Introduction to Machine Learning and Data Science

2. Review of Probability and Statistics

Pierre Vandenhove
Course material by Souhaib Ben Taieb

Université de Mons

UMONS
Université de Mons

Faculté
des Sciences

References

- **Introduction to Probability for Data Science**, Stanley H. Chan. [Link] (Book, slides and videos)
- **Probability Theory Review for Machine Learning (CS229)**, Samuel Ieong. [Link]
- **All of Statistics**, Larry Wasserman. [Link]

Outline

Sample space and events

- When we speak about probability, we often refer to the probability of **an event of uncertain nature** taking place.
- We first need to clarify what the **possible events** to which we want to attach probability are.
- We often conduct an experiment, i.e., take some measurements of a **random (stochastic) process**.
- Our measurements take values in some set Ω, the **sample space** (or the outcome space), which defines *all possible outcomes* of our measurements.

Sample space and events

- We toss one coin. Two outcomes: *heads* (H) or *tails* (T).
  - $\Omega = \{H, T\}$.
- We toss two coins:
  - $\Omega = \{HH, HT, TH, TT\}$.
- We measure the reaction time to some stimulus:
  - $\Omega = (0, \infty)$.

Sample space and events

An **event** A is a subset of Ω ($A \subseteq \Omega$), i.e., it is a subset of possible outcomes of our experiment. We say that an event A **occurs** if the outcome of our experiment belongs to the set A.

- Let $\Omega = \{HH, HT, TH, TT\}$, and consider the following events:
  - $A_1 = \{HH, TH, TT\}$ and $A_2 = \{TH, TT\}$.

  We observe $\omega = HH$. Which events did occur?

- Let $\Omega = (0, \infty)$, and consider the following events:
  - $A_1 = (3, 6)$, $A_2 = (1, 2)$ and $A_3 = (2, 8)$.

  We observe $\omega = 4$. Which events did occur?

Probability space

A **probability space** is defined by the triple $(\Omega, \mathcal{F}, \mathbb{P})$ where

- Ω is the **sample space**,
- $\mathcal{F}$ is the **space of events** (or event space); we abusively[1] assume here that $\mathcal{F} = 2^{\Omega}$,
- $\mathbb{P}$ is the **probability measure/distribution** that maps an event $A \in \mathcal{F}$ to a real value between 0 and 1.

[1] 2^S is the set of all subsets of S, including S and the empty set $\emptyset$. Note that $\mathcal{F} = 2^{\Omega}$ is not mathematically sound in general, but it is sufficient for our practical purposes.

Probability axioms

A **probability distribution** is a mapping from events to real numbers that satisfy certain **axioms**:

1. *Non-negativity*:

$$\forall A \in \mathcal{F}, \mathbb{P}(A) \geq 0.$$

2. *Unity of* Ω:

$$\mathbb{P}(\Omega) = 1.$$

3. σ*-additivity*. For all *disjoint* events $A_1, A_2, \ldots \in \mathcal{F}$ (i.e., for all $i \neq j$, $A_i \cap A_j = \emptyset$), we have that

$$\mathbb{P}\left(\bigcup_{i=1}^{+\infty} A_i\right) = \sum_{i=1}^{+\infty} \mathbb{P}(A_i).$$

Probability properties

Using set theory and the probability axioms, we can show several useful and intuitive properties of probability distributions.

- $\mathbb{P}(\emptyset) = 0$.
- For all $A, B \in \mathcal{F}$, $A \subset B \implies \mathbb{P}(A) \leq \mathbb{P}(B)$.
- For all $A \in \mathcal{F}$, $0 \leq \mathbb{P}(A) \leq 1$.
- For all $A \in \mathcal{F}$, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
- For all $A, B \in \mathcal{F}$, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

All of these properties can be understood via a **Venn diagram**.

Probability properties

- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

$$\begin{aligned}
&\mathbb{P}(\Omega) = 1 \quad (\textbf{Axiom } 2) \\
&\iff \mathbb{P}(A \cup A^c) = 1, \quad \forall A \subseteq \Omega \\
&\iff \mathbb{P}(A^c) + \mathbb{P}(A) = 1 \quad (\textbf{Axiom } 3) \\
&\iff \mathbb{P}(A^c) = 1 - \mathbb{P}(A).
\end{aligned}$$

- $A \subseteq B \implies \mathbb{P}(A) \leq \mathbb{P}(B)$.

$$\begin{aligned}
&A \subseteq B \\
&\implies B = A \cup (B \backslash A) \\
&\implies \mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \backslash A) \quad (\textbf{Axiom } 3, \text{ using that } A \cap (B \backslash A) = \emptyset) \\
&\implies \mathbb{P}(B) \geq \mathbb{P}(A) \quad (\textbf{Axiom } 1).
\end{aligned}$$

Probability of an event (discrete case)

- The probability of any event $A = \{\omega_1, \omega_2, \ldots, \omega_k\}$ ($\omega \in \Omega$) is the sum of the probabilities of its elements:

$$\mathbb{P}(A) = \mathbb{P}(\{\omega_1, \omega_2, \ldots, \omega_k\}) = \sum_{i=1}^{k} \mathbb{P}(\{\omega_i\}).$$

- If Ω consists of n equally likely outcomes (i.e., a uniform distribution), then the probability of any event A is

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{n}.$$

- Suppose we toss a **fair** dice twice. The sample space is $\Omega = \{(t_1, t_2) : t_1, t_2 = 1, 2, \ldots, 6\}$. Let A be the event that the sum of two tosses being < 5. What is $\mathbb{P}(A)$?

Conditional probability

If $\mathbb{P}(B) > 0$, the **conditional probability** of A *given* B is

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Note that $\mathbb{P}(A \mid B) \neq \mathbb{P}(B \mid A)$ (in general).

The **chain rule** can be obtained by rewriting the above expression as follows:

$$\mathbb{P}(A \cap B) = \mathbb{P}(B)\mathbb{P}(A \mid B) = \mathbb{P}(A)\mathbb{P}(B \mid A).$$

More generally, we have

$$\mathbb{P}(A_1 \cap A_2 \cap A_3 \dots) = \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1)\mathbb{P}(A_3 \mid A_1 \cap A_2) \dots$$

Independence of events

Two events A and B are **independent** if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

A finite set of events $(A_j)_{j \in J}$ are **mutually independent** if for all subsets $I \subseteq J$,

$$\mathbb{P}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbb{P}(A_i).$$

Conditional probability gives another interpretation of independence: A and B are independent if the *unconditional probability* is the **same** as the *conditional probability*.

When combined with other properties of probability, independence can often **simplify the calculation** of the probability of certain events.

Example

Consider a **fair** coin. What is the probability of *at least one head in the first 10 (**independent**) tosses*?

Let A be the event "at least one head in 10 tosses". Then, A^c is the event "No heads in 10 tosses" (all 10 tosses being tails).

We have

$$\begin{aligned}
\mathbb{P}(A) &= 1 - \mathbb{P}(A^c) \\
&= 1 - \mathbb{P}(\underbrace{T_1 \cap T_2 \cap \cdots \cap T_{10}}) \\
&= 1 - \prod_{i=1}^{10} \mathbb{P}(T_i) \qquad \text{(independent tosses)} \\
&= 1 - \left(\frac{1}{2}\right)^{10} \qquad \text{(fair coin)}.
\end{aligned}$$

Exercise

Consider tossing a **fair** dice. Let A be the event that the result is an odd number, and $B = \{1, 2, 3\}$.

- Compute $\mathbb{P}(A \mid B)$.
- Compute $\mathbb{P}(A)$.
- Are A and B independent?

Law of total probability

Let $A_1, A_2, \ldots, A_n \in \mathcal{F}$ be a *partition* of Ω. Then, for any $B \in \mathcal{F}$, we have that

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B \cap A_i) = \sum_{i=1}^{n} \mathbb{P}(B \mid A_i)\mathbb{P}(A_i).$$

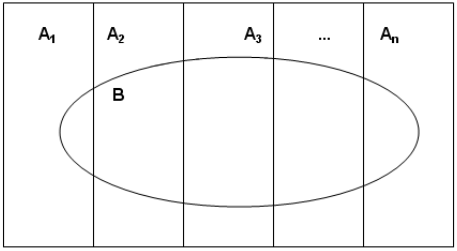


Image source: https://mathwiki.cs.ut.ee/probability/04_total_probability

Law of total probability: proof

The **law of total probability** is a combination of additivity and conditional probability. We have

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}((B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_k)) \\ &= \sum_{i=1}^{n} \mathbb{P}(B \cap A_i) \qquad \text{(since events } (B \cap A_i)_i \text{ are disjoint)} \\ &= \sum_{i=1}^{n} \mathbb{P}(B \mid A_i)\mathbb{P}(A_i).\end{aligned}$$

Bayes' Rule

Let $A, B \in \mathcal{F}$ be two events with positive probability. **Bayes' rule** states that

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Roughly, Bayes' rule allows us to calculate $\mathbb{P}(A \mid B)$ from $\mathbb{P}(B \mid A)$. This is useful when $\mathbb{P}(A \mid B)$ is *not obvious to calculate* but $\mathbb{P}(B \mid A)$ is easy to obtain.

Bayes' rule follows from the definition of conditional probability (two uses):

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Example

Suppose there are exactly three types of emails:

- $A_1 =$ Spam,
- $A_2 =$ Low Priority,
- $A_3 =$ High Priority.

Based on previous experience, we have

$$\mathbb{P}(A_1) = 0.85,\ \mathbb{P}(A_2) = 0.1,\ \mathbb{P}(A_3) = 0.05.$$

Let B be the event that an email contains the word "free", then

$$\mathbb{P}(B \mid A_1) = 0.9,\ \mathbb{P}(B \mid A_2) = 0.1,\ \mathbb{P}(B \mid A_3) = 0.1.$$

When we receive an email containing the word "free", **what is the probability that it is a spam**?

Outline

Random variables

Often we are interested in dealing with *summaries of experiments* rather than the actual *outcome*.

For instance, suppose we toss a coin three times. But we may only be interested in a "summary", such as the number of heads. We have

$$\begin{array}{c} \Omega \\ \downarrow \\ \mathbb{R} \end{array} = \{\underbrace{HHH}_{\downarrow 3}, \underbrace{HHT}_{\downarrow 2}, \underbrace{HTH}_{\downarrow 2}, \underbrace{THH}_{\downarrow 2}, \underbrace{TTH}_{\downarrow 1}, \underbrace{THT}_{\downarrow 1}, \underbrace{HTT}_{\downarrow 1}, \underbrace{TTT}_{\downarrow 0}\}.$$

These summary statistics are called **random variables**. Specifically, a random variable X is a (measurable) *function* from the sample space Ω to the real numbers:

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

Random variables

A random variable (r.v.) can be seen as a **mapping** between

- a distribution on Ω,

to

- a distribution on the reals (or the *range* $\mathcal{X} \subseteq \mathbb{R}$ of the r.v.).

Formally, we have that for some subset $S \subseteq \mathcal{X}$,

$$\mathbb{P}_X(X \in S) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in S\}).$$

Random variables

$$\begin{array}{c} \Omega = \{\underbrace{HHH}_{3}, \underbrace{HHT}_{2}, \underbrace{HTH}_{2}, \underbrace{THH}_{2}, \underbrace{TTH}_{1}, \underbrace{THT}_{1}, \underbrace{HTT}_{1}, \underbrace{TTT}_{0}\}. \\ \downarrow \\ \mathbb{R} \end{array}$$

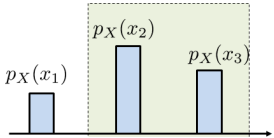
For the previous example, we have

$$\mathbb{P}_X(X = 0) = 1/8, \quad \mathbb{P}_X(X = 1) = 3/8,$$

$$\mathbb{P}_X(X = 2) = 3/8, \quad \mathbb{P}_X(X = 3) = 1/8.$$

Outline

Probability mass function



A random variable is **discrete** if its image $\mathcal{X}$ is a countable set.

The **probability mass function** (PMF) of a discrete random variable X is a function which specifies the *probability* of obtaining a number x. We denote the PMF as

$$p_X(x) = \mathbb{P}(X = x).$$

A function p_X is a PMF if and only if

1. $\forall x \in \mathcal{X}, p_X(x) \geq 0$,
2. $\sum_{x \in \mathcal{X}} p_X(x) = 1$.

What is the PMF of the previous coin flip example?

Some important discrete distributions

- The discrete **uniform** distribution on K categories. The PMF of $X \in \{C_1, C_2, \ldots, C_K\}$ is given by
$$p_X(x) = \frac{1}{k}, \qquad \forall x \in \{C_1, C_2, \ldots, C_K\}.$$
- The **Bernoulli** distribution with parameter $p \in [0, 1]$. The PMF of $X \in \{0, 1\}$ is given by
$$p_X(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases} = p^x(1-p)^{1-x}.$$
It can represent a coin toss when the coin has bias p, where 1 denotes *heads* and 0 denotes *tails*.
- Other important distributions: Binomial, Geometric, Poisson, etc.
- The symbol "$\sim$" denotes "distributed as", i.e., $X \sim \text{Ber}(p)$ means that X has a Bernoulli distribution with parameter p.

Expectation

The **expectation** of a random variable X is

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \; p_X(x).$$

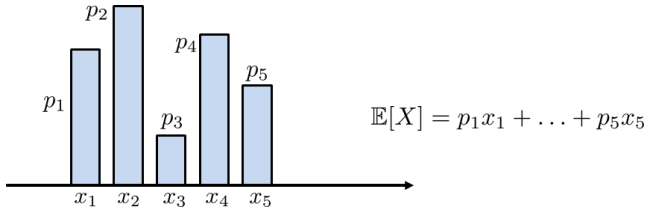


Image source: Introduction to Probability for Data Science, Stanley H. Chan.

Expectation and its properties

For any function $g\colon \mathbb{R} \to \mathbb{R}$, we have

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x)\ p_X(x).$$

For any functions $g\colon \mathbb{R} \to \mathbb{R}$ and $h\colon \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[g(X) + h(X)] = \mathbb{E}[g(X)] + \mathbb{E}[h(X)].$$

For any constant $c \in \mathbb{R}$,

$$\mathbb{E}[cX] = c\mathbb{E}[X].$$

For any constant $c \in \mathbb{R}$,

$$\mathbb{E}[X + c] = \mathbb{E}[X] + c.$$

Moments and variance

The k-th **moment** of a random variable X is

$$\mathbb{E}[X^k] = \sum_{x \in \mathcal{X}} x^k \; p_X(x).$$

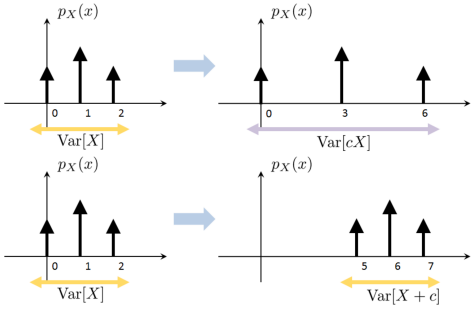
The **variance** of a random variable X is

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

The **standard deviation** of X, often denoted σ, is $\sqrt{\mathrm{Var}(X)}$.

Useful properties of the variance include:

- $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$,
- $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$,
- $\mathrm{Var}(X + c) = \mathrm{Var}(X)$.



Outline

Probability density function

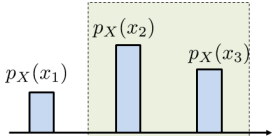


Image source: Introduction to Probability for Data Science, Stanley H. Chan.

A random variable is **continuous** if its image $\mathcal{X}$ is an uncountable set.

The **probability density function** (PDF) of a continuous random variable X is a function $f_X : \mathcal{X} \to [0, +\infty)$, when integrated over an interval $[a, b]$, yields the probability of obtaining $\{x \in \mathcal{X} \mid a \leq x \leq b\}$:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)dx.$$

A PDF must satisfy the following properties:

$$(1) \quad \forall x \in \mathcal{X}, f_X(x) \geq 0, \qquad (2) \quad \int_{\mathcal{X}} f_X(x)dx = 1.$$

Note: $f_X(x)$ is not the probability that $X = x$ (e.g., we can have $f_X(x) > 1$).

Some important continuous distributions

- The continuous **uniform** distribution on interval $[a, b]$. The PDF is given by

$$f_X(x) = \frac{1}{b-a} \quad (x \in [a, b]).$$

  We write $X \sim \mathcal{U}[a, b]$.
- The **Gaussian** or **normal** distribution. With a location (mean) μ and scale (standard deviation) σ, the PDF is given by

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

  We write $X \sim \mathcal{N}(\mu, \sigma^2)$.

Expectation and its properties

The **expectation** of a continuous random variable X is given by

$$\mathbb{E}[X] = \int_{\mathcal{X}} x f_X(x) dx.$$

For any function $g\colon \mathcal{X} \to \mathbb{R}$, we have

$$\mathbb{E}[g(X)] = \int_{\mathcal{X}} g(x) f_X(x) dx.$$

Let $I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$. Then, we have

$$\mathbb{E}[I_A(X)] = \int_{\mathcal{X}} I_A(x) f_X(x) dx = \int_A f_X(x) dx = \mathbb{P}(X \in A).$$

Moments and variance

The k-th **moment** of a continuous random variable X is

$$\mathbb{E}[X^k] = \int_{\mathcal{X}} x^k \; f_X(x) dx.$$

The **variance** of a continuous random variable X is

$$\text{Var}(X) = \mathbb{E}[(X - \mu_X)^2] = \int_{\mathcal{X}} (x - \mu_X)^2 \; f_X(x) dx,$$

where $\mu_X = \mathbb{E}[X]$. The **standard deviation** of X is $\sqrt{\text{Var}(X)}$.

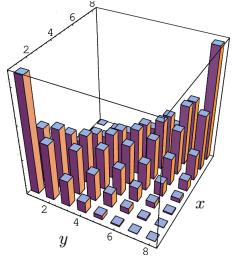
The properties of variance introduced previously still hold in this case.

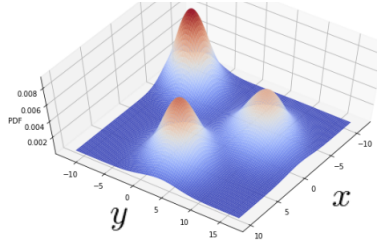
Exercise. If $X \sim \mathcal{U}[a, b]$, what is $\mathbb{E}[X]$ and $\text{Var}(X)$?

Outline

More than one random variable?

- **Multivariate** random variables or **random vectors** are ubiquitous in modern data analysis.
- The uncertainty in the random vector is characterized by a **joint** PMF or PDF.
- We first focus in this section on **bivariate** random variables.





Important concepts

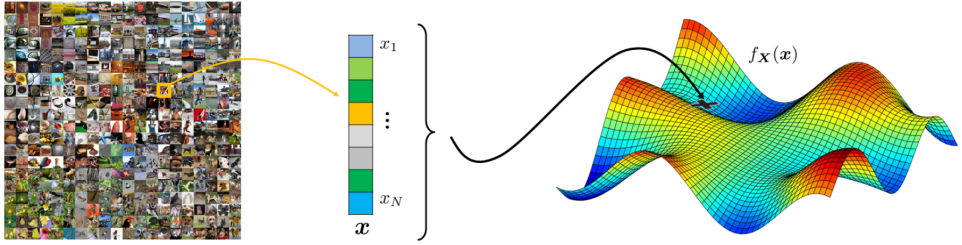
- Joint distribution
- Marginal distribution
- Independence
- Joint expectations
- Covariance and correlation
- Conditional distribution
- Conditional expectations

What are joint distributions?

One random variable is characterized by a PDF f_X (or PMF p_X), whose input x is single-dimensional.

We can characterize N random variables $X_1, \ldots, X_N$ using a PDF $f_{X_1,\ldots,X_N}$ that takes as an input an N-dimensional vector $\mathbf{x} = (x_1, \ldots, x_N)$.

Why do we need more than one dimension?



For instance, the ImageNet dataset consists of images, which can be represented as a 3-dimensional array of pixels of size $224 \times 224 \times 3$. Each pixel is a random variable, and the joint distribution of all pixels characterizes the distribution over all images.

Joint PMF

Let X and Y be two **discrete** random variables. The **joint PMF** of X and Y is defined as

$$p_{X,Y}(x,y) = \mathbb{P}(X = x \text{ and } Y = y).$$

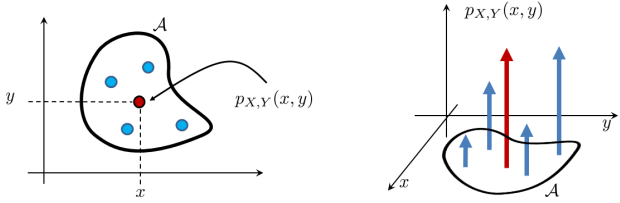


Image source: Introduction to Probability for Data Science, Stanley H. Chan.

For any $A \subseteq \mathcal{X} \times \mathcal{Y}$, we have

$$\mathbb{P}((X,Y) \in A) = \sum_{(x,y)\in A} p_{X,Y}(x,y).$$

Example

Let X be a coin flip, Y be a dice. Find the **joint PMF**.

The *sample space* of X is $\{0, 1\}$. The *sample space* of Y is $\{1, 2, 3, 4, 5, 6\}$. The joint PMF is

			Y			
	1	2	3	4	5	6
X = 0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$
X = 1	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$

Equivalently, we have

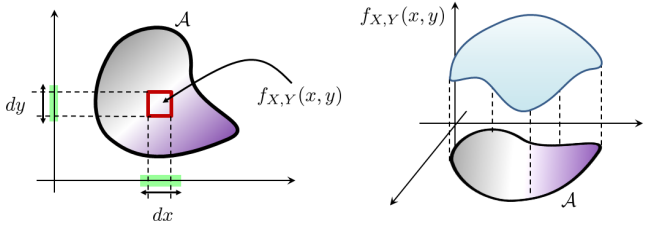
$$p_{X,Y}(x, y) = \frac{1}{12} \text{ for all } x \in \{0, 1\}, y \in \{1, 2, 3, 4, 5, 6\}.$$

Joint PDF

Let X and Y be two **continuous** random variables. The **joint PDF** of X and Y is a function $f_{X,Y}(x,y)$ that can be integrated to yield a probability:

$$\mathbb{P}((X,Y) \in A) = \int_A f_{X,Y}(x,y)dx\ dy,$$

for any (measurable) $A \subseteq \mathcal{X} \times \mathcal{Y}$.



Exercise. Consider a uniform joint PDF $f_{X,Y}(x,y)$ defined on the square $[0,2]^2$ with $f_{X,Y}(x,y) = 1/4$ for all $(x,y) \in [0,2]^2$. Find (i) $\mathbb{P}([a,b] \times [c,d])$ for $0 \leq a \leq b \leq 2$ and $0 \leq c \leq d \leq 2$, and (ii) $\mathbb{P}(X + Y \leq 2)$.

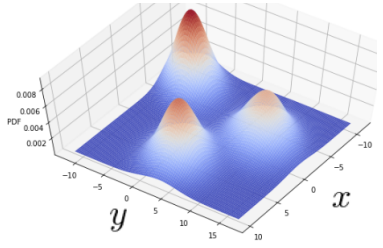
Marginal distribution

The **marginal PMFs** of discrete r.v. are defined as

$$p_X(x) = \sum_{y \in \mathcal{Y}} p_{X,Y}(x, y) \text{ and } p_Y(y) = \sum_{x \in \mathcal{X}} p_{X,Y}(x, y),$$

and the **marginal PDFs** of continuous r.v. are defined as

$$f_X(x) = \int_{\mathcal{Y}} f_{X,Y}(x, y) dy \text{ and } f_Y(y) = \int_{\mathcal{X}} f_{X,Y}(x, y) dx.$$



Marginal PDFs: exercise

Assume the joint PDF of X and Y is given by

$$f_{X,Y}(x,y) = \begin{cases} x + \frac{3}{2}y^2 & \text{if } 0 \leq x \leq 1, 0 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Find the marginal PDFs of X and Y.

Independence

Two random variables X and Y are **independent** if for all x, y, we have

$$p_{X,Y}(x,y) = p_X(x)p_Y(y) \text{ (discrete case) or } f_{X,Y}(x,y) = f_X(x)f_Y(y) \text{ (continuous case).}$$

A sequence of random variables $X_1, \ldots, X_N$ are **independent** if their joint PDF (or joint PMF) can be **factorized**:

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{j=1}^{n} f_{X_j}(x_j).$$

Exercise. Consider two random variables X and Y with joint PDF

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}.$$

Show that X and Y are independent.

Independent and Identically Distributed (i.i.d.)

A collection of random variables $X_1, \ldots, X_N$ are called **independent and identically distributed (i.i.d.)** if

1. all $X_1, \ldots, X_N$ are **independent**, and
2. all $X_1, \ldots, X_N$ have the **same distribution**.

Example: N distinct coin tosses ($X_1, \ldots, X_N \sim \text{Ber}(p)$).

Example: Let θ be a deterministic number that was sent through a noisy channel. We model the noise as an additive Gaussian random variable with mean 0 and variance σ^2. Supposing we have observed measurements $X_i = \theta + W_i$, for $i = 1, \ldots, N$, where $W_i \sim \mathcal{N}(0, \sigma^2)$. Then, the joint PDF of $X_1, \ldots, X_N$ is

$$\begin{aligned} f_{X_1,\ldots,X_N}(x_1, \ldots, x_N) &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\theta)^2}{2\sigma^2}} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\theta)^2}. \end{aligned}$$

Multivariate expectations

Recall that the expectation of a discrete random variable X is given by

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \; p_X(x).$$

How about the expectation for two variables?

Let X and Y be two *discrete* random variables. For any function $g \colon \mathbb{R}^2 \to \mathbb{R}$, the **joint expectation** is

$$\mathbb{E}[g(X, Y)] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} g(x, y) \; p_{X,Y}(x, y).$$

If X and Y are *continuous*, we have

$$\mathbb{E}[g(X, Y)] = \int_{\mathcal{Y}} \int_{\mathcal{X}} g(x, y) \; f_{X,Y}(x, y) \; dx \; dy.$$

Joint expectation

Let $g(X, Y) = XY$. If X and Y are **discrete**, the **joint expectation** is

$$\mathbb{E}[XY] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} xy \; p_{X,Y}(x, y).$$

If X and Y are **continuous**, we have

$$\mathbb{E}[XY] = \int_{\mathcal{Y}} \int_{\mathcal{X}} xy \; f_{X,Y}(x, y) dx dy.$$

Remark. Why do we consider $\mathbb{E}[XY]$ in particular, and not, e.g., $\mathbb{E}[X + Y]$? It has good properties (it defines an *inner product* in the space of random variables) and is related to the notion of *covariance*.

Covariance

Let X and Y be two random variables. Then the **covariance of X and Y** is

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y],\end{aligned}$$

where $\mu_X = E[X]$ and $\mu_Y = E[Y]$.

Note that $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$. Hence, the covariance is a **generalization** of the variance.

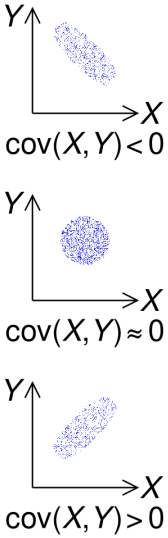


Image source: https://en.wikipedia.org/wiki/Covariance#/media/File:Covariance_trends.svg

Useful properties

For any X and Y, we have

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y],$$

and

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + 2\mathrm{Cov}(X, Y) + \mathrm{Var}[Y].$$

If X and Y are **independent**, then

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

Correlation coefficient

Let X and Y be two random variables. Assume $\text{Var}[X], \text{Var}[Y] \neq 0$ (*why is this a reasonable assumption?*). The **correlation coefficient** is

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}},$$

and always lies between -1 and 1 (which can be shown using Cauchy-Schwarz inequality).

- When $X = Y$ (**perfect positive correlation**), $\rho = 1$.
- When $X = -Y$ (**perfect negative correlation**), $\rho = -1$.
- When X and Y are **uncorrelated** then $\rho = 0$.

Independence vs correlation

Consider the following two statements:

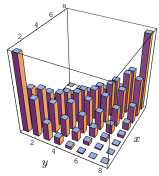
1. X and Y are independent;
2. X and Y are uncorrelated (i.e., $\mathrm{Cov}(X, Y) = 0$; the correlation coefficient is 0).

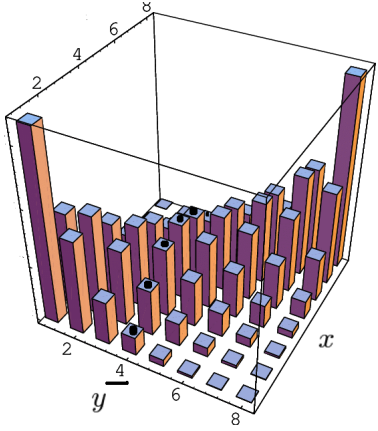
We have

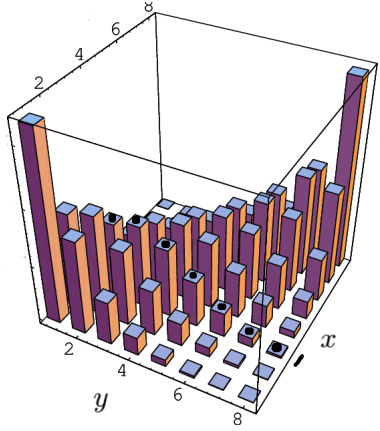
- (1) $\implies$ (2) (independence $\implies$ uncorrelated)
- (2) $\not\implies$ (1) (uncorrelated $\not\implies$ independence)
- Independence is a **stronger** condition than uncorrelation.

Outline

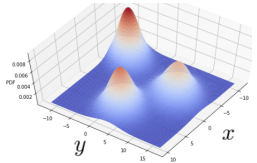
Conditional distributions

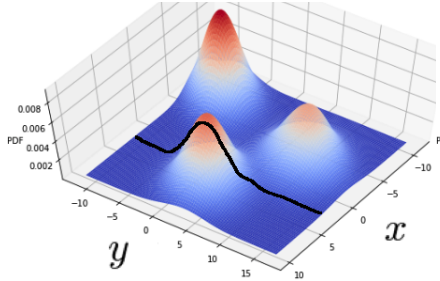


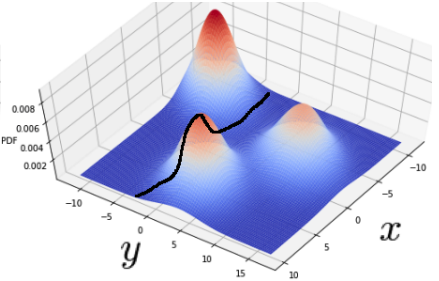




Conditional distributions







Conditional distributions

Let X and Y be two **discrete** random variables. The **conditional PMF** of Y given X is

$$p_{Y|X}(y \mid x) = \frac{p_{X,Y}(x,y)}{p_X(x)}.$$

We can see it as $\mathbb{P}(Y = y \mid X = x)$.

Let X and Y be two **continuous** random variables. The **conditional PDF** of Y given X is

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

Example

Consider **two coins** which can take values in $\{0, 1\}$. Let X be the *sum* of the two coins and Y be the *value of the first coin*.

- What is $p_X(\cdot)$?
- What is $p_{X|Y}(\cdot \mid y = 1)$?

Conditional distributions

Let X and Y be two **discrete** random variables. For any (measurable) $A \subseteq \mathcal{Y}$, we have

$$\mathbb{P}(Y \in A \mid X = x) = \sum_{y \in A} p_{Y|X}(y \mid x),$$

and

$$\mathbb{P}(Y \in A) = \sum_{x \in \mathcal{X}} \mathbb{P}(Y \in A \mid X = x) p_X(x).$$

Let X and Y be two **continuous** random variables. For any $A \subseteq \mathcal{Y}$, we have

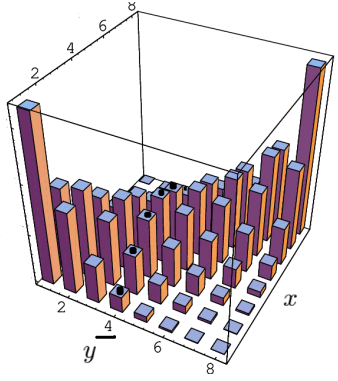
$$\mathbb{P}(Y \in A \mid X = x) = \int_A f_{Y|X}(y \mid x) dy,$$

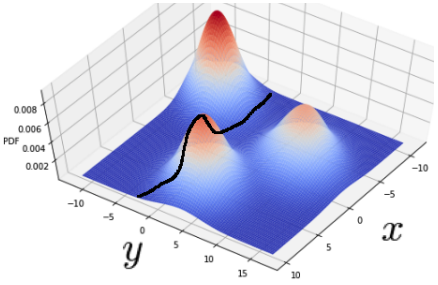
and

$$\mathbb{P}(Y \in A) = \int_{\mathcal{X}} \mathbb{P}(Y \in A \mid X = x) f_X(x) dx.$$

Outline

Conditional expectations





Conditional expectations

For a **discrete** random variable X, the **conditional expectation** of X given Y is

$$\mathbb{E}[X \mid Y = y] = \sum_{x \in \mathcal{X}} x \, p_{X|Y}(x \mid y).$$

For a **continuous** random variable X, the conditional expectation of X given Y is

$$\mathbb{E}[X \mid Y = y] = \int_{\mathcal{X}} x \, f_{X|Y}(x \mid y) dx.$$

The summation/integration is taken w.r.t. x, because $Y = y$ is **fixed**.

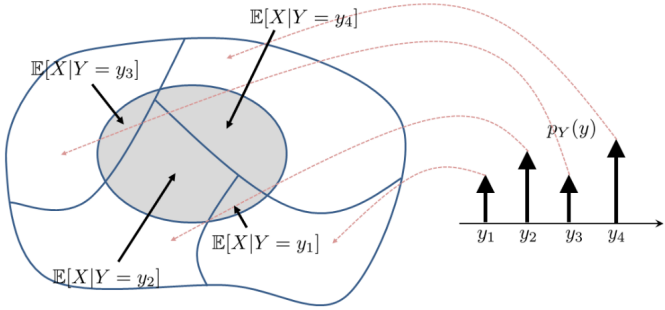
Law of Total Expectation

The **law of total expectation**[2] is a decomposition rule which allows to decompose the computation of $\mathbb{E}[X]$ into **conditional expectations** that are smaller/easier to compute.

$$\mathbb{E}[X] = \sum_{y \in \mathcal{Y}} \mathbb{E}[X \mid Y = y] \; p_Y(y),$$

or

$$\mathbb{E}[X] = \int_{\mathcal{Y}} \mathbb{E}[X \mid Y = y] \; f_Y(y) dy.$$



[2] Also known as the **law of iterated expectations** and the **tower rule**.

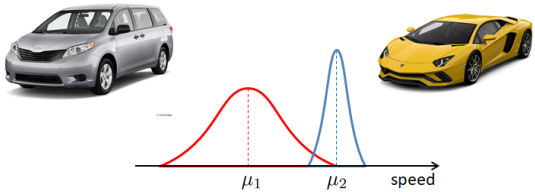
Example

Suppose there are two classes of cars. Let $C \in \{1, 2\}$ be the **class** and $S \in \mathbb{R}$ be the **speed**.

We know that

- $\mathbb{P}(C = 1) = p$,
- When $C = 1$, $S \sim \mathcal{N}(\mu_1, \sigma_1^2)$,
- When $C = 2$, $S \sim \mathcal{N}(\mu_2, \sigma_2^2)$.

You see a car on the freeway, what is its **average speed**?



Outline

Random vectors

Random vector:

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \text{ and } \boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

We assume that, by default, vectors are column vectors; we use $\boldsymbol{x}^T$ to denote the transpose of $\boldsymbol{x}$, which is a row vector

Joint PDF:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n).$$

Probability:

$$\mathbb{P}(\boldsymbol{X} \in A) = \int_A f_{\boldsymbol{X}}(\boldsymbol{x}) dx.$$

Mean vector and covariance matrix

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ be a random vector. The **expectation** is

$$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}] = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}.$$

The **covariance** matrix is

$$\boldsymbol{\Sigma} = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \ldots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \ldots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \ldots & \mathrm{Var}(X_n) \end{pmatrix},$$

which can be written in a more compact way as

$$\boldsymbol{\Sigma} = \mathbb{E}[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T].$$

Covariance matrix: properties

The covariance matrix is **symmetric**: $\mathbf{\Sigma} = \mathbf{\Sigma}^T$. This is due to the symmetry of the covariance operator: $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.

If the coordinates $X_1, X_2, \ldots, X_n$ are *uncorrelated* (or pairwise independent, which is stronger), the covariance matrix is a **diagonal** matrix:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathrm{Var}(X_1) & 0 & \ldots & 0 \\ 0 & \mathrm{Var}(X_2) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathrm{Var}(X_n) \end{pmatrix}.$$

Multivariate Gaussian distribution

A multivariate Gaussian distribution is characterized by a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$.[3] The PDF of a d-dimensional **multivariate Gaussian** is given by

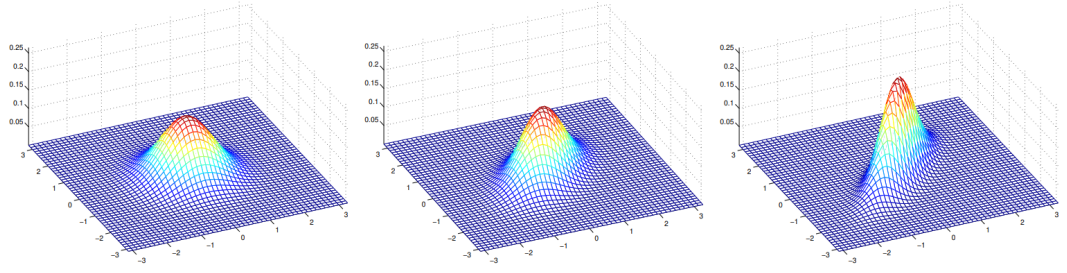
$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right\},$$

where $\boldsymbol{d}$ is the dimension of the random vector $\boldsymbol{X}$, and $|\boldsymbol{\Sigma}|$ is the determinant of the covariance matrix $\boldsymbol{\Sigma}$.

Exercise. Check that it gives the correct PDF when the components X_i are independent!

[3] The covariance matrix must be symmetric and *positive semi-definite*.

Multivariate Gaussian: examples

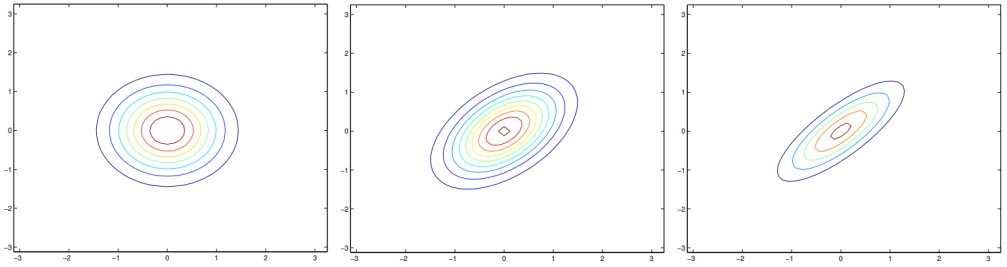


The figures above show the PDFs of Gaussians with mean 0 and with covariance matrix respectively

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}.$$

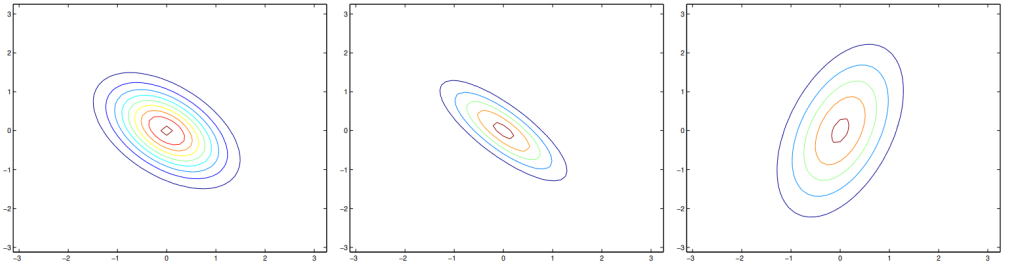
Multivariate Gaussian: examples

Other point of view: the *contours* of the above Gaussians are



Multivariate Gaussian: examples

Here are the contours of the PDF of other multivariate Gaussians of mean 0:

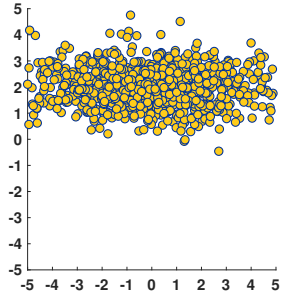


Try to guess (roughly) their covariance matrix!
Answer:

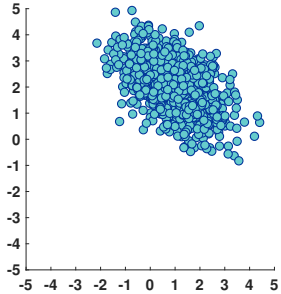
$$\Sigma = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 3 & 0.8 \\ 0.8 & 1 \end{pmatrix}.$$

Multivariate Gaussian: examples

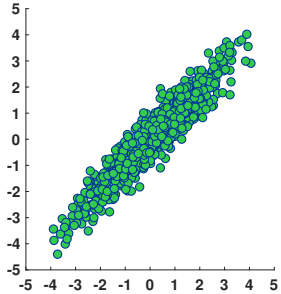
Third type of plot, obtained by sampling many points from Gaussians:



$$(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 5 & 0 \\ 0 & 0.5 \end{bmatrix}$$



$$\begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$



$$\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 & 1.9 \\ 1.9 & 2 \end{bmatrix}$$

Outline

Estimators

A central concept of machine learning (or statistics) is to **learn (or estimate)** certain properties about some underlying (stochastic) process on the basis of *samples* (data).

Point estimation refers to calculating a single "best guess" of the value of an unknown quantity of interest, which could be a **parameter** or a **density function**. We typically use $\hat{\theta}$ to denote a point estimator for a fixed parameter θ.

Given $X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_X$, a (point) **estimator** is a function of the observed sample, i.e.

$$\hat{\theta} = T(X_1, X_2, \ldots, X_n),$$

so that $\hat{\theta}$ is a *random variable*.

For example, the *sample mean* $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is an estimator for the expectation ($\theta = \mathbb{E}[X]$).

Bias of an estimator

The **bias of an estimator** is given by

$$b(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta = \mathbb{E}[\hat{\theta} - \theta].$$

An estimator is **unbiased** if $b(\hat{\theta}) = 0$.

Other properties of estimators

The **variance of an estimator** is given by

$$v(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2].$$

The **standard error of an estimator** is given by

$$\mathsf{se}(\hat{\theta}) = \sqrt{v(\hat{\theta})},$$

i.e., its standard deviation.

The **sampling distribution** of an estimator is the probability distribution of the estimator.

Example – The sample mean

Let $X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_X$, with $\mathbb{E}[X] = \mu_X$ and $\text{Var}(X) = \sigma_X^2$. The *sample mean* estimator is defined as

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

What are the bias and variance of $\overline{X}_n$?

- Since $\mathbb{E}[\overline{X}_n] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n} \sum_{i=1}^{n} \mu_X = \mu_X$, $\bar{X}_n$ is unbiased, i.e., the bias is equal to zero.
- Also, using the fact that $\text{Var}(\frac{1}{n} \sum_{i=1}^{n} X_i) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}(X_i, X_j)$, we can show (how?) that

  $$\text{Var}(\overline{X}_n) = \frac{\sigma_X^2}{n}.$$

  The variance decreases as the sample size n increases!

More on the sample mean

- The variance of the *average* is **much smaller** that the variance of the *individual* random variables.
  - This is one of the core principles of statistics and help us learn various quantities reliably by making **repeated independent measurements**.
- Why are *independent* measurements **essential**?
  - The extreme case of non-independence is when $X_1 = X_2 = \cdots = X_n$, for which we have (how?)

$$\mathrm{Var}(\overline{X}_n) = \sigma_X^2.$$

Inference

Let $y_1, y_2, \ldots, y_n \overset{\text{i.i.d.}}{\sim} p_Y$. How can we estimate p_Y?

- We often **assume** that the sample was generated from some (parametric) model.
- When we specify a model, we hope that it can provide a useful **approximation** to the data generation mechanism.
- The George Box quote is worth remembering in this context: "**all models are wrong; the practical question is how wrong do they have to be to not be useful**".

Maximum likelihood estimation

Let $y_1, y_2, \ldots, y_n \overset{\text{i.i.d.}}{\sim} p_Y$. How can we estimate p_Y?

- Let us restrict ourselves to a set of **possible distributions** $p(y; \theta)$, described by a finite number of parameters $\theta \in \mathbb{R}^d$.
- The goal of **maximum likelihood estimation** is to select the distribution $p(y; \theta)$ that is **most likely** to have generated the sample $y_1, y_2, \ldots, y_n$.

Maximum likelihood estimation

An example of a set of possible distributions for a **continuous variable** ($y \in \mathbb{R}$) is

$$\left\{ p(y; \mu; \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(y-\mu)^2}{2\sigma^2} \right\} \mid \mu \in \mathbb{R}, \sigma > 0 \right\},$$

where $\boldsymbol{\theta} = (\mu, \sigma)^T$ and $d = 2$. What is this distribution?

An example of a set of possible distributions for a **discrete variable** ($y \in \{0, 1\}$) is

$$\left\{ p(y; \alpha) = \alpha^y (1-\alpha)^{1-y} : 0 \leq \alpha \leq 1 \right\},$$

where $\boldsymbol{\theta} = \alpha$ and $d = 1$. What is this distribution?

Maximum likelihood estimation

The **likelihood function** is defined as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &\equiv \mathcal{L}(\boldsymbol{\theta}; y_1, y_2, \ldots, y_n) \\ &= p(y_1, y_2, \ldots, y_n; \boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} p(y_i; \boldsymbol{\theta}).\end{aligned}$$

The **maximum likelihood estimator**, or MLE — denoted by $\hat{\boldsymbol{\theta}}$ — is the value of $\boldsymbol{\theta}$ that maximizes $\mathcal{L}(\boldsymbol{\theta})$.

Log-likelihood

Working with a big product may not be convenient for practical computations. But note that $\hat{\theta}$ also maximizes the **log-likelihood function** $\log(\mathcal{L}(\theta))$! This is due to the fact that the logarithm is a **non-decreasing** function. This can be used to make a product into a sum, which is, e.g., easier to differentiate.

We write

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ \mathcal{L}(\theta) = \underset{\theta \in \Theta}{\operatorname{argmax}}\ \log(\mathcal{L}(\theta)),$$

where Θ is the parameter space.

Example

We observe $y_1, \ldots, y_n$ where $y_i \in \{0, 1\}$ with unknown PMF p_Y. If we assume

$$y_1, \ldots, y_n \overset{\text{i.i.d.}}{\sim} p(y; \alpha),$$

where

$$p(y; \alpha) = \alpha^y (1 - \alpha)^{1-y}$$

with $0 \leq \alpha \leq 1$.

What is the maximum likelihood estimate $\hat{\alpha}$?

Example

The **likelihood function** is given by

$$\begin{aligned}\mathcal{L}(\alpha; y_1, \ldots, y_n) &= p(y_1, \ldots, y_n; \alpha) \\ &= \prod_{i=1}^{n} p(y_i; \alpha) \\ &= \prod_{i=1}^{n} \alpha^{y_i}(1-\alpha)^{1-y_i} \\ &= \alpha^{\sum_{i=1}^{n} y_i}(1-\alpha)^{\sum_{i=1}^{n}(1-y_i)},\end{aligned}$$

and the **log-likelihood function** is given by

$$\begin{aligned}\log\, \mathcal{L}(\alpha; y_1, \ldots, y_n) &= \sum_{i=1}^{n} y_i \log(\alpha) + (1-y_i)\log(1-\alpha) \\ &= n\bar{y}\log(\alpha) + n(1-\bar{y})\log(1-\alpha),\end{aligned}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Example

The first derivative of the log-likelihood is given by

$$(\log\mathcal{L})'(\alpha) = n\bar{y}\frac{1}{\alpha} - n(1-\bar{y})\frac{1}{1-\alpha}.$$

A necessary condition for a maximum is given by

$$(\log\mathcal{L})'(\alpha) = 0 \iff \hat{\alpha} = \bar{y}.$$

We can verify that it is indeed a maximum by checking that the second derivative of the log-likelihood at $\hat{\alpha}$ is indeed negative, i.e., $(\log\mathcal{L})''(\hat{\alpha}) < 0$.

Overview

Probability

Random variables

Discrete random variables

Continuous random variables

Multivariate random variables

Conditional distributions

Conditional expectations

Random vectors (more than two variables)

Inference