Probability and density estimation

Sandro Cumani

sandro.cumani@polito.it

Politecnico di Torino

Random events and probability

Our goal: making predictions that allow us to take actions

Complex phenomena: too many factors that can influence the outcomes to account for all of them

Solution: model phenomena in terms of random events

- Deterministic event: its occurrence can be predicted exactly
- Random event: its occurrence is uncertain

Random events and probability

We describe random events in terms of their probability

- Classical interpretation: probability as fraction of favorable outcomes
- Frequentist interpretation: probability as frequency of an outcome under a large (infinite) number of repeated trials
- Bayesian interpretation: probability as (subjective) measure of belief that an event will occur

The axiomatic treatment defines the quantity that we will use

Interpretations provide guidance on how to employ the axiomatic theory to describe reality

Random events and probability

The next set of slides (5 to 51) recalls basic notions of probability

- Probability spaces
- Random variables (discrete and continuous) and random vectors
- Cumulative distribution functions
- Probability density functions
- Expectations

These are provided as a quick reference. Specific notions will be discussed when needed.

Axioms

Let

i) Ω be a set (possible outcomes)

ii) $\mathcal{A}$ be σ-field over Ω

a) $\mathcal{A}$ is a collection of subsets of Ω (events)

b) $\Omega \in \mathcal{A}$

c) If $A \in \mathcal{A}$ then $A^C \in \mathcal{A}$

d) If $A_1, A_2, \cdots \in \mathcal{A}$ then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$

Note that b), c) and d) imply that

e) $\emptyset \in \mathcal{A}$

f) $\bigcap_{n=1}^{\infty} A_n \in \mathcal{A}$

Axioms

We define probability as a function P

$$P : \mathcal{A} \longrightarrow \mathbb{R}^+$$

which has the properties:

1) $P(\Omega) = 1$, and

2) *(countable additivity)* Given a sequence $A_1, A_2, \ldots$ of mutually exclusive elements of $\mathcal{A}$ (i.e., $i \neq j \implies A_i \cap A_j = \emptyset$),

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{i=n}^{\infty} P(A_n)$$

The triplet $(\Omega, \mathcal{A}, P)$ is called a probability space

Properties

Some properties of probability spaces:

- $P(\emptyset) = 0$
- $A \subset B \implies P(A) \leq P(B)$
- $P(A) \leq 1$
- $P(A^C) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P\left(\bigcup_n A_n\right) = 1 - P\left(\bigcap_n A_n^C\right)$
- If $A_1 \subset A_2 \subset A_3 \subset \ldots$ is an increasing sequence of sets with $A = \bigcup_n^\infty A_n$ (with $A = \bigcap_n^\infty A_n$). Then $P(A) = \lim_{n\to\infty} P(A_n)$ (the same applies for a decreasing sequence of sets, with $A = \bigcap_n A_n$)

An example

Consider rolling a six-sided die

The set of possible outcomes is the set of possible values that we can roll

- $\Omega = \{1, 2, 3, 4, 5, 6\}$

Events can consist of single outcomes, e.g. “rolling a 6”

- $A = \{6\}$

Events can consider more outcomes, e.g. “rolling an even number”

- $A = \{2, 4, 6\}$

An example

We can assign a probability to the different events corresponding to the outcomes

If we assume that these events have the same probability p:

$$P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = P(\{6\}) = p$$

Since the events are mutually exclusives and their union is Ω, it follows that the value of p should be $p = \frac{1}{6}$

We can then compute the probability of rolling an even number:

$$P(\{2,4,6\}) = P(\{2\}) + P(\{4\}) + P(\{6\}) = \frac{1}{2}$$

Conditional probability

Let $A, B \in \mathcal{A}$, with $P(B) > 0$

We define the conditional probability of A given B as

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

The conditioned probability represents the probability of A, when we know that B has happened

Note: $A|B$ is not an event per se, we are rather defining a new probability measure $P(\cdot|B)$

Conditional probability

Consider again rolling a six-sided die

We don't know the result, but somebody tells us that it's not 1

We can compute the probability of rolling a number lower than 4 ($A = \{1, 2, 3\}$) conditioned on the roll being larger than one ($B = \{2, 3, 4, 5, 6\}$):

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{2, 3\})}{P(\{2, 3, 4, 5, 6\})} = \frac{2}{5}$$

Conditional probability

Bayes formula allows expressing the conditional probability $P(B|A)$

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

We say that two events are independent if and only if

$$P(A \cap B) = P(A)P(B)$$

In the following we will often denote the *joint* probability of two events $P(A \cap B)$ simply as $P(A, B)$

Conditional probability

Considering again the previous example, we can compute the probability that the roll is greater than one (event B) given that the roll is lower than 4 (event A):

$$P(A|B) = \frac{2}{5}, \quad P(A) = \frac{1}{2}, \quad P(B) = \frac{5}{6}$$

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{2}{3}$$

We can also verify that

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(\{2,3\})}{P(\{1,2,3\})} = \frac{2}{3}$$

Random variables

Random variables allow extending probabilistic reasoning to quantities that depend on events

A random variable X is defined as a function

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

such that, for any real value $x \in \mathbb{R}$ the event

$$\{\omega \in \Omega : X(\omega) \leq x\}$$

belongs to $\mathcal{A}$.

Essentially X is a function of the outcomes ω, for which we can compute the probability that it takes values no greater than x:

$$P(X \leq x) = P(\{\omega \in \Omega : X(\omega) \leq x\})$$

Random variables

Consider rolling two six-sided dices

The outcome space Ω consists of pairs of values

$$\Omega = \{(1,1),(1,2),\ldots,(6,6)\}$$

We are interested in computing the probability that the sum of the two dice is lower or equal to some value

Let an outcome be denoted as $\omega = (r_1, r_2)$. We define the R.V. X that maps each outcome to the sum of its components:

$$X(\omega) = r_1 + r_2$$

Random variables

We can then compute the probability that $X \leq x$:

$$P(X \leq x) = P(\{(r_1, r_2) \in \Omega : r_1 + r_2 \leq x\})$$

For example, if we set $x = 4$, then we have

$$\begin{aligned} P(X \leq 4) &= P(\{(r_1, r_2) \in \Omega : r_1 + r_2 \leq 4\}) \\ &= P(\{(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)\}) = \frac{1}{6} \end{aligned}$$

Random variables

A R.V. induces a probability space over the real line

In general, we will be interested in assigning probabilities to subset of the real line A in the form $P(X \in A)$

Note that not all subsets A of the real line may be assigned a probability, as the set $\{\omega \in \Omega : X(\omega) \in A\}$ may not be an event

The subset of the real line for which we can compute the probability $P(X \in A)$ is a *Borel* σ-field $\mathcal{B}$

The probability space induced by X is $(\mathbb{R}, \mathcal{B}, P)$

Note: the same symbol P is used for both for the original and the induced probability space, its meaning will be clear from the context

Random variables

For example, the definition of a R.V. implies that $\{\omega : X(\omega) \leq b\}$ is an event

It follows that

$$\{\omega : X(\omega) > a\}$$

is also an event. It follows that

$$\{\omega : a < X(\omega) \leq b\} = \{\omega : X(\omega) > a\} \cap \{\omega : X(\omega) \leq b\}$$

and

$$\{\omega : X(\omega) = x\} = \bigcup_n \{\omega : x - \frac{1}{n} < X(\omega) \leq x\}$$

are also an event

We can thus compute the probability that $X \in (a, b]$, i.e. $P(a < X \leq b)$, or the probability that $X = x$, i.e. $P(X = x)$

Random variables

Given R.V. X, we define its cumulative distribution function (c.d.f.) as:

$$F_X(x) = P(X \leq x)$$

The c.d.f. has the following properties:

1) $0 \leq F_X(x) \leq 1 \quad \forall x \in \mathbb{R}$

2) F_X is non-decreasing, i.e. $x_1 < x_2 \implies F_X(x_1) \leq F_X(x_2)$

3) $\lim_{x \to -\infty} F_X(x) = 0$, $\lim_{x \to \infty} F_X(x) = 1$

4) F_X is right-continuous: $\lim_{x \downarrow x_0} F_X(x) = F_X(x_0)$

It follows that

i) $P(a < X \leq b) = F_X(b) - F_X(a)$

ii) $P(X = x_0) = F_X(x_0) - \lim_{x \uparrow x_0} F_X(x)$

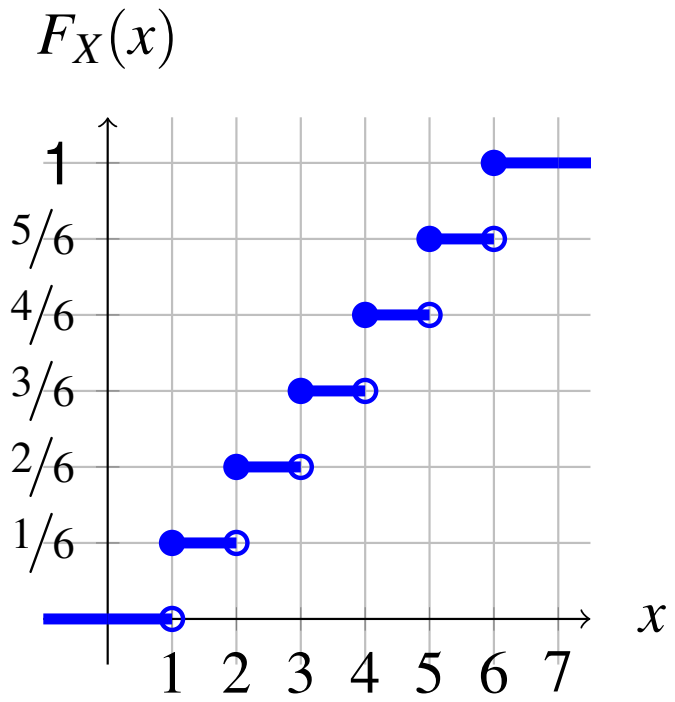
Random variables

Consider again rolling a six-sided die

Consider the R.V. that gives the rolled value $X(\omega) = \omega$

The c.d.f. is given by

$$F_X(x) = \begin{cases} 0 & \text{if } x < 1 \\ 1/6 & \text{if } 1 \leq x < 2 \\ 2/6 & \text{if } 2 \leq x < 3 \\ 3/6 & \text{if } 3 \leq x < 4 \\ 4/6 & \text{if } 4 \leq x < 5 \\ 5/6 & \text{if } 5 \leq x < 6 \\ 1/ & \text{if } x \geq 6 \end{cases} \tag{1}$$



Discrete random variables

A R.V. is said to be discrete if it takes a finite or a countably infinite number of values

For a discrete R.V. X we can define the probability mass function (p.m.f.) or discrete density:

$$f_X(x) = P(X = x)$$

Discrete random variables

The discrete density has the following properties:

1) $f_X(x) \geq 0$

2) $f_X(x) = 0$ for all x except at most a countably infinite number of values (those taken by X)

3) $\sum_{x \in \mathcal{S}} f_X(x) = 1$

where $\mathcal{S}$ is the support of X, i.e. the set of values for which $f_X(x) > 0$

If we know the density. we can compute the c.d.f.

$$F_X(x) = \sum_{t \leq x} f_X(t)$$

Discrete random variables

The examples that we have considered so far involve discrete R.V. with a finite support

- Rolled value for a six-sided die:

$$\mathcal{S} = 1, 2, 3, 4, 5, 6$$

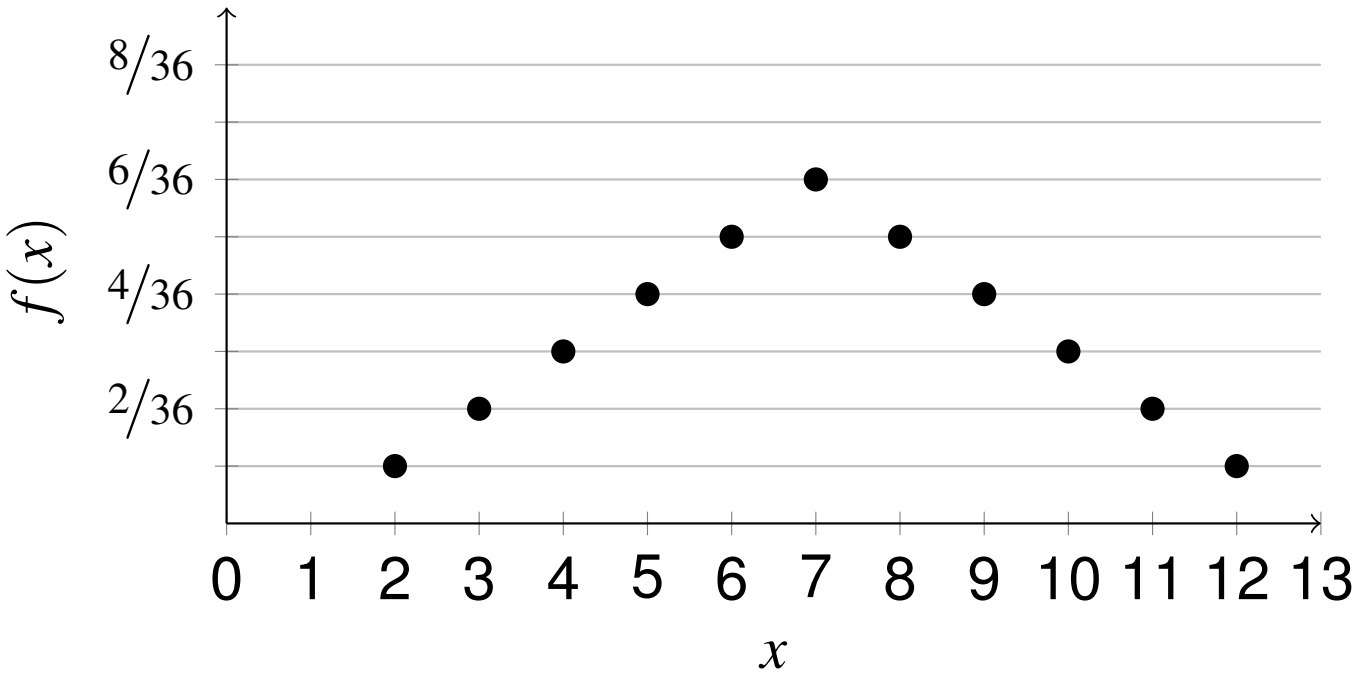
$$f_X(x) = \frac{1}{6} \quad \forall x \in \mathcal{S}$$

Discrete random variables

The examples that we have considered so far involve discrete R.V. with a finite support

- Sum of the values rolled by two dice:

$$\mathcal{S} = 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$$



Discrete random variables

We consider now an example where the support is not finite

We start considering the problem of tossing a coin K times

For example, with $K = 3$ the outcomes are the possible sequences

$$\Omega = (HHH), (HHT), (HTH), (HTT), (THH), (THT), (TTH), (TTT)$$

In general, if the coin is fair, we can assign to each sequence the same probability $p = \frac{1}{2^K}$

Discrete random variables

We now consider the problem of tossing a coin until it comes up head

The outcome consists of a countable number of elements (sequences of tails followed by one head)

$$\Omega = \{(H), (TH), (TTH), (TTTH), \dots\}$$

If the coin is fair, it's reasonable to assume that the events "the first toss is a head", $A = \{(H)\}$, and its complement "the first toss is a tail" have both probability $P(A) = P(A^C) = 1/2$

Assuming tosses are independent, we can now consider the events "the second toss is a head", $B = \{(TH)\}$, and its complement "the second toss is a tail" (both imply that the first toss was tail as well)

Discrete random variables

We note that $B \cup B^C = A^C$, so that $P(B \cup B^C) = 1/2$

Again, if the coin is fair, it's reasonable to assume that $P(B) = P(B^C) = 1/4$

By induction, we can assign to any event ω a probability $P(\{\omega\}) = 1/2^K$, where K is the length of the sequence ω.

In general, if the coin is biased and the probability of head is p, then

$$P(\{\omega\}) = p(1-p)^{K-1}$$

Discrete random variables

We now consider the R.V. X that maps an outcome to the number of *tails* required to get a head (note that this is the sequence length *minus 1*)

$$X((H)) = 0, \quad X((TH)) = 1, \quad X((TTH)) = 2, \ldots$$

The support of X is the set of positive integers $\mathbb{N}^+$
The density (or distribution) of X is the geometric distribution:

$$f_X(x) = \begin{cases} p(1-p)^x & x \in \mathbb{N}^+ \\ 0 & \text{otherwise} \end{cases}$$

Continuous random variables

In many cases, a R.V. can take any value in $\mathbb{R}$ (or in a subset of $\mathbb{R}$)

Nevertheless, we have shown that also in this cases $\{X \leq x\}$ are events, and we can define the c.d.f. of X as $F_X(x) = P(X \leq x)$

A R.V. X for which F_X is a continuous function will be referred to as continuous Random Variable

If F_X is continuous, then $\lim_{x \uparrow x_0} F_X(x) = \lim_{x \downarrow x_0} F_X(x)$, thus

$$P(X = x) = 0$$

It follows that

$$P(a < X \leq b) = P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b)$$

Continuous random variables

We have seen that for discrite R.V.s we can express the cumulative function in terms of probability mass functions

Similarly, for a continuous R.V. we introduce the concept of probability density

We say that a function $f : \mathbb{R} \longrightarrow \mathbb{R}$ is a density if and only if

1) $f(x) \geq 0 \quad \forall x \in \mathbb{R}$

2) f is integrable over $\mathbb{R}$

3) $\int_{-\infty}^{+\infty} f(x)dx = 1$

Continuous random variables

Let F_X be the c.d.f. of continuous R.V. X, and let there exist a density $f_X(x)$ such that

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

We will say that f_X is a probability density function (p.d.f.) of X

Note that f_X is not unique (e.g. if f_X and g_X differ on a non-measurable subset of $\mathbb{R}$, they are both densities for X)

If F_X is differentiable, then

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

is a p.d.f. of X

Continuous random variables

In contrast with the density of discrete R.V.s, the p.d.f. of a continuous R.V. can, in general, take values greater than 1

The p.d.f. allows computing the probability that X is in an interval $[a, b]$:

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

More in general, if $X \in A$ is an event, we can compute its probability as

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

Random vectors

We define a m-dimensional Random Vector as a vector whose components $X_1 \dots X_m$ of $\boldsymbol{X}$ are Random Variables

Let $\boldsymbol{x} = (x_1, \dots, x_m)$.

The cumulative distribution function of $\boldsymbol{X}$ is defined as:

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = P(X_1 \leq x_1, \dots, X_m \leq x_m)$$

i.e., the joint cumulative distribution for $X_1 \dots X_m$

The distributions of each component (or subset of components) are called marginal distributions

Random vectors

Discrete R.V.:

$$f_X(\boldsymbol{x}) = P(\boldsymbol{X} = \boldsymbol{x}) = P(X_1 = x_1, \ldots, X_m = x_m)$$

$$F_X(\boldsymbol{x}) = \sum_{y_1 \leq x_1} \cdots \sum_{y_m \leq x_m} f_X(\boldsymbol{y})$$

Continuous R.V.:

$$f_X(\boldsymbol{x}) = \frac{\partial^m}{\partial x_1 \cdots \partial x_m} F_X(\boldsymbol{x})$$

$$F_X(\boldsymbol{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_m} f_X(\boldsymbol{y}) dy_1 \cdots dy_m$$

Marginals

$$f_{X_1}(x_1) = \sum_{y_2 \in \mathcal{S}(X_2)} \cdots \sum_{y_m \in \mathcal{S}(X_m)} f_X(x_1, y_2, \ldots, y_m)$$

$$f_{X_1}(x_1) = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f_X(x_1, y_2, \ldots, y_m) dy_2 \cdots dy_m$$

Statistical independence: $X \perp\!\!\!\perp Y \iff F_{X,Y}(x,y) = F_X(x) F_Y(y)$

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) \iff X \perp\!\!\!\perp Y$$

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) \implies X \perp\!\!\!\perp Y$$

Given two R.V.s X and Y, we define the conditional density of X given $Y = y$:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}, \quad \forall y \in \mathcal{S}(Y)$$

where $\mathcal{S}(Y)$ is the support of Y

Bayes rule:

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

Also,

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{\int f_{X,Y}(x,y)dy} = \frac{f_{X|Y}(x|y)f_Y(y)}{\int f_{X|Y}(x|y)f_Y(y)dy}$$

Coinditional independence: X and Y are independent given Z,

$$(X \perp\!\!\!\perp Y)|Z \iff F_{X,Y|Z}(x,y|z) = F_{X|Z}(x|z)F_{Y|Z}(y|z)$$

Transformations of random variables

In many cases we are interested in the distribution of *functions* of our R.V.s

For example, we may want to know the distribution of $Y = X^2$, or, more in general, of $Y = g(X)$

For discrete R.V.s we can simply compute

$$f_Y(y) = \sum_{x|f(x)=y} f_X(x)$$

Transformations of random variables

For continuous R.V.s, we can compute the c.d.f. of Y as

$$F_Y(y) = P(g(X) \leq y) = P(X \in \{x | g(x) \leq y\})$$

If g is monotonic, differentiable, and with differentiable inverse, we can write

$$F_Y(y) = P\left(g(X) \leq y\right) = P\left(X \leq g^{-1}(y)\right) = F_X\left(g^{-1}(y)\right)$$

Taking the derivative w.r.t. y:

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X\left(g^{-1}(y)\right) = \frac{d}{dg^{-1}(y)} F_X\left(g^{-1}(y)\right) \cdot \frac{d}{dy} g^{-1}(y) \\ &= f_X\left(g^{-1}(y)\right) \cdot \frac{d}{dy} g^{-1}(y) \end{aligned}$$

Transformations of random variables

In general, let $g : \mathcal{S}_X \longrightarrow \mathcal{S}_Y$ be invertible and differentiable, with differentiable inverse. Then

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right|$$

For example, let X be a R.V. with support $\mathbb{R}^+$, and $Y = X^3$. We have

$$g(x) = x^3, \quad g^{-1}(y) = \sqrt[3]{y} \quad \frac{d}{dy}g^{-1}(y) = \frac{1}{3}y^{-2/3}$$

The density of Y is thus

$$f_Y(y) = \frac{1}{3} f_X\left(\sqrt[3]{y}\right) y^{-2/3}$$

Transformations of random variables

We can extend the change of variables rule to continuous Random Vectors

Let $\boldsymbol{Y} = g(\boldsymbol{X})$, invertible, differentiable and with differentiable inverse

Then

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}\left(g^{-1}(\boldsymbol{y})\right)\left|\det \boldsymbol{D}g^{-1}(\boldsymbol{y})\right|$$

$\boldsymbol{D}g^{-1}(\boldsymbol{y})$ is the Jacobian matrix of g^{-1} (i.e. matrix of partial derivatives $\frac{\partial}{\partial y_j}\left[g^{-1}\right]_i(\boldsymbol{y})$

Transformations of random variables

In some cases we will consider the sum or R.V.s $Z = X + Y$

In this case, the c.d.f. is given by

$$F_Z(z) = P(Z \leq z) = P(X + Y \leq z) = \int_{\{(x,y)|x+y\leq z\}} f_{X,Y}(x,y)dxdy$$

and corresponds to the density

$$f_Z(z) = \int f_{X,Y}(x, z-x)dx = \int f_{X,Y}(z-y, y)dy$$

Transformations of random variables

The same result can be obtained from the conditional density of $Z|Y$:

$$(Z|Y=y) = (X|Y=y) + y$$

Applying the change of variable

$$f_{Z|Y}(z|y) = f_{X|Y}(z-y|y)$$

and finally

$$f_Z(z) = \int f_{Z|Y}(z|y) f_Y(y) dy = \int f_{X|Y}(z-y|y) f_Y(y) dy = \int f_{X,Y}(z-y,y) dy$$

Expectations

We define the mean or expected value of a R.V. as:[1]

- Discrete R.V.:

$$\mathbb{E}_X[X] = \sum_{x \in \mathcal{S}} x\, f_X(x)$$

- Continuous R.V.:

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

[1] Here and in the following definitions we assume that the integrals exist and are finite, otherwise the corresponding quantities are undefined. We will discuss the cases where this happens when needed.

Expectations

The variance of a R.V. is given by

$$\mathrm{var}(X) = \mathbb{E}_X\left[(X - \mathbb{E}_X[X])^2\right] = \mathbb{E}_X\left[X^2\right] - \mathbb{E}_X[X]^2$$

where $\mathbb{E}_X\left[X^2\right]$ is given by:

- Discrete R.V.:

$$\mathbb{E}_X\left[X^2\right] = \sum_{x \in \mathcal{S}} x^2 f_X(x)$$

- Continuous R.V.:

$$\mathbb{E}_X\left[X^2\right] = \int_{\mathcal{S}} x^2 f_X(x)\, dx$$

The variance is a measure of "spread" of the distribution around the mean

We define the standard deviation as $\mathrm{std}(X) = \sqrt{\mathrm{var}(X)}$

Expectations

We define the k-th moment of X as $\mathbb{E}_X\left[X^k\right]$

- Discrete R.V.:

$$\mathbb{E}_X\left[X^k\right] = \sum_{x \in \mathcal{S}} x^k \, f_X(x)$$

- Continuous R.V.:

$$\mathbb{E}_X\left[X^k\right] = \int_{\mathcal{S}} x^k \, f_X(x) \, dx$$

The central k-th moment is defined as $\mathbb{E}_X\left[\left(X - \mathbb{E}_X\left[X\right]\right)^k\right]$.

Expectations

In general, we define the expectation of function $g(X)$ as:

- Discrete R.V.:

$$\mathbb{E}_X\left[g(X)\right] = \sum_{x \in \mathcal{S}} g(x)\, f_X(x)$$

- Continuous R.V.:

$$\mathbb{E}_X\left[g(X)\right] = \int_{\mathcal{S}} g(x)\, f_X(x)\, dx$$

Expectations

For a Random Vector $\boldsymbol{X}$ we define the expectation as:

$$\mathbb{E}_{\boldsymbol{X}}[\boldsymbol{X}] = \begin{bmatrix} \mathbb{E}_{X_1}[X_1] \\ \vdots \\ \mathbb{E}_{X_m}[X_m] \end{bmatrix}$$

Since summation and integration are linear operators, also the expectation operator $\mathbb{E}[\cdot]$ is linear:

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b} \implies \mathbb{E}_{\boldsymbol{Y}}[\boldsymbol{Y}] = \mathbb{E}_{\boldsymbol{X}}[\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}] = \boldsymbol{A}\mathbb{E}_{\boldsymbol{X}}[\boldsymbol{X}] + \boldsymbol{b}$$

and

$$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{Z} \implies \mathbb{E}_{\boldsymbol{Y}}[\boldsymbol{Y}] = \mathbb{E}_{\boldsymbol{X}}[\boldsymbol{X}] + \mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{Z}]$$

In general, we will write just $\mathbb{E}$, unless the considered distribution is not clear.

Expectations

Let X, Y be two R.V.s

We define the covariance of X and Y as

$$\operatorname{cov}(X, Y) = \mathbb{E}_{X,Y}[XY] - \mathbb{E}_X[X]\mathbb{E}_Y[Y] = \operatorname{cov}(Y, X)$$

If $\operatorname{cov}(X, Y) = 0$ we say that X and Y are uncorrelated

$X \perp\!\!\!\perp Y \implies \operatorname{cov}(X, Y) = 0$ (but the opposite is not true in general!)

Expectations

We also define the Pearson correlation coefficient between X and Y as

$$\operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X) \operatorname{var}(Y)}}$$

We have $-1 \leq \operatorname{corr}(X, Y) \leq 1$

Also, $\operatorname{corr}(X, Y) = 1 \iff Y = aX + b$ for some $a \neq 0, b \in \mathbb{R}$

Expectations

For a Random Vector we define the covariance matrix

$$\boldsymbol{\Sigma} = \mathrm{cov}(\boldsymbol{X}) = \mathbb{E}\left[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^T\right]$$

$$= \begin{bmatrix} \mathrm{var}(\boldsymbol{X}_1) & \mathrm{cov}(\boldsymbol{X}_1, \boldsymbol{X}_2) & \cdots & \mathrm{cov}(\boldsymbol{X}_1, \boldsymbol{X}_m) \\ \mathrm{cov}(\boldsymbol{X}_1, \boldsymbol{X}_2) & \mathrm{var}(\boldsymbol{X}_2) & \cdots & \mathrm{cov}(\boldsymbol{X}_2, \boldsymbol{X}_m) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(\boldsymbol{X}_m, \boldsymbol{X}_1) & \mathrm{cov}(\boldsymbol{X}_m, \boldsymbol{X}_2) & \cdots & \mathrm{var}(\boldsymbol{X}_m) \end{bmatrix}$$

Expectations

As for the variance, we can express the covariance matrix as

$$\boldsymbol{\Sigma} = \operatorname{cov}(\boldsymbol{X}) = \mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^T\right] - \mathbb{E}\left[\boldsymbol{X}\right]\mathbb{E}\left[\boldsymbol{X}\right]^T$$

$\boldsymbol{\Sigma}$ is symmetric

$\boldsymbol{\Sigma}$ is positive semi-definite: all eigenvalues are $\lambda_i \geq 0$, and $\forall \boldsymbol{v} \in \mathbb{R}^m, \boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v} \geq 0$

If $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$, then

$$\operatorname{cov}(\boldsymbol{Y}) = \boldsymbol{A}\operatorname{cov}(\boldsymbol{X})\boldsymbol{A}^T$$

For 1-dimensional R.V. the covariance matrix becomes a scalar, and corresponds to the variance

Notation

In the following we will denote the density of a continuous R.V. as $f_X(x)$

For discrete R.V.s, we will also use $P_X(x)$ or $P(X = x)$

For conditional continuous p.d.f., we will use $f_{X|Y}(x|y)$, $f_{X|Y=y}(x)$, or $f_{X|y}(x)$ if there is on ambiguity on the conditioning R.V.

When densities depend on unknown values y, we use the notation $f_{X|y}(x|y)$ or $P_X(x|y)$ for those cases where y is not treated as a random value (note that there is no R.V. Y involved in the notation)

For conditional discrete densities, we may alternatively use $P_{X|Y}(x|y)$, $P_{X|Y=y}(x)$, $P(X = x|Y = y)$, or, if there is on ambiguity on the conditioning R.V., simply $P_{X|y}(x)$ and $P(X = x|y)$

When referring to known distributions, e.g. the normal distribution $\mathcal{N}$, we will use the same symbol both for the distribution and its density: $\mathcal{N}(\mu, \sigma^2)$ is a normal distribution, with density given by $\mathcal{N}(x|\mu, \sigma^2)$

Discrete R.V.: Bernoulli distribution

The Bernoulli distribution can be used to model the outcome of a binary event, e.g. tossing a single (potentially biased) coin that can result in a (H)ead or a (T)ail

Let $X \in \{0, 1\}$ denote the $R.V.$ that assigns value 1 (also called success) to head

The distribution of X is

$$X \sim \text{Ber}(p)$$

$$P_X(x) = \text{Ber}(x|p) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

$$= p^x (1 - p)^{1-x}$$

In many cases we will simply write that $P(H) = p$, implicitly assuming that symbol H has been mapped to 1

Discrete R.V.: Binomial distribution

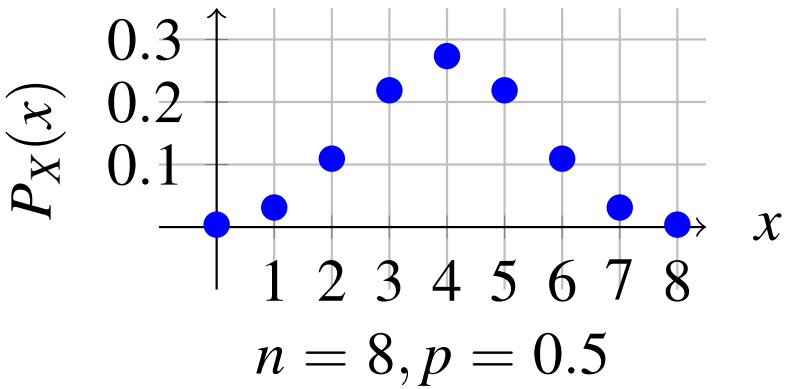
Suppose we want to count the number of successes in n repeated trials (e.g. the number of heads over n coin flips)

Let p denote the probability of success for a single trial

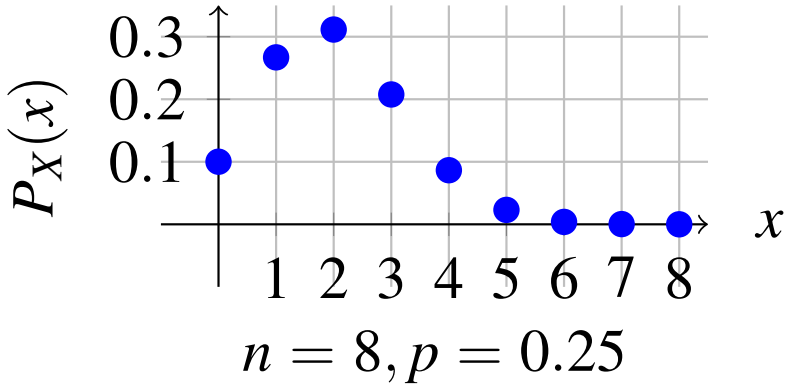
The distribution of X is the Binomial distribution

$$X \sim \text{Bin}(n, p)$$

$$P_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$$



$n = 8, p = 0.5$



$n = 8, p = 0.25$

Discrete R.V.: Binomial distribution

- $\text{Ber}(p) \sim \text{Bin}(1, p)$
- $X_1 \perp\!\!\!\perp \ldots \perp\!\!\!\perp X_n \sim \text{Ber}(p) \implies Y = \sum_i X_i \sim \text{Bin}(n, p)$
- $X \sim \text{Ber}(p) \implies \mathbb{E}[X] = p, \text{var}(X) = p(1-p)$
- $X \sim \text{Bin}(n, p) \implies \mathbb{E}[X] = np, \text{var}(X) = np(1-p)$

Discrete R.V.: Categorical and multinomial distribution

The Bernoulli and Binomial distributions can be extended to events that have K possible outcomes (e.g. rolling a die)

Categorical distribution: $X \in \{1, 2, \ldots K\}$[2]

$$X \sim \text{Cat}(\boldsymbol{p})$$

$$f_X(x) = P(X = x) = p_x = \prod_i p_i^{\mathbb{I}[x=i]}$$

where $\boldsymbol{p} = (p_1, \ldots p_K)$, with $\sum_{i=1}^{K} p_i = 1$. p_i is the probability of outcome i, and $\mathbb{I}$ is the indicator function

$$\mathbb{I}[C] = \begin{cases} 1 & \text{if } C \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

[2]The actual labels are in many practical cases irrelevant. For example, if we want to use the same encoding as for the Bernoulli distribution, we can assume $X \in \{0, \ldots, K-1\}$

Discrete R.V.: Categorical and multinomial distribution

In many cases it's convenient to represent outcomes with a 1-of-K encoding (vector):

$$
\begin{aligned}
X = 1 &\implies \boldsymbol{X} = (1, 0, \ldots, 0) \\
X = 2 &\implies \boldsymbol{X} = (0, 1, \ldots, 0) \\
&\ldots \\
X = K &\implies \boldsymbol{X} = (0, 0, \ldots, 1)
\end{aligned}
$$

The density can then be expressed as

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_i p_i^{x_i}$$

In the following we will adopt this approach

Discrete R.V.: Categorical and multinomial distribution

We can consider a set of n trials, encoded as $\boldsymbol{x} = (x_1 \dots x_K)$, where x_i denotes the number of occurrencies of outcome i, and $n = \sum_{i=1}^{m} x_i$.

Let $\boldsymbol{p} = (p_1, \dots p_K)$ be the vector of probabilities for a single trial

$\boldsymbol{X}$ follows a Multinomial distribution

$$\boldsymbol{X} \sim \text{Mul}(n, \boldsymbol{p})$$

$$f_X(\boldsymbol{x}) = \frac{n!}{x_1! \dots x_K!} \prod_{i=1}^{K} p_i^{x_i}$$

- $\text{Mul}(1, \boldsymbol{p}) \sim \text{Cat}(\boldsymbol{p})$
- $\boldsymbol{X}_1 \perp\!\!\!\perp \dots \perp\!\!\!\perp \boldsymbol{X}_n \sim Cat(\boldsymbol{p}) \implies \boldsymbol{Y} = \sum_{i=1}^{n} \boldsymbol{X}_i \sim \text{Mul}(n, \boldsymbol{p})$

Continuous R.V.: Gaussian distribution

The Gaussian or normal distribution is probably the most employed example of continuous distributions

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

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

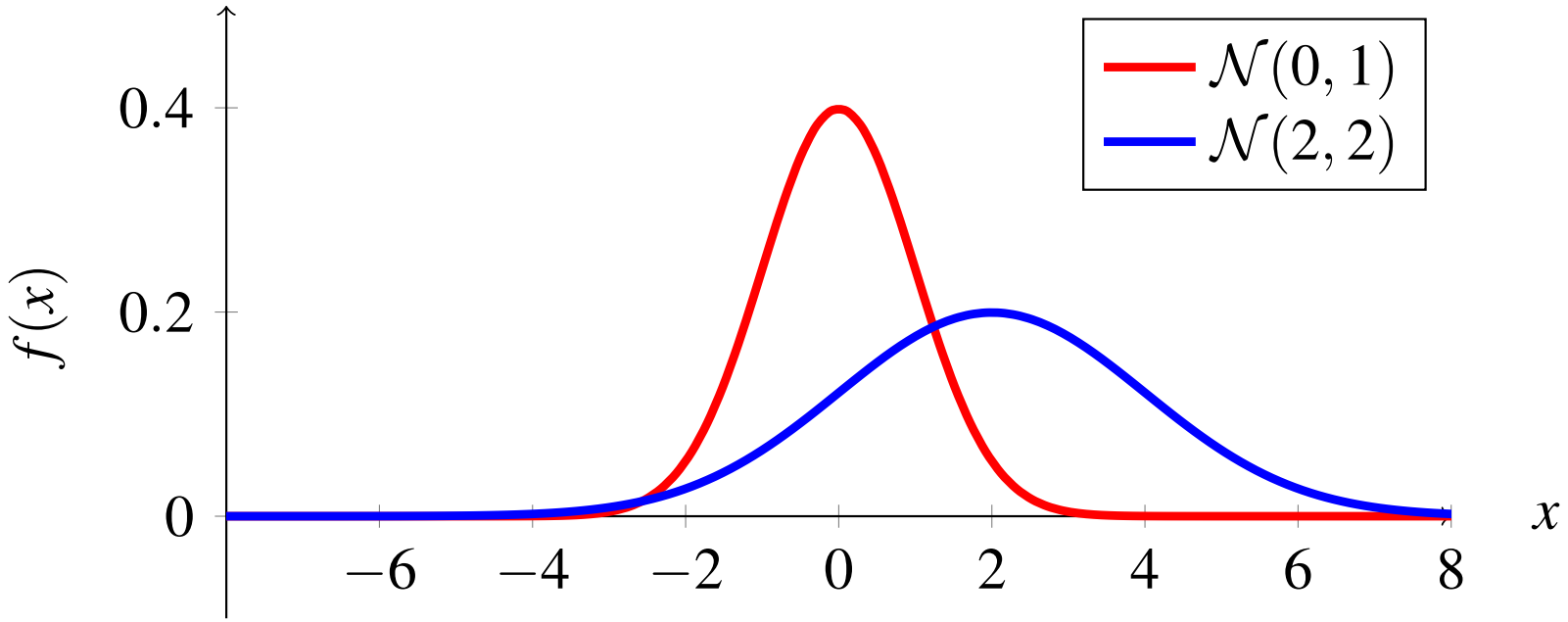
- μ is the mean, and as the name suggests $\mathbb{E}[X] = \mu$
- σ is called standard deviation, and $\mathrm{var}(X) = \sigma^2$
- The inverse of the variance $\lambda = \frac{1}{\sigma^2}$ is called precision

if $X \sim \mathcal{N}(0, 1)$ we say that X follows a standard normal distribution

Continuous R.V.: Gaussian distribution

The distribution is symmetric, centered around μ

Higher variance corresponds to a "flatter" density



Continuous R.V.: Gaussian distribution

Theorem (Central limit)

Let $X_1, \ldots X_N$ be a sequence of independent and identically distributed (i.i.d.) random variables, with mean μ and variance $\sigma^2 > 0$. Let $S_N = \sum_{i=1}^{N} X_i$. Then

$$\frac{S_N - N\mu}{\sigma\sqrt{N}}$$

converges in distribution to $\mathcal{N}(0, 1)$

Convergence in distribution means that

$$\lim_{n\to\infty} F_{S_N}(x) = F_{\mathcal{N}(0,1)}(x)$$

for each x

Multivariate Gaussian Distribution

We can extend the Gaussian distribution to random vectors

Let $\boldsymbol{X}$ be a random vector $\boldsymbol{X} = [X_1, \ldots, X_N]^T$ where X_i are i.i.d., standard normal distributed R.V.s $X_i \sim \mathcal{N}(0, 1)$

The distribution of $\boldsymbol{X}$ is given by the joint distribution of $X_1, \ldots, X_N$

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{N} f_{X_i}(x_i)$$

or, equivalently,

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = (2\pi)^{-\frac{N}{2}} e^{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}}$$

We say that $\boldsymbol{X}$ follows a standard multivariate normal distribution

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$$

Multivariate Gaussian Distribution

$\boldsymbol{X}$ follows a multivariate Gaussian (MVG) distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ if $\boldsymbol{X}$ can be written as a linear transformation of a standard multivariate normal distributed random vector $\boldsymbol{Y}$:

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Y} + \boldsymbol{\mu}$$

where $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T$

We will write $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

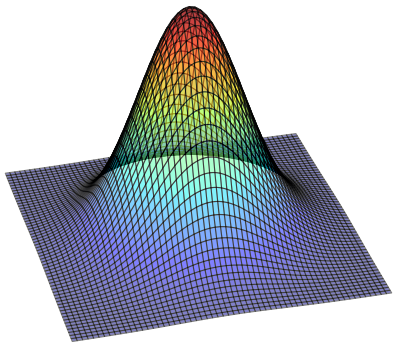
The p.d.f. of $\boldsymbol{X}$ is given by

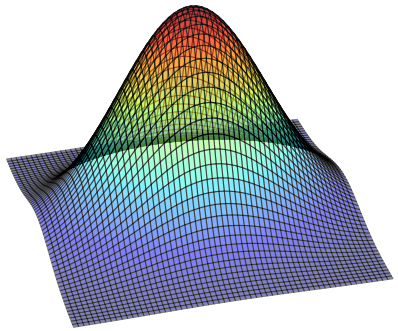
$$f_{\boldsymbol{X}}(\boldsymbol{x}) = (2\pi)^{-\frac{N}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

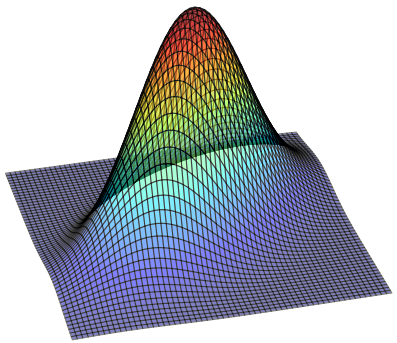
Often, rather than working with a covariance matrix, it's easier to work with its inverse, the precision matrix $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$

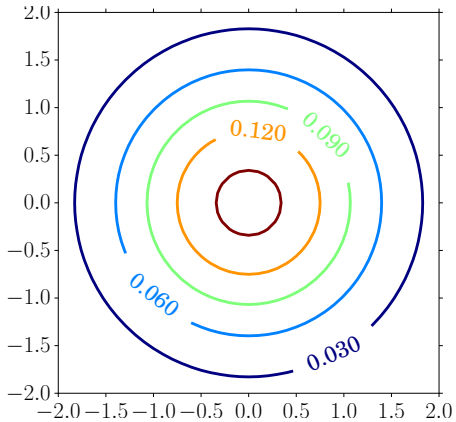
Multivariate Gaussian Distribution

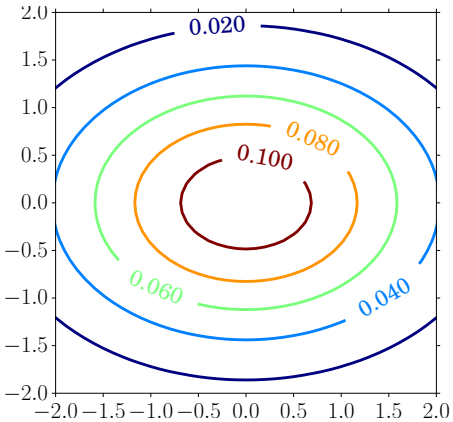
$$\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \quad \boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \begin{bmatrix} 1.5 & 0.7 \\ 0.7 & 1 \end{bmatrix}$$

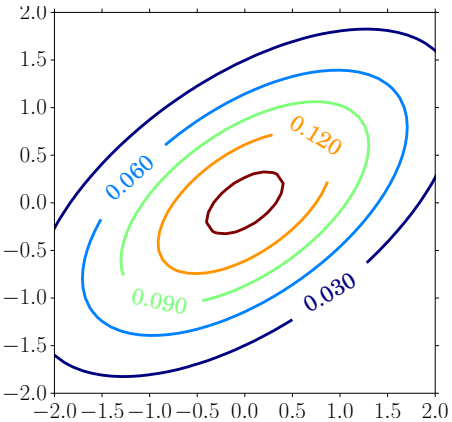












Density estimation

We start considering a simple example

We would like to predict whether a coin flip will result in a head (H) or tail (T)

We do not know whether the coin is biased or not

However, we have observed a number of tosses n

Density estimation

Let's denote the observed results as $(x_1 \dots x_n)$

For a head toss, we have $x_i = 1$, while we represent a tail as $x_i = 0$

These can be considered outcomes of R.V.s $(X_1 \dots X_n)$

Our goal is to predict the probability that a new toss X_t will result in head

We want to model the distribution $X_t | X_1 = x_1 \dots X_n = x_n$

Density estimation

Let's, for a moment, assume that we knew the probability that the coin lands a head, $P(H) = \pi$

We also assume that the coin tosses are independent and identically distributed

Thus we can model the R.V.s X_i as Bernoulli R.V.s with parameter π:

$$X_i \sim X \sim \text{Ber}(\pi)$$

Since we know the probability of head π, the R.V. X_t is independent of $X_1 \dots X_n$ and is also Bernoulli distributed:

$$X_t \sim X \sim \text{Ber}(\pi)$$

Density estimation

Unfortunately, we do not know π

The first approach that we follow, the frequentist approach, assumes that there exists a "true" value π_T that explains the observed data and can be used to predict new values

Since we do not know this true value, we have to estimate it

A possible way to estimate a "good" value for π consists in looking for the π that best explains the observed tosses $x_1 \dots x_n$

We thus define the likelihood function

$$\mathcal{L}(\pi) = f_{X_1 \dots X_n|\pi}(x_1 \dots x_n|\pi) = P(X_1 = x_1 \dots X_n = x_n|\pi)$$

Note that π is not treated as a random value, but only as an unknown value

Density estimation

We assume again that, given the value of π, the tosses are independent:

$$P(X_1 = x_1 \dots X_n = x_n|\pi) = \prod_{i=1}^{n} P(X_i = x_i|\pi)$$

and since

$$P(X_i = x_i|\pi) = \text{Ber}(x_i|\pi) = \pi^{x_i}(1-\pi)^{1-x_i}$$

the likelihood becomes

$$\mathcal{L}(\pi) = \prod_{i=1}^{n} \pi^{x_i}(1-\pi)^{1-x_i}$$

Density estimation

We can compute the Maximum Likelihood estimate (ML) for π as the value that maximizes the likelihood:

$$\pi^*_{ML} = \arg\max_{\pi} \mathcal{L}(\pi) = \arg\max_{\pi} \prod_{i=1}^{n} \pi^{x_i}(1-\pi)^{1-x_i}$$

π^* is the parameter π which maximizes the probability of the observed sequence of toss

To solve for π, we consider the logarithm of the likelihood[3]

$$\ell(\pi) = \log \mathcal{L}(\pi)$$

Since the logarithm is monotonically increasing, maximizing ℓ is equivalent to maximizing $\mathcal{L}$

[3] We will often work with logarithms of densities, for two reasons: (i) many expressions will be simpler in the log domain, (ii) as we will see in the laboratories, working directly with densities often leads to numerical issues

The expression for $\ell(\pi)$ is

$$\ell(\pi) = \log \mathcal{L}(\pi) = \sum_{i=1}^{n} [x_i \log \pi + (1 - x_i) \log(1 - \pi)]$$

We set the derivative of $\ell(\pi)$ equal to zero:

$$\frac{d\ell}{d\pi} = \sum_{i=1}^{n} \left[\frac{x_i}{\pi} - \frac{(1 - x_i)}{1 - \pi} \right] = 0$$

and we obtain

$$\pi^*_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Density estimation

We can finally predict the outcome for a new toss:

$$P(X_t = 1 | X_1 = x_1 \dots X_n = x_n) = \pi^*_{ML}$$

We have "condensed" the knowledge of the observed outcomes in the estimate π^*_{ML} (which is a function of $x_1 \dots x_n$)

Given enough observations, our estimation will often be "good".

Consider, however, a simple scenario where our observations are just 3 heads. The ML estimate would be $\pi^*_{ML} = 1$, i.e., we predict that the coin will never land a tail

The issue raises from the fact that we use a single value (the estimate π^*_{ML} of π_T), and we do not consider other possible, even if less likely, values — we are neglecting our uncertainty over the estimated value

Density estimation

The Bayesian approach addresses this issue considering the unknown parameters as random values, that can be described in terms of probability distributions

In contrast with the frequentist approach, we do not require the notion of a "true" value for π

What we need is to specify a distribution for π

The Bayesian approach thus requires us to specify our knowledge about the possible values that π may assume in terms of a prior distribution $f_{\Pi}(\pi)$

The prior distribution reflects our "knowledge of the world" before we observed the data

Density estimation

We can combine the likelihood and the prior to compute the posterior distribution for the model parameters

The posterior distribution is the distribution for the model parameters given the observed values

It can be computed from Bayes rule:

$$f_{\Pi|X_1 \ldots X_n}(\pi|x_1 \ldots x_n) = \frac{P(X_1 = x_1 \ldots X_n = x_n|\pi) f_\Pi(\pi)}{P(X_1 = x_1 \ldots X_n = x_n)}$$
$$= \frac{P(X_1 = x_1 \ldots X_n = x_n|\pi) f_\Pi(\pi)}{\int P(X_1 = x_1 \ldots X_n = x_n|\pi) f_\Pi(\pi) d\pi}$$

The posterior distribution reflects our knowledge of the parameters once we have observed the data

It combines both our prior information and the knowledge provided by the likelihood

Density estimation

Assuming again that observations are independent, given the model parameter π, the posterior distribution allows computing the predictive distribution as

$$
\begin{aligned}
&P(X_t = x_t | X_1 = x_1 \dots X_n = x_n) \\
&\quad = \int P(X_t, \Pi = \pi | X_1 = x_1 \dots X_n = x_n) d\pi \\
&\quad = \int P(X_t | \Pi = \pi, X_1 = x_1 \dots X_n = x_n) f_{\Pi | X_1 \dots X_n}(\pi | x_1 \dots x_n) d\pi \\
&\quad = \int P(X_t | \Pi = \pi) f_{\Pi | X_1 \dots X_n}(\pi | x_1 \dots x_n) d\pi
\end{aligned}
$$

i.e., by marginalization of the joint distribution for x_t, π given the observed samples

Density estimation

The Bayesian approach has two main limitations

- The choice of the prior is somewhat arbitrary
- The computation of the posterior may be intractable in practice

On the other hand, Bayesian models tend to provide better results for scenarios where data is scarce, since they can account for parameter uncertainty

Density estimation

When modeling continuous values, Gaussian distributions arise naturally in a wide variety of contexts

For example, we have seen that the distribution of sums of i.i.d. R.V.s converges to Gaussian distributions

In many cases we observe that data histogram present Gaussian-like shapes

Furthermore, the Gaussian density is easy to work with

For these reasons, in many cases it's reasonable to assume that our data have been generated by Gaussian R.V.s

Density estimation

Let's assume we have some data $\mathcal{D} = (x_1, \ldots, x_n)$

We decide to model the data as samples of a Gaussian distribution, with mean μ and variance v (and precision $\lambda = v^{-1}$)

We assume that the points have been generated by independent, identically distributed R.V.s $X_i \sim X$

Given the values of the model parameter $\theta = (\mu, v)$, the distribution of X is

$$f_{X|\theta}(x) = \mathcal{N}(x|\mu, v)$$

As for the discrete case, we can express the likelihood for θ as

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f_{X_i|\theta}(x_i) = \prod_{i=1}^{n} \mathcal{N}(x_i|\mu, v)$$

Density estimation

We again consider a frequentist approach, which assumes the existence of "true", but unknown values, for the model parameters $\theta = (\mu, v)$

If we knew the values of these parameters μ_T, v_T, then the density for unseen samples X_t would be

$$f_{X_t}(x_t) = \mathcal{N}(x_t|\mu_T, v_T)$$

Since we don't not know the model parameters, we again need a way to estimate them

Density estimation

As for the discrete case, we want to find an estimator of μ and v, possibly close to the true values μ_T, v_T

We have already seen an example of estimator, the Maximum Likelihood estimator (we will show the ML solution for Gaussians in a moment)

In general, an estimator is a function T of the data, that maps our dataset $\mathcal{D}$ to values for the model parameters θ^*

$$\theta^* = T(\mathcal{D})$$

Density estimation

We would like estimators that are consistent, i.e. that converge (in probability) to the "true" distribution parameter θ_T as the sample size n grows to infinity

A simple way to produce (under mild assumptions) consistent estimators consists in matching the moments of the assumed distribution to those of the data

This method is called method of moments (MOM)

For the Gaussian distribution, the first two moments are μ (first order moment) and v (centered second order moment)

Density estimation

We can write the equations that match the moments to the empirical mean and covariance of the data, obtaining

$$\mu^*_{MOM} = \frac{1}{n}\sum_i x_i$$

$$v^*_{MOM} = \frac{1}{n}\sum_i (x_i - \mu^*_{MOM})^2$$

The MOM produces, in this case, consistent estimators

Density estimation

The MOM approach does not, in general, produce very accurate estimators

We have already encountered another estimator, that is widely used in practice: the Maximum Likelihood (ML) estimator

The Maximum Likelihood estimator is the value that maximizes the likelihood

$$\theta^*_{ML} = \arg\max_{\theta} \mathcal{L}(\theta)$$

We can regard the ML solution as the parameter that best explains the observed dataset $\mathcal{D}$, i.e. the value for which it's most likely that we observe $\mathcal{D}$.

Density estimation

We can derive the ML estimator for the Gaussian distribution

Very often it's more practical to work with the logarithm of the likelihood[4]

$$\ell(\theta) = \log \mathcal{L}(\theta)$$

Since the logarithm is monotonically increasing, maximizing ℓ is equivalent to maximizing $\mathcal{L}$:

$$\arg\max_{\theta} \ell(\theta) = \arg\max_{\theta} \mathcal{L}(\theta)$$

[4] We will shortly see that the log-pdf of the Gaussian has a simple expression. Furthermore, product of densities transform to sum of log-densities. Also, as we will see in the laboratory, working with densities rather than log-densities often results in numerical problems

Density estimation

Since we assumed that X_i are independent, and $X_i \sim X$, then

$$f_{X_1 \ldots X_n}(x_1 \ldots x_n|\theta) = \prod_{i=1}^{n} f_{X_i}(x_i|\theta) = \prod_{i=1}^{n} f_X(x_i|\theta)$$

The log-likelihood is then

$$\ell(\theta) = \log \mathcal{L}(\theta) = \log \prod_{i=1}^{n} f_X(x_i|\theta) = \sum_{i=1}^{n} \log f_X(x_i|\theta)$$

Plugging in the Gaussian density of $X \sim \mathcal{N}(\mu, v)$:

$$\ell(\theta) = \sum_{i=1}^{n} \log \mathcal{N}(x|\mu, v) = \sum_{i=1}^{n} \log \mathcal{N}(x|\mu, \lambda^{-1})$$

In the following we parametrize the density in terms of precision λ, since this allows for simpler expressions

Density estimation

In the log domain the Gaussian log-pdf has a simple expression:

$$\log \mathcal{N}(x|\mu, \lambda^{-1}) = \xi_1 + \frac{1}{2}\log\lambda - \frac{\lambda}{2}(x-\mu)^2$$

where ξ_1 collects constant terms that are irrelevant for the optimization (they do not depend on the parameters)

The log-likelihood is then

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^{n} \log \mathcal{N}(x|\mu, \lambda^{-1}) = \xi_2 + \sum_{i=1}^{n}\left[\frac{1}{2}\log\lambda - \frac{1}{2}\lambda(x_i-\mu)^2\right] \\ &= \xi_2 + \frac{1}{2}n\log\lambda - \frac{1}{2}\lambda\sum_{i=1}^{n}x_i^2 - \lambda\mu\sum_{i=1}^{n}x_i + \frac{1}{2}n\lambda\mu^2\end{aligned}$$

and ξ_2 collects all constant terms

Density estimation

The ML estimate can be obtained by taking solving for

$$\begin{cases} \frac{\partial \ell}{\partial \mu} = 0 \\ \frac{\partial \ell}{\partial \lambda} = 0 \end{cases}$$

The first derivative is

$$\frac{\partial \ell}{\partial \mu} = n\lambda\mu - \lambda \sum_{i=1}^{n} x_i$$

thus

$$\mu^*_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Density estimation

The derivative with respect to the precision is

$$\frac{\partial \ell}{\partial \lambda} = \frac{n}{2\lambda} - \frac{1}{2}\left[\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

thus

$$v_{ML}^* = (\lambda_{ML}^*)^{-1} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{ML}^*)^2$$

The solution thus corresponds to the empirical mean and empirical covariance matrix of the data

In this case, we can observe that the solution is the same as the one obtained by the MOM approach (this does not hold in general)

Density estimation

As in the binary case, we can use the ML estimates to form the predictive distribution

$$f_{X_t|X_1 \ldots X_n}(x_t|x_1 \ldots x_n) \approx \mathcal{N}\left(x_t \,|\mu^*_{ML}, v^*_{ML}\right)$$