Intelligent distributed systems

Daniele Fontanelli

Department of Industrial Engineering
University of Trento
E-mail address: *daniele.fontanelli@unitn.it*

2022/2023

Outline

1. Background on Statistics
   - Probability
   - Random variables
   - Multivariate Pdfs
   - Propagation of errors
   - Take home message

Outline

Outline

Probability

Axioms of probability

Consider an *experiment* with random outcomes $\omega_i \in \Omega$, where Ω is the *sample set*.

Definition (Event)

An *event* is a set of the outcomes of the experiment. The *event occurs* if the outcome is one of the element of the set.

Definition (Probability (Kolmogorov))

The *probability* $\Pr[\cdot]$ of an event A is a *measure* that satisfies the *three axioms of probability*:

(1) $\Pr[A] \geq 0$;

(2) $\Pr[S] = 1$ if S is the *sure event*;

(3) If $A \cap B = \emptyset$, then $\Pr[A \cup B] = \Pr[A + B] = \Pr[A] + \Pr[B]$, i.e., additivity with respect to *mutually exclusive events*.

Probability

Axioms of probability

As a consequence, $\Pr\left[\bar{A}\right] = 1 - \Pr\left[A\right] \leq 1$, where $\bar{A}$ is the *complementary event*.

Moreover, the *impossible event* is the complementary of S, i.e., $\Pr\left[\emptyset\right] = 0$.

Remark

A possible event is a subset of S.

Remark

The additivity of mutually exclusive events, i.e., countable additivity, *is necessary when the experiments has an infinite number of outcomes.*

Probability

Frequentist vs Bayesian

There are usually two different views of probability:

- The *frequentist* interpretation: the probability of something corresponds to what fraction of the time it happens “in the long run”. Although engineering sound, there are some systems in which this approach fails. For example, it is not clear how to apply the frequency approach to the Earth’s gravitational field since there is only one field;
- The *Bayesian* interpretation: the probability of something corresponds to how likely we “think” it is to happen, aka *measure-of-belief*.

Probability

Measure of belief

The *relative-frequency* concept is a *non-rigorous* way to introduce the concept of probability, which are widely used in engineering systems. For example, we can explain the probability of each face of the dice by means of empirical measures.

Probability

Conditional probability: Axioms of probability

For two events A and B we have

$$\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B],$$

where $\Pr[A \cap B]$ is the *joint probability*.

Probability

Conditional probability: Axioms of probability

Using the axioms of probability, we can define the *conditional probability* as

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]}.$$

Indeed, $\Pr[A|B]$ satisfies Axiom 1, since $\Pr[A \cap B] \geq 0$ and $\Pr[B] > 0$ (since it happened).
Since $\Pr[S|A] = \Pr[A] / \Pr[A] = 1$, it also satisfies Axiom 2.
Finally, if $A_1, A_2, \ldots$ are *pair-wise mutually exclusive* events, one has

$$\Pr[A_1 \cup A_2 \cup \ldots |B] = \frac{\Pr[(A_1 \cup A_2 \cup \ldots) \cap B]}{\Pr[B]} = \frac{\Pr[(A_1 \cap B) \cup (A_2 \cap B) \cup \ldots)]}{\Pr[B]} =$$
$$= \frac{\Pr[(A_1 \cap B)]}{\Pr[B]} + \frac{\Pr[(A_2 \cap B)]}{\Pr[B]} + \cdots = \Pr[A_1|B] + \Pr[A_2|B] + \ldots$$

which satisfies Axiom 3.

Probability

Conditional probability

In general, since

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]} \text{ and } \Pr[B|A] = \frac{\Pr[A \cap B]}{\Pr[A]}$$

one has

$$\Pr[A \cap B] = \Pr[A|B]\Pr[B] = \Pr[B|A]\Pr[A].$$

As a consequence, if A and B are *statistically independent*, $\Pr[A|B] = \Pr[A]$ and hence

$$\Pr[A \cap B] = \Pr[A]\Pr[B].$$

Probability

Total Probability Law

n *disjoint* (*mutually exclusive*) events A_i, $i = 1, \dots, n$, satisfies

$$A_j \cap A_i = \emptyset, \forall i, j$$

Moreover, if

$$\cup_{i=1}^{n} A_i = \Omega$$

we have that for any event B

$$\Pr[B] = \sum_{j=1}^{n} \Pr[B \cap A_j] = \sum_{j=1}^{n} \Pr[B|A_j] \Pr[A_j],$$

which is the *Total Probability Law*.

Probability

Bayes Theorem

For the conditional probability, we can rewrite

$$\Pr[A \cap B] = \Pr[A|B]\Pr[B] = \Pr[B|A]\Pr[A],$$

as

$$\Pr[A|B] = \frac{\Pr[B|A]\Pr[A]}{\Pr[B]}.$$

Generalising to n *disjoint* events A_i, $i = 1, \ldots, n$, we have

$$\Pr[A_i|B] = \frac{\Pr[B|A_i]\Pr[A_i]}{\Pr[B]} = \frac{\Pr[B|A_i]\Pr[A_i]}{\sum_{j=1}^{n}\Pr[B|A_j]\Pr[A_j]},$$

which is the *Bayes Theorem*.

Probability

Bayesian inference

The *Bayes Theorem*

$$\Pr[A_i|B] = \frac{\Pr[A_i]\Pr[B|A_i]}{\Pr[B]} = \frac{\Pr[A_i]\Pr[B|A_i]}{\sum_{j=1}^{n}\Pr[B|A_j]\Pr[A_j]},$$

has three different actors playing a crucial role:

- $\Pr[A_i]$ is the *prior probability* of A_i (the probability of A_i known upfront);
- $\Pr[B|A_i]$ is the *likelihood* of B given A_i (how likely B happens when A_i happens);
- $\Pr[A_i|B]$ is the *posterior probability* of A_i given B (the probability of A_i knowing that B has actually happened).

This is the basis of the *Bayesian inference*.

Example: Bayes Theorem

Is Joe sick?

Example: Bayes Theorem

Joe carries out a set of health tests and it turns out that he results positive to a test about a very rare and bad disease affecting 0.1% of the population.
The test *correctly* identifies 99% of the population having the disease and only 1% *wrongly* identifies persons who don't have the disease.
So, what is the probability that Joe actually has the disease?

Example: Bayes Theorem

If you answered 99%, you are *wrong*!
Let us define the event "Joe has the disease" with J, and with T "The test is positive".
It is now clear that $\Pr[T|J] = 99\%$!
Instead we are interested in computing $\Pr[J|T]$!
So, to figure out this probability we need *Bayes theorem*:

$$\Pr[J|T] = \frac{\Pr[T|J]\Pr[J]}{\Pr[T]} = \frac{\Pr[T|J]\Pr[J]}{\Pr[T|J]\Pr[J] + \Pr\left[T|\overline{J}\right]\Pr\left[\overline{J}\right]}.$$

The prior $\Pr[J]$ is the most difficult probability to figure out. However, you can assume that $\Pr[J] = 0.1\%$.

Example: Bayes Theorem

Plugging all the numbers, there are good news for Joe, since we have:

$$\Pr[J|T] = \frac{\Pr[T|J]\Pr[J]}{\Pr[T|J]\Pr[J] + \Pr\left[T|\overline{J}\right]\Pr\left[\overline{J}\right]} = 9\%,$$

which is quite below the 99%!
If you think about it is not strange at all. Indeed, only 1 person of 1000 may have the disease, however the test executed on the entire population is wrongly positive for approximately 10 persons: 9% is about 1 person out of the 11 eleven persons having the test positive!

Example: Bayes Theorem

But in the *Bayes theorem* there is even more: *the knowledge increases as more evidence is accumulated*.

This was the basic idea of Bayes, whose main result came out *by trowing balls on a table*...

The mathematician and philosopher *Richard Price*, who edited and published in *Philosophical Transactions* the work of *Thomas Bayes* in "An Essay towards solving a Problem in the Doctrine of Chances" (1763), introduced the *Bayes Philosophy* with an analogy: *the man coming out from the cave*.

How this analogy maps onto the Joe's health conditions?

Example: Bayes Theorem

When Joe repeated the test in a different laboratory (to have *independency* among the events) he got another positive result (sigh!).
What is now $\Pr[J|T]$?
Of course we use the same equations, but with a different prior, which is now $\Pr[J] = 9\%$, i.e. the *posterior of the previous test*.
Hence:

$$\Pr[J|T] = \frac{\Pr[T|J]\Pr[J]}{\Pr[T|J]\Pr[J] + \Pr\left[T|\overline{J}\right]\Pr\left[\overline{J}\right]} = 91\%,$$

which is quite high now, even though still less than 99%.
Notice that we have considered only the *false positives*, i.e. $\Pr\left[T|\bar{J}\right] = 1\%$, but we can also consider the *false negatives*, i.e. $\Pr\left[\bar{T}|J\right]$, and hence infer $\Pr\left[J|\bar{T}\right]$ using the same approach!

Probability

Bayesian inference: A historic view

This concept of *degree of belief* has been then rediscovered by *Laplace* in 1812. Laplace adopted the *Bayesian philosophy* in combination with observations and the law of mechanics to compute *the mass of Saturn*. Incredibly, after 150 years his results were changed less than 1%!

Outline

Background on statistics

Random variables

In the majority of the engineering cases the variables are not described by sets but by *numbers*

Definition (Random variable)

A *random variable* is a (real-valued) function that assumes a certain *value* according to the outcome of a certain *random experiment*.

A random variable is *not* a *conventional variable*. For example, the number we got *after* rolling a dice is a variable, but the outcome we *expect* from the rolling is a random variable (**rv**).

Background on statistics

Random variables

Formally, a **rv** maps the set containing all the possible outcomes of a certain experiment, i.e., the *event space* or *sample space* Ω, to the space of numbers.

If the mapping is onto some continuous set (e.g., the $\mathbb{R}$ set), the **rv** is *continuous*.

If the mapping is onto some discrete set (e.g., the $\mathbb{Z}$ set), the **rv** is *discrete*.

In practice $\forall \omega \in \Omega$, we get that $x(\omega)$ is an element of the arriving set (e.g., a number).

Background on statistics

Random variables

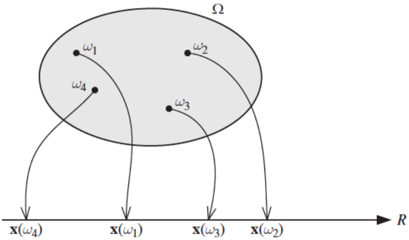


Figure: A random variable can be seen as a mapping from a sample space Ω to a continuous (discrete) set, e.g., the set $\mathbb{R}$ ($\mathbb{Z}$) of real (integer) numbers (Courtesy of Ha H. Nguyen Ed Shwedyk - A First Course in Digital Communications, Cambridge University Press, 2009).

Background on statistics

Random variables, pdfs and pmfs

To be completely described a **rv** needs a *probabilistic description*, that is given in terms of the *probability mass function*.

Definition (Probability mass function)

The *probability mass function* (pmf) of a *discrete-valued* random variable x taking values in the set $\{a_i, i = 1, \dots, n\}$ is

$$\pi(a_i) = \mathsf{Pr}\left[x = a_i\right] = \pi_i,$$

where π_i are the *point masses*. In order to be *proper* and satisfy the Kolmogorov axioms,

$$\sum_{i=1}^{n} \pi_i = 1.$$

Background on statistics

Random variables, pdfs and pmfs

For a *continuous* **rv**, the *probabilistic description* is given in terms of the *probability density function*.

Definition (Probability density function)

The *probability density function* (pdf) of a scalar random variable x at a certain value $x = a$ is

$$p(a) = \lim_{\delta_a \to 0} \frac{\mathsf{Pr}\left[a - \delta_a < x \leq a\right]}{\delta_a} \geq 0.$$

Notice the similarity with the *ratio of the increments* as the approximation of the derivative.

Background on statistics

Random variables, pdfs and pmfs

Since a **rv** maps disjoint events into different portions of the real axis, the values taken by the **rv** are *mutually exclusive*.

Example

Rolling a fair dice. Hence, $\Omega = \{1, 2, 3, 4, 5, 6\}$ and the pmf is $\pi(a_i) = \frac{1}{6}$, $\forall a_i \in \Omega$.
Consider $A_1 = \{x \leq 3\}$. So, being mutually exclusive, by the third axiom of probability

$$\text{Pr}\left[A_1\right] = \text{Pr}\left[x = a_1 \cup x = a_2 \cup x = a_3\right] = \sum_{i=1}^{3} \text{Pr}\left[x = a_i\right]\frac{1}{2}.$$

Background on statistics

Random variables, pdfs and pmfs

The results of the previous example can be extended to pdfs and continuous **rv**s, i.e.

$$\Pr\left[a < x \leq b\right] = \int_a^b p(x)dx,$$

we have

Definition (Cumulative distribution function)

The *cumulative distribution function* (cdf) of x at b is given by

$$P(b) = \Pr\left[x \leq b\right] = \int_{-\infty}^b p(x)dx.$$

Therefore, in order to be a proper pdf, we have for the second axiom

$$\Pr\left[x \leq \infty\right] = \int_{-\infty}^{\infty} p(x)dx = 1.$$

Background on statistics

Random variables, pdfs and pmfs

Notice how

$$p(x) = \left.\frac{dP(a)}{da}\right|_{a=x}$$

Moreover,

$$\mathsf{Pr}\left[a < x \leq a + \delta_a\right] = \int_a^{a+\delta_a} p(x)dx,$$

that for sufficiently small δ_a means that

$$\mathsf{Pr}\left[a < x \leq a + \delta_a\right] = p(a)\delta_a + O(\delta_a^2),$$

so $p(x)$ represents the *density of probability* per unit value of a in the neighbourhood of a.

Background on statistics

Random variables, pdfs and pmfs

Hence, $p(x) \geq 0$, $\forall x$ (first axiom of probability).
Furthermore, for a continuous **rv**

$$\mathsf{Pr}\left[x = a\right] = \int_a^a p(x)dx = 0.$$

Background on statistics

Random variables, pdfs and pmfs

For the *cumulative distribution function*, we have that

$$\Pr\left[a < x \leq a + \delta_a\right] = P(a + \delta_a) - P(a),$$

which is verified by

$$P(a) = \Pr\left[x \leq a\right] = \int_{-\infty}^{a} p(x)dx.$$

The cdf has the following properties:

- $0 \leq P(a) \leq 1$;
- $\lim_{a \to -\infty} P(a) = 0$ and $\lim_{a \to +\infty} P(a) = 1$;
- $P(\cdot)$ is non-decreasing;
- $P(\cdot)$ is right continuous, i.e., $\lim_{\delta_a \to 0^+} P(a + \delta_a) = P(a)$.

Background on statistics

Random variables, pdfs and pmfs

Usually, all the pdfs of continuous **rv**s have the following form:

$$p(x) = \frac{1}{cN(s)} p\left(\frac{x-l}{c}\right),$$

where:

- l is a *location parameter*, that has the role to translate the pdf;
- c is a *scale parameter*, having the role of expanding or contracting the pdf. For normalisation purposes, its reciprocal multiplies the pdf;
- s is a *shape parameter*, that governs the shape of the pdf, and it is also an argument of the normalising function $N(s)$.

Let us now consider a couple of examples...

Background on statistics

Examples of pdfs: Uniform

$x \sim \mathcal{U}(a, b)$ defines x as a *uniformly distributed random variable* in the interval $[a, b]$.
The *pdf* is given by

$$p(x) = \mathcal{U}(x; a, b) \triangleq \begin{cases} \dfrac{1}{b-a} & \text{if } x \in [a, b], \\ 0 & \text{otherwise.} \end{cases}$$

The *cdf* is instead given by

$$P(c) = \begin{cases} 0 & \text{if } c < a, \\ \dfrac{c-a}{b-a} & \text{if } c \in [a, b], \\ 1 & \text{if } c > b. \end{cases}$$

Background on statistics

Examples of pdfs: Gaussian or Normal

$x \sim \mathcal{N}(l, c, N(s))$ defines x as a *Gaussian (or normal) random variable* with parameters $N(s)$, l and c.
The *pdf* is given by

$$p(x) = \mathcal{N}(x; l, c, N(s)) \triangleq \frac{1}{\sqrt{2\pi}c} e^{-\frac{(x-l)^2}{2c^2}},$$

where $N(s) = \sqrt{2\pi}$ comes from the normalisation of the pdf.
The *cdf* is instead given by

$$P(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}c} e^{-\frac{(x-l)^2}{2c^2}} dx = \int_{-\infty}^{\frac{z-l}{c}} \frac{1}{\sqrt{2\pi}} e^{-y^2} dy,$$

which is a widely studied integral, tabulated numerically, called *cumulative standard Gaussian distribution*.

Background on statistics

Random variables, pdfs and pmfs

Using the property of the *Dirac delta function*, i.e.,

$$\delta(x) = 0, \;\; \forall x \neq 0, \;\; \text{and} \;\; \int_{-\infty}^{\infty} \delta(x) dx = 1,$$

the pdf corresponding to the pmf can be written as

$$p(x) = \sum_{i=1}^{n} \pi_i \delta(x - a_i).$$

Notice that the previous relation holds whenever the pdf is considered as a *generalised function*.

If standard functions are used, the *Lebesgue's decomposition theorem* should be considered.

With a slight abuse of notation, we can state that a deterministic variable, i.e. a generic value b, can be represented with Dirac delta pdf, i.e. $p(x) = \delta(x - b)$.

Background on statistics

Random variables, pdfs and pmfs

Definition (Cumulative probability mass function)

The *cumulative probability mass function* (cmf) is a staircase function, i.e.,

$$\Pr[x \le a] = \sum_{i=1}^{n} \pi_i \mathbf{1}(a - a_i),$$

where $\mathbf{1}(\cdot)$ is the *unitary step function*.

Background on statistics

Examples of pmfs: Poisson

$x \sim \mathcal{P}(\lambda)$ defines x as a *Poisson distributed random variable* with *rate* λ.
The Poisson distribution describes the probability of a certain number of random points in an interval T.
The *pmf* is given by

$$p(x) = e^{-\lambda T} \frac{(\lambda T)^x}{x!},$$

or

$$\pi(a) = \Pr\left[x = a\right] = e^{-\lambda T} \frac{(\lambda T)^a}{a!},$$

where $a \in \mathbb{N}_0$.

Background on statistics

Random variables attributes

It is sometimes needed to characterise a **rv** with certain attributes describing, e.g., the *typical* value, the *spread* or variability, etc.
These attributes are computed on the pdf and *not* on the data.
For example we can be interested on the *central value*.

Definition (Mode)

The *mode* is the value in which the pdf attains its maximum value.

This is not a very good measure of the central value since the pdf can have multiple peaks.

Background on statistics

Random variables attributes

A good measure is instead given by the *moments*, a concept derived from standard mechanics.

r–th moment	Discrete	Continuous
about the origin	$\mu'_r = \sum_i a_i^r \pi_i$	$\mu'_r = \int_{-\infty}^{+\infty} x^r p(x) dx$
about μ'_1	$c'_r = \sum_i (a_i - \mu'_1)^r \pi_i$	$c'_r = \int_{-\infty}^{+\infty} (x - \mu'_1)^r p(x) dx$

The r–th *moment* about the origin is called *r–th order raw statistical moment* of a **rv**.
The r–th *moment* about μ'_1 is called *r–th order central statistical moment* of a **rv**.

Background on statistics

Random variables attributes

More precisely, the *moments* are defined using the *expected value* operator:

$$\mathsf{E}\left\{x\right\} \triangleq \int_{-\infty}^{+\infty} x p(x) dx \text{ or } \mathsf{E}\left\{x\right\} \triangleq \sum_i a_i \pi_i .$$

Therefore

r–th moment	Discrete	Continuous
raw	$\mu_r' = \mathsf{E}\left\{x^r\right\}$	$\mu_r' = \mathsf{E}\left\{x^r\right\}$
central	$c_r' = \mathsf{E}\left\{(x - \mu_1')^r\right\}$	$c_r' = \mathsf{E}\left\{(x - \mu_1')^r\right\}$

Background on statistics

Random variables attributes

Notice that:

- For $r = 0$, $\mu'_0 = 1$, $\forall x$. Hence no information can be retrieved about x. The same for c'_0;
- For $r = 1$, $\mu'_1 \triangleq \mu$ is called the *mean*, i.e. the long run average;
- For $r = 1$, $c'_1 = 0$, $\forall x$. Hence no information can be retrieved about x;
- For $r = 2$, $c'_2 \triangleq \sigma^2$ is called the *variance*.

Background on statistics

Random variables attributes

To summarise:

Moments	Discrete	Continuous
mean	$\mu = \sum_i a_i \pi_i$	$\mu = \int_{-\infty}^{+\infty} x p(x) dx$
variance	$\sigma^2 = \sum_i (a_i - \mu)^2 \pi_i$	$\sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 p(x) dx$

Moments of order greater than the second are called the *skewness* (third moment) and the *kurtosis* or *flatness* (fourth moment).

Background on statistics

Random variables attributes

Let us suppose that $y = g(x)$, where x is a **rv** and $g(\cdot)$ is a certain function.
By definition, the *expected value* of $y = g(x)$ is defined as

$$\mathsf{E}\{y\} \triangleq \int_{-\infty}^{+\infty} g(x)p(x)dx.$$

We further define:

$$\mathsf{V}\{x\} \triangleq \mathsf{E}\left\{(x-\mu)^2\right\} = \sigma^2,$$

where $\mathsf{V}\{x\}$ stands for "the variance of x".

Background on statistics

Random variables attributes

For the *expected value*, we further know that:

- If α is a deterministic value

$$\mathsf{E}\left\{\alpha\right\} = \alpha.$$

  (You can prove it by assuming that a deterministic value can be considered a **rv** with pdf equals to a Dirac δ function);

- Moreover, if α_1 and α_2 are two deterministic values and x_1 and x_2 two **rv**s, we have

$$\mathsf{E}\left\{\alpha_1 x_1 + \alpha_2 x_2\right\} = \alpha_1 \mathsf{E}\left\{x_1\right\} + \alpha_2 \mathsf{E}\left\{x_2\right\},$$

  i.e. the *expected value* is a *linear operator*.

Background on statistics

Mean and variance of some typical pdfs

For a *Uniform* pdf $\mathcal{U}(a, b)$, we have:

$$\mu = \mathsf{E}\left\{x\right\} = \frac{b+a}{2},$$

and

$$\sigma^2 = \mathsf{V}\left\{x\right\} = \mathsf{E}\left\{(x-\mu)^2\right\} = \frac{(b-a)^2}{12}.$$

Background on statistics

Mean and variance of some typical pdfs

For a *Gaussian* pdf $\mathcal{N}(l, c)$, we have:

$$
\begin{aligned}
\mathsf{E}\{x\} &= \int_{-\infty}^{+\infty} x \frac{1}{\sqrt{2\pi}c} e^{-\frac{(x-l)^2}{2c^2}} dx = \\
&= \int_{-\infty}^{+\infty} (x-l) \frac{1}{\sqrt{2\pi}c} e^{-\frac{(x-l)^2}{2c^2}} dx + l \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}c} e^{-\frac{(x-l)^2}{2c^2}} dx = \\
&= \int_{-\infty}^{+\infty} y \frac{1}{\sqrt{2\pi}c} e^{-\frac{y^2}{2c^2}} dy + l = \\
&= \int_{-\infty}^{0} y \frac{1}{\sqrt{2\pi}c} e^{-\frac{y^2}{2c^2}} dy + \int_{0}^{+\infty} y \frac{1}{\sqrt{2\pi}c} e^{-\frac{y^2}{2c^2}} dy + l =
\end{aligned}
$$

Background on statistics

Mean and variance of some typical pdfs

$$
\begin{aligned}
&= \int_{-\infty}^{0} y \frac{1}{\sqrt{2\pi}c} e^{-\frac{y^2}{2c^2}} dy + \int_{0}^{+\infty} y \frac{1}{\sqrt{2\pi}c} e^{-\frac{y^2}{2c^2}} dy + l = \\
&= -\int_{0}^{-\infty} y \frac{1}{\sqrt{2\pi}c} e^{-\frac{y^2}{2c^2}} dy + \int_{0}^{+\infty} y \frac{1}{\sqrt{2\pi}c} e^{-\frac{y^2}{2c^2}} dy + l = \\
&= -\int_{0}^{+\infty} (-z) \frac{1}{\sqrt{2\pi}c} e^{-\frac{(-z)^2}{2c^2}} d(-z) + \int_{0}^{+\infty} y \frac{1}{\sqrt{2\pi}c} e^{-\frac{y^2}{2c^2}} dy + l = \\
&= -\int_{0}^{+\infty} z \frac{1}{\sqrt{2\pi}c} e^{-\frac{z^2}{2c^2}} dz + \int_{0}^{+\infty} y \frac{1}{\sqrt{2\pi}c} e^{-\frac{y^2}{2c^2}} dy + l = l = \mu = \mathsf{E}\{x\}
\end{aligned}
$$

Background on statistics

Mean and variance of some typical pdfs

For a *Gaussian* pdf $\mathcal{N}(\mu, c)$, we have:

$$\mathsf{V}\left\{x\right\} = \mathsf{E}\left\{(x-\mu)^2\right\} = \int_{-\infty}^{+\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi}c} e^{-\frac{(x-\mu)^2}{2c^2}} dx =$$
$$= \int_{-\infty}^{+\infty} y^2 \frac{1}{\sqrt{2\pi}c} e^{-\frac{y^2}{2c^2}} dy.$$

To solve this integral we first notice that

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}c} e^{-\frac{y^2}{2c^2}} dy = 1 \Rightarrow \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2c^2}} dy = c.$$

Background on statistics

Mean and variance of some typical pdfs

Differentiating both sides with respect to c, we get

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2c^2}} dy = c \Rightarrow \int_{-\infty}^{+\infty} y^2 \frac{1}{\sqrt{2\pi} c^3} e^{-\frac{y^2}{2c^2}} dy = 1.$$

Therefore

$$\int_{-\infty}^{+\infty} y^2 \frac{1}{\sqrt{2\pi} c} e^{-\frac{y^2}{2c^2}} dy = c^2 \Rightarrow \mathsf{V}\{x\} = c^2 = \sigma^2.$$

For a Gaussian pdf, $\sigma = \sqrt{\sigma^2}$ is called the *standard deviation*.
It has to be noted that *only* the first two moments are sufficient to characterise a Gaussian pdf, i.e. $\mathcal{N}(\mu, \sigma)$.

Background on statistics

Mean and variance of some typical pdfs

The mean of the *Poisson discrete* random variable is given by

$$\mu = \mathsf{E}\left\{x\right\} = \sum_{x=0}^{+\infty} x e^{-\lambda T} \frac{(\lambda T)^x}{x!} = e^{-\lambda T} \sum_{x=0}^{+\infty} x \frac{(\lambda T)^x}{x!}.$$

Since for $x = 0$ the last term is 0, we can write

$$\mu = e^{-\lambda T} \sum_{x=1}^{+\infty} x \frac{(\lambda T)^x}{x!} = e^{-\lambda T} \sum_{x=1}^{+\infty} \frac{(\lambda T)^x}{(x-1)!} = e^{-\lambda T} \lambda T \sum_{x=1}^{+\infty} \frac{(\lambda T)^{(x-1)}}{(x-1)!}.$$

Background on statistics

Mean and variance of some typical pdfs

By changing the variable $x - 1 = y$, we have

$$\mu = e^{-\lambda T} \lambda T \sum_{x=1}^{+\infty} \frac{(\lambda T)^{(x-1)}}{(x-1)!} = e^{-\lambda T} \lambda T \sum_{y=0}^{+\infty} \frac{(\lambda T)^{y}}{y!}.$$

Notice that the summation on the right hand side is the *Taylor expansion of the exponential function*, thus yielding to

$$\mu = e^{-\lambda T} \lambda T e^{\lambda T} \Rightarrow \mu = \lambda T.$$

Background on statistics

Mean and variance of some typical pdfs

For the variance of the *Poisson discrete* random variable, we first derive a relation of general validity:

Remark

$V\{x\} \triangleq E\{(x-\mu)^2\} = E\{x^2+\mu^2-2\mu x\} = E\{x^2\} - \mu^2.$

Hence:

$$\mathrm{V}\{x\} = \mathrm{E}\{x^2\} - \mu^2 = \sum_{x=0}^{+\infty} x^2 e^{-\lambda T}\frac{(\lambda T)^x}{x!} - (\lambda T)^2.$$

We then notice that $x^2 = x(x-1)+x$, thus yielding to

$$\mathrm{V}\{x\} = \sum_{x=0}^{+\infty} x(x-1)e^{-\lambda T}\frac{(\lambda T)^x}{x!} + \sum_{x=0}^{+\infty} xe^{-\lambda T}\frac{(\lambda T)^x}{x!} - (\lambda T)^2.$$

Background on statistics

Mean and variance of some typical pdfs

From

$$\mathsf{V}\left\{x\right\} = \sum_{x=0}^{+\infty} x(x-1)e^{-\lambda T}\frac{(\lambda T)^x}{x!} + \sum_{x=0}^{+\infty} x e^{-\lambda T}\frac{(\lambda T)^x}{x!} - (\lambda T)^2.$$

we recognise μ in the second term.
Moreover, we can apply the previous trick for *two terms* noticing that for $x = 0$ and $x = 1$ the first term is 0, hence

$$\mathsf{V}\left\{x\right\} = (\lambda T)^2 + \mu - (\lambda T)^2 = \mu,$$

which is an interesting result!

Background on statistics

Random variables attributes

The *mean* value gives an idea of the central value of the **rv**, while the *variance* gives an idea about the spread.
Additional measures are given by the *quantiles*, which are based on the fact that the cdf is always *invertible*.

Definition (Quantile)

The *r–th quantile* is the value x_r such that $x_r = P^{-1}(r)$.

Usually, the quartiles $r = 0.25$ and $r = 0.75$ are considered. The difference $x_{0.75} - x_{0.25}$ is the *interquartile range*.

Background on statistics

Random variables attributes

Definition (Median)

The *median* $\tilde{\mu}$ is the quantile for $r = \frac{1}{2}$.

In practice:

Median	Discrete	Continuous
$\frac{1}{2} =$	$\Pr[x \leq \tilde{\mu}] = \Pr[x \geq \tilde{\mu}]$	$\int_{-\infty}^{\tilde{\mu}} p(x)dx = \int_{\tilde{\mu}}^{+\infty} p(x)dx$

The median is in general preferable to the mean and the variance as a measure of the pdf because the latter are more influenced by the pdf tails or when the pdf is multimodal.
For this reason, the median is defined as a *robust* attribute of the pdf.

Background on statistics

Example of Median Robustness

Yesterday, 8 individuals payed these amounts of Euros to have a dinner in a certain restaurant:

$$x = \begin{bmatrix} 10 & 10 & 20 & 360 & 30 & 10 & 40 & 30 \end{bmatrix}$$

What is the typical amount of money you are expecting to pay for a dinner in the same restaurant tonight?
If you compute the *mean*, it is 60 Euros.
If you compute the *mode*, it is 10 Euros.
If you compute the *median*, it is 25 Euros.
What is the most appropriate?

Background on statistics

Random variables attributes

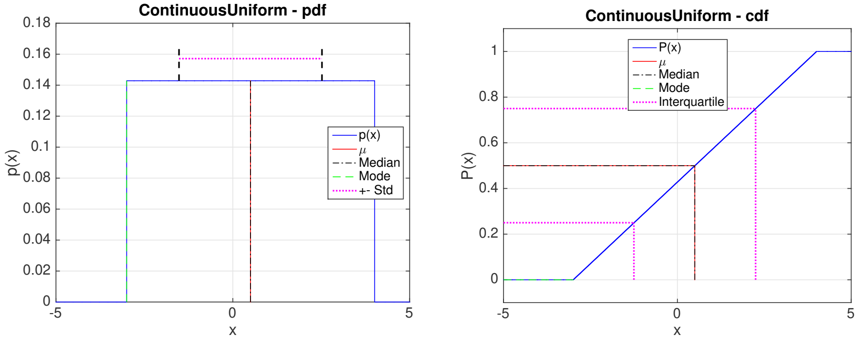
Sometimes it is also used the *mean deviation* with respect to the *mean* μ

$$\overline{\mu} = \mathsf{E}\left\{|x-\mu|\right\}\int_{-\infty}^{+\infty}|x-\mu|p(x)dx,$$

or with respect to the *median* $\tilde{\mu}$, i.e. $\underline{\mu} = \mathsf{E}\left\{|x-\tilde{\mu}|\right\}$.
Since the *median* robustly replaces the *mean*, the *mean deviation* is sometimes considered as the robust version of the *variance*.

Background on statistics

Random variables: Uniform continuous pdf and cdf



(a) (b)

Figure: Uniform continuous pdf (a) and cdf (b) with attributes.

Background on statistics

Random variables: Gaussian continuous pdf and cdf

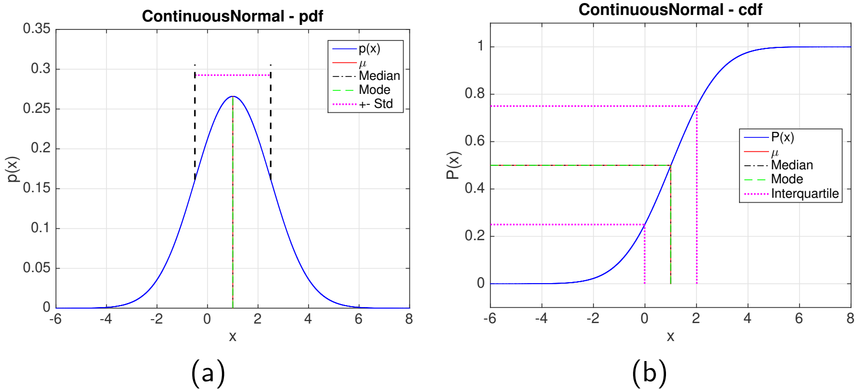
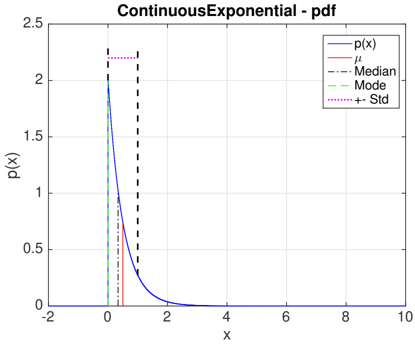


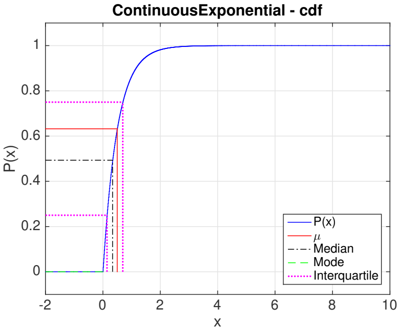
Figure: Gaussian continuous pdf (a) and cdf (b) with attributes.

Background on statistics

Random variables: Exponential continuous pdf and cdf



(a)

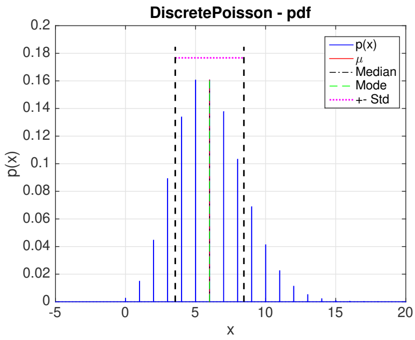


(b)

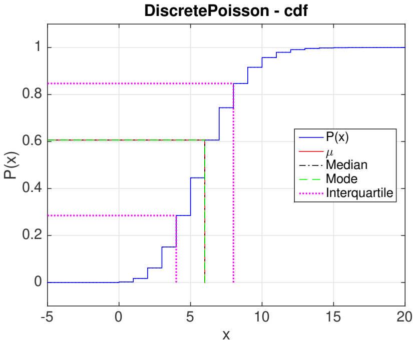
Figure: Exponential continuous pdf (a) and cdf (b) with attributes.

Background on statistics

Random variables: Poisson discrete pdf and cdf



(a)



(b)

Figure: Poisson discrete pdf (a) and cdf (b) with attributes.

Outline

Background on statistics

Suppose we have a vector of random variables

$$x = [x_1, x_2, \dots, x_n]^T.$$

The **rv**s are then distributed according to a *joint probability density function*

$$p(x) \triangleq p(x_1, x_2, \dots, x_n).$$

Moreover, if the domain of $p(x)$ is all the n-dimensional space (if it is not, the integrals should be computed with the appropriate limits)

$$P(a) = P(a_1, a_2, \dots, a_n) = \int_{-\infty}^{a_1} \int_{-\infty}^{a_2} \cdots \int_{-\infty}^{a_n} p(x) d^n x,$$

that is the *multivariate cumulative distribution function*.

Background on statistics

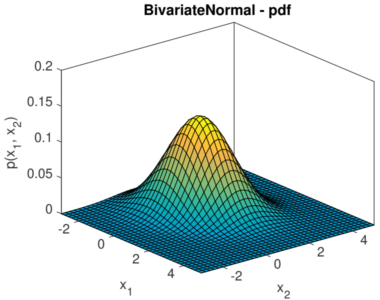
Bivariate distributions

A particular case of the multivariate pdfs are the *bivariate* pdfs, for which the graphical representation is possible.
The main concept around the *bivariate* pdfs is that the probability of x_1 of falling in a certain range is not unrelated to the probability of x_2 of falling in a certain (possibly different) range.
The main point here is that the multivariate pdfs allow to express *relationships* between variables.
For a bivariate pdf is also practical writing the definition of the *joint pdf* as

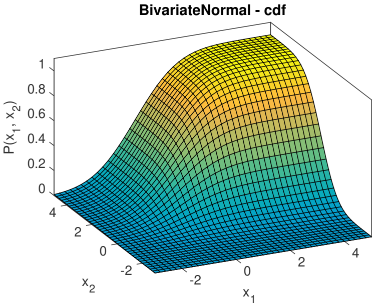
$$p(x_1, x_2) \triangleq \lim_{\delta_a \to 0, \delta_b \to 0} \frac{\text{Pr}\left[(a \leq x_1 \leq a + \delta_a) \cap (b \leq x_2 \leq b + \delta_b)\right]}{\delta_a \delta_b}$$

Background on statistics

Example: Bivariate Gaussian Pdf



(a)



(b)

Figure: Bivariate Gaussian pdf (a) and cdf (b).

Background on statistics

Moments of bivariate pdfs

The *joint statistical moments* of order $r + k$ of two **rvs** x_1 and x_2 are

Moments	Discrete
raw	$\mu'_{r,k}(x_1, x_2) = \sum_i \sum_j a_i^r a_j^k \pi_{x_1,x_2}$
central	$c'_{r,k}(x_1, x_2) = \sum_i \sum_j (a_i - \mu_1)^r (a_j - \mu_2)^k \pi_{x_1,x_2}$
	Continuous
raw	$\mu'_{r,k}(x_1, x_2) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x_1^r x_2^k p(x_1, x_2) dx_1 dx_2$
central	$c'_{r,k}(x_1, x_2) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x_1 - \mu_1)^r (x_2 - \mu_2)^k p(x_1, x_2) dx_1 dx_2$

Background on statistics

Moments of bivariate pdfs

And using the expected value operator

Moments	Discrete
raw	$\mu'_{r,k}(x_1, x_2) = \mathsf{E}\left\{x_1^r x_2^k\right\}$
central	$c'_{r,k}(x_1, x_2) = \mathsf{E}\left\{(x_1 - \mu_1)^r (x_2 - \mu_2)^k\right\}$
	Continuous
raw	$\mu'_{r,k}(x_1, x_2) = \mathsf{E}\left\{x_1^r x_2^k\right\}$
central	$c'_{r,k}(x_1, x_2) = \mathsf{E}\left\{(x_1 - \mu_1)^r (x_2 - \mu_2)^k\right\}$

Background on statistics

Moments of bivariate pdfs

Notice that:

- For $r = k = 0$, $\mu'_{0,0} = 1$, $\forall x_1, x_2$. Hence no information can be retrieved about x_1 and x_2. The same for $c'_{0,0}$;
- For $r = 1$ and $k = 0$ or $r = 0$ and $k = 1$, $\mu'_{1,0} \triangleq \mu_1$ or $\mu'_{0,1} \triangleq \mu_2$, respectively, i.e. the *mean* of the single **rv**;
- For $r = 1$ and $k = 0$ or $r = 0$ and $k = 1$, $c'_{1,0} = c'_{0,1} = 0$, $\forall x_1, x_2$. Hence no information can be retrieved about x_1 and x_2;
- For $r = 1$ and $k = 1$, $\mu'_{1,1} \triangleq \phi_{x_1,x_2}$ is the *correlation*;
- For $r = 1$ and $k = 1$, $c'_{1,1} \triangleq \mathsf{C}\left\{x_1, x_2\right\}$ is the *covariance*.

Background on statistics

Conditionals and marginals

Given a multivariate pdf it is possible to infer a certain set of *lower dimensional* pdfs.
Consider a *bivariate* cdf. Basically, the *marginal* cdf consists of considering the probability only in one direction or for *one variable*, i.e.,

$$P(a_1, \infty) = \mathsf{Pr}\left[x_1 \leq a_1, x_2 \leq +\infty\right] = \mathsf{Pr}\left[x_1 \leq a_1\right].$$

So from the *bivariate cumulative distribution function* it is possible to obtain the *marginal cumulative distribution function* of x_1.

Background on statistics

Conditionals and marginals

The same can be obtained using the pdfs, and hence the *marginal probability density function* of x_1 is obtained by integrating the bivariate density function over all the possible values of x_2, i.e.

$$p_1(x_1) = \int_{-\infty}^{+\infty} p(x_1, x_2) dx_2.$$

This operation corresponds in collapsing the density on one axis.

Background on statistics

Bivariate multimodal Gaussian

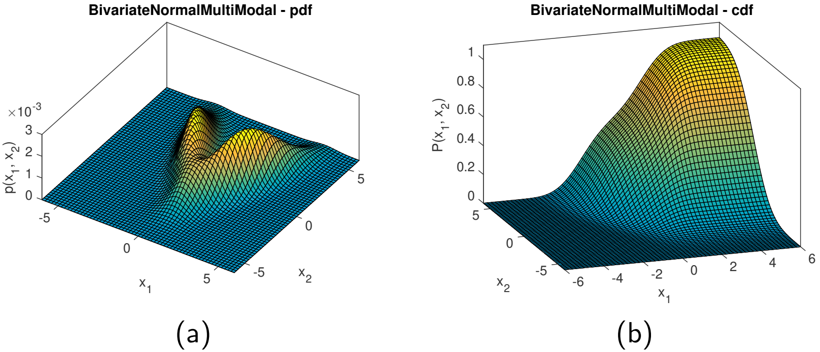


Figure: Bivariate multimodal Gaussian pdf (a) and cdf (b).

Background on statistics

Marginals: Bivariate multimodal Gaussian

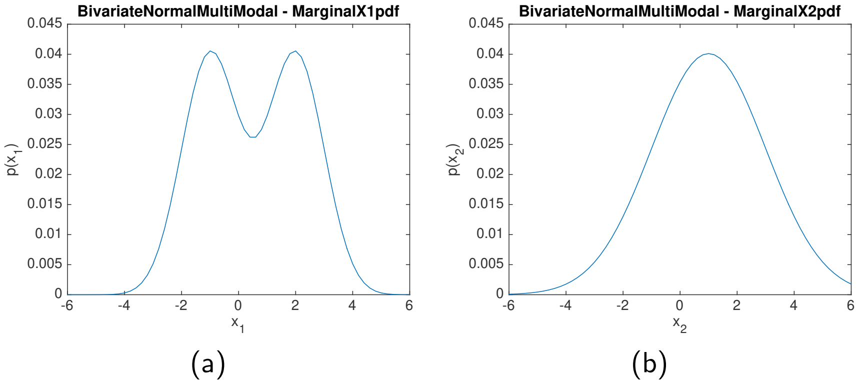


Figure: Marginals of the Bivariate multimodal Gaussian pdf with respect to x_2 (a) and x_1 (b).

Background on statistics

Conditional pdf

As the conditional probability gives the probability of A given B, i.e., $\mathsf{Pr}\left[A|B\right]$, the *conditional* pdf gives the pdf of x_1 given that $x_2 = a_2$. Let us recall that

$$\mathsf{Pr}\left[A|B\right] = \frac{\mathsf{Pr}\left[B|A\right]\mathsf{Pr}\left[A\right]}{\mathsf{Pr}\left[B\right]} = \frac{\mathsf{Pr}\left[A \cap B\right]}{\mathsf{Pr}\left[B\right]}.$$

For a cdf it becomes

$$\mathsf{Pr}\left[a_1|B\right] = \frac{\mathsf{Pr}\left[(x_1 \leq a_1) \cap B\right]}{\mathsf{Pr}\left[B\right]} = P_1(a_1|B).$$

Background on statistics

Conditional pdf

Let us suppose that the event B is $\{a_2 < x_2 \leq a_2 + \delta_{a_2}\}$.
Hence, by substitution

$$P_1(a_1|a_2 < x_2 \leq a_2 + \delta_{a_2}) = \frac{\mathsf{Pr}\left[(x_1 \leq a_1) \cap (a_2 < x_2 \leq a_2 + \delta_{a_2})\right]}{\mathsf{Pr}\left[a_2 < x_2 \leq a_2 + \delta_{a_2}\right]},$$

and therefore, since $\mathsf{Pr}\left[a_2 < x_2 \leq a_2 + \delta_{a_2}\right] = P_2(a_2 + \delta_{a_2}) - P_2(a_2)$, we have

$$P_1(a_1|a_2 < x_2 \leq a_2 + \delta_{a_2}) = \frac{P(a_1, a_2 + \delta_{a_2}) - P(a_1, a_2)}{P_2(a_2 + \delta_{a_2}) - P_2(a_2)}.$$

Background on statistics

Conditional pdf

Recall that the cdf is the integral of the pdf, hence

$$P_1(a_1|a_2 < x_2 \leq a_2 + \delta_{a_2}) = \frac{\int_{-\infty}^{a_1} \int_{a_2}^{a_2+\delta_{a_2}} p(x_1, x_2) dx_2 dx_1}{\int_{a_2}^{a_2+\delta_{a_2}} p_2(x_2) dx_2}.$$

Assuming δ_{a_2} small enough, we get

$$P_1(a_1|a_2 < x_2 \leq a_2 + \delta_{a_2}) \approx \frac{\int_{-\infty}^{a_1} p(x_1, x_2 = a_2) dx_1 \delta_{a_2}}{p_2(x_2 = a_2)\delta_{a_2}},$$

and in the limit

$$\begin{aligned} P_1(a_1|x_2 = a_2) &= \lim_{\delta_{a_2} \to 0} P_1(a_1|a_2 < x_2 \leq a_2 + \delta_{a_2}) = \\ &= \frac{\int_{-\infty}^{a_1} p(x_1, x_2 = a_2) dx_1}{p_2(x_2 = a_2)}. \end{aligned}$$

Background on statistics

Conditional pdf

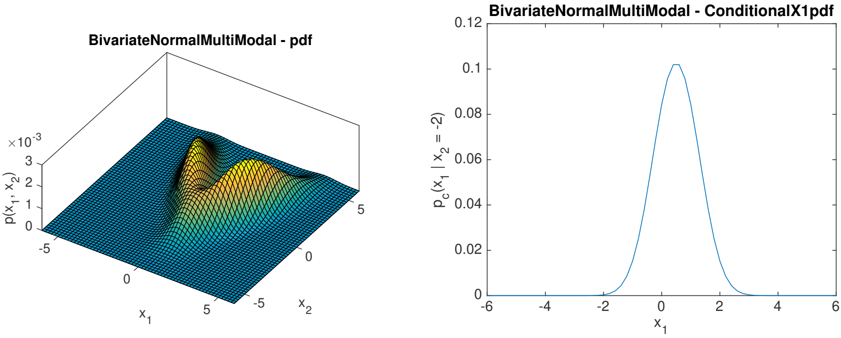
Finally, by differentiating both sides with respect to x_1, we finally have:

$$p_{c_1}(x_1|x_2 = a_2) = \frac{p(x_1, x_2 = a_2)}{p_2(x_2 = a_2)} = \frac{p(x_1, x_2 = a_2)}{\int_{-\infty}^{+\infty} p(x_1, x_2 = a_2) dx_1},$$

which has the marginal at the denominator.
In practice, the *conditional pdf* $p_c(x_1|x_2 = a_2)$ is a *slice* of the joint pdf $p(x_1, x_2)$ with x_2 held constant and equal to a_2.
The conditional pdf can be computed for *any arbitrary slice*, i.e. $\forall a_2$, but also *for any* closed curve on the domain of $p(x_1, x_2)$.
Of course, the conditional pdf and the marginal pdf can be extended to *any dimension* for the joint pdf.

Background on statistics

Conditional pdf: Bivariate multimodal Gaussian



(a) (b)

Figure: Bivariate multimodal Gaussian pdf (a) and its conditional with respect to $x_2 = -2$ (b).

Background on statistics

Conditional pdf: Bivariate multimodal Gaussian

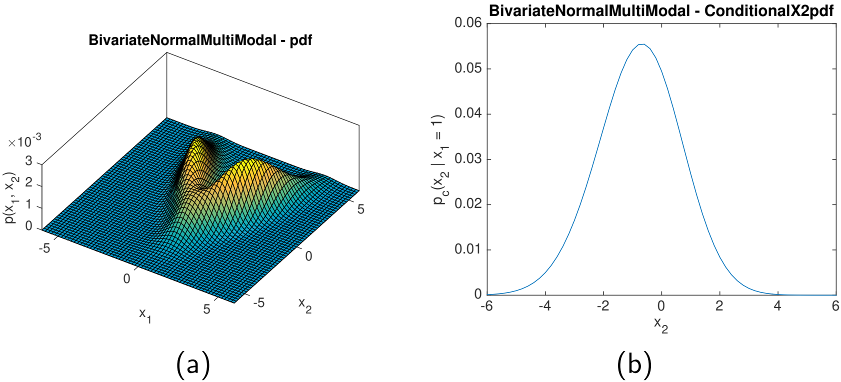


Figure: Bivariate multimodal Gaussian pdf (a) and its conditional with respect to $x_1 = 1$ (b).

Background on statistics

Conditional pdf

From

$$p_{c_1}(x_1|x_2) = \frac{p(x_1, x_2)}{p_2(x_2)} = \frac{p(x_1, x_2)}{\int_{-\infty}^{+\infty} p(x_1, x_2) dx_1},$$

we can derive that:

- $p(x_1, x_2) = p_{c_1}(x_1|x_2) p_2(x_2) = p_{c_2}(x_2|x_1) p_1(x_1)$;
- $p_1(x_1) = \int_{-\infty}^{+\infty} p(x_1, x_2) dx_2 = \int_{-\infty}^{+\infty} p_{c_1}(x_1|x_2) p_2(x_2) dx_2$.

Therefore:

$$p_{c_1}(x_1|x_2) = \frac{p(x_1, x_2)}{\int_{-\infty}^{+\infty} p(x_1, x_2) dx_1} = \frac{p_{c_2}(x_2|x_1) p_1(x_1)}{\int_{-\infty}^{+\infty} p_{c_2}(x_2|x_1) p_1(x_1) dx_1},$$

which is the *Bayes theorem* for pdfs.

Background on statistics

Conditional pdf

Let us show that the *conditional pdf* is actually a well posed pdf.
To this end, we first notice that

$$p_{c_1}(x_1|x_2) = \frac{p(x_1, x_2)}{p_2(x_2)} \geq 0.$$

Moreover, since

$$p_{c_1}(x_1|x_2) = \frac{p_{c_2}(x_2|x_1)p_1(x_1)}{\int_{-\infty}^{+\infty} p_{c_2}(x_2|x_1)p_1(x_1)dx_1},$$

it is obvious that

$$\int_{-\infty}^{+\infty} p_{c_1}(x_1|x_2)dx_1 = 1,$$

i.e. the denominator turns to be a *normalisation factor*.

Background on statistics

Conditional pdf: Example 1

Given

$$p(x_1, x_2) = \begin{cases} k, & 0 \leq x_1 < x_2 \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

compute the conditional pdfs.

First, we have to determine the value of the constant k, which is a function of the *normalisation of the pdf*, i.e.

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x_1, x_2) dx_1 dx_2 = 1.$$

Therefore

$$\int_0^1 \int_0^{x_2} k dx_1 dx_2 = \int_0^1 k x_2 dx_2 = \frac{k}{2} = 1 \Rightarrow k = 2.$$

Background on statistics

Conditional pdf: Example 1

For the marginals we have:

$$p_1(x_1) = \int_{-\infty}^{+\infty} p(x_1, x_2) dx_2 = \int_{x_1}^{1} k dx_2 = k(1 - x_1), \;\; 0 < x_1 < 1,$$

and

$$p_2(x_2) = \int_{-\infty}^{+\infty} p(x_1, x_2) dx_1 = \int_{0}^{x_2} k dx_1 = kx_2, \;\; 0 < x_2 < 1.$$

Background on statistics

Conditional pdf: Example 1

As a consequence:

$$p_{c_1}(x_1|x_2) = \frac{p(x_1, x_2)}{p_2(x_2)} = \frac{1}{x_2}, \;\; x_1 \in [0, x_2),$$

and

$$p_{c_2}(x_2|x_1) = \frac{p(x_1, x_2)}{p_1(x_1)} = \frac{1}{1 - x_1}, \;\; x_2 \in (x_1, 1].$$

We can now compute *any slice* of this *bivariate* pdf by imposing a value to x_2 in $p_{c_1}(x_1|x_2)$ or a value x_1 in $p_{c_2}(x_2|x_1)$: we have then a *set of parametric uniform pdfs*.

Notice that now each *conditional pdf* is defined for a *generic* value of x_2 (resp., x_1)!

We can now derive the generic form of the *Bayes Theorem*...

Background on statistics

Bayes postulate

The *Bayes theorem*

$$p_{c_1}(x_1|x_2) = \frac{p(x_1, x_2)}{\int_{-\infty}^{+\infty} p(x_1, x_2) dx_1} = \frac{p_{c_2}(x_2|x_1) p_1(x_1)}{\int_{-\infty}^{+\infty} p_{c_2}(x_2|x_1) p_1(x_1) dx_1},$$

is the basis of the *Bayesian inference*.

The *prior* $p_1(x_1)$ reflects the *degree of a-priori belief* about x_1.

The *likelihood function* $p_{c_2}(x_2|x_1)$ is the *evidence from the data*, i.e., the *measurements*.

The *posterior* $p_{c_1}(x_1|x_2)$ is the combination of the prior with the likelihood function by means of the *Bayes theorem*.

Background on statistics

Conditional pdf: Example 2

Let us consider the problem of estimating the angular position of a motor shaft, with angle $\theta \in [0, 2\pi]$, using a measure $\theta_m = \theta + \varepsilon$, with *sensing error noise* $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

This problem can be solved with conditional pdfs and *Bayes theorem*.

Indeed, the *pdf of the estimate* $\hat{\theta}$ of the actual value θ can be derived if the *posterior* $p_{c_1}(\hat{\theta}|\theta_m)$ is known.

The *posterior* gives a stochastic description of the uncertainty associated with the estimates $\hat{\theta}$.

Background on statistics

Conditional pdf: Example 2

To this end, we apply the *Bayes theorem* and we write

$$p_{c_1}(\hat{\theta}|\theta_m) = \frac{p_{c_2}(\theta_m|\hat{\theta})p(\hat{\theta})}{\int_{-\infty}^{+\infty} p_{c_2}(\theta_m|\hat{\theta})p(\hat{\theta})d\hat{\theta}}.$$

Since before the first measure, there is *no knowledge* about the angle position, define $p(\hat{\theta})$ is a $\mathcal{U}(0, 2\pi)$, i.e. the *prior*.
Therefore

$$p_{c_1}(\hat{\theta}|\theta_m) = \frac{p_{c_2}(\theta_m|\hat{\theta})\frac{1}{2\pi}}{\int_{0}^{2\pi} p_{c_2}(\theta_m|\hat{\theta})\frac{1}{2\pi}d\hat{\theta}}.$$

Background on statistics

Conditional pdf: Example 2

Since $\theta_m = \theta + \varepsilon$, we immediately have that the measurement takes place when θ *is deterministic* (i.e. a value), hence

$$\mathsf{E}\left\{\theta_m\right\} = \mathsf{E}\left\{\theta + \varepsilon\right\} = \mathsf{E}\left\{\theta\right\} + \mathsf{E}\left\{\varepsilon\right\} = \theta,$$

and

$$\mathsf{V}\left\{\theta_m\right\} = \mathsf{E}\left\{(\theta_m - \mathsf{E}\left\{\theta_m\right\})^2\right\} = \mathsf{E}\left\{\varepsilon^2\right\} = \mathsf{V}\left\{\varepsilon\right\} = \sigma^2.$$

Therefore, since θ is a value

$$p_{c_2}(\theta_m|\theta) \sim \mathcal{N}(\theta, \sigma^2),$$

which is the *measurement uncertainty*.

Background on statistics

Conditional pdf: Example 2

Unfortunately, the pdf $\mathcal{N}(\theta, \sigma^2)$ is centred in the *actual value* θ, which is unknown (thus useless in an engineering problem), while σ^2 is known due to, e.g., *Type A* analysis.
However, for the sake of the *posterior conditional pdf* $p_{c_1}(\hat{\theta}|\theta_m)$, we can build the *likelihood function*, by assuming $\hat{\theta}$, the *known prior*, as the mean value, i.e.

$$p_{c_2}(\theta_m|\hat{\theta}) \sim \mathcal{N}(\hat{\theta}, \sigma^2),$$

This way we can compute *how likely* the *measurement* θ_m would be measured given the prior knowledge $\hat{\theta}$, i.e. *evidence from the data*.

Background on statistics

Conditional pdf: Example 2

Therefore:

$$p_{c_1}(\hat{\theta}|\theta_m) = \frac{\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(\theta_m-\hat{\theta})^2}{2\sigma^2}}\frac{1}{2\pi}}{\int_0^{2\pi}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(\theta_m-\hat{\theta})^2}{2\sigma^2}}\frac{1}{2\pi}d\hat{\theta}} = \frac{e^{-\frac{(\theta_m-\hat{\theta})^2}{2\sigma^2}}}{\int_0^{2\pi}e^{-\frac{(\theta_m-\hat{\theta})^2}{2\sigma^2}}d\hat{\theta}} =$$
$$= f(\theta_m)e^{-\frac{(\hat{\theta}-\theta_m)^2}{2\sigma^2}}$$

which is a *unimodal pdf* defined in the set $[0, 2\pi]$ (because of the presence of the prior $p(\hat{\theta})$ at numerator) and with a mean value in θ_m.
Hence, *the posterior gives more information than the prior*!

Background on statistics

Example: Marginal mean and conditional mean

Consider a population of n_m males and n_f females.
What is the *average height among the men* or *among the females*?
Denoting with h_{m_i} the *height of the i–th male* and with h_{f_i} the *height of the i–th female*, we get

$$\mathsf{E}\left\{\text{height among the men}\right\} = \frac{\sum_i h_{m_i}}{n_m}$$
$$\mathsf{E}\left\{\text{height among the female}\right\} = \frac{\sum_i h_{f_i}}{n_f}.$$

If we define with the **rv** x the *height* of the population, while with the **rv** y the sex (m or f), we have immediately

$$\mathsf{E}\left\{\text{height among the men}\right\} = \mathsf{E}\left\{x|y=m\right\}$$
$$\mathsf{E}\left\{\text{height among the female}\right\} = \mathsf{E}\left\{x|y=f\right\}.$$

Background on statistics

Example: Marginal mean and conditional mean

We first noticed that $\mathsf{E}\{x|y=m\}$ and $\mathsf{E}\{x|y=f\}$ are *two numbers* representing the *mean of the conditional pdf*.
We are now interested in computing the average height $\mathsf{E}\{x\}$.

Background on statistics

Example: Marginal mean and conditional mean

To this end, we first noticed that

$$\sum_i h_{m_i} = n_m \mathsf{E}\{x|y=m\} \text{ and } \sum_i h_{f_i} = n_f \mathsf{E}\{x|y=f\}.$$

Hence, the *mean height* is given by

$$\begin{aligned}\mathsf{E}\{x\} &= \frac{\sum_i h_{m_i} + \sum_i h_{w_i}}{n_m + n_f} = \\ &= \frac{n_m}{n_m + n_f}\mathsf{E}\{x|y=m\} + \frac{n_f}{n_m + n_f}\mathsf{E}\{x|y=f\},\end{aligned}$$

or, in other words:

$$\mathsf{E}\{x\} = \Pr[y=m]\,\mathsf{E}\{x|y=m\} + \Pr[y=f]\,\mathsf{E}\{x|y=f\}.$$

Background on statistics

Example: Marginal mean and conditional mean

The previous relation can be generalised to *any number n* of the conditioning event y (e.g it can be the **rv** *age*) and hence assume the following form

$$\mathsf{E}\{x\} = \sum_{i=1}^{n} \mathsf{Pr}[y = y_i]\, \mathsf{E}\{x|y = i\} = \sum_{i=1}^{n} \pi_{y_i} \mathsf{E}\{x|y = i\} \triangleq \mathsf{E}\{\mathsf{E}\{x|y\}\},$$

i.e. the *mean* for discrete **rv**s.

For continuous **rv** we have the relation given previously, i.e.

$$\mathsf{E}\{x\} = \int_{-\infty}^{+\infty} \mathsf{E}\{x|y\}\, p_y(y) dy = \mathsf{E}\{\mathsf{E}\{x|y\}\}.$$

Of course the previous relation can be extended to a generic $g(x)$ function...

Background on statistics

Conditional mean

In general, we may be interested in computing the parameters of the *conditional pdf* $p_c(x|y)$, e.g. the *mean of the conditional pdf*.
The mean of the conditional pdf is called the *conditional mean* and, given a slice $y = a$, it is given by

$$\mu_{x|y=a} \triangleq \mathsf{E}\{x|y=a\} \triangleq \int_{-\infty}^{+\infty} x p_c(x|y=a)dx,$$

which is a function of a.
The *conditional mean*, given the slice $y = a$, can be similarly expressed for a generic function $g(x)$, i.e.

$$\mathsf{E}\{g(x)|y=a\} = \int_{-\infty}^{+\infty} g(x)p_c(x|y=a)dx.$$

In both coses, the *conditional mean* is a number because *we restrict to a slice* $y = a$.

Background on statistics

Conditional mean

However, if we *do not restrict* to a specific value of y, the *conditional mean* $\mu_{g(x)|y}$ is a **rv**, i.e.

$$\mu_{g(x)|y} = \mathsf{E}\left\{g(x)|y\right\} = \int_{-\infty}^{+\infty} g(x)p_c(x|y)dx$$

is a **rv** related to the pdf $p_y(y)$ (which is now *not* restricted to a number). Therefore, we can compute the *mean* of the **rv** $\mu_{g(x)|y}$ using the definition

$$\mathsf{E}\left\{\mathsf{E}\left\{g(x)|y\right\}\right\} = \int_{-\infty}^{+\infty} \mathsf{E}\left\{g(x)|y\right\} p_y(y)dy,$$

where $\mathsf{E}\left\{\mathsf{E}\left\{g(x)|y\right\}\right\} = \mathsf{E}\left\{\mu_{g(x)|y}\right\}$.

Background on statistics

Marginal mean

In particular:

$$\begin{aligned}
\mathsf{E}\left\{\mathsf{E}\left\{g(x)|y\right\}\right\} &= \int_{-\infty}^{+\infty} \mathsf{E}\left\{g(x)|y\right\} p_y(y)dy = \\
&= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x)p_c(x|y)dx \; p_y(y)dy = \\
&= \int_{-\infty}^{+\infty} g(x) \int_{-\infty}^{+\infty} p_c(x|y)p_y(y)dy \; dx = \\
&= \int_{-\infty}^{+\infty} g(x) \int_{-\infty}^{+\infty} p(x,y)dy \; dx = \\
&= \int_{-\infty}^{+\infty} g(x)p_x(x)dx = \mathsf{E}\left\{g(x)\right\}.
\end{aligned}$$

Hence, the *mean* of the *conditional mean* is called *marginal mean* because it is obtained through *marginalisation* of the *joint pdf*.

Background on statistics

Variance of the conditional mean

Since the *conditional mean* is a **rv**, we can also compute its *variance*

$$\mathsf{V}\left\{\mathsf{E}\left\{g(x)|y\right\}\right\} = \mathsf{E}\left\{\mathsf{E}\left\{g(x)|y\right\}^2\right\} - \mathsf{E}\left\{\mathsf{E}\left\{g(x)|y\right\}\right\}^2 = \\ = \mathsf{E}\left\{\mu^2_{g(x)|y}\right\} - \mathsf{E}\left\{g(x)\right\}^2,$$

obtained using the *marginal mean*

$$\mathsf{E}\left\{\mu_{g(x)|y}\right\} = \mathsf{E}\left\{\mathsf{E}\left\{g(x)|y\right\}\right\} = \mathsf{E}\left\{g(x)\right\}.$$

Background on statistics

Conditional variance

Similarly, we may be interested in computing the *variance of the conditional pdf*.

The variance of the conditional pdf is called the *conditional variance* and, given $y = a$, it is given by

$$\sigma^2_{x|y=a} = \mathsf{V}\left\{x|y=a\right\} = \mathsf{E}\left\{(x-\mu_{x|y})^2|y=a\right\}.$$

We first recall that in general, for a **rv** x we have

$$\sigma^2_x = \mathsf{V}\left\{x\right\} = \mathsf{E}\left\{(x-\mu_x)^2\right\} = \mathsf{E}\left\{x^2\right\} - \mu^2_x.$$

Background on statistics

Conditional variance

The *conditional variance*, given $y = a$, it is then given by

$$\begin{aligned}\sigma^2_{x|y=a} &= \mathsf{V}\{x|y=a\} = \mathsf{E}\left\{(x-\mu_{x|y=a})^2|y=a\right\} = \\ &= \mathsf{E}\left\{x^2+\mu^2_{x|y=a}-2x\mu_{x|y=a}|y=a\right\} = \mathsf{E}\left\{x^2|y=a\right\} - \mu^2_{x|y=a}.\end{aligned}$$

This relation *is very important* in estimation theory!
The *conditional variance* $\sigma^2_{x|y}$ is a **rv** if we let y to be unspecified, i.e.

$$\sigma^2_{x|y} = \mathsf{V}\{x|y\} = \mathsf{E}\left\{(x-\mu_{x|y})^2|y\right\} = \mathsf{E}\left\{x^2|y\right\} - \mu^2_{x|y}.$$

Background on statistics

Conditional variance

Again, since the *conditional variance* $\sigma^2_{x|y}$ is a **rv**, we can compute the *mean of the conditional variance*, i.e.

$$\mathsf{E}\left\{\mathsf{V}\left\{x|y\right\}\right\} = \mathsf{E}\left\{\mathsf{E}\left\{x^2|y\right\}\right\} - \mathsf{E}\left\{\mu^2_{x|y}\right\} = \mathsf{E}\left\{x^2\right\} - \mathsf{E}\left\{\mu^2_{x|y}\right\},$$

where we make use of $\mathsf{E}\left\{\mathsf{E}\left\{g(x)|y\right\}\right\} = \mathsf{E}\left\{g(x)\right\}$ with $g(x) = x^2$.

Background on statistics

Conditional variance

Recalling the *variance* of the *conditional mean*

$$\mathsf{V}\left\{\mathsf{E}\left\{x|y\right\}\right\} = \mathsf{E}\left\{\mu_{x|y}^2\right\} - \mathsf{E}\left\{x\right\}^2,$$

and summing it with the *mean* of the *conditional variance*

$$\mathsf{E}\left\{\mathsf{V}\left\{x|y\right\}\right\} = \mathsf{E}\left\{x^2\right\} - \mathsf{E}\left\{\mu_{x|y}^2\right\},$$

we got

$$\mathsf{V}\left\{\mathsf{E}\left\{x|y\right\}\right\} + \mathsf{E}\left\{\mathsf{V}\left\{x|y\right\}\right\} = \mathsf{E}\left\{x^2\right\} - \mathsf{E}\left\{x\right\}^2.$$

Background on statistics

Marginal variance

Since

$$\mathsf{V}\{x\} = \mathsf{E}\left\{x^2\right\} - \mathsf{E}\{x\}^2 \,,$$

we finally have the expression of the *marginal variance*

$$\mathsf{V}\{x\} = \mathsf{V}\{\mathsf{E}\{x|y\}\} + \mathsf{E}\{\mathsf{V}\{x|y\}\} \,.$$

Background on statistics

Marginals

Hence the *marginal mean* and the *marginal variance* are given by

$$\mathsf{E}\left\{x\right\} = \mathsf{E}\left\{\mathsf{E}\left\{x|y\right\}\right\},$$

and

$$\mathsf{V}\left\{x\right\} = \mathsf{V}\left\{\mathsf{E}\left\{x|y\right\}\right\} + \mathsf{E}\left\{\mathsf{V}\left\{x|y\right\}\right\}.$$

Therefore, knowing the *conditional pdf* $p_c(x|y)$ (usually a *posterior* pdf for *Bayesian estimators*), we can retrieve the first and second moments of the **rv** of interest x!

Background on statistics

Conditional pmf

Sometimes, the *marginal mean* is called the *Total Expectation Law*, i.e.

$$\mathsf{E}\{x\} = \int_{-\infty}^{+\infty} \mathsf{E}\{x|y\}\, p(y)dy = \mathsf{E}\{\mathsf{E}\{x|y\}\},$$

while the *marginal variance* is called the *Total Variance Law*, i.e.

$$\mathsf{V}\{x\} = \mathsf{V}\{\mathsf{E}\{x|y\}\} + \mathsf{E}\{\mathsf{V}\{x|y\}\}.$$

Both are inspired by the *Total Probability Law*, i.e.

$$p_x(x) = \int_{-\infty}^{+\infty} p(x,y)dy = \int_{-\infty}^{+\infty} p(x|y)p_y(y)dy.$$

For pmfs, the same derivations can be adapted to the discrete case.

Background on statistics

Moments

Recall that, as for *univariate* pdfs, it is possible to compute the moments for *multivariate* pdfs using similar definitions.
In practice, for the *mean* of x_i, which is *jointly distributed* with $x = [x_1, x_2, \dots, x_n]^T$, we have:

$$\mu_i = \mathsf{E}\{x_i\} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} x_i p(x) d^n x$$

Background on statistics

Moments

For the second moments (about the mean), the matter is a little bit more complicated since we have n^2 moments, one for *each possible combination of variables*.

Hence,

$$\mathsf{Cov}\{x_i, x_j\} = \mathsf{E}\{(x_i - \mu_i)(x_j - \mu_j)\} = \\ = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}(x_i - \mu_i)(x_j - \mu_j)p(x)d^n x.$$

Moreover, if we select the same i–th element:

$$\sigma_i^2 = \mathsf{E}\left\{(x_i - \mu_i)^2\right\} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}(x_i - \mu_i)^2 p(x)d^n x.$$

Background on statistics

Moments

Since

$$\sigma_i^2 = \mathsf{E}\left\{(x_i - \mu_i)^2\right\} = \mathsf{E}\left\{x_i^2\right\} - \mu_i^2,$$

we have immediately

$$\begin{aligned}\mathsf{Cov}\left\{x_i, x_j\right\} &= \mathsf{E}\left\{(x_i - \mu_i)(x_j - \mu_j)\right\} = \mathsf{E}\left\{x_i x_j\right\} - \mu_i \mu_j = \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} x_i x_j p(x) d^n x - \mu_i \mu_j,\end{aligned}$$

that is the difference between the *correlation* and the product of the *means*.

Background on statistics

Moments

We are now in a position to draw the following definitions:

- The term $\text{Cov}\{x_i, x_j\} = \text{Cov}\{x_j, x_i\}$ is called the *covariance* between x_i and x_j;
- The term σ_i^2 is the *variance* of x_i, which is a special case of the *covariance* .

Notice that while $\sigma_i^2 \geq 0$, the *covariance* $\text{Cov}\{x_i, x_j\}$ can be positive or negative.
For example, if $\mu_x = 0$ and $y = -x$ (hence $\mu_y = 0$), we have that

$$\text{Cov}\{x, y\} = \text{E}\{xy\} = \text{E}\{x(-x)\} = -\text{E}\left\{x^2\right\} = -\sigma_x^2.$$

Background on statistics

Moments

Covariances play a fundamental role for joint pdfs since they express the degree of *linear dependence* between the variables of the multidimensional **rv** x.

In practice:

- If x_j tends to *increase linearly* away from μ_j when x_i does so with respect to μ_i, the covariance $\text{Cov}\{x_i, x_j\}$ will be large and positive;
- If x_j tends to *decrease linearly* away from μ_j when x_i *increases linearly* away with respect to μ_i, the covariance $\text{Cov}\{x_i, x_j\}$ will be large and negative;
- If there is very little linear dependency between x_i and x_j, then $\text{Cov}\{x_i, x_j\}$ will be small.

Background on statistics

Moments

Usually, the *variance* $\sigma_i^2 \geq 0$ is normalised taking the square root, which yields the *standard deviation* $\sigma_i \geq 0$.
This solution is not viable for the *covariance* $\mathsf{Cov}\left\{x_i, x_j\right\}$, which is undefined in sign.
Moreover, we know from the *Hölder inequality* that

$$|\mathsf{Cov}\left\{x_i, x_j\right\}| \leq \sqrt{\mathsf{V}\left\{x_i\right\}\mathsf{V}\left\{x_j\right\}}.$$

Background on statistics

Moments

This yields to the *correlation coefficient*:

$$\rho(x_i, x_j) = \frac{\mathsf{Cov}\{x_i, x_j\}}{\sqrt{\mathsf{V}\{x_i\}\,\mathsf{V}\{x_j\}}} = \frac{\mathsf{Cov}\{x_i, x_j\}}{\sigma_i \sigma_j} \in [-1, 1].$$

Hence:

- If x_j tends to *increase linearly* away from μ_j when x_i does so with respect to μ_i, then $\rho(x_i, x_j) \approx 1$;
- If x_j tends to *decrease linearly* away from μ_j when x_i *increases linearly* away from μ_i, then $\rho(x_i, x_j) \approx -1$;
- If there is very little linear dependency between x_i and x_j, then $\rho(x_i, x_j) \approx 0$ (the variables are *uncorrelated*).

Background on statistics

Covariance properties

It is easy to prove the following properties:

- The *covariance* and the correlation coefficient are symmetric, i.e. $\text{Cov}\{x, y\} = \text{Cov}\{y, x\}$ and $\rho(x, y) = \rho(y, x)$;
- $\text{Cov}\{ax, by\} = ab\text{Cov}\{x, y\}$, $a, b \in \mathbb{R}$;
- $\rho(ax, by) = \frac{ab}{|ab|}\rho(x, y)$, $a, b \in \mathbb{R}$;
- $\text{Cov}\{a_1x_1 + a_2x_2 + a_0, x_3\} = a_1\text{Cov}\{x_1, x_3\} + a_2\text{Cov}\{x_2, x_3\}$.

Background on statistics

Independence

We saw that the joint probability of independent events is given by the product of the probabilities.
Similarly, for $x = [x_1, x_2, \ldots, x_n]^T$ where $x_1, x_2, \ldots, x_n$ are *independent random variables*:

$$p(x) \triangleq p_1(x_1)p_2(x_2)p_3(x_3)\ldots p_n(x_n).$$

Alternatively, we can say that the n **rv**s in x are independent if *their joint pdf equals the product of the corresponding marginal densities*.
A set of random variables is called *independent and identically distributed* (or *i.i.d.*) if they are *independent* and *the marginal densities are identical*.

Background on statistics

Independence and conditional pdf

Recall that the conditional pdf is

$$p_{c_1}(x_1|x_2) = \frac{p(x_1, x_2)}{p_2(x_2)}.$$

If the **rv**s are independent

$$p_{c_1}(x_1|x_2) = \frac{p(x_1, x_2)}{p_2(x_2)} = \frac{p_1(x_1)p_2(x_2)}{p_2(x_2)} = p_1(x_1),$$

or equivalently

$$p_{c_2}(x_2|x_1) = \frac{p(x_1, x_2)}{p_1(x_1)} = \frac{p_1(x_1)p_2(x_2)}{p_1(x_1)} = p_2(x_2).$$

Background on statistics

Joint probability and independence

Suppose we have two **rv**s, we can say that

$$
\begin{aligned}
&\mathsf{Pr}\left[(a_1 \leq x_1 \leq a_1 + \delta_{a_1}) \cap (a_2 \leq x_2 \leq a_2 + \delta_{a_2})\right] = \\
&= \int_{a_1}^{a_1+\delta_{a_1}} \int_{a_2}^{a_2+\delta_{a_2}} p(x_1, x_2) dx_1 dx_2,
\end{aligned}
$$

hence $x_1, x_2 \sim p(x_1, x_2)$, i.e., x_1 and x_2 are *jointly distributed* with pdf $p(x_1, x_2)$.

Background on statistics

Joint probability and independence

Let us suppose that we want to compute the pdf of $y = x_1 + x_2$.
Then

$$P_y(a) = \mathsf{Pr}\left[y \le a\right] = \mathsf{Pr}\left[x_1 + x_2 \le a\right] = \int_{x_1+x_2\le a} p(\xi_1, \xi_2) d\xi_1 d\xi_2,$$

which defines a half plane.
For the *joint* pdf of two *independent* **rv**s we have

$$\mathsf{Pr}\left[y \le a\right] = \int_{x_1+x_2\le a} p(\xi_1, \xi_2) d\xi_1 d\xi_2 = \int_{x_1+x_2\le a} p_1(\xi_1) p_2(\xi_2) d\xi_1 d\xi_2.$$

Background on statistics

Joint probability and independence

Letting $z = \xi_1 + \xi_2$, we have

$$P_y(a) = \Pr[y \leq a] = \int_{-\infty}^{a} \int_{-\infty}^{+\infty} p_1(\xi_1) p_2(z - \xi_1) d\xi_1 dz.$$

Since $P_y(a)$ is the cdf of y, for its pdf we have (by differentiation)

$$p_y(y) = \left. \frac{dP_y(a)}{da} \right|_{a=y} = \int_{-\infty}^{+\infty} p_1(\xi_1) p_2(z - \xi_1) d\xi_1 = p_1 * p_2,$$

that is $x_1 + x_2 \sim p_1 * p_2$.
For a generic number of random variables we have

$$x_1 + x_2 + x_3 + \cdots + x_n \sim p_1 * p_2 * p_3 * \cdots * p_n,$$

that is: the pdf of the sum of n *statistically independent* **rv**s is the *convolution* of the n pdfs.

Background on statistics

Characteristic function

We know that the convolution integral in the Fourier domain is given by the product of the *Fourier transform*.

The Fourier transform of a pdf is given by

$$\bar{p}(f) \triangleq \int_{-\infty}^{+\infty} p(x) e^{-j2\pi f x} dx \leftrightarrow p(x) \triangleq \int_{-\infty}^{+\infty} \bar{p}(f) e^{j2\pi f x} df,$$

that is the *characteristic function* of the pdf $p(x)$.

Notice that

$$\bar{p}(0) = \int_{-\infty}^{+\infty} p(x) dx = 1.$$

Furthermore

$$\bar{p}(f) = \mathsf{E}\left\{ e^{-j2\pi f x} \right\}.$$

Background on statistics

Independence

If two random variables of a *bivariate* pdf are independent then:

$$\begin{aligned}\mathrm{Cov}\{x_1, x_2\} &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x_1-\mu_1)(x_2-\mu_2)p(x_1,x_2)dx_1dx_2 = \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x_1-\mu_1)(x_2-\mu_2)p_1(x_1)p_2(x_2)dx_1dx_2 = \\ &= \int_{-\infty}^{+\infty}(x_1-\mu_1)p_1(x_1)dx_1\int_{-\infty}^{+\infty}(x_2-\mu_2)p_2(x_2)dx_2 = 0.\end{aligned}$$

which states that the two variables are *uncorrelated*, i.e., $\rho = 0$.
This is true for any dimension of the **rv** vector.

Background on statistics

Independence

So, *two independent* **rv**s are also *uncorrelated*. The converse is *not* true!
There exists variables that are *uncorrelated but not independent*.
For example, consider $x_1 \sim \mathcal{N}(0,1)$ and $x_2 = x_1^2$. They cannot be independent by definition.
However:

$$\text{Cov}\{x_1, x_2\} = \text{E}\left\{(x_1 - \text{E}\{x_1\})(x_1^2 - \text{E}\left\{x_1^2\right\})\right\} = 0,$$

(prove for exercise - Hint: use integration by parts).
This apparent incongruence simply state that there is *no linear dependence* between the two variables (indeed, the dependence is quadratic).

Background on statistics

Uncorrelated variables

It turns out that for two *uncorrelated* **rv**s, we have

$$\mathsf{E}\{x_1, x_2\} = \mathsf{E}\{x_1\}\,\mathsf{E}\{x_2\}$$

Indeed, since they are *uncorrelated* their *covariance* is

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x_1-\mu_1)(x_2-\mu_2)p(x_1,x_2)dx_1dx_2 = 0,$$

hence

$$\begin{aligned}&\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}x_1x_2p(x_1,x_2)dx_1dx_2 = \\ &= \mu_2\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}x_1p(x_1,x_2)dx_1dx_2 + \\ &+ \mu_1\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}x_2p(x_1,x_2)dx_1dx_2 - \mu_1\mu_2\end{aligned}$$

Background on statistics

Uncorrelated variables

Since the first term is the *correlation* $\mathsf{E}\{x_1, x_2\}$, it follows, using the marginals that

$$\mathsf{E}\{x_1, x_2\} = \mu_1\mu_2$$

as desired.
Alternatively, we can simply notice that

$$\mathsf{Cov}\{x_1, x_2\} = \mathsf{E}\{x_1, x_2\} - \mu_1\mu_2,$$

and that $\mathsf{Cov}\{x_1, x_2\} = 0$.

Background on statistics

Uncorrelated variables: Example

Let $z = ax + by$, where x and y are two *uncorrelated* **rv**s.
The mean of z is

$$\mathsf{E}\left\{z\right\} = \mathsf{E}\left\{ax + by\right\} = a\mathsf{E}\left\{x\right\} + b\mathsf{E}\left\{y\right\} = a\mu_x + b\mu_y = \mu_z.$$

The variance of z is

$$\begin{aligned}\mathsf{V}\left\{z\right\} &= \mathsf{E}\left\{(z-\mu_z)^2\right\} = \mathsf{E}\left\{(ax + by - a\mu_x - b\mu_y)^2\right\} = \\ &= \mathsf{E}\left\{[a(x-\mu_x) + b(y-\mu_y)]^2\right\} = \\ &= a^2\mathsf{E}\left\{(x-\mu_x)^2\right\} + b^2\mathsf{E}\left\{(y-\mu_y)^2\right\} + 2ab\mathsf{E}\left\{(x-\mu_x)(y-\mu_y)\right\} = \\ &= a^2\sigma_x^2 + b^2\sigma_y^2\end{aligned}$$

Background on statistics

Uncorrelated variables: Example

Furthermore, the covariance of $z = ax + by$ and x is

$$\begin{aligned}\mathsf{E}\left\{(z-\mu_z)(x-\mu_x)\right\} &= \mathsf{E}\left\{[a(x-\mu_x)+b(y-\mu_y)](x-\mu_x)\right\} = \\ &= a\mathsf{E}\left\{(x-\mu_x)^2\right\} + b\mathsf{E}\left\{(x-\mu_x)(y-\mu_y)\right\} = \\ &= a\sigma_x^2\end{aligned}$$

hence a full linear dependence.

Background on statistics

Total probability law for pdfs

We finally present the extension to pdfs of the *Total Probability Law*, i.e.,

$$\mathsf{Pr}\left[B\right] = \sum_{i=1}^{n} \mathsf{Pr}\left[B \cap A_i\right] = \sum_{i=1}^{n} \mathsf{Pr}\left[B|A_i\right] \mathsf{Pr}\left[A_i\right]$$

where key is that the A_i are *mutually exclusive* and exhaustive with respect to the event set Ω.
The extension is based on the *marginal* and the *conditional* pdfs for bivariate, i.e.

$$p_x(x) = \int_{-\infty}^{+\infty} p(x,y)dy = \int_{-\infty}^{+\infty} p_c(x|y)p_y(y)dy.$$

Background on statistics

Total probability law for pdfs

The generalisation to an arbitrary number of pdfs is given by

$$p_{x|z}(x|z) = \int_{-\infty}^{+\infty} p(x,y|z)dy = \int_{-\infty}^{+\infty} p_c(x|y,z)p_{y|z}(y|z)dy.$$

Background on statistics

Compact form of mean and covariances

For multivariate pdfs, there is a compact form in which we can represent the mean and the covariances.
Consider a vector of random variables

$$x = [x_1, x_2, \ldots, x_n]^T.$$

The *mean* can be efficiently represented by

$$\mu = \mathsf{E}\{x\} = \begin{bmatrix} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} x_1 p(x) d^n x \\ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} x_2 p(x) d^n x \\ \vdots \\ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} x_n p(x) d^n x \end{bmatrix}$$

Background on statistics

Compact form of mean and covariances

For the *covariance* can be efficiently represented by

$$
\mathsf{E}\left\{(x-\mu)(x-\mu)^T\right\} =
$$

$$
\mathsf{E}\left\{\begin{bmatrix} (x_1-\mu_1)^2 & (x_1-\mu_1)(x_2-\mu_2) & \dots & (x_1-\mu_1)(x_n-\mu_n) \\ (x_2-\mu_2)(x_1-\mu_1) & (x_2-\mu_2)^2 & \dots & (x_2-\mu_2)(x_n-\mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ (x_n-\mu_n)(x_1-\mu_1) & (x_n-\mu_n)(x_2-\mu_2) & \dots & (x_n-\mu_n)^2 \end{bmatrix}\right\}
$$

Such a matrix is *symmetric* and *positive definite*, unless there is a linear dependence among the elements of x.

Background on statistics

Compact form of mean and covariances

For example, let $z = ax + by$, where x and y are two uncorrelated **rv**s. We already know that:

$$\mathsf{E}\{z\} = \mathsf{E}\{ax + by\} = a\mathsf{E}\{x\} + b\mathsf{E}\{y\} = a\mu_x + b\mu_y = \mu_z,$$

$$\mathsf{V}\{z\} = a^2\sigma_x^2 + b^2\sigma_y^2,$$

and

$$\mathsf{E}\{(z-\mu_z)(x-\mu_x)\} = a\sigma_x^2 \text{ and } \mathsf{E}\{(z-\mu_z)(y-\mu_y)\} = b\sigma_y^2$$

hence a full linear dependence.

Background on statistics

Compact form of mean and covariances

In such a case the *covariance matrix* of the vector $w = [x, y, z]^T$ is given by

$$\mathsf{E}\left\{(w - \mu_w)(w - \mu_w)^T\right\} = \begin{bmatrix} \sigma_x^2 & 0 & a\sigma_x^2 \\ 0 & \sigma_y^2 & b\sigma_y^2 \\ a\sigma_x^2 & b\sigma_y^2 & a^2\sigma_x^2 + b^2\sigma_y^2 \end{bmatrix}$$

which is symmetric but *positive semidefinite*.

Background on statistics

Compact form of mean and covariances

The *covariance matrix* is given by

$$
\mathsf{C}\{x\} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{bmatrix} =
$$

$$
= \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \dots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \dots & \rho_{2n}\sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1}\sigma_n\sigma_1 & \rho_{n2}\sigma_n\sigma_2 & \dots & \sigma_n^2 \end{bmatrix}
$$

where we make use of the notation $\sigma_{ij} = \mathsf{Cov}\{x_i, x_j\}$.

Background on statistics

Multivariate Gaussian pdf

For a *multivariate normally distributed pdf*, the expression of a the *vector-valued Gaussian density* is given by

$$p(x) = \frac{1}{\sqrt{(2\pi)^n |\mathsf{C}\{x\}|}} e^{-\frac{1}{2}(x-\mu)^T \mathsf{C}\{x\}^{-1}(x-\mu)} = \\ = \frac{1}{\sqrt{|2\pi \mathsf{C}\{x\}|}} e^{-\frac{1}{2}(x-\mu)^T \mathsf{C}\{x\}^{-1}(x-\mu)}$$

The components of the vector x are then said to be *jointly Gaussian*.
As for the *univariate* pdfs, the *vector of means* and the *covariance matrix* completely describe the pdf.
Notation: $|M|$ for a generic matrix M stands for the *determinant* of M.

Background on statistics

Multivariate Gaussian pdf

The following properties hold for the Gaussian density function:

- All *marginals* distributions are Normal;
- The *conditional* distribution is still Normal;
- If all the variables are mutually *uncorrelated*, so that $\mathsf{C}\{x\}$ is diagonal, the variables are also *independent*. In other words, *independence* and *zero correlation* are equivalent.

The last property can be shown by substituting a diagonal $\mathsf{C}\{x\}$ in the Gaussian pdf definition.

Background on statistics

Central limit theorem

It is evident how the Gaussian pdf has a set of relevant properties. Interestingly, the summation of **rv**s tends towards a Gaussian pdf, as stated by the *central limit theorem*:

Definition (Central limit theorem)

If a random variable y is the sum of a *sufficiently large number* of **rv**s, then y will be approximately distributed as a Gaussian, irrespective of the distribution of the components.

Background on statistics

Linear combinations of random vectors

In general, a *linear combination* of random variables can be expressed as

$$y = Lx.$$

We therefore have

$$\mathsf{E}\left\{y\right\} = \mathsf{E}\left\{Lx\right\} = L\mathsf{E}\left\{x\right\} = L\mu.$$

Moreover

$$\begin{aligned}
\mathsf{C}\left\{y\right\} &= \mathsf{E}\left\{(y - L\mu)(y - L\mu)^T\right\} = \mathsf{E}\left\{(Lx - L\mu)(Lx - L\mu)^T\right\} = \\
&= \mathsf{E}\left\{L(x-\mu)(x-\mu)^T L^T\right\} = L\mathsf{E}\left\{(x-\mu)(x-\mu)^T\right\} L^T = \\
&= L\mathsf{C}\left\{x\right\} L^T = L\Sigma_x L^T
\end{aligned}$$

Outline

Propagation of errors

Sometimes we assume, for simplicity, that the **rv**s are distributed according to a Normal pdf.
Sometimes this is implicit, by propagating *only* the mean and the variance.
This is what happens when we have the following situation: suppose that sensor readings $y \in \mathbb{R}^n$ of certain quantities $x \in \mathbb{R}^n$ are corrupted by a **rv** $\varepsilon \in \mathbb{R}^n$, i.e.,

$$y = x + \varepsilon.$$

It follows that y is a **rv** having:

$$\mu_y = \mathsf{E}\left\{y\right\} = \mathsf{E}\left\{x + \varepsilon\right\} = x + \mathsf{E}\left\{\varepsilon\right\} = x + \mu_\varepsilon$$

and

$$\mathsf{C}\left\{y\right\} = \mathsf{C}\left\{x + \varepsilon\right\} = \mathsf{E}\left\{(\varepsilon - \mu_\varepsilon)(\varepsilon - \mu_\varepsilon)^T\right\} = \mathsf{C}\left\{\varepsilon\right\} = \Sigma_\varepsilon.$$

Propagation of errors

Suppose now that

$$y = g(x) + f(\varepsilon),$$

where $g : \mathbb{R}^n \mapsto \mathbb{R}^m$ and $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ are *nonlinear* vector functions.
Now $y \in \mathbb{R}^m$, where in general $m \leq n$.
It is hard to understand what would be the pdf of y for a generic nonlinear function, even in the case of a normal pdf for ε.
However, we can approximate *locally* the propagation of the *mean* and the *covariance* through the nonlinear transformation $f(\cdot)$ using the so called *propagation of errors*.

Propagation of errors

This is possible using the *Taylor expansion* of $f(\cdot)$ about the pdf *mean*, i.e.,

$$f(\varepsilon) = f(\mu_\varepsilon) + \left.\frac{df(\varepsilon)}{d\varepsilon}\right|_{\varepsilon=\mu_\varepsilon} (\varepsilon - \mu_\varepsilon) + \mathcal{O}(\varepsilon^2)$$

Using the expansion up to the first order, we have

$$f(\varepsilon) \approx f(\mu_\varepsilon) + J(\varepsilon - \mu_\varepsilon)$$

where $J \in \mathbb{R}^{m\times n}$ is the so called *Jacobian* of $f(\cdot)$ about ε.
Therefore

$$y \approx g(x) + f(\mu_\varepsilon) + J(\varepsilon - \mu_\varepsilon).$$

Propagation of errors

It then follows that

$$\mu_y = \mathsf{E}\{y\} = \mathsf{E}\{g(x) + f(\mu_\varepsilon) + J(\varepsilon - \mu_\varepsilon)\} =$$
$$= g(x) + f(\mu_\varepsilon) + J\mathsf{E}\{(\varepsilon - \mu_\varepsilon)\} = g(x) + f(\mu_\varepsilon).$$

For the covariance matrix

$$\Sigma_y = \mathsf{C}\{y\} =$$
$$= \mathsf{E}\left\{(g(x) + f(\mu_\varepsilon) + J(\varepsilon - \mu_\varepsilon) - \mu_y)(g(x) + f(\mu_\varepsilon) + J(\varepsilon - \mu_\varepsilon) - \mu_y)^T\right\} =$$
$$= \mathsf{E}\left\{J(\varepsilon - \mu_\varepsilon)(\varepsilon - \mu_\varepsilon)^T J^T\right\} = J\Sigma_\varepsilon J^T,$$

that is, in practice, equal to the linear transformation.
Notice that this is an *approximated relation*.

Outline

Estimation and Random Variables

Random variables are a mapping from an *event set* to the *real numbers* if *continuous*, to a *discrete set of numbers* if *discrete*.

Random variables are described by their *joint pdfs*.

Given a *joint pdf*, it is always possible to reduce the *range space* by *marginalisation* and to compute *conditional pdf* to derive the relation among the **rv**s.

Knowing the *conditional pdf*, it is possible to derive the first two moments of the *conditioned* **rv**.

Jointly distributed **rv**s can be represented with the *mean* vector and the *covariance* matrix.

Independency implies *uncorrelatedness*, but the *converse is not true*!