# Selected Topics in Mathematics of Learning

## High-Dimensional Statistics

Lecturer: Marius Yamakou

Winter Semester 2024/25
Department of Data Science, FAU

October 22, 2024

# 3: Continuous distributions

**Continuous random variable:** A continuous random variable takes values within an interval of real numbers.

# 3: Continuous distributions

**Continuous random variable:** A continuous random variable takes values within an interval of real numbers.

**Probability density function (PDF):** Let $X$ be a continuous random variable with cumulative distribution function (CDF) $F_X(x)$. The PDF denoted $f_X(x)$ is the derivative of the CDF:

$$f_X(x) = \frac{d}{dx} F_X(x),$$

with $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$.

# 3: Continuous distributions

**Continuous random variable:** A continuous random variable takes values within an interval of real numbers.

**Probability density function (PDF):** Let $X$ be a continuous random variable with cumulative distribution function (CDF) $F_X(x)$. The PDF denoted $f_X(x)$ is the derivative of the CDF:

$$f_X(x) = \frac{d}{dx} F_X(x),$$

with $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$.

**Expected value:** The expected value of $X$, $\mathbb{E}(X)$ or $\mu_X$, is defined by: $\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx$.

# 3: Continuous distributions

**Continuous random variable:** A continuous random variable takes values within an interval of real numbers.

**Probability density function (PDF):** Let $X$ be a continuous random variable with cumulative distribution function (CDF) $F_X(x)$. The PDF denoted $f_X(x)$ is the derivative of the CDF:

$$f_X(x) = \frac{d}{dx} F_X(x),$$

with $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$.

**Expected value:** The expected value of $X$, $\mathbb{E}(X)$ or $\mu_X$, is defined by: $\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx$. For any function $g(X)$, the expected value of $g(X)$ is given by: $\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$.

# 3: Continuous distributions

**Continuous random variable:** A continuous random variable takes values within an interval of real numbers.

**Probability density function (PDF):** Let $X$ be a continuous random variable with cumulative distribution function (CDF) $F_X(x)$. The PDF denoted $f_X(x)$ is the derivative of the CDF:

$$f_X(x) = \frac{d}{dx} F_X(x),$$

with $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$.

**Expected value:** The expected value of $X$, $\mathbb{E}(X)$ or $\mu_X$, is defined by: $\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx$. For any function $g(X)$, the expected value of $g(X)$ is given by: $\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$.

**Variance:** $Var(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right]$;

**Standard deviation:** $\sigma_X = \sqrt{Var(X)}$.

# Useful properties of expectation and variance (Proofs: Left as Exercise)

**For any constants** $a, b \in \mathbb{R}$:

- **Linearity of Expectation:** $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$
  Scaling and shifting a random variable affects its expected value linearly.

# Useful properties of expectation and variance (Proofs: Left as Exercise)

**For any constants** $a, b \in \mathbb{R}$:

- **Linearity of Expectation:** $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$
  Scaling and shifting a random variable affects its expected value linearly.
- **Additivity of Expectation:** $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$
  The expectation of a sum of random variables is the sum of their individual expectations.

# Useful properties of expectation and variance (Proofs: Left as Exercise)

**For any constants** $a, b \in \mathbb{R}$:

- **Linearity of Expectation:** $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$
  Scaling and shifting a random variable affects its expected value linearly.
- **Additivity of Expectation:** $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$
  The expectation of a sum of random variables is the sum of their individual expectations.
- **Variance Formula:** $Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$
  The variance measures the spread by comparing the expectation of the square to the square of the expectation.

# Useful properties of expectation and variance (Proofs: Left as Exercise)

**For any constants** $a, b \in \mathbb{R}$:

- **Linearity of Expectation:** $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$
  Scaling and shifting a random variable affects its expected value linearly.
- **Additivity of Expectation:** $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$
  The expectation of a sum of random variables is the sum of their individual expectations.
- **Variance Formula:** $Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$
  The variance measures the spread by comparing the expectation of the square to the square of the expectation.
- **Variance of Independent Variables:** If $X$ and $Y$ are independent, then:
  $Var(X + Y) = Var(X) + Var(Y)$
  Independence allows variances to add up when summing random variables.

# Useful properties of expectation and variance (Proofs: Left as Exercise)

**For any constants** $a, b \in \mathbb{R}$:

- **Linearity of Expectation:** $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$
  Scaling and shifting a random variable affects its expected value linearly.

- **Additivity of Expectation:** $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$
  The expectation of a sum of random variables is the sum of their individual expectations.

- **Variance Formula:** $Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$
  The variance measures the spread by comparing the expectation of the square to the square of the expectation.

- **Variance of Independent Variables:** If $X$ and $Y$ are independent, then:
  $Var(X + Y) = Var(X) + Var(Y)$
  Independence allows variances to add up when summing random variables.

- **Scaling Variance:** $Var(aX + b) = a^2 Var(X)$
  Scaling a random variable by a constant $a$ scales its variance by $a^2$.

# 3. Continuous distributions: Exponential $(\lambda)$

**Exponential Distribution:** A random variable $X$ has an exponential distribution with parameter $\lambda > 0$ (rate parameter), denoted $X \sim \text{Exponential}(\lambda)$, if its PDF is given by:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

# 3. Continuous distributions: Exponential $(\lambda)$

**Exponential Distribution:** A random variable $X$ has an exponential distribution with parameter $\lambda > 0$ (rate parameter), denoted $X \sim \text{Exponential}(\lambda)$, if its PDF is given by:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

This describes the time between events in a Poisson process, with the support $\mathcal{F} = [0, \infty)$ and parameter space $\Omega = \{\lambda \mid \lambda > 0\}$.

# 3. Continuous distributions: Exponential $(\lambda)$

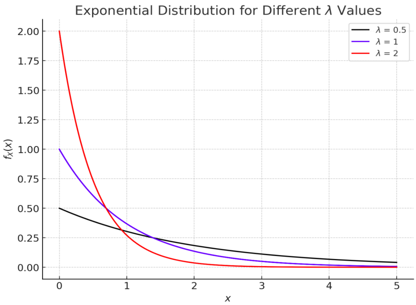
**Exponential Distribution:** A random variable $X$ has an exponential distribution with parameter $\lambda > 0$ (rate parameter), denoted $X \sim \text{Exponential}(\lambda)$, if its PDF is given by:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

This describes the time between events in a Poisson process, with the support $\mathcal{F} = [0, \infty)$ and parameter space $\Omega = \{\lambda \mid \lambda > 0\}$.

**Properties:**

- $\mathbb{E}(X) = \int_0^\infty x\lambda e^{-\lambda x}\, dx \stackrel{?}{=} \frac{1}{\lambda}$
- $\mathbb{E}(X^2) \stackrel{?}{=} \frac{2}{\lambda^2}$
- $Var(X) \stackrel{?}{=} \frac{1}{\lambda^2}$



# 3. Continuous distributions: Gaussian (Normal) $(\mu, \sigma^2)$

**Gaussian Distribution:** A random variable $X$ has a Gaussian (normal) distribution with mean $\mu$ and variance $\sigma^2 > 0$, denoted $X \sim \mathcal{N}(\mu, \sigma^2)$, if its PDF is given by:

$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$

# 3. Continuous distributions: Gaussian (Normal) $(\mu, \sigma^2)$

**Gaussian Distribution:** A random variable $X$ has a Gaussian (normal) distribution with mean $\mu$ and variance $\sigma^2 > 0$, denoted $X \sim \mathcal{N}(\mu, \sigma^2)$, if its PDF is given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

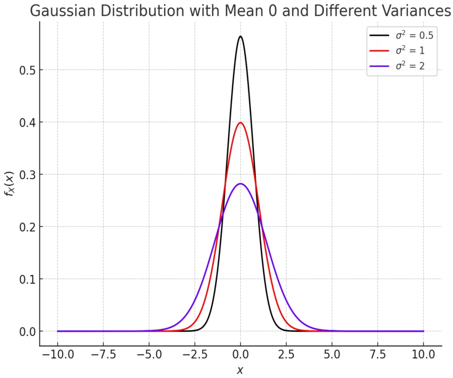
This describes a symmetric distribution around the mean $\mu$, with spread determined by $\sigma^2$. $X$ has support $\mathcal{F} = (-\infty, \infty)$ and parameter space $\Omega = \{(\mu, \sigma^2) \mid \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_{>0}\}$.

# 3. Continuous distributions: Gaussian (Normal) $(\mu, \sigma^2)$

**Gaussian Distribution:** A random variable $X$ has a Gaussian (normal) distribution with mean $\mu$ and variance $\sigma^2 > 0$, denoted $X \sim \mathcal{N}(\mu, \sigma^2)$, if its PDF is given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

This describes a symmetric distribution around the mean $\mu$, with spread determined by $\sigma^2$. $X$ has support $\mathcal{F} = (-\infty, \infty)$ and parameter space $\Omega = \{(\mu, \sigma^2) \mid \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_{>0}\}$.



# 3. Continuous distributions: Gaussian (Normal) $(\mu, \sigma^2)$

**Example:** Consider a normal distribution with $\mu = 0$ and $\sigma = 1$ (standard normal distribution). The PDF simplifies to:
$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

# 3. Continuous distributions: Gaussian (Normal) $(\mu, \sigma^2)$

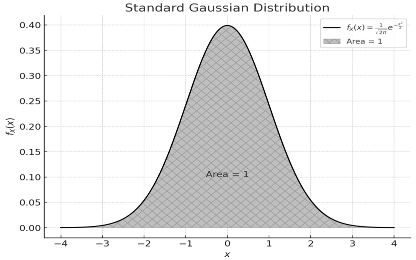
**Example:** Consider a normal distribution with $\mu = 0$ and $\sigma = 1$ (standard normal distribution). The PDF simplifies to:
$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

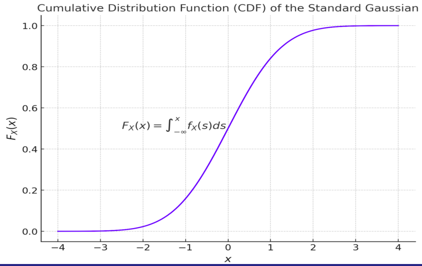
This distribution has zero mean and unit variance and is commonly used in many applications, including hypothesis testing and confidence intervals.

# 3. Continuous distributions: Gaussian (Normal) $(\mu, \sigma^2)$

**Example:** Consider a normal distribution with $\mu = 0$ and $\sigma = 1$ (standard normal distribution). The PDF simplifies to:
$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

This distribution has zero mean and unit variance and is commonly used in many applications, including hypothesis testing and confidence intervals.



# 3. Continuous distributions: Gaussian (Normal) $(\mu, \sigma^2)$

**Gaussian Distribution Example:** Consider the diameters of apples in a large orchard. The diameters are normally distributed. Let the random variable $X_a \sim \mathcal{N}(\mu_a = 7, \sigma_a^2 = 0.25)$ represent the diameter (in cm) of apples. Assume that $X_a$ is independent of any other factors in the orchard.

# 3. Continuous distributions: Gaussian (Normal) $(\mu, \sigma^2)$

**Gaussian Distribution Example:** Consider the diameters of apples in a large orchard. The diameters are normally distributed. Let the random variable $X_a \sim \mathcal{N}(\mu_a = 7, \sigma_a^2 = 0.25)$ represent the diameter (in cm) of apples. Assume that $X_a$ is independent of any other factors in the orchard.

To find the probability that a randomly picked apple has a diameter less than or equal to 7.5 cm, we compute:

$P(X_a \leq 7.5) = \int_{-\infty}^{7.5} \frac{1}{\sqrt{2\pi\sigma_a^2}} \exp\left(-\frac{(x-\mu_a)^2}{2\sigma_a^2}\right) dx.$

# 3. Continuous distributions: Gaussian (Normal) $(\mu, \sigma^2)$

**Gaussian Distribution Example:** Consider the diameters of apples in a large orchard. The diameters are normally distributed. Let the random variable $X_a \sim \mathcal{N}(\mu_a = 7, \sigma_a^2 = 0.25)$ represent the diameter (in cm) of apples. Assume that $X_a$ is independent of any other factors in the orchard.
To find the probability that a randomly picked apple has a diameter less than or equal to 7.5 cm, we compute:
$P(X_a \leq 7.5) = \int_{-\infty}^{7.5} \frac{1}{\sqrt{2\pi\sigma_a^2}} \exp\left(-\frac{(x-\mu_a)^2}{2\sigma_a^2}\right) dx.$

Alternatively, using the cumulative distribution function (CDF) of the normal distribution, we can find: $P(X_a \leq 7.5) = \Phi\left(\frac{7.5-\mu_a}{\sigma_a}\right)$, where $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{t^2}{2}\right) dt$ is the standard normal CDF. Notice that, since this integral does not have a closed-form solution, it is typically computed numerically.

# 3. Continuous distributions: Gaussian (Normal) $(\mu, \sigma^2)$

**Some properties of Gaussian distribution**

- Linear transformation preserves normality:
  If $X \sim \mathcal{N}(\mu, \sigma^2)$, then for any constants $a$ and $b$:
  $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

# 3. Continuous distributions: Gaussian (Normal) $(\mu, \sigma^2)$

**Some properties of Gaussian distribution**

- Linear transformation preserves normality:
  If $X \sim \mathcal{N}(\mu, \sigma^2)$, then for any constants $a$ and $b$:
  $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.
- Sum of independent normal random variables:
  If $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent, then their sum is normally distributed: $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

# 4. Independent random variables

Two random variables $X$ and $Y$ are **independent** if the occurrence of one does not affect the probability distribution of the other.

# 4. Independent random variables

Two random variables $X$ and $Y$ are **independent** if the occurrence of one does not affect the probability distribution of the other.
Formally, $X$ and $Y$ are independent if, for all events $A$ and $B$:
$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$

# 4. Independent random variables

Two random variables $X$ and $Y$ are **independent** if the occurrence of one does not affect the probability distribution of the other.
Formally, $X$ and $Y$ are independent if, for all events $A$ and $B$:
$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$

**Key Properties:**

- The joint probability distribution factors as the product of the individual distributions: $f_{X,Y}(x, y) = f_X(x)f_Y(y)$

# 4. Independent random variables

Two random variables $X$ and $Y$ are **independent** if the occurrence of one does not affect the probability distribution of the other.
Formally, $X$ and $Y$ are independent if, for all events $A$ and $B$:
$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$

**Key Properties:**

- The joint probability distribution factors as the product of the individual distributions: $f_{X,Y}(x, y) = f_X(x)f_Y(y)$
- For independent random variables, the expectation of their product is the product of their expectations: $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$

# 4. Independent random variables

Two random variables $X$ and $Y$ are **independent** if the occurrence of one does not affect the probability distribution of the other.
Formally, $X$ and $Y$ are independent if, for all events $A$ and $B$:
$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$

**Key Properties:**

- The joint probability distribution factors as the product of the individual distributions: $f_{X,Y}(x, y) = f_X(x)f_Y(y)$
- For independent random variables, the expectation of their product is the product of their expectations: $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
- Variance: $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \quad$ (if independent)

# 5. Estimators

**Estimator:** An estimator is a rule or function that provides an estimate of an unknown parameter based on observed data. It is typically denoted as $\widehat{\theta}$ for estimating a parameter $\theta$.

# 5. Estimators

**Estimator:** An estimator is a rule or function that provides an estimate of an unknown parameter based on observed data. It is typically denoted as $\widehat{\theta}$ for estimating a parameter $\theta$.

**Bias:** The bias of an estimator is the difference between the expected value of the estimator and the true value of the parameter being estimated. It's given as:

$$\text{Bias}(\widehat{\theta}) = E(\widehat{\theta}) - \theta$$

- If $\text{Bias}(\widehat{\theta}) < 0$, the estimator $\widehat{\theta}$ underestimates $\theta$.
- If $\text{Bias}(\widehat{\theta}) = 0$, the estimator $\widehat{\theta}$ is unbiased.
- If $\text{Bias}(\widehat{\theta}) > 0$, the estimator $\widehat{\theta}$ overestimates $\theta$.

# 5. Estimators

**Estimator:** An estimator is a rule or function that provides an estimate of an unknown parameter based on observed data. It is typically denoted as $\widehat{\theta}$ for estimating a parameter $\theta$.

**Bias:** The bias of an estimator is the difference between the expected value of the estimator and the true value of the parameter being estimated. It's given as:

$$\text{Bias}(\widehat{\theta}) = E(\widehat{\theta}) - \theta$$

- If $\text{Bias}(\widehat{\theta}) < 0$, the estimator $\widehat{\theta}$ underestimates $\theta$.
- If $\text{Bias}(\widehat{\theta}) = 0$, the estimator $\widehat{\theta}$ is unbiased.
- If $\text{Bias}(\widehat{\theta}) > 0$, the estimator $\widehat{\theta}$ overestimates $\theta$.

**Example:** Consider the sample size $n$. The estimator for the mean is $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$, where $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

# 5. Estimators

**Estimator:** An estimator is a rule or function that provides an estimate of an unknown parameter based on observed data. It is typically denoted as $\widehat{\theta}$ for estimating a parameter $\theta$.

**Bias:** The bias of an estimator is the difference between the expected value of the estimator and the true value of the parameter being estimated. It's given as:

$$\text{Bias}(\widehat{\theta}) = E(\widehat{\theta}) - \theta$$

- If $\text{Bias}(\widehat{\theta}) < 0$, the estimator $\widehat{\theta}$ underestimates $\theta$.
- If $\text{Bias}(\widehat{\theta}) = 0$, the estimator $\widehat{\theta}$ is unbiased.
- If $\text{Bias}(\widehat{\theta}) > 0$, the estimator $\widehat{\theta}$ overestimates $\theta$.

**Example:** Consider the sample size $n$. The estimator for the mean is $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$, where $X_i \sim \mathcal{N}(\mu, \sigma^2)$. $\text{Bias}(\hat{\mu}) = \mathbb{E}(\hat{\mu}) - \mu$.

# 5. Estimators

**Estimator:** An estimator is a rule or function that provides an estimate of an unknown parameter based on observed data. It is typically denoted as $\widehat{\theta}$ for estimating a parameter $\theta$.

**Bias:** The bias of an estimator is the difference between the expected value of the estimator and the true value of the parameter being estimated. It's given as:

$$\text{Bias}(\widehat{\theta}) = E(\widehat{\theta}) - \theta$$

- If $\text{Bias}(\widehat{\theta}) < 0$, the estimator $\widehat{\theta}$ underestimates $\theta$.
- If $\text{Bias}(\widehat{\theta}) = 0$, the estimator $\widehat{\theta}$ is unbiased.
- If $\text{Bias}(\widehat{\theta}) > 0$, the estimator $\widehat{\theta}$ overestimates $\theta$.

**Example:** Consider the sample size $n$. The estimator for the mean is $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$, where $X_i \sim \mathcal{N}(\mu, \sigma^2)$. $\text{Bias}(\hat{\mu}) = \mathbb{E}(\hat{\mu}) - \mu$.
We have $\mathbb{E}(\hat{\mu}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$

# 5. Estimators

**Estimator:** An estimator is a rule or function that provides an estimate of an unknown parameter based on observed data. It is typically denoted as $\widehat{\theta}$ for estimating a parameter $\theta$.

**Bias:** The bias of an estimator is the difference between the expected value of the estimator and the true value of the parameter being estimated. It's given as:

$$\text{Bias}(\widehat{\theta}) = E(\widehat{\theta}) - \theta$$

- If $\text{Bias}(\widehat{\theta}) < 0$, the estimator $\widehat{\theta}$ underestimates $\theta$.
- If $\text{Bias}(\widehat{\theta}) = 0$, the estimator $\widehat{\theta}$ is unbiased.
- If $\text{Bias}(\widehat{\theta}) > 0$, the estimator $\widehat{\theta}$ overestimates $\theta$.

**Example:** Consider the sample size $n$. The estimator for the mean is $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$, where $X_i \sim \mathcal{N}(\mu, \sigma^2)$. $\text{Bias}(\hat{\mu}) = \mathbb{E}(\hat{\mu}) - \mu$.
We have $\mathbb{E}(\hat{\mu}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i)$

# 5. Estimators

**Estimator:** An estimator is a rule or function that provides an estimate of an unknown parameter based on observed data. It is typically denoted as $\widehat{\theta}$ for estimating a parameter $\theta$.

**Bias:** The bias of an estimator is the difference between the expected value of the estimator and the true value of the parameter being estimated. It's given as:

$$\text{Bias}(\widehat{\theta}) = E(\widehat{\theta}) - \theta$$

- If $\text{Bias}(\widehat{\theta}) < 0$, the estimator $\widehat{\theta}$ underestimates $\theta$.
- If $\text{Bias}(\widehat{\theta}) = 0$, the estimator $\widehat{\theta}$ is unbiased.
- If $\text{Bias}(\widehat{\theta}) > 0$, the estimator $\widehat{\theta}$ overestimates $\theta$.

**Example:** Consider the sample size $n$. The estimator for the mean is $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$, where $X_i \sim \mathcal{N}(\mu, \sigma^2)$. $\text{Bias}(\hat{\mu}) = \mathbb{E}(\hat{\mu}) - \mu$.
We have $\mathbb{E}(\hat{\mu}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\cdot n \cdot \mu = \mu$.

# 5. Estimators

**Estimator:** An estimator is a rule or function that provides an estimate of an unknown parameter based on observed data. It is typically denoted as $\widehat{\theta}$ for estimating a parameter $\theta$.

**Bias:** The bias of an estimator is the difference between the expected value of the estimator and the true value of the parameter being estimated. It's given as:

$$\text{Bias}(\widehat{\theta}) = E(\widehat{\theta}) - \theta$$

- If $\text{Bias}(\widehat{\theta}) < 0$, the estimator $\widehat{\theta}$ underestimates $\theta$.
- If $\text{Bias}(\widehat{\theta}) = 0$, the estimator $\widehat{\theta}$ is unbiased.
- If $\text{Bias}(\widehat{\theta}) > 0$, the estimator $\widehat{\theta}$ overestimates $\theta$.

**Example:** Consider the sample size $n$. The estimator for the mean is $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$, where $X_i \sim \mathcal{N}(\mu, \sigma^2)$. $\text{Bias}(\hat{\mu}) = \mathbb{E}(\hat{\mu}) - \mu$.
We have $\mathbb{E}(\hat{\mu}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) = \frac{1}{n} \cdot n \cdot \mu = \mu$.
Thus, $\text{Bias}(\hat{\mu}) = \mu - \mu = 0$, which implies that $\hat{\mu}$ is an unbiased estimator of $\mu$.

# 5. Estimators

**Mean-Squared Error (MSE):** The MSE of an estimator is the expected value of the squared difference between the estimator and the true parameter value. It measures the accuracy of the estimator by combining both variance and bias. $\mathsf{MSE}(\widehat{\theta}) = E[(\widehat{\theta} - \theta)^2]$.

# 5. Estimators

**Mean-Squared Error (MSE):** The MSE of an estimator is the expected value of the squared difference between the estimator and the true parameter value. It measures the accuracy of the estimator by combining both variance and bias. $\text{MSE}(\widehat{\theta}) = E[(\widehat{\theta} - \theta)^2]$.

**Lemma**

$MSE(\widehat{\theta}) = Var(\widehat{\theta}) + Bias(\widehat{\theta})^2$, *where* $Var(\widehat{\theta}) = E[(\widehat{\theta} - E(\widehat{\theta}))^2]$.

# 5. Estimators

**Mean-Squared Error (MSE):** The MSE of an estimator is the expected value of the squared difference between the estimator and the true parameter value. It measures the accuracy of the estimator by combining both variance and bias. MSE$(\widehat{\theta}) = E[(\widehat{\theta} - \theta)^2]$.

**Lemma**

*MSE*$(\widehat{\theta})$ = *Var*$(\widehat{\theta})$ + *Bias*$(\widehat{\theta})^2$, *where* $Var(\widehat{\theta}) = E[(\widehat{\theta} - E(\widehat{\theta}))^2]$.

**Proof:** MSE$(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}(\hat{\theta}) + \mathbb{E}(\hat{\theta}) - \theta)^2\right]$

# 5. Estimators

**Mean-Squared Error (MSE):** The MSE of an estimator is the expected value of the squared difference between the estimator and the true parameter value. It measures the accuracy of the estimator by combining both variance and bias. $\text{MSE}(\widehat{\theta}) = E[(\widehat{\theta} - \theta)^2]$.

**Lemma**

$MSE(\widehat{\theta}) = Var(\widehat{\theta}) + Bias(\widehat{\theta})^2$, *where* $Var(\widehat{\theta}) = E[(\widehat{\theta} - E(\widehat{\theta}))^2]$.

**Proof:**
$$\begin{aligned}\text{MSE}(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}(\hat{\theta}) + \mathbb{E}(\hat{\theta}) - \theta)^2\right] \\ &= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2 + 2(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta) + (\mathbb{E}(\hat{\theta}) - \theta)^2\right] \\ &= \text{Var}(\hat{\theta}) + 2\mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta)] + \mathbb{E}[(\mathbb{E}(\hat{\theta}) - \theta)^2].\end{aligned}$$

# 5. Estimators

**Mean-Squared Error (MSE):** The MSE of an estimator is the expected value of the squared difference between the estimator and the true parameter value. It measures the accuracy of the estimator by combining both variance and bias. MSE$(\widehat{\theta}) = E[(\widehat{\theta} - \theta)^2]$.

**Lemma**

*MSE*$(\widehat{\theta})$ = *Var*$(\widehat{\theta})$ + *Bias*$(\widehat{\theta})^2$, *where* $Var(\widehat{\theta}) = E[(\widehat{\theta} - E(\widehat{\theta}))^2]$.

**Proof:**

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}(\hat{\theta}) + \mathbb{E}(\hat{\theta}) - \theta)^2\right] \\
&= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2 + 2(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta) + (\mathbb{E}(\hat{\theta}) - \theta)^2\right] \\
&= \text{Var}(\hat{\theta}) + 2\mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta)] + \mathbb{E}[(\mathbb{E}(\hat{\theta}) - \theta)^2].
\end{aligned}$$

By the linearity of expectation and the fact that $\mathbb{E}(\hat{\theta})$ and $\mathbb{E}(\hat{\theta}) - \theta$ are deterministic, we find that: $\mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta)] = 0$.

# 5. Estimators

**Mean-Squared Error (MSE):** The MSE of an estimator is the expected value of the squared difference between the estimator and the true parameter value. It measures the accuracy of the estimator by combining both variance and bias. $\text{MSE}(\widehat{\theta}) = E[(\widehat{\theta} - \theta)^2]$.

**Lemma**

*MSE*$(\widehat{\theta}) =$ *Var*$(\widehat{\theta}) +$ *Bias*$(\widehat{\theta})^2$, *where* $Var(\widehat{\theta}) = E[(\widehat{\theta} - E(\widehat{\theta}))^2]$.

**Proof:**
$$\begin{aligned}\text{MSE}(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}(\hat{\theta}) + \mathbb{E}(\hat{\theta}) - \theta)^2\right] \\ &= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2 + 2(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta) + (\mathbb{E}(\hat{\theta}) - \theta)^2\right] \\ &= \text{Var}(\hat{\theta}) + 2\mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta)] + \mathbb{E}[(\mathbb{E}(\hat{\theta}) - \theta)^2].\end{aligned}$$

By the linearity of expectation and the fact that $\mathbb{E}(\hat{\theta})$ and $\mathbb{E}(\hat{\theta}) - \theta$ are deterministic, we find that: $\mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta)] = 0$.

Thus, the MSE simplifies to: $\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2$. □

# 6. Convergence

**1. Convergence in Distribution:** A sequence of random variables $\{X_n\}$ converges in distribution to a random variable $X$ if:

$$\lim_{n\to\infty} F_n(x) = F(x), \quad \text{for all continuity points of } F(x),$$

where $F_n(x)$ and $F(x)$ are the CDFs of $X_n$ and $X$, respectively, i.e., $F_n(x) = P(X_n \leq x)$ and $F(x) = P(X \leq x)$.

# 6. Convergence

**1. Convergence in Distribution:** A sequence of random variables $\{X_n\}$ converges in distribution to a random variable $X$ if:

$$\lim_{n\to\infty} F_n(x) = F(x), \quad \text{for all continuity points of } F(x),$$

where $F_n(x)$ and $F(x)$ are the CDFs of $X_n$ and $X$, respectively, i.e., $F_n(x) = P(X_n \leq x)$ and $F(x) = P(X \leq x)$.

**2. Convergence in Probability:** A sequence of random variables $\{X_n\}$ converges in probability to a random variable $X$ if:

$$\lim_{n\to\infty} P\left(|X_n - X| > \varepsilon\right) = 0, \quad \text{for all} \quad \varepsilon > 0,$$

# 6. Convergence

**1. Convergence in Distribution:** A sequence of random variables $\{X_n\}$ converges in distribution to a random variable $X$ if:

$$\lim_{n\to\infty} F_n(x) = F(x), \quad \text{for all continuity points of } F(x),$$

where $F_n(x)$ and $F(x)$ are the CDFs of $X_n$ and $X$, respectively, i.e., $F_n(x) = P(X_n \leq x)$ and $F(x) = P(X \leq x)$.

**2. Convergence in Probability:** A sequence of random variables $\{X_n\}$ converges in probability to a random variable $X$ if:

$$\lim_{n\to\infty} P\left(|X_n - X| > \varepsilon\right) = 0, \quad \text{for all} \quad \varepsilon > 0,$$

**3. Almost Sure Convergence:** A sequence of random variables $\{X_n\}$ converges almost surely to a random variable $X$ if:

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

# 6.1 General concepts: Probabilistic Big-O Notation

For a sequence of random variables $\{X_n\}$ and a corresponding sequence of deterministic constants $\{a_n\}$, we define:

**Big-O: Stochastic Boundedness**
The notation $X_n = O_p(a_n)$ indicates that $X_n$ is *stochastically bounded* by $a_n$. Formerly, for any $\varepsilon > 0$, there exist finite constants $M > 0$ and $N > 0$ such that $P\left(\left|\frac{X_n}{a_n}\right| > M\right) < \varepsilon \quad \text{for all} \quad n > N$. This ensures that $X_n/a_n$ does not grow unbounded with high probability.

# 6.1 General concepts: Probabilistic Big-O Notation

For a sequence of random variables $\{X_n\}$ and a corresponding sequence of deterministic constants $\{a_n\}$, we define:

**Big-O: Stochastic Boundedness**

The notation $X_n = O_p(a_n)$ indicates that $X_n$ is *stochastically bounded* by $a_n$. Formerly, for any $\varepsilon > 0$, there exist finite constants $M > 0$ and $N > 0$ such that $P\left(\left|\frac{X_n}{a_n}\right| > M\right) < \varepsilon \quad \text{for all} \quad n > N$. This ensures that $X_n/a_n$ does not grow unbounded with high probability.

**Small-o: Convergence in Probability**

The notation $X_n = o_p(a_n)$ means that $X_n/a_n$ *converges to zero in probability*. Formally: $\lim_{n\to\infty} P\left(\left|\frac{X_n}{a_n}\right| > \varepsilon\right) = 0 \quad \text{for any} \quad \varepsilon > 0$. This implies that the set of values $X_n/a_n$ becomes arbitrarily small with increasing $n$.

# 6.1 General concepts: Probabilistic Big-O notation

**Example:** Let $\{X_n\}_{n\geq 1}$ be a sequence of random variables. $E[X_n] = \mu_n$, $\sigma_n^2 = \text{Var}(X_n) < \infty$ for all $n$. Show that $X_n - \mu_n = O_P(\sigma_n)$.

- We say that $X_n - \mu_n = O_P(g(n))$, where $g(n)$ is some function of $n$, if for every $\epsilon > 0$, there exists a constant $M > 0$ and $N \geq 1$ such that: $\mathbb{P}\left(|X_n - \mu_n| > Mg(n)\right) < \epsilon, \quad$ for all $n \geq N$. This means that the probability that $X_n - \mu_n$ exceeds $Mg(n)$ in absolute value can be made arbitrarily small for large enough $n$.

# 6.1 General concepts: Probabilistic Big-O notation

**Example:** Let $\{X_n\}_{n\geq 1}$ be a sequence of random variables. $E[X_n] = \mu_n$, $\sigma_n^2 = \text{Var}(X_n) < \infty$ for all $n$. Show that $X_n - \mu_n = O_P(\sigma_n)$.

- We say that $X_n - \mu_n = O_P(g(n))$, where $g(n)$ is some function of $n$, if for every $\epsilon > 0$, there exists a constant $M > 0$ and $N \geq 1$ such that: $\mathbb{P}\left(|X_n - \mu_n| > Mg(n)\right) < \epsilon, \quad$ for all $n \geq N$. This means that the probability that $X_n - \mu_n$ exceeds $Mg(n)$ in absolute value can be made arbitrarily small for large enough $n$.
- To bound $|X_n - \mu_n|$, we can apply Chebyshev's inequality, which relates the variance of a random variable to the probability that it deviates from its mean. Specifically, Chebyshev's inequality states: $\mathbb{P}\left(|X_n - \mu_n| > k\sigma_n\right) \leq \frac{1}{k^2}$, where $k > 0$. This suggests that $X_n - \mu_n$ is typically of the order of $\sigma_n$.

# Probabilistic Big-O Notation

- Since the variance $\sigma_n^2$ measures the typical size of fluctuations of $X_n$ around its mean $\mu_n$, it is natural to consider the scaling $g(n) = \sigma_n$. Thus, we hypothesize that $X_n - \mu_n = O_P(\sigma_n)$.

# Probabilistic Big-O Notation

- Since the variance $\sigma_n^2$ measures the typical size of fluctuations of $X_n$ around its mean $\mu_n$, it is natural to consider the scaling $g(n) = \sigma_n$. Thus, we hypothesize that $X_n - \mu_n = O_P(\sigma_n)$.
- By Chebyshev's inequality, for any $\epsilon > 0$, we can take $M = \frac{1}{\sqrt{\epsilon}}$ to find that: $\mathbb{P}\left(|X_n - \mu_n| > M\sigma_n\right) < \epsilon$, for all $n \geq 1$. This verifies that $X_n - \mu_n$ is probabilistically bounded by $\sigma_n$, meaning that: $X_n - \mu_n = O_P(\sigma_n)$.

# Probabilistic Big-O Notation

- Since the variance $\sigma_n^2$ measures the typical size of fluctuations of $X_n$ around its mean $\mu_n$, it is natural to consider the scaling $g(n) = \sigma_n$. Thus, we hypothesize that $X_n - \mu_n = O_P(\sigma_n)$.
- By Chebyshev's inequality, for any $\epsilon > 0$, we can take $M = \frac{1}{\sqrt{\epsilon}}$ to find that: $\mathbb{P}\left(|X_n - \mu_n| > M\sigma_n\right) < \epsilon$, for all $n \geq 1$. This verifies that $X_n - \mu_n$ is probabilistically bounded by $\sigma_n$, meaning that: $X_n - \mu_n = O_P(\sigma_n)$.
- Thus, the probabilistic Big-O of $X_n - \mu_n$ is: $X_n - \mu_n = O_P(\sigma_n)$, which means that the deviations of $X_n$ from its mean $\mu_n$ are of the order of the standard deviation $\sigma_n$ with high probability as $n \to \infty$.

## 6.2.1 The weak law of large numbers

**Lemma (Weak Law of Large Numbers)**

Let $X_1, X_2, \dots$ be a sequence of i.i.d. (independent and identically distributed) random variables with mean $\mu$. Consider the sum:

$$S_n = X_1 + X_2 + \cdots + X_n.$$

As $n \to \infty$, the sample mean $\bar{X}_n = \frac{S_n}{n}$ converges to the population mean $\mu$ *in probability*. Formally:

$$\frac{S_n}{n} \xrightarrow{\mathsf{P}} \mu.$$

# 6.2.1 The weak law of large numbers

**Lemma (Weak Law of Large Numbers)**

Let $X_1, X_2, \dots$ be a sequence of i.i.d. (independent and identically distributed) random variables with mean $\mu$. Consider the sum:

$$S_n = X_1 + X_2 + \cdots + X_n.$$

As $n \to \infty$, the sample mean $\bar{X}_n = \frac{S_n}{n}$ converges to the population mean $\mu$ *in probability*. Formally:

$$\frac{S_n}{n} \xrightarrow{\mathsf{P}} \mu.$$

- This is a fundamental result in probability theory, underpinning the idea that averages of large samples tend to be close to the expected value.

# 6.2.2 The strong law of large numbers

## Lemma (Strong Law of Large Numbers)

Let $X_1$ , $X_2$, ... be a sequence of i.i.d. random variables with mean $\mu$. Consider the sum:

$$S_n = X_1 + X_2 + \cdots + X_n.$$

As $n \to \infty$, the sample mean $\bar{X}_n = \frac{S_n}{n}$ converges to the population mean $\mu$ *almost surely*. Formally:

$$\frac{S_n}{n} \xrightarrow{\text{a.s.}} \mu.$$

# 6.2.2 The strong law of large numbers

**Lemma (Strong Law of Large Numbers)**

Let $X_1$ , $X_2$, ... be a sequence of i.i.d. random variables with mean $\mu$. Consider the sum:

$$S_n = X_1 + X_2 + \cdots + X_n.$$

As $n \to \infty$, the sample mean $\bar{X}_n = \frac{S_n}{n}$ converges to the population mean $\mu$ *almost surely*. Formally:

$$\frac{S_n}{n} \xrightarrow{\text{a.s.}} \mu.$$

- This is a stronger form of the Law of Large Numbers compared to the Weak Law because it ensures convergence for every sample path, not just in probability.

# 6.3 Central Limit Theorem (CLT)

**Lemma (Central Limit Theorem)**

*Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with mean $\mu$ and variance $\sigma^2/n$. The sum (or average) of these variables, normalized by subtracting the mean and dividing by the standard deviation, converges in distribution to a standard normal distribution $\mathcal{N}(0,1)$ as $n \to \infty$.*

*Formally, if $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, then:*

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

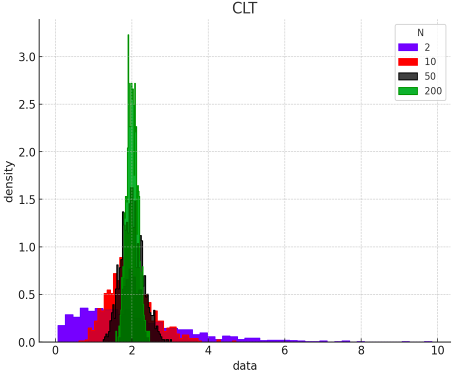
# 6.3 Central Limit Theorem (CLT)

**Lemma (Central Limit Theorem)**

*Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with mean $\mu$ and variance $\sigma^2/n$. The sum (or average) of these variables, normalized by subtracting the mean and dividing by the standard deviation, converges in distribution to a standard normal distribution $\mathcal{N}(0,1)$ as $n \to \infty$.*

*Formally, if $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, then:*

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




This theorem is fundamental, as it justifies using normal distribution in many real-world applications, even when the underlying data is not normally distributed.

# 7. The multivariate normal distribution

- The **multivariate normal distribution** is one of the most important distributions in multivariate statistics.
- It extends the properties of the one-dimensional Gaussian (Normal) distribution to higher dimensions, retaining its central role due to properties such as linear transformations and marginal distributions remaining Gaussian.
- In higher dimensions, the multivariate normal distribution is even more crucial because many distributions that work well in one dimension are not easily generalized to multiple dimensions.

# 7. The multivariate normal distribution

- The **multivariate normal distribution** is one of the most important distributions in multivariate statistics.
- It extends the properties of the one-dimensional Gaussian (Normal) distribution to higher dimensions, retaining its central role due to properties such as linear transformations and marginal distributions remaining Gaussian.
- In higher dimensions, the multivariate normal distribution is even more crucial because many distributions that work well in one dimension are not easily generalized to multiple dimensions.
- Understanding the multivariate normal distribution requires key results from **linear algebra**, including:
  - Covariance matrices
  - Eigenvalues and eigenvectors
  - Positive definiteness
- It is foundational in many statistical methods, including regression, PCA, and classification algorithms.

# Detour in Linear Algebra

The **inverse** of a square $p \times p$ matrix $A$, denoted $A^{-1}$, is the matrix that satisfies:

$$AA^{-1} = A^{-1}A = I_p,$$

where $I_p$ is the identity matrix of size $p \times p$.

- Only square matrices can have inverses.
- $A$ is **singular** or **non-invertible** if $A^{-1}$ doesn't exist.

# Detour in Linear Algebra

The **inverse** of a square $p \times p$ matrix $A$, denoted $A^{-1}$, is the matrix that satisfies:

$$AA^{-1} = A^{-1}A = I_p,$$

where $I_p$ is the identity matrix of size $p \times p$.

- Only square matrices can have inverses.
- $A$ is **singular** or **non-invertible** if $A^{-1}$ doesn't exist.

A symmetric $p \times p$ matrix $A$ is said to be **positive definite** if for all non-zero vectors $x \in \mathbb{R}^p$, $x^T Ax > 0$. If the inequality is non-strict (i.e., $x^T Ax \geq 0$), then $A$ is **positive semi-definite**.

# Detour in Linear Algebra

The **inverse** of a square $p \times p$ matrix $A$, denoted $A^{-1}$, is the matrix that satisfies:

$$AA^{-1} = A^{-1}A = I_p,$$

where $I_p$ is the identity matrix of size $p \times p$.

- Only square matrices can have inverses.
- $A$ is **singular** or **non-invertible** if $A^{-1}$ doesn't exist.

A symmetric $p \times p$ matrix $A$ is said to be **positive definite** if for all non-zero vectors $x \in \mathbb{R}^p$, $x^T Ax > 0$. If the inequality is non-strict (i.e., $x^T Ax \geq 0$), then $A$ is **positive semi-definite**.

- The **rank** of a matrix is the dimension of the largest nonsingular (invertible) submatrix it contains.

# Detour in Linear Algebra

The **inverse** of a square $p \times p$ matrix $A$, denoted $A^{-1}$, is the matrix that satisfies:

$$AA^{-1} = A^{-1}A = I_p,$$

where $I_p$ is the identity matrix of size $p \times p$.

- Only square matrices can have inverses.
- $A$ is **singular** or **non-invertible** if $A^{-1}$ doesn't exist.

A symmetric $p \times p$ matrix $A$ is said to be **positive definite** if for all non-zero vectors $x \in \mathbb{R}^p$, $x^T Ax > 0$. If the inequality is non-strict (i.e., $x^T Ax \geq 0$), then $A$ is **positive semi-definite**.

- The **rank** of a matrix is the dimension of the largest nonsingular (invertible) submatrix it contains.
- A square matrix of size $p \times p$ is **full rank** if its rank is $p$, meaning it is nonsingular (invertible).

# Detour in Linear Algebra

### Expectation and Variance Formulas for Linear Combinations

Let $A \in \mathbb{R}^{p\times p}$ be a matrix and $X = (X_1, X_2, \ldots, X_p)^{\mathrm{T}}$ a random vector with mean vector $\mu = (\mu_1, \mu_2, \ldots, \mu_p)^{\mathrm{T}}$ and covariance matrix $\Sigma = \mathrm{Cov}(X) = \mathbb{E}[(X - \mathbb{E}(X))(X - \mathbb{E}(X))^{\mathrm{T}}]$ with entries $\mathrm{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))]_{i,j=1,2,\ldots,p}$.

# Detour in Linear Algebra

**Expectation and Variance Formulas for Linear Combinations**

Let $A \in \mathbb{R}^{p \times p}$ be a matrix and $X = (X_1, X_2, \ldots, X_p)^{\mathrm{T}}$ a random vector with mean vector $\mu = (\mu_1, \mu_2, \ldots, \mu_p)^{\mathrm{T}}$ and covariance matrix $\Sigma = \mathrm{Cov}(X) = \mathbb{E}[(X - \mathbb{E}(X))(X - \mathbb{E}(X))^{\mathrm{T}}]$ with entries $\mathrm{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))]_{i,j=1,2,\ldots,p}$.

Then the following properties hold:

- $\mathbb{E}(A^{\mathrm{T}}X) = A^{\mathrm{T}}\mathbb{E}(X) = A^{\mathrm{T}}\mu$ (Linear transformation of expectation)

# Detour in Linear Algebra

**Expectation and Variance Formulas for Linear Combinations**

Let $A \in \mathbb{R}^{p \times p}$ be a matrix and $X = (X_1, X_2, \dots, X_p)^{\mathrm{T}}$ a random vector with mean vector $\mu = (\mu_1, \mu_2, \dots, \mu_p)^{\mathrm{T}}$ and covariance matrix $\Sigma = \mathsf{Cov}(X) = \mathbb{E}[(X - \mathbb{E}(X))(X - \mathbb{E}(X))^{\mathrm{T}}]$ with entries $\mathsf{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))]_{i,j=1,2,\dots,p}$.

Then the following properties hold:

- $\mathbb{E}(A^{\mathrm{T}} X) = A^{\mathrm{T}} \mathbb{E}(X) = A^{\mathrm{T}} \mu$ (Linear transformation of expectation)
- $\mathrm{Var}(A^{\mathrm{T}} X) \overset{?}{=} A^{\mathrm{T}} \Sigma A$ (Variance of a linear transformation)
- $\mathbb{E}(X^{\mathrm{T}} A X) \overset{?}{=} \mu^{\mathrm{T}} A \mu + \mathrm{tr}(A\Sigma)$ (Quadratic form expectation)
- $\mathsf{Cov}(A^{\mathrm{T}} X) \overset{?}{=} A^{\mathrm{T}} \mathbb{E}[(X - \mu)(X - \mu)^{\mathrm{T}}] A$ (Covariance of a linear transformation)

# Detour in Linear Algebra

For an $n \times n$ matrix $A$, the trace is given by: $\operatorname{tr}(A) = \sum_{i=1}^{n} A_{ii}$ where $A_{ii}$ is the element in the $i$-th row and $i$-th column.

# Detour in Linear Algebra

For an $n \times n$ matrix $A$, the trace is given by: $\mathrm{tr}(A) = \sum_{i=1}^{n} A_{ii}$ where $A_{ii}$ is the element in the $i$-th row and $i$-th column.

**Important Properties of Traces:**

- $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ (Cyclic property of the trace)
- $\mathrm{tr}(A + B) = \mathrm{tr}(A) + \mathrm{tr}(B)$ (Additivity of the trace)
- $\mathrm{tr}(cA) = c\,\mathrm{tr}(A)$ (Scaling factor can be pulled out of the trace)
- $\mathrm{tr}(A) = \mathrm{Rank}(A)$ if $A^2 = A$ (For idempotent matrices, trace = rank)

# Detour in Linear Algebra

For an $n \times n$ matrix $A$, the trace is given by: $\mathrm{tr}(A) = \sum_{i=1}^{n} A_{ii}$ where $A_{ii}$ is the element in the $i$-th row and $i$-th column.

**Important Properties of Traces:**

- $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ (Cyclic property of the trace)
- $\mathrm{tr}(A + B) = \mathrm{tr}(A) + \mathrm{tr}(B)$ (Additivity of the trace)
- $\mathrm{tr}(cA) = c\,\mathrm{tr}(A)$ (Scaling factor can be pulled out of the trace)
- $\mathrm{tr}(A) = \mathrm{Rank}(A)$ if $A^2 = A$ (For idempotent matrices, trace = rank)

**Key Properties of Ranks:**

- For $A$ and $B$ with appropriate dimensions:

$$\mathrm{Rank}(AB) \leq \min\{\mathrm{Rank}(A), \mathrm{Rank}(B)\}$$

  (Rank of a product is bounded by the rank of each factor)
- $\mathrm{Rank}(A^{\mathrm{T}}A) = \mathrm{Rank}(AA^{\mathrm{T}}) = \mathrm{Rank}(A)$

  (For any matrix $A$, the rank is preserved in these forms)

# 7.1 Standard normal distribution

A real random vector $\mathbf{Z}^{\mathrm{T}} = (z_1, \dots, z_p)$ is called a $p$-**variate standard normal random vector** if the components $Z_1, \dots, Z_p$ are mutually independent and each follows a standard normal distribution $\mathcal{N}(0, 1)$.

## 7.1 Standard normal distribution

A real random vector $\mathbf{Z}^{\mathrm{T}} = (z_1, \ldots, z_p)$ is called a $p$-**variate standard normal random vector** if the components $Z_1, \ldots, Z_p$ are mutually independent and each follows a standard normal distribution $\mathcal{N}(0, 1)$.

We write this as $\mathbf{Z} \sim \mathcal{N}_p(\mathbf{0}, I_p)$, where $\mathbf{0} = \mathbb{E}(\mathbf{Z})$ is a $p$-dimensional null vector, $I_p$ is a $p \times p$ identity matrix, and $\mathbf{Z}$ has the following probability density function (pdf):

$$f_{\mathbf{Z}}(z_1, \ldots, z_p) = \frac{1}{\sqrt{(2\pi)^p}} \exp\left(-\frac{1}{2}\mathbf{z}^T\mathbf{z}\right),$$

where $\mathbf{z}^T\mathbf{z} = z_1^2 + z_2^2 + \ldots + z_p^2$.

# 7.2 Multivariate normal distribution

A real random vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)^{\mathrm{T}}$ is called a **multivariate normal** random vector if there exists:

- A standard normal random vector $\mathbf{Z} \in \mathbb{R}^p$
- A mean vector $\boldsymbol{\mu} \in \mathbb{R}^p$
- A matrix $A \in \mathbb{R}^{p \times p}$

such that $\mathbf{X} = A\mathbf{Z} + \boldsymbol{\mu}$.

# 7.2 Multivariate normal distribution

A real random vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)^{\mathrm{T}}$ is called a **multivariate normal** random vector if there exists:

- A standard normal random vector $\mathbf{Z} \in \mathbb{R}^p$
- A mean vector $\boldsymbol{\mu} \in \mathbb{R}^p$
- A matrix $A \in \mathbb{R}^{p \times p}$

such that $\mathbf{X} = A\mathbf{Z} + \boldsymbol{\mu}$.

If $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ full rank, the PDF of $\mathbf{X}$ is given by:

$$f_{\mathbf{X}}(x_1, \ldots, x_p) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right),$$

where $|\boldsymbol{\Sigma}|$ is the determinant of the covariance matrix $\boldsymbol{\Sigma}$.

# 7.2 Multivariate normal distribution

## Key Properties

- **Covariance and Independence:**
  - Let $X \sim \mathcal{N}_p(\mu, \Sigma)$, where $X$ is a multivariate normal random vector.
  - If $\Sigma_{ij} = 0$ (i.e., the covariance between $X_i$ and $X_j$ is zero), then $X_i$ and $X_j$ are independent.

# 7.2 Multivariate normal distribution

## Key Properties

- **Covariance and Independence:**
  - Let $X \sim \mathcal{N}_p(\mu, \Sigma)$, where $X$ is a multivariate normal random vector.
  - If $\Sigma_{ij} = 0$ (i.e., the covariance between $X_i$ and $X_j$ is zero), then $X_i$ and $X_j$ are independent.
- **Linear Combinations:**
  - Linear combinations of multivariate normal variables are also normally distributed.
  - If $X \sim \mathcal{N}_p(\mu, \Sigma)$, $b$ is a $p \times 1$ vector of constants, and $B$ is a $p \times p$ matrix of constants, then $b + BX \sim \mathcal{N}_p(b + B\mu, B\Sigma B^T)$.

    Proof:

    - $\mathbb{E}(b + BX) = \mathbb{E}(b) + \mathbb{E}(BX) = b + B\mathbb{E}(X) = b + B\mu$
    - $Var(b + BX) \overset{?}{=} B\Sigma B^T$

# Summary

- What is the difference between discrete and continuous distributions?
- How does one calculate a random variable's expected value and variance?
- What are some basic properties of the expected value and the variance?
- independence of random variables
- How is convergence in probability defined?
- How is convergence in distribution defined?
- What do the CLT and LLN tell us?
- What is the probabilistic big-O notation?
- How to calculate mean and covariance of multivariate normal distribution?
- How is the rank of a matrix defined?