# Machine Learning I
## Review of Probability and Statistics

Souhaib Ben Taieb

University of Mons

# Overview

Probability

Random variables

Discrete random variables

Continuous random variables

Multivariate random variables

Conditional distributions

Conditional expectations

Random vectors (more than two variables)

Inference

# References

- **Introduction to Probability for Data Science**, Stanley H. Chan. [Link] (Book, slides and videos)
- Probability Theory Review for Machine Learning, 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** are to which we want to attach probability.

▶ We often conduct an experiment, i.e. take some measurements of a **random (stochastic) process**.

▶ Our measurements take values in some set $\Omega$, the **sample space** (or the outcome space)., which defines *all possbile outcomes* of our measurements.

## Sample space and events

▶ We toss one coin 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 $\Omega$ ($A \subseteq \Omega$), 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 $\omega = HT$. 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

- $\Omega$ is the **sample space**
- $\mathcal{F} = 2^{\Omega}$ is the **space of events** (or event space)[1]
- $\mathbb{P}$ is the **probability measure/distribution** that maps an event $A \in \mathcal{F}$ to a real value between zero and one

---

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

# Probability axioms

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

1. *Non-negativity*: $\mathbb{P}(A) \geq 0, \forall A \subseteq \Omega$
2. *Unity of $\Omega$*: $\mathbb{P}(\Omega) = 1$
3. *Additivity*: For all disjoint events $A, B \in \mathcal{F}$ (i.e. $A \cap B = \varnothing$), we have that, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

# Probability properties

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

- $\mathbb{P}(\varnothing) = 0$
- $A \subset B \implies \mathbb{P}(A) \leq \mathbb{P}(B)$
- $0 \leq \mathbb{P}(A) \leq 1$
- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

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

## Probability properties

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

$$\mathbb{P}(\Omega) = 1 \quad \text{(Axiom 2)}$$
$$\iff \mathbb{P}(A \cup A^c) = 1, \quad \forall A \subseteq \Omega$$
$$\iff \mathbb{P}(A^c) + \mathbb{P}(A) = 1 \quad \text{(Axiom 3)}$$
$$\iff \mathbb{P}(A^c) = 1 - \mathbb{P}(A)$$

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

$$A \subseteq B$$
$$\implies B = A \cup (B \backslash A) \quad (A \cap (B \backslash A) = \varnothing)$$
$$\implies \mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \backslash A) \quad \text{(Axiom 3)}$$
$$\implies \mathbb{P}(B) \geq \mathbb{P}(A) \quad \text{(Axiom 1)}$$

## 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:

$$\mathbb{P}(A) = \mathbb{P}(\{w_1, w_2, \ldots, w_k\}) = \sum_{i=1}^{k} \mathbb{P}(\{w_i\})$$

▶ If $\Omega$ 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 less than five. What is $\mathbb{P}(A)$?

## Conditional probability

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

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

Note: $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$  (in general)

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

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

More generally, we have

$$\mathbb{P}(A_1 \cap A_2 \cap A_3 \dots) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1A_2)\dots$$

## Independence of events

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

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

A set of events $A_j (j \in J)$ are called **mutually independent** if

$$\mathbb{P}\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} \mathbb{P}(A_j).$$

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 sometimes 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 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

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

$$= 1 - \mathbb{P}(T \cap T \cap T \cap \cdots \cap T) \tag{2}$$

$$= 1 - \prod_{i=1}^{10} \mathbb{P}(T) \tag{3}$$

$$= 1 - (1/2)^{10} \tag{4}$$

## Exercise

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

▶ Compute $\mathbb{P}(A|B)$

▶ Compute $P(A)$

▶ Are $A$ and $B$ independent?

# Law of total probability

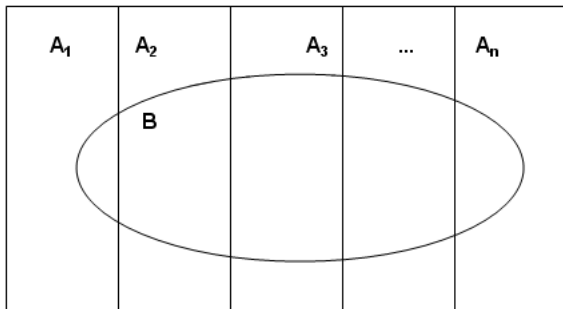Let $A_1, A_2, \ldots, A_n$ be a partition of $\Omega$. What is the probability of $B$?

## Law of total probability

Let $A_1, A_2, \ldots, A_n$ be a partition of $\Omega$. Then, for any $B \subseteq \Omega$, we have that

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

The **law of total probability** is a combination of **additivity** and **conditional probability**. In fact, we have

$$\begin{aligned}
\mathbb{P}(B) &= \mathbb{P}((B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_k)) \\
&= \sum_{i=1}^{n} \mathbb{P}(B \cap A_i) \\
&= \sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i)
\end{aligned}$$

## Bayes' Rule

(**Bayes' Rule**) Let $A_1, A_2, \ldots, A_n$ be a partition of $\Omega$. Then, we have that

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

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

Bayes' Rule is a combination of **conditional probability** and the **law of total probability**:

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)}.$$

## Example

Suppose there are three types of emails: $A_1 =$ SPAM, $A_2 =$ Low Priority and $A_3 =$ 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 $B$ the event that an email contains the word "free", then $\mathbb{P}(B|A_1) = 0.9, \mathbb{P}(B|A_2) = 0.1, \mathbb{P}(B|A_3) = 0.1$. 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

$$\Omega = \{\underbrace{HHH}_{3}, \underbrace{HHT}_{2}, \underbrace{HTH}_{2}, \underbrace{THH}_{2}, \underbrace{TTH}_{1}, \underbrace{THT}_{1}, \underbrace{HTT}_{1}, \underbrace{TTT}_{0}\}$$

These summary statistics are called **random variables**. Specifically, a random variable is a function from the sample space $\Omega$ to the reals.

## Random variables

A random variable can be seen as a **mapping** between a distribution on $\Omega$ to a distribution on the reals (or the range of the random variable, $\mathcal{X} \subseteq \mathbb{R}$). Formally, we have that for some subset $S \subseteq \mathcal{X}$,

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

For the previous example, we have

$$\Omega = \{\underbrace{HHH}_{3}, \underbrace{HHT}_{2}, \underbrace{HTH}_{2}, \underbrace{THH}_{2}, \underbrace{TTH}_{1}, \underbrace{THT}_{1}, \underbrace{HTT}_{1}, \underbrace{TTT}_{0}\}$$

and

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

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

In the following, we will use $\mathbb{P}$ to denote probability.

# Outline

# Probability mass function

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

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

What is the PMF of the previous coin flip example?

A function $p_X$ is a PMF if and only if

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

# Some important discrete distributions

▶ 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 **Bernouilli** 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

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



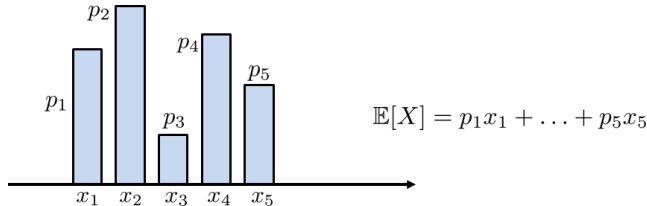$$\mathbb{E}[X] = p_1 x_1 + \ldots + p_5 x_5$$

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

# Expectation and its properties

For any function $g$, we have

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

For any function g and h,

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

For any constant c,

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

For any constant c,

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

## 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

$$\text{Var}(X) = \mathbb{E}[(X - \mu_X)^2],$$

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

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

# Outline

# Probability density function

The **probability density function** (PDF) of a continuous random variable X is a function $f_X$, when integrated over an interval $[a, b]$, yields the probability of obtaining $a \leq X \leq b$:

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

A PDF has the following properties:

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

Note that $f_X(x)$ is not the probability of having $X = x$. In fact, we can have $f_X(x) > 1$.

## Probability density function

A PDF has the following properties:

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

Note that $f_X(x)$ is **not** the probability of having $X = x$ since we can have $f_X(x) > 1$. Also, for a continuous random variable, we have $\mathbb{P}(X = x) = 0$ $\forall x \in \mathcal{X}$.

## Some important continuous distributions

▶ 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 $X \sim \mathcal{U}[a, b]$.

▶ **Gaussian** distribution. With a location (mean) $\mu$ and scale (standard deviation) $\sigma$, 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 $g$, we have

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

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

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

## 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

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

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

See the useful properties of the variance introduced previously.

# 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** PDF or PMF.
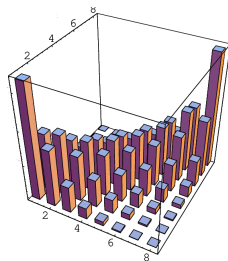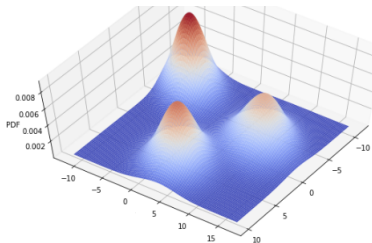
# Important concepts

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

In the following, we will mainly focus on **bivariate random variables**.

## Joint distributions

- $f_X(x)$
- $f_{X_1, X_2}(x_1, x_2)$
- $f_{X_1, X_2, X_3}(x_1, x_2, x_3)$
- $\ldots$
- $f_{X_1, \ldots, X_n}(x_1, \ldots, x_n)$
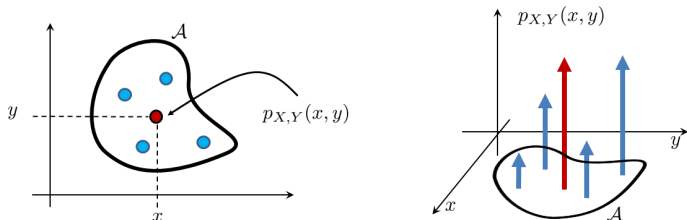- We often just write $f_X(x)$ when the dimensionality is clear from context.

# Joint PMF

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

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

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

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



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

## 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

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

# 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:

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

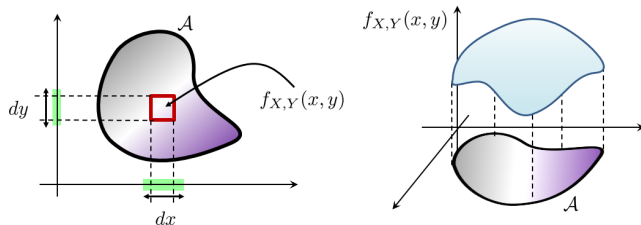for any $A \subseteq \mathcal{X} \times \mathcal{Y}$.



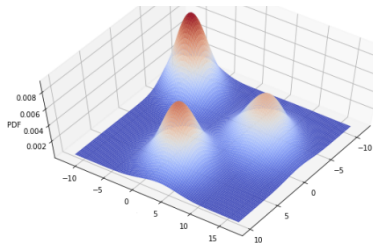Image source:: Introduction to Probability for Data Science, Stanley H. Chan.

# Marginal distribution

The **marginal PMF** is 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 PDF** is defined as

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

# Independence

If two random variables X and Y are **independent**, then

$$p_{X,Y}(x,y) = p_X(x)p_Y(y), \qquad \text{and } f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

If a sequence of random variables $X_1, \ldots, X_N$ are independent, then their joint PDF (or joint PMF) can be factorized:

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

# 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.
2. All $X_1, \ldots, X_N$ have the same distribution.

## Joint 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 $g$, 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

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

## Joint expectations

Let $g(X, Y) = XY$, we have

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

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

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

## Covariance

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

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \tag{5}$$
$$= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y], \tag{6}$$

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

Note that $\text{Cov}(X, X) = \text{Var}(X)$.

## Useful properties

For any $X$ and $Y$, we have

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

and

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

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

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

## Correlation

Let X and Y be two random variables. The **correlation coefficient** is

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

where $-1 \leq \rho \leq 1$ .

- ▶ When $X = Y$ (fully correlated), $\rho = 1$.
- ▶ When $X = -Y$ (fully correlated), $\rho = -1$.
- ▶ When $X$ and $Y$ are uncorrelated then $\rho = 0$.

## Independence vs correlation

Consider the following two statements:

1. $X$ and $Y$ are independent;
2. $\text{Cov}(X, Y) = 0$.

We have

- ▶ (1) $\implies$ (2) (independence $\implies$ uncorrelated)
- ▶ (2) $\not\implies$ (1) (uncorrelated $\not\implies$ independence)
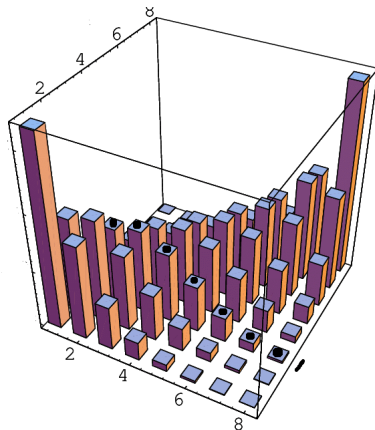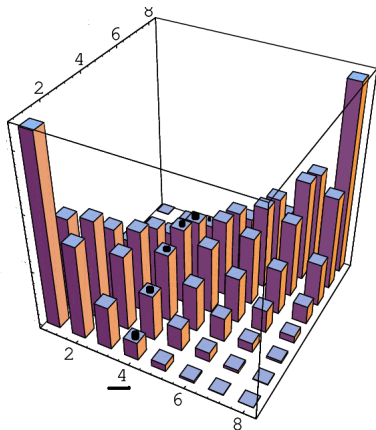- ▶ Independence is a stronger condition than correlation

# Outline

# Conditional distributions
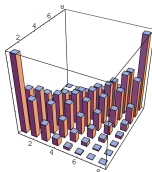
# Conditional distributions
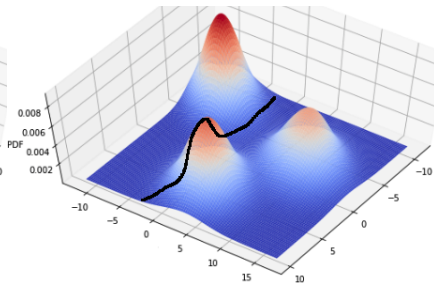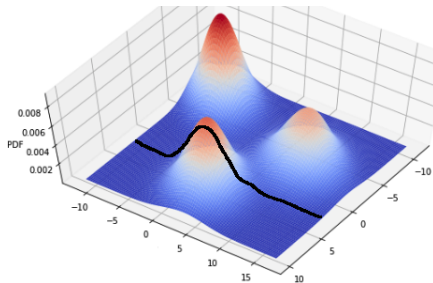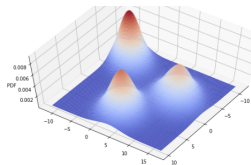
## Conditional distributions

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

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

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

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

# Example

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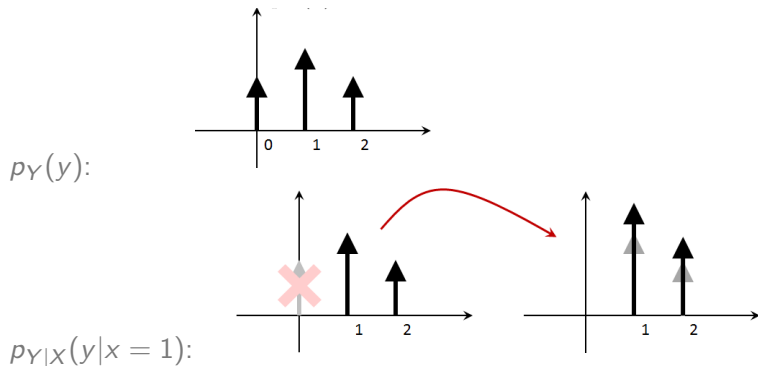


$p_Y(y)$:

$p_{Y|X}(y|x = 1)$:

## Conditional distributions

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

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

and

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

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

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

and

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

# Outline

## Conditional expectations

For a discrete random variable $Y$, the **conditional expectation** of $Y$ given $X$ is
$$\mathbb{E}[Y|X = x] = \sum_{y \in \mathcal{Y}} y \; p_{Y|X}(y|x).$$

For a continuous random variable $Y$, the conditional expectation of $Y$ given $X$ is
$$\mathbb{E}[Y|X = x] = \int_{\mathcal{Y}} y \; f_{Y|X}(y|x) dy$$

The summation/integration is taken w.r.t. y, because $X = x$ is given and fixed.

## Law of Total Expectation

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

$$\mathbb{E}[Y] = \sum_{x \in \mathcal{X}} \mathbb{E}[Y|X = x]p_X(x) \text{ or } \mathbb{E}[Y] = \int_{\mathcal{X}} \mathbb{E}[Y|X = x]f_X(x)dx$$

Note the difference between

$$h(x) = \mathbb{E}_{Y|X}[Y|X = x], \quad \text{(A deterministic function in } x\text{)}$$

and

$$h(X) = \mathbb{E}_{Y|X}[Y|X]. \quad \text{(A function of the random variable } X\text{)}$$

The **law of total expectation** can also be written as

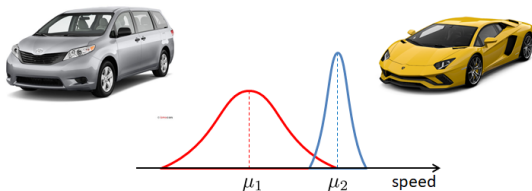$$\mathbb{E}[Y] = \mathbb{E}_X[\mathbb{E}_{Y|X}[Y|X]].$$

# Example

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

An image from a dataset can be represented by a **high-dimensional** vector.
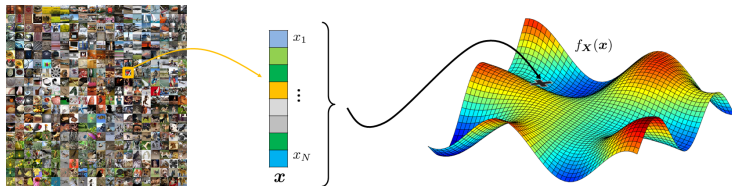


Image source:: Introduction to Probability for Data Science, Stanley H. Chan.

# Random vectors

Random vector:

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

Joint PDF:

$$f_X(x) = f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)$$

Probability:

$$\mathbb{P}(X \in A) = \int_A f_X(x)dx$$

## Mean vector and covariance matrix

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

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

The **covariance** matrix is

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

which can be written in a more compact way as

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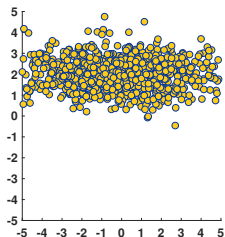
## Diagonal covariance matrix

If the coordinates $X_1, X_2, \ldots, X_n$ are *uncorrelated*, the covariance matrix is a **diagonal** matrix:

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

# Multivariate Gaussian

A $d$-dimensional **joint Gaussian** has a PDF:

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



$$(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 5 & 0 \\ 0 & 0.5 \end{bmatrix} \qquad \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix} \qquad \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 & 1.9 \\ 1.9 & 2 \end{bmatrix}$$
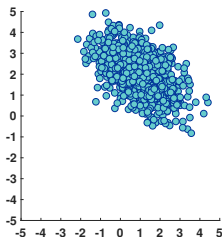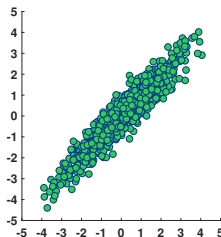
# Outline

# 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 $\theta$.

Given $X_1, X_2, \ldots, X_n \sim p_X$, 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* $\hat{\theta} = \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ is an estimator for the expectation ($\theta = \mathbb{E}[X]$).

# Properties of estimators

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

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

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

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

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

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

i.e., its standard deviation.

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

## Example - The sample mean

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

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

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

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

## Example - The sample mean

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

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

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

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

## 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 independent measurements are **essential**? The extreme case of
non-independence is when $X_1 = X_2 = \cdots = X_n$, for which we have

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

## Inference

Let $y_1, y_2, \ldots, y_n \sim p_Y$. 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, but some are useful**".

## Maximum likelihood estimation

Let us restrict ourselves to a set of possible distributions $p(y; \boldsymbol{\theta})$, described by a finite number of parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

An example for $y \in \mathbb{R}$ is

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

where $\boldsymbol{\theta} = (\mu, \sigma)^T$, and, for $y \in \{0, 1\}$,

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

where $\boldsymbol{\theta} = \alpha$.

The goal of maximum likelihood estimation is to select the distribution $p(y; \boldsymbol{\theta})$ that is **most likely** to have generated the sample $y_1, y_2, \ldots, y_n$.

## Maximum likelihood estimation

The **likelihood function** is defined as

$$\mathcal{L}(\boldsymbol{\theta}) \equiv \mathcal{L}(\boldsymbol{\theta}; y_1, y_2, \ldots, y_n) \tag{7}$$

$$= p(y_1, y_2, \ldots, y_n; \boldsymbol{\theta}) \tag{8}$$

$$= \Pi_{i=1}^n p(y_i; \boldsymbol{\theta}). \tag{9}$$

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

We write

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

where $\Theta$ 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

$$y_1, \ldots, y_n \sim p(y; \alpha),$$

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

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

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-ikelihood at $\hat{\alpha}$ is indeed negative, i.e. $(\log \mathcal{L})^{''}(\hat{\alpha}) < 0$.