Probability and Distribution

AI ToolKit

Fall Semester, 2023

Reference: mml-book.github.io && d2l.ai

UNIST ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY 2007
Ahri Lab
AI & Human-Robot Interaction Laboratory

Coin Toss example

- How can we "quantify" the likelihood of observing heads?
  - Assuming the "fair" coin, both outcomes are equally likely.
- If we get n_t tails and n_h heads, let $n = n_h + n_t$.
  - We "expect" to see $n_t/n = 1/2$ and $n_h/n = 1/2$.
  - But can we say observed frequencies as "probabilities"? isn't it "statistics"?
- Probability : theoretical quantities that underlie the data generation process.
- Statistics : empirical quantities computed as functions of the observed data.

Sample space, event space, and probability.

- Sample space Ω : a set of all possible outcomes of the experiment.
    - i.e., if we toss coins for two times,
        - Ω = {  , , , }
- Event $\mathcal{A}$: subsets of the sample space.
    - i.e., the event of “the first coin toss comes up heads” is, $\mathcal{A}$ ={  , }
- Probability function: maps events onto real values, $P : \mathcal{A} \subseteq \Omega \rightarrow [0, 1]$.
    - The probability of any event is a non-negative real number, $P(\mathcal{A}) \geq 0$.
    - The probability of the entire sample space is 1, $P(\Omega) = 1$.
    - For any countable event sequence $\mathcal{A}_1, \mathcal{A}_2 \ldots$, if $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$ for $i \neq j$,

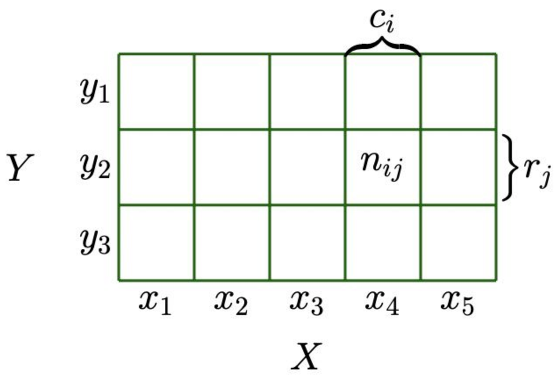
$$\Rightarrow P(\cup_{i=1}^{\infty} \mathcal{A}_i) = \sum_{i=1}^{\infty} P(\mathcal{A}_i)$$

Random Variables

- Mapping from an underlying sample space to a set of values.
- For coin-tossing example, assume that you consider “number of heads”.
  - A random variable X maps $X($$)=2$, $X($$)=1$, $X($$)=1$, $X($$)=0$.
  - Here, if $X : \Omega \to \mathcal{T}$, $\mathcal{T} = \{0, 1, 2\}$.
  - For any subset $S \subseteq \mathcal{T}$, $P_X(S) \in [0, 1]$.
  - $P_X(S) = P(X \in S) = P(X^{-1}(S)) = P(\{\omega \in \Omega : X(\omega) \in S\})$
  - Remember $\Omega = \{$, , , $\}$ in this example.
  - The function P_X is the *distribution of random variable* X.
    - If $\mathcal{T}$ is finite or countably infinite (i.e., integer), X is discrete random variable.
    - If $\mathcal{T} \in \mathbb{R}^n$, X is continuous random variable.

Discrete Probabilities

- When target space $\mathcal{T}$ is discrete (i.e., integer)
- We define “Probability mass function” (p.m.f) of a discrete probability distribution.



- For two random variables X and Y, $p(X = x, Y = y) = \frac{n_{ij}}{N}$

  where n_{ij} denote the number of events for states x_i and y_j,

  where N denote the total number of events.
- This **joint probability** is often written as $p(x, y)$ for $X = x$ and $Y = y$.
- The **marginal probability**: $p(x) = \sum_y p(x, y)$
- The **conditional probability of y given x** : $p(y|x)$

Continuous Probabilities

- A function $f : \mathbb{R}^D \to \mathbb{R}$ is called a probability density function (pdf) if
  1. $\forall \mathbf{x} \in \mathbb{R}^D : f(\mathbf{x}) \geq 0$
  2. $\int_{\mathbb{R}^D} f(\mathbf{x}) d\mathbf{x} = 1$
- $P(a \leq X \leq b) = \int_a^b f(x)dx = 1$, $P(X = x) = 0$
- A cumulative distribution function (c.d.f) of a multivariate real-valued random variable X with states $\mathbf{x} \in \mathbb{R}^D$ is:
  - $F_X(\mathbf{x}) = P(X_1 \leq x_1, ..., X_D \leq x_D)$
  - $F_X(\mathbf{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_D} f(z_1, ...z_D) dz_1 \cdots dz_D$

Rules of Probability

- Sum rule : $p(\boldsymbol{x}) = \begin{cases} \sum_{\boldsymbol{y}\in\mathcal{Y}} p(\boldsymbol{x},\boldsymbol{y}) & \text{if } \boldsymbol{y} \text{ is discrete} \\ \int_{\mathcal{Y}} p(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\boldsymbol{y} & \text{if } \boldsymbol{y} \text{ is continuous} \end{cases}$

- Product Rule : $p(\boldsymbol{x},\boldsymbol{y}) = p(\boldsymbol{y}\,|\,\boldsymbol{x})p(\boldsymbol{x})$

- Baye's Rule : $\underbrace{p(\boldsymbol{x}\,|\,\boldsymbol{y})}_{\text{posterior}} = \frac{\overbrace{p(\boldsymbol{y}\,|\,\boldsymbol{x})}^{\text{likelihood}}\,\overbrace{p(\boldsymbol{x})}^{\text{prior}}}{\underbrace{p(\boldsymbol{y})}_{\text{evidence}}}$

: making an inference of unobserved (latent) random variable $\mathbf{x}$,
from observed values of $\mathbf{y}$.

Means

- The expected value of a function $g : \mathbb{R} \to \mathbb{R}$ of a univariate continuous random variable $X \sim p(x)$ is given by:
  - $\mathbb{E}_X[g(x)] = \int_{\mathcal{X}} g(x)p(x)dx$, and for discrete random variable, $\mathbb{E}_X[g(x)] = \sum_{x \in \mathcal{X}} g(x)p(x)$.
  - This a linear operator.
- The mean of a random variable X with states $\mathbf{x} \in \mathbb{R}^D$ is defined as:
  - $\mathbb{E}_X[\boldsymbol{x}] = \begin{bmatrix} \mathbb{E}_{X_1}[x_1] \\ \vdots \\ \mathbb{E}_{X_D}[x_D] \end{bmatrix} \in \mathbb{R}^D$, where $\mathbb{E}_{X_d}[x_d] := \begin{cases} \int_{\mathcal{X}} x_d p(x_d) \mathrm{d}x_d & \text{if } X \text{ is a continuous random variable} \\ \sum_{x_i \in \mathcal{X}} x_i p(x_d = x_i) & \text{if } X \text{ is a discrete random variable} \end{cases}$
- Modes : most frequent value.
- Median : the middle value.

Covariance

- For univariate random variables, $X, Y \in \mathbb{R}$, the covariance between two is given as the expected product of their deviations from their respective means:

  $\Rightarrow \mathrm{Cov}_{X,Y}[x,y] := \mathbb{E}_{X,Y}\left[(x-\mathbb{E}_X[x])(y-\mathbb{E}_Y[y])\right].$

  ⇒ The covariance of itself is called as variance, and its root is standard deviation.

  ⇒ By using the linearity of expectations, $\mathrm{Cov}[x,y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]$.

Covariance

- For multivariate random variables, X, Y where $\mathbf{x} \in \mathbb{R}^D, \mathbf{y} \in \mathbb{R}^E$, covariance is a matrix as:

$\Rightarrow \mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] = \mathbb{E}[\boldsymbol{x}\boldsymbol{y}^\top] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{y}]^\top = \mathrm{Cov}[\boldsymbol{y}, \boldsymbol{x}]^\top \in \mathbb{R}^{D \times E}$

⇒ The covariance of itself is also called a variance, as below.

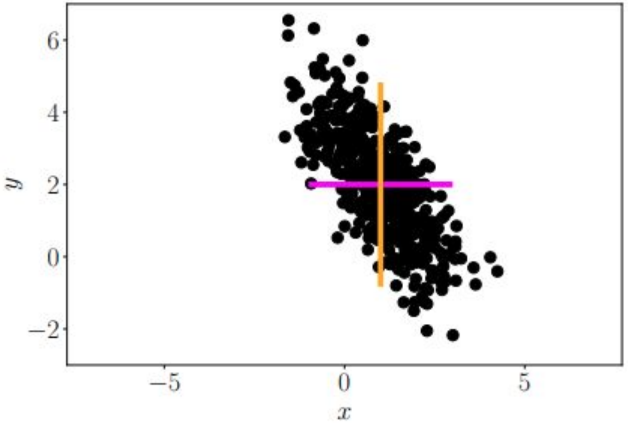
$$\begin{aligned}
\mathbb{V}_X[\boldsymbol{x}] &= \mathrm{Cov}_X[\boldsymbol{x}, \boldsymbol{x}] \\
&= \mathbb{E}_X[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top] = \mathbb{E}_X[\boldsymbol{x}\boldsymbol{x}^\top] - \mathbb{E}_X[\boldsymbol{x}]\mathbb{E}_X[\boldsymbol{x}]^\top \\
&= \left.\begin{bmatrix} \mathrm{Cov}[x_1, x_1] & \mathrm{Cov}[x_1, x_2] & \dots & \mathrm{Cov}[x_1, x_D] \\ \mathrm{Cov}[x_2, x_1] & \mathrm{Cov}[x_2, x_2] & \dots & \mathrm{Cov}[x_2, x_D] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[x_D, x_1] & \dots & \dots & \mathrm{Cov}[x_D, x_D] \end{bmatrix} . \right\rbrace \text{A symmetric and positive-semidefinite Covariance matrix.}
\end{aligned}$$

Correlation

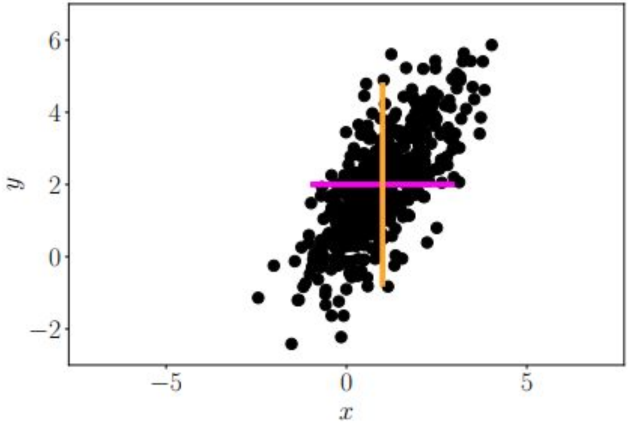
- The correlation between two random variables $X, Y \in \mathbb{R}$ is given by

$$\Rightarrow \operatorname{corr}[x, y] = \frac{\operatorname{Cov}[x, y]}{\sqrt{\mathbb{V}[x]\mathbb{V}[y]}} \in [-1, 1]$$

⇒ Indicates how two random variables are related.



(a) x and y are negatively correlated.



(b) x and y are positively correlated.

Empirical Mean and Covariance

- Given a finite dataset of size N, we can obtain an estimate of the mean.
- The empirical mean / sample mean vector is the arithmetic average observations for each variable, defined as $\bar{\boldsymbol{x}} := \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n$. where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ are observations.
- The empirical covariance matrix is $\boldsymbol{\Sigma} := \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})^\top$.

Sums and Transformations of Random Variables

- For two random variables X, Y with states $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$,

$$
\begin{aligned}
\mathbb{E}[\boldsymbol{x}+\boldsymbol{y}] &= \mathbb{E}[\boldsymbol{x}]+\mathbb{E}[\boldsymbol{y}] \\
\mathbb{E}[\boldsymbol{x}-\boldsymbol{y}] &= \mathbb{E}[\boldsymbol{x}]-\mathbb{E}[\boldsymbol{y}] \\
\mathbb{V}[\boldsymbol{x}+\boldsymbol{y}] &= \mathbb{V}[\boldsymbol{x}]+\mathbb{V}[\boldsymbol{y}]+\operatorname{Cov}[\boldsymbol{x}, \boldsymbol{y}]+\operatorname{Cov}[\boldsymbol{y}, \boldsymbol{x}] \\
\mathbb{V}[\boldsymbol{x}-\boldsymbol{y}] &= \mathbb{V}[\boldsymbol{x}]+\mathbb{V}[\boldsymbol{y}]-\operatorname{Cov}[\boldsymbol{x}, \boldsymbol{y}]-\operatorname{Cov}[\boldsymbol{y}, \boldsymbol{x}]
\end{aligned}
$$

- What would happen for $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$?

$$
\begin{aligned}
\Rightarrow\ & \mathbb{E}_Y[\boldsymbol{y}] = \mathbb{E}_X[\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}] = \boldsymbol{A}\mathbb{E}_X[\boldsymbol{x}]+\boldsymbol{b} = \boldsymbol{A}\boldsymbol{\mu}+\boldsymbol{b}\,, \\
& \mathbb{V}_Y[\boldsymbol{y}] = \mathbb{V}_X[\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}] = \mathbb{V}_X[\boldsymbol{A}\boldsymbol{x}] = \boldsymbol{A}\mathbb{V}_X[\boldsymbol{x}]\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{Cov}[\boldsymbol{x}, \boldsymbol{y}] &= \mathbb{E}[\boldsymbol{x}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})^\top] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}]^\top \\
&= \mathbb{E}[\boldsymbol{x}]\boldsymbol{b}^\top + \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]\boldsymbol{A}^\top - \boldsymbol{\mu}\boldsymbol{b}^\top - \boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{A}^\top \\
&= \boldsymbol{\mu}\boldsymbol{b}^\top - \boldsymbol{\mu}\boldsymbol{b}^\top + \left(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top\right)\boldsymbol{A}^\top \\
&\overset{(6.38\text{b})}{=} \boldsymbol{\Sigma}\boldsymbol{A}^\top\,,
\end{aligned}
$$

Independence

- Two random variables X, Y are independent if and only if $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$, and below properties are hold:
  - $p(\boldsymbol{y} \mid \boldsymbol{x}) = p(\boldsymbol{y})$
  - $p(\boldsymbol{x} \mid \boldsymbol{y}) = p(\boldsymbol{x})$
  - $\mathbb{V}_{X,Y}[\boldsymbol{x} + \boldsymbol{y}] = \mathbb{V}_X[\boldsymbol{x}] + \mathbb{V}_Y[\boldsymbol{y}]$
  - $\mathrm{Cov}_{X,Y}[\boldsymbol{x}, \boldsymbol{y}] = \mathbf{0}$

- Two random variables X, Y are conditionally independent given Z if and only if:

$$p(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{z}) = p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{z})p(\boldsymbol{y} \mid \boldsymbol{z}) = p(\boldsymbol{x} \mid \boldsymbol{z})p(\boldsymbol{y} \mid \boldsymbol{z}) \quad \text{for all} \quad \boldsymbol{z} \in \mathcal{Z}$$

- Here, what would $p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{z}) = p(\boldsymbol{x} \mid \boldsymbol{z})$ imply?

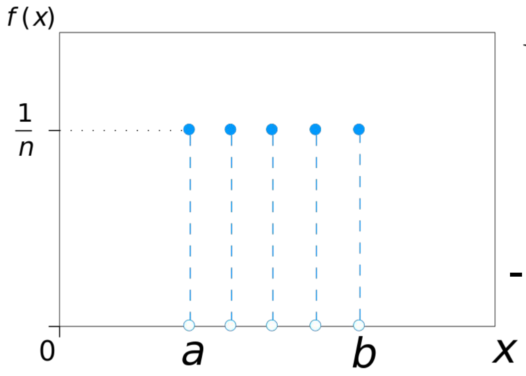
Bernoulli distribution

- For a single binary random variable X with state $x = \{0, 1\}$ (i.e., coin flip) assume that we observe $X = 1$ with probability μ.
- We write this as $X \sim \text{Bernoulli}(\mu)$, and
  - $p(x|\mu) = \mu^x(1-\mu)^{1-x}, \; x \in \{0, 1\}$
  - $\mathbb{E}[x] = \mu$
  - $\mathbb{V}[x] = \mu(1-\mu)$
- How we get mean and variance?

  $\Rightarrow$

Uniform Distribution

- When every possible values is equally likely.
- For discrete random variable X with states $x \in \{a, a+1, ..., b-1, b\}$, its p.m.f is :



where $n = b - a + 1$.

- $\mathbb{E}[x] = \dfrac{a+b}{2}, \quad \mathbb{V}[x] = \dfrac{n^2-1}{12}.$

Uniform Distribution

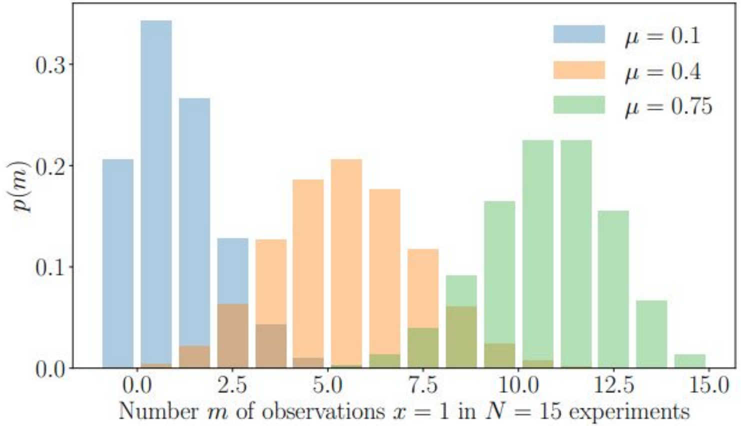
- When every possible values is equally likely.
- For continuous random variable X with states $x \in [a, b]$, its p.d.f is :
  - $p(x) \begin{cases} \frac{1}{b-a} & x \in [a,b], \\ 0 & x \notin [a,b] \end{cases}$
  - $\mathbb{E}[x] = \dfrac{a+b}{2}, \quad \mathbb{V}[x] = \dfrac{(b-a)^2}{12}$

$\Rightarrow$

Binomial Distribution

- Performing a sequence of n independent experiments, where the probability of each success is μ.
- Each experiment is an independent random variable X_i , which encode success with 1, failure with 0, such that $X_i \sim \text{Bernoulli}(\mu)$.
- Then, if $X = \sum_{i=1}^{n} X_i$, we write $X \sim \text{Binomial}(n, \mu)$ and $p(x) = \binom{n}{k} \mu^k (1-\mu)^{n-k}$,

  where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.
- $\mathbb{E}[x] = np$, $\mathbb{V}[x] = np(1-p)$.



Poisson Distribution

- The probability of a given number of events what can occur in a fixed interval of time/space.
- Let us assume that “event” is “one bus arrives bus stop”, in a fixed “1 minute” time window.
    - Denote this with $X^{(1)} \sim \mathrm{Bernoulli}(p)$.
    - If you want to model “max. two bus arrive bus stop”, in a fixed “1 minute”, you can…

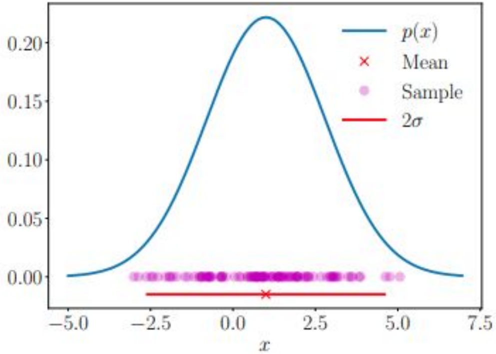
$$X_1^{(2)} \sim \mathrm{Bernoulli}(p/2),\ X_2^{(2)} \sim \mathrm{Bernoulli}(p/2),$$

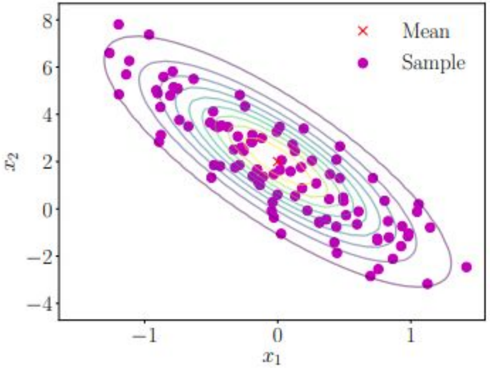
$$X^{(2)} \sim X_1^{(2)} + X_2^{(2)} \quad , X^{(2)} \sim \mathrm{Binomial}(2, p/2),$$

- Consider $X^{(n)} \sim \mathrm{Binomial}(n, p/n)$, and try to $n \to \infty$. Then, $\mathbb{E}[X^{(\infty)}] = p$ and $\mathbb{V}[X^{(\infty)}] = p$.
- We say $X \sim \mathrm{Poisson}(\lambda)$,
  if it is a random variable which takes the non-negative integer values with probability $p(X = k) = \dfrac{\lambda^k e^{-\lambda}}{k!}$.
- λ is a rate/shape denoting the average number of events we expect in one unit of time.

Gaussian Distribution

- For univariate random variable, the Gaussian distribution has a density as:
- $p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, where μ is mean and σ^2 is variance.
- For D-dimensional random variable, the Gaussian distribution has a density as:
- $p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}}|\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$
    - We write $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ or $X \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$.





Properties of Gaussian Distribution

- For two multivariate random variables X and Y, If joint distribution is defined as:

$$p(\mathbf{x}, \mathbf{y}) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{bmatrix}\right)$$

, where

$$\begin{aligned} \boldsymbol{\Sigma}_{xx} &= \mathrm{Cov}[\mathbf{x}, \mathbf{x}] \\ \boldsymbol{\Sigma}_{yy} &= \mathrm{Cov}[\mathbf{y}, \mathbf{y}]. \\ \boldsymbol{\Sigma}_{xy} &= \mathrm{Cov}[\mathbf{x}, \mathbf{y}] \end{aligned}$$

- Then, conditional distribution is also Gaussian, such that:

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \mathcal{N}(\boldsymbol{\mu}_{x \mid y}, \boldsymbol{\Sigma}_{x \mid y})$$

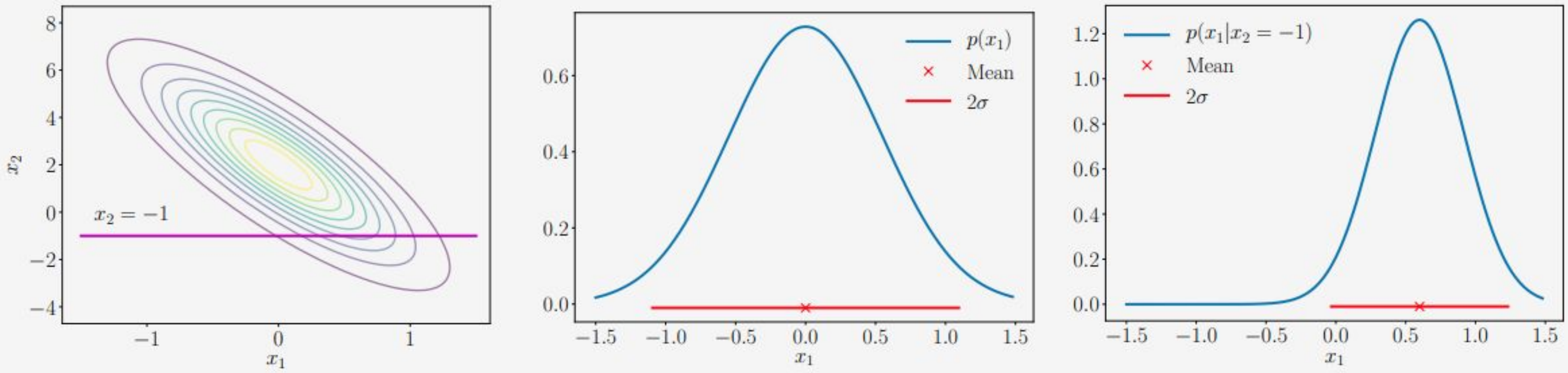
$$\boldsymbol{\mu}_{x \mid y} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_y)$$

$$\boldsymbol{\Sigma}_{x \mid y} = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1} \boldsymbol{\Sigma}_{yx} \,.$$

- The marginal distribution is also Gaussian,

where $p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{y} = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx})$, same for Y.

Properties of Gaussian Distribution

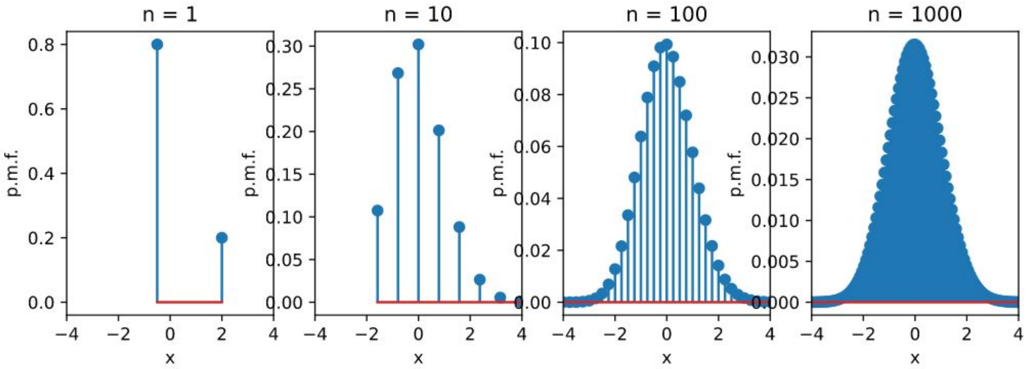


(a) Bivariate Gaussian. (b) Marginal distribution. (c) Conditional distribution.

Central Limit Theorem (CLT)

- Let be $X_1, X_2, \ldots$ i.i.d (independent, identically distributed) random variables with finite mean μ and finite variance σ^2.
- The sample average is $\bar{X}_n \equiv \dfrac{X_1 + \cdots + X_n}{n}$.

- By the law of large numbers, the sample average converge almost surely to μ as $n \to \infty$.
- Central Limit Theorem:
  - For large enough n, the distribution of $\bar{X}_n$ gets arbitrarily close to the Gaussian distribution with mean μ and variance σ^2/n.



Ahri Lab

Any Questions?