\left(1 - \frac{a}{n}\right)^n = e^{-a}$:

$$\lim_{n\to\infty} F_{Z_n}(z) = 1 - e^{-z/\boldsymbol{\theta_0}}$$

This is the CDF of an Exponential distribution with rate $1/\boldsymbol{\theta_0}$.

Therefore, we have established the convergence:

$$n(\boldsymbol{\theta_0} - \hat{\theta}_{n,1}) \xrightarrow{d} Exp(1/\theta_0)$$

This confirms that $\hat{\theta}_{n,1}$ has a convergence rate of $\mathbf{n^1}$.

2. For the estimator $\hat{\theta}_{n,2}$, the convergence rate is determined by the Central Limit Theorem (CLT). For $X_i \sim U(0, \boldsymbol{\theta_0})$, the true population mean and variance are:

$$E(X_i) = \frac{\theta_0}{2} \quad \textit{and} \quad \mathrm{Var}(X_i) = \frac{\theta_0^2}{12}$$

The CLT states that the scaled sample mean converges to a Normal distribution:

$$\sqrt{n}(\bar{X}_n - E(X_i)) \xrightarrow{d} N\left(0, \mathrm{Var}(X_i)\right)$$

Substituting the true parameter values:

$$\sqrt{n}\left(\bar{X}_n - \frac{\theta_0}{2}\right) \xrightarrow{d} N\left(0, \frac{\theta_0^2}{12}\right)$$

Since $\hat{\theta}_{n,2} = g(\bar{X}_n) = 2\bar{X}_n$ is a linear transformation, we can apply the properties of limiting distributions. We multiply both sides of the convergence by 2:

$$2 \cdot \sqrt{n}\left(\bar{X}_n - \frac{\theta_0}{2}\right) \xrightarrow{d} N\left(0, 2^2 \cdot \frac{\theta_0^2}{12}\right)$$

Rearranging the term on the left:

$$\sqrt{n}\left(2\bar{X}_n - \boldsymbol{\theta_0}\right) \xrightarrow{d} N\left(0, \frac{4\theta_0^2}{12}\right)$$

Therefore:

$$\sqrt{n}(\hat{\theta}_{n,2} - \boldsymbol{\theta_0}) \xrightarrow{d} N\left(0, \frac{\theta_0^2}{3}\right)$$

This confirms that $\hat{\theta}_{n,2}$ has a convergence rate of $\mathbf{n^{1/2}}$.

Since n^1 is a faster rate of convergence than $n^{1/2}$, we prefer $\hat{\theta}_{n,1} = X_{(n)}$ over $\hat{\theta}_{n,2} = 2\bar{X}_n$ for estimating $\boldsymbol{\theta_0}$ based on this asymptotic criterion.

2.5.1 Asymptotically Normal Estimators

For many commonly used estimators $\hat{\theta}_n$, the convergence rate is $\mathbf{n^{1/2}}$ and the limiting distribution G_{θ_0} is a Normal distribution. This means:

$$\sqrt{n}(\hat{\theta}_n - \boldsymbol{\theta_0}) \xrightarrow{d} N(0, V(\theta_0)) \tag{2.2}$$

where $V(\theta_0)$ is a positive definite matrix (or scalar variance).

$V(\theta_0)$ is commonly called the "asymptotic variance" of $\hat{\theta}_n$. This denomination is technically imprecise for two reasons:

1. $V(\theta_0)$ is the variance of the **asymptotic distribution**, not necessarily the limit of the sequence of variances $\lim_{n\to\infty} \mathrm{Var}_{\theta_0}[\sqrt{n}(\hat{\theta}_n - \boldsymbol{\theta_0})]$.
2. $V(\theta_0)$ is the variance of the limiting distribution of the scaled term $\sqrt{n}(\hat{\theta}_n - \boldsymbol{\theta_0})$, not the asymptotic distribution of $\hat{\theta}_n$ itself, which has an asymptotic variance of 0.

2.5.2 MOM ASYM

The next theorem states that the method of moments estimator is consistent and asymptotically normally distributed. In order to state the result we need some additional notation. Let $\mathbf{Y}_1 = (X_1, X_1^2, \ldots, X_1^d)^T$ and $\psi(\theta_0) = (m_1(\theta_0), \ldots, m_d(\theta_0))^T$ denote its expectation and $\Sigma(\theta_0) = \mathrm{Var}(\mathbf{Y}_1)$ its variance.

Theorem 14. *If ψ^{-1} is continuously differentiable at $\psi(\theta_0)$ then*

$$\sqrt{n}(\hat{\theta}_n^{MM} - \theta_0) \xrightarrow{D} \mathcal{N}(0, V(\theta_0));$$

where $V(\theta_0) = [\nabla\psi^{-1}(\psi(\theta_0))]\Sigma(\theta_0)[\nabla\psi^{-1}(\psi(\theta_0))]^T$.

2.5.3 Asymptotic Normality of the Maximum Likelihood Estimator (MLE)

The following is a fundamental result in Maximum Likelihood Estimation (MLE) theory: The expected value of the Score Function (the first derivative of the log-likelihood) evaluated at the true parameter $\boldsymbol{\theta_0}$ is zero.

$$E_{\boldsymbol{\theta_0}}\left[\frac{\partial \log f(X_i, \boldsymbol{\theta_0})}{\partial\theta}\right] = 0 \quad \text{(Expected Score is Zero)}$$

The proof relies on the basic property that the probability density (or mass) function must integrate to one, $\int f(x;\theta)dx = 1$. Assuming the necessary regularity conditions allow differentiation under the integral sign, we use the identity $\frac{\partial f}{\partial\theta} = \left(\frac{\partial \log f}{\partial\theta}\right) f$:

$$\int \frac{\partial f(x;\theta)}{\partial\theta} dx = \frac{\partial}{\partial\theta}(1) = 0$$

Substituting the log-derivative identity and recognizing the integral as the expected value (evaluated at $\boldsymbol{\theta_0}$) concludes the proof:

$$\int \left[\frac{\partial \log f(x;\boldsymbol{\theta_0})}{\partial\theta}\right] f(x;\boldsymbol{\theta_0})dx = E_{\boldsymbol{\theta_0}}\left[\frac{\partial \log f(X;\boldsymbol{\theta_0})}{\partial\theta}\right] = 0$$

With the expected score proven to be zero, we introduce the crucial notion of the **Fisher Information**, $I(\theta)$, which measures the amount of information the sample provides about the parameter θ. The Fisher Information is defined as the variance of the Score Function:

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\log L(\theta; X_1)\right)^2\right] = -E\left[\frac{\partial^2}{\partial\theta^2}\log L(\theta; X_1)\right];$$

where X_1 means one single observation and $L(\theta; X_1)$ is its likelihood function. When it comes to the whole sample (of n i.i.d. random variables), the Fisher Information of the whole sample, $I_n(\theta)$, is additive:

$$I_n(\theta) = E\left[-\frac{\partial^2}{\partial\theta^2}\sum_{i=1}^{n}\log L(\theta; X_i)\right] = \sum_{i=1}^{n} E\left[-\frac{\partial^2}{\partial\theta^2}\log L(\theta; X_i)\right] = nI(\theta).$$

We will revisit this definition later in more detail, explaining its importance and its properties.

Theorem 15. *Under regularity conditions we have*

$$\hat{\theta}_n^{MLE} \xrightarrow{P} \theta_0; \quad \text{as } n \to \infty$$

$$\sqrt{n}(\hat{\theta}_n^{MLE} - \theta_0) \xrightarrow{D} \mathcal{N}(0, I(\theta_0)^{-1})$$

where $I(\theta_0)$ is the Fisher Information.

A careful proof and assumptions can be found in Zacks, S. (1971, The Theory of Statistical Inference. J. Wiley & Sons).

Proof.

The MLE, $\hat{\theta}_n$, is defined as the solution to the Score Equation, $\ell'(\hat{\theta}_n) = 0$. We expand $\ell'(\hat{\theta}_n)$ around the true parameter $\boldsymbol{\theta_0}$ using a second-order Taylor series:

$$0 = \ell'(\hat{\theta}_n) = \ell'(\boldsymbol{\theta_0}) + (\hat{\theta}_n - \boldsymbol{\theta_0})\ell''(\boldsymbol{\theta_0}) + \frac{1}{2}(\hat{\theta}_n - \boldsymbol{\theta_0})^2\ell'''(\xi_n)$$

where ξ_n is an intermediate point between $\hat{\theta}_n$ and $\boldsymbol{\theta_0}$.

Rearranging to isolate the term $(\hat{\theta}_n - \boldsymbol{\theta_0})$:

$$\hat{\theta}_n - \boldsymbol{\theta_0} = \frac{-\ell'(\boldsymbol{\theta_0})}{\ell''(\boldsymbol{\theta_0}) + \frac{1}{2}(\hat{\theta}_n - \boldsymbol{\theta_0})\ell'''(\xi_n)}$$

Now, we scale by $\sqrt{n}$ to obtain the required convergence term:

$$\sqrt{n}(\hat{\theta}_n - \boldsymbol{\theta_0}) = \frac{n^{-1/2}\ell'(\boldsymbol{\theta_0})}{-n^{-1}\ell''(\boldsymbol{\theta_0}) - \frac{1}{2}n^{-1}(\hat{\theta}_n - \boldsymbol{\theta_0})\ell'''(\xi_n)} \tag{2.3}$$

The numerator is the scaled sum of the Score function evaluated at the true param-

eter $\boldsymbol{\theta_0}$:

$$n^{-1/2}\ell'(\boldsymbol{\theta_0}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(X_i, \boldsymbol{\theta_0})}{\partial \theta}$$

The terms in the summation are i.i.d. random variables with:

$$E_{\boldsymbol{\theta_0}}\left[\frac{\partial \log f(X_i, \boldsymbol{\theta_0})}{\partial \theta}\right] = 0 \quad \text{(Expected Score is Zero)}$$

and variance equal to the Fisher Information $I(\theta_0)$:

$$\text{Var}_{\boldsymbol{\theta_0}}\left[\frac{\partial \log f(X_i, \boldsymbol{\theta_0})}{\partial \theta}\right] = E_{\boldsymbol{\theta_0}}\left[\left(\frac{\partial \log f(X_i, \boldsymbol{\theta_0})}{\partial \theta}\right)^2\right] = I(\theta_0)$$

By the **Central Limit Theorem (CLT)**, the numerator converges in distribution:

$$n^{-1/2}\ell'(\boldsymbol{\theta_0}) \xrightarrow{d} N(0, I(\theta_0))$$

The first term of the denominator is the negative scaled second derivative of the log-likelihood (Hessian):

$$-n^{-1}\ell''(\boldsymbol{\theta_0}) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(X_i, \boldsymbol{\theta_0})}{\partial \theta^2}$$

By the **Law of Large Numbers (LLN)** and the regularity condition $E_{\boldsymbol{\theta_0}}[-\ell''(\boldsymbol{\theta_0})] = I(\theta_0)$:

$$-n^{-1}\ell''(\boldsymbol{\theta_0}) \xrightarrow{p} E_{\boldsymbol{\theta_0}}\left[-\frac{\partial^2 \log f(X_i, \boldsymbol{\theta_0})}{\partial \theta^2}\right] = I(\theta_0)$$

The second (remainder) term of the denominator, $\frac{1}{2}n^{-1}(\hat{\theta}_n - \boldsymbol{\theta_0})\ell'''(\xi_n)$, converges in probability to zero because $\hat{\theta}_n \xrightarrow{p} \boldsymbol{\theta_0}$ (consistency of the MLE) and under typical regularity conditions on the third derivative $\ell'''(\theta)$.

$$\frac{1}{2}n^{-1}(\hat{\theta}_n - \boldsymbol{\theta_0})\ell'''(\xi_n) \xrightarrow{p} 0$$

Therefore, the entire denominator converges in probability:

$$-n^{-1}\ell''(\boldsymbol{\theta_0}) - \frac{1}{2}n^{-1}(\hat{\theta}_n - \boldsymbol{\theta_0})\ell'''(\xi_n) \xrightarrow{p} I(\theta_0) + 0 = I(\theta_0)$$

Applying Slutsky's Theorem to equation (2.3) (where the numerator converges in distribution and the denominator converges in probability to a constant $I(\theta_0)$):

$$\sqrt{n}(\hat{\theta}_n - \boldsymbol{\theta_0}) \xrightarrow{d} \frac{N(0, I(\theta_0))}{I(\theta_0)}$$

Using the properties of the Normal distribution, this simplifies to the final result:

$$\sqrt{n}(\hat{\theta}_n - \boldsymbol{\theta_0}) \xrightarrow{d} N\left(0, \frac{I(\theta_0)}{I(\theta_0)^2}\right) = N\left(0, \frac{1}{I(\theta_0)}\right)$$

This demonstrates that the MLE is $\sqrt{n}$-consistent and Asymptotically Normal.

The previous result provides the asymptotic distribution for $\hat{\theta}_n$ itself. However, we are often interested in estimating a transformed parameter, $\beta = q(\boldsymbol{\theta_0})$, where $q(\cdot)$ is a differentiable function (e.g., estimating the variance θ^2 when θ is the mean).

The key tool for finding the asymptotic distribution of the transformed estimator, $\hat{\beta}_n = q(\hat{\theta}_n)$, is the Delta Method. It allows us to transfer the asymptotic normality of $\hat{\theta}_n$ to the asymptotic normality of $q(\hat{\theta}_n)$.

Proposition 3 (Asymptotic Distribution of $q(\hat{\theta}_n)$)**.** *Under regularity conditions. Let $\hat{\theta}_n$ be a consistent MLE of $\boldsymbol{\theta_0}$, and let $q(\theta)$ be a differentiable function such that $q'(\theta) \neq 0$ for all θ.*

Then, $\hat{q}_n = q(\hat{\theta}_n)$ is asymptotically normal for estimating $q(\boldsymbol{\theta_0})$, specifically:

$$\sqrt{n}\left(q(\hat{\theta}_n) - q(\boldsymbol{\theta_0})\right) \xrightarrow{d} N\left(0, \frac{[q'(\boldsymbol{\theta_0})]^2}{I(\boldsymbol{\theta_0})}\right)$$

Proof.

We use a first-order Taylor series expansion of $q(\hat{\theta}_n)$ around $q(\boldsymbol{\theta_0})$:

$$q(\hat{\theta}_n) = q(\boldsymbol{\theta_0}) + q'(\boldsymbol{\theta_0})(\hat{\theta}_n - \boldsymbol{\theta_0}) + R_n$$

where R_n is the remainder term, which is typically of a smaller order (e.g., $o_p(|\hat{\theta}_n - \boldsymbol{\theta_0}|)$).

Rearranging the terms to isolate the desired expression and scaling by $\sqrt{n}$:

$$\sqrt{n}\left(q(\hat{\theta}_n) - q(\boldsymbol{\theta_0})\right) \approx q'(\boldsymbol{\theta_0})\sqrt{n}(\hat{\theta}_n - \boldsymbol{\theta_0})$$

The approximation comes from ignoring the scaled remainder term, $\sqrt{n}R_n$, which can be shown to converge to zero in probability.

We substitute the known asymptotic distribution of the MLE (from Theorem **??**) into the right-hand side. Since $q'(\boldsymbol{\theta_0})$ is a constant:

$$\sqrt{n}(\hat{\theta}_n - \boldsymbol{\theta_0}) \xrightarrow{d} N\left(0, \frac{1}{I(\boldsymbol{\theta_0})}\right)$$

Therefore:

$$q'(\boldsymbol{\theta_0})\sqrt{n}(\hat{\theta}_n - \boldsymbol{\theta_0}) \xrightarrow{d} q'(\boldsymbol{\theta_0}) \cdot N\left(0, \frac{1}{I(\boldsymbol{\theta_0})}\right)$$

Using the property that $c \cdot N(\mu, \sigma^2) = N(c\mu, c^2\sigma^2)$, the final asymptotic distribution is:

$$\sqrt{n}\left(q(\hat{\theta}_n) - q(\boldsymbol{\theta_0})\right) \xrightarrow{d} N\left(0, [q'(\boldsymbol{\theta_0})]^2 \cdot \frac{1}{I(\boldsymbol{\theta_0})}\right)$$

$$\sqrt{n}\left(q(\hat{\theta}_n) - q(\boldsymbol{\theta_0})\right) \xrightarrow{d} N\left(0, \frac{[q'(\boldsymbol{\theta_0})]^2}{I(\boldsymbol{\theta_0})}\right)$$

This demonstrates the asymptotic normality of the transformed estimator, with the variance scaled by the square of the derivative of the transformation function.

Chapter 3

Confidence Intervals

In the previous chapter, we dedicated our efforts to understanding Point Estimation. Our goal was simple: given a sample $X_1, \ldots, X_n$, we wanted to find the single best value, our estimator $\hat{\theta}$, to serve as the best guess for the true unknown parameter θ_0. We analyzed properties like unbiasedness and efficiency (MSE) to evaluate the quality of this estimator.

However, as we have already seen, a point estimator $\hat{\theta}$ is a random variable. Consider this: if $\hat{\theta}$ is a continuous random variable (like the sample mean), what is the probability that our estimate $\hat{\theta}$ is exactly equal to the true value θ_0? That probability is practically zero. This leads us to a key conclusion: even if our point estimator is the 'best,' it is almost certainly wrong. What we do know is that if our estimator is good (for example, it has a low mean squared error), then the true value of θ_0 should not be very far from our estimate $\hat{\theta}$.

This forces us to ask the next question in Statistical Inference:

> ***Instead of finding a single value (which is likely incorrect), can we find a range of values that has a high probability of containing the true parameter*** θ_0***?***

This is the essence of Interval Estimation and the core concept we will explore in this chapter: the Confidence Interval (CI). Instead of reporting $\hat{\theta}$, we will report an interval $[L, U]$ that has a high, predefined probability (for example, 95% or 99%) of 'capturing' the true value θ_0.

3.1 Exact confidence interval

Definition 18 (Two-Sided Confidence Interval)**.** *Let $X_1, \ldots, X_n$ be a collection of* ***independent and identically distributed*** *(i.i.d.) observations drawn from a population governed by the parameter θ_0, where $\theta_0 \in \Theta \subset \mathbb{R}$. For a chosen constant $\alpha \in (0, 1)$, let $L(X_1, \ldots, X_n)$ and $U(X_1, \ldots, X_n)$ be two statistics, representing the* ***lower bound*** *and* ***upper bound*** *of the interval, respectively, such that the following condition is satisfied for all $\theta \in \Theta$:*

$$P_\theta \left[L(X_1, \ldots, X_n) \leq \theta \leq U(X_1, \ldots, X_n) \right] \geq 1 - \alpha$$

The resulting random interval $[L(X_1, \ldots, X_n), U(X_1, \ldots, X_n)]$ is then formally defined as a ***two-sided confidence interval*** *(or confidence interval) for θ with a confidence level of $(1 - \alpha)$.*

Since the interval's construction depends entirely on our observed sample $X_1, \ldots, X_n$, any candidate interval we propose must inherently be a **random interval**. Its endpoints, L and U, are statistics—functions of the sample—meaning the realized interval will vary with each new collection of data.

For this interval to truly represent a likely region for the parameter θ_0, we demand that the probability of the event $\{L \leq \theta \leq U\}$ be at least $1 - \alpha$, where α is a small probability of error. Crucially, this coverage probability must hold **robustly** across the entire parameter space Θ, regardless of the true underlying value of θ_0.

While two-sided intervals are the most common, there are circumstances where we are solely interested in establishing a lower or upper bound on the true value of a parameter θ_0. In such cases, we utilize the concept of a **one-sided confidence interval**.

Definition 19. *Assume we have an **independent and identically distributed** (i.i.d.) random sample $X_1, \ldots, X_n$ drawn from a distribution characterized by the parameter $\theta \in \Theta \subset \mathbb{R}$. Let $\alpha \in (0, 1)$ be a predefined constant.*

1. ***Left-Sided Interval:** If $L(X_1, \ldots, X_n)$ is a statistic (a function of the data) such that:*

$$P_\theta \left[L(X_1, \ldots, X_n) \leq \theta \right] \geq 1 - \alpha$$

   then the random interval $[L(X_1, \ldots, X_n), \infty)$ is termed a lower confidence bound or a left-sided confidence interval for θ with a confidence level of $(1 - \alpha)$.

2. ***Right-Sided Interval:** Similarly, if $U(X_1, \ldots, X_n)$ is a statistic such that:*

$$P_\theta \left[\theta \leq U(X_1, \ldots, X_n) \right] \geq 1 - \alpha$$

   then the random interval $(-\infty, U(X_1, \ldots, X_n)]$ is termed an upper confidence bound or a right-sided confidence interval for θ with a confidence level of $(1 - \alpha)$.

It is vital to recognize that $[L(X_1, \ldots, X_n), U(X_1, \ldots, X_n)]$ is random whereas θ represents a **fixed, non-random magnitude**. By convention, most analyses employ confidence intervals at the 95 percent level, which implies setting $\alpha = 0.05$.

The Interpretation of Confidence Intervals is Frequently Misunderstood. A confidence interval does not provide a probability statement concerning θ, precisely because θ itself is not subject to randomness. A common but misleading explanation suggests that if the same experiment were endlessly replicated, the interval would encompass the parameter 95 percent of the time. While factually correct, this interpretation is generally unhelpful, as actual repetition of the identical experiment is rare. A more insightful and pragmatic interpretation is as follows:

Imagine carrying out distinct statistical investigations on consecutive days. On Day 1, you generate data and compute a 95 percent confidence interval for parameter θ_1. On Day 2, you collect new, unrelated data to calculate an interval for a different parameter θ_2. You continue this procedure for a sequence of independent parameters $\theta_1, \theta_2, \ldots$ The core guarantee is that, over the entire sequence of intervals generated, 95 percent of those intervals will successfully enclose the respective true parameter value. This interpretation correctly isolates the probability to the performance of the interval-generating procedure, rather than requiring the impractical notion of endlessly repeating a single study.

Example 23 (Confidence Interval for the Mean μ (Variance σ^2 Known))**.** *Consider an **independent and identically distributed** (i.i.d.) random sample $X_1, \ldots, X_n$ originating from a Normal distribution $N(\mu, \sigma^2)$, where the mean μ is the unknown parameter*

of interest, but the variance σ^2 is assumed to be known. Our objective is to determine a two-sided confidence interval for μ. We rely on the pivotal quantity derived from known properties of the sample mean for Normal data:

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Let $z_{\alpha/2}$ and $z_{1-\alpha/2}$ represent the $\alpha/2$ and $1-\alpha/2$ quantiles (percentiles) of the standard normal distribution, respectively. The probability of the standardized statistic Z falling between these two quantiles is exactly $1-\alpha$:

$$P\left[z_{\alpha/2} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right] = 1-\alpha$$

The next step involves algebraic isolation of the parameter μ:

$$\begin{aligned} P\left[z_{\alpha/2} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right] &= 1-\alpha \\ P\left[z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{X}_n - \mu \leq z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] &= 1-\alpha \\ P\left[-\bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X}_n + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] &= 1-\alpha \end{aligned}$$

Multiplying the inequality inside the brackets by -1 reverses the direction of the inequalities:

$$P\left[\bar{X}_n - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$

Since the $N(0,1)$ density is symmetric, we know that $z_{\alpha/2} = -z_{1-\alpha/2}$. Substituting this simplifies the interval: Thus, the lower and upper limits are:

$$L(\mathbf{X}) = \bar{X}_n - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \textit{and} \quad U(\mathbf{X}) = \bar{X}_n + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

The resulting $(1-\alpha)$ confidence interval for μ is:

$$\left[\bar{X}_n - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \bar{X}_n + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

This interval is centered around the Maximum Likelihood Estimator of μ, $\bar{X}_n$, and is often written as $\bar{X}_n \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$.

Example 24 (Confidence Interval for the Mean μ (Variance σ^2 Unknown))**.** *Let $X_1, \ldots, X_n$ be an i.i.d. random sample from $N(\mu, \sigma^2)$, where both the mean μ and the variance σ^2 are unknown. Let $S^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2/(n-1)$ be the unbiased sample variance, and let $t_{k;\alpha}$ be the α-quantile of Student's t_k distribution (with k degrees of freedom). Since the population variance σ^2 is unknown, we must use the sample standard deviation S to standardize the sample mean $\bar{X}_n$. This yields the T-statistic:*

$$T = \frac{\bar{X}_n - \mu}{S/\sqrt{n}}$$

This T-statistic follows a Student's t-distribution with $k = n - 1$ degrees of freedom ($T \sim t_{n-1}$). This quantity is pivotal because its distribution does not depend on the unknown parameters μ or σ^2. Let $t_{n-1;1-\alpha/2}$ be the $(1-\alpha/2)$-quantile of the t_{n-1} distribution. Due to the symmetry of the t-distribution, we establish the central $1-\alpha$ area:

$$P\left[-t_{n-1;1-\alpha/2} \leq T \leq t_{n-1;1-\alpha/2}\right] = 1 - \alpha$$

Substitute the expression for T back into the inequality:

$$P\left[-t_{n-1;1-\alpha/2} \leq \frac{\bar{X}_n - \mu}{S/\sqrt{n}} \leq t_{n-1;1-\alpha/2}\right] = 1 - \alpha$$

Now, we perform the algebraic steps to isolate μ:

$$P\left[-t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}} \leq \bar{X}_n - \mu \leq t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}}\right] = 1 - \alpha$$

$$P\left[-\bar{X}_n - t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}} \leq -\mu \leq -\bar{X}_n + t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}}\right] = 1 - \alpha$$

Multiplying by -1 reverses the direction and signs, yielding the final interval form:

$$P\left[\bar{X}_n - t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X}_n + t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}}\right] = 1 - \alpha$$

The resulting random interval has a coverage probability of $1-\alpha$ for all μ and σ^2, which satisfies the definition of a two-sided confidence interval.

3.2 Pivots

The derivation of the confidence interval for the mean parameter of a Normal distribution (as seen in the previous example) appears notably straightforward and lucid. However, the methodology deployed in that construction seems highly specialized and particular to that singular scenario. This raises a fundamental question: How can we transpose the insights from that specific case into universal strategies for building confidence intervals in more complex or general statistical settings?

To address this, we must establish general tools for constructing such regions. The pivotal step in the previous example involved leveraging the property that the standardized sample mean:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

This relationship allowed us to articulate a precise probability statement,

$$P\left[z_{\alpha/2} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right] = 1 - \alpha$$

which was critically valid regardless of the true value of μ. We were then able to algebraically isolate the parameter μ within the inequality. The reason this technique was successful is that the quantity $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ exemplifies what is known as a pivotal quantity.

In the case where the population variance σ^2 is also unknown, the appropriate quantity is the **T**-statistic:

$$T = \frac{\bar{X}_n - \mu}{S/\sqrt{n}}$$

Even though the denominator contains the random sample standard deviation S, the distribution of T is the known Student's t_{n-1} distribution, which is still independent of both unknown parameters (μ and σ^2). Therefore, the T-statistic is also an exact pivot, which enables the construction of the exact t-confidence interval.

Definition 20 (Pivotal Quantity or Pivot)**.** *Let $X_1, \ldots, X_n$ be an i.i.d. sample from the density $f(x, \theta)$. A function $g : \mathcal{X}^n \times \Theta \to \mathbb{R}$ is designated a pivot if it satisfies two conditions:*

1. *Continuity: The mapping $\theta \mapsto g(x_1, \ldots, x_n, \theta)$ is continuous for all possible sample realizations $(x_1, \ldots, x_n) \in \mathcal{X}^n$.*
2. *Parameter-Free Distribution: The cumulative distribution function of the resulting statistic, $P\left[g(X_1, \ldots, X_n, \theta) \leq x\right]$, is independent of the unknown parameter θ.*

Remark 8. *A pivot $g(X_1, \ldots, X_n, \theta)$ is, by its nature, not a statistic because its definition depends on the unknown parameter θ. However, the crucial point is that its sampling distribution is fully known and does not vary with θ. The continuity requirement will become significant when dealing with interval boundaries.*

If we successfully identify a pivot for θ whose probability distribution is known, we can immediately determine quantiles q_1 and q_2 such that:

$$P\left[q_1 \leq g(X_1, \ldots, X_n, \theta) \leq q_2\right] = 1 - \alpha$$

If the function g permits algebraic isolation of θ (as we saw in the Normal example), we obtain an explicit confidence interval.

Even if algebraic isolation is impossible, we can still numerically define the confidence set as the collection of all θ values that satisfy the inequality for the observed data:

$$C_n = \{\theta \in \Theta \mid q_1 \leq g(X_1, \ldots, X_n, \theta) \leq q_2\}$$

Due to the continuity requirement on g, this set C_n will typically be a single interval (especially if g is monotonic in θ) or a union of intervals.

Example 25. *Let $X_1, \ldots, X_n$ be a random sample from a Uniform distribution $\mathcal{U}(0, \theta)$, where $\theta > 0$. The probability density function (PDF) is $f(x|\theta) = \frac{1}{\theta}$ for $0 < x < \theta$. The **Maximum Likelihood Estimator (MLE)** is the maximum order statistic:*

$$\hat{\theta}_n = X_{(n)}$$

Since the distribution of the MLE, $\hat{\theta}_n = X_{(n)}$, can be derived exactly, we can construct an exact confidence interval (CI) for θ. The Cumulative Distribution Function (CDF) of $X_{(n)}$ is $F_{X_{(n)}}(t) = \left(\frac{t}{\theta}\right)^n$ for $0 < t < \theta$. We define the pivot quantity Y by normalizing the estimator:

$$Y = \frac{X_{(n)}}{\theta}$$

*The CDF of Y is $F_Y(y) = y^n$ for $0 < y < 1$. Thus, Y follows a **Beta distribution** $Beta(n, 1)$. For a $(1-\alpha)100\%$ CI, we find the quantiles $y_{\alpha/2}$ and $y_{1-\alpha/2}$ of the distribution of Y:*

$$P(Y \leq y_{\alpha/2}) = \frac{\alpha}{2} \implies y_{\alpha/2} = \left(\frac{\alpha}{2}\right)^{1/n}$$

$$P(Y \leq y_{1-\alpha/2}) = 1 - \frac{\alpha}{2} \implies y_{1-\alpha/2} = \left(1 - \frac{\alpha}{2}\right)^{1/n}$$

The probability statement is $P(y_{\alpha/2} < Y < y_{1-\alpha/2}) = 1 - \alpha$. *Substituting* $Y = X_{(n)}/\theta$:

$$P\left(y_{\alpha/2} < \frac{X_{(n)}}{\theta} < y_{1-\alpha/2}\right) = 1 - \alpha$$

Inverting the inequality to isolate θ:

$$CI_{EXACT}(\theta) = \left[\frac{X_{(n)}}{y_{1-\alpha/2}}, \frac{X_{(n)}}{y_{\alpha/2}}\right] = \left[\frac{X_{(n)}}{\left(1 - \frac{\alpha}{2}\right)^{1/n}}, \frac{X_{(n)}}{\left(\frac{\alpha}{2}\right)^{1/n}}\right]$$

Finding an exact pivot and analytically determining its distribution is generally challenging and relies heavily on the specific parametric family. Therefore, there is often no general, explicit formula for confidence intervals.

3.3 Asymptotics Intervals

We often address this challenge by employing an approximate pivot. This is a function that may not satisfy the pivot criteria for small sample sizes (n), but whose distribution converges to a known, parameter-free distribution as $n \to \infty$.

Definition 21 (Asymptotic or Approximate Pivot)**.** *Let* $X_1, \ldots, X_n$ *be an i.i.d. sample from* $f(x, \theta)$. *A function* $g : \mathcal{X}^n \times \Theta \to \mathbb{R}$ *is an approximate pivot if:*

1. *Continuity: For all* n, *the mapping* $\theta \mapsto g(x_1, \ldots, x_n, \theta)$ *remains continuous.*
2. *Asymptotic Distribution: The sequence of statistics converges in distribution to a random variable* Y *whose distribution is independent of* θ:

$$g(X_1, \ldots, X_n, \theta) \xrightarrow{d} Y$$

If the asymptotic distribution of an approximate pivot, F_Y, is known, we can construct an approximate confidence interval. If Y is a continuous random variable, we select quantiles q_1 and q_2 of F_Y such that $P[q_1 \leq Y \leq q_2] = 1 - \alpha$. By the definition of convergence in distribution, this implies that:

$$P\left[q_1 \leq g(X_1, \ldots, X_n, \theta) \leq q_2\right] \approx 1 - \alpha \quad \text{for large } n$$

We can consequently utilize the approximate pivot to establish a confidence interval that is asymptotically valid.

Example 26 (Asymptotic Interval for the Population Mean (Distribution Unknown))**.** *Consider an i.i.d. collection of random variables* $X_1, \ldots, X_n$ *drawn from a distribution with an unknown expected value* $\mu = E[X]$ *and a finite, yet unknown, variance* $\sigma^2 < \infty$. *Our goal is to derive a large-sample,* $(1 - \alpha)$ *confidence region for* μ.

1. ***Central Limit Theorem (CLT):*** *The standardized sample average converges in distribution:* $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$.

2. ***Variance Substitution:*** *Since σ^2 is unknown, we substitute its consistent estimator, the sample variance $S^2 = \sum_{i=1}^{n}(X_i - \bar{X}_n)^2/(n-1)$. The Strong Law of Large Numbers (SLLN) ensures $S^2 \xrightarrow{P} \sigma^2$.*

3. ***Slutsky's Combination:*** *By applying Slutsky's Theorem, the substitution of S for σ does not alter the limiting distribution, yielding the final approximate pivot:*

$$g(\mathbf{X}, \mu) = \frac{\bar{X}_n - \mu}{S/\sqrt{n}} \xrightarrow{d} N(0,1)$$

By setting the asymptotic probability statement using the standard normal quantiles ($z_{1-\alpha/2}$) and isolating μ, we establish that the interval:

$$\left[\bar{X}_n - z_{1-\alpha/2}\frac{S}{\sqrt{n}}, \quad \bar{X}_n + z_{1-\alpha/2}\frac{S}{\sqrt{n}}\right]$$

is an approximate two-sided $(1-\alpha)$ confidence interval for μ, valid for sufficiently large n.

Example 27 (Approximate Interval for Binomial Proportion)**.** *Let $X_1, \ldots, X_n$ be an i.i.d. sample of Bernoulli trials, $X_i \sim$ Bernoulli(p), where p is the unknown probability of success ($\theta = p$). We seek an approximate $(1-\alpha)$ confidence interval for p.*

The maximum likelihood estimator (MLE) for p is the sample proportion: $\hat{p} = \bar{X}_n = \frac{1}{n}\sum X_i$.

1. Constructing the Approximate Pivot:

- *By the Central Limit Theorem (CLT), we know the standardized mean converges to the standard normal distribution:*

$$\frac{\bar{X}_n - p}{\sqrt{p(1-p)/n}} \xrightarrow{d} N(0,1)$$

This function is an approximate pivot, but the true standard deviation $\sqrt{p(1-p)}$ depends on the unknown parameter p.

- *To obtain a usable statistic, we replace the unknown true standard deviation with its consistent estimator, $\sqrt{\hat{p}(1-\hat{p})}$. By Slutsky's Theorem, this substitution does not change the asymptotic distribution:*

$$g(\mathbf{X}, p) = \frac{\bar{X}_n - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \xrightarrow{d} N(0,1)$$

This resulting quantity is the operational approximate pivot.

2. Deriving the Interval: *We use the quantiles of the standard normal distribution, $z_{1-\alpha/2}$:*

$$P\left[-z_{1-\alpha/2} \leq \frac{\bar{X}_n - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \leq z_{1-\alpha/2}\right] \underset{n\to\infty}{\longrightarrow} 1-\alpha$$

Algebraically isolating the parameter p yields the well-known approximate confidence interval (sometimes called the Wald interval) for large n:

$$\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

3.3.1 Confidence Intervals Based on Maximum Likelihood Estimators (MLEs)

The construction of confidence intervals that are asymptotically valid, as demonstrated in the previous Binomial example, finds its most general theoretical foundation in the asymptotic stability of Maximum Likelihood Estimators (MLEs). This methodology provides a systematic procedure for any statistical model that satisfies certain regularity conditions.

We known that under appropriate regularity conditions (related to the smoothness of the likelihood function), if $\hat{\theta}_n$ is the MLE for the parameter θ, its distribution converges to a Normal distribution. Specifically, the centered and rescaled MLE converges as follows:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\left(0, [I_1(\theta)]^{-1}\right)$$

where $I_1(\theta)$ represents the Fisher Information contained in a single observation.

This asymptotic convergence allows for the creation of a generalized approximate pivot. To obtain a quantity whose limiting distribution is fully known (i.e., independent of θ), we substitute the theoretical Fisher Information $I_1(\theta)$ with the sample Fisher Information $I_n(\hat{\theta}_n)$ evaluated at the MLE. The resulting statistic, known as the Wald Statistic, is our asymptotic pivot:

$$Z_{\text{Wald}} = \frac{\hat{\theta}_n - \theta}{\sqrt{\left[I_n(\hat{\theta}_n)\right]^{-1}}} \xrightarrow{d} N(0,1)$$

Here, $I_n(\hat{\theta}_n) = n \cdot I_1(\hat{\theta}_n)$, where $I_n(\hat{\theta}_n)^{-1}$ represents the estimated asymptotic variance of the MLE $\hat{\theta}_n$.

By using the $z_{1-\alpha/2}$ quantiles of the standard Normal distribution as the capture region for the Wald pivot, the general approximate $(1-\alpha)$ Confidence Interval for θ takes the form:

$$\left[\hat{\theta}_n - z_{1-\alpha/2}\sqrt{\frac{1}{I_n(\hat{\theta}_n)}}, \quad \hat{\theta}_n + z_{1-\alpha/2}\sqrt{\frac{1}{I_n(\hat{\theta}_n)}}\right]$$

This formula provides a standardized and powerful method for obtaining approximate CIs for any parameter, depending solely on the estimate $\hat{\theta}_n$ and the computation of the Fisher Information.

Example 28. *Let $X_1, \ldots, X_n$ be a random sample from a Uniform distribution $\mathcal{U}(0,\theta)$, where $\theta > 0$. The probability density function (PDF) is $f(x|\theta) = \frac{1}{\theta}$ for $0 < x < \theta$. Since the Uniform distribution does not satisfy the standard regularity conditions (because its support depends on θ), the standard asymptotic normality theorem for the MLE does not apply. Instead, a specialized result for the asymptotic distribution of the maximum order statistic must be used:*

$$n(\theta - X_{(n)}) \xrightarrow{d} W, \quad \textit{where } W \sim \textit{Exponential}(1/\theta)$$

The limiting variable W has the CDF $F_W(w) = 1 - e^{-w/\theta}$ for $w > 0$. For large n, we use the approximation $W \approx n(\theta - X_{(n)})$. We define the quantiles $w_{\alpha/2}$ and $w_{1-\alpha/2}$ such that $P(W > w_\gamma) = 1 - \gamma$:

$$P(W > w_\gamma) = 1 - F_W(w_\gamma) = e^{-w_\gamma/\theta} = 1 - \gamma$$

Solving for w_γ: $w_\gamma = -\theta \ln(1-\gamma)$.

We seek $P(w_{\alpha/2} < W < w_{1-\alpha/2}) \approx 1 - \alpha$. *Using the derived quantiles:*

$$P\left(-\theta \ln(1-\alpha/2) < n(\theta - X_{(n)}) < -\theta \ln(\alpha/2)\right) \approx 1 - \alpha$$

We rearrange the terms to solve for θ. *Let* $C_1 = -\ln(1-\alpha/2)$ *and* $C_2 = -\ln(\alpha/2)$.

$$P\left(C_1 < n\left(1 - \frac{X_{(n)}}{\theta}\right) < C_2\right) \approx 1 - \alpha$$

Inverting the inner inequalities leads to the asymptotic CI:

$$CI_{ASYM}(\theta) = \left[\frac{X_{(n)}}{1 - \frac{C_1}{n}}, \frac{X_{(n)}}{1 - \frac{C_2}{n}}\right] = \left[\frac{X_{(n)}}{1 + \frac{\ln(1-\alpha/2)}{n}}, \frac{X_{(n)}}{1 + \frac{\ln(\alpha/2)}{n}}\right]$$

This asymptotic interval provides a good approximation for the true CI when the sample size n *is large.*

3.4 Constructing Confidence Intervals using Hoeffding's Inequality

Hoeffding's Inequality is a powerful non-parametric tool that provides strict probability bounds for the sum of bounded, independent random variables. This inequality is especially useful for constructing confidence intervals when complete information about the underlying data distribution is lacking.

Hoeffding's Inequality

Let $X_1, X_2, \ldots, X_n$ be independent random variables. Suppose each variable is bounded within a known interval, $X_i \in [a_i, b_i]$. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean and $\mu = E[\bar{X}_n]$ be its expected mean (if all X_i have the same mean μ, then $E[\bar{X}_n] = \mu$).

Hoeffding's inequality bounds the probability that the sample mean $\bar{X}_n$ deviates from its expected mean μ by more than an amount $\epsilon > 0$:

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

A particular case is where all variables X_i are **i.i.d.** and bounded in the same interval $[a, b]$, then $b_i - a_i = b - a$ for all i, and the sum in the denominator is $n(b-a)^2$. The inequality simplifies to:

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right).$$

We can rearrange Hoeffding's inequality to obtain a confidence interval for the mean μ, where the probability that μ lies within the interval is at least $1 - \alpha$. Let $1 - \alpha$ be the desired confidence level (e.g., 0.95, where $\alpha = 0.05$). We want to find ϵ such that the probability of error is less than or equal to α:

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \alpha.$$

Using Hoeffding's inequality (in its simplified form, assuming $X_i \in [a, b]$), we set the right side equal to α and solve for ϵ:

$$2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right) = \alpha$$

$$\begin{aligned}
\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right) &= \frac{\alpha}{2} \\
-\frac{2n\epsilon^2}{(b-a)^2} &= \ln\left(\frac{\alpha}{2}\right) \\
\epsilon^2 &= -\frac{(b-a)^2}{2n}\ln\left(\frac{\alpha}{2}\right) \\
\epsilon &= \sqrt{-\frac{(b-a)^2}{2n}\ln\left(\frac{\alpha}{2}\right)}
\end{aligned}$$

With this value of ϵ, the confidence interval for the true mean μ, with a confidence level of at least $1-\alpha$, is:

$$\text{C.I.}_{1-\alpha}(\mu) = \left[\bar{X}_n - \epsilon, \quad \bar{X}_n + \epsilon\right]$$

Or formally:

$$P\left(\mu \in \left[\bar{X}_n - \epsilon, \bar{X}_n + \epsilon\right]\right) \geq 1 - \alpha.$$

The main advantage is that the inequality does not require the variables X_i to follow any specific distribution (e.g., normal), only that they are bounded and independent. The bound depends only on the range amplitude $(b-a)$ and the sample size (n), not on the true value of μ. It provides valid probability limits even for small sample sizes (unlike the Central Limit Theorem, which is asymptotic). The main limitation is that the bound is often wider (and therefore more conservative) than intervals based on the Central Limit Theorem (like Z or T intervals) when normality holds.

Example 29 (Bernoulli Case: Estimating a Probability)**.** *The most common case is estimating the probability of success p of a Bernoulli distribution.*

- *Variables: $X_1, \ldots, X_n$ are i.i.d. $\sim$ Bernoulli(p).*
- *Bounds: The variable X_i is bounded in $[0,1]$. Therefore, $a = 0$, $b = 1$, and $b - a = 1$.*
- *Sample Mean: $\bar{X}_n$ is the sample proportion of successes $\hat{p}$, and its expected mean is $\mu = E[\bar{X}_n] = p$.*

Substituting $a = 0$ and $b = 1$ into the simplified Hoeffding's inequality, we obtain a bound for the probability that the sample proportion deviates from the true probability p:

$$P(|\hat{p} - p| \geq \epsilon) \leq 2\exp\left(-2n\epsilon^2\right).$$

To construct a confidence interval for p with a level $1-\alpha$, we find ϵ such that $2\exp(-2n\epsilon^2) = \alpha$:

$$\epsilon = \sqrt{\frac{1}{2n}\ln\left(\frac{2}{\alpha}\right)}.$$

The confidence interval for p is:

$$C.I._{1-\alpha}(p) = \left[\hat{p} - \epsilon, \quad \hat{p} + \epsilon\right].$$

Example 30 (Confidence Interval for the Cumulative Distribution Function (CDF))**.** *Hoeffding's Inequality can also be used to construct a confidence interval for the Cumulative Distribution Function (CDF)* $F(x) = P(X \leq x)$.

For a fixed point x_0*, the Empirical CDF (ECDF) at* x_0 *is defined as:*

$$\hat{F}_n(x_0) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x_0\}}$$

Where $\mathbf{1}_{\{X_i \leq x_0\}}$ *is an indicator variable.*

We note that the indicator variable $Y_i = \mathbf{1}_{\{X_i \leq x_0\}}$ *is a Bernoulli variable with* $p = F(x_0)$.

- $Y_i \in \{0, 1\}$*, thus it is bounded in* $[0, 1]$.
- *The sample mean of* Y_i *is* $\bar{Y}_n = \hat{F}_n(x_0)$.
- *The expected mean of* $\bar{Y}_n$ *is* $E[\bar{Y}_n] = F(x_0)$.

Applying Hoeffding's inequality (Bernoulli case), we obtain the bound for $F(x_0)$*:*

$$P(|\hat{F}_n(x_0) - F(x_0)| \geq \epsilon) \leq 2 \exp\left(-2n\epsilon^2\right).$$

Using the same ϵ *derived in the Bernoulli case:*

$$\epsilon = \sqrt{\frac{1}{2n} \ln\left(\frac{2}{\alpha}\right)}.$$

The confidence interval for the CDF at point x_0*, with a confidence level of at least* $1 - \alpha$*, is:*

$$C.I._{1-\alpha}(F(x_0)) = \left[\hat{F}_n(x_0) - \epsilon, \quad \hat{F}_n(x_0) + \epsilon\right].$$

It is crucial to clarify that this interval is calculated for each single point x_0; this is not sufficient to provide a confidence band for the entire function $F(x)$. Constructing a confidence band for the entire CDF requires a uniform concentration inequality like the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality.

Chapter 4

The Rao-Cramer Inequality and Optimality.

We typically assess estimators by comparing their mean squared error (MSE). While knowing the relative performance between two estimators is useful, establishing an absolute standard, the best possible MSE, for any given estimation problem provides a powerful benchmark. Identifying a uniformly optimal estimator, T^*, which minimizes the MSE for all parameter values θ, is generally a challenging task that often requires restricting the class of estimators under consideration.

A less ambitious, but highly valuable, objective is to determine the theoretical floor for an estimator's variance, given a specific level of bias. If an estimator is unbiased (i.e., its bias is zero), is there a theoretical minimum variance it can achieve? The answer is provided by the following central result.

Theorem 16. ***(Cramér-Rao Lower Bound)*** *Let $X_1, \ldots, X_n$ be an independent and identically distributed (i.i.d.) sample drawn from a regular parametric family $f(x;\theta)$, where the parameter space $\Theta \subseteq \mathbb{R}$. Let $\hat{\beta}_n$ be an estimator for θ based on the sample. Assume the following regularity conditions hold:*

1. *The variance of the estimator is finite: $Var(\hat{\beta}_n) < \infty$, for all $\theta \in \Theta$.*
2. *The order of integration and differentiation with respect to θ can be exchanged for the integral of the probability density function (PDF).*
3. *The order of integration of the estimator multiplied by the PDF and differentiation with respect to θ can be exchanged.*

If we define the bias of $\hat{\beta}_n$ as $\beta(\theta) = E_\theta[\hat{\beta}_n] - \theta$, and assume $\beta(\theta)$ is differentiable, then the variance of $\hat{\beta}_n$ must satisfy the inequality:

$$Var(\hat{\beta}_n) \geq \frac{[1+\beta'(\theta)]^2}{nE\left[\left(\frac{\partial}{\partial\theta}\log f(X_1;\theta)\right)^2\right]} = \frac{[1+\beta'(\theta)]^2}{nI(\theta)}$$

Remark 9. *For discrete random variables, the integrals above will be replaced by sums.*

Even in the most favorable scenario where the estimator is unbiased ($\beta(\theta) = 0$), the variance is still constrained from below. The denominator, $nI(\theta)$, represents the

total Fisher Information contributed by the sample of size n. This theorem reveals that the minimum achievable variance (and consequently, the minimum MSE for unbiased estimators) is $\mathbf{1/nI}(\theta)$. The factor n^{-1} emphasizes that estimation accuracy improves proportionally with the sample size.

If an unbiased estimator achieves the variance $1/nI(\theta)$, it is known to be the Minimum Variance Unbiased Estimator (MVUE), achieving the best possible MSE among all unbiased estimators.

The proof relies on the properties of the score function, $\frac{\partial}{\partial\theta}\log f(X_1 \mid \theta)$, and the Cauchy–Schwarz inequality applied to the covariance between the estimator T and the score function.

Remark 10. *The assumption regarding the interchangeability of differentiation and integration (Condition 3) is guaranteed under several practical conditions, notably when the underlying distribution belongs to the* ***one-parameter exponential family*** *and T is the natural sufficient statistic.*

Recall: Properties of the Score Function

Before proceeding with the proof of the Cramér-Rao Lower Bound, we recall the essential properties of the score function, and its connection to the Fisher Information, $I(\theta)$. The score function for a single observation X is defined as the partial derivative of the log-likelihood (logarithm of the probability density or mass function) with respect to the parameter θ:

$$\frac{\partial}{\partial\theta}\log f(X;\theta).$$

Key Properties (Under Regularity)

1. Zero Expectation: The expected value of the score function is always zero:

$$E\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right] = 0.$$

2. Variance Equals Fisher Information: The variance of the score function is equal to the Fisher Information, $I(\theta)$:

$$\operatorname{Var}\left[\frac{\partial}{\partial\theta}\log f(x;\theta)\right] = E\left[\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2\right] = I(\theta).$$

Alternative Calculation for Fisher Information

A crucial and often simpler property for computing the Fisher Information involves the second derivative of the log-likelihood.

If we assume that the order of integration/summation and the second derivative can be interchanged, the Fisher Information can also be computed as the negative of the expectation of the second derivative of the log-likelihood:

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)\right].$$

We start from the zero expectation property, $E[U(\theta)] = 0$:

$$\int_{\mathcal{X}} \left(\frac{\partial}{\partial \theta} \log f(x;\theta) \right) f(x;\theta) dx = 0.$$

Differentiating both sides with respect to θ:

$$\frac{\partial}{\partial \theta} \left[\int_{\mathcal{X}} \left(\frac{\partial}{\partial \theta} \log f(x;\theta) \right) f(x;\theta) dx \right] = 0.$$

Interchanging the order of derivative and integral (by regularity) and applying the product rule:

$$\int_{\mathcal{X}} \left[\left(\frac{\partial^2}{\partial \theta^2} \log f(x;\theta) \right) f(x;\theta) + \left(\frac{\partial}{\partial \theta} \log f(x;\theta) \right) \left(\frac{\partial}{\partial \theta} f(x;\theta) \right) \right] dx = 0.$$

Using the identity $\frac{\partial f}{\partial \theta} = \frac{\partial \log f}{\partial \theta} \cdot f$:

$$\int_{\mathcal{X}} \left(\frac{\partial^2}{\partial \theta^2} \log f(x;\theta) \right) f(x;\theta) dx + \int_{\mathcal{X}} \left(\frac{\partial}{\partial \theta} \log f(x;\theta) \right)^2 f(x;\theta) dx = 0.$$

Rewriting in terms of expectation:

$$E\left[\frac{\partial^2}{\partial \theta^2} \log f(x;\theta) \right] + E\left[\left(\frac{\partial}{\partial \theta} \log f(x;\theta) \right)^2 \right] = 0.$$

Since the second term is $I(\theta)$, we conclude:

$$I(\theta) = -E\left[\frac{\partial^2}{\partial \theta^2} \log f(x;\theta) \right].$$

Proof. Denote the total score function for the sample of size n:

$$U_n(\theta) = \frac{\partial}{\partial \theta} \log f_{X_1,\ldots,X_n}(X_1, \ldots, X_n; \theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i;\theta)$$

The expected value of the total score function is also zero:

$$E[U_n(\theta)] = 0.$$

Since $X_1, \ldots, X_n$ are i.i.d., the scores are independent. Therefore, the variance of the sum is the sum of the variances:

$$\mathrm{Var}[U_n(\theta)] = \sum_{i=1}^{n} \mathrm{Var}\left[\frac{\partial}{\partial \theta} \log f(X_i;\theta) \right] = nI(\theta).$$

The bias of the estimator $\hat{\beta}_n$ is $\beta(\theta) = E_\theta[\hat{\beta}_n] - \theta$. Differentiating the expected value of

$\hat{\beta}_n$:

$$\begin{aligned}
\beta'(\theta) + 1 = \frac{\partial}{\partial\theta} E[\hat{\beta}_n] &= \frac{\partial}{\partial\theta} \int_{\mathcal{X}^n} \hat{\beta}_n(\mathbf{x}) f_{X_1,\ldots,X_n}(\mathbf{x};\theta) d\mathbf{x} \\
&= \int_{\mathcal{X}^n} \hat{\beta}_n(\mathbf{x}) \frac{\partial}{\partial\theta} f_{X_1,\ldots,X_n}(\mathbf{x};\theta) d\mathbf{x} \\
&= \int_{\mathcal{X}^n} \hat{\beta}_n(\mathbf{x}) \left(\frac{\partial}{\partial\theta} \log f_{X_1,\ldots,X_n}(\mathbf{x};\theta) \right) f_{X_1,\ldots,X_n}(\mathbf{x};\theta) d\mathbf{x} \\
&= E[\hat{\beta}_n \cdot U_n(\theta)]
\end{aligned}$$

Now, we look at the covariance between $\hat{\beta}_n$ and $U_n(\theta)$:

$$\mathrm{Cov}[U(\theta), \hat{\beta}_n] = E[\hat{\beta}_n \cdot U_n(\theta)] - E[\hat{\beta}_n]E[U_n(\theta)]$$

Since $E[U_n(\theta)] = 0$, we have:

$$\mathrm{Cov}[U_n(\theta), \hat{\beta}_n] = E[\hat{\beta}_n \cdot U_n(\theta)] = \beta'(\theta) + 1$$

For any two random variables, the square of their covariance is bounded by the product of their variances:

$$\Big(\mathrm{Cov}[U_n(\theta), \hat{\beta}_n]\Big)^2 \leq \mathrm{Var}[U_n(\theta)] \cdot \mathrm{Var}[\hat{\beta}_n]$$

Substituting the expressions

$$[\beta'(\theta) + 1]^2 \leq [nI(\theta)] \cdot \mathrm{Var}[\hat{\beta}_n]$$

Solving for $\mathrm{Var}(\hat{\beta}_n)$ yields the Cramér-Rao Lower Bound:

$$\mathrm{Var}(\hat{\beta}_n) \geq \frac{[1 + \beta'(\theta)]^2}{nI(\theta)}$$

Theorem 17. ***(Cramér-Rao Lower Bound)*** *Let $X_1, \ldots, X_n$ be an independent and identically distributed (i.i.d.) sample drawn from a regular parametric family $f(x;\theta)$, where the parameter space $\Theta \subseteq \mathbb{R}$. Let $\hat{\beta}_n$ be an estimator for $q(\theta)$ based on the sample. Assume the following regularity conditions hold:*

1. *The variance of the estimator is finite: $Var(\hat{\beta}_n) < \infty$, for all $\theta \in \Theta$.*
2. *The order of integration and differentiation with respect to θ can be exchanged for the integral of the probability density function (PDF).*
3. *The order of integration of the estimator multiplied by the PDF and differentiation with respect to θ can be exchanged.*

If we define the bias of $\hat{\beta}_n$ as $\beta(\theta) = E_\theta[\hat{\beta}_n] - q(\theta)$, and assume $\beta(\theta)$ is differentiable,

then the variance of $\hat{\beta}_n$ must satisfy the inequality:

$$Var(\hat{\beta}_n) \geq \frac{[q'(\theta) + \beta'(\theta)]^2}{nI(\theta)}$$

Corollary 1. ***(CRLB for Unbiased Estimators)*** *If the estimator $\hat{\beta}_n$ for $q(\theta)$ is **unbiased**, then the bias is zero, $\beta(\theta) = 0$, and thus $\beta'(\theta) = 0$.*

In this case, the Cramér-Rao Lower Bound simplifies to:

$$Var(\hat{\beta}_n) \geq \frac{[q'(\theta)]^2}{nI(\theta)}$$

Furthermore, if $\hat{\beta}_n$ is an unbiased estimator for the parameter θ itself (i.e., $q(\theta) = \theta$), then $q'(\theta) = 1$, and the lower bound is:

$$Var(\hat{\beta}_n) \geq \frac{1}{nI(\theta)}$$

4.1 Asymptotic Efficiency

Let us begin by recalling the definition of asymptotically normal.

Definition 22. *$\hat{\beta}$ is said to be an **asymptotically normal** estimator of a parameter $q(\theta)$ if it satisfies:*

$$\sqrt{n}\left(\hat{\beta} - q(\theta)\right) \xrightarrow{L(F_\theta)} N(0, W(\theta)) \quad \text{for some } W(\theta).$$

Note that with large samples, clearly, among all asymptotically normal estimators, we prefer the one that has the **smallest** $W(\theta)$. This suggests the following definition:

Definition 23. *Suppose two estimators $\hat{\beta}$ and $\tilde{\beta}$ are such that*

$$\sqrt{n}\left(\hat{\beta} - q(\theta)\right) \xrightarrow{L(F_\theta)} N(0, V_1(\theta)) \tag{4.1}$$

$$\sqrt{n}\left(\tilde{\beta} - q(\theta)\right) \xrightarrow{L(F_\theta)} N(0, V_2(\theta)) \tag{4.2}$$

*Then we say that $\hat{\beta}$ is **more efficient** than $\tilde{\beta}$ if*

$$V_1(\theta) < V_2(\theta)$$

Furthermore, given two estimators $\hat{\beta}$ and $\tilde{\beta}$ of a scalar parameter $q(\theta)$, that satisfy the previous definition, the quantity

$$\tau(\theta) = \frac{V_2(\theta)}{V_1(\theta)}$$

is called the **relative asymptotic efficiency** of $\hat{\beta}$ with respect to $\tilde{\beta}$ (note that the variance of $\tilde{\beta}$ is in the numerator). We interpret $\tau(\theta)$ as an indicator of how much larger or smaller the sample size must be when using $\tilde{\beta}$ to obtain the same precision as if we had used $\hat{\beta}$. For example, if $\tau(\theta) = 2$, then we must use a sample twice as large if we use $\tilde{\beta}$ than if we use $\hat{\beta}$ to obtain the same precision in the estimation.

This is shown as follows. If n_1 is the sample size with which we calculate $\hat{\beta}$ and n_2 is the sample size with which we calculate $\tilde{\beta}$, then the variance of $\hat{\beta}$ will be approximately $V_1(\theta)/n_1$ and that of $\tilde{\beta}$ will be approximately $V_2(\theta)/n_2$. If we want to have the same precision, we must obtain sample sizes such that

$$\frac{V_1(\theta)}{n_1} = \frac{V_2(\theta)}{n_2}$$

or equivalently

$$\frac{n_2}{n_1} = \frac{V_2(\theta)}{V_1(\theta)} = \tau(\theta)$$

Hence

$$n_2 = \tau(\theta) n_1.$$

The larger $\tau(\theta)$ is, the more efficient $\hat{\beta}$ will be with respect to $\tilde{\beta}$.

4.1.1 Efficiency of the Maximum Likelihood Estimator (MLE)

It is possible to prove that in a very large class of statistical models, the **Maximum Likelihood Estimator (MLE)** for $q(\theta)$, denoted $\hat{q}_{\text{MLE}} = q(\hat{\theta}_{\text{MLE}})$, is **asymptotically efficient**. Asymptotic efficiency is one of the fundamental reasons why the MLE is one of the preferred estimation procedures.

The heuristic argument for this fact is as follows. Suppose $\hat{\beta}_n$ is an estimator for $q(\theta)$ that is asymptotically normal. Then there exists $W(\theta) > 0$ such that, for any θ, under $f(x;\theta)$:

$$\sqrt{n}\left(\hat{\beta}_n - q(\theta)\right) \xrightarrow{L(F_\theta)} N\left(0, W(\theta)\right) \quad \text{when } n \text{ is large.}$$

Equivalently,

$$\hat{\beta}_n \approx N\left(q(\theta), \frac{W(\theta)}{n}\right) \quad \text{for any } \theta \text{ when } n \text{ is large.}$$

Then $\hat{\beta}_n$ is an **approximately unbiased** estimator of $q(\theta)$. Therefore, by the Cramér-Rao inequality for unbiased estimators (Corollary 1), one would expect the asymptotic variance to satisfy:

$$\frac{W(\theta)}{n} \geq \frac{[q'(\theta)]^2}{nI(\theta)}.$$

Then, canceling the factor $1/n$ on both sides of the last inequality, we arrive at the asymptotic lower bound:

$$W(\theta) \geq \frac{[q'(\theta)]^2}{I(\theta)}$$

Since the Maximum Likelihood Estimator (MLE) of $q(\theta)$ is asymptotically normal (under regularity conditions, by the Delta Method) and the variance of its limiting distribution is precisely

$$W_{\text{MLE}}(\theta) = \frac{[q'(\theta)]^2}{I(\theta)},$$

the last inequality implies that the MLE is **asymptotically efficient**.

A particularly interesting point is that although it is possible (and even common) that:

1. no unbiased estimator of $q(\theta)$ exists, or

2. unbiased estimators of $q(\theta)$ exist but none have variance equal to the Cramér-Rao Lower Bound,

the asymptotic normality and efficiency of the MLE imply that under those models, with large samples, it is possible to obtain a "nearly unbiased" estimator of $q(\theta)$ whose variance is "nearly" equal to the Cramér-Rao bound. This estimator is precisely the Maximum Likelihood Estimator.

4.2 Sufficient Statistics

Consider a random vector $\mathbf{X}$ of dimension n whose distribution belongs to the family $\mathcal{F} = \{F(\mathbf{x}, \theta) : \theta \in \Theta\}$. The vector $\mathbf{X}$ is of interest to us because it provides information about the true value of θ. It may happen that some of the information contained in $\mathbf{X}$ is irrelevant for the knowledge of θ, and consequently, it is convenient to eliminate it, thus simplifying the available information.

By performing this simplification, eliminating all irrelevant information from $\mathbf{X}$, we obtain another vector $\mathbf{T}$, which may have a dimension smaller than n.

Definition 24. *A **statistic** is any function $T = t(\mathbf{X})$ of the random vector $\mathbf{X}$ that represents the data to be measured in the sample. Any statistic $t(\mathbf{X})$ is a form of data reduction of the random data $\mathbf{X}$.*

If the function $t(\cdot)$ is not one-to-one, the value of $\mathbf{X}$ cannot be reconstructed from the knowledge of $\mathbf{T}$, so $\mathbf{T}$ retains only a part of the information contained in $\mathbf{X}$. The statistic $T = t(\mathbf{X})$ is **sufficient** when it retains all the relevant information for the knowledge of θ.

Definition 25. *A statistic $T = t(\mathbf{X})$ is said to be **sufficient** for θ if the distribution of $\mathbf{X}$ conditional on $T = t$ is independent of θ for all t.*

$$\mathbf{X} \mid T(\mathbf{X}) = t \rightarrow \text{independent of } \theta \text{ for all } t$$

This can be interpreted as: once the value t of T is known, the distribution of $\mathbf{X}$ is independent of θ, and therefore $\mathbf{X}$ contains no supplementary information about θ. In other words: once the value of T is known, we can forget the value of $\mathbf{X}$, since T contains all the information that $\mathbf{X}$ has about θ.

Example 31 (Trivial Example)**.** *Let $t(\mathbf{X}) = \mathbf{X}$ (we are given all the data, no reduction is done). To prove that T is sufficient, we must show that $f_{\mathbf{X}|T}(\mathbf{x} \mid t)$ does not depend on θ.*

$$f_{\mathbf{X}|T}(X_1 = x_1, \ldots, X_n = x_n \mid T(x_1, \ldots, x_n) = t)$$

For $t = (x_1, \ldots, x_n)$, this is:

$$\frac{f(X_1 = x_1, \ldots, X_n = x_n \text{ and } T(\mathbf{x}) = \mathbf{x})}{f_T(t)} = \frac{f(X_1 = x_1, \ldots, X_n = x_n)}{f(X_1 = x_1, \ldots, X_n = x_n)} = 1$$

The value 1 clearly does not depend on θ, thus $T(\mathbf{X}) = \mathbf{X}$ is trivially a sufficient statistic.

Example 32. *Let* $X_1, X_2, \ldots, X_n \overset{iid}{\sim} Be(\theta)$.

The joint point probability function is equal to:

$$p_{\mathbf{X}}(\mathbf{x};\theta) = P(X_1 = x_1, \ldots, X_n = x_n;\theta) = \prod_{i=1}^{n} P(X_i = x_i)$$

$$= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}\mathbb{I}_{\{0,1\}}(x_i)$$

$$= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\prod_{i=1}^{n}\mathbb{I}_{\{0,1\}}(x_i)$$

Let $T = t(\mathbf{X}) = \sum_{i=1}^n X_i$. *Let us prove that this statistic is sufficient for* θ. *For this, we must calculate the distribution of* $\mathbf{X} = (X_1, \ldots, X_n)$ *conditional on* $T = t$*:*

$$p_{\mathbf{X}|T}(\mathbf{x};\theta \mid t) = \frac{p_{\mathbf{X},T}(\mathbf{x}, t;\theta)}{p_T(t;\theta)}$$

The numerator of this quotient is the joint probability of $\mathbf{X}$ *and* T*:*

$$p_{\mathbf{X},T}(\mathbf{x}, t;\theta) = P(X_1 = x_1, \ldots, X_n = x_n \text{ and } T = t)$$

$$= \begin{cases} \theta^t(1-\theta)^{n-t}\prod_{i=1}^n \mathbb{I}_{\{0,1\}}(x_i) & \text{if } \sum_{i=1}^n x_i = t \\ 0 & \text{if } \sum_{i=1}^n x_i \neq t \end{cases}$$

Since $T = \sum_{i=1}^n X_i \to Bin(n,\theta)$, *the denominator is:*

$$p_T(t;\theta) = \binom{n}{t}\theta^t(1-\theta)^{n-t}$$

Thus:

$$p_{\mathbf{X}|T}(\mathbf{x};\theta \mid t) = \frac{\theta^t(1-\theta)^{n-t}\prod_{i=1}^n \mathbb{I}_{\{0,1\}}(x_i)}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{\prod_{i=1}^n \mathbb{I}_{\{0,1\}}(x_i)}{\binom{n}{t}}$$

This result is independent of θ. *Therefore,* $T = \sum_{i=1}^n X_i$ *is sufficient for* θ.

The following theorem greatly facilitates the search for sufficient statistics: (The Factorization Theorem is implied here)

Theorem 18 (Theorem of Factorization)**.** *Let* $\mathbf{X}$ *be a random vector with probability density function* $f(\mathbf{x},\theta)$, $\theta \in \Theta$. *Then, the statistic* $T = t(\mathbf{X})$ *is sufficient for* θ *if and only if there exist functions* $k_1(\cdot)$ *and* $k_2(\cdot)$ *such that:*

$$f(\mathbf{x},\theta) = k_1(t(\mathbf{x}),\theta)k_2(\mathbf{x})$$

where $k_1(t(\mathbf{x}),\theta)$ *depends on* $\mathbf{x}$ *only through the statistic* $t(\mathbf{x})$ *and the parameter* θ, *and* $k_2(\mathbf{x})$ *does not depend on* θ.

Example 33. *Let* $X_1, X_2, \ldots, X_n \overset{iid}{\sim} U[\theta_1, \theta_2]$. *The joint density function is written as:*

$$\begin{aligned} f_{\mathbf{X}}(x_1, \ldots, x_n; \theta_1, \theta_2) &= \prod_{i=1}^{n} f_{X_i}(x_i; \theta_1, \theta_2) \\ &= \prod_{i=1}^{n} \frac{1}{\theta_2 - \theta_1} \mathbb{I}_{[\theta_1, \theta_2]}(x_i) \\ &= (\theta_2 - \theta_1)^{-n} \prod_{i=1}^{n} \mathbb{I}_{[\theta_1, \theta_2]}(x_i) \end{aligned}$$

The product of indicator functions is 1 if and only if $x_i \in [\theta_1, \theta_2]$ *for all* i, *which is equivalent to* $\min\{x_i\} \geq \theta_1$ *and* $\max\{x_i\} \leq \theta_2$.

$$f_{\mathbf{X}}(x_1, \ldots, x_n; \theta_1, \theta_2) = (\theta_2 - \theta_1)^{-n} \mathbb{I}_{[\theta_1, \infty)}(\min\{x_i\}) \mathbb{I}_{(-\infty, \theta_2]}(\max\{x_i\})$$

In this form:

- $k_1(t(\mathbf{x}), \theta) = (\theta_2 - \theta_1)^{-n} \mathbb{I}_{[\theta_1, \infty)}(\min\{x_i\}) \mathbb{I}_{(-\infty, \theta_2]}(\max\{x_i\})$
- $k_2(\mathbf{x}) = 1$

Thus, $T = (\min\{X_i\}, \max\{X_i\}) = (X_{(1)}, X_{(n)})$ *(the first and the last order statistic) is a sufficient statistic for* $\theta = (\theta_1, \theta_2)$. *Note that here the sufficient statistic is a two-dimensional vector. The sufficient statistic is not just* $\min\{X_i\}$ *or just* $\max\{X_i\}$, *but the vector composed of both order statistics.*

Exponential Families

Let us assume the density function for a k-parameter exponential family is given by (3.18):

$$f(x, \boldsymbol{\theta}) = A(\boldsymbol{\theta}) h(x) \exp\left\{ \sum_{i=1}^{k} c_i(\boldsymbol{\theta}) r_i(x) \right\}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is the parameter vector.

The following theorem establishes the most important property of exponential families regarding samples.

Theorem 19. *Let* $X_1, X_2, \ldots, X_n$ *be a random sample from a distribution that belongs to a* k*-parameter exponential famil. Then the joint distribution of* $X_1, \ldots, X_n$ *also belongs to a* k*-parameter exponential family, and the sufficient statistic for* $\boldsymbol{\theta}$ *is the vector* $\mathbf{T}^* = (T_1^*, \ldots, T_k^*)$, *where:*

$$T_i^* = \sum_{j=1}^{n} r_i(X_j), \quad 1 \leq i \leq k.$$

Proof. The proof is immediate, since by (3.18), the joint density $f(\mathbf{x}, \boldsymbol{\theta}) = f(x_1, \ldots, x_n; \boldsymbol{\theta})$

is the product of the individual densities:

$$\begin{aligned} f(x_1,\ldots,x_n;\boldsymbol{\theta}) &= \prod_{j=1}^{n} f(x_j,\boldsymbol{\theta}) \\ &= \prod_{j=1}^{n}\left[A(\boldsymbol{\theta})h(x_j)\exp\left\{\sum_{i=1}^{k} c_i(\boldsymbol{\theta})r_i(x_j)\right\}\right] \\ &= (A(\boldsymbol{\theta}))^n\left(\prod_{j=1}^{n} h(x_j)\right)\exp\left\{\sum_{j=1}^{n}\sum_{i=1}^{k} c_i(\boldsymbol{\theta})r_i(x_j)\right\} \\ &= \underbrace{(A(\boldsymbol{\theta}))^n}_{\mathbf{A}^*(\boldsymbol{\theta})}\underbrace{\left(\prod_{j=1}^{n} h(x_j)\right)}_{\mathbf{h}^*(\mathbf{x})}\exp\left\{\sum_{i=1}^{k} c_i(\boldsymbol{\theta})\left(\sum_{j=1}^{n} r_i(x_j)\right)\right\} \\ &= \mathbf{A}^*(\boldsymbol{\theta})\mathbf{h}^*(\mathbf{x})\exp\left\{\sum_{i=1}^{k} c_i(\boldsymbol{\theta})T_i^*\right\} \end{aligned}$$

where $\mathbf{A}^*(\boldsymbol{\theta}) = (A(\boldsymbol{\theta}))^n$, $T_i^* = \sum_{j=1}^{n} r_i(x_j)$, and $\mathbf{h}^*(\mathbf{x}) = \prod_{j=1}^{n} h(x_j)$. This joint density has the canonical form of a k-parameter exponential family.

Therefore, applying the Factorization Theorem, the statistic $\mathbf{T}^* = (\sum_{j=1}^{n} r_1(X_j),\ldots,\sum_{j=1}^{n} r_k(X_j))$ is sufficient for $\boldsymbol{\theta}$.

This last theorem affirms that for k-parameter exponential families, regardless of the sample size n, there always exists a sufficient statistic with only k components. That is, all the information can be summarized into k random variables.

Estimators Based on Sufficient Statistics

Suppose $\mathbf{X}$ is a vector corresponding to a sample from a distribution that belongs to the family $F(\mathbf{x},\theta)$ with $\theta \in \Theta$. Suppose that $T = r(\mathbf{X})$ is a sufficient statistic for θ. Then, according to the intuitive concept we have of a sufficient statistic, to estimate a function $q(\theta)$, it should be enough to use estimators that depend *only* on T, since T contains all the information that $\mathbf{X}$ holds about the parameter θ. This is precisely what the following theorem states.

Theorem 20 (Rao–Blackwell)**.** *Let $\mathbf{X}$ be a vector from a distribution belonging to the family $F(\mathbf{x},\theta)$ with $\theta \in \Theta$. Let T be a sufficient statistic for θ and $\delta(\mathbf{X})$ be an estimator of $q(\theta)$. Define a new estimator:*

$$\delta^*(\mathbf{T}) = E(\delta(\mathbf{X}) \mid \mathbf{T}).$$

Then we have:

1. *$MSE_\theta(\delta^*) \leq MSE_\theta(\delta)$, for all $\theta \in \Theta$.*
2. *Equality in (i) holds if and only if $P_\theta(\delta^*(\mathbf{T}) = \delta(\mathbf{X})) = 1$ for all $\theta \in \Theta$.*
3. *If $\delta(\mathbf{X})$ is unbiased, then $\delta^*(\mathbf{T})$ is also unbiased.*

Proof. We start with the Mean Squared Error (MSE) of δ:

$$MSE_\theta(\delta) = E_\theta((\delta(\mathbf{X}) - q(\theta))^2)$$

We add and subtract $\delta^*(\mathbf{T})$ inside the parentheses:

$$\begin{aligned} MSE_\theta(\delta) &= E_\theta\left([(\delta^*(\mathbf{T}) - q(\theta)) + (\delta(\mathbf{X}) - \delta^*(\mathbf{T}))]^2\right) \\ &= E_\theta((\delta^*(\mathbf{T}) - q(\theta))^2) + E_\theta((\delta(\mathbf{X}) - \delta^*(\mathbf{T}))^2) \\ &\quad +2E_\theta((\delta^*(\mathbf{T}) - q(\theta))(\delta(\mathbf{X}) - \delta^*(\mathbf{T}))) \end{aligned} \tag{4.3}$$

We analyze the cross-term:

$$E_\theta((\delta^*(\mathbf{T}) - q(\theta))(\delta(\mathbf{X}) - \delta^*(\mathbf{T}))) = E_\theta[E((\delta^*(\mathbf{T}) - q(\theta))(\delta(\mathbf{X}) - \delta^*(\mathbf{T})) \mid \mathbf{T})]$$

Since $\delta^*(\mathbf{T})$ and $q(\theta)$ depend only on $\mathbf{T}$ (or are constant), they are treated as constants inside the inner expectation $E(\cdot \mid \mathbf{T})$:

$$= E_\theta[(\delta^*(\mathbf{T}) - q(\theta))E(\delta(\mathbf{X}) - \delta^*(\mathbf{T}) \mid \mathbf{T})]$$

Now, we look at the inner term:

$$E(\delta(\mathbf{X}) - \delta^*(\mathbf{T}) \mid \mathbf{T}) = E(\delta(\mathbf{X}) \mid \mathbf{T}) - \delta^*(\mathbf{T}) = \delta^*(\mathbf{T}) - \delta^*(\mathbf{T}) = 0$$

Substituting this result back, we get:

$$E_\theta((\delta^*(\mathbf{T}) - q(\theta))(\delta(\mathbf{X}) - \delta^*(\mathbf{T}))) = 0.$$

Then, 4.3 simplifies to:

$$MSE_\theta(\delta) = MSE_\theta(\delta^*) + E_\theta((\delta(\mathbf{X}) - \delta^*(\mathbf{T}))^2)$$

Since the expectation of a squared term is always non-negative, $E_\theta((\delta(\mathbf{X})-\delta^*(\mathbf{T}))^2) \geq 0$, which yields:

$$MSE_\theta(\delta) \geq MSE_\theta(\delta^*).$$

This proves part (i). Furthermore, equality holds only if $E_\theta((\delta(\mathbf{X})-\delta^*(\mathbf{T}))^2) = 0$, which is true if and only if $P_\theta(\delta(\mathbf{X}) = \delta^*(\mathbf{T})) = 1$ for all $\theta \in \Theta$. This proves part (ii).

To show (iii), suppose δ is unbiased, so $E_\theta(\delta(\mathbf{X})) = q(\theta)$. We calculate the expected value of $\delta^*(\mathbf{T})$:

$$E_\theta(\delta^*(\mathbf{T})) = E_\theta(E(\delta(\mathbf{X}) \mid \mathbf{T})) = E_\theta(\delta(\mathbf{X})) = q(\theta).$$

Thus, $\delta^*(\mathbf{T})$ is also unbiased. This proves part (iii).

Remark: The estimator $\delta^*(\mathbf{T}) = E(\delta(\mathbf{X}) \mid \mathbf{T})$ is indeed an estimator because it depends only on $\mathbf{T}$ (and therefore on $\mathbf{X}$) and **not on** θ. This is because, since $\mathbf{T}$ is a sufficient statistic, the conditional distribution of $\delta(\mathbf{X})$ given $\mathbf{T} = t$ is independent of θ. Consequently, the conditional expectation is also independent of θ.

Example 34. *Let $X_1, X_2, \ldots, X_n$ be a random sample from a $Bi(\theta, 1)$ distribution (Bernoulli). Then $\delta(\mathbf{X}) = X_1$ is an unbiased estimator of θ. A sufficient statistic for θ is $T = \sum_{i=1}^n X_i$.*

Therefore, according to the Rao–Blackwell theorem, $\delta^*(\mathbf{T}) = E(\delta(X_1, \ldots, X_n) \mid \mathbf{T})$ *will be another unbiased estimator of* θ*, and* $Var_\theta(\delta^*) \leq Var_\theta(\delta)$*. Let us calculate* $\delta^*(\mathbf{T})$*.*

Since $X_1, X_2, \ldots, X_n$ *are identically distributed, therefore,* $E(X_i \mid \mathbf{T})$ *will be independent of* i*. Thus:*

$$E(X_i \mid \mathbf{T}) = E(X_1 \mid \mathbf{T}) = \delta^*(\mathbf{T}), \quad 1 \leq i \leq n.$$

Summing over i *yields:*

$$\sum_{i=1}^{n} E(X_i \mid \mathbf{T}) = n\delta^*(\mathbf{T}).$$

But it is also true that, by the linearity of conditional expectation and the tower property $E(Y \mid Y) = Y$*:*

$$\sum_{i=1}^{n} E(X_i \mid \mathbf{T}) = E\left(\sum_{i=1}^{n} X_i \middle| \mathbf{T}\right) = E(\mathbf{T} \mid \mathbf{T}) = \mathbf{T}.$$

Equating the two results:

$$n\delta^*(\mathbf{T}) = \mathbf{T}$$

Hence:

$$\delta^*(\mathbf{T}) = \frac{\mathbf{T}}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n.$$

$$Var_\theta(\delta^*(\mathbf{T})) = \frac{\theta(1-\theta)}{n} \quad \textit{and} \quad Var_\theta(\delta(\mathbf{X})) = Var_\theta(X_1) = \theta(1-\theta).$$

Thus, $Var_\theta(\delta^*(\mathbf{T})) \leq Var_\theta(\delta(\mathbf{X}))$*, illustrating the variance reduction achieved by the Rao-Blackwell theorem.*

4.3 Complete Statistics

So far, we have seen that by taking unbiased estimators of a function $\beta(\theta)$ based on sufficient statistics, the estimation efficiency is improved (by the Rao-Blackwell Theorem). What we do not yet know is whether there can be more than one unbiased estimator based on a given sufficient statistic $\mathbf{T}$. We will see that under certain conditions, there is only one.

Definition 26. *A statistic* $\mathbf{T} = T(\mathbf{X})$ *is* ***complete*** *for* θ *when the following holds:*

$$E_\theta[g(\mathbf{T})] = 0 \textit{ for all } \theta \in \Theta \quad \Longrightarrow \quad P_\theta(g(\mathbf{T}) = 0) = 1.$$

Proving completeness by definition can only be achieved in some simple cases, such as the Binomial, Poisson, or Uniform $[0, \theta]$ distributions. In most other cases, it is often a quite complex task. We will study a family of distributions (the exponential family) where determining the complete statistic is a task that is greatly simplified.

Exponential Families

Complete statistics exist in a large and important class of model families $\mathcal{F}$, called exponential families , where the density function $f(\mathbf{x}; \boldsymbol{\theta})$ is of the form:

$$f(\mathbf{x}, \boldsymbol{\theta}) = h(\mathbf{x})A(\boldsymbol{\theta}) \exp\left\{\sum_{j=1}^{k} c_j(\boldsymbol{\theta}) r_j(\mathbf{x})\right\}$$

An exponential family is called **full-rank** (or canonical) if:

1. $r_1(\mathbf{x}), r_2(\mathbf{x}), \ldots, r_k(\mathbf{x})$ satisfy no linear restrictions (they are linearly independent).
2. $c_1(\boldsymbol{\theta}), c_2(\boldsymbol{\theta}), \ldots, c_k(\boldsymbol{\theta})$ satisfy no linear restrictions on the parameters.
3. The set $\Lambda = \{c(\boldsymbol{\theta}) = (c_1(\boldsymbol{\theta}), c_2(\boldsymbol{\theta}), \ldots, c_k(\boldsymbol{\theta})) : \ \theta \in \Theta\}$ contains an open set (an k-dimensional sphere) in $\mathbb{R}^k$.

Theorem 21. *In a full-rank exponential family, the statistic $\mathbf{T}(\mathbf{X}) = (T_1(\mathbf{X}), T_2(\mathbf{X}), \ldots, T_k(\mathbf{X}))$, where $T_j(\mathbf{X}) = \sum_{i=1}^n r_j(X_i)$, is a **complete** statistic (and therefore also a minimal sufficient statistic).*

Example 35. *Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} Be(\theta)$, with $\Theta = (0,1)$. The individual density function is:*

$$f_X(x;\theta) = \theta^x(1-\theta)^{1-x}\mathbb{I}_{\{0,1\}}(x)\mathbb{I}_{(0,1)}(\theta)$$

The sample density function can be written as:

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x};\theta) &= \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\prod \mathbb{I}_{\{0,1\}}(x_i)\mathbb{I}_{(0,1)}(\theta) \\ &= \underbrace{\prod \mathbb{I}_{\{0,1\}}(x_i)}_{h(\mathbf{x})}\underbrace{(1-\theta)^n\mathbb{I}_{(0,1)}(\theta)}_{c(\theta)}\exp\left\{\left[\ln\left(\frac{\theta}{1-\theta}\right)\right]\sum_{i=1}^n x_i\right\} \end{aligned}$$

This is a one-parameter exponential family ($k = 1$).

- $r_1(\mathbf{x}) = \sum_{i=1}^n x_i$
- $c_1(\theta) = \ln(\theta/(1-\theta))$

*Since this is a full-rank exponential family (as $\Theta = (0,1)$ is an open interval and then Λ too), $T = \sum_{i=1}^n X_i$ is a **complete** statistic for θ.*

If a sufficient and complete statistic is known, a method exists to calculate UMVUE estimators.

Theorem 22. *Let $\mathbf{T}$ be a sufficient and complete statistic for θ. Then, given a function $q(\theta)$, we have:*

1. *There is **at most one** unbiased estimator of $q(\theta)$ based on $\mathbf{T}$.*
2. *If $\delta^*(\mathbf{T})$ is an unbiased estimator of $q(\theta)$, then $\delta^*(\mathbf{T})$ is the **UMVUE** (Uniformly Minimum Variance Unbiased Estimator).*
3. *If $\delta(\mathbf{X})$ is any unbiased estimator of $q(\theta)$, then $\delta^*(\mathbf{T}) = E_\theta(\delta(\mathbf{X}) \mid \mathbf{T})$ is the **UMVUE**.*

Remark 11. *Regarding (1): "At most one" implies that if another unbiased estimator $\delta'(\mathbf{T})$ exists for $\beta(\theta)$, then $P_\theta(\delta^*(\mathbf{T}) = \delta'(\mathbf{T})) = 1$. Regarding (3): This gives us the recipe to obtain $\delta^*(\mathbf{T})$ starting from any unbiased estimator $\delta\mathbf{X})$ of $q(\theta)$. $\delta^*(\mathbf{T})$ is obtained using the Rao-Blackwell method, with the key difference that we now condition on a statistic that is not only sufficient but also **complete**.*

Remark 12. *As a conclusion to the chapter, if we are looking for optimal estimators in the sense that they have the least possible variance among the set of estimators, we must start with an unbiased estimator and a sufficient and complete statistic, and construct the UMVUE (Uniformly Minimum Variance Unbiased Estimator). Although this estimator is the UMVUE, it does not always reach the Rao-Cramér Lower Bound (RCLB). On the other hand, if we have an unbiased estimator that reaches the Rao-Cramér Bound, we are certain that it is the UMVUE, since the bound guarantees that there can be no unbiased estimators with a variance lower than that limit.*

Example 36. *Let $X_1, X_2, \ldots, X_n$ be a simple random sample of size n drawn from a population $N(\mu, \sigma^2)$, with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ unknown.*
$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ are UMVU for μ and σ^2.
We known that the Normal distribution is a two Exponential Family

$$f(x; \mu, \sigma^2) = \exp\left\{c_1(\theta)T_1(x) + c_2(\theta)T_2(x) - A(\boldsymbol{\theta})\right\} h(x)$$

Where the canonical parameters are $c_1(\theta) = \mu/\sigma^2$ and $c_2(\theta) = -1/(2\sigma^2)$, and the base statistics are $T_1(x) = x$ and $T_2(x) = x^2$.
The sufficient statistic $T(\mathbf{X}) = (\sum X_i, \sum X_i^2)$ is complete because the space Λ is an open set in $\mathbb{R}^2$.

$$\Lambda = \{(\eta_1, \eta_2) \mid \eta_1 \in \mathbb{R}, \eta_2 < 0\}$$

Since Λ contains an open, two-dimensional rectangle, the Completeness Theorem for Exponential Families guarantees that the sufficient statistic is complete.
If T is a complete and sufficient statistic for $\boldsymbol{\theta}$, and if $W = g(T)$ is an unbiased estimator of θ (i.e., $E[W] = \theta$), then W is the Uniformly Minimum Variance Unbiased Estimator (UMVUE) for θ.
*Since $\bar{X}$ and S^2 are **unbiased functions** of the complete and sufficient statistic , the theorem guarantees that:*

1. *$\bar{X}$ is the UMVUE for μ.*
2. *S^2 is the UMVUE for σ^2.*

Example 37. *Let $X_1, \ldots, X_n$ be an i.i.d. random sample from a Bernoulli distribution with probability of success p, denoted $X_i \sim$ Bernoulli(p), where $0 < p < 1$.*
The parameter function of interest is $q(p) = p^2$.
The sum of Bernoulli trials $T = \sum_{i=1}^{n} X_i$ follows a Binomial distribution $T \sim$ Binomial(n, p). Since the Bernoulli/Binomial distribution is in the one-parameter Exponential Family and T is the natural sufficient statistic, T is the Complete and Sufficient statistic for p.
We seek an unbiased estimator $h(T)$ such that $E[h(T)] = p^2$. We use the method of factorization for the moments of T.

$E[T] = np$

$E[T(T-1)] = E[T^2] - E[T]$. *Since* $Var(T) = np(1-p)$, *we have* $E[T^2] = np(1-p) + (np)^2 = np - np^2 + n^2p^2$.

$$E[T(T-1)] = (np - np^2 + n^2p^2) - np = n^2p^2 - np^2 = p^2(n^2 - n) = p^2n(n-1)$$

Therefore, to isolate p^2, we define the UMVUE as:

$$\widehat{p^2}_{UMVUE} = \frac{T(T-1)}{n(n-1)}$$

Since this estimator is unbiased and is a function of the Complete Sufficient Statistic T, by the Lehmann-Scheffé Theorem, it is the UMVUE for p^2.

The calculation of $Var[\widehat{p^2}_{UMVUE}]$ is tedious, but the final, known result is:

$$Var_{UMVUE}(p^2) = \frac{4p^3(1-p)}{n} + \frac{2p^2(1-p)^2}{n(n-1)}$$

For the Bernoulli distribution, the Fisher Information for a single observation $I_1(p)$ is $\frac{1}{p(1-p)}$. The Fisher Information for the sample of size n is:

$$I_n(p) = nI_1(p) = \frac{n}{p(1-p)}$$

The CRLB for an unbiased estimator of $g(p)$ is $CRLB(q(p)) = \frac{[q'(p)]^2}{I_n(p)}$.

1. *Derivative of the function: $q(p) = p^2 \implies q'(p) = 2p$.*
2. *Applying the CRLB formula:*

$$CRLB(p^2) = \frac{(2p)^2}{n/(p(1-p))} = \frac{4p^2}{n} \cdot p(1-p) = \frac{4p^3(1-p)}{n}$$

We compare the variance of the UMVUE and the CRLB:

$$Var_{UMVUE}(p^2) = \frac{4p^3(1-p)}{n} + \frac{2p^2(1-p)^2}{n(n-1)}$$

$$CRLB(p^2) = \frac{4p^3(1-p)}{n}$$

For any sample size $n > 1$ and any parameter $0 < p < 1$, the second term of the UMVUE variance is strictly positive:

$$\frac{2p^2(1-p)^2}{n(n-1)} > 0$$

Therefore, the UMVUE variance is strictly greater than the CRLB:

$$Var_{UMVUE}(p^2) > CRLB(p^2)$$

This is a classic example of an unbiased estimator of minimum variance (UMVUE) that fails to achieve the Rao-Cramér Lower Bound, even though all regularity conditions are satisfied. This occurs because the necessary condition for equality in the CRLB (the proportionality between the estimator and the score function) is not met for the UMVUE $\widehat{p^2}$.

Chapter 5

The Theory of Regression.

Regression analysis serves as a statistical tool for examining the dependence between a dependent variable Y (the response) and an explanatory variable X (the covariate). The explanatory variable X is also known as a regressor, or predictor.

A fundamental method for summarizing the statistical link between X and Y is through the **conditional mean function, $r(x)$**, which represents the expected value of the response variable Y when the predictor variable takes the specific value x:

$$r(x) = \mathbb{E}(Y|X = x) = \int y f(y|x)\, dy \tag{5.1}$$

where $f(y|x)$ denotes the conditional density. The primary objective in this field is to derive an estimate for the regression function $r(x)$ using an observed dataset comprising n paired measurements: $(Y_1, X_1), \ldots, (Y_n, X_n) \sim F_{X,Y}$.

In this discussion, we adopt a parametric framework, specifically postulating structures for the prediction function $g(X)$ that minimizes the Mean Squared Error (MSE):

$$MSE(g) = \mathbb{E}[(Y - g(X))^2]$$

5.1 Best Predictor

5.1.1 Best Constant Predictor (BCP)

The predictor is restricted to $g(X) = c$.

Derivation for $\hat{c}$

Minimize $MSE(c) = \mathbb{E}[(Y - c)^2]$.

$$\frac{d}{dc} MSE(c) = \mathbb{E}\left[\frac{\partial}{\partial c}(Y - c)^2\right] = \mathbb{E}[-2(Y - c)] = 0$$

$$\mathbb{E}[Y] - c = 0 \quad \Rightarrow \quad \tilde{c} = \mathbb{E}[Y]$$

$$\tilde{Y}_{\text{BCP}} = \mathbb{E}[Y]$$

Calculation of Minimum MSE

Substitute $\tilde{c}$ back into the MSE formula:

$$MSE(\tilde{Y}_{\text{BCP}}) = \mathbb{E}[(Y - \mathbb{E}[Y])^2]$$

$$MSE(\tilde{Y}_{\text{BCP}}) = \text{Var}(Y)$$

5.1.2 Best Linear Predictor (BLP)

The predictor is restricted to $g(X) = a + bX$. This framework is often adopted when $r(x)$ is assumed to be linear.

Derivation for $\tilde{a}$ and $\tilde{b}$

We minimize $MSE(a, b) = \mathbb{E}[(Y - a - bX)^2]$.

Partial Derivative w.r.t. a:

$$\frac{\partial MSE}{\partial a} = \mathbb{E}[-2(Y - a - bX)] = 0$$

$$a = \mathbb{E}[Y] - b\mathbb{E}[X]$$

Partial Derivative w.r.t. b:

$$\frac{\partial MSE}{\partial b} = \mathbb{E}[-2X(Y - a - bX)] = 0$$

Substitute a and simplify:

$$0 = \mathbb{E}[X(Y - (\mathbb{E}[Y] - b\mathbb{E}[X]) - bX)]$$

$$0 = \text{Cov}(X, Y) - b \cdot \text{Var}(X)$$

$$\tilde{b} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

$$\tilde{Y}_{\text{BLP}} = \mathbb{E}[Y] + \frac{\text{Cov}(X, Y)}{\text{Var}(X)}(X - \mathbb{E}[X])$$

Calculation of Minimum MSE

$$MSE(\tilde{Y}_{\text{BLP}}) = \text{Var}(Y) - \frac{(\text{Cov}(X, Y))^2}{\text{Var}(X)}$$

Using the correlation coefficient ρ, this simplifies to:

$$MSE(\tilde{Y}_{\text{BLP}}) = \text{Var}(Y)(1 - \rho^2)$$

5.1.3 Best General Predictor (BP)

The predictor $g(X)$ is any measurable function of X. This optimal predictor is the **Conditional Mean Function** $r(x)$ itself.

Derivation for $\tilde{g}(X)$

We minimize $MSE(g) = \mathbb{E}[(Y - g(X))^2]$. Using the Law of Total Expectation: $MSE(g) = \mathbb{E}_X[\mathbb{E}[(Y - g(X))^2|X]]$. For a fixed value $X = x$, the inner term $\mathbb{E}[(Y - c)^2|X = x]$ is minimized when c is the conditional mean.

$$\tilde{g}(x) = \mathbb{E}[Y|X = x] = r(x)$$

$$\tilde{Y}_{\text{BP}} = \mathbb{E}[Y|X]$$

This confirms that the regression function $r(x)$ defined in equation (5.1) is the optimal predictor under the MSE criterion.

Calculation of Minimum MSE

Substitute $\tilde{Y}_{\text{BP}}$ back into the MSE formula:

$$MSE(\tilde{Y}_{\text{BP}}) = \mathbb{E}[(Y - \mathbb{E}[Y|X])^2]$$

Using the variance decomposition formula $\text{Var}(Y) = \mathbb{E}[\text{Var}(Y|X)] + \text{Var}(\mathbb{E}[Y|X])$, the MSE simplifies to:

$$MSE(\tilde{Y}_{\text{BP}}) = \mathbb{E}[\text{Var}(Y|X)]$$

This term, $\mathbb{E}[\text{Var}(Y|X)]$, represents the **expected unexplained variance** of Y given X.

5.1.4 Summary and Error Hierarchy

Predictor	**Constraint**	**Optimal Predictor ($\tilde{Y}$)**	**Minimum MSE**
BCP	$g(X) = c$	$\mathbb{E}[Y]$	$\text{Var}(Y)$
BLP	$g(X) = a + bX$	$\mathbb{E}[Y] + \tilde{b}(X - \mathbb{E}[X])$	$\text{Var}(Y) \cdot (1 - \rho^2)$
BP	$g(X)$ any function	$\mathbb{E}[Y\|X]$	$\mathbb{E}[\text{Var}(Y\|X)]$

Error Relationship (Hierarchy):

$$MSE(\tilde{Y}_{\text{BP}}) \leq MSE(\tilde{Y}_{\text{BLP}}) \leq MSE(\tilde{Y}_{\text{BCP}})$$

The most flexible predictor (BP) always yields the lowest or equal error. The BLP is only equal to the BP if the conditional expectation $\mathbb{E}[Y|X]$ (the true regression function $r(x)$) is, in fact, a linear function of X.

5.2 Simple Linear Regression

In this initial discussion, we adopt a parametric framework, specifically postulating that the function r follows a linear structure. Later chapters will delve into nonparametric estimation techniques when we do not consider any restriccion for the funcion r.

The most basic configuration of regression occurs when the covariates X_i are one-dimensional and the conditional mean function $r(x)$ is assumed to follow a straight line:

$$r(x) = \beta_0 + \beta_1 x. \tag{5.2}$$

Definition 27 (The Simple Linear Regression Model)**.** *The linear relationship between the variables, incorporating an unobserved random disturbance, is specified as:*

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{5.3}$$

where the error term has a conditional mean of zero $\mathbb{E}(\varepsilon_i|X_i) = 0$ and the conditional variance, $\mathbb{V}(\varepsilon_i|X_i) = \sigma^2$, is constant for all x (homoscedasticity).

This structure is termed the simple linear regression model. The unknown parameters to be estimated from the available data are the intercept (β_0), the slope (β_1), and the residual variance (σ^2).

5.3 Least Squares Estimation in Simple Linear Regression

The fit of the linear model to the observed data is commonly assessed by quantifying the residual variation.

Let $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the estimated values for the parameters β_0 and β_1. The estimated linear relationship, $\hat{r}(x)$, is given by:

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x.$$

The predicted values (or fitted values) for the response are $\hat{Y}_i = \hat{r}(X_i)$. The residuals are defined as the vertical discrepancies between the observed data points and the estimated line:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i).$$

Residual Sums of Squares (RSS) The residual sums of squares (or RSS) measures how effectively the estimated straight line matches the data. It is defined as the sum of the squared residuals ($\hat{\varepsilon}_i = Y_i - \hat{Y}_i$) for all n data points:

$$\text{RSS} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2.$$

Definition 28 (Least Squares Estimates)**.** *The least squares estimates (LSE), denoted $\hat{\beta}_0$ and $\hat{\beta}_1$, are the specific values of the intercept and slope that minimize the Residual Sums of Squares (RSS).*

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\textit{argmin}} \left[\sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2 \right]$$

Theorem 23. *Given a dataset $(Y_1, X_1), \ldots, (Y_n, X_n)$, the least squares estimates for the slope (β_1) and the intercept (β_0) of the simple linear regression model are uniquely determined by:*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \tag{5.4}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{5.5}$$

where $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ and $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$ are the sample means of the covariate and the response, respectively.

Proof of Theorem 23. Our goal is to find the values of β_0 and β_1 that minimize the RSS function:

$$\text{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

To find the minimum, we take the partial derivatives of RSS with respect to β_0 and β_1 and set them equal to zero (the first-order conditions).

$$\frac{\partial \text{RSS}}{\partial \beta_0} = \sum_{i=1}^{n} 2(Y_i - \beta_0 - \beta_1 X_i)(-1) = 0$$

Dividing by -2 and rearranging the terms yields:

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \beta_0 - \sum_{i=1}^{n} \beta_1 X_i = 0$$

Since $\sum_{i=1}^{n} \beta_0 = n\beta_0$ and β_1 is a constant:

$$\sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i$$

then

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$$

Rearranging to solve for $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

This proves Equation (5.5). This equation confirms that the estimated line must pass through the sample mean point $(\bar{X}, \bar{Y})$.

Now the Partial Derivative with respect to β_1

$$\frac{\partial \text{RSS}}{\partial \beta_1} = \sum_{i=1}^{n} 2(Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0$$

Substituting the expression for β_0 from the first step ($\beta_0 = \bar{Y} - \beta_1 \bar{X}$):

$$\sum_{i=1}^{n} X_i(Y_i - (\bar{Y} - \beta_1 \bar{X}) - \beta_1 X_i) = 0$$

$$\sum_{i=1}^{n} X_i(Y_i - \bar{Y} - \beta_1 X_i + \beta_1 \bar{X}) = 0$$

$$\sum_{i=1}^{n} X_i((Y_i - \bar{Y}) - \beta_1 (X_i - \bar{X})) = 0$$

Expanding the sum:

$$\sum_{i=1}^{n} X_i(Y_i - \bar{Y}) - \beta_1 \sum_{i=1}^{n} X_i(X_i - \bar{X}) = 0$$

We need to simplify the term $\sum X_i(Y_i - \bar{Y})$. Since $\sum \bar{X}(Y_i - \bar{Y}) = \bar{X} \sum (Y_i - \bar{Y}) = \bar{X} \cdot 0 = 0$, we can substitute $\sum X_i(Y_i - \bar{Y})$ with $\sum (X_i - \bar{X})(Y_i - \bar{Y})$ (due to properties of covariance).

Also, $\sum X_i(X_i - \bar{X}) = \sum(X_i - \bar{X})^2$ (due to properties of variance). Substituting these simplified terms:

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) - \hat{\beta}_1 \sum_{i=1}^{n}(X_i - \bar{X})^2 = 0$$

Solving for $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

This proves Equation (5.4).

Estimation of the Variance (σ^2) An unbiased estimate of the error variance σ^2 is the mean squared error (MSE), which uses the Residual Sums of Squares (RSS) and corrects for the degrees of freedom used to estimate the two parameters (β_0 and β_1):

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{\text{RSS}}{n-2}$$

The divisor $n-2$ is used because two degrees of freedom are lost in estimating $\hat{\beta}_0$ and $\hat{\beta}_1$.

5.3.1 Connection between Least Squares and Maximum Likelihood

Up to this point, the determination of the Least Squares Estimates (LSE) did not necessitate any assumptions regarding the probability distribution of the error term ε_i. We now introduce a specific distributional assumption:

$$\text{Suppose we assume that } \varepsilon_i \mid X_i \sim N(0, \sigma^2).$$

This assumption implies that the response variable Y_i, conditional on the predictor X_i, also follows a Normal distribution:

$$Y_i \mid X_i \sim N(\mu_i, \sigma^2)$$

where the conditional mean is $\mu_i = \beta_0 + \beta_1 X_i$.

The joint probability density function for the observed data (X_i, Y_i) is given by the product of the marginal density of X and the conditional density of Y given X:

$$f(X_i, Y_i) = f_X(X_i) f_{Y|X}(Y_i|X_i)$$

Assuming the observations are independent and identically distributed (i.i.d.), the joint likelihood function $\mathcal{L}$ for all n observations is:

$$\mathcal{L}(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} f(X_i, Y_i) = \left(\prod_{i=1}^{n} f_X(X_i)\right) \times \left(\prod_{i=1}^{n} f_{Y|X}(Y_i|X_i)\right)$$

We can express this as $\mathcal{L} = \mathcal{L}_1 \times \mathcal{L}_2$.

The first term, $\mathcal{L}_1 = \prod_{i=1}^{n} f_X(X_i)$, is independent of the parameters of interest, β_0 and β_1. Therefore, maximizing the full likelihood $\mathcal{L}$ is equivalent to maximizing the second term, $\mathcal{L}_2$, which is known as the conditional likelihood given X:

$$\mathcal{L}_2 \equiv \mathcal{L}(\beta_0, \beta_1, \sigma^2 \mid \mathbf{X}) = \prod_{i=1}^{n} f_{Y|X}(Y_i|X_i)$$

Since $Y_i \mid X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$, the conditional density function is:

$$f_{Y|X}(Y_i|X_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(Y_i - (\beta_0 + \beta_1 X_i))^2\right)$$

Substituting this into $\mathcal{L}_2$ and ignoring the constant factor $(2\pi)^{-n/2}$:

$$\mathcal{L}_2 \propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - (\beta_0 + \beta_1 X_i))^2\right)$$

The conditional log-likelihood, $\ell(\beta_0, \beta_1, \sigma^2) = \log(\mathcal{L}_2)$, is obtained by taking the natural logarithm

$$\ell(\beta_0, \beta_1, \sigma^2) = -n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - (\beta_0 + \beta_1 X_i))^2$$

To find the Maximum Likelihood Estimators (MLEs) for β_0 and β_1, we must maximize $\ell(\beta_0, \beta_1, \sigma^2)$. We observe that the only term depending on β_0 and β_1 is the summation term:

$$\sum_{i=1}^{n}(Y_i - (\beta_0 + \beta_1 X_i))^2.$$

This sum of squares term is exactly the Residual Sums of Squares (RSS).

$$\text{Maximizing } \ell(\beta_0, \beta_1, \sigma^2) \iff \text{Minimizing } \sum_{i=1}^{n}(Y_i - (\beta_0 + \beta_1 X_i))^2 = \text{RSS}$$

Therefore, we prove that under the supplementary assumption that the errors (ε_i) are independently and identically Normally distributed, the Least Squares Estimator for the regression coefficients $(\hat{\beta}_0, \hat{\beta}_1)$ is also the Maximum Likelihood Estimator.

We can also find the MLE for σ^2 by maximizing the log-likelihood ℓ with respect to σ^2. This yields:

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{\text{RSS}}{n}$$

This estimator differs slightly from the unbiased estimator $\hat{\sigma}^2_{\text{Unbiased}} = \frac{\text{RSS}}{n-2}$ presented earlier, as it uses n instead of $n-2$ in the denominator. In statistical practice, the unbiased estimator $(\hat{\sigma}^2_{\text{Unbiased}})$ is generally preferred.

5.3.2 Properties of the Least Squares Estimators

We now examine the distributional characteristics, standard errors, and the limiting behavior of the least squares estimators. In regression analysis, it is standard practice to focus on the properties of these estimators conditional on the observed covariate values, $\mathbf{X}^n = (X_1, \ldots, X_n)$. Consequently, the mean and variance are stated as conditional moments.

Theorem 24 (Conditional Moments of LSE)**.** *Let $\hat{\boldsymbol{\beta}}^T = (\hat{\beta}_0, \hat{\beta}_1)$ denote the Least Squares Estimators. Assuming the errors are uncorrelated with mean zero and common variance σ^2 (i.e., $\mathbb{E}(\varepsilon_i|\mathbf{X}^n) = 0$ and $\mathbb{V}(\varepsilon_i|\mathbf{X}^n) = \sigma^2$):*

1. *Conditional Mean (Unbiasedness): The estimators are conditionally unbiased:*

$$\mathbb{E}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}^n) = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

2. *Conditional Variance-Covariance Matrix: The matrix representing the variances and covariance of the estimators is:*

$$\mathbb{V}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}^n) = \frac{\sigma^2}{n s_X^2} \begin{pmatrix} \bar{X^2} & -\bar{X} \\ -\bar{X} & 1 \end{pmatrix}$$

where $s_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ *is the sample variance of* X *and* $\bar{X^2} = \frac{1}{n}\sum_{i=1}^n X_i^2$.

The estimated standard errors (SE) of the estimators are obtained by taking the square roots of the corresponding diagonal entries of the conditional variance matrix $\mathbb{V}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}^n)$ and replacing the unknown population variance σ with its unbiased estimate, $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ (where $\hat{\sigma}^2 = \text{RSS}/(n-2)$).

1. Standard Error of $\hat{\beta}_0$:

$$\hat{\text{se}}(\hat{\beta}_0) = \frac{\hat{\sigma}}{\sqrt{n} s_X} \sqrt{\bar{X^2}}$$

2. Standard Error of $\hat{\beta}_1$:

$$\hat{\text{se}}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{n} s_X}$$

We often use the abbreviated notation $\hat{\text{se}}(\hat{\beta}_0)$ and $\hat{\text{se}}(\hat{\beta}_1)$ for simplicity.

Asymptotic Properties and Inference

Theorem 25 (Asymptotic Properties)**.** *Under certain regularity conditions (e.g., moments exist and the sample variance of* X *converges to a positive limit):*

1. *Consistency: The estimators converge in probability to the true parameter values:*

$$\hat{\beta}_0 \xrightarrow{P} \beta_0 \quad \textit{and} \quad \hat{\beta}_1 \xrightarrow{P} \beta_1.$$

2. *Asymptotic Normality: The standardized estimators converge in distribution to the Standard Normal distribution (*$N(0,1)$*). This is a consequence of the Central Limit Theorem applied to the LSE formulas:*

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{se}(\hat{\beta}_0)} \xrightarrow{D} N(0,1) \quad \textit{and} \quad \frac{\hat{\beta}_1 - \beta_1}{\hat{se}(\hat{\beta}_1)} \xrightarrow{D} N(0,1).$$

3. *Confidence Intervals: Approximate* $1-\alpha$ *confidence intervals for the regression coefficients* β_0 *and* β_1 *can be constructed using the quantiles of the Standard Normal*

distribution ($z_{\alpha/2}$):

$$CI(\beta_0) \approx \hat{\beta}_0 \pm z_{\alpha/2}\hat{se}(\hat{\beta}_0)$$
$$CI(\beta_1) \approx \hat{\beta}_1 \pm z_{\alpha/2}\hat{se}(\hat{\beta}_1)$$

(Note: For small samples, the Student's t-distribution with $n-2$ degrees of freedom is typically used instead of the Normal distribution.)

Prediction

Suppose we have fit a regression model $\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ using an observed dataset $(X_1, Y_1), \ldots, (X_n, Y_n)$. If we obtain a new observation of the covariate, $X = x^*$, and wish to forecast the corresponding outcome Y^*, the point prediction is given by substituting the new covariate value into the estimated model:

$$\hat{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*.$$

5.4 Multiple Regression Model

We now consider the scenario where the response variable Y is modeled as a function of multiple covariates. Suppose the predictor is a vector $\mathbf{X}_i$ of length k. The observed data consist of n independent observations of the form:

$$(\mathbf{Y}_1, \mathbf{X}_1), \ldots, (\mathbf{Y}_i, \mathbf{X}_i), \ldots, (\mathbf{Y}_n, \mathbf{X}_n)$$

where each covariate vector for the i-th observation is:

$$\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ik})$$

The linear regression mode with multiple predictors is expressed as:

$$Y_i = \sum_{j=1}^{k} \beta_j X_{ij} + \varepsilon_i \quad \text{for } i = 1, \ldots, n \tag{5.6}$$

where $\mathbb{E}(\varepsilon_i \mid X_{i1}, \ldots, X_{ik}) = 0$ (the error term has zero conditional mean).

To include an intercept term (β_0) in the model, we typically define the first component of the covariate vector for all observations as a constant: $X_{i1} = 1$ for all $i = 1, \ldots, n$. In this structure, the number of parameters is k, including the intercept.

The model is most conveniently expressed using matrix notation.

Response Vector (Y): The collection of all observed outcomes is written as a column vector of dimension $n \times 1$:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

Design Matrix (X): The covariates are collected into an $n \times k$ matrix, where each row represents one observation and each column corresponds to a different covariate (or parameter):

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1k} \\ X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nk} \end{pmatrix}$$

Parameter Vector ($\boldsymbol{\beta}$) and Error Vector ($\boldsymbol{\varepsilon}$): The parameters and the random errors are column vectors of dimension $k \times 1$ and $n \times 1$, respectively:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

The entire system of n equations in Equation 5.6 can be concisely written as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{5.7}$$

The estimated regression function is $\hat{r}(\mathbf{x}) = \sum_{j=1}^{k} \hat{\beta}_j x_j$. The vector of residuals is $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$.

An unbiased estimator of the error variance σ^2 is the Mean Squared Error (MSE):

$$\hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{1}{n-k} \hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}}$$

Here, $n - k$ are the degrees of freedom, since k parameters were estimated.

For the following theorem, we rely on the standard Gauss-Markov assumptions for the linear model :

1. The linear model is correctly specified.
2. The errors have zero conditional mean: $\mathbb{E}(\varepsilon_i \mid \mathbf{X}) = 0$.
3. The errors are homoscedastic and uncorrelated: $\mathbb{V}(\varepsilon_i \mid \mathbf{X}) = \sigma^2$ and $\mathbb{C}ov(\varepsilon_i, \varepsilon_j \mid \mathbf{X}) = 0$ for $i \neq j$.

Theorem 26 (Least Squares Estimator in Multiple Regression)**.** *Assuming that the $(k \times k)$ matrix $\mathbf{X}^T\mathbf{X}$ is invertible (i.e., the covariates are not perfectly collinear), the Least Squares Estimator for the coefficient vector $\boldsymbol{\beta}$ is:*

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Under the Gauss-Markov assumptions ($\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathbb{V}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$), the conditional variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is:

$$\mathbb{V}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

Under the assumption of Normal errors, or asymptotically:

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

Proof.

The Multiple Linear Regression Model is given in matrix notation as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{Y}$ is $n \times 1$, $\mathbf{X}$ is $n \times k$, $\boldsymbol{\beta}$ is $k \times 1$, and $\boldsymbol{\varepsilon}$ is $n \times 1$. The Least Squares Estimator $\hat{\boldsymbol{\beta}}$ is the vector of coefficients that minimizes the Residual Sums of Squares (RSS).

The RSS in matrix form is:

$$\text{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

Using the properties of matrix transposition,

$$\text{RSS}(\boldsymbol{\beta}) = \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T(\mathbf{X}\boldsymbol{\beta}) - (\mathbf{X}\boldsymbol{\beta})^T\mathbf{Y} + (\mathbf{X}\boldsymbol{\beta})^T(\mathbf{X}\boldsymbol{\beta})$$

Since $\mathbf{Y}^T(\mathbf{X}\boldsymbol{\beta})$ is a scalar, its transpose is equal to itself: $(\mathbf{X}\boldsymbol{\beta})^T\mathbf{Y} = \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y}$. Thus, the two middle terms are equal:

$$\text{RSS}(\boldsymbol{\beta}) = \mathbf{Y}^T\mathbf{Y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

We find the minimum by taking the derivative (gradient) of $\text{RSS}(\boldsymbol{\beta})$ with respect to the vector $\boldsymbol{\beta}$ and setting it equal to the zero vector ($\mathbf{0}$). We use the following vector calculus rules:

$$\frac{\partial(\boldsymbol{\beta}^T\mathbf{A})}{\partial\boldsymbol{\beta}} = \mathbf{A} \quad \text{and} \quad \frac{\partial(\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = 2\mathbf{A}\boldsymbol{\beta} \quad (\text{if } \mathbf{A} \text{ is symmetric})$$

Since $\mathbf{X}^T\mathbf{X}$ is symmetric:

$$\frac{\partial\text{RSS}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = \mathbf{0} - 2(\mathbf{X}^T\mathbf{Y}) + 2(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta}$$

Setting the gradient to zero:

$$-2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}$$

Rearranging the terms:

$$2(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\beta}} = 2\mathbf{X}^T\mathbf{Y}$$
$$(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}$$

Assuming $\mathbf{X}^T\mathbf{X}$ is invertible (i.e., there is no perfect multicollinearity), we premultiply both sides by $(\mathbf{X}^T\mathbf{X})^{-1}$:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

We rely on the Gauss-Markov assumptions, specifically $\mathbb{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}$ and $\mathbb{V}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \sigma^2\mathbf{I}$.

First, we must confirm that $\hat{\boldsymbol{\beta}}$ is unbiased. Substitute the true model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ into the estimator formula:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}$$

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}$$

Taking the conditional expectation: $\mathbb{E}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \boldsymbol{\beta} + \mathbf{0} = \boldsymbol{\beta}$.

The variance of a random vector $\mathbf{W}$ is $\mathbb{V}(\mathbf{W}) = \mathbb{E}[(\mathbf{W} - \mathbb{E}[\mathbf{W}])(\mathbf{W} - \mathbb{E}[\mathbf{W}])^T]$. Using the result $\hat{\boldsymbol{\beta}} - \mathbb{E}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$:

$$\mathbb{V}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) = \mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mid \mathbf{X}\right]$$

Substituting $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}$:

$$\mathbb{V}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) = \mathbb{E}\left[\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\right)\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\right)^T \mid \mathbf{X}\right]$$

Using the property $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$:

$$\mathbb{V}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) = \mathbb{E}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \mid \mathbf{X}\right]$$

Since $\mathbf{X}$ is fixed conditionally, and $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T \mid \mathbf{X}) = \mathbb{V}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \sigma^2\mathbf{I}$:

$$\mathbb{V}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$

Factoring out the scalar σ^2:

$$\mathbb{V}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}$$

Since $(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X}) = \mathbf{I}$:

$$\mathbb{V}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

If we add the explicit assumption that the errors follow a Normal distribution, $\boldsymbol{\varepsilon} \mid \mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, then the LSE is a linear combination of Normal random variables (the $\mathbf{Y}$ vector):

$$\hat{\boldsymbol{\beta}} = \underbrace{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T}_{\text{Constant Matrix}}\mathbf{Y}$$

A linear transformation of a multivariate Normal vector is also multivariate Normal. Therefore, conditionally on $\mathbf{X}$:

$$\hat{\boldsymbol{\beta}} \mid \mathbf{X} \sim N(\mathbb{E}[\hat{\boldsymbol{\beta}}], \mathbb{V}[\hat{\boldsymbol{\beta}}])$$

Substituting the results from Parts 1 and 2:

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

Even without the assumption of Normal errors, the asymptotic distribution holds true by application of the Central Limit Theorem (specifically, the multivariate Central Limit Theorem) to the LSE estimator, provided that certain regularity conditions (e.g., moments exist and $\frac{1}{n}\mathbf{X}^T\mathbf{X}$ converges to a non-singular matrix) are met as the sample size $n \to \infty$.

An approximate $1-\alpha$ confidence interval for a specific coefficient β_j is given by:

$$\text{CI}(\beta_j) \approx \hat{\beta}_j \pm z_{\alpha/2}\hat{\text{se}}(\hat{\beta}_j)$$

where $\hat{\text{se}}^2(\hat{\beta}_j)$ is the j-th diagonal element of the estimated covariance matrix, $\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$.

5.5 The Bias-Variance Trade-off

Example 38. *We simulate from a known non-linear model, which represents the True Underlying Relationship:*

$$E[Y|X] = \exp(\beta_0^* + \beta_1^* X^2)$$

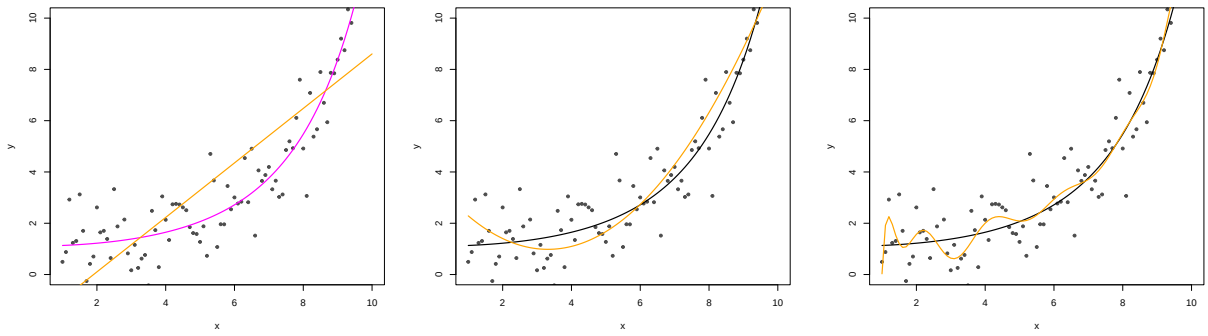
The observed response Y includes a random error component ϵ, assumed to be normally distributed:

$$Y = \exp(0.1 + 0.025 \cdot X^2) + \epsilon, \quad \epsilon \sim N(0, \sqrt{1.3})$$

The True Function is an upward-curving exponential quadratic line (represented by the magenta line in the plots).

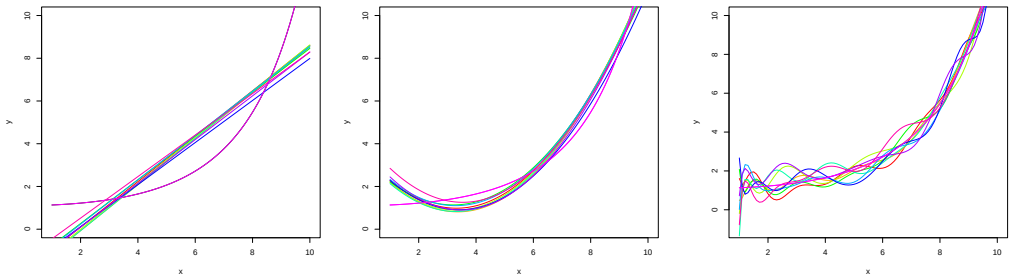
We perform three models of increasing polynomial degree ($d = 1, 2, 15$) to the same initial dataset:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2 + \cdots + \hat{\beta}_d X^d$$



1. *Degree 1 (Simple Linear):*
   - *High Bias (Underfitting): The straight line cannot capture the true curvature. The model is too simple for the data.*
2. *Degree 2 (Quadratic):*
   - *Optimal Fit: Aligns closely with the true exponential quadratic curve, achieving a good balance between bias and variance.*
3. *Degree 15 (High Order):*
   - *Overfitting: The curve is excessively flexible and attempts to pass through every noise point in the sample. This fit is excellent on the training data but performs poorly on new data.*

Now we repeats the data generation and fitting process 9 times to visualize the stability of the models.

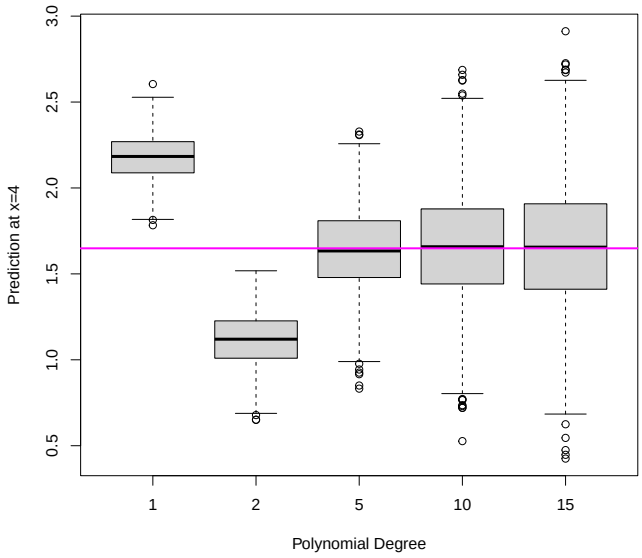


- *Linear Model (Degree 1): The 9 fitted lines are tightly clustered (low variance) but are consistently far from the true curve (high bias). Underfitting is driven by bias.*
- *Quadratic Model (Degree 2): The 9 fitted curves cluster closely around the true curve. This model exhibits both Low Bias and Low Variance.*
- *High-Degree Model (Degree 15): The 9 fitted curves vary wildly, especially near the edges of the data (extrapolation). This exhibits Extremely High Variance, indicating Overfitting. A small change in the training data (due to new random error) results in a large change in the fitted model.*

Finally we repeats the simulation $N_{rep} = 1000$ *times to quantify the prediction variability (Variance) and accuracy (Bias) at a specific point,* $x_0 = 4$.

- *Boxplot Visualization: The boxplot displays the distribution of the 1000 predicted values for each model degree (d=1, 2, 5, 10, 15).*
- *Bias Check: The difference between the true value at* $x = 4$ *(magenta line) and the center of the boxplot (median or mean prediction) measures the prediction bias. The linear model shows the greatest bias.*
- *Variance Check: The height (spread) of the boxplot measures the prediction variance. The highest-degree models (10 and 15) show the greatest height/spread, confirming their high variance and poor generalization to new data.*

The plot clearly demonstrates that selecting a model of intermediate complexity (Degree 2) provides the best trade-off, minimizing the total error by balancing the contributions of bias and variance.



In the context of statistical learning, the expected prediction error (or generalization error) of a model $\widehat{m}(\mathbf{X})$ can be decomposed into three primary components, leading to the trade-off principle:

- Bias of $\widehat{m}$: Measures the proximity of the model's average prediction to the true, underlying complex function $m(\mathbf{X})$.

  $$\text{Bias} = E\left[\widehat{m}(\mathbf{X})\right] - m(\mathbf{X})$$

  A simple (rigid) model (e.g., a linear fit to non-linear data) exhibits High Bias because its functional form is fundamentally too inflexible to capture the true relationship.

- Variance of $\widehat{m}$: Refers to the variability or instability of our model's estimates when trained on different random samples (datasets) drawn from the same population.

  $$\text{Variance} = E\left[(\widehat{m}(\mathbf{X}) - E[\widehat{m}(\mathbf{X})])^2\right]$$

  A complex (flexible) model (e.g., a high-degree polynomial) exhibits High Variance because it adapts excessively to the noise in each specific training set.

- Irreducible Error (ϵ): The noise inherent in the data-generating process that no model can eliminate.

The relationship between complexity, bias, and variance is inverse:

$$\text{Expected Prediction Error} \propto \text{Bias}^2 + \text{Variance}$$

Simple Models ($\downarrow$ Complexity) $\implies$ $\uparrow$ Bias, $\downarrow$ Variance (Underfitting)

Complex Models ($\uparrow$ Complexity) $\implies$ $\downarrow$ Bias, $\uparrow$ Variance (Overfitting)

The optimal model complexity lies at the point where the sum of squared bias and variance is minimized.

5.5.1 The Problem with Training Error

To evaluate a model $m(\mathbf{X})$, we often consider the Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{m}(\mathbf{X}_i))^2$$

However, computing the MSE on the same observations used to fit the model ($\widehat{m}$) leads to an underestimation of the true error on new data. This is because the model has already "seen" and minimized the error for those specific points, leading to a phenomenon known as optimism or overestimation of model goodness.

Contemporary strategies resolve this issue by ensuring that the model is evaluated on data it has not seen. The primary approach involves splitting the dataset $\mathcal{M}$:

- $\mathcal{T}$ (Training Set): Used exclusively to fit the model (i.e., estimate the parameters $\widehat{\boldsymbol{\beta}}$).
- $\mathcal{V}$ (Validation Set): Used exclusively to evaluate the model's performance and select the best model among competitors.

Case: Large Datasets (e.g., 80/20 Split)

When data is abundant, a simple split (e.g., 80% Training, 20% Validation) is sufficient. The evaluation criterion is the validation error:

$$\text{Validation Error} = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\left(y_i - \widehat{m}_{\mathcal{T}}(\mathbf{X}_i)\right)^2$$

We choose the competing model m that minimizes this validation error:

$$\min_{m}\left\{\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\left(y_i - \widehat{m}_{\mathcal{T}}(\mathbf{X}_i)\right)^2\right\}$$

Case: Limited Data - K-Fold Cross-Validation (CV)

When data is scarce, a single split is inefficient as it leaves a large portion unused for training or validation. K-Fold Cross-Validation is employed to use all data points for both roles.

1. The dataset $\mathcal{M}$ is randomly divided into K equally sized, non-overlapping subsets (or *folds*) $\mathcal{T}_1^c, \mathcal{T}_2^c, \ldots, \mathcal{T}_K^c$.
2. For $k = 1$ to K:
   - The k-th fold, $\mathcal{T}_k^c$, is reserved as the Validation Set.
   - The remaining $K-1$ folds, denoted $\mathcal{T}_k$ (the union of all other folds), form the Training Set.
   - The model is fitted using $\mathcal{T}_k$, yielding $\widehat{m}_{\mathcal{T}_k}$.
   - The prediction error is computed on the reserved validation set $\mathcal{T}_k^c$.

The final CV score for the model $\widehat{m}$ is the average of the K validation errors:

$$\mathrm{CV}(\widehat{m}) = \frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{|\mathcal{T}_k^c|}\sum_{j\in\mathcal{T}_k^c}(y_j - \widehat{m}_{\mathcal{T}_k}(\mathbf{X}_j))^2\right]$$

The model with the lowest $\mathrm{CV}(\widehat{m})$ score is selected as the optimal choice.

—

5.6 Model Selection Techniques

Methods for choosing the best set of predictors (ℓ) or model form m can be classified as follows:

- Best All Subsets:
  - Considers every possible combination of $\ell = 1, 2, \ldots, p$ predictors.
  - Feasible only when the total number of predictors (p) is small or moderate ($p \approx 20$). The number of models grows exponentially (2^p models).
- Stepwise Methods: Variables are incorporated or eliminated sequentially based on a criterion (e.g., AIC, BIC, or minimizing CV error).
  - Forward Selection: Starts with no predictors. Iteratively adds the single predictor that yields the best improvement until the chosen criterion stops improving. The number of models is manageable, order $p(p+1)/2$.
  - Backward Elimination: Starts with the full model (all p predictors). Iteratively removes the least significant variable until the criterion indicates the best model is found.
- Regularization (Penalty) Methods:
  - Techniques like LASSO (Least Absolute Shrinkage and Selection Operator) and Elastic Net automatically perform variable selection by shrinking the coefficients of less relevant predictors towards zero.

Regularization: Ridge and LASSO

Prediction Accuracy and Instability

In standard Ordinary Least Squares (OLS) regression, issues arise when the model complexity is high relative to the data size:

- **Overfitting and High Variance:** When the sample size n is not much larger than the number of predictors p, the OLS estimator can suffer from overfitting, leading to high variability (variance) in the estimated coefficients.
- **Non-Uniqueness (The $p > n$ Case):** If the number of predictors p is greater than the number of observations n ($p > n$), the estimation of OLS coefficients is no longer unique, and the variance of the estimators technically increases to infinity.

- **The Solution: Shrinkage:** A strategy to combat this is to **constrain the estimators**, resulting in estimates that, although **biased**, effectively achieve a significant **reduction in variance** (Bias-Variance Trade-off).
- **Sparsity and Interpretability:** Furthermore, when predictors are unrelated to the response, it is desirable to exclude them (i.e., set their coefficient to zero) to lower model complexity and enhance interpretability.

5.7 Regularization (Penalization)

Regularization is a technique used to stabilize models and prevent overfitting by adding a penalty term to the OLS objective function. In this approach, a model containing all covariates is fitted. However, the fitting method forces the coefficient estimates to **shrink** (move) towards zero. This procedure is called **regularization** or **penalization**, and its primary purpose and effect is to **reduce the estimated variance**. Depending on the specific penalty chosen, some coefficient estimates may become **exactly zero**, thereby achieving the secondary goal of **variable selection** (sparsity).

5.7.1 Ridge Regression (L2 Penalty)

The Ridge estimator $\widehat{\boldsymbol{\beta}}^{\text{Ridge}}$ is defined as the solution to a constrained optimization problem:

$$\arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \mathbf{x}_i^\top \boldsymbol{\beta} \right]^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} b_j^2 \leq c$$

Alternatively, Ridge Regression is typically presented in its penalized form:

$$\arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \mathbf{x}_i^\top \boldsymbol{\beta} \right]^2 + \lambda \sum_{j=1}^{p} b_j^2 \right\}$$

for some tuning parameter $\lambda > 0$.

The penalty $\lambda \sum b_j^2$ uses the L_2 norm (squared Euclidean distance). Ridge shrinks coefficients towards zero, but it **never sets them exactly to zero**. Data is usually standardized (mean 0, variance 1) so that coefficients are on a comparable scale, preventing the penalty from unfairly favoring variables with smaller units of measurement. The intercept is typically not included in the penalty term.

5.7.2 LASSO: Least Absolute Selection and Shrinkage Operator (L1 Penalty)

LASSO is a method that also penalizes coefficients for taking large values, but uses the L_1 norm (absolute values).

The LASSO estimator $\widehat{\boldsymbol{\beta}}^{\text{LASSO}}$ is defined by minimizing:

$$\widehat{\boldsymbol{\beta}}^{\text{LASSO}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left[y_i - \mathbf{x}_i^\top \boldsymbol{\beta} \right]^2 + \lambda \sum_{j=1}^{p} |b_j| \right\}$$

for some $\lambda > 0$.

The objective function $\mathcal{S}^{\text{LASSO}}(\boldsymbol{\beta})$ has two clear components:

$$\mathcal{S}^{\text{LASSO}}(\boldsymbol{\beta}) = \underbrace{\frac{1}{2n}\sum_{i=1}^{n}\left[y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right]^2}_{\text{Penalizes poor fit}} + \underbrace{\lambda \sum_{j=1}^{p} |b_j|}_{\text{Penalizes non-zero coefficients}}$$

The first term controls the goodness-of-fit (data fidelity). The second term controls the parsimony and complexity of the model.

Under regularity conditions, LASSO automatically performs variable selection: it sets the coefficients of variables unrelated to the response to **exactly zero**. Like Ridge, LASSO can be expressed as a constrained optimization problem:

$$\widehat{\boldsymbol{\beta}}^{\text{LASSO}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\sum_{i=1}^{n}\left[y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right]^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |b_j| \leq s$$

The constraint region defined by $\sum |\boldsymbol{b}_j| \leq s$ (the L_1 norm) is a **diamond or rhombus** in $\mathbb{R}^p$. . Because the OLS error contours (ellipses) often intersect the diamond constraint at its sharp **corners** (where one coefficient is zero), LASSO inherently favors sparse solutions.

5.7.3 Selecting the Regularization Parameter (λ)

The parameter λ controls the strength of the penalty: $\lambda = 0$ gives OLS, and $\lambda = \infty$ sets all coefficients to zero.

λ is chosen by finding the value that minimizes the estimated prediction error, typically using K-Fold Cross-Validation:

1. **Grid Definition:** A grid of potential λ values is fixed.
2. **K-Fold Split:** The sample is randomly divided into K folds (e.g., $K = 5$ or $K = 10$).
3. **Iteration:** For each fold k and every λ:
   - The LASSO estimator is calculated using the $K - 1$ training folds.
   - The Prediction Error $\text{MSE}_k(\lambda)$ is computed on the validation fold (the k-th fold).
4. **CV Score:** The cross-validation loss for a given λ is the average of the errors:

$$\text{CV}(\lambda) = \frac{1}{K}\sum_{k=1}^{K} \text{MSE}_k(\lambda)$$

5. **Optimal λ:** The value that minimizes this average error is chosen: $\lambda^{CV} = \arg\min \text{CV}(\lambda)$.

5.8 Logistic Regression

While preceding discussions have focused on models where the response variable Y is continuous (real-valued), Logistic Regression is a parametric method specifically designed

for situations where the outcome variable Y_i is binary (dichotomous), taking values in $\{0, 1\}$.

For a k-dimensional vector of covariates $\mathbf{X} = (x_1, \ldots, x_k)$, the goal is to model the probability of success, p, which is the conditional probability of $Y = 1$ given the predictors.

The model defines the conditional probability of a successful outcome ($Y = 1$) using the logistic function to ensure that the probability p is constrained between 0 and 1:

$$p \equiv p(\boldsymbol{\beta}) \equiv \mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x}) = \frac{e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}}}$$

The parameters are $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^T$, where β_0 is the intercept.

The logistic regression model can also be expressed in a linear form by applying the logit transformation to the probability p_i:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

This transformation maps the probability scale $[0, 1]$ to the entire real line $(-\infty, \infty)$. Applying this transformation to the model yields:

$$\text{logit}(p_i) = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}$$

The quantity $\text{logit}(p)$ is known as the log-odds (logarithm of the odds of success).

Since the outcomes Y_i are binary, the data follows a Bernoulli distribution conditionally on the covariates:

$$Y_i \mid \mathbf{X}_i = \mathbf{x}_i \sim \text{Bernoulli}(p_i(\boldsymbol{\beta}))$$

The probability mass function for a single observation Y_i is $p_i^{Y_i}(1 - p_i)^{1-Y_i}$. Assuming the observations are independent, the conditional likelihood function $\mathcal{L}(\boldsymbol{\beta})$ for the entire dataset is the product of these individual probabilities:

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i(\boldsymbol{\beta})^{Y_i}(1 - p_i(\boldsymbol{\beta}))^{1-Y_i}$$

The Maximum Likelihood Estimates (MLE) for the parameters $\boldsymbol{\beta}$ are found by maximizing the logarithm of this likelihood function.

Maximum Likelihood Estimation: Iterative Solution

The Maximum Likelihood Estimator (MLE) for the parameter vector $\boldsymbol{\beta}$ in logistic regression cannot be solved in closed form (analytically) like the Least Squares Estimator. The solution, $\hat{\boldsymbol{\beta}}$, must be obtained by maximizing the log-likelihood function $\ell(\boldsymbol{\beta}) = \log(\mathcal{L}(\boldsymbol{\beta}))$ numerically.

A common and efficient iterative numerical procedure used to solve these non-linear equations is based on the Newton-Raphson algorithm, which is often implemented in the context of Generalized Linear Models as the Iteratively Reweighted Least Squares (IRLS) method.

Newton-Raphson Update Rule The Newton-Raphson method finds the maximum of $\ell(\boldsymbol{\beta})$ by finding the root of its first derivative, the **Score Vector** (or Gradient), $\mathbf{g}(\boldsymbol{\beta}) = \nabla\ell(\boldsymbol{\beta})$, using the following iterative update formula:

$$\boldsymbol{\beta}^{(s+1)} = \boldsymbol{\beta}^{(s)} - \mathbf{H}(\boldsymbol{\beta}^{(s)})^{-1}\mathbf{g}(\boldsymbol{\beta}^{(s)})$$

Where:

- $\boldsymbol{\beta}^{(s)}$ is the current parameter vector estimate at iteration s.
- $\mathbf{g}(\boldsymbol{\beta}^{(s)}) = \nabla\ell(\boldsymbol{\beta}^{(s)})$ is the Score Vector (vector of first derivatives).
- $\mathbf{H}(\boldsymbol{\beta}^{(s)}) = \nabla^2\ell(\boldsymbol{\beta}^{(s)})$ is the **Hessian Matrix** (matrix of second derivatives).

Components for Logistic Regression For the log-likelihood function $\ell(\boldsymbol{\beta})$ of the logistic model, the components required for the NR update are:

1. **Score Vector (Gradient):** The partial derivative with respect to β_j is given by:

$$\frac{\partial\ell(\boldsymbol{\beta})}{\partial\beta_j} = \sum_{i=1}^{n}(Y_i - p_i(\boldsymbol{\beta}))X_{ij}$$

   In matrix form, the Score Vector $\mathbf{g}(\boldsymbol{\beta})$ is:

$$\mathbf{g}(\boldsymbol{\beta}) = \mathbf{X}^T(\mathbf{Y} - \mathbf{p})$$

   where $\mathbf{p} = [p_1, \ldots, p_n]^T$ is the vector of predicted probabilities.

2. **Hessian Matrix:** The matrix of second partial derivatives is:

$$\frac{\partial^2\ell(\boldsymbol{\beta})}{\partial\beta_j\partial\beta_k} = -\sum_{i=1}^{n} X_{ij}X_{ik}p_i(\boldsymbol{\beta})(1 - p_i(\boldsymbol{\beta}))$$

   In matrix form, the Hessian $\mathbf{H}(\boldsymbol{\beta})$ is:

$$\mathbf{H}(\boldsymbol{\beta}) = -\mathbf{X}^T\mathbf{W}\mathbf{X}$$

   where $\mathbf{W} = \text{diag}(p_i(1 - p_i))$ is the diagonal weight matrix defined previously.

Equivalence with Fisher Scoring / IRLS In Logistic Regression, the Newton-Raphson method is mathematically equivalent to the Fisher Scoring method because the Hessian $\mathbf{H}(\boldsymbol{\beta})$ does not depend on the observed response $\mathbf{Y}$, making the observed information matrix $(-\mathbf{H})$ equal to the expected information matrix (Fisher Information, $\mathbf{I}(\boldsymbol{\beta})$).

The Newton-Raphson update can be shown to simplify exactly to the IRLS formula used in your notes:

$$\boldsymbol{\beta}^{(s+1)} = (\mathbf{X}^T\mathbf{W}^{(s)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{(s)}\mathbf{Z}^{(s)}$$

This confirms that the two procedures are solving the same problem.

The algorithm requires starting values and iteratively refines the coefficients until they converge.

Initialization:

1. Choose initial parameter values $\hat{\boldsymbol{\beta}}^{(0)} = (\hat{\beta}_0^{(0)}, \ldots, \hat{\beta}_k^{(0)})^T$. A common starting point is $\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$.

2. Use these initial estimates to calculate the initial predicted probabilities $p_i^{(0)}$ for $i = 1, \ldots, n$, using the logistic function:

$$p_i^{(0)} = \frac{e^{\hat{\beta}_0^{(0)} + \sum_{j=1}^{k} \hat{\beta}_j^{(0)} X_{ij}}}{1 + e^{\hat{\beta}_0^{(0)} + \sum_{j=1}^{k} \hat{\beta}_j^{(0)} X_{ij}}}$$

3. Set the iteration counter $s = 0$.

Iterative Steps (Repeat until convergence):

1. **Step 1: Calculate the Working Response (Z_i).** The algorithm linearizes the problem by constructing a modified dependent variable, Z_i, which is a combination of the current log-odds and the observed error term. This is often called the "working response" or "adjusted dependent variable":

$$Z_i^{(s)} = \text{logit}(p_i^{(s)}) + \frac{Y_i - p_i^{(s)}}{p_i^{(s)}(1 - p_i^{(s)})}, \quad i = 1, \ldots, n$$

2. **Step 2: Calculate the Weight Matrix (W).** Let $\mathbf{W}^{(s)}$ be an $n \times n$ diagonal matrix where the (i, i)-th element represents the weight assigned to the i-th observation. This weight is the estimated variance of the logit transformation evaluated at $p_i^{(s)}$:

$$\mathbf{W}^{(s)} = \text{diag}(p_1^{(s)}(1 - p_1^{(s)}), \ldots, p_n^{(s)}(1 - p_n^{(s)}))$$

3. **Step 3: Update the Coefficients ($\hat{\boldsymbol{\beta}}$).** The new estimate $\hat{\boldsymbol{\beta}}^{(s+1)}$ is calculated by performing a Weighted Least Squares (WLS) regression of the working response vector $\mathbf{Z}^{(s)}$ on the design matrix $\mathbf{X}$, using the weight matrix $\mathbf{W}^{(s)}$:

$$\hat{\boldsymbol{\beta}}^{(s+1)} = (\mathbf{X}^T \mathbf{W}^{(s)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(s)} \mathbf{Z}^{(s)}$$

   This step involves solving a standard WLS problem, hence the name "reweighted least squares."

4. **Step 4: Update and Repeat.** Set $s = s + 1$ and use the new estimate $\hat{\boldsymbol{\beta}}^{(s+1)}$ to compute updated probabilities $p_i^{(s+1)}$ for the next iteration. The process continues until the change in the coefficient vector $\hat{\boldsymbol{\beta}}$ falls below a predefined tolerance level (convergence criterion).

Example 39. *Dependent Variable (Y): Default No ($Y = 0$) or Default Yes ($Y = 1$). Independent Variable (X): A continuous predictor, such as the Balance: The average balance that the customer has remaining on their credit card after making their monthly payment (x). Objective: To estimate the probability $P(Y = 1|X = x)$, which is the probability that a customer defaults, given their credit card balance (x).*

The logit is defined as the logarithm of the odds (the odds ratio of success to failure):

$$logit[P(Y=1|X=x)] = \ln\left(\frac{P(Y=1|X=x)}{1-P(Y=1|X=x)}\right) = \beta_0 + \beta_1 x$$

β_0: The intercept. It is the log-odds of defaulting when the balance (x) is zero. β_1: The coefficient associated with the variable X. It measures the change in the log-odds for a one-unit change in x (balance).

By applying the inverse of the logit function, we directly obtain the probability of default:

$$P(Y=1|X=x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

This function, which always produces a characteristic S-shaped curve, is crucial. The Sigmoid curve starts close to 0 for very small values of x (or $\beta_0+\beta_1 x$), smoothly transitions near the center (where $P = 0.5$), and asymptotically approaches 1 for very large values of x. This mathematically ensures that the estimated probability that a client enters default ($Y = 1$) given their credit card balance x remains valid.

*The interpretation of the coefficients (β) in logistic regression is based on the **odds** and the **Odds Ratio**. The odds represent the ratio between the probability of the event of interest occurring (default) and the probability of it not occurring (no default).*

$$Odds(x) = \frac{P(Y=1|X=x)}{1-P(Y=1|X=x)} = e^{\beta_0+\beta_1 x}$$

The Odds Ratio (OR) is the key metric for interpreting β_1 and is calculated as e^{β_1}. It measures the multiplicative change in the odds when the variable X (credit card balance) increases by one unit.

$$Odds\ Ratio = OR = e^{\beta_1}$$

$$OR = \frac{Odds\ of\ default\ with\ x+1\ unit\ of\ balance}{Odds\ of\ default\ with\ x\ unit\ of\ balance}$$

Suppose the model produces an Odds Ratio of $e^{\beta_1} = 1.50$.

1. *β_1 is positive (> 0), indicating a positive relationship: a higher balance increases the likelihood of default.*

2. *The OR is 1.50. This means that for **each additional unit of balance**, the **odds of default** are **multiplied by 1.50**. If a client currently has odds of default of $1:4$ (i.e., 0.25), increasing their balance by one unit would raise their odds to $1.50 \times 0.25 = 0.375$ (or $0.375:1$).*

Table 5.1: Key Values of the Odds Ratio (OR)

Value of $OR = e^{\beta_1}$	**Interpretation**
$OR > 1$	Variable X is a **risk factor** for the event (default).
$OR = 1$	Variable X has **no effect** on the odds of default.
$OR < 1$	Variable X is a **protective factor** for the event (default).

The complete model $P(Y = 1|X = x)$ estimates the probability. The S-shaped (sigmoid) curve generated by this function shows that:

1. *For low values of x (low balance), the probability of default is close to zero.*

2. *For high values of x (high balance), the probability of default is close to one.*

3. *The crossover point (where $P(Y = 1|X = x) = 0.5$) is reached when the logit is zero ($\beta_0 + \beta_1 x = 0$). This value of x indicates the balance required to have an equal chance of defaulting or not defaulting (odds of 1:1).*

Chapter 6

The Bootstrap Method

The **Bootstrap**, originally introduced by Bradley Efron in his seminal 1979 paper *"Bootstrap methods: another look at the jackknife"*, refers to the fundamental idea of approximating the sampling distribution and precision of an estimator via computational simulation (specifically, resampling).

This technique has had a revolutionary impact on modern statistics, granting Efron significant academic recognition. Before the Bootstrap, statistical inference heavily relied on strict theoretical assumptions (like asymptotic normality) that were often difficult to verify or derive analytically for complex estimators.

It is important to note that there is no single way to apply the Bootstrap. Several variants exist (non-parametric bootstrap, parametric, Bayesian, wild bootstrap, etc.), making it more accurate to refer to it as a **collection of resampling methods** rather than a single, rigid technique.

6.1 The Empirical Distribution

To understand the operation of the Bootstrap, we must first formalize the object it operates on. Let X be a random variable with Cumulative Distribution Function (CDF) $F : \mathbb{R} \to [0, 1]$, defined by:

$$F(x) = P(X \leq x)$$

In practice, F is unknown, and our goal is to infer its properties from data.

6.1.1 The Empirical Distribution Function

Given a random sample $X_1, \ldots, X_n$ i.i.d. (independent and identically distributed) drawn from the distribution F, we define the **Empirical Distribution Function**, denoted as $\widehat{F}_n$, as follows:

$$\widehat{F}_n(x) = \frac{\sum_{i=1}^{n} I\left(X_i \leq x\right)}{n} \tag{6.1}$$

where $I(\cdot)$ is the indicator function, which equals 1 if the condition is true and 0 otherwise.

We can analyze $\widehat{F}_n$ from two perspectives:

1. Theoretical Perspective: Before observing the data, $\widehat{F}_n(x)$ is a *random function*, as it depends on the random variables X_i.

2. Sample Perspective: For a concrete realization $x_1, \ldots, x_n$ of the sample, $\widehat{F}_{n,\,obs}(x)$ is a step function (a staircase CDF). It assigns a probability (or mass) of $1/n$ to each observed point x_i.

Properties of $\widehat{F}_n(x)$

Let us fix a value $x \in \mathbb{R}$. We are interested in studying the statistical properties of the random variable $Y = I(X_i \leq x)$. Note that Y is a Bernoulli variable with success probability $p = P(X_i \leq x) = F(x)$.

Therefore, the variable $n\widehat{F}_n(x)$ follows a Binomial distribution:

$$n\widehat{F}_n(x) = \sum_{i=1}^{n} I(X_i \leq x) \sim \text{Binomial}(n, F(x))$$

From this, we can answer questions about its moments:

Expectation and Bias

$$\mathbb{E}[\widehat{F}_n(x)] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[I(X_i \leq x)]$$

Since $E[I(X_i \leq x)] = F(x)$:

$$\mathbb{E}[\widehat{F}_n(x)] = \frac{1}{n} \cdot nF(x) = F(x)$$

Then $\widehat{F}_n(x)$ is an unbiased estimator of $F(x)$ for all x.

Variance

$$\mathbb{V}(\widehat{F}_n(x)) = \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}(I(X_i \leq x))$$

The variance of a Bernoulli variable with probability $p = F(x)$ is $F(x)(1 - F(x))$. By independence:

$$\mathbb{V}(\widehat{F}_n(x)) = \frac{1}{n^2} \cdot nF(x)(1 - F(x)) = \frac{F(x)(1 - F(x))}{n}$$

Note that as $n \to \infty$, the variance tends to 0. This suggests weak consistency for each fixed x.

Furthermore, by the Strong Law of Large Numbers, for each fixed x, $\widehat{F}_n(x) \xrightarrow{c.s.} F(x)$. However, in statistics, we need a stronger result than pointwise convergence: we need uniform convergence over the entire domain. This result is provided by the following theorem.

Theorem 27 (Glivenko - Cantelli)**.** *Let $X_1, \ldots, X_n \sim F$. Then:*

$$\sup_{x \in \mathbb{R}} \left|\widehat{F}_n(x) - F(x)\right| \xrightarrow{c.s.} 0$$

This theorem, sometimes called the *Fundamental Theorem of Statistics*, assures us that, if the sample size is sufficiently large, the empirical distribution almost surely reconstructs the true theoretical distribution uniformly across the entire domain.

6.2 Statistical Functionals and The Plug-in Estimator

A **statistical functional** is a function T that takes a distribution function F as its argument and returns a real number (or vector). Formally, $T : \mathcal{F} \to \mathbb{R}$, where $\mathcal{F}$ is a space of distribution functions.

Classic Examples:

- **The Expectation:** The functional is $T(F) = \mathbb{E}_F(X) = \int x \, dF(x)$.
- **The Variance:** The functional is $T(F) = \mathbb{V}_F(X) = \int (x - \mu)^2 \, dF(x)$, where $\mu = \int x \, dF(x)$.
- **The Median:** The functional is $T(F) = F^{-1}(1/2) = \inf\{x : F(x) \geq 1/2\}$.

The **Plug-in Principle** is a simple yet powerful method for constructing estimators. If we want to estimate a parameter θ that can be written as a functional of the true distribution, $\theta = T(F)$, the plug-in estimator $\hat{\theta}$ is obtained by simply substituting the unknown distribution F with the empirical distribution $\widehat{F}_n$.

$$\widehat{T} = T(\widehat{F}_n)$$

Example 40. *Plug-in Estimator of the Expectation*

Let $X_1, \ldots, X_n \sim F$ and let the parameter of interest be the population mean $T(F) = \mathbb{E}_F(X)$. The plug-in estimator is:

$$\widehat{T} = \mathbb{E}_{\widehat{F}_n}(X) = \int x \, d\widehat{F}_n(x)$$

How do we calculate this integral with respect to the empirical distribution? Recall that $\widehat{F}_n$ assigns a probability mass of $1/n$ to each observed point x_i. Therefore, the expectation under this distribution is simply the weighted average of the observed values:

$$\mathbb{E}_{\widehat{F}_n}(X) = \sum_{i=1}^{n} x_i \cdot P_{\widehat{F}_n}(X = x_i) = \sum_{i=1}^{n} x_i \cdot \frac{1}{n} = \overline{X}$$

This demonstrates that the sample mean $\overline{X}$ is the plug-in estimator for the population mean μ. The logic of the Bootstrap is built upon this principle: if $\widehat{F}_n$ is a good approximation of F (thanks to Glivenko-Cantelli), then simulating samples from $\widehat{F}_n$ allows us to estimate the properties of $T(F)$.

The estimation of quantiles presents a slight notational complication when applying the plug-in principle, due to the non-invertibility of the empirical distribution function.

The p-th quantile of a distribution F, denoted Q_p, is the value x such that the probability of the random variable being less than or equal to x is at least p.

If F is continuous and strictly increasing, the p-th quantile is uniquely defined by the inverse function:

$$T(F) = F^{-1}(p)$$

However, this definition is often too restrictive, as many distributions (including discrete distributions like the empirical one) are not strictly increasing or invertible in the standard sense.

Therefore, the general and robust definition of the **Quantile Function** (or generalized inverse of F) is used:

Definition 29 (Quantile Function $F^{-1}(p)$)**.** *For $p \in (0,1)$, the p-th quantile of F is defined as:*

$$T(F) = F^{-1}(p) = \inf\{x : F(x) \geq p\}$$

This definition ensures a unique result even when the distribution F has flat sections or jumps.

Example 41. *The Plug-in Estimator for Quantiles*

Following the plug-in principle, the estimator for the p-th quantile is obtained by substituting the unknown theoretical distribution F with the observable Empirical Distribution Function, $\widehat{F}_n$:

$$\widehat{T} = T(\widehat{F}_n) = \widehat{F}_n^{-1}(p)$$

*This estimator $\widehat{F}_n^{-1}(p)$ is called the **Sample p-th Quantile**.*

The Empirical Distribution Function, $\widehat{F}_n$, is a step function . It is not strictly increasing and therefore does not have a standard inverse. The generalized inverse definition (the inf *operator) resolves this issue.*

*We can explicitly show how the sample p-th quantile relates to the **Order Statistics** of the sample.*

Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ be the order statistics of the sample $X_1, \ldots, X_n$.

The empirical distribution $\widehat{F}_n$ has jumps of size $1/n$ at each observation $X_{(i)}$.

- *$\widehat{F}_n(x)$ equals $(i-1)/n$ for $x < X_{(i)}$.*
- *$\widehat{F}_n(x)$ jumps to i/n at $x = X_{(i)}$.*

To find $\widehat{F}_n^{-1}(p) = \inf\{x : \widehat{F}_n(x) \geq p\}$, we look for the smallest x that satisfies the condition.

The smallest value of x for which $\widehat{F}_n(x)$ exceeds or equals p occurs at one of the order statistics. Specifically, the condition $\widehat{F}_n(x) \geq p$ is satisfied by the first order statistic $X_{(i)}$ such that the jump in $\widehat{F}_n$ at $X_{(i)}$ reaches or surpasses p.

Since $\widehat{F}_n(X_{(i)}) = i/n$, we must find the smallest index i such that:

$$\frac{i}{n} \geq p \quad \Longrightarrow \quad i \geq np$$

Therefore, the sample p-th quantile is the order statistic corresponding to the index $i = \lceil np \rceil$, where $\lceil \cdot \rceil$ is the ceiling function (rounding up to the nearest integer).

$$\widehat{F}_n^{-1}(p) = X_{(\lceil np \rceil)}$$

This establishes the direct link between the theoretical plug-in principle and the commonly used Order Statistics in non-parametric statistics.

Example 42. *Total Family Income in Argentina (2023)*

This section illustrates the use of the Empirical Distribution Function ($\widehat{F}_n$) and Plug-in Estimators (like the median and mean) using real-world economic data. (Source: INDEC, National Institute of Statistics and Census of Argentina, [www.indec.gob.ar])

To visualize the CDF, we evaluate $\widehat{F}_n$ at every observed data point and plot the result.

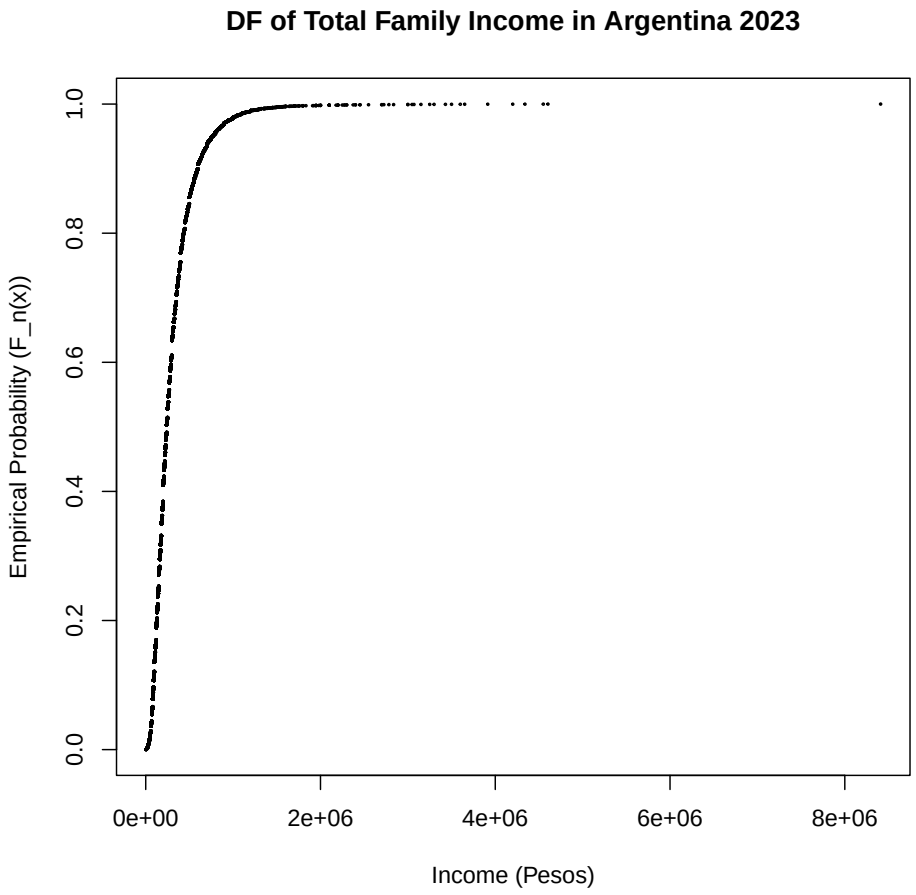


Figure 6.1: Empirical Cumulative Distribution Function of Total Family Income (ITF) in Argentina, 2023.

Using the $\widehat{F}_n$ (the plug-in estimator for F), we can estimate probabilities for intervals and quantiles.

What proportion of Argentine families had an income between 40,000 and 50,000 pesos in 2023?

$$P(40000 < X \leq 50000) \approx \widehat{F}_n(50000) - \widehat{F}_n(39999)$$

What proportion of Argentine families had an income greater than 100,000 pesos in 2023?

$$P(X > 100000) \approx 1 - \widehat{F}_n(100000)$$

We estimate the median ($Q_{0.5}$) and the mean (μ) using their respective plug-in estimators: the sample median ($\widehat{F}_n^{-1}(0.5)$) and the sample mean ($\overline{X}$).

We estimate the 0.5 *quantile of the total family income. This value separates the bottom 50% of the population from the top 50%.*

The sample mean is the plug-in estimator for the population mean ($\mathbb{E}_F(X)$).

In skewed distributions, like income data, the mean is often significantly higher than the median, pulled by high-income outliers. This is clearly shown when comparing the two plug-in estimates.

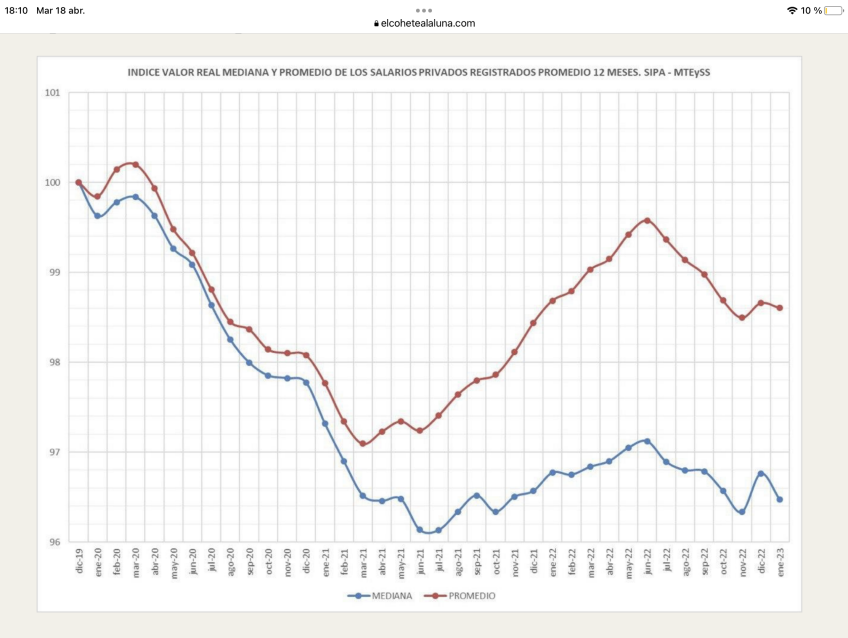


Figure 6.2: Illustration of Mean vs. Median for a Right-Skewed Distribution (typical of income data). The mean is pulled to the right by high values.

Example 43. *Lifetime of Lamps*

In industrial statistics, we often analyze ***survival data*** *(like the lifetime of a component). Here, we study the time-to-failure (in days) for a sample of 30 lamps.*

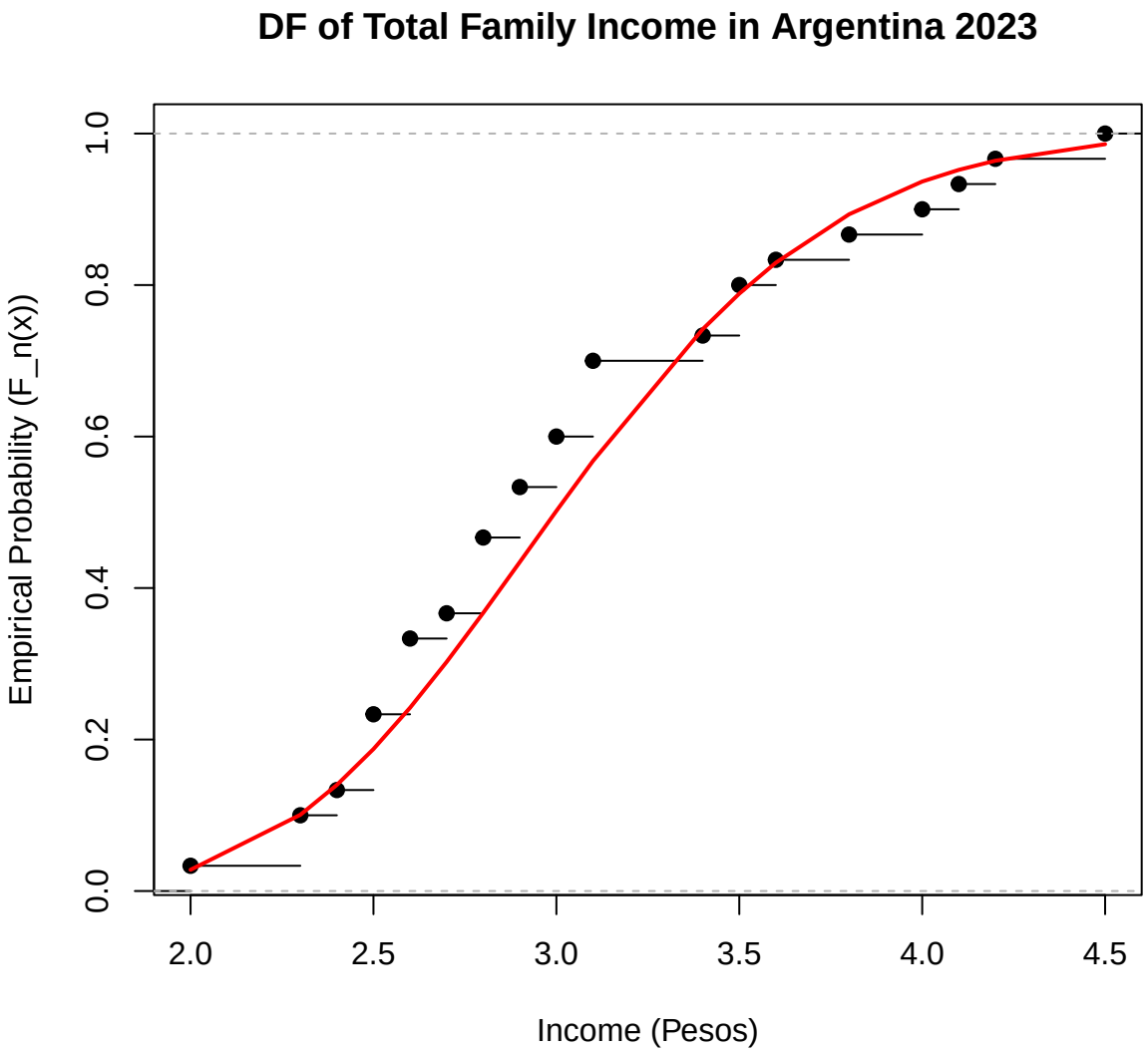
3.0	2.6	4.6	2.7	3.4	4.5	3.3	3.0	2.7	3.2
2.5	2.1	3.6	2.3	3.6	2.6	4.1	3.1	3.0	2.9
2.9	3.3	2.9	2.4	2.8	3.2	3.8	3.4	2.5	2.8

Table 6.1: Observed Lamp Lifetimes (in days)

We are primarily interested in estimating the population median lifetime, m.
We have two primary ways to estimate m:

1. ***Parametric Assumption:*** *Assuming the underlying distribution (F) belongs to a known family (e.g., Exponential, Weibull, Gamma).*

2. ***Non-Parametric Approach:*** *Using the sample median (the plug-in estimator $\widehat{F}_n^{-1}(0.5)$).*



Case 1: Parametric Assumption (Example: Exponential Distribution)

If we assume the distribution of lifetimes is ***Exponential*** *($\mathcal{E}(\lambda)$), the CDF is $F(t) = 1 - e^{-\lambda t}$. The population median m is found by solving $F(m) = 0.5$:*

$$1 - e^{-\lambda m} = 0.5 \quad \Longrightarrow \quad m = -\frac{\log(0.5)}{\lambda}$$

The value of m depends entirely on estimating the parameter λ from the sample.

Case 2: Non-Parametric Case

For the non-parametric approach , we use the sample median as the estimator.

The sample median is a consistent and robust estimator, but answering questions about its precision is difficult without knowing F or relying on asymptotic theory:

1. *What is the* ***median lifetime*** *of the lamps? (Answered by the plug-in estimator $\widehat{F}_n^{-1}(0.5)$).*
2. *How* ***precise*** *is the estimator? How do we estimate its* ***Standard Error*** *(SE)?*
3. *How do we calculate a* ***Confidence Interval*** *for the population median m?*

6.3 The Bootstrap: Handling Complex Estimators and Unknown Distributions

When dealing with a sample $X_1, \ldots, X_n$ from a distribution F, and an estimator $\widehat{T}_n$ for a functional $T(F)$, we face difficulties in determining the precision of $\widehat{T}_n$ (i.e., its standard error and confidence intervals) under two main scenarios:

1. **Known but Complex F:** We know the theoretical distribution F, but the estimator $\widehat{T}_n$ is a complicated function of the sample data. This complexity makes it mathematically difficult or impossible to derive the estimator's exact sampling distribution or its standard deviation.

2. **Unknown F:** The true underlying distribution F is completely unknown, precluding the use of standard asymptotic formulas that rely on properties of F (e.g., population variance).

The Bootstrap method offers a universal, computer-intensive solution to approximate the sampling distribution of $\widehat{T}_n$ in both cases. To frame the problem, let's consider the population median, $m = F^{-1}(0.5)$. As we explain, we can approach its estimation in two ways:

a) Estimate the population median by **assuming a parametric model** (e.g., Normal, Gamma).

b) Estimate the population median **without assuming any model** (using the sample median, a non-parametric plug-in estimator).

A crucial step for inference is answering: **How do we estimate the standard error ($\widehat{\text{se}}$) of this estimator?** The answer lies in the BOOTSTRAP!

6.3.1 Toy Scenario

To understand the core idea of the Bootstrap, imagine a theoretical scenario where the true distribution F is known and we can generate data from it as many times as we want.

1. **Generate Bootstrap Samples ($\mathbf{X}^*$):** We generate B large independent sets of data (e.g., $B = 1000$), each of size n, from the known true distribution F. Let $\mathbf{X}_b^* = X_{b,1}^*, \ldots, X_{b,n}^*$ denote the b-th sample.

2. **Calculate Bootstrap Replicates ($\widehat{T}^*$):** For each generated dataset $\mathbf{X}_b^*$, we calculate the value of the estimator of interest: $\widehat{T}_{n,b}^* = \delta(\mathbf{X}_b^*)$.

3. **Approximate the Distribution:** The empirical distribution of the resulting values:
$$\widehat{T}_{n,1}^*, \widehat{T}_{n,2}^*, \ldots, \widehat{T}_{n,B}^*$$
is an approximation to the true sampling distribution of $\widehat{T}_n$.

The key insight of the Bootstrap is to replace the unknown F with an estimate ($\widehat{F}_n$) and conduct this simulation in the "bootstrap world."

6.3.2 Parametric Bootstrap

In reality, we only possess one sample: $\mathbf{X} = X_1, \ldots, X_n$, drawn from the unknown distribution F.

If we assume the underlying distribution belongs to a parametric family, $F = F(\boldsymbol{\theta})$, we can use the Parametric Bootstrap.

1. **Estimate Parameters:** Use the original sample $\mathbf{X}$ to estimate the unknown parameter $\boldsymbol{\theta}$ (e.g., $\widehat{\boldsymbol{\theta}}_n$) and the functional $\widehat{T}_n = T(F(\widehat{\boldsymbol{\theta}}_n))$.
2. **Generate Bootstrap Samples:** Generate B resamples of size n, denoted $\mathbf{X}_b^*$, from the estimated parametric distribution $F(\widehat{\boldsymbol{\theta}}_n)$.
3. **Calculate Replicates:** For each resample, calculate the estimator of interest: $\widehat{T}_{n,b}^*$.
4. **Approximate Distribution:** The empirical distribution of the resulting values $(\widehat{T}_{n,1}^*, \ldots, \widehat{T}_{n,B}^*)$ approximates the sampling distribution of $\widehat{T}_n$.

Example 44. *Lamp Lifetime*

We are interested in the population median lifetime (m) and assume $X_i \sim \Gamma(\alpha, \lambda)$.

1. *Estimate the parameters: $\widehat{\alpha}_n$ and $\widehat{\lambda}_n$ (often via Maximum Likelihood Estimation or MLE).*
2. *Generate B resamples $\mathbf{X}^* = X_1^*, \ldots X_n^*$ from $F(\widehat{\alpha}_n, \widehat{\lambda}_n)$. For each sample, estimate the median, $\widehat{T}_n^* = \widehat{F}_n^{-1}(0.5)$.*
3. *The distribution of the B replicates approximates the sampling distribution of the median.*

6.3.3 Non-Parametric Bootstrap

If we make no assumptions about F (Case 2), the solution is the standard Non-Parametric Bootstrap. Since the empirical CDF ($\widehat{F}_n$) is the consistent estimator for F (by the Glivenko-Cantelli theorem), the bootstrap solution is to generate samples from $\widehat{F}_n$. The empirical distribution $\widehat{F}_n$ assigns a probability mass of $1/n$ to each observation X_i.

A bootstrap sample $\mathbf{X}^* = X_1^*, \ldots X_n^*$ from the distribution $\widehat{F}_n$ is equivalent to drawing a sample of size n with replacement from the original data set $X_1, \ldots, X_n$.

Example 45. *Lamp Lifetime*

We estimate the median life, $\widehat{T}_n = median(\mathbf{X})$, and its standard error, $\widehat{se}_{boot}$.

1. *Generate a bootstrap sample $X_1^*, \ldots, X_n^*$ from the empirical distribution $\widehat{F}_n$. This means sampling with replacement from the original data $\mathbf{X}$.*
2. *Calculate the estimator on the bootstrap sample: $\widehat{T}_n^* = median\,(X_1^*, \ldots, X_n^*) = \widehat{F}_n^{-1}(0.5)$.*
3. *Repeat: Repeat Steps 1 and 2, B times, to obtain $\widehat{T}_{n,1}^*, \ldots, \widehat{T}_{n,B}^*$.*
4. *Calculate Standard Error: The Bootstrap Standard Error ($\widehat{se}_{boot}$) is the standard deviation of the replicates:*

$$\widehat{se}_{boot} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\widehat{T}_{n,b}^* - \frac{1}{B}\sum_{r=1}^{B}\widehat{T}_{n,r}^*\right)^2}$$

6.3.4 The General Bootstrap Algorithm and $\widehat{\text{se}}_{\text{boot}}$

The following general algorithm allows us to estimate the standard error of any estimator $\widehat{T}_n = \delta(X_1, \ldots, X_n)$ by approximating its sampling distribution.

1. Generate Bootstrap Sample ($\mathbf{X}^*$): Generate a bootstrap sample $X_1^*, \ldots, X_n^*$ from the estimated distribution ($\widehat{F}_n$). (This means sampling with replacement from the original data $\mathbf{X}$ in the non-parametric case or from the parametric estimated distribution).

2. Calculate Bootstrap Replicate ($\widehat{T}^*$): Calculate the estimator on the bootstrap sample: $\widehat{T}_n^* = \delta(X_1^*, \ldots, X_n^*)$.

3. **Repeat:** Repeat Steps 1 and 2, B times, to obtain $\widehat{T}_{n,1}^*, \ldots, \widehat{T}_{n,B}^*$.

4. **Calculate Standard Error:** The Bootstrap Standard Error ($\widehat{\text{se}}_{\text{boot}}$) is the standard deviation of the replicates:

$$\widehat{\text{se}}_{\text{boot}} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\widehat{T}_{n,b}^* - \frac{1}{B}\sum_{r=1}^{B}\widehat{T}_{n,r}^*\right)^2}$$

Example 46. *Cauchy Distribution and the Median*

This example highlights the power of the bootstrap when analytical solutions are difficult or when standard methods fail (as with the sample mean in the Cauchy case).

Let $X_1, \ldots, X_n$ be i.i.d. random variables following a Cauchy distribution with median m. The Cauchy distribution has no finite mean or variance. The median is measure of centrality.

For n being an odd number, say $n = 2k + 1$, the exact variance of the sample median ($\widetilde{X}_n$), $\mathbb{V}(\widetilde{X}_n)$ is complicated:

$$\mathbb{V}\left(\widetilde{X}_n\right) = \frac{2n!}{(k!)^2\pi^n}\int_0^{\pi/2} x^k(\pi - x)^k(\cot x)^2 dx$$

Using asymptotic approximation theory (which requires knowledge of the underlying density f), we obtain an approximation for the variance:

$$\widehat{\mathbb{V}\left(\widetilde{X}_n\right)}_{Asymptotic} = \pi^2/4n$$

The bootstrap provides a numerical approximation to the true variance without needing the complex analytical derivation. For a sample size of $n = 21$:

- *Exact Variance:* $\mathbb{V}\left(\widetilde{X}_n\right) \approx 0.1367$
- *Asymptotic Approximation:* $\widehat{\mathbb{V}\left(\widetilde{X}_n\right)}_{Asymptotic} \approx 0.1175$
- *Bootstrap Approximation: A bootstrap estimator of $\mathbb{V}(\widetilde{X}_n)$ using $B = 5000$ iterations yields:*

$$\widehat{\mathbb{V}\left(\widetilde{X}_n\right)}_{boot} \approx 0.1356$$

The bootstrap result (≈ 0.1356) is remarkably close to the true exact variance (≈ 0.1367), illustrating that it gives highly accurate results with very little analytical effort, even when dealing with distributions that challenge classical statistical methods.

Remark 13. ***Parametric Bootstrap:*** *The distribution from which we resample is obtained by* ***fitting a known parametric model*** *(e.g., Gamma, Normal) to the original data x_i. This is more efficient if the model assumption is correct, but highly misleading if the model is wrong.* ***Non-Parametric Bootstrap:*** *The true unknown distribution F is approximated by the Empirical Distribution ($\widehat{F}_n$), which is discrete and places mass $1/n$ on each observation x_i. We resample directly from the observed data with replacement. This method is more robust as it makes minimal assumptions about F.*

The bootstrap procedure conceptually maps the real-world inference problem (where F is unknown) onto a solvable problem in the "bootstrap world" (where the distribution $\widehat{F}_n$ is known).

$$\begin{array}{lccccc} \text{Real World} & F & \xrightarrow{\text{Sample}} & X_1, \ldots, X_n & \xrightarrow{\text{Estimate}} & \hat{T} \\ \text{Bootstrap World} & \widehat{F}_n & \xrightarrow{\text{Resample}} & X_1^*, \ldots, X_n^* & \xrightarrow{\text{Replicate}} & \hat{T}^* \end{array}$$

The Bootstrap provides a powerful computational approach to solving complex inference problems, particularly estimating the precision of an estimator and gives us a way to estimate the **standard error (se)** of an estimator.

Given the estimator $\widehat{T}_n = \delta(X_1, \ldots, X_n)$, the core idea of the non-parametric bootstrap is twofold:

1. **Substitution Principle:** Estimate the true variance $\mathbb{V}_F(\widehat{T}_n)$ by substituting the unknown distribution F with the known Empirical Distribution Function $\widehat{F}_n$:

$$\mathbb{V}_F(\widehat{T}_n) \approx \mathbb{V}_{\widehat{F}_n}(\widehat{T}_n)$$

2. **Simulation:** Approximate the bootstrap variance $\mathbb{V}_{\widehat{F}_n}(\widehat{T}_n)$ through Monte Carlo simulation (resampling).

Example 47. *Illustrative Example: Variance of the Sample Median (Toy Case)*

To demonstrate the theoretical foundation of the Non-Parametric Bootstrap, we consider a small "toy case" where the true distribution F is unknown, and the sample size is very small, allowing us to compute the exact bootstrap sampling distribution by hand (without Monte Carlo simulation).

Let the original sample be $\mathbf{X} = \{3, 7, 9\}$, drawn from an unknown distribution F. The sample size is $n = 3$.

- *Estimator of Interest: The population median m.*
- *Plug-in Estimator: The sample median, $\widehat{T}_n = \text{median}(\mathbf{X}) = 7$.*

The Empirical Distribution Function ($\widehat{F}_n$) places a probability mass of $1/3$ on each observation $\{3, 7, 9\}$.

A bootstrap sample $\mathbf{X}^* = (X_1^*, X_2^*, X_3^*)$ is obtained by drawing $n = 3$ observations with replacement from the original sample $\{3, 7, 9\}$.

Since we draw $n = 3$ items from $k = 3$ unique values with replacement, there are a total of $k^n = 3^3 = 27$ possible ordered bootstrap samples (X_1^*, X_2^*, X_3^*). In simple random sampling, all 27 ordered samples are equally likely, each having a probability of $1/27$.

The probability mass function $p_{\widehat{F}}(\mathbf{X}^* = \{x_1^*, x_2^*, x_3^*\})$ of obtaining a specific set of values (e.g., $\{3, 3, 7\}$) is the number of distinct ordered permutations of those values divided by 27.

The following table summarizes all unique combinations of values and the corresponding bootstrap replicate of the median, $\widehat{T}_n^*$.

Values in Sample $\{x_1^*, x_2^*, x_3^*\}$	$\widehat{\mathbf{T}}_\mathbf{n}^* = median(\mathbf{X}^*)$	***Number of Permutations***	$\mathbf{P}_{\widehat{\mathbf{F}}}(Values)$
$\{3, 7, 9\}$	7	$3! = 6$	$6/27$
$\{3, 3, 7\}$	3	$3!/2! = 3$	$3/27$
$\{3, 3, 9\}$	3	$3!/2! = 3$	$3/27$
$\{7, 7, 3\}$	7	$3!/2! = 3$	$3/27$
$\{7, 7, 9\}$	7	$3!/2! = 3$	$3/27$
$\{9, 9, 3\}$	9	$3!/2! = 3$	$3/27$
$\{9, 9, 7\}$	9	$3!/2! = 3$	$3/27$
$\{3, 3, 3\}$	3	1	$1/27$
$\{7, 7, 7\}$	7	1	$1/27$
$\{9, 9, 9\}$	9	1	$1/27$

By summing the probabilities of all bootstrap samples that result in the same median value, we obtain the exact probability mass function of the bootstrap estimator $\widehat{T}_n^*$.

$$P_{\widehat{F}}(\widehat{T}_n^* = 3) = \frac{1}{27} \times (3 + 3 + 1) = \frac{7}{27}$$
$$P_{\widehat{F}}(\widehat{T}_n^* = 7) = \frac{1}{27} \times (6 + 3 + 3 + 1) = \frac{13}{27}$$
$$P_{\widehat{F}}(\widehat{T}_n^* = 9) = \frac{1}{27} \times (3 + 3 + 1) = \frac{7}{27}$$

The cumulative distribution function (CDF) of the bootstrap estimator, $P_{\widehat{F}}\left(\widehat{T}_n^* \leq u\right)$, is a step function:

$$P_{\widehat{F}}\left(\widehat{T}_n^* \leq u\right) = \begin{cases} 0 & \text{if } u < 3 \\ 7/27 & \text{if } u \in [3, 7) \\ 20/27 & \text{if } u \in [7, 9) \\ 1 & \text{if } u \geq 9 \end{cases}$$

The ideal bootstrap estimator for the variance of the sample median, $\mathrm{Var}_F(\widehat{T}_n)$, is the exact variance of the replicates $\widehat{T}_n^*$ under the empirical distribution $\widehat{F}_n$:

$$\mathrm{Var}_{\widehat{F}}\left(\widehat{T}_n^*\right) = E_{\widehat{F}}\left(\widehat{T}_n^{*2}\right) - \left[E_{\widehat{F}}\left(\widehat{T}_n^*\right)\right]^2$$

First, calculate the expected value of $\widehat{T}_n^*$:

$$E_{\widehat{F}}\left(\widehat{T}_n^*\right) = \sum_t t \cdot P_{\widehat{F}}(\widehat{T}_n^* = t) = 3 \cdot \frac{7}{27} + 7 \cdot \frac{13}{27} + 9 \cdot \frac{7}{27} = \frac{21 + 91 + 63}{27} = \frac{175}{27} \approx 6.481$$

Then, calculate the expected value of $(\widehat{T}_n^*)^2$*:*

$$E_{\widehat{F}}\left(\widehat{T}_n^{*2}\right) = \sum_t t^2 \cdot P_{\widehat{F}}(\widehat{T}_n^* = t) = 3^2 \cdot \frac{7}{27} + 7^2 \cdot \frac{13}{27} + 9^2 \cdot \frac{7}{27} = \frac{63 + 637 + 567}{27} = \frac{1267}{27} \approx 46.926$$

Finally, the ideal bootstrap variance is:

$$\mathrm{Var}_{\widehat{F}}\left(\widehat{T}_n^*\right) = E_{\widehat{F}}\left(\widehat{T}_n^{*2}\right) - \left[E_{\widehat{F}}\left(\widehat{T}_n^*\right)\right]^2 \approx 46.926 - (6.481)^2 \approx 46.926 - 42.003 = 4.923$$

Thus, the ideal bootstrap variance estimate is $\mathrm{Var}_{\widehat{F}}\left(\widehat{T}_n^*\right) = \mathbf{4.923}$*, and the ideal bootstrap standard error is* $\widehat{se}_{boot} = \sqrt{4.923} \approx 2.219$.

The preceding toy example, while theoretically insightful, demonstrates a critical limitation of the ideal bootstrap:

The distribution of the ideal bootstrap estimator, $G_{\widehat{F}}(u) = P_{\widehat{F}}\left(\widehat{T}_n^* \leq u\right)$, of the true sampling distribution $G_F(u) = P_F\left(\widehat{T}_n \leq u\right)$, is in principle calculable because it depends only on the known empirical distribution $\widehat{F}$. However, in practice, it is **infeasible** due to the computational burden: it requires the enumeration of an exorbitant number of possible bootstrap samples ($\mathbf{X}^*$) for any realistic sample size n.

$$\text{Total possible bootstrap samples} = n^n$$

For instance, if $n = 20$, the number of samples is 20^{20}, which is astronomically large.

To overcome the infeasibility of full enumeration, we approximate the ideal bootstrap distribution $G_{\widehat{F}}$ using Monte Carlo simulation.

The (practical) bootstrap estimator, $G_{\text{boot}}(u)$, of $G_F(u)$ is an empirical estimator, $\widehat{G}_{\widehat{F}}(u)$, of $G_{\widehat{F}}(u)$, based on B independent realizations (replicates) $\widehat{T}_{n,1}^*, \ldots, \widehat{T}_{n,B}^*$, calculated as:

$$G_{\text{boot}}(u) = \widehat{G}_{\widehat{F}}(u) = \frac{\#\left\{\widehat{T}_{n,b}^* : \widehat{T}_{n,b}^* \leq u\right\}}{B}$$

Here, $\widehat{T}_{n,b}^*$ are the bootstrap statistics generated by resampling with replacement B times.

The bootstrap procedure involves two distinct approximations to reach the final practical estimate of the sampling distribution:

$$\underbrace{G_F}_{\text{True Sampling Dist.}} \overbrace{\approx}^{\substack{\text{Approximation 1:}\\ \text{Substitution}\\ (\text{requires } n \gg 0)}} \underbrace{G_{\widehat{F}}}_{\substack{\text{Ideal Bootstrap Dist.}\\ \text{(Non-parametric Plug-in)}}} \overbrace{\approx}^{\substack{\text{Approximation 2:}\\ \text{Simulation}\\ (\text{requires } B \gg 0)}} \underbrace{\widehat{G}_{\widehat{F}}}_{\substack{\text{Practical Bootstrap Estimator}\\ \text{(Monte Carlo)}}}$$

The value of n (the sample size) is given by the data and dictates the quality of Approximation 1 (the better $\widehat{F}_n$ approximates F). The value of B (the number of bootstrap iterations) is under the control of the analyst and dictates the quality of Approximation 2 (the Monte Carlo precision).

The required number of bootstrap replicates, B, depends on the goal of the analysis:

- If the goal is simply to estimate the **variance** $\text{Var}_F(\widehat{T}_n)$ (i.e., the standard error), typically $B \approx 200$ to 500 will be sufficient.
- If the goal is to construct accurate **percentile confidence intervals** or estimate small tail probabilities, a much larger number of replicates, often $B \approx 1,000$ to $10,000$ or more, is recommended to ensure stability in the tails of the $\widehat{G}_{\widehat{F}}$ distribution.

6.4 The Bootstrap for Confidence Interval

We will see that the bootstrap method also provides multiple ways to calculate Confidence Intervals (CIs).

6.4.1 Asymptotic Normality and the Standard Error

The Normal Bootstrap Interval relies on the assumption that the estimator $\widehat{T}_n$ is **asymptotically normal**.

Under regularity conditions, the standardized estimator follows an approximate standard normal distribution:

$$\frac{\widehat{T}_n - T(F)}{\text{se}(\widehat{T}_n)} \approx \mathcal{N}(0,1)$$

By substituting the true standard error, $\text{se}(\widehat{T}_n)$, with its bootstrap estimate, $\widehat{\text{se}}_{\text{boot}}$, we can construct a confidence interval.

Let $z_{\alpha/2}$ be the critical value from the standard normal distribution corresponding to the desired level $1-\alpha$ (e.g., $z_{0.025} \approx 1.96$ for a 95% CI).

Normal Bootstrap Interval, approximate level $1-\alpha$:

$$\widehat{T}_n \pm z_{\alpha/2}\, \widehat{\text{se}}_{\text{boot}}$$

Example 48. *The Median Under certain regularity conditions, the sample median ($\widehat{M}_n$) is asymptotically normal. If m is the true median and $f(m)$ is the probability density function evaluated at m, then:*

$$\sqrt{n}(\widehat{M}_n - m) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{4f^2(m)}\right)$$

The bootstrap avoids the complex task of estimating $f(m)$.

6.4.2 The Percentile Bootstrap Confidence Interval

The Percentile Bootstrap Interval is often preferred because it does not rely on the assumption of asymptotic normality, nor does it require estimating the standard error explicitly. It is robust against non-normal or skewed sampling distributions.

1. Run the bootstrap algorithm (B iterations) to obtain the sorted bootstrap replicates: $\widehat{T}^*_{n,(1)} \le \widehat{T}^*_{n,(2)} \le \cdots \le \widehat{T}^*_{n,(B)}$.
2. Determine the quantiles corresponding to $\alpha/2$ and $1-\alpha/2$.

Let $\widehat{T}^*_\beta$ be the sample β-quantile of the bootstrap statistics $(\widehat{T}^*_{n,1}, \ldots, \widehat{T}^*_{n,B})$.

Percentile Bootstrap Interval, level $1-\alpha$:

$$\left(\widehat{T}^*_{\alpha/2}\ ,\ \widehat{T}^*_{1-\alpha/2}\right)$$

For a 95% CI ($\alpha = 0.05$), we use the 2.5-th percentile and the 97.5-th percentile of the bootstrap replicates.

6.5 Bootstrap For Linear Regression

Having previously explored the methodology of Linear Regression and the fundamental concept of the Bootstrap Method, this section merges both tools to address one of the central problems of statistical inference: the uncertainty of parameter estimates.

Regression, typically fitted using Ordinary Least Squares (OLS), provides point estimates for the coefficients ($\hat{\beta}$), which form the basis for prediction and model interpretation.

The Bootstrap in Regression allows us to move beyond the restrictive parametric assumptions (such as normality and homoscedasticity) often required to analytically derive standard errors and confidence intervals. By employing resampling with replacement techniques, the Bootstrap offers a robust, non-parametric approach to achieve the following goals:

1. Estimate the Empirical Sampling Distribution of the coefficients.
2. Calculate Stable and Unbiased Standard Errors.
3. Construct Confidence Intervals (such as the percentile interval) with greater accuracy, especially in complex models or when the error distribution is unknown.

We will subsequently examine the different implementations of the bootstrap, including the Paired Bootstrap, the Residual Bootstrap, and the Wild Bootstrap, understanding when each is appropriate based on the model's properties (e.g., the presence of heteroskedasticity).

6.5.1 Core Principle of the Bootstrap in Regression

The fundamental idea is to treat the sample data as if it were the population and repeatedly resample from it to understand the variability of the regression coefficients (β).

1. **Original Data:** We have the original dataset $\mathcal{D} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. The linear model is fitted:
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
This yields the Ordinary Least Squares (OLS) estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, and the residuals $e_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$.
2. **Resampling Procedure:** Repeat the resampling process B times (e.g., $B = 1000$). Each repetition b yields a bootstrap sample $\mathcal{D}^{*(b)} = \{(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)\}$ of size n.
3. **Estimation:** For each $\mathcal{D}^{*(b)}$, the regression model is refitted to obtain the bootstrap estimates $\hat{\beta}_0^{*(b)}$ and $\hat{\beta}_1^{*(b)}$.

4. **Inference:** The collection of B estimates $\{\hat{\beta}_j^{*(1)}, \ldots, \hat{\beta}_j^{*(B)}\}$ approximates the sampling distribution of $\hat{\beta}_j$.

The bootstrap standard error for a coefficient $\hat{\beta}_j$ is the standard deviation of its bootstrap distribution:

$$\text{SE}_B(\hat{\beta}_j) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}_j^{*(b)} - \bar{\hat{\beta}}_j^*\right)^2}, \quad \text{where } \bar{\hat{\beta}}_j^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_j^{*(b)}.$$

A simple and highly effective method for constructing the $100(1-\alpha)\%$ CI is to use the percentiles of the bootstrap distribution. For $\hat{\beta}_j$:

$$\text{CI}_{1-\alpha}(\hat{\beta}_j) = \left[Q_{\alpha/2}(\{\hat{\beta}_j^{*(b)}\}), Q_{1-\alpha/2}(\{\hat{\beta}_j^{*(b)}\})\right],$$

where Q_p is the p-th quantile (percentile) of the bootstrap estimates.

6.5.2 Bootstrap Sample

The choice of how to generate the bootstrap sample $\mathcal{D}^{*(b)}$ determines the type of bootstrap.

Paired Bootstrap (Empirical Bootstrap)

This method is the most straightforward, universally applicable, and least restrictive among the bootstrap techniques used for regression analysis. It is also often referred to simply as the **Empirical Bootstrap**.

The Paired Bootstrap treats the observed data pairs $\mathbf{z}_i = (X_i, Y_i)$ as the fundamental sampling units. It aims to mimic the original data generation process, where new observations are assumed to be drawn independently and identically distributed (IID) from the true, unknown population distribution F.

Procedure:

1. The original sample $\mathcal{D} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$ consists of n independent data points.
2. A bootstrap sample $\mathcal{D}^{*(b)}$ is formed by drawing n pairs from the set $\mathcal{D}$ with replacement.
3. The resulting bootstrap sample is $\mathcal{D}^{*(b)} = \{\mathbf{z}_1^*, \mathbf{z}_2^*, \ldots, \mathbf{z}_n^*\}$, where each $\mathbf{z}_i^* = (X_i^*, Y_i^*)$ is a resampled pair. The key here is that the structural relationship between X_i and Y_i is maintained in the resampled pairs.

By resampling the entire data pair (X_i, Y_i) together, the method implicitly captures the joint distribution of the predictor and the response. It simultaneously resamples both the $\mathbf{X}$ component and the underlying random error ε_i. This preserves the correlation and dependence structure present in the original data.

t does not require knowledge of the error distribution (e.g., normal errors).

It naturally handles issues like heteroskedasticity (non-constant error variance) and potential serial correlation in the errors because it doesn't assume the errors are IID; the entire observation unit is retained.

Due to its minimal assumptions, the Paired Bootstrap is the standard and most recommended approach for many regression models, including complex Generalized Linear Models (GLMs) like Logistic and Poisson Regression, where other bootstrap variations often fail.

Residual Bootstrap

This method assumes that the fitted model is correctly specified and that the errors are Independent and Identically Distributed (IID).

Procedure:

1. Fit the model to obtain the residuals $e_1, \ldots, e_n$ with $e_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$.
2. Generate n bootstrap errors $\{\varepsilon_1^*, \ldots, \varepsilon_n^*\}$ by sampling with replacement from the centered or uncentered residuals $\{e_1, \ldots, e_n\}$.
3. Construct the new bootstrap response variable Y_i^* by adding the bootstrap error to the fitted value:
$$Y_i^* = \hat{Y}_i + \varepsilon_i^* = \hat{\beta}_0 + \hat{\beta}_1 X_i + \varepsilon_i^*.$$
4. The bootstrap sample is $\{(X_1, Y_1^*), \ldots, (X_n, Y_n^*)\}$. Note: The covariates X_i are fixed and not resampled.

This approach has a limitation. It breaks down severely in the presence of heteroskedasticity because the errors are swapped indiscriminately across observations, violating the dependence structure of $\text{Var}(\varepsilon_i|X_i)$.

Wild Bootstrap

The Wild Bootstrap is specifically designed to provide robust standard errors and CIs in the presence of heteroskedasticity.

textbfProcedure:

1. Compute the residuals $e_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$ from the original model fit, $e_1, \ldots, e_n$.
2. Generate n IID random variables $V_1, \ldots, V_n$ with mean $\mathbb{E}(V_i) = 0$ and variance $\text{Var}(V_i) = 1$. A common choice is the distribution where V_i takes value 1 and -1 with equal probability 0.5.
3. Construct the new response Y_i^* by scaling the residual e_i by its unique perturbation V_i:
$$Y_i^* = \hat{\beta}_0 + \hat{\beta}_1 X_i + V_i \cdot e_i.$$
4. The bootstrap sample is $\{(X_1, Y_1^*), \ldots, (X_n, Y_n^*)\}$.

By multiplying e_i by V_i, the variance of the bootstrap error for observation i becomes $\text{Var}(V_i \cdot e_i|X_i) = \text{Var}(V_i) \cdot e_i^2 = e_i^2$. Since e_i^2 is an estimate of the local variance, the method successfully captures the heteroskedastic structure without requiring the error distribution to be known.

6.6 Bootstrap for Logistic Regression

For Logistic Regression, the Residual and Wild Bootstraps are generally inappropriate because the response Y_i is non-Gaussian (is binary 0 and 1).

Paired Bootstrap

The Paired Bootstrapresampling pairs $\{X_i, Y_i\}$) remains fully valid and is the default choice for GLMs.

Parametric/Model-Based Bootstrap

This specialized approach leverages the structure of the binary response:

1. Fit the logistic model to the original data to obtain estimates $\hat{\beta}_0, \hat{\beta}_1$.
2. For each observation i, calculate the estimated probability of success:

$$\hat{P}_i = \hat{P}(Y_i = 1|X_i) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_i}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 X_i}}.$$

3. Generate the bootstrap response Y_i^* from a Bernoulli distribution with this probability $\hat{P}_i$:

$$X_i^* = X_i, \quad Y_i^* \sim \text{Bernoulli}(\hat{P}_i).$$

This guarantees that the bootstrap response Y_i^* is correctly binary (0 or 1), adhering to the constraints of the model.

Chapter 7

Nonparametric Estimation.

Statistical inference commonly centers on distribution functions that are either purely parametric or purely nonparametric.

In parametric models, we begin by making rigid assumptions about the structure of the data and then proceed to estimate the parameters defining that structure as efficiently as possible. A well-specified parametric model yields accurate inferences, while an incorrect model will likely lead to mistaken conclusions.

However, in most applications, parametric models constitute only an approximation to the underlying true model, and identifying an adequate model is often not straightforward. This is where nonparametric estimation techniques emerge as more flexible alternatives to their parametric counterparts.

A commonality among nonparametric methods is the exploitation of the concept of **local smoothing**, which only utilizes the properties of continuity or local differentiability of the function being estimated. The success of local smoothing depends on the presence of a sufficient number of observations around each point of interest to provide adequate information for the estimation. Furthermore, nonparametric procedures can assist early in an investigation to uncover the probabilistic structure governing the data, thereby ensuring that the assumptions for subsequent parametric analysis are well-founded.

The basic premise in nonparametric estimation is to use the data to perform inference while making the fewest possible assumptions. In the context of this course, we will refer to nonparametric inference as a set of techniques that attempt to keep the number of assumptions as low as possible. We will focus on two key problems:

1. Density Estimation.
2. Regression Estimation.

7.1 Density Estimation

A fundamental characteristic describing the behavior of a random variable $\mathbf{X}$ is its density function, f. Knowledge of the density function is beneficial in many ways. For instance, if we have a set of observations generated from the density f and we wish to know how many observations fall within a set A, we can calculate the probability of the random variable $\mathbf{X}$ belonging to that set as an integral over A:

$$\mathbb{P}(\mathbf{X} \in A) = \int_A f(x)dx.$$

If this value is high for a certain set A compared to the probability over another set B, we can informally say that, given a set of observations, there is a high probability of finding an observation in region A and a low probability in region B. That is, the density function tells us where observations occur most frequently.

In most practical studies, the true density function f of $\mathbf{X}$ is unknown. Instead, we only have a set of observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$, which we assume are independent and identically distributed (i.i.d.) from the unknown density f. Our objective is to study how to estimate this density function based on the random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

Nonparametric estimation methods have emerged to address these issues and have been widely studied. In this Chapter, we will explore various proposals for density function estimation and analyze their properties and implementation.

7.1.1 Histogram

One of the most fundamental and intuitive approaches to nonparametric density estimation is the histogram. Unlike parametric methods that assume a specific functional form (e.g., Gaussian), histograms allow the data to dictate the shape of the density function f.

The histogram is the oldest and most popular density estimator. Its calculation requires an origin and a bin width h to specify the intervals $I_j = (x_0 + jh, x_0 + (j+1)h]$, where $(j = \ldots, -1, 0, 1, \ldots)$. The histogram counts the number of observations falling into each interval. It is then drawn such that the area under each bar is proportional to the number of observations in the bin.

Definition 30 (The Histogram Estimator)**.** *Assume a probability density function f with support on a compact interval, typically $[0, 1]$. Let the interval be partitioned into m disjoint bins $\{B_j\}_{j=1}^m$. The bins are defined as:*

$$B_j = \left[\frac{j-1}{m}, \frac{j}{m}\right), \quad j = 1, \ldots, m. \tag{7.1}$$

The width of each bin is denoted by $h = 1/m$. Let $X_1, \ldots, X_n$ be a random sample from f. Let Y_j be the count of observations falling into B_j. The estimator $\hat{f}_n(x)$ is defined as:

$$\hat{f}_n(x) = \sum_{j=1}^{m} \frac{\hat{p}_j}{h} I(x \in B_j) \tag{7.2}$$

where $\hat{p}_j = Y_j/n$ represents the empirical proportion of data in bin j, and $I(\cdot)$ is the indicator function.

The performance of the histogram is critically dependent on the choice of the binwidth h. This relationship is characterized by the trade-off between the estimation bias and the sampling variance.

Theorem 28 (Expectation and Local Variance)**.** *For a fixed $x \in B_j$, the expected value*

and variance of the histogram estimator are given by:

$$E[\hat{f}_n(x)] = \frac{p_j}{h} = \frac{1}{h}\int_{B_j} f(u)du \tag{7.3}$$

$$Var(\hat{f}_n(x)) = \frac{p_j(1-p_j)}{nh^2} \tag{7.4}$$

The Smoothing Trade-off

The choice of m (number of bins) dictates the degree of smoothing:

- **Oversmoothing (Small m, Large h):** High bias occurs because the estimator fails to capture local fluctuations of f, although the variance is low.
- **Undersmoothing (Large m, Small h):** Low bias is achieved, but the estimator becomes highly unstable (high variance), resulting in a "jagged" appearance.

Optimal Binwidth and Risk Analysis

To objectively evaluate the quality of the histogram estimator $\hat{f}_n$, we utilize the **Integrated Mean Squared Error (IMSE)**, also referred to as the Risk $R(\hat{f}_n, f)$. This metric captures the balance between the approximation error (bias) and the estimation error (variance).

Theorem 29 (Asymptotic Risk of the Histogram)**.** *Assume f' is absolutely continuous and $\int (f'(u))^2 du < \infty$. The risk of the histogram estimator $\hat{f}_n$ with binwidth h is:*

$$R(\hat{f}_n, f) = \frac{h^2}{12}\int (f'(u))^2 du + \frac{1}{nh} + o(h^2) + o\left(\frac{1}{nh}\right) \tag{7.5}$$

Proof. The Risk can be decomposed into the integral of the bias squared and the variance:

$$R(\hat{f}_n, f) = \int B^2(x)dx + \int V(x)dx$$

1. Bias Analysis: For $x \in B_j = [g_j, g_{j+1})$, where $h = g_{j+1} - g_j$, the expected value is:

$$E[\hat{f}_n(x)] = \frac{p_j}{h} = \frac{1}{h}\int_{B_j} f(u)du$$

Using a Taylor expansion of $f(u)$ around x:

$$f(u) \approx f(x) + (u-x)f'(x)$$

Integrating over the bin B_j:

$$\frac{1}{h}\int_{B_j} f(u)du \approx f(x) + f'(x)\left[\frac{1}{h}\int_{B_j}(u-x)du\right]$$

Let x_j be the midpoint of bin B_j. The bias term $b(x) = E[\hat{f}_n(x)] - f(x)$ integrated over

the bin leads to the squared bias:

$$\int B^2(x)dx \approx \int \frac{h^2}{4}\left(1-\frac{2(x-g_j)}{h}\right)^2 (f'(x))^2 dx = \frac{h^2}{12}\int (f'(x))^2 dx$$

2. Variance Analysis: The variance at point x is:

$$V(\hat{f}_n(x)) = \frac{p_j(1-p_j)}{nh^2}$$

Since $p_j \approx f(x)h$ and for large n, $(1-p_j) \approx 1$:

$$V(\hat{f}_n(x)) \approx \frac{f(x)h}{nh^2} = \frac{f(x)}{nh}$$

Integrating the variance over the support:

$$\int V(x)dx \approx \int \frac{f(x)}{nh}dx = \frac{1}{nh}\int f(x)dx = \frac{1}{nh}$$

Combining both results, we obtain the expression for $R(\hat{f}_n, f)$.
The optimal binwidth h^* is found by differentiating the Risk with respect to h:

$$\frac{d}{dh}\left(\frac{h^2}{12}\int (f'(u))^2 du + \frac{1}{nh}\right) = 0 \implies h^* = \left(\frac{6}{n\int (f'(u))^2 du}\right)^{1/3}$$

This leads to the characteristic convergence rate of $n^{-2/3}$ for histograms.

The histogram presents several disadvantages:

- It is constant over intervals (piecewise constant).
- The results depend on the chosen origin (x_0).
- The choice of the bin width (h) is crucial.
- It exhibits a slow rate of convergence.
- The discontinuities in the estimator are due to the procedure, not necessarily to the underlying distribution.

7.1.2 Kernel Estimation

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample from a population with density function $f(x)$. As mentioned, the problem is to estimate $f(x)$ from the observations. We first provide an intuitive idea of kernel density estimation.

If $\mathbf{X}$ is a random variable with a density f that is continuous at x:

$$\begin{aligned} f(x) = F'(x) &= \lim_{h\to 0}\frac{F(x+h)-F(x-h)}{2h} \\ &= \lim_{h\to 0}\frac{\mathbb{P}(x-h<\mathbf{X}<x+h)}{2h} \end{aligned}$$

A natural estimator for $\mathbb{P}(x-h < \mathbf{X} < x+h)$ is simply the sample proportion of observations that fall into the interval $(x-h, x+h)$. Thus, for a sufficiently small h, we can deduce the following estimator for $f(x)$:

$$\widetilde{f}(x) = \frac{1}{2h} \frac{\#\{\mathbf{X}_i : \mathbf{X}_i \in (x-h, x+h)\}}{n}.$$

Essentially, this estimator counts the number of observations that "fall" within a neighborhood of radius h around x. Moreover, if we let F_n be the empirical distribution function, we can write $\widetilde{f}(x)$ as:

$$\widetilde{f}(x) = \frac{F_n(x+h) - F_n(x-h)}{2h}.$$

Note that this estimator differs from the histogram because the histogram starts with a fixed partition (grid) of the real line, and to estimate $f(x)$, it calculates the proportion of observations in the interval containing x divided by the length of that interval. Consequently, the density of two points x and x' located in the same interval is estimated by the same value. The estimator $\widetilde{f}(x)$, however, calculates the proportion of observations in a neighborhood centered at x. Therefore, even if x' is within the neighborhood of x when estimating $f(x)$, the estimation of $f(x')$ may differ because the neighborhood of x' varies.

Another way to express the estimator $\widetilde{f}(x)$ is as follows:

$$\widetilde{f}(x) = \frac{1}{2h} \frac{\#\{\mathbf{X}_i : \mathbf{X}_i \in (x-h, x+h)\}}{n} = \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} I(|x - \mathbf{X}_i| < h).$$

If we define the function w as $w(u) = \frac{1}{2} I_{(|u|<1)}$, where I is the indicator function, then $\widetilde{f}(x)$ is equivalent to:

$$\widetilde{f}(x) = \sum_{i=1}^{n} \frac{1}{nh} w\Big(\frac{x - \mathbf{X}_i}{h}\Big). \tag{7.6}$$

Note that $w \geq 0$ and $\int w(u)du = 1$. Furthermore, for each $1 \leq i \leq n$, $w\Big(\frac{x-\mathbf{X}_i}{h}\Big) = \frac{1}{2}$ if and only if $\mathbf{X}_i \in (x-h, x+h)$. That is, the function w assigns a uniform weight to each observation $\mathbf{X}_i$ within the neighborhood $(x-h, x+h)$ and a weight of 0 to observations outside the neighborhood. The function w is called the uniform or Parzen kernel.

However, one might be interested in giving greater weight to observations closer to x. This is easily achieved by replacing the weight function or kernel w with a non-negative function K that satisfies the condition $\int K(u)du = 1$. Moreover, if we consider a smoother weight function K, we would obtain a smoother estimator. In general, the weights utilized decrease smoothly, thereby giving less weight to observations further away from the point x. Possible kernel options are shown in the figures below.

Funciones de Núcleo o Kernel	
Kernel	$K(u)$
Uniforme	$\frac{1}{2}\mathrm{I}(\lvert u\rvert \leq 1)$
Triangular	$(1-\lvert u\rvert)\,\mathrm{I}(\lvert u\rvert \leq 1)$
Epanechnikov	$\frac{3}{4}(1-u^2)\,\mathrm{I}(\lvert u\rvert \leq 1)$
Quartic (Biweight)	$\frac{15}{16}(1-u^2)^2\,\mathrm{I}(\lvert u\rvert \leq 1)$
Triweight	$\frac{35}{32}(1-u^2)^3\,\mathrm{I}(\lvert u\rvert \leq 1)$
Gaussiano	$\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2)$
Coseno	$\frac{\pi}{4}\cos(\frac{\pi}{2}u)\,\mathrm{I}(\lvert u\rvert \leq 1)$

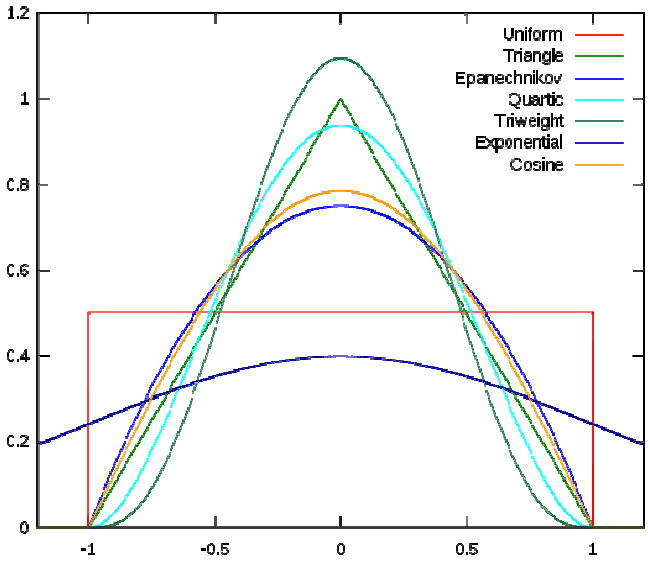


Figure 7.1.2: Kernel functions.

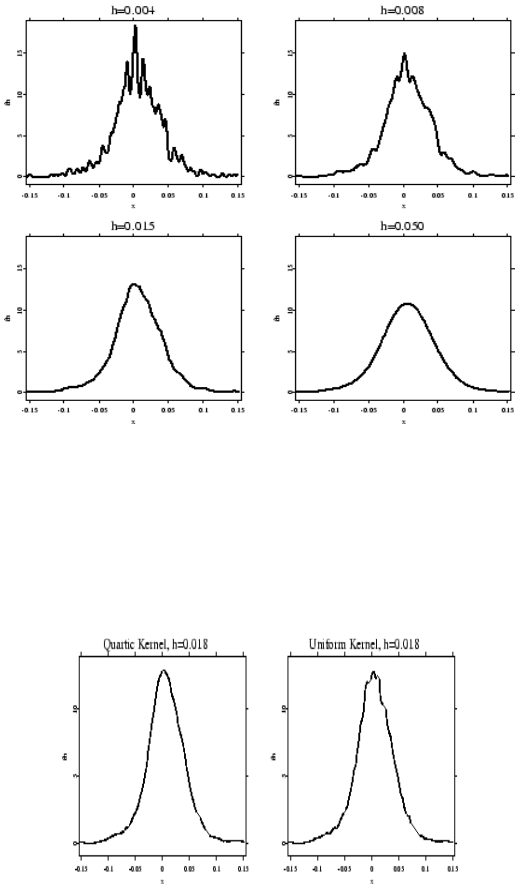
This leads to the estimator, one of the most studied nonparametric estimators, defined by Rosenblatt (1956):

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x-\mathbf{X}_i}{h}\Big) \tag{7.7}$$

where K is a kernel function, $h = h_n$ is called the *smoothing parameter* or *bandwidth*, and satisfies $h_n \to 0$ as $n \to \infty$.

These estimators are constructed at every point on the real axis based on the sample values closest to that point. That is, a neighborhood around the point where the density is to be estimated is considered, and based on the observations found in that neighborhood, the estimator is constructed, giving greater weight to observations closer to the point and lesser weight to those further away, within the neighborhood. The weighting is typically established using various weighting functions called kernels. The neighborhoods are defined by a smoothing parameter or window; to visualize them, one can imagine a sphere centered at the point of estimation whose radius corresponds precisely to the bandwidth.

The smoothing parameter h is often a crucial factor in the estimation process, as its name indicates, it is highly related to the level of smoothing introduced in the estimation. Figure 7.1.2a) illustrates the influence of the window width choice for a given dataset, and Figure 7.1.2b) shows the influence of the kernel on the estimation.



a) b)

Figure 7.1.2: a)Density estimator for different bandwidths. Data corresponding to stock returns. b)Density estimator for different kernels. Data corresponding to stock returns.

The properties of the density estimator depend on the choice of the kernel K and the bandwidth h. The combination of the weighting function, the window width, and the sample size determines the quality of the resulting estimation.

- Too small a bandwidth will lead to highly variable estimators, as the neighborhoods at each point will lack sufficient observations to base the estimation on.

- Conversely, too large a bandwidth will produce overly smooth estimators that fail to capture the local structure of the density, resulting in biased estimators.

Note that if $\int_{-\infty}^{+\infty} K(x)dx = 1$ and $K \geq 0$, then the estimator $\hat{f}(x)$ is also a density function, meaning $\int_{-\infty}^{+\infty} \hat{f}(x)dx = 1$. This is proven as follows:

$$\int_{-\infty}^{+\infty} \hat{f}(x)dx = \int_{-\infty}^{+\infty} \sum_{i=1}^{n} \frac{1}{nh} K\Big(\frac{x - \mathbf{X}_i}{h}\Big)dx =$$

$$\frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{+\infty} \frac{1}{h} K\Big(\frac{x - \mathbf{X}_i}{h}\Big)dx = \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{+\infty} K(s)ds = 1$$

by the change of variables $s = (x - \mathbf{X}_i)/h$.

Moreover, the smoothness conditions imposed on the kernel will be inherited by the resulting density function. That is, if the kernel is a continuous function, the associated density estimator will also be continuous.

Observe that once the bandwidth and kernel are fixed, the density estimator is unique for the given dataset and does not depend on the "origin," unlike the histogram. The choice of the kernel is usually a positive function to ensure the estimator is indeed a density; however, in some circumstances, kernels with some negative values may be considered, which does not always imply that the resulting estimator will also take negative values.

7.1.3 Properties of the Kernel Density Estimator

One of the first properties analyzed for any estimator is the decomposition into bias and variance.

Theorem 30. *Under the following assumptions:*

i) The density f is twice continuously differentiable such that $\int |f''(s)|ds < \infty$.

ii) The kernel K is a symmetric density function such that $\int K(s)ds = 1$, $\int K(s)sds = 0$, and $\mu_2(K) = \int K(s)s^2ds < \infty$.

The expected value of the kernel estimator $\widetilde{f}(x)$ is:

$$\mathbb{E}\left[\widetilde{f}(x)\right] = f(x) + \frac{h^2}{2} f''(x)\mu_2(K) + o(h^2) \quad \text{as } h \to 0 \text{ for each } x.$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\widetilde{f}(x)\right] &= \mathbb{E}\Big(\frac{1}{hn}\sum_{i=1}^{n} K\Big(\frac{x-\mathbf{X}_i}{h}\Big)\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\Big(\frac{1}{h}K\Big(\frac{x-\mathbf{X}_i}{h}\Big)\Big) \\
&= \mathbb{E}\Big(\frac{1}{h}K\Big(\frac{x-\mathbf{X}_1}{h}\Big)\Big) \\
&= \frac{1}{h}\int_{-\infty}^{+\infty} K\Big(\frac{x-u}{h}\Big) f(u)du \quad \text{(using } u \text{ as the integration variable)}
\end{aligned}$$

Using the change of variables $y = (x-u)/h$ (which implies $u = x - hy$ and $du = -hdy$):

$$\mathbb{E}\left[\widetilde{f}(x)\right] = \int_{-\infty}^{+\infty} K(y) f(x-hy)dy$$

We apply the second-order Taylor expansion of $f(x-hy)$ around x: $f(x-hy) = f(x) - f'(x)hy + \frac{f''(x)}{2}h^2y^2 + o(h^2)$.

$$\begin{aligned}
\mathbb{E}\left[\widetilde{f}(x)\right] &= \int_{-\infty}^{+\infty} K(y)\left[f(x) - f'(x)hy + \frac{f''(x)}{2}h^2y^2 + o(h^2)\right]dy \\
&= f(x)\int_{-\infty}^{+\infty} K(y)dy - f'(x)h\int_{-\infty}^{+\infty} K(y)ydy \\
&\quad + \frac{f''(x)}{2}h^2\int_{-\infty}^{+\infty} K(y)y^2dy + o(h^2)\int_{-\infty}^{+\infty} K(y)dy
\end{aligned}$$

By the assumptions on the kernel: $\int K(y)dy = 1$ and $\int K(y)ydy = 0$.

$$\mathbb{E}\left[\widetilde{f}(x)\right] = f(x) + 0 + \frac{h^2}{2}f''(x)\mu_2(K) + o(h^2)$$

Consequently, the asymptotic bias of the estimator is:

$$\text{Bias}\left(\widetilde{f}(x)\right) = \mathbb{E}\left[\widetilde{f}(x)\right] - f(x) = h^2\frac{f''(x)}{2}\mu_2(K) + o(h^2)$$

This result shows that larger bandwidths h increase the bias, and that the bias depends on $f''(x)$, the curvature of the true function. For example, the bias will be negative if the second derivative is negative (i.e., the function has a local maximum), meaning the estimator is expected to undershoot the true density at that point.

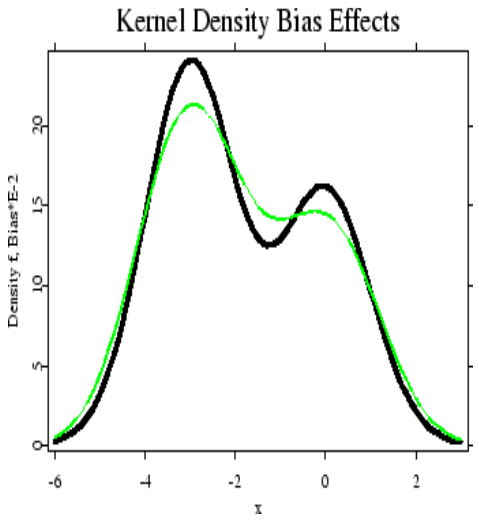


Figure 7.1.3: Density estimator (in green) and true density (in black).

Theorem 31. *Under the same assumptions, the asymptotic variance of the estimator is:*

$$Var(\widetilde{f}(x)) = \frac{1}{nh}\|K\|^2 f(x) + o\left(\frac{1}{nh}\right)$$

where $\|K\|^2 = \int K^2(s)ds$ *and*

$$MSE(\widetilde{f}(x)) = h^4 \frac{(f''(x))^2}{4}\mu_2^2(K) + o(h^4) + \frac{1}{nh}\|K\|^2 f(x) + o\left(\frac{1}{nh}\right)$$

Proof. Let $K_h(u) = \frac{1}{h}K(u/h)$. Since the $\mathbf{X}_i$'s are i.i.d., $\text{Var}(\sum Z_i) = \sum \text{Var}(Z_i)$:

$$\begin{aligned}\text{Var}(\widetilde{f}(x)) &= n^{-2}\text{Var}\Big(\sum_{i=1}^{n} K_h(x - \mathbf{X}_i)\Big) = n^{-2}\sum_{i=1}^{n}\text{Var}(K_h(x - \mathbf{X}_i)) \\ &= n^{-1}\text{Var}(K_h(x - \mathbf{X}_1)) = n^{-1}\left[\mathbb{E}(K_h^2(x - \mathbf{X}_1)) - \mathbb{E}^2(K_h(x - \mathbf{X}_1))\right]\end{aligned}$$

The second term $\mathbb{E}^2(K_h(x - \mathbf{X}_1))$ is $O(1)$. Using a change of variable and Taylor expansion (similar to the bias derivation), we find the first term:

$$\mathbb{E}(K_h^2(x - \mathbf{X}_1)) = h^{-1}\int K^2(y)f(x - hy)dy = \frac{f(x)}{h}\|K\|^2 + O(1)$$

Substituting back:

$$\mathrm{Var}(\widetilde{f}(x)) = \frac{1}{n}\left[\frac{f(x)}{h}\|K\|^2 + O(1) - O(1)\right] = \frac{1}{nh}\|K\|^2 f(x) + O\left(\frac{1}{n}\right)$$

This result indicates that larger values of nh lead to smaller variance. Similarly, the variance is smaller if $\|K\|^2$ is small, which corresponds to a flatter kernel.

We combine the bias and variance to obtain the asymptotic Mean Squared Error (MSE) for each x:

$$\mathrm{MSE}(\widetilde{f}(x)) = \mathrm{Bias}^2\left(\widetilde{f}(x)\right) + \mathrm{Var}(\widetilde{f}(x))$$

$$\mathrm{MSE}(\widetilde{f}(x)) = h^4\frac{(f''(x))^2}{4}\mu_2^2(K) + o(h^4) + \frac{1}{nh}\|K\|^2 f(x) + o\left(\frac{1}{nh}\right)$$

This reveals the fundamental bias-variance trade-off:

- Small h reduces the bias ($O(h^2)$) but increases the variance ($O(1/nh)$).
- Large h reduces the variance but increases the bias, leading to an overly smooth estimator.

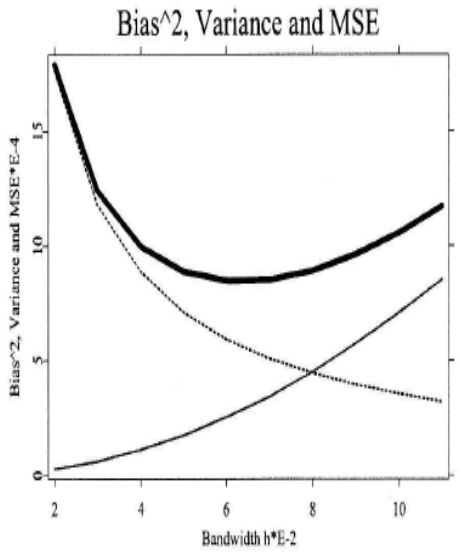


Figure 7.1.3: Squared Bias (solid line); Variance (dashed line) and Mean Squared Error (thick solid line).

A corollary is the weak consistency of the estimator.

Corollary 2. *If $h \to 0$ and $nh \to \infty$ as $n \to \infty$, then $\widetilde{f}(x) \xrightarrow{\mathbb{P}} f(x)$ for each x.*

Selection of the Kernel and the Bandwidth

Figure 7.1.3 illustrates the bias-variance trade-off, underscoring the importance of selecting the appropriate bandwidth h. A natural choice for the bandwidth would be the one that minimizes the Mean Squared Error (MSE).

Recall that the MSE for the estimator is:

$$\text{MSE}(\widetilde{f}(x)) = h^4 \frac{(f''(x))^2}{4} \mu_2^2(K) + \frac{1}{nh} \|K\|^2 f(x) + o\left(\frac{1}{nh}\right) + o(h^4)$$

Assuming $h \to 0$ and $nh \to \infty$ as $n \to \infty$, we can disregard the lower-order terms and seek the value of h that minimizes the approximate MSE:

$$\text{MSE}(\widetilde{f}(x)) \approx h^4 \frac{(f''(x))^2}{4} \mu_2^2(K) + \frac{1}{nh} \|K\|^2 f(x)$$

By taking the derivative with respect to h and setting it to 0, we obtain the **locally optimal bandwidth**:

$$\begin{aligned} h_{\text{opt}}(x) &= \left(\frac{\|K\|^2 f(x)}{(f''(x))^2 \mu_2^2(K) n} \right)^{1/5} \\ &= \left(\frac{\|K\|^2 f(x)}{(f''(x))^2 \mu_2^2(K)} \right)^{1/5} n^{-1/5} \end{aligned}$$

This optimal bandwidth $h_{\text{opt}}(x)$ depends on unknown quantities, namely $f(x)$ and $f''(x)$, and thus cannot be computed in practice. Furthermore, the obtained bandwidth is local, as it depends on the specific point x at which the density is being estimated.

These drawbacks can be addressed by considering the **Mean Integrated Squared Error (MISE)**:

$$\text{MISE}(\widetilde{f}) = \int \text{MSE}(\widetilde{f}(x)) dx$$

The integral of the error over the entire support is:

$$\text{MISE}(\widetilde{f}) = h^4 \frac{\int (f''(x))^2 dx}{4} \mu_2^2(K) + \frac{1}{nh} \|K\|^2 \int f(x) dx + o\left(\frac{1}{nh}\right) + o(h^4)$$

Since $\int f(x)dx = 1$, neglecting the smaller order terms gives the Asymptotic Mean Integrated Squared Error (AMISE):

$$\text{AMISE}(\widetilde{f}) = h^4 \frac{\int (f''(x))^2 dx}{4} \mu_2^2(K) + \frac{1}{nh} \|K\|^2.$$

Minimizing the AMISE analogously yields the globally optimal bandwidth:

$$h_{\text{opt}} = \left(\frac{\|K\|^2}{\|f''\|^2 \mu_2^2(K)} \right)^{1/5} n^{-1/5}$$

where $\|f''\|^2 = \int (f''(x))^2 dx$. This global bandwidth no longer depends on the specific point x or the value of $f(x)$ but still depends on the unknown integral $\|f''\|^2$.

If we substitute the optimal bandwidth h_{opt} back into the AMISE expression, we find the minimum error rate:

$$\text{AMISE}(\widetilde{f})(h_{\text{opt}}) = \frac{5}{4} \left(\|f''\| \mu_2(K) \right)^{2/5} \|K\|^{8/5} n^{-4/5}$$

This confirms the convergence rate of $O(n^{-4/5})$. For comparison, the optimal histogram estimator achieves a slower rate of $O(n^{-2/3})$, providing a strong argument for the superiority of the kernel density estimator.

Methods for Bandwidth Selection

To select the smoothing parameter in practice, two main approaches are presented: Cross-Validation and the Plug-in method.

Plug-in Method

The plug-in method is a standard estimation technique that involves replacing the unknown parameters in an expression with estimators. To obtain a computable estimator for the optimal bandwidth h_{opt}, it suffices to estimate $\|f''\|^2$, as the constants depending on the kernel ($\|K\|^2$ and $\mu_2(K)$) are known once the kernel is fixed.

Silverman's Rule of Thumb Silverman proposed a method under the strong assumption that the density f is Normal. In this case, $\|f''\|^2$ can be calculated analytically: $\|f''\|^2 = \sigma^{-5}\frac{3}{8\sqrt{\pi}}$. σ is then estimated from the data. For a Gaussian kernel, this yields $\hat{h}_{\text{opt}} = 1.06\widehat{\sigma}n^{-1/5}$. While assuming normality is generally ill-advised in a nonparametric context, this rule provides reasonable results for unimodal and nearly symmetric densities in practice.

Refined Plug-in A more sophisticated approach considers a nonparametric estimator of the second derivative, $\hat{f}''(x)$, which can be calculated by deriving the kernel density estimator twice with a preliminary bandwidth (which could be the one derived from Silverman's rule).

Cross-Validation Method

Cross-validation methods avoid making assumptions about the parametric family of f. The idea is to consider a measure of error between f and its estimator $\widetilde{f}$, in this case, the **Integrated Squared Error (ISE)**:

$$\begin{aligned}\text{ISE}(h) &= \int (\widetilde{f}(x) - f(x))^2 dx \\ &= \int \widetilde{f}^2(x)dx - 2\int \widetilde{f}(x)f(x)dx + \int f^2(x)dx\end{aligned}$$

We seek to minimize the expected value of ISE(h), which is MISE(h). Since $\int f^2(x)dx$ does not depend on h, and $\int \widetilde{f}^2(x)dx$ can be calculated directly from the data, the core challenge is estimating the term $\int \widetilde{f}(x)f(x)dx = \mathbb{E}(\widetilde{f}(\mathbf{X}))$.

This expectation can be estimated using the **leave-one-out** principle:

$$\widehat{\mathbb{E}(\widetilde{f}(\mathbf{X}))} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{f}_{-i}(\mathbf{X}_i)$$

where $\widetilde{f}_{-i}(\mathbf{X}_i)$ is the density estimator calculated **excluding** the i-th observation, evaluated at $\mathbf{X}_i$:

$$\widetilde{f}_{-i}(x) = \frac{1}{h(n-1)}\sum_{j=1;i\neq j}^{n} K\Big(\frac{x - \mathbf{X}_j}{h}\Big).$$

The optimal bandwidth is then estimated as $\widehat{h}_{\text{cv}} = \arg\min_h \text{CV}(h)$, where the cross-validation function $\text{CV}(h)$ is:

$$\text{CV}(h) = \int \widetilde{f}^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} \widetilde{f}_{-i}(\mathbf{X}_i)$$

This expression can be algebraically simplified for computational purposes to:

$$\text{CV}(h) = \frac{1}{n^2 h}\sum_i \sum_j K * K\left(\frac{\mathbf{X}_i - \mathbf{X}_j}{h}\right) - \frac{2}{n(n-1)h}\sum_{i=1}^{n}\sum_{j=1;i\neq j}^{n} K\left(\frac{\mathbf{X}_i - \mathbf{X}_j}{h}\right).$$

where $K * K(u) = \int K(u-v)K(v)dv$ is the convolution of the kernel with itself.

Selecting the Optimal Kernel

Regarding kernel selection, the objective is to choose the kernel K that minimizes the factor $T(K) = (\mu_2(K)^2 \|K\|^8)^{1/5}$ in the AMISE formula.

Epanechnikov (1969) showed that, among all non-negative kernels with compact support, the optimal kernel is the Epanechnikov kernel:

$$K(u) = \frac{3}{4}\left(1 - u^2\right) I(|u| \leq 1).$$

(The formula provided in the source text uses a rescaling factor which is not the standard definition of the Epanechnikov kernel, but the underlying concept of finding the kernel that minimizes $T(K)$ holds.)

Despite the existence of an optimal kernel, the ratios comparing the efficiency of common kernels (as shown in the table below) are very close to 1. The largest difference is seen with the Uniform kernel, which results in only a 6% increase in the AMISE.

Nucleo	T(K)	T(K)/T(K_{epan})
Uniform	0.3701	1.0602
Triangle	0.3531	1.0114
Epanechnikov	0.3491	1.0000
Quartic	0.3507	1.0049
Triweight	0.3699	1.0595
Gaussian	0.3633	1.0408
Cosine	0.3494	1.0004

Comparison of kernels (Efficiency Ratios).

Therefore, in practice, the choice of the kernel is far less important than the choice of the bandwidth h. For implementation, it is generally best to try several methods (Plug-in and Cross-Validation) and compare the resulting estimates.

7.1.4 Extension to the Multivariate Case

In certain situations, one may be interested in estimating the density in a multivariate context rather than in a single dimension. Therefore, it is interesting to extend the previous proposal to when we work with higher dimensions.

Consider a density function $f(\mathbf{x})$ over $\mathbb{R}^d$. We observe a sample of size n of random vectors $\mathbf{X}_i$, where:

$$\mathbf{X}_i = \begin{pmatrix} X_{i1} \\ \vdots \\ X_{id} \end{pmatrix} \qquad i = 1, \ldots, n.$$

The objective is to estimate the density $f(\mathbf{x}) = f(x_1, \ldots, x_d)$. The natural extension of the univariate proposal is to consider:

$$\begin{aligned} \widehat{f}(\mathbf{x}) &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{x_1 - X_{i1}}{h}, \ldots, \frac{x_d - X_{id}}{h}\right) \end{aligned}$$

where $K : \mathbb{R}^d \to \mathbb{R}$ is a multivariate kernel.

In this formulation, the same bandwidth h is chosen for all components, but this is not strictly necessary. We could select a distinct bandwidth for each dimension. If we let $\mathbf{h} = (h_1, \ldots, h_d)'$, the estimator becomes:

$$\widehat{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_1 \ldots h_d} K\left(\frac{x_1 - X_{i1}}{h_1}, \ldots, \frac{x_d - X_{id}}{h_d}\right).$$

Multivariate Kernels A multivariate kernel can be chosen as a **multiplicative kernel** (or product kernel), meaning:

$$K(\mathbf{u}) = K_1(u_1) \dots K_d(u_d)$$

where K_j, $1 \leq j \leq d$, is a univariate kernel. In this case, the density estimator is:

$$\widehat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\prod_{j=1}^{d} h_j} K_1 \left(\frac{x_1 - X_{i1}}{h_1} \right) \dots K_d \left(\frac{x_d - X_{id}}{h_d} \right).$$

Another alternative is to use a true multivariate kernel, such as the spherically symmetric kernel, which is obtained from a univariate function K_0:

$$K(\mathbf{u}) \propto K_0(\|\mathbf{u}\|).$$

An example is the multivariate Epanechnikov kernel:

$$K(\mathbf{u}) \propto (1 - \mathbf{u}^T \mathbf{u}) I(\mathbf{u}^T \mathbf{u} \leq 1).$$

Generalized Bandwidth Matrix A more general approach proposes using a non-singular bandwidth matrix $\mathbf{H}$ (a $d \times d$ matrix):

$$\widehat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\det(\mathbf{H})} K \left(\mathbf{H}^{-1}(\mathbf{x} - \mathbf{X}_i) \right).$$

The case of equal bandwidths in all components corresponds to $\mathbf{H} = h\mathbf{I}_d$, where $\mathbf{I}_d$ is the identity matrix, and $\det(\mathbf{H}) = h^d$.

The Curse of Dimensionality Similar to the univariate case ($d = 1$), the bias and variance of the multivariate estimator can be computed, leading to an expression for the AMISE which allows for the calculation of the optimal bandwidth. The resulting optimal bandwidth h_{opt} and the minimum AMISE scale with dimension d as follows:

$$h_{\text{opt}} \sim n^{-1/(4+d)} \quad \text{and} \quad \text{AMISE}(h_{\text{opt}}) \sim n^{-4/(d+4)}.$$

When $d = 1$, the optimal rate of convergence is $O(n^{-4/5})$. As the dimension d increases, the exponent $\frac{4}{d+4}$ decreases sharply towards zero, meaning the rate of convergence slows down dramatically. This phenomenon is known as the **curse of dimensionality**.

For example, to maintain the same statistical accuracy achieved with n observations in $d = 1$, the number of observations required for a density in d dimensions is $n^{1+d/4}$. The table below illustrates the rapid increase in required sample size:

$n^{-4/(4+d)}$	$d=1$	$d=2$	$d=3$	$d=5$	$d=10$
$n=100$	0.025	0.046	0.072	0.129	0.268
$n=1000$	0.004	0.010	0.019	0.046	0.139
$n=100'000$	$1.0\,10^{-4}$	$4.6\,10^{-4}$	$13.9\,10^{-4}$	0.006	0.037

Comparison of dimension and error (The Curse of Dimensionality).

For this reason, kernel density estimators are typically used only for very low dimensions, such as $d = 2$ or $d = 3$.

The criteria for bandwidth selection (Plug-in and Cross-Validation) introduced for the univariate case can be extended to the multivariate setting to select the optimal bandwidth matrix $\mathbf{H}$.

7.1.5 Confidence Intervals and Bands

To obtain confidence intervals, it is necessary to determine the distribution of the estimator. While the exact distribution remains unknown, it is possible to derive the asymptotic behavior of the estimator. Under certain regularity conditions:

1. $h_n \to 0$
2. $nh_n \to \infty$
3. x has a density f that is continuous at x and twice differentiable.
4. $K : R \to R$ is bounded, $\int K = 1$, $\int u^2 K(u) > 0$, and has compact support.

It can be proven that if $h_n = cn^{-1/5}$, then:

$$\sqrt{nh}(\widetilde{f}(x) - f(x)) \xrightarrow{\mathcal{D}} N\left(\frac{c^{5/2}}{2} f''(x)\mu_2(K), f(x)\|K\|^2\right)$$

This yields the following confidence interval with an approximate level of $1 - \alpha$:

$$\left[\widetilde{f}(x) - \frac{h^2}{2} f''(x)\mu_2(K) - z_{\alpha/2}\sqrt{\frac{f(x)\|K\|^2}{nh}}, \widetilde{f}(x) - \frac{h^2}{2} f''(x)\mu_2(K) + z_{\alpha/2}\sqrt{\frac{f(x)\|K\|^2}{nh}}\right]$$

If h is sufficiently small, the term involving the second derivative can be neglected, resulting in the following interval:

$$\left[\widetilde{f}(x) - z_{\alpha/2}\sqrt{\frac{\widetilde{f}(x)\|K\|^2}{nh}}, \widetilde{f}(x) + z_{\alpha/2}\sqrt{\frac{\widetilde{f}(x)\|K\|^2}{nh}}\right]$$

Otherwise, the second derivative can be estimated by differentiating a kernel estimator using a pilot bandwidth g.

It is important to note that this interval is valid only for $f(x)$ at a specific point and not for the entire density function. To derive confidence bands for the entire function,

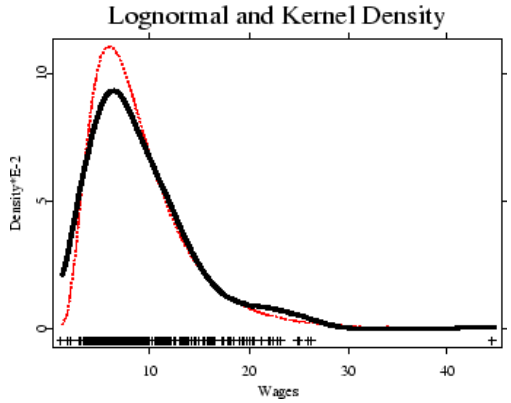
other techniques must be employed. Bickel and Rosenblatt (1973) proved the following result: let f be a density function defined on $(0,1)$, $h_n = n^{-\delta}$ with $\delta \in (1/5, 1/2)$, then for all $x \in (0,1)$:

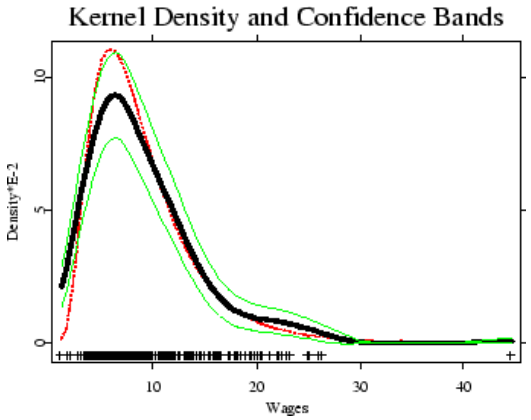
$$\lim_{n\to\infty} P\left(\widetilde{f}(x) - \sqrt{\frac{\widetilde{f}(x)\|K\|^2}{nh}} \left\{ \frac{z}{2\delta \log n} + d_n \right\}^{-1/2} \leq f(x) \leq \widetilde{f}(x) + \sqrt{\frac{\widetilde{f}(x)\|K\|^2}{nh}} \left\{ \frac{z}{2\delta \log n} + d_n \right\}^{-1/2} \right)$$
$$= \exp\{-2\exp\{-z\}\}$$

where $d_n = (2\delta \log n)^{1/2} + (2\delta \log n)^{-1/2} \log\left\{ \frac{\|K'\|_2}{2\pi \|K\|_2} \right\}$.

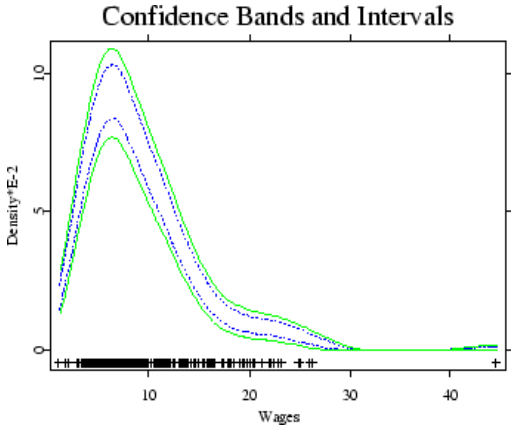
Thus, to find a confidence band of level α, it suffices to find the value of z that satisfies $\exp(-2\exp(-z)) = 1 - \alpha$. For example, if $\alpha = 0.05$, then $z \approx 3.663$.

The following example corresponds to data on the average hourly earnings of 534 randomly selected workers in the United States during May 1985.





We can observe that near the mode, the parametric estimator lies outside the confidence band; therefore, we would reject the hypothesis that the true distribution is lognormal. However, the parametric estimation appears to capture the shape of the distribution fairly well. Non-parametric tests or intervals often suffer from a loss of efficiency, but it is possible to find non-parametric tests with better convergence rates.



7.1.6 k-Nearest Neighbors Density Estimation

If we apply kernel density estimators to data from distributions with heavy tails, using a bandwidth small enough to accurately estimate the central part of the distribution will fail to correctly estimate the tails. Conversely, a large bandwidth suitable for estimating the tails will obscure the details occurring in the main part of the distribution.

To overcome these flaws, an estimator was proposed that is conceptually similar to the one studied by Rosenblatt, but whose neighborhoods are not fixed. Instead, they adapt to the point at which the estimation is being performed. These are known as **generalized k-nearest neighbor kernel estimators**.

As previously mentioned, choosing the bandwidth value is a non-trivial problem. An h that is too small causes the variance of the estimator to increase significantly, as very few observations are considered at each point. On the other hand, an excessively high value yields results with high bias, as it averages too many observations that fail to capture the trend or shape of the curve. This trade-off in the selection of h is known as the **bias-variance trade-off**.

One way to solve this problem is to consider variable neighborhoods. That is, instead of fixing a window width and estimating the density function based on the sample values that fall within it, the idea is to construct, at each point where we wish to estimate, neighborhoods that contain a fixed number of observations. More precisely, let $d(x, y) = |x - y|$ be the distance between two points x and y. For each value of x, consider the distances $d(x, X_i)$ for $1 \leq i \leq n$ and let $d_i(x)$ be the ordered distances; that is, $d_i(x) = (d(x, X_i))^{(i)}$, the i-th order statistic of the distances to point x.

We define the density estimator using the **k-th nearest neighbor method** as:

$$\widehat{f}(x) = \frac{k}{2nd_k(x)} \tag{7.8}$$

To better understand this definition, recall that for a sample of size n, one would expect approximately $2hnf(x)$ observations within the interval $[x-h, x+h]$ for any $h > 0$. On the other hand, exactly k observations will fall within the interval $[x-d_k(x), x+d_k(x)]$. Thus, it is reasonable to expect k to be approximately $2d_k(x)nf(x)$. From this relationship, we obtain the k-nearest neighbor estimator proposed in (7.8).

While standard fixed-bandwidth estimators are based on the number of observations lying within an interval of fixed length centered at the point of interest, the **k-th nearest neighbor estimator** is inversely proportional to the size of the interval that contains a given number k of observations.

It is possible to generalize the k-th nearest neighbor estimator by combining fixed kernels with variable neighborhoods, obtaining the following estimator:

$$\widehat{f}(x) = \frac{1}{nd_k(x)} \sum_{i=1}^{n} K\left(\frac{x - X_i}{d_k(x)}\right) \tag{7.9}$$

where K is a kernel function with the same properties as defined previously, $k = k_n$ is a sequence such that $k_n \to \infty$ as $n \to \infty$, and $d_k(x)$ is the distance between x and its k-th nearest neighbor.

7.2 Nonparametric Regression: Nonparametric Models

The regression curve describes the relationship between two variables: an explanatory variable $\mathbf{X}$ and a response variable Y. Once $\mathbf{X}$ is observed, the mean value of Y is given

by the regression function. In many situations, having knowledge of this relationship is of great interest.

Given a sample $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, the goal is to estimate the conditional expectation, i.e., $m(\mathbf{X}_i) = \mathbb{E}(Y_i|\mathbf{X}_i)$, without making any rigid assumption about the function m, such as linearity, monotonicity, or a quadratic relationship. This relationship is commonly modeled as:

$$Y_i = m(\mathbf{X}_i) + \varepsilon_i \qquad i = 1, \ldots, n$$

where ε_i are independent random variables with zero mean, representing the variation of Y_i around $m(\mathbf{X}_i)$.

Before proceeding, let us recall how to calculate the conditional expectation in the case of a joint density. Let $\mathbf{X}$ and Y be two random variables with joint density $f(x, y)$. The conditional expectation of Y given $\mathbf{X} = x$ can be calculated as:

$$\begin{aligned} \mathbb{E}(Y|\mathbf{X} = x) &= \int y f(y|x) dy = \int y \frac{f(x,y)}{f_{\mathbf{X}}(x)} dy \\ &= \frac{r(x)}{f_{\mathbf{X}}(x)} = m(x) \end{aligned}$$

where $r(x) = \int y f(x, y) dy$ and $f_{\mathbf{X}}(x)$ is the marginal density of $\mathbf{X}$.

Example 49. *If we consider a simple joint density $f(x, y) = x + y$ for $0 < x < 1$ and $0 < y < 1$, the marginal density is $f_X(x) = x + \frac{1}{2}$ for $0 < x < 1$. The conditional expectation is:*

$$\mathbb{E}(Y|X = x) = \int y \frac{x + y}{x + \frac{1}{2}} dy = \frac{\frac{1}{2}x + \frac{1}{3}}{x + \frac{1}{2}} = m(x).$$

As seen in this example, the dependency structure given by the conditional expectation is not linear.

The objective of this section is to provide estimation mechanisms for the function m with the minimum number of assumptions.

7.2.1 Kernel Estimation (Nadaraya–Watson)

We will study the estimator proposed by Nadaraya (1964) and Watson (1964). Based on the previous derivation, if $(\mathbf{X}, Y)$ has a joint density, the regression function is:

$$m(x) = \int y \frac{f(x, y)}{f_{\mathbf{X}}(x)} dy.$$

Since $f(x, y)$ and $f_{\mathbf{X}}(x)$ are unknown, a simple idea is to use a **plug-in** approach, replacing these density functions with the estimators studied in the previous chapter.

We consider the kernel density estimators for the joint and marginal densities:

$$\widehat{f}_{h,g}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - \mathbf{X}_i}{h}\right) \frac{1}{g} K\left(\frac{y - Y_i}{g}\right)$$

and

$$\widehat{f}_{\mathbf{X}}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - \mathbf{X}_i}{h}\right).$$

The numerator of the regression function estimator is obtained by integrating $y\widehat{f}_{h,g}(x,y)$ with respect to y:

$$\begin{aligned}
\int y\widehat{f}_{h,g}(x,y)dy &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-\mathbf{X}_i}{h}\right)\int y\frac{1}{g}K\left(\frac{y-Y_i}{g}\right)dy \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-\mathbf{X}_i}{h}\right)\int (sg+Y_i)K(s)ds \quad (\text{with } s=(y-Y_i)/g) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-\mathbf{X}_i}{h}\right)\left(g\int sK(s)ds + Y_i\int K(s)ds\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-\mathbf{X}_i}{h}\right)Y_i \quad \left(\text{since } \int sK(s)ds = 0 \text{ and } \int K(s)ds = 1\right)
\end{aligned}$$

Thus, the Nadaraya–Watson estimator $\widehat{m}_h(x)$ is defined as the ratio of the estimated conditional mean and the marginal density estimate:

$$\widehat{m}_h(x) = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-\mathbf{X}_i}{h}\right)Y_i}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-\mathbf{X}_i}{h}\right)} = \frac{\sum_{i=1}^{n}K\left(\frac{x-\mathbf{X}_i}{h}\right)Y_i}{\sum_{i=1}^{n}K\left(\frac{x-\mathbf{X}_i}{h}\right)}.$$

Interpretation as a Local Average

The Nadaraya–Watson estimator can be interpreted as a local average of the response variables:

$$\widehat{m}_h(x) = \sum_{i=1}^{n} W_{ni}(x)Y_i$$

where the weights $W_{ni}(x)$ are:

$$W_{ni}(x) = \frac{K\left(\frac{x-\mathbf{X}_i}{h}\right)}{\sum_{j=1}^{n}K\left(\frac{x-\mathbf{X}_j}{h}\right)}.$$

Note that the weights sum to one: $\sum_{i=1}^{n} W_{ni}(x) = 1$.

More precisely, the estimator locally averages the observations Y_i with weights that depend on the proximity of the explanatory variables $\mathbf{X}_i$ to the point x where we want to estimate the function.

Interpretation of the Nadaraya–Watson Estimator as a Local Average

Let us see how the estimator works with an example. The following graph shows a set of simulated data.

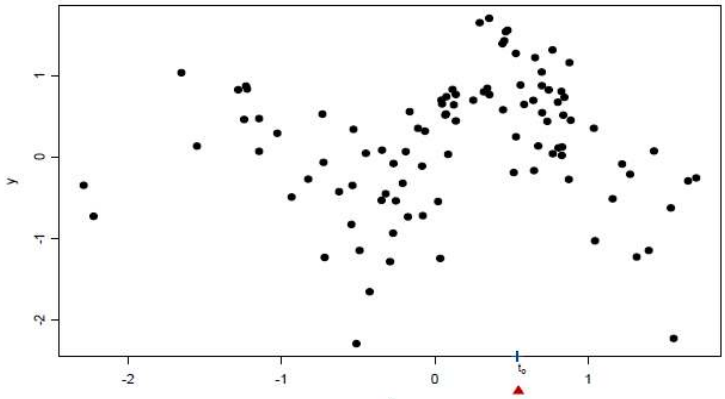


Figure 7.2.1: Example of simulated data.

At point t_0, we estimate $m(t_0) = \mathbb{E}(Y|X = t_0)$. The Nadaraya–Watson estimator can be interpreted as a local average, specifically:

$$\widehat{m}_h(x) = \sum_{i=1}^{n} W_{ni}(t_0) Y_i$$

where the weights $W_{ni}(t_0)$ are:

$$W_{ni}(t_0) = \frac{K\left(\frac{t_0 - \mathbf{X}_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{t_0 - \mathbf{X}_j}{h}\right)}$$

and $\sum_{i=1}^{n} W_{ni}(t_0) = 1$.

More precisely, the estimator works by locally averaging the observations Y_i with weights that depend on the proximity of the $\mathbf{X}_i$ variables to the point t_0 where we want to estimate the function.

If we consider the uniform kernel $K(u) = \frac{1}{2} I_{[-1,1]}(u)$ and a bandwidth $h = 0.3$, then at t_0 we look at the neighborhood $(t_0 - 0.3, t_0 + 0.3)$ and average the observations Y_i such that their respective $\mathbf{X}_i$ belong to this neighborhood.

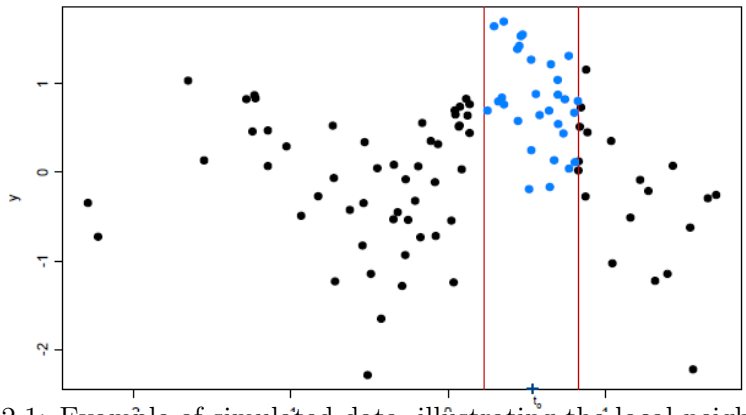


Figure 7.2.1: Example of simulated data, illustrating the local neighborhood.

Just as in the case of density estimators, the role of the bandwidth is very important in the estimation process, and as seen in the following figure, it determines the degree of smoothness of the estimated function.

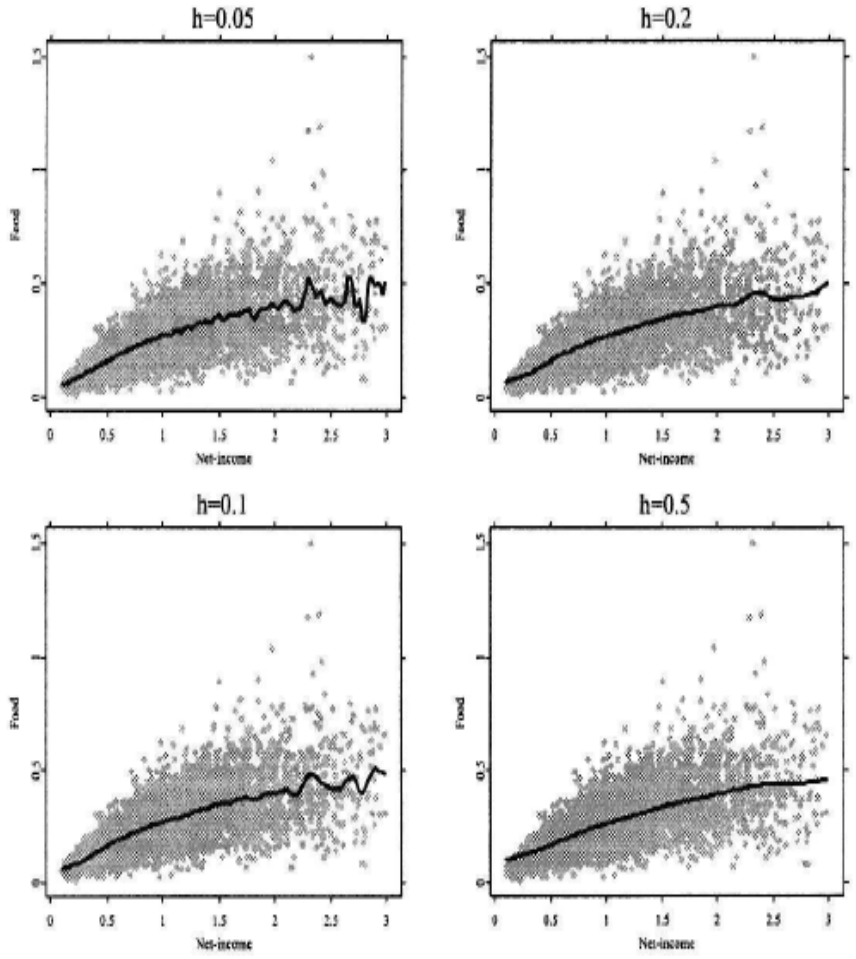


Figure 7.2.1: Average income vs. average food expenditure in England in 1973 ($n = 7125$).

In general, it can be seen that with very small bandwidths, the estimator tends to **interpolate** the data at the sample points, while very large bandwidths tend towards **constant estimators** around $\bar{Y}$.

Interpretation as Weighted Least Squares

The Nadaraya–Watson estimator can also be viewed through the lens of local optimization. If we consider the following objective function for a fixed point x:

$$M(\theta) = \sum_{i=1}^{n} W_{ni}(x)(Y_i - \theta)^2$$

and we seek the value of θ that minimizes $M(\theta)$ for each x, it is easy to see that the solution is the weighted mean:

$$\arg\min_{\theta} M(\theta) = \sum_{i=1}^{n} W_{ni}(x) Y_i = \widehat{m}_h(x).$$

That is, the nonparametric regression estimator is a **weighted least squares** estimator, where the weights are calculated locally at the point x being estimated.

7.2.2 Properties and Optimal Bandwidth

Under certain regularity hypotheses, an expression for the Mean Squared Error (MSE) of the Nadaraya–Watson estimator can be calculated:

$$\mathrm{MSE}(\widehat{m}(x)) = \frac{1}{nh}\frac{\sigma^2(x)}{f(x)}\|K\|^2 + \frac{h^4}{4}\left[m''(x) + 2\frac{m'(x)f'(x)}{f(x)}\right]^2 \mu_2^2(K) + o((nh)^{-1}) + o(h^4)$$

where $\mu_2^2(K) = \int u^2 K(u)du$ and $\sigma^2(x) = \mathrm{Var}(Y|\mathbf{X} = x)$.

Consistency and Convergence Rate It can be verified that if $h \to 0$ and $nh \to \infty$ as $n \to \infty$, then $\widehat{m}(x) \xrightarrow{\mathbb{P}} m(x)$, confirming weak consistency.

Optimal Bandwidth As in density estimation, the optimal bandwidth can be calculated by minimizing the Asymptotic Mean Squared Error (AMSE) with respect to h. In this case, the optimal bandwidth h_{opt} is proportional to $n^{-1/5}$:

$$h_{\text{opt}} \propto n^{-1/5}$$

Substituting this optimal bandwidth back into the AMSE expression, we find that the $\mathrm{AMSE}(h_{\text{opt}})$ is of the order:

$$\mathrm{AMSE}(h_{\text{opt}}) = O(n^{-4/5}).$$

As expected, the nonparametric estimator has a slower convergence rate than parametric linear regression estimators (which converge at $O(n^{-1})$) but shares the same optimal order as the nonparametric density estimator.

7.2.3 k-Nearest Neighbors (k-NN)

The kernel estimators defined previously can be viewed as a weighted average of the response variable within a fixed interval of width h around x.

The k-nearest neighbors estimator can also be seen as a weighted average of the response, but within a neighborhood of variable width: the values involved in the average now correspond to the k observations whose X values are the k closest to the point of interest x.

$$\widehat{m}_k(x) = \frac{1}{n}\sum_{i=1}^{n} W_{ki}(x) Y_i$$

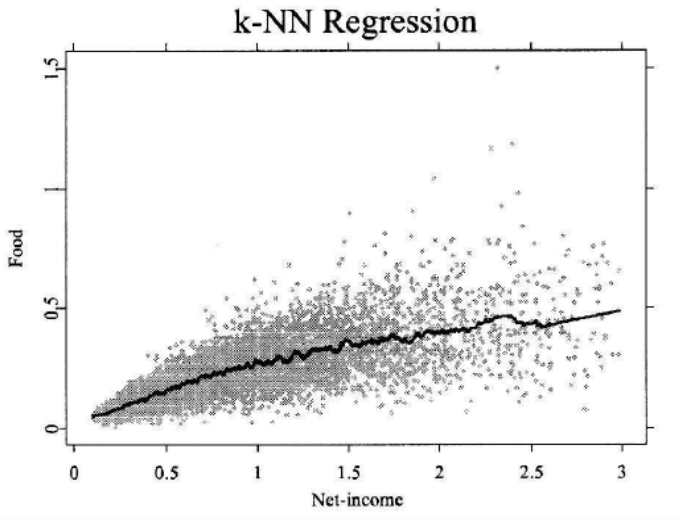
where $W_{ki}(x) = \frac{n}{k}$ if $\mathbf{x}_i$ is one of the k observations closest to x, and 0 otherwise. The parameter k is related to the smoothness of the estimation; increasing k will lead to

a smoother estimator. When x is located in a sparse region, the points falling into the interval may be far from x, resulting in estimators with high bias.

The previous estimator can be thought of as a kernel estimator with a uniform kernel. If we let $d_k(x)$ be the largest distance between x and its k-th nearest neighbor, we can write the estimator as follows:

$$\widehat{m}_k(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-\mathbf{x}_i}{d_k(x)}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{x-\mathbf{x}_i}{d_k(x)}\right)}.$$

This can be generalized by using other kernels instead of just the uniform one, giving rise to what are known as k-nearest neighbor kernel estimators.



Expressions for the bias and variance of these estimators can be obtained. The following table compares both expressions:

	núcleo	k- NN
sesgo	$h^2 \frac{(m''f+2m'f')(x)}{2f(x)} \mu_2(K)$	$\left(\frac{k}{n}\right)^2 \frac{(m''f+2m'f')(x)}{8f^3(x)} \mu_2(K)$
varianza	$\frac{\sigma^2(x)}{nhf(x)} \|\|K\|\|_2^2$	$\frac{2\sigma^2(x)}{k} \|\|K\|\|_2^2$

Note that if $\frac{h^2}{2f(x)} = \left(\frac{k}{n}\right)^2 \frac{1}{8f^3(x)}$, i.e., if $k = 2nhf(x)$, the biases coincide. However, this depends on the marginal distribution of x, which is unknown. Furthermore, under this same restriction, the variances would also coincide, making the estimators equivalent. The key point to highlight here is that the number of neighbors k must have the same order as nh.