Probability Theory

Review

Faculty of Computer Science
University of Information Technology (UIT)
Vietnam National University - Ho Chi Minh City (VNU-HCM)

Maths for Computer Science, Fall 2023

References

The contents of this document are taken mainly from the follow sources:

- John Tsitsiklis. Massachusetts Institute of Technology. Introduction to Probability.[1]
- Marek Rutkowski. University of Sydney. Probability Review.[2]
- https://www.probabilitycourse.com/

[1]https://ocw.mit.edu/resources/res-6-012-introduction-to-probability-spring-2018/index.htm

[2]http://www.maths.usyd.edu.au/u/UG/SM/MATH3075/r/Slides_1_Probability.pdf

Table of Contents

Table of Contents

Sample Space

- List (set) of all possible states of the world, Ω. The states are called **samples** or **elementary events**.
- List (set) of possible **outcomes**, Ω.
- E.g., flipping a coin twice, the sample space will be a set of 4 elements:

$$\Omega = \{(heads, heads), (heads, tails), (tails, heads), (tails, tails)\}$$

- List must be:
  - Mutually exclusive
  - Collectively exhaustive
  - At the "right" granularity
- The sample space Ω is either **countable** or **uncountable**.

Probability

A discrete sample space $\Omega = (\omega_k)_{k \in I}$, where the set I is countable.

Definition (Probability)

A map $P : \Omega \mapsto [0, 1]$ is called a **probability** on a discrete sample space Ω if the following conditions are satisfied:

- $P(\omega_k) \geq 0$ for all $k \in I$
- $\sum_{k \in I} P(\omega_k) = 1$

Probability Measure

- Let $\mathcal{F} = 2^{\Omega}$ be the set of all subsets of the sample space Ω.
- $\mathcal{F}$ contains the **empty set** $\emptyset$ and Ω.
- Any set $A \in \mathcal{F}$ is called an **event** (or a **random event**).
- The set $\mathcal{F}$ is called the **event space**.
- Probability is assigned to **events**.

Definition (Probability Measure)

A map $P : \mathcal{F} \mapsto [0, 1]$ is called a **probability measure** on $(\Omega, \mathcal{F})$ if

- For any sequence $A_i \in \mathcal{F}, i = 1, 2, \ldots$ of events such that $A_i \cap A_j = \emptyset$ for all $i \neq j$ we have

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

- $P(\Omega) = 1$

Probability Measure

- A probability $P : \Omega \mapsto [0, 1]$ on a discrete sample space Ω uniquely specifies probability of all events $A_k = \{\omega_k\}$.
- $P(\{\omega_k\}) = P(\omega_k) = p_k$.

Theorem

Let $P : \Omega \mapsto [0, 1]$ be a probability on a discrete sample space Ω. Then the unique probability measure on $(\Omega, \mathcal{F})$ generated by P satisfies for all $A \in \mathcal{F}$

$$P(A) = \sum_{\omega_k \in A} P(\omega_k)$$

Some properties of probability

- If $A \subset B$, then $P(A) \leq P(B)$.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cup B) \leq P(A) + P(B)$
- $P(A \cup B \cup C) = P(A) + P(A^c \cap B) + P(A^c \cap B^c \cap C)$

Table of Contents

Random Variables

Definition (Random Variable)

A random variable X on a sample space Ω is a real-valued function on Ω:

$$X : \Omega \to \mathbb{R}$$

- A r.v. associates a value (a number) to every possible outcome.
- It can take discrete or continuous values.

Notation

Random variable X	Numerical value x

- Different r.v.'s can be defined on the same sample space.
- A function of one or several random variables is also a r.v.

Random Variables

- E.g., flipping a coin twice. The sample space Ω will be:

$$\Omega = \{(heads, heads), (heads, tails), (tails, heads), (tails, tails)\}$$

- If the first coin toss returns heads, you win \$10. Else, you lose \$5.
- If the second coin toss returns heads, you win \$15. Else, you lose \$12.
- We can use a random variable to model the above game:

$$X : \Omega \to \mathbb{R}$$

as

$$\begin{aligned} X(heads, heads) &= 10 + 15 = 25 \\ X(heads, tails) &= 10 - 12 = -2 \\ X(tails, heads) &= -5 + 15 = 10 \\ X(tails, tails) &= -5 - 12 = -17 \end{aligned}$$

Probability Mass Function (pmf)

Probability mass function (pmf) of a discrete random variable X.

- It is the "probability law" or "probability distribution" of X.
- If we fix some x, then "$X = x$" is an event.

Definition

$$p_X(x) = P(X = x) = P(\{\omega \in \Omega \text{ s.t. } X(\omega) = x\})$$

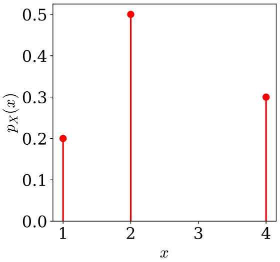
Properties

- $p_X(x) \geq 0$
- $\sum_x p_X(x) = 1$

Expectation

- Example: Play a game 1000 times. Random gain at each game is described by

$$X = \begin{cases} 1, & \text{with probability } 2/10 \sim 200 \\ 2, & \text{with probability } 5/10 \sim 500 \\ 4, & \text{with probability } 3/10 \sim 300 \end{cases}$$



- "Average" gain:

$$\frac{1 \cdot 200 + 2 \cdot 500 + 4 \cdot 300}{1000} = 2.4$$

- Definition: $E[X] = \sum_x x p_X(x)$

Expectation

- $E[X] = \sum_{x} x p_X(x)$
- $E(\cdot)$ is called the expectation operator.
- Average in a large number of independence experiments.
- Expectation of a r.v. can be seen as the weighted average.
- It is impossible to know the exact event to happen in the future and thus expectation is useful in making decisions when the probabilities of future outcomes are known.
- Any random variable defined on a finite set Ω admits the expectation.
- When the set Ω is countable but infinite, we need $\sum_{x} |x| p_X(x) < \infty$ so that $E[X]$ is well-defined.

Expectation

Definition

The **expectation** (**expected value** or **mean value**) of a random variable X on a discrete sample space Ω is given by

$$E_P(X) = \mu := \sum_{k \in I} X(\omega_k) P(\omega_k) = \sum_{k \in I} x_k p_k$$

where P is a probability measure on Ω.

Definition

The **expectation** (**expected value** or **mean value**) of a discrete random variable X with range $R_X = \{x_1, x_2, x_3, \ldots\}$ (finite or countably infinite) is defined as

$$E(X) = \mu := \sum_{x_k \in R_X} x_k P(X = x_k) = \sum_{x_k \in R_X} x_k P_X(x_k)$$

Elementary Properties of Expectation

Definition

$$E[X] = \sum_x x p_X(x)$$

- If $X \geq 0$ then $E[X] \geq 0$.
  For all outcomes $w : X(w) \geq 0$.

Elementary Properties of Expectation

Definition

$$E[X] = \sum_x x p_X(x)$$

- If $X \geq 0$ then $E[X] \geq 0$.
  For all outcomes $w : X(w) \geq 0$.
- If $a \leq X \leq b$ then $a \leq E[X] \leq b$.
  For all outcomes $w : a \leq X(w) \leq b$.
  $E[X] = \sum_x x p_X(x) \geq \sum a p_X(x) = a \sum_x p_X(x) = a \cdot 1 = a$

Elementary Properties of Expectation

Definition

$$E[X] = \sum_x x p_X(x)$$

- If $X \geq 0$ then $E[X] \geq 0$.
  For all outcomes $w : X(w) \geq 0$.
- If $a \leq X \leq b$ then $a \leq E[X] \leq b$.
  For all outcomes $w : a \leq X(w) \leq b$.
  $E[X] = \sum_x x p_X(x) \geq \sum a p_X(x) = a \sum_x p_X(x) = a \cdot 1 = a$
- If c is a constant, $E[c] = c$
  $E[c] = c \cdot p(c) = c$

Expected value rule, to compute $E[g(X)]$

- If X is a r.v. and $Y = g(X)$, then Y itself is a r.v.
- Average over y:

$$E[Y] = \sum_y y p_Y(y)$$

- Average over x:

Theorem (Law of the unconscious statistician (LOTUS))

$$E[Y] = E[g(X)] = \sum_x g(x) p_X(x)$$

$$\sum_y \sum_{x:g(x)=y} g(x) p_X(x) = \sum_y \sum_{x:g(x)=y} y p_X(x) = \sum_y y \sum_{x:g(x)=y} p_X(x)$$

$$= \sum_y y p_Y(y) = E[Y]$$

- $E[X^2] = \sum x^2 p_X(x) \neq (E[X])^2$. Caution: $E[g(X)] \neq g(E[X])$.

Linearity of Expectation

Theorem

$$E[aX + b] = aE[X] + b$$

Example: X = salary $\quad E[X]$ = average salary.
Y = new salary $= 2X + 100 \quad E[Y] = E[2X + 100] = 2E[X] + 100$.

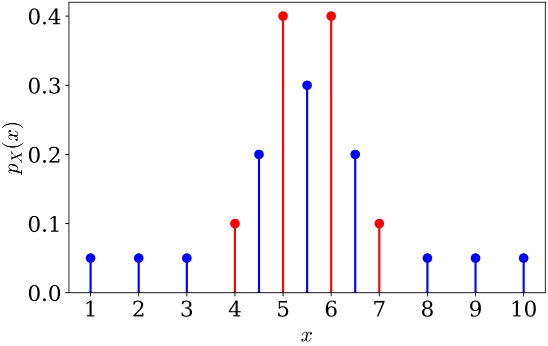
Proof

Based on the expected value rule: $g(x) = ax + b; \quad Y = g(X)$

$$\begin{aligned} E[Y] &= \sum_x g(x)p_X(x) \\ &= \sum_x (ax + b)p_X(x) \\ &= a\sum_x xp_X(x) + b\sum_x p_X(x) \\ &= aE[X] + b \end{aligned}$$

Variance

- Variance is a measure of the spread of a random variable about its mean and also a measure of uncertainty.



- R.v. X with $\mu = E[X]$. Average distance from the mean?

$$\mathrm{E}[X - \mu] = \mathrm{E}[X] - \mu = \mu - \mu = 0$$

Variance

- Variance is a measure of the spread of a random variable about its mean and also a measure of uncertainty.
- R.v. X with $\mu = E[X]$. Average distance from the mean?

$$\mathrm{E}\left[X - \mu\right] = \mathrm{E}\left[X\right] - \mu = \mu - \mu = 0$$

- Average of the squared distance from the mean.

Definition (Variance)

The variance of a random variable X on a discrete sample space Ω is defined as

$$Var(X) = \sigma^2 = \mathrm{E}_P[(X - \mu)^2],$$

where P is a probability measure on Ω.

Variance

- $Var(X) = \sigma^2 = \mathrm{E}\left[(X-\mu)^2\right]$
- $g(x) = (x-\mu)^2$
- To calculate, use the expected value rule, $E[g(X)] = \sum_x g(x)p_X(x)$

$$Var(X) = \mathrm{E}\left[g(X)\right] = \sum_x (x-\mu)^2 p_X(x)$$

- Variance is non-negative: $Var(X) = \sigma^2 \geq 0$.
- $Var(X) = 0$ iff X is deterministic.

Definition (Standard Deviation)

The **standard deviation** of a random variable X is defined as

$$SD(X) = \sigma_X = \sqrt{Var(X)}$$

Properties of the variance

Theorem

For a random variable X and real numbers a and b,

$$Var(aX + b) = a^2 Var(X)$$

- Notation $\mu = E[X]$
- Let $Y = X + b$, $\gamma = E[Y] = \mu + b$.

$$Var(Y) = E[(Y-\gamma)^2] = E[(X+b-(\mu+b))^2] = E[(X-\mu)^2] = Var(X)$$

- Let $Y = aX$, $\gamma = E[Y] = a\mu$

$$\begin{aligned} Var(Y) &= E[(aX - a\mu)^2] = E[a^2(X - \mu)^2] \\ &= a^2 E[(X - \mu)^2] = a^2 Var(X) \end{aligned}$$

Computational formula for the variance

$$Var(X) = E(X^2) - [E(X)]^2$$

$$\begin{aligned}
Var(X) &= E[(X-\mu)^2] \\
&= E[X^2 - 2\mu X + \mu^2] \\
&= E[X^2] - 2\mu E[X] + \mu^2 \\
&= E[X^2] - (E[X])^2
\end{aligned}$$

Independence and Expectation

- In general: $\mathrm{E}[g(X,Y)] \neq g(\mathrm{E}[X], \mathrm{E}[Y])$
- Exceptions:
  $\mathrm{E}[aX+b] = a\mathrm{E}[X] + b \qquad \mathrm{E}[X+Y+Z] = \mathrm{E}[X] + \mathrm{E}[Y] + \mathrm{E}[Z]$

Theorem

If X, Y are independent: $\mathrm{E}[X,Y] = \mathrm{E}[X]\mathrm{E}[Y]$,
$g(X)$ and $h(Y)$ are also independent: $\mathrm{E}[g(X), h(Y)] = \mathrm{E}[g(X)]\mathrm{E}[h(Y)]$

Independence and Variances

- Always true: $\mathrm{Var}\left(aX\right) = a^2\mathrm{Var}\left(X\right) \qquad \mathrm{Var}\left(X+a\right) = \mathrm{Var}\left(X\right)$
- In general: $\mathrm{Var}\left(X+Y\right) \neq \mathrm{Var}\left(X\right) + \mathrm{Var}\left(Y\right)$
- However

Theorem

If X, Y are independent: $\mathrm{Var}\left(X+Y\right) = \mathrm{Var}\left(X\right) + \mathrm{Var}\left(Y\right)$

Proof.

Assume $\mathrm{E}\left[X\right] = \mathrm{E}\left[Y\right] = 0 \qquad \mathrm{E}\left[XY\right] = \mathrm{E}\left[X\right]\mathrm{E}\left[Y\right] = 0.$

$$\begin{aligned} \mathrm{Var}\left(X+Y\right) &= \mathrm{E}\left[(X+Y)^2\right] = \mathrm{E}\left[X^2 + 2XY + Y^2\right] \\ &= \mathrm{E}\left[X^2\right] + 2\mathrm{E}\left[XY\right] + \mathrm{E}\left[Y^2\right] = \mathrm{Var}\left(X\right) + Var(Y) \end{aligned}$$

$\square$

Table of Contents

Bernoulli Random Variables

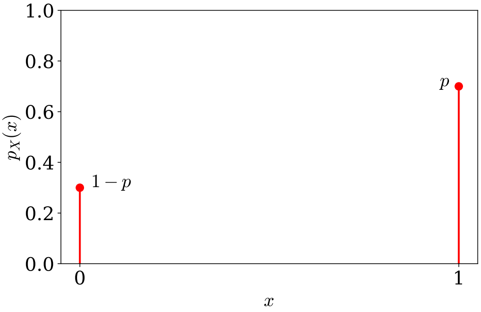
- A Bernoulli r.v. X takes two possible values, usually 0 and 1, modeling random experiments that have two possible outcomes (e.g., "success" and "failure").
  - e.g., tossing a coin. The outcome is either Head or Tail.
  - e.g., taking an exam. The result is either Pass or Fail.
  - e.g., classifying images. An image is either Cat or Non-cat.

Bernoulli Random Variables

Definition

A random variable X is a Bernoulli random variable with parameter $p \in [0,1]$, written as $X \sim Bernoulli(p)$ if its PMF is given by

$$P_X(x) = \begin{cases} p, & \text{for } x = 1 \\ 1-p, & \text{for } x = 0. \end{cases}$$



Bernoulli & Indicator Random Variables

- A Bernoulli r.v. X with parameter $p \in [0, 1]$ can also be described as

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$$

- A Bernoulli r.v. is associated with a certain event A. If event A occurs, then $X = 1$; otherwise, $X = 0$.
- Bernoulli r.v. is also called the indicator random variable of an event.

Definition

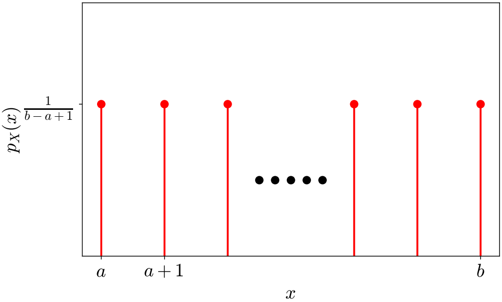
The indicator random variable of an event A is defined by

$$I_A = \begin{cases} 1 & \text{if the event } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

The indicator r.v. for an event A has Bernoulli distribution with parameter $p = P(I_A = 1) = P_{I_A}(1) = P(A)$. We can write $I_A \sim Bernoulli((P(A))$.

Discrete Uniform Random Variables

- Parameters: integer a, b; $\quad a \le b$
- Experiment: Pick one of $a, a+1, \ldots, b$ at random; all equally likely.
- Sample space: $\{a, a+1, \ldots, b\}$
- Random variable X : $X(\omega) = \omega$
- $b - a + 1$ possible values, $P_X(x) = 1/(b-a+1)$ for each value.
- Model of: complete ignorance.



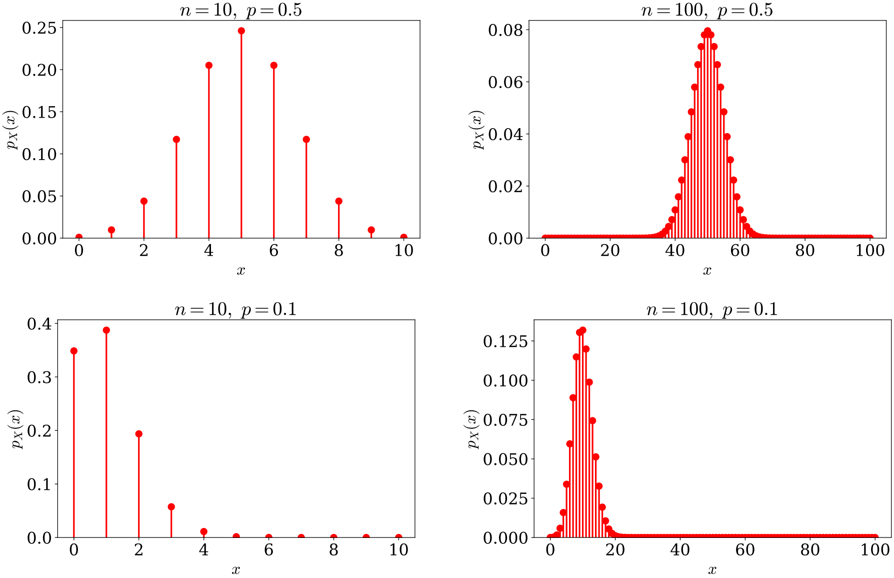
Binomial Random Variables

- Parameters: Probability $p \in [0, 1]$, positive integer n.
- Experiment: e.g., n independent tosses of a coin with $P(\text{Head}) = p$
- Sample space: Set of sequences of H and T of length n
- Random variable X : number of Heads observed.
- Model of: number of successes in a given number of independent trials.

Examples

$$\begin{aligned} P_X(2) &= P(X = 2) \\ &= P(\text{HHT}) + P(\text{HTH}) + P(\text{THH}) \\ &= 3p^2(1 - p) \\ &= \binom{3}{2} p^2 (1 - p) \end{aligned}$$

Binomial Random Variables



Binomial Random Variables

- Let $\Omega = \{0, 1, 2, \ldots, n\}$ be the sample space and let X be the number of successes in n independent trials where p is the probability of success in a single Bernoulli trial.
- The probability measure P is called the binomial distribution if

$$P_X(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, \ldots, n$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

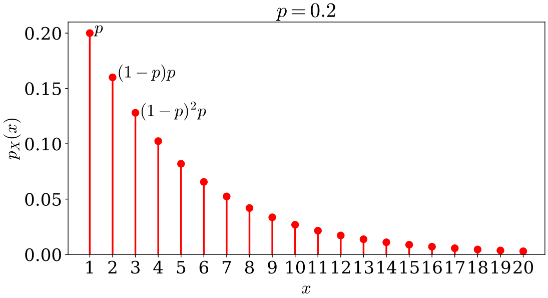
- Then

$$E[X] = np \quad \text{and} \quad Var(X) = np(1-p)$$

Geometric Random Variables

- Parameters: Probability $p \in (0, 1]$.
- Experiment: infinitely many independent tosses of a coin; $P(\text{Head}) = p$.
- Sample space: Set of infinite sequences of H and T.
- Random variable X : number of tosses until the first Head.
- Model of: waiting times, number of trials until a success.

$$P_X(k) = P(X = k) = P(\underbrace{\text{T} \ldots \text{T}}_{k-1}\text{H}) = (1-p)^{k-1}p$$



Geometric Random Variables

- Let $\Omega = \{1, 2, 3, \ldots\}$ be the sample space and X be the number of independent trials to achieve the first success.
- Let p stand for the probability of a success in a single trial.
- The probability measure P is called the geometric distribution if

$$P_X(k) = (1-p)^{k-1}p \quad \text{for } k = 1, 2, 3\ldots$$

- Then

$$E[X] = \frac{1}{p} \quad \text{and} \quad Var(X) = \frac{1-p}{p^2}$$

- $P(\texttt{no Heads}) \leq P(\underbrace{\mathsf{T}\ldots\mathsf{T}}_{k}) = (1-p)^k$. As $k \longrightarrow \infty$, $(1-p)^k \longrightarrow 0$

Table of Contents

Continuous Random Variables

Definition

A random variable X on the sample space Ω is said to have a continuous distribution if there exists a real-valued function f such that

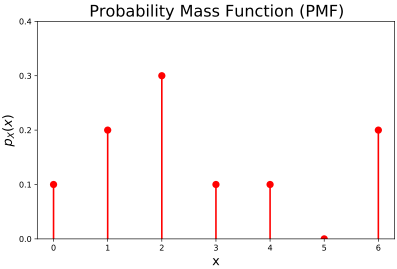
$$f(x) \geq 0,$$
$$\int_{-\infty}^{\infty} f(x)\, dx = 1,$$

and for all real numbers $a < b$:

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx.$$

Then $f: \mathbb{R} \mapsto \mathbb{R}_+$ is called the **probability density function (PDF)** of a **continuous random variable** X.

Probability Density Function (PDF)

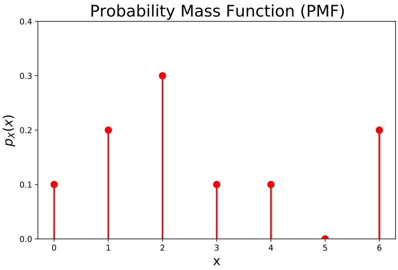


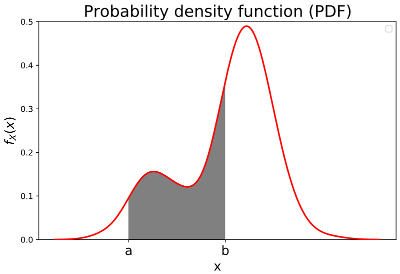
$$P(a \leq X \leq b) = \sum_{x:\ a \leq x \leq b} p_X(x)$$

$$p_X(x) \geq 0$$

$$\sum_x p_X(x) = 1$$

Probability Density Function (PDF)





$$P(a \leq X \leq b) = \sum_{x:\ a \leq x \leq b} p_X(x)$$

$$p_X(x) \geq 0$$

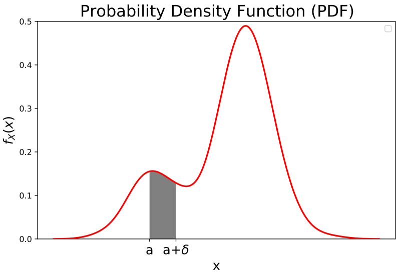
$$\sum_x p_X(x) = 1$$

$$P(a \leq X \leq b) = \int_a^b f_X(x)\ dx$$

$$f_X(x) \geq 0$$

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

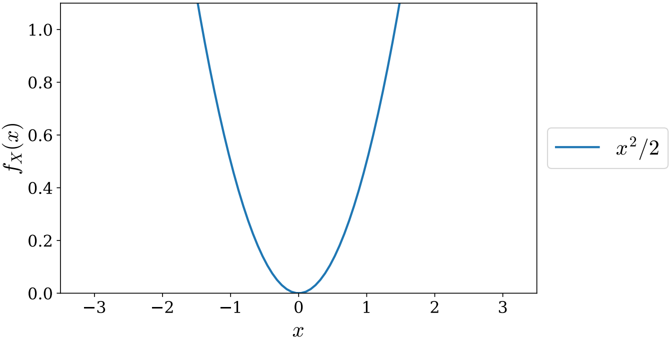
Probability Density Function (PDF)



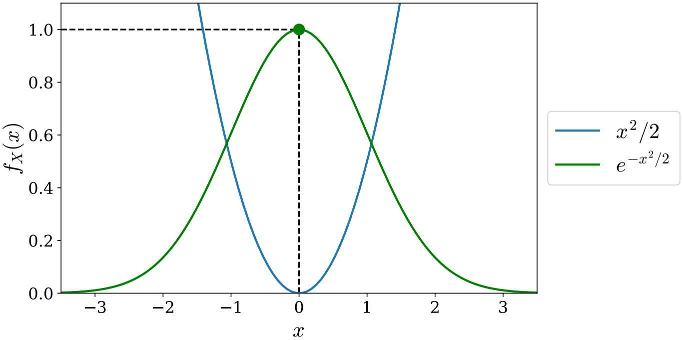
$$P(a \leq X \leq b) = \int_a^b f_X(x)\ dx$$

- $\delta > 0$, small
- $P(a \leq X \leq a + \delta) \approx f_X(a).\delta$
- $P(X = a) = 0$
- Just like, a single point has zero length.
- But, a set of lots of points has a positive length.

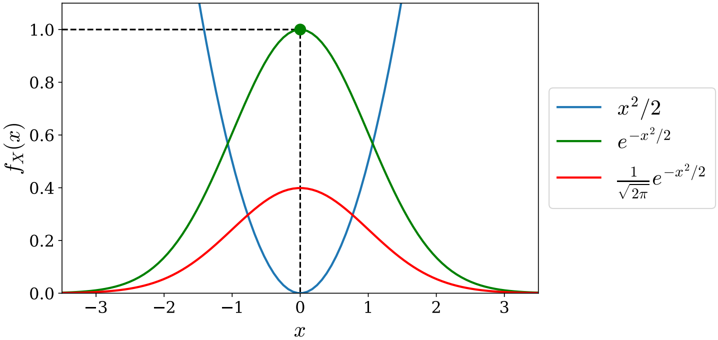
Standard Normal (Gaussian) Random Variable $N(0,1)$



Standard Normal (Gaussian) Random Variable $N(0, 1)$



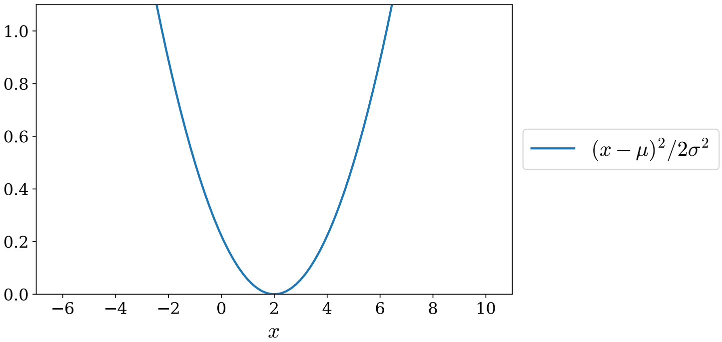
Standard Normal (Gaussian) Random Variable $N(0,1)$



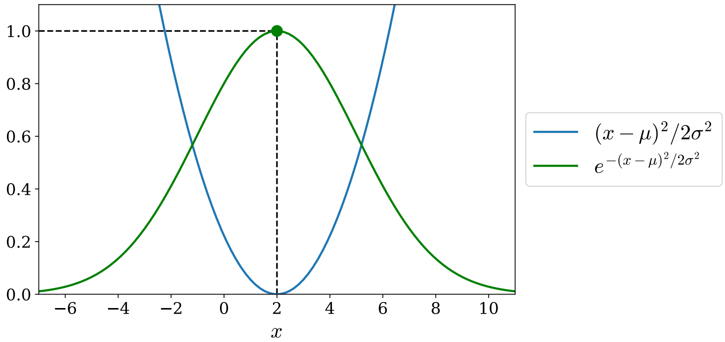
$$\int_{-\infty}^{\infty} e^{-x^2/2}\, dx = \sqrt{2\pi}$$

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

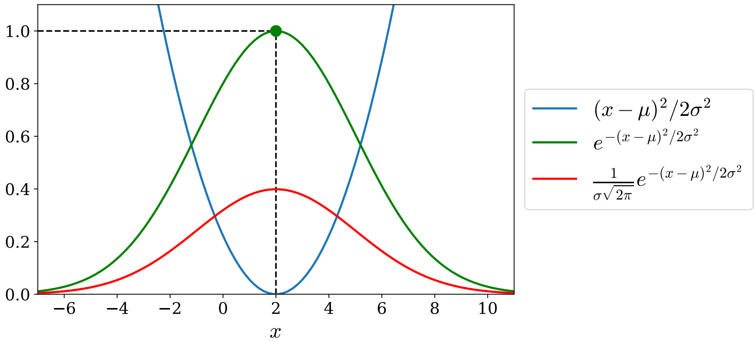
General Normal (Gaussian) Random Variable $N(\mu, \sigma^2)$



General Normal (Gaussian) Random Variable $N(\mu, \sigma^2)$



General Normal (Gaussian) Random Variable $N(\mu, \sigma^2)$

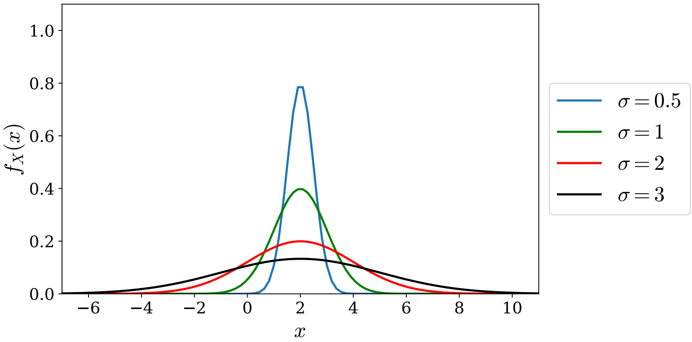


$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

$$\mathrm{E}\left[X\right] = \mu$$

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

General Normal (Gaussian) Random Variable $N(\mu, \sigma^2)$

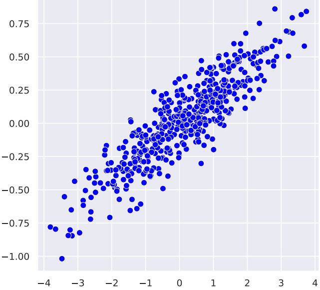


- Smaller σ, narrower PDF.
- Let $Y = aX + b \qquad N \sim N(\mu, \sigma^2)$
- Then, $\mathrm{E}[Y] = aE[X] + b \qquad \mathrm{Var}(Y) = a^2\sigma^2$ (always true)
- But also, $Y \sim N(a\mu + b, a^2\sigma^2)$

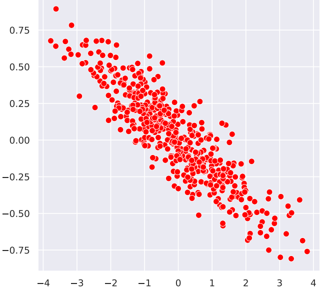
Table of Contents

Covariance

- Consider zero-mean, discrete random variables X and Y
- If they are independent, $E[XY] = E[X]E[Y]$=0.
- But if their joint PDF is as follows.



$$E[XY] > 0$$



$$E[XY] < 0$$

Covariance

- Definition for general case:

$$\texttt{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

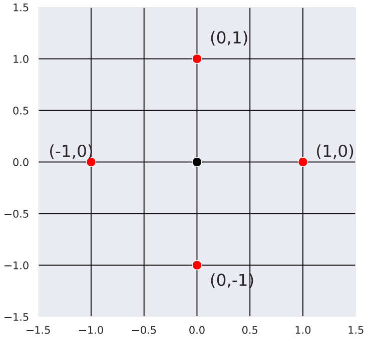
- The deviation of X from its mean value $X - E[X]$.
- The deviation of Y from its mean value $Y - E[Y]$.
- Whether these two deviations tend to have the same sign or not.
- In general, the covariance tells us whether two random variables tend to move together, both being high or both being low, on average.
- Whether they move in same direction or not.
- If the covariance is positive, it indicates that whenever the quantity $X - E[X]$ is positive (X is above its mean), the deviation of Y from its mean will also tend to be positive $Y - E[Y]$.

Covariance

- If X and Y are independent, then

$$\begin{aligned}\texttt{Cov}(X,Y) &= E[(X-E[X])(Y-E[Y])] \\ &= E[(X-E[X])]E[(Y-E[Y])] \\ &= 0\end{aligned}$$

- However, the inverse is not true.



- $E[X]=0, E[Y]=0$
- $XY=0$
- $\texttt{Cov}(X,Y)=0$
- $X=1 \implies Y=0$. Knowing the value of X tells a lot about Y. Therefore, X and Y are dependent.

Covariance properties

- $$\begin{aligned}\texttt{Cov}(X, X) &= E[(X - E[X])^2] = \texttt{Var}(X) \\ &= E[X^2] - (E[X])^2\end{aligned}$$

Covariance properties

- $$\begin{aligned}\texttt{Cov}(X,X) &= E[(X-E[X])^2] = \texttt{Var}(X)\\ &= E[X^2]-(E[X])^2\end{aligned}$$

- $$\begin{aligned}\texttt{Cov}(X,Y) &= E[(X-E[X])(Y-E[Y])]\\ &= E[XY - XE[Y] - E[X]Y + E[X]E[Y]]\\ &= E[XY] - E[XE[Y]] - E[E[X]Y] + E[E[X]E[Y]]\\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y]\\ &= E[XY] - E[X]E[Y]\end{aligned}$$

Covariance properties

- $$\texttt{Cov}(X,X) = E[(X - E[X])^2] = \texttt{Var}(X)$$
  $$= E[X^2] - (E[X])^2$$

- $$\begin{aligned}
\texttt{Cov}(X,Y) &= E[(X - E[X])(Y - E[Y])] \\
&= E[XY - XE[Y] - E[X]Y + E[X]E[Y]] \\
&= E[XY] - E[XE[Y]] - E[E[X]Y] + E[E[X]E[Y]] \\
&= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\
&= E[XY] - E[X]E[Y]
\end{aligned}$$

- Assume 0 means

$$\texttt{Cov}(aX + b, Y) = E[(aX + b)Y] = aE[XY] + bE[Y] = a\texttt{Cov}(X,Y)$$

Joint distribution for multiple random variables

- The **covariance** between two random variables X and Y measures the degree to which X and Y are related

$$\texttt{Cov}[X,Y] \triangleq E[(X-E[X])(Y-E[Y])] = E[XY]-E[X]E[Y]$$

- If $\boldsymbol{x}$ is a D-dimensional random vector, its **covariance matrix** is defined to be the following symmetric, positive semi definite matrix

$$\begin{aligned}\texttt{Cov}[\boldsymbol{x}] &\triangleq E[(\boldsymbol{x}-E[\boldsymbol{x}])(\boldsymbol{x}-E[\boldsymbol{x}])^{\mathsf{T}}] \triangleq \boldsymbol{\Sigma} \\ &= \begin{bmatrix} V[X_1] & \texttt{Cov}[X_1,X_2] & \dots & \texttt{Cov}[X_1,X_D] \\ \texttt{Cov}[X_2,X_1] & V[X_2] & \dots & \texttt{Cov}[X_2,X_D] \\ \vdots & \vdots & \ddots & \vdots \\ \texttt{Cov}[X_D,X_1] & \texttt{Cov}[X_D,X_2] & \dots & V[X_D] \end{bmatrix}\end{aligned}$$

from which we get

$$E[\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}] = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}$$

Multivariate Gaussian (normal) distribution

- The **multivariate normal (MVN)** density is defined

$$\mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right]$$

where $\boldsymbol{\mu} = E[\boldsymbol{y}] \in \mathbb{R}^D$, $\boldsymbol{\Sigma} = \texttt{Cov}[\boldsymbol{y}]$ is the $D \times D$ **covariance matrix**

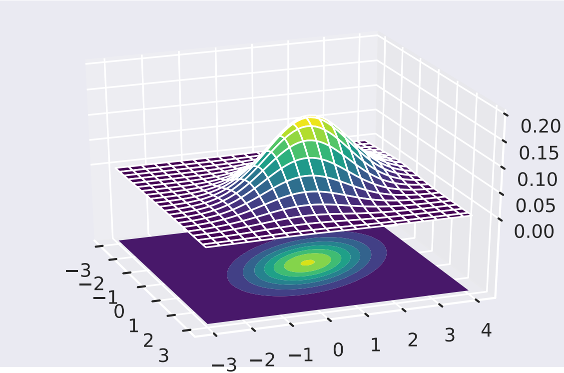
$$\begin{aligned}\texttt{Cov}[\boldsymbol{y}] &\triangleq E[(\boldsymbol{y}-E[\boldsymbol{y}])(\boldsymbol{y}-E[\boldsymbol{y}])^{\mathsf{T}}] \triangleq \boldsymbol{\Sigma} \\ &= \begin{bmatrix} V[Y_1] & \texttt{Cov}[Y_1, Y_2] & \dots & \texttt{Cov}[Y_1, Y_D] \\ \texttt{Cov}[Y_2, Y_1] & V[Y_2] & \dots & \texttt{Cov}[Y_2, Y_D] \\ \vdots & \vdots & \ddots & \vdots \\ \texttt{Cov}[Y_D, Y_1] & \texttt{Cov}[Y_D, Y_2] & \dots & V[X_D] \end{bmatrix}\end{aligned}$$

- A **full covariance matrix** had $D(D+1)/2$ parameters.

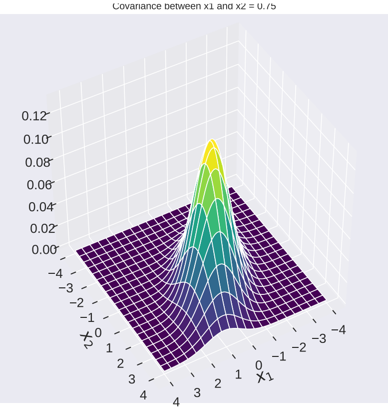
Multivariate Gaussian (normal) distribution

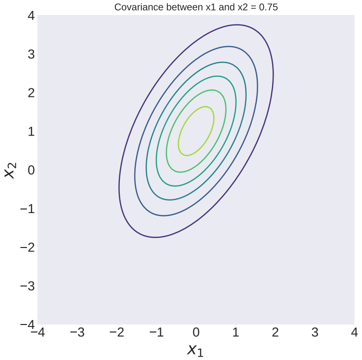
$$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right]$$

where $\boldsymbol{\mu} = E[\boldsymbol{y}] \in \mathbb{R}^D$, $\boldsymbol{\Sigma} = \texttt{Cov}[\boldsymbol{y}]$ is the $D \times D$ **covariance matrix**.



Multivariate Gaussian - Full Covariance Matrix

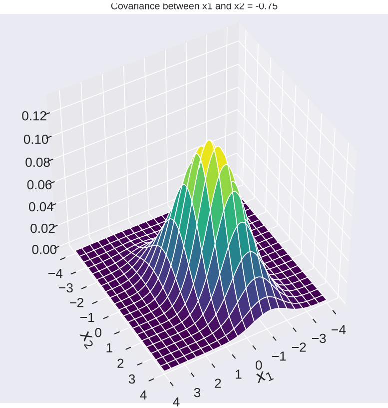


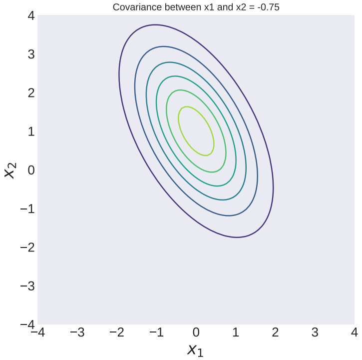


$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.75 \\ 0.75 & 2 \end{bmatrix}$$

A **full covariance matrix** had $D(D+1)/2$ parameters.

Multivariate Gaussian - Full Covariance Matrix

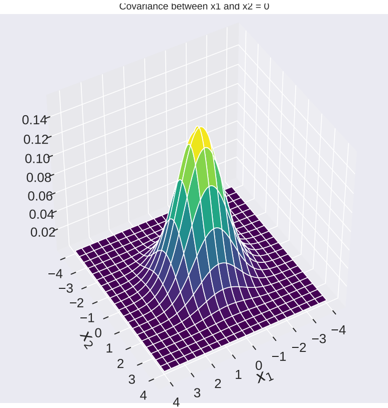


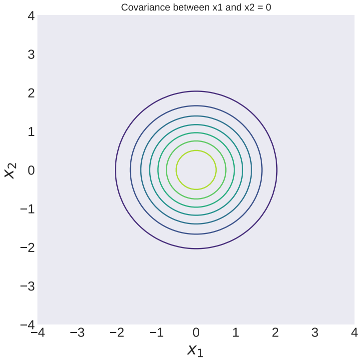


$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & -0.75 \\ -0.75 & 2 \end{bmatrix}$$

A **full covariance matrix** had $D(D+1)/2$ parameters.

Standard Multivariate Gaussian - Spherical Cov. Matrix

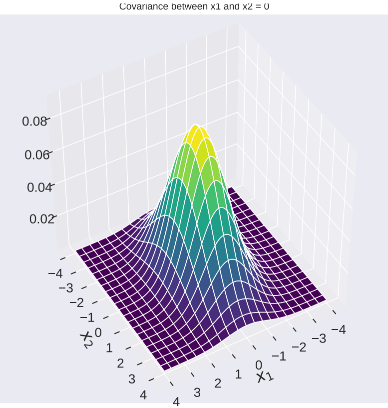


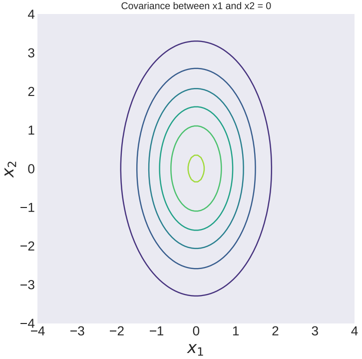


$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

A spherical (isotropic) covariance matrix $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_D$ has one free parameter σ^2.

Uncorrelated Multivariate Gaussian - Diagonal Cov. Matrix





$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}$$

A diagonal covariance matrix has D parameters.

Bivariate Gaussian (normal) distribution

- In 2D, the MVN is known as the **bivariate Gaussian** distribution. Its PDF can be represented as $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{y} \in \mathbb{R}^2, \boldsymbol{\mu} \in \mathbb{R}^2$ and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

where ρ is the **correlation coefficient**

$$\texttt{corr}[Y_1, Y_2] \triangleq \frac{\texttt{Cov}[Y_1, Y_2]}{\sqrt{V[Y_1]V[Y_2]}} = \frac{\sigma_{12}^2}{\sigma_1\sigma_2}$$

Correlation coefficient

- Dimensionless version of covariance:

$$\rho(X,Y) = \texttt{corr}[X,Y] \triangleq E\left[\frac{(X - E[X])}{\sigma_X}\frac{(Y - E[Y])}{\sigma_Y}\right] = \frac{\texttt{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

Correlation coefficient

- Dimensionless version of covariance:

$$\rho(X,Y) = \texttt{corr}[X,Y] \triangleq E\left[\frac{(X-E[X])}{\sigma_X}\frac{(Y-E[Y])}{\sigma_Y}\right] = \frac{\texttt{Cov}(X,Y)}{\sigma_X\sigma_Y}$$

- $1 \le \rho \le 1$: measure of the degree of "association" between X and Y.

Correlation coefficient

- Dimensionless version of covariance:

$$\rho(X,Y) = \texttt{corr}[X,Y] \triangleq E\left[\frac{(X-E[X])}{\sigma_X}\frac{(Y-E[Y])}{\sigma_Y}\right] = \frac{\texttt{Cov}(X,Y)}{\sigma_X\sigma_Y}$$

- $1 \le \rho \le 1$: measure of the degree of "association" between X and Y.
- Independent $\Longrightarrow \rho = 0$: "uncorrelated" (the converse is not true).

Correlation coefficient

- Dimensionless version of covariance:

$$\rho(X,Y) = \mathtt{corr}[X,Y] \triangleq E\left[\frac{(X-E[X])}{\sigma_X}\frac{(Y-E[Y])}{\sigma_Y}\right] = \frac{\mathtt{Cov}(X,Y)}{\sigma_X\sigma_Y}$$

- $1 \le \rho \le 1$: measure of the degree of "association" between X and Y.
- Independent $\Longrightarrow \rho = 0$: "uncorrelated" (the converse is not true).
- $\rho(X,X) = \frac{\mathtt{Var}(X)}{\sigma_X^2} = 1$.

Correlation coefficient

- Dimensionless version of covariance:

$$\rho(X, Y) = \texttt{corr}[X, Y] \triangleq E\left[\frac{(X - E[X])}{\sigma_X}\frac{(Y - E[Y])}{\sigma_Y}\right] = \frac{\texttt{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

- $1 \leq \rho \leq 1$: measure of the degree of "association" between X and Y.
- Independent $\Longrightarrow \rho = 0$: "uncorrelated" (the converse is not true).
- $\rho(X, X) = \frac{\texttt{Var}(X)}{\sigma_X^2} = 1$.
- $|\rho| = 1 \Longleftrightarrow (X - E[X]) = c(Y - E[Y])$: linearly related, deterministic relation between the two rv.

Correlation coefficient

- Dimensionless version of covariance:

$$\rho(X,Y) = \mathtt{corr}[X,Y] \triangleq E\left[\frac{(X-E[X])}{\sigma_X}\frac{(Y-E[Y])}{\sigma_Y}\right] = \frac{\mathtt{Cov}(X,Y)}{\sigma_X\sigma_Y}$$

- $1 \leq \rho \leq 1$: measure of the degree of "association" between X and Y.
- Independent $\implies \rho = 0$: "uncorrelated" (the converse is not true).
- $\rho(X,X) = \frac{\mathtt{Var}(X)}{\sigma_X^2} = 1$.
- $|\rho| = 1 \iff (X - E[X]) = c(Y - E[Y])$: linearly related, deterministic relation between the two rv.
- We have $\mathtt{Cov}(aX+b,Y) = a\mathtt{Cov}(X,Y)$. Therefore,

$$\rho(aX+b,Y) = \frac{a\mathtt{Cov}(X,Y)}{|a|\sigma_X\sigma_Y} = \mathtt{sign}(a)\cdot\rho(X,Y)$$

$\implies$ Changing the unit of X does not change correlation coefficient between X and Y.

Table of Contents

Example

- A bag contains 3 balls, each ball is either red or blue.
- The number of blue balls θ can be $0, 1, 2, 3$.

Example

- A bag contains 3 balls, each ball is either red or blue.
- The number of blue balls θ can be $0, 1, 2, 3$.
- Choose 4 balls randomly **with replacement**.

Example

- A bag contains 3 balls, each ball is either red or blue.
- The number of blue balls θ can be $0, 1, 2, 3$.
- Choose 4 balls randomly **with replacement**.
- Random variables X_1, X_2, X_3, X_4 are defined as

$$X_i = \begin{cases} 1, & \text{if the } i\text{-th chosen ball is blue} \\ 0, & \text{if the } i\text{-th chosen ball is red} \end{cases}$$

Example

- A bag contains 3 balls, each ball is either red or blue.
- The number of blue balls θ can be $0, 1, 2, 3$.
- Choose 4 balls randomly **with replacement**.
- Random variables X_1, X_2, X_3, X_4 are defined as

$$X_i = \begin{cases} 1, & \text{if the } i\text{-th chosen ball is blue} \\ 0, & \text{if the } i\text{-th chosen ball is red} \end{cases}$$

- After doing the experiment, the following values for X_i's are observed: $x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 1$.

Example

- A bag contains 3 balls, each ball is either red or blue.
- The number of blue balls θ can be $0, 1, 2, 3$.
- Choose 4 balls randomly **with replacement**.
- Random variables X_1, X_2, X_3, X_4 are defined as

$$X_i = \begin{cases} 1, & \text{if the } i\text{-th chosen ball is blue} \\ 0, & \text{if the } i\text{-th chosen ball is red} \end{cases}$$

- After doing the experiment, the following values for X_i's are observed: $x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 1$.
- Note that X_i's are i.i.d. (independent and identically distributed) and $X_i \sim Bernoulli(\frac{\theta}{3})$. For which value of θ is the probability of the observed sample is the largest?

Example

$$P_{X_i}(x) = \begin{cases} \frac{\theta}{3}, & \text{for } x = 1 \\ 1 - \frac{\theta}{3}, & \text{for } x = 0 \end{cases}$$

Example

$$P_{X_i}(x) = \begin{cases} \frac{\theta}{3}, & \text{for } x = 1 \\ 1 - \frac{\theta}{3}, & \text{for } x = 0 \end{cases}$$

X_i's are independent, the joint PMF of X_1, X_2, X_3, X_4 can be written

$$P_{X_1X_2X_3X_4}(x_1, x_2, x_3, x_4) = P_{X_1}(x_1)P_{X_2}(x_2)P_{X_3}(x_3)P_{X_4}(x_4)$$

$$P_{X_1X_2X_3X_4}(1, 0, 1, 1) = \frac{\theta}{3} \cdot \left(1 - \frac{\theta}{3}\right) \cdot \frac{\theta}{3} \cdot \frac{\theta}{3} = \left(\frac{\theta}{3}\right)^3 \left(1 - \frac{\theta}{3}\right)$$

Example

$$P_{X_i}(x) = \begin{cases} \frac{\theta}{3}, & \text{for } x = 1 \\ 1 - \frac{\theta}{3}, & \text{for } x = 0 \end{cases}$$

X_i's are independent, the joint PMF of X_1, X_2, X_3, X_4 can be written

$$P_{X_1X_2X_3X_4}(x_1, x_2, x_3, x_4) = P_{X_1}(x_1)P_{X_2}(x_2)P_{X_3}(x_3)P_{X_4}(x_4)$$

$$P_{X_1X_2X_3X_4}(1, 0, 1, 1) = \frac{\theta}{3} \cdot \left(1 - \frac{\theta}{3}\right) \cdot \frac{\theta}{3} \cdot \frac{\theta}{3} = \left(\frac{\theta}{3}\right)^3 \left(1 - \frac{\theta}{3}\right)$$

θ	$P_{X_1X_2X_3X_4}(1, 0, 1, 1; \theta)$
0	0
1	0.0247
2	0.0988
3	0

Example

$$P_{X_i}(x) = \begin{cases} \frac{\theta}{3}, & \text{for } x = 1 \\ 1 - \frac{\theta}{3}, & \text{for } x = 0 \end{cases}$$

X_i's are independent, the joint PMF of X_1, X_2, X_3, X_4 can be written

$$P_{X_1X_2X_3X_4}(x_1, x_2, x_3, x_4) = P_{X_1}(x_1)P_{X_2}(x_2)P_{X_3}(x_3)P_{X_4}(x_4)$$

$$P_{X_1X_2X_3X_4}(1,0,1,1) = \frac{\theta}{3} \cdot \left(1 - \frac{\theta}{3}\right) \cdot \frac{\theta}{3} \cdot \frac{\theta}{3} = \left(\frac{\theta}{3}\right)^3 \left(1 - \frac{\theta}{3}\right)$$

θ	$P_{X_1X_2X_3X_4}(1,0,1,1;\theta)$
0	0
1	0.0247
2	0.0988
3	0

The observed data is most likely to occur for $\theta = 2$.
We may choose $\hat{\theta} = 2$ as our estimate of θ.

Maximum Likelihood Estimation (MLE)

Definition

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with a parameter θ.
Given that we have observed $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, a maximum likelihood estimate of θ, denoted as $\hat{\theta}_{ML}$, is a value of θ that maximizes the likelihood function

$$L(x_1, x_2, \ldots, x_n; \theta)$$

A maximum likelihood estimator (MLE) of the parameter θ, denoted as $\hat{\Theta}_{ML}$, is a random variable $\hat{\Theta}_{ML} = \hat{\Theta}(X_1, X_2, \ldots, X_n)$ whose values $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ is given by $\hat{\theta}_{ML}$.