Mathematical Statistics

Daniela Rodriguez

Schedule and Grading Policy

Homework Schedule and Deadlines

Please note: **All dates listed are tentative and subject to change.** Any modifications will be announced promptly.

Homework Policy

The dates listed below indicate when each homework assignment will become **available** on the course platform. The deadline for each assignment is the day immediately preceding the start date of the subsequent assignment. Please ensure you submit your work according to the due dates provided below:

- **Homework 1: Models - Metrics**
  - Available: October 20
  - Due: October 29
- **Homework 2: Point Estimation**
  - Available: October 30
  - Due: November 10
- **Homework 3: Confidence Intervals**
  - Available: November 11
  - Due: December 3
- **Homework 4: Fisher Information + Rao-Cramér**
  - Available: December 4
  - Due: December 17
- **Homework 5: Regression + Bootstrap**
  - Available: December 18
  - Due: December 28
- **Homework 6: Nonparametric Methods**
  - Available: December 29
  - Due: January 13

Grading Policy

The final course grade (FC) is determined by the following policy, where F is the Final Exam grade, M is the Midterm grade, and HW is the Homework grade (calculated as the average of your best four HW scores). The passing grade is 55.

Let $P = \max(M, F) \times 0.2 + F \times 0.8$ be the weighted passing threshold.

$$\text{FC} = \begin{cases} \max(HW, F) \times 0.15 + \max(M, F) \times 0.15 + F \times 0.7 & \text{if } P \geq 55 \\ \max(M, F) \times 0.2 + F \times 0.8 & \text{if } P < 55 \end{cases}$$

Where:

- F: Final Exam
- M: Midterm
- HW: Homework

Exam Dates

- **Midterm Exam:** Date and time To Be Determined (TBD).
- **Exam A:**
  - Date: January 30, 2026
  - Time: 14:00 (2:00 PM)
- **Exam B:**
  - Date: February 10, 2026
  - Time: 14:00 (2:00 PM)

Formulation of The Problem of Statistical Estimation.

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.

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 $F_\theta(x)$.

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:*

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

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. Often, the standard parameters used for a distribution don't align with the natural parameters. However, there is typically a smooth, one-to-one transformation between them, so the density can also be written in this form:

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

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

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

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

Define:

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

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

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

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

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

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

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

Define:

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

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

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

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

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

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

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 $c'(\theta)$:

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

Sampling from a k-Parameter Exponential Family

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

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

where $\boldsymbol{\eta} = (\eta_1,\ldots,\eta_k)^\top$ is the vector of **natural parameters**, $T_j(x)$ are the component statistics, and $A(\boldsymbol{\eta})$ is the log-normalizer (or cumulant-generating function).

Let $\mathbf{X} = (X_1,\ldots,X_n)$ be a random sample of size n, where $X_1,\ldots,X_n$ are independent and identically distributed (i.i.d.) according to $f(x;\boldsymbol{\eta})$.

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

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

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

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

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

the vector of summed component statistics

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

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

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

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

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

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

Review of some useful probability tools

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

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

We have no control over the first term on the right-hand side of (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,

$$\min_b E[(X-b)^2] = E[(X-E[X])^2]. \tag{3}$$

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 $\varepsilon > 0$,*

$$P(|X_n - X| > \varepsilon) \to 0 \quad \text{as } n \to \infty. \tag{4}$$

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

$$\lim_{n\to\infty} F_n(t) = F(t) \tag{5}$$

   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*

$$E(X_n - X)^2 \to 0 \tag{6}$$

   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

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

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

Convergence in distribution does not imply convergence in probability.

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

Warning.

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

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

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

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

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

Example 1: Poisson Distribution

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

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

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

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

The CLT states:

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

Example 2: Normal (Gaussian) Distribution

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

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

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

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

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

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

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

Some convergence properties are preserved under transformations.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Example 4

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

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

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

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

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

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

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

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

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

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

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

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

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

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