Mathematical Statistics

Daniela Rodriguez

Universidad Torcuato Di Tella and CONICET

November 24, 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:*

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

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 $X \in \{0, 1, 2, \dots, n\}$ 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:

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

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(θ_1, θ_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 $\mathbf{X} = (X_1, \dots, X_q)$ be a random vector whose distribution belongs to a one-parameter exponential family with density given by*

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

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 $\boldsymbol{\eta} = (\eta_1,\dots,\eta_k)^\top$ 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 $\mathbf{X} = (X_1,\dots,X_n)$ be a random sample of size n, where $X_1,\dots,X_n$ 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*

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

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 $X_1, \dots, X_n$ 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 $X_1, \dots, X_n$ 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 $X_1, \dots, X_n$ 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 $X_1, \ldots, X_n \sim \text{iidPoisson}(\lambda)$. 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, $X_i \sim \text{Poisson}(\lambda)$. 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

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

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 $X_1, \dots, X_n$ 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 $X_1, \dots, X_n$ 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 $X_1, \dots, X_n$ 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 $\theta_1, \dots, \theta_p$, 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:*

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

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

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

provided it exists.

Example 14. *Suppose that $X_1, X_2, \dots, X_n$ 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, $x_1, x_2, \dots, x_n$, 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:

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

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

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

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 $X_1, \dots, X_n$ 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*

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

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

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

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.,

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

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:

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

Then **maximum likelihood estimator** of θ_0 is

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

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:

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

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

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

and the log-likelihood is

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

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

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

Solving for p, we get the MLE:

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

Also,

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

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

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

The log-likelihood is

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

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

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

Solving for λ, we get the MLE:

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

And

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

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

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

The likelihood function is

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

and thus the log-likelihood is

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

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 $X_1, \dots, X_n \overset{iid}{\sim} N(\eta, \sigma^2)$. 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 $X_1, \dots, X_n$ 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 $X_1, \dots, X_n$, 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 $X_1, X_2, \dots, X_n \overset{i.i.d.}{\sim} U(0, \theta_0)$, where $\theta_0 > 0$ is the true parameter. We compare two consistent estimators for θ_0:*

1. $\hat{\theta}_{n,1} = X_{(n)} = \max\{X_1, \dots, X_n\}$ *(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 $\boldsymbol{\theta_0} - z/n$.

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

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 $\lim_{n\to\infty} \text{Var}_{\theta_0}[\sqrt{n}(\hat{\theta}_n - \boldsymbol{\theta_0})]$.

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 $\Sigma(\theta_0) = \text{Var}(\mathbf{Y}_1)$ 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 parameter $\boldsymbol{\theta_0}$:

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

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 $I(\theta_0)$:

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

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 $X_1, \dots, X_n$ 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 $L(X_1, \dots, X_n)$ is a statistic (a function of the data) such that:*

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

then the random interval $[L(X_1, \dots, X_n), \infty)$ 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 $U(X_1, \dots, X_n)$ is a statistic such that:*

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

then the random interval $(-\infty, U(X_1, \dots, X_n)]$ 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 $[L(X_1, \dots, X_n), U(X_1, \dots, X_n)]$ 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 $\theta_1, \theta_2, \dots$ 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 $X_1, \dots, X_n$ 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:

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

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

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

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:

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

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 $X_1, \dots, X_n$ 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:*

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

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

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 $X_1, \dots, X_n$ 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 $\theta \mapsto g(x_1, \dots, x_n, \theta)$ remains continuous.*
2. *Asymptotic Distribution: The sequence of statistics converges in distribution to a random variable Y whose distribution is independent of θ:*

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

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 $P[q_1 \le Y \le q_2] = 1 - \alpha$. By the definition of convergence in distribution, this implies that:

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

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: $X_1, \dots, X_n$ 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 $\frac{\partial f}{\partial\theta}=\frac{\partial\log f}{\partial\theta}\cdot f$:

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

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:

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

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:*

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

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

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

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:*

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

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:

$$\begin{aligned} MSE_\theta(\delta) \quad &= E_\theta\left([(\delta^*(\mathbf{T}) - q(\theta)) + (\delta(\mathbf{X}) - \delta^*(\mathbf{T}))]^2\right) \\ &= E_\theta((\delta^*(\mathbf{T}) - q(\theta))^2) + E_\theta((\delta(\mathbf{X}) - \delta^*(\mathbf{T}))^2) \\ &\quad +2E_\theta((\delta^*(\mathbf{T}) - q(\theta))(\delta(\mathbf{X}) - \delta^*(\mathbf{T}))) \end{aligned} \tag{4.3}$$

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:

$$\delta^*(\mathbf{T}) = \frac{\mathbf{T}}{n} = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n.$$

$$Var_\theta(\delta^*(\mathbf{T})) = \frac{\theta(1-\theta)}{n} \quad \text{and} \quad Var_\theta(\delta(\mathbf{X})) = Var_\theta(X_1) = \theta(1-\theta).$$

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:*

$$E_\theta[g(\mathbf{T})] = 0 \text{ for all } \theta \in \Theta \quad \Longrightarrow \quad P_\theta(g(\mathbf{T}) = 0) = 1.$$

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:

$$f(\mathbf{x}, \boldsymbol{\theta}) = h(\mathbf{x}) c(\boldsymbol{\theta}) \exp\left\{ \sum_{j=1}^{k} \eta_j(\boldsymbol{\theta}) t_j(\mathbf{x}) \right\}$$

An exponential family is called **full-rank** (or canonical) if:

1. $t_1(\mathbf{x}), t_2(\mathbf{x}), \ldots, t_k(\mathbf{x})$ satisfy no linear restrictions (they are linearly independent).
2. $\eta_1(\boldsymbol{\theta}), \eta_2(\boldsymbol{\theta}), \ldots, \eta_k(\boldsymbol{\theta})$ satisfy no linear restrictions on the parameters.
3. The parameter space Θ 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 $T_j(\mathbf{X}) = \sum_{i=1}^{n} t_j(X_i)$, 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 ($k = 1$).

- $t_1(\mathbf{x}) = \sum_{i=1}^{n} x_i$
- $\eta_1(\theta) = \ln(\theta/(1-\theta))$

*Since this is a full-rank exponential family (as $\Theta = (0, 1)$ is an open interval), $T = \sum_{i=1}^{n} X_i$ 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 $\beta(\theta)$, we have:*

1. *There is **at most one** unbiased estimator of $\beta(\theta)$ based on $\mathbf{T}$.*
2. *If $\delta^*(\mathbf{T})$ is an unbiased estimator of $\beta(\theta)$, then $\delta^*(\mathbf{T})$ is the **UMVUE** (Uniformly Minimum Variance Unbiased Estimator).*
3. *If $\delta(\mathbf{X})$ is any unbiased estimator of $\beta(\theta)$, 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 $\beta(\theta)$. $\delta^*(\mathbf{T})$ 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.*

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:

$$r(x) = \mathbb{E}(Y|X = x) = \int yf(y|x)\,dy \tag{5.1}$$

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 $MSE(c) = \mathbb{E}[(Y - c)^2]$.

$$\frac{d}{dc}MSE(c) = \mathbb{E}\left[\frac{\partial}{\partial c}(Y - c)^2\right] = \mathbb{E}[-2(Y - c)] = 0$$

$$\mathbb{E}[Y] - c = 0 \quad \Rightarrow \quad \hat{c} = \mathbb{E}[Y]$$

$$\hat{Y}_{\text{BCP}} = \mathbb{E}[Y]$$

Calculation of Minimum MSE

Substitute $\hat{c}$ back into the MSE formula:

$$MSE(\hat{Y}_{\text{BCP}}) = \mathbb{E}[(Y - \mathbb{E}[Y])^2]$$

$$MSE(\hat{Y}_{\text{BCP}}) = \text{Var}(Y)$$

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 $\hat{a}$ and $\hat{b}$

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:

$$0 = \mathbb{E}[X(Y - (\mathbb{E}[Y] - b\mathbb{E}[X]) - bX)]$$

$$0 = \text{Cov}(X, Y) - b \cdot \text{Var}(X)$$

$$\hat{b} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

$$\hat{Y}_{\text{BLP}} = \mathbb{E}[Y] + \frac{\text{Cov}(X, Y)}{\text{Var}(X)}(X - \mathbb{E}[X])$$

Calculation of Minimum MSE

$$MSE(\hat{Y}_{\text{BLP}}) = \text{Var}(Y) - \frac{(\text{Cov}(X, Y))^2}{\text{Var}(X)}$$

Using the correlation coefficient ρ, this simplifies to:

$$MSE(\hat{Y}_{\text{BLP}}) = \text{Var}(Y)(1 - \rho^2)$$

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 $\hat{g}(X)$

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.

$$\hat{g}(x) = \mathbb{E}[Y|X = x] = r(x)$$

$$\hat{Y}_{\text{BP}} = \mathbb{E}[Y|X]$$

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 $\hat{Y}_{\text{BP}}$ back into the MSE formula:

$$MSE(\hat{Y}_{\text{BP}}) = \mathbb{E}[(Y - \mathbb{E}[Y|X])^2]$$

Using the variance decomposition formula $\text{Var}(Y) = \mathbb{E}[\text{Var}(Y|X)] + \text{Var}(\mathbb{E}[Y|X])$, the MSE simplifies to:

$$MSE(\hat{Y}_{\text{BP}}) = \mathbb{E}[\text{Var}(Y|X)]$$

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 ($\hat{Y}$)**	**Minimum MSE**
BCP	$g(X) = c$	$\mathbb{E}[Y]$	$\text{Var}(Y)$
BLP	$g(X) = a + bX$	$\mathbb{E}[Y] + \hat{b}(X - \mathbb{E}[X])$	$\text{Var}(Y) \cdot (1 - \rho^2)$
BP	$g(X)$ any function	$\mathbb{E}[Y\|X]$	$\mathbb{E}[\text{Var}(Y\|X)]$

Error Relationship (Hierarchy):

$$MSE(\hat{Y}_{\text{BP}}) \leq MSE(\hat{Y}_{\text{BLP}}) \leq MSE(\hat{Y}_{\text{BCP}})$$

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).*

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{argmin} \left[\sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2 \right]$$

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 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 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:

$$p \equiv p(\boldsymbol{\beta}) \equiv \mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x}) = \frac{e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}}} \quad (13.32)$$

The parameters are $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)^T$, 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:

$$\text{logit}(p_i) = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} \quad (13.33)$$

The quantity 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:

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i(\boldsymbol{\beta})^{Y_i}(1 - p_i(\boldsymbol{\beta}))^{1-Y_i} \quad (13.34)$$

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.

The algorithm requires starting values and iteratively refines the coefficients until they converge.

Initialization:

1. Choose initial parameter values $\hat{\boldsymbol{\beta}}^{(0)} = (\hat{\beta}_0^{(0)}, \dots, \hat{\beta}_k^{(0)})^T$. 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 $i = 1, \dots, n$, 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":

$$Z_i^{(s)} = \text{logit}(p_i^{(s)}) + \frac{Y_i - p_i^{(s)}}{p_i^{(s)}(1 - p_i^{(s)})}, \quad i = 1, \dots, n$$

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 $p_i^{(s)}$:

$$\mathbf{W}^{(s)} = \text{diag}(p_1^{(s)}(1 - p_1^{(s)}), \dots, p_n^{(s)}(1 - p_n^{(s)}))$$

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).