Week 12

Lecture 1

Standard Normal Vector

Let

$$z_1, z_2, \ldots, z_d \sim \mathcal{N}(0,1)$$

be independent standard normal random variables.
Define the random vector

$$z = \begin{bmatrix} z_1 \\ \vdots \\ z_d \end{bmatrix}.$$

Since the components are independent, the joint density is

$$f_Z(z) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z_i^2\right).$$

Using

$$\sum_{i=1}^{d} z_i^2 = \|z\|^2,$$

we obtain the compact form

$$f_Z(z) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\|z\|^2\right).$$

Simple Linear Transform of a 2D Standard Normal

Let

$$z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, \quad z_1, z_2 \sim \mathcal{N}(0,1), \text{ independent.}$$

Define

$$x_1 = z_1,$$

$$x_2 = \rho z_1 + \sqrt{1-\rho^2}\, z_2,$$

where $-1 < \rho < 1$.
In matrix form,

$$x = Az,$$

with

$$A = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix}.$$

The inverse is

$$A^{-1} = \begin{bmatrix} 1 & 0 \\ -\frac{\rho}{\sqrt{1-\rho^2}} & \frac{1}{\sqrt{1-\rho^2}} \end{bmatrix}.$$

Determinants:

$$\det(A) = \sqrt{1-\rho^2},$$

$$\det(A^{-1}) = \frac{1}{\sqrt{1-\rho^2}}.$$

Mean and Covariance

Mean:

$$\mathbb{E}[x_1] = 0, \quad \mathbb{E}[x_2] = 0.$$

Variance:

$$\mathrm{Var}(x_1) = 1,$$

$$\mathrm{Var}(x_2) = \rho^2 + (1-\rho^2) = 1.$$

Covariance:

$$\mathrm{Cov}(x_1, x_2) = \mathbb{E}[x_1 x_2] = \rho.$$

Covariance matrix:

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Observe:

$$\Sigma = AA^\top.$$

Determinant:

$$\det(\Sigma) = 1 - \rho^2.$$

Inverse:

$$\Sigma^{-1} = \frac{1}{1-\rho^2} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}.$$

Density via Change of Variables

Using change of variables,

$$f_X(x) = f_Z(A^{-1}x) |\det(A^{-1})|.$$

Since

$$f_Z(z) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\|z\|^2\right),$$

we obtain

$$f_X(x) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2} x^\top \Sigma^{-1} x\right).$$

Explicitly,

$$f_X(x_1, x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left(x_1^2 - 2\rho x_1 x_2 + x_2^2\right)\right).$$

Factorization and Conditional Distributions

The joint density factorizes as

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1).$$

Marginal:

$$X_1 \sim \mathcal{N}(0, 1).$$

Conditional:

$$X_2 \mid X_1 = x_1 \sim \mathcal{N}(\rho x_1,\, 1 - \rho^2).$$

By symmetry,

$$X_2 \sim \mathcal{N}(0, 1),$$

$$X_1 \mid X_2 = x_2 \sim \mathcal{N}(\rho x_2,\, 1-\rho^2).$$

General Bivariate Normal

Let

$$x = Az + \mu,$$

where

$$z \sim \mathcal{N}(0, I),$$

and

$$\Sigma = AA^\top.$$

Then

$$x \sim \mathcal{N}(\mu, \Sigma),$$

with density

$$f_X(x) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right).$$

General covariance form:

$$\Sigma = \begin{bmatrix} a^2 & \rho ab \\ \rho ab & b^2 \end{bmatrix},$$

with $a > 0$, $b > 0$, $|\rho| < 1$.

Multivariate Normal in Dimension d

Let

$$x = Az + \mu,$$

where

$$z \sim \mathcal{N}(0, I_d),$$

and

$$\Sigma = AA^\top.$$

Then

$$x \sim \mathcal{N}(\mu, \Sigma),$$

with density

$$f_X(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right).$$

Mean:

$$\mathbb{E}[x] = \mu.$$

Covariance:

$$\mathrm{Cov}(x) = \Sigma.$$

Important Properties

Let

$$x \sim \mathcal{N}(\mu, \Sigma).$$

Linear Scalar Transform

If

$$y = a^\top x,$$

then

$$y \sim \mathcal{N}(a^\top \mu,\ a^\top \Sigma a).$$

Linear Vector Transform

If

$$y = Bx,$$

then

$$y \sim \mathcal{N}(B\mu,\ B\Sigma B^\top).$$

Independence Property

For components x_i and x_j of a multivariate normal:

$$x_i \text{ and } x_j \text{ independent} \iff \Sigma_{ij} = 0.$$

Uncorrelated components of a multivariate normal are independent.

```
*********************************************************************************
```

Lecture 2

Parameter Estimation

Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a family of probability distributions indexed by parameter θ.
Given data

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} P_{\theta_0}$$

for some unknown $\theta_0 \in \Theta$,
Goal: Estimate the true parameter θ_0 from data.

Maximum Likelihood Estimation

Likelihood Function

Given observations $x_1, \ldots, x_n$,

$$L(\theta) = P(X_1 = x_1, \ldots, X_n = x_n \mid \theta)$$

Under independence,

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i)$$

where f_θ denotes the pmf or pdf.

Log-Likelihood

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f_\theta(x_i)$$

Negative Log-Likelihood

Define the risk function

$$R(\theta) = -\ell(\theta) = -\sum_{i=1}^{n} \log f_\theta(x_i)$$

Maximum likelihood estimate

$$\hat{\theta}_{ML} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \ell(\theta) = \arg\min_{\theta} R(\theta)$$

Example 1: Bernoulli Bias

Model

$$\mathcal{P} = \{\mathrm{Bern}(\theta) : \theta \in 0, 1\}$$

Data:

$$X_i \in \{0, 1\}$$

pmf

$$P_\theta(X = x) = \begin{cases} \theta & x = 1 \\ 1 - \theta & x = 0 \end{cases}$$

Compact form:

$$P_\theta(x) = \theta^x (1 - \theta)^{1-x}$$

Negative Log-Likelihood

$$R(\theta) = -\sum_{i=1}^{n} \log\left(\theta^{x_i} (1 - \theta)^{1-x_i}\right)$$

$$= -\sum_{i=1}^{n} (x_i \log\theta + (1 - x_i)\log(1 - \theta))$$

Let

$$a = \sum_{i=1}^{n} x_i$$

Then

$$R(\theta) = a \log\frac{1}{\theta} + (n - a)\log\frac{1}{1 - \theta}$$

Minimization

Setting derivative to zero gives

$$\hat{\theta}_{ML} = \frac{a}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Important observation:
ML estimate equals sample mean.

Example 2: Uniform Distribution

Model

$$\mathcal{P} = \{\mathrm{Unif}(a,b) : a, b \in \mathbb{R}, a < b\}$$

Density:

$$f_\theta(x) = \begin{cases} \frac{1}{b-a} & x \in a, b \\ 0 & \text{otherwise} \end{cases}$$

Equivalent form:

$$f_\theta(x) = \frac{1}{b-a}\mathbf{1}(x \in a, b)$$

Negative Log-Likelihood

$$R(\theta) = -\sum_{i=1}^{n} \log\left(\frac{1}{b-a}\mathbf{1}(x_i \in a, b)\right)$$

If any $x_i \notin a, b$, then

$$R(\theta) = \infty$$

Thus require

$$a \le \min_i x_i, \quad b \ge \max_i x_i$$

Then

$$R(\theta) = n \log(b-a)$$

Minimization

To minimize $ \log(b-a) $, choose smallest interval containing data:

$$\hat{a}_{ML} = \min_i x_i$$

$$\hat{b}_{ML} = \max_i x_i$$

Example 3: Normal Mean, Variance Known

Model

$$\mathcal{P} = \{\mathcal{N}(\mu, 1) : \mu \in \mathbb{R}\}$$

Density:

$$f_\mu(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\mu)^2\right)$$

Negative Log-Likelihood

Ignoring constants independent of $ \mu $,

$$R(\mu) = \sum_{i=1}^{n} \frac{1}{2}(x_i - \mu)^2 + C$$

First-Order Condition

$$\frac{\partial R}{\partial \mu} = -\sum_{i=1}^{n}(x_i - \mu)$$

Setting to zero,

$$\sum_{i=1}^{n}(x_i - \mu) = 0$$

$$\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Important observation:
ML estimate equals sample mean.

Example 4: Normal Mean and Variance Unknown

Model

$$\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$$

Density:

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

Negative Log-Likelihood

Ignoring constants:

$$R(\mu, \sigma^2) = \frac{n}{2}\log(\sigma^2) + \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - \mu)^2}{\sigma^2}$$

Derivative with Respect to μ

$$\frac{\partial R}{\partial \mu} = \sum_{i=1}^{n}\frac{x_i - \mu}{\sigma^2}$$

Setting to zero:

$$\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Derivative with Respect to σ^2

Rewrite:

$$R = \frac{n}{2}\log(\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Derivative:

$$\frac{\partial R}{\partial \sigma^2} = \frac{n}{2}\frac{1}{\sigma^2} - \frac{1}{2}\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{(\sigma^2)^2}$$

Setting to zero:

$$\frac{n}{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{(\sigma^2)^2}$$

Solving:

$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_{ML})^2$$

Important observation:
ML variance uses denominator n, not $n - 1$.

Extension: Multivariate Normal

Model:

$$\mathcal{P} = \{\mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^d, \Sigma \succ 0\}$$

Data:

$$x_1, \ldots, x_N \in \mathbb{R}^d$$

ML estimates:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

Linear Regression with Gaussian Noise

Model

$$X \in \mathbb{R}^d, \quad Y \in \mathbb{R}$$

$$Y = w^T X + \varepsilon$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Thus,

$$Y \mid X \sim \mathcal{N}(w^T X, \sigma^2)$$

Data:

$$(x_1, y_1), \ldots, (x_n, y_n)$$

Likelihood

$$L(w) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y_i - w^T x_i)^2\right)$$

Negative Log-Likelihood

Ignoring constants:

$$R(w) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - w^T x_i)^2 + C$$

Important observation:
Maximum likelihood estimation reduces to minimizing

$$\sum_{i=1}^{n} (y_i - w^T x_i)^2$$

Thus ML for linear regression with Gaussian noise is equivalent to least squares.

****

Lecture 3

Gaussian Mixture Models and Expectation Maximization

Motivation: Multi Modal Data

Consider data generated from K different groups. Each group produces observations following a Gaussian distribution with different parameters.

Example in one dimension with $K = 3$:
Component 1:
Mean -4, variance 0.5, weight 0.4

Component 2:
Mean 0, variance 1, weight 0.3

Component 3:
Mean 5, variance 1, weight 0.3

Observed data exhibits multiple peaks. A single Gaussian cannot model such multi modal behavior.

Definition: Gaussian Mixture Model

Let $X \in \mathbb{R}^d$.
A Gaussian Mixture Model with K components has density

$$f_X(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)$$

where

$$\sum_{k=1}^{K} \pi_k = 1, \quad \pi_k \geq 0$$

and

$$\mathcal{N}(x \mid \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right)$$

Parameters

For each component k:
Mixing weight π_k
Mean $\mu_k \in \mathbb{R}^d$
Covariance $\Sigma_k \in \mathbb{R}^{d \times d}$
Total parameter set

$$\theta = \{\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K\}$$

Important Observation: Label Non Identifiability

Permutation of component indices does not change the density.
If parameters are swapped between components, the mixture density remains identical.

Latent Variable Formulation

Introduce hidden variable Z.
Definition:

$$P(Z = k) = \pi_k$$

Conditional distribution:

$$X \mid Z = k \sim \mathcal{N}(\mu_k, \Sigma_k)$$

Marginal Distribution

Using marginalization:

$$P(X) = \sum_{k=1}^{K} P(X, Z = k)$$

$$= \sum_{k=1}^{K} P(X \mid Z = k) P(Z = k)$$

$$= \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)$$

Thus latent formulation is equivalent to mixture density.

Maximum Likelihood Estimation

Given data

$$\{x_1, x_2, \ldots, x_N\}$$

Likelihood:

$$P(\text{Data} \mid \theta) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)$$

Negative log likelihood:

$$\mathcal{L}(\theta) = -\sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \right)$$

Direct maximization is difficult because of the summation inside the logarithm.

Chicken and Egg Problem

If component assignments were known:
Parameter estimation reduces to computing sample means and covariances per cluster.
If parameters were known:
Cluster assignments can be inferred probabilistically.
But neither is known.
This motivates the Expectation Maximization algorithm.

E Step: Cluster Responsibilities

Compute posterior probability that point x_n belongs to component k.
Definition:

$$\gamma_{nk} = P(Z = k \mid X = x_n)$$

Using Bayes rule:

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n \mid \mu_j, \Sigma_j)}$$

Properties:

$$\sum_{k=1}^{K} \gamma_{nk} = 1$$

γ_{nk} is called responsibility of component k for data point n.

Effective Cluster Size

Define

$$N_k = \sum_{n=1}^{N} \gamma_{nk}$$

Interpreted as soft number of points assigned to component k.

M Step: Parameter Updates

Update means:

$$\mu_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} x_n$$

Update covariances:

$$\Sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (x_n - \mu_k^{new})(x_n - \mu_k^{new})^T$$

Update mixing weights:

$$\pi_k^{new} = \frac{N_k}{N}$$

EM Algorithm Summary

1. Initialize parameters π_k, μ_k, Σ_k randomly.
2. E step:
   Compute γ_{nk} for all n, k.
3. M step:
   Update μ_k, Σ_k, π_k using responsibilities.
4. Repeat E and M steps until convergence.

Conceptual Flow

Latent variable model introduces hidden component index.
Likelihood maximization is difficult due to log of sum structure.
EM alternates between:
Estimating hidden variables using current parameters.
Estimating parameters using expected hidden variables.
This procedure iteratively improves the likelihood.

Application: Clustering

Gaussian mixture models provide probabilistic clustering.
In high dimensional settings, visualization is impossible.
EM based GMM remains a fundamental tool for unsupervised learning.
Gaussian mixture models form a foundational building block in machine learning.

```
********************************************************************************
```

Lecture 4

Tail Bounds and Concentration

Two types of results:

1. Bounds on deviation of a random variable from its mean.
2. Behavior of averages of many random variables.

Markov Inequality

Setup

Let X be a positive random variable such that

$$X \geq 0$$

Let

$$\mathbb{E}[X] = \mu$$

Statement

For any $t > 0$,

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$$

Interpretation

If $t < \mu$, then

$$\frac{\mu}{t} > 1$$

and the bound is vacuous since probability is at most 1.
The bound is meaningful for $t \geq \mu$.

Proof

Since $X \geq 0$,

$$\mathbb{E}[X] = \int_0^\infty x f_X(x) dx$$

Split the integral at t:

$$mathbbE[X] = \int_0^t x f_X(x) dx + \int_t^\infty x f_X(x) dx$$

Since $x \geq t$ on $[t, \infty)$,

$$\int_t^\infty x f_X(x) dx \geq \int_t^\infty t f_X(x) dx = t \int_t^\infty f_X(x) dx$$

Thus,

$$\mathbb{E}[X] \geq t\mathbb{P}(X \geq t)$$

Rearranging,

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$$

Tightness Example

Let X be discrete:

$$\mathbb{P}(X = 0) = \frac{4}{5}$$

$$\mathbb{P}(X = 50) = \frac{1}{5}$$

Then,

$$\mathbb{E}[X] = \frac{1}{5} \cdot 50 = 10$$

Compute:

$$\mathbb{P}(X \geq 50) = \frac{1}{5}$$

Markov bound:

$$\frac{\mathbb{E}[X]}{50} = \frac{10}{50} = \frac{1}{5}$$

Equality holds.

Chebyshev Inequality

Setup

Let

$$\mathbb{E}[X] = \mu$$

$$\mathrm{Var}(X) = \sigma^2$$

Statement

For any $t > 0$,

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

Proof via Markov

Observe:

$$|X - \mu| \geq t \iff (X - \mu)^2 \geq t^2$$

Thus,

$$\mathbb{P}(|X - \mu| \geq t) = \mathbb{P}((X - \mu)^2 \geq t^2)$$

Apply Markov to $(X - \mu)^2$:

$$\leq \frac{\mathbb{E}[(X - \mu)^2]}{t^2} = \frac{\mathrm{Var}(X)}{t^2}$$

Sample Mean and Concentration

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed.
Assume

$$\mathbb{E}[X_i] = \mu$$

Define the sample mean:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Then,

$$\mathbb{E}[\overline{X}_n] = \mu$$

Variance of Sample Mean

Since variables are independent,

$$\mathrm{Var}(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

Chebyshev Bound for Sample Mean

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

Decay rate:

$$O\left(\frac{1}{n}\right)$$

Hoeffding Inequality

Additional Assumption

Assume

$$a \leq X_i \leq b$$

Then,

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

Decay rate:

$$O(e^{-n})$$

Exponential decay is much faster than $1/n$.

Convergence Concepts

Convergence in Probability

A sequence X_n converges to X in probability if

$$\lim_{n \to \infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0 \quad \forall \epsilon > 0$$

Convergence in Distribution

X_n converges to X in distribution if

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x) \quad \forall x$$

Convergence in probability implies convergence in distribution.

Law of Large Numbers

Let $X_1, \ldots, X_n$ be iid with

$$\mathbb{E}[X_i] = \mu$$

Define

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Weak Law of Large Numbers

$$\overline{X}_n \to \mu \quad \text{in probability}$$

Proof Using Chebyshev

If variance is finite,

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

As $n \to \infty$,

$$\frac{\sigma^2}{n\epsilon^2} \to 0$$

Thus convergence in probability holds.

Central Limit Theorem

Let $X_1, X_2, \ldots$ be iid with

$$\mathbb{E}[X_i] = \mu$$

$$\mathrm{Var}(X_i) = \sigma^2$$

Define

$$Y_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)$$

Statement

$$Y_n \to \mathcal{N}(0, \sigma^2) \quad \text{in distribution}$$

Interpretation

Scaled sums of independent random variables converge to a normal distribution.
This explains ubiquity of the normal distribution in additive phenomena.
Even for moderate n, the approximation is accurate.

Summary of Results

Markov:

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$$

Chebyshev:

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

Hoeffding:

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

Weak Law:

$$\overline{X}_n \to \mu \quad \text{in probability}$$

Central Limit Theorem:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \to \mathcal{N}(0, \sigma^2) \quad \text{in distribution}$$

```
********************************************************************************
```
