Statistical Data Analysis for HEP

PART-2 of the course (*core* part + *in-depth* part)

Prof. Alexis Pompili (University of Bari Aldo Moro)*

Erasmus+ Teaching Mobility Program / 16-20 October 2023 @ Sofia Physics Faculty

* alexis.pompili@ba.infn.it (or alexis.pompili@cern.ch)
https://persone.ict.uniba.it/rubrica/alexis.pompili

PART 2A - CORE

alexis.pompili@ba.infn.it

PROBABILITY DENSITY FUNCTIONS

Probability Density Function (p.d.f.) - I

- ***Probability distribution function*** (aka **p.d.f.**): distribution of the probability for a RV to assume a certain value among those allowed

  In other words: **the p.d.f. of a RV is the law which rules the assumption of a certain value by the RV in one measurement/experiment**

  We will see during this course that: **the link between experiment and theoretical model indeed happens through the p.d.f., that is predicted by the model to describe (the result of) an experiment**

- Consider a discrete random variable x having more than one possible elementary result, that is $(x_1, \dots, x_N)$ each occurring with a probability $P(x_i)$, where $i = 1, \dots, N$, thus *associated* to each of the possible results.
  The function that associates the probability $P(x_i)$ to each possible value x_i is called **probability distribution**.
  Note : the result of an event is not predictable but - instead - the probability distribution of the results can be known.

Probability Density Function (p.d.f.) - I

- ***Probability distribution function*** (aka **p.d.f.**): distribution of the probability for a RV to assume a certain value among those allowed

  In other words: **the p.d.f. of a RV is the law which rules the assumption of a certain value by the RV in one measurement/experiment**

  We will see during this course that: **the link between experiment and theoretical model indeed happens through the p.d.f., that is predicted by the model to describe (the result of) an experiment**

- Consider a discrete random variable x having more than one possible elementary result, that is $(x_1, \ldots, x_N)$ each occurring with a probability $P(x_i)$, where $i = 1, \ldots, N$, thus *associated* to each of the possible results.
  The function that associates the probability $P(x_i)$ to each possible value x_i is called **probability distribution**.
  Note : the result of an event is not predictable but - instead - the probability distribution of the results can be known.

  The probability of a random event ***E*** corresponding to a set of distinct possible elementary results $(x_{E_1}, \ldots, x_{E_K})$ where $x_{E_j} \in \Omega = (x_1, \ldots, x_N)$ for all $j = 1, \ldots, K$, is, according to the 3rd Kolmogorov's axiom, given by:

$$P\left(\bigcup_{j=1}^{K}\{x_{E_j}\}\right) = P(\{x_{E_1}, \ldots, x_{E_K}\}) = P(E) = \sum_{j=1}^{K} P(x_{E_j})$$

  From the 2nd Kolmogorov's axiom, the probability of the event Ω corresponding to the set of **all** possible values must be: $\sum_{i=1}^{N} P(x_i) = 1$

  (normalization condition)

  From the 1st Kolmogorov's axiom: $P\left(x_{E_j}\right) \geq 0 \ \forall j \Longrightarrow P(E \subset \Omega) \geq 0$

Probability Density Function (p.d.f.) - II

- Most quantities of interest to us are continuous, thus we will treat mainly the continuous case.
  The discrete probability introduced in the previous slide can be generalized to the continuous case with the replacement … $\sum_{\Omega} \Rightarrow \int_{\Omega}$

In the discrete case we deal with a **genuine probability function**; in the continuous case we must introduce a **probability density function**!

- Let us consider a sample space $\Omega \subseteq \mathbb{R}^n$. Each random experiment will lead to a measurement corresponding to one point $\vec{x} \in \Omega$.
  We can associate a probability density $f(\vec{x}) = f(x_1, \ldots, x_n)$ to any point $\vec{x} \in \Omega$. Of course, $f(\vec{x}) \geq 0$ (1^{st} $axiom$).

  The probability of an event A with $\mathrm{A} \subseteq \Omega$, namely the probability that $\vec{x} \in A$ is given by : $P(A) = \int_A f(x_1, \ldots, x_n) d^n x$

  The function $f(\vec{x})$ is called **probability density function p.d.f.** ! The function $f(x_1, \ldots, x_n) d^n x$ can be interpreted as differential probability.

  The normalization condition can be expressed as: $\int_{\Omega} f(x_1, \ldots, x_n) d^n x = 1$

-

Probability Density Function (p.d.f.) - II

- Most quantities of interest to us are continuous, thus we will treat mainly the continuous case.
  The discrete probability introduced in the previous slide can be generalized to the continuous case with the replacement … $\sum_{\Omega} \Rightarrow \int_{\Omega}$

In the discrete case we deal with a **genuine probability function**; in the continuous case we must introduce a **probability density function**!

- Let us consider a sample space $\Omega \subseteq \mathbb{R}^n$. Each random experiment will lead to a measurement corresponding to one point $\vec{x} \in \Omega$.
  We can associate a probability density $f(\vec{x}) = f(x_1, \dots, x_n)$ to any point $\vec{x} \in \Omega$. Of course, $f(\vec{x}) \geq 0$ $(1^{st}\ axiom)$.

The probability of an event A with A $\subseteq \Omega$, namely the probability that $\vec{x} \in A$ is given by : $P(A) = \int_A f(x_1, \dots, x_n) d^n x$

The function $f(\vec{x})$ is called **probability density function p.d.f.** ! The function $f(x_1, \dots, x_n) d^n x$ can be interpreted as differential probability.

The normalization condition can be expressed as: $\int_{\Omega} f(x_1, \dots, x_n) d^n x = 1$

- In 1 dim: Probability of the outcome X to be within the continuous interval of possible values $[x, x+dx]$ is $P(x \leq X \leq x+dx) = f(x) \cdot dx$

  The **p.d.f.** $f(x)$ is of course normalized by the condition : $\int_{-\infty}^{+\infty} f(x) dx = 1$

It can be verified that :

the p.d.f. corresponds to an histogram of the RV x normalized to the unity area in the limit for which … - the bin width → 0
- the total # of entries →∞

Cumulative Distribution Function (c.d.f.)

The cumulative distribution function (c.d.f.) is the probability that the value of a r.v. will be $\leq$ a specific value. The c.d.f. is denoted by the capital letter corresponding to the small letter signifying the p.d.f. The c.d.f. is thus given by

$$\boxed{F(x) = \int_{-\infty}^{x} f(x')\,dx' = P(X \leq x)}$$

Clearly, $F(-\infty) = 0$ and $F(+\infty) = 1$.

Properties of the c.d.f.:

- $0 \leq F(x) \leq 1$
- $F(x)$ is monotone and not decreasing.
- $P(a \leq X \leq b) = F(b) - F(a)$
- $F(x)$ discontinuous at x implies

  $$P(X = x) = \lim_{\delta x \to 0} [F(x + \delta x) - F(x - \delta x)] \ \text{, i.e., the size of the jump.}$$

- $F(x)$ continuous at x implies $P(X = x) = 0$.

The c.d.f. can be considered to be more fundamental than the p.d.f. since the c.d.f. is an actual probability rather than a probability density. However, in applications we usually need the p.d.f. Sometimes it is easier to derive first the c.d.f. from which you get the p.d.f. by

$$f(x) = \frac{\partial F(x)}{\partial x} \tag{2.4}$$

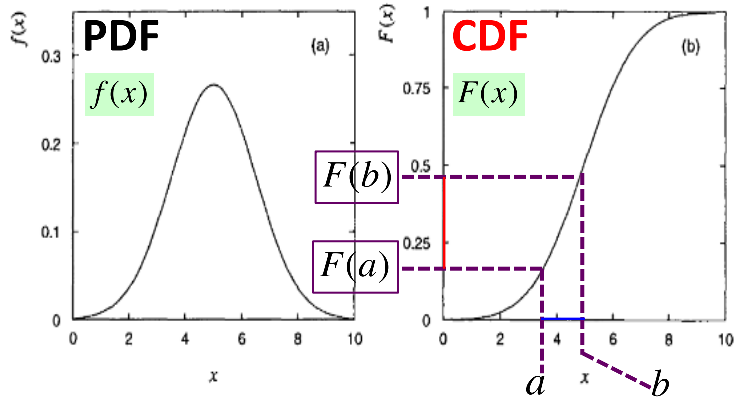
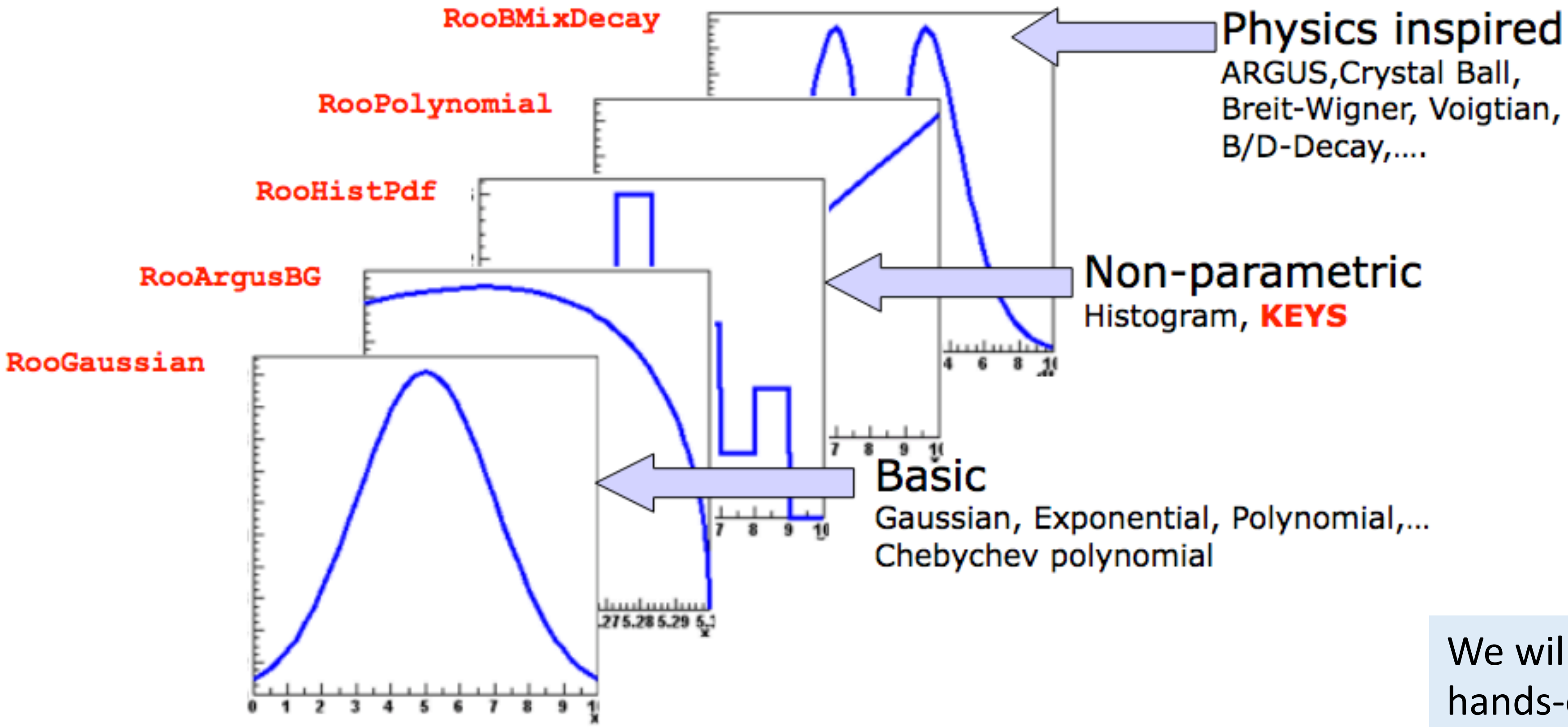


Fig. 1.3 (a) A probability density function $f(x)$. (b) The corresponding cumulative distribution function $F(x)$.

Note: the p.d.f. for F is **uniformly distributed** in [0,1]: $\dfrac{dP}{dF} = \dfrac{dP}{dx} \cdot \dfrac{dx}{dF} = \dfrac{f(x)}{f(x)} = 1$

- RooFit provides a collection of compiled standard PDF classes



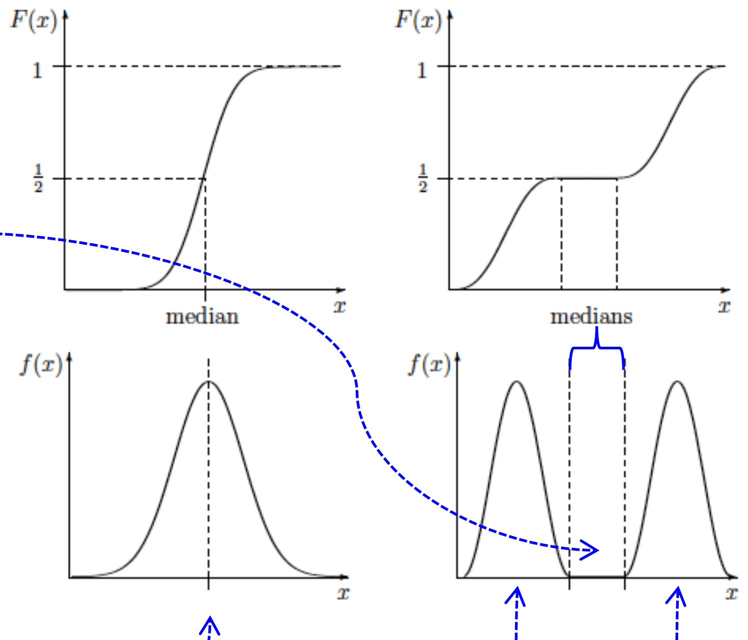
We will use them in the hands-on exercises in the lab

Easy to extend the library: each p.d.f. is a separate C++ class

Attributes of a p.d.f. : mode & median

- **Median of a p.d.f.** : value of ***x*** for which $F(x)=1/2$
  (it divides the distribution in 2 parts with the same area)

Note : **the median is not always well defined**
since there can be more than one such value of x



- **Mode of a p.d.f.** : the location of a maximum of f(x)
  (value of x that in an infinite sampling would appear the highest number of times)

Note : **a p.d.f. can be *multimodal*** !

Note : in this example … mode and median coincide

- **Expectation value of a p.d.f.** (sometimes called "***Mean***" which is very misleading actually! Better ***population mean***): represents the **central value of a p.d.f.** and it is defined as: $$\mu \equiv E[x] = \int_{-\infty}^{+\infty} x\, f(x)\, dx$$

Note: $E[x]$ is not a function of $\boldsymbol{x}$ (there is an integral on $\boldsymbol{x}$!) but depends on the distribution of the values taken by x (that is on the shape of the p.d.f.)

The mean is often a good measure of location, *i.e.*, it frequently tells roughly where the most probable region is, but not always.

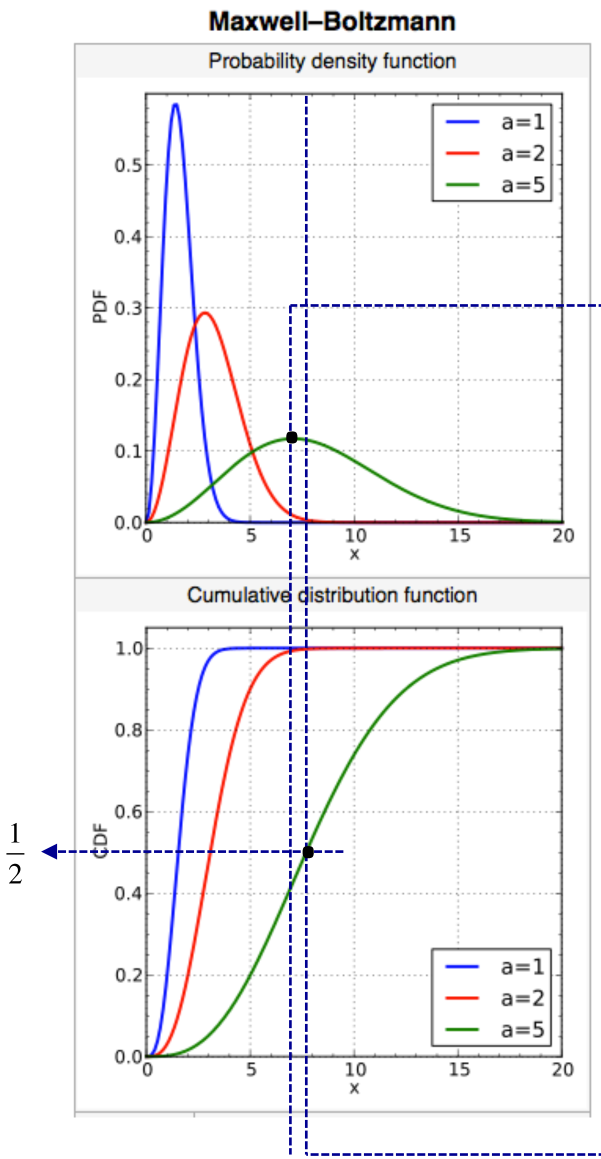
it can even happen that it is a value never taken by the x !

Properties: $a = \text{cost} \Rightarrow E[a] = a$ & $E[ax] = a \cdot E[x]$

if $\boldsymbol{u}$ is a function of $\boldsymbol{x}$: $E[au(x)] = a \cdot E[u(x)]$ where $E[u(x)] = \int_{-\infty}^{+\infty} u(x)\, f(x)\, dx$

$\boldsymbol{E}$ is a linear operator: $E[a_1 u(x) + a_2 \upsilon(x)] = a_1 E[u(x)] + a_2 E[\upsilon(x)]$

Attributes of a p.d.f. : example of the Maxwell-Boltzmann distributuion



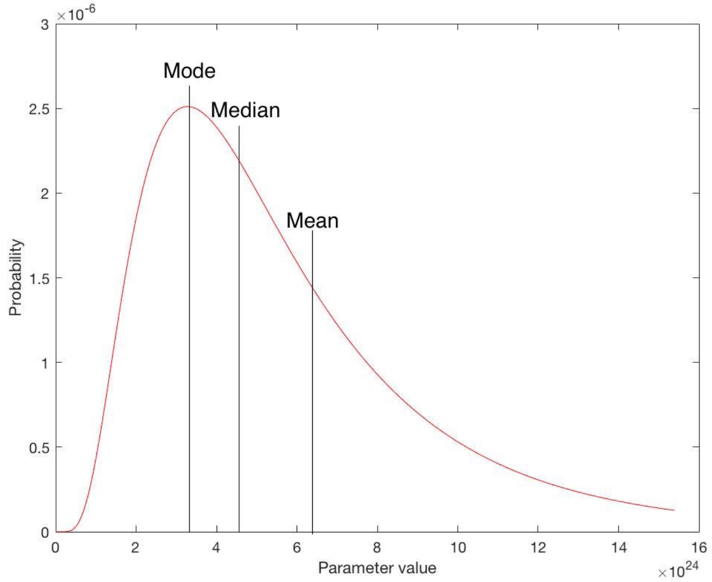
Mode

***Median* > Mode**

Example: distribution of the squared velocity of the gas molecules exiting the hole of a cavity/container

For this distribution:

the expectation value ("Mean") > Median



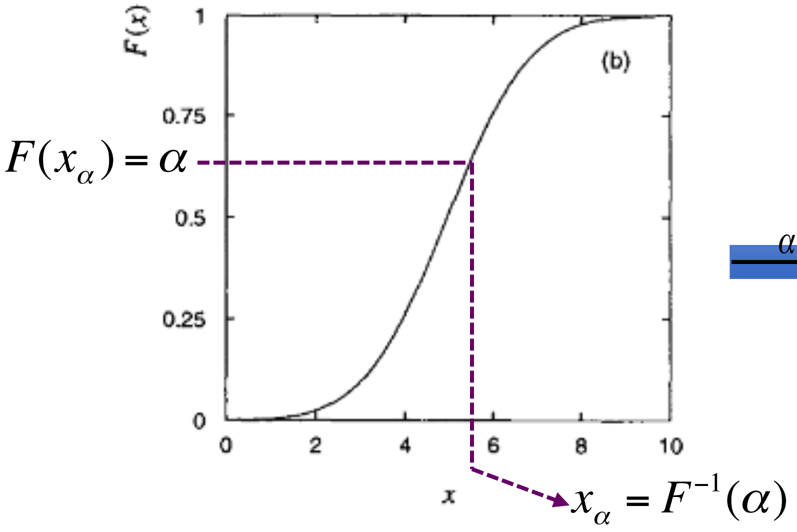
(note: this is the effect of the large tail on the right)

A useful concept related to the cumulative distribution is the so-called **quantile of order α** or **α-point**. The quantile x_α is defined as the value of the random variable x such that $F(x_\alpha) = \alpha$, with $0 \leq \alpha \leq 1$. That is, the quantile is simply the inverse function of the cumulative distribution,

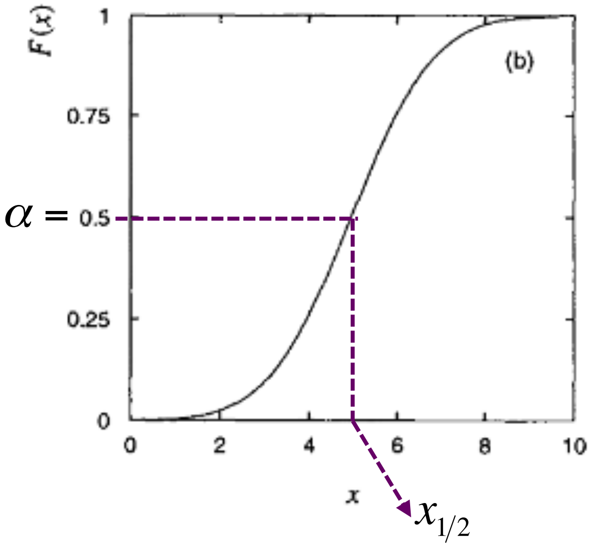
$$x_\alpha = F^{-1}(\alpha). \tag{1.17}$$

A commonly used special case is $x_{1/2}$, called the **median** of x. This is often used as a measure of the typical 'location' of the random variable, in the sense that there are equal probabilities for x to be observed greater or less than $x_{1/2}$.

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



α=1/2



Attribute of a p.d.f. : central moments

- The moments are particular expectation values. The **moments of order m** are defined as: $E[x^m] = \int_{-\infty}^{+\infty} x^m f(x)dx$.

  Therefore: **moment of order 1 $\equiv$ expectation value**

- It is possible to introduce also the ***central*** **moments of order m**, defined as: $E[(x-\mu)^m] = \int_{-\infty}^{+\infty} (x-\mu)^m f(x)dx$.

  Note: if μ is finite … the central moment of order 1 is null for any μ :

  $$E[(x-\mu)^{m=1}] = \int_{-\infty}^{+\infty} (x-\mu) f(x)dx = \int_{-\infty}^{+\infty} x f(x)dx - \mu \underbrace{\boxed{\int_{-\infty}^{+\infty} f(x)dx}}_{=1 \text{ (normalization)}} = \int_{-\infty}^{+\infty} x f(x)dx - \mu = E[x] - \mu = \mu - \mu = 0$$

  Note also: if $f(x)$ is symmetric … **the central moments of odd orders ($m = 1, 3, 5, \ldots$) are null** !

- The **central moment of order 2 is called variance and represents the spread of the $f(x)$ around the expectation value.**

  See details next slide!

- **Variance of a p.d.f.** is defined as: $\sigma_x^2 = \mathrm{V}[x] = E[(x-\mu)^2] = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx$

$$= \int_{-\infty}^{+\infty} x^2 f(x)dx - 2\mu \underbrace{\int_{-\infty}^{+\infty} x f(x)dx}_{=\mu} + \mu^2 \underbrace{\int_{-\infty}^{+\infty} f(x)dx}_{=1 \text{ (norm.)}}$$

$$= E[x^2] - 2\mu^2 + \mu^2 = E[x^2] - \mu^2 = \boxed{E[x^2] - (E[x])^2}$$

- The **squared root of the variance** is called standard deviation of x **and denoted by** $\boldsymbol{\sigma_x}$.

  It is often useful because **it has the same dimentional units of** $\boldsymbol{x}$ and thus …

  … it represents the *spread* of the p.d.f. around its expectation value.

Property: $V[ax] = a^2 \cdot V[x]$, with $a = cost$.

Indeed: $\mathrm{V}[ax] = E[a^2x^2] - (E[ax])^2 = a^2\, E[x^2] - (aE[x])^2 = a^2 \cdot (E[x^2] - (E[x])^2) = a^2 \cdot V[x]$

Note: other attibutes like **skewness** (***asymmetry* indicator**) and **kurtosis** **(*sharpness* indicator)** are defined in the in-depth part.

Mixture of subsamples - I

- Often a data sample under analysis is the **sum of two (or more) subsamples distributed according to different p.d.f.s;** an obvious example is the sum of a certain signal and one (or more) background(s).

Let's express the **fractions** of events belonging to each subsample as φ_i and the $f_i(x)$ are the corresponding p.d.f.s of the r.v. x ; then:

overall p.d.f. : $f(x) = \sum_i \varphi_i f_i(x)$

An obvious example is ($i \equiv 1$ for signal, $i \equiv 2$ for background): $f_{tot}(x) = \varphi_{sig} f_{sig}(x) + \varphi_{bkg} f_{bkg}(x) = \varphi_{sig}\, f_{sig}(x) + (1 - \varphi_{sig}) f_{bkg}(x)$

- The (overall) expectation value of x is (where μ_i represents the expectation value of x for each subsample):

$$\mu \equiv E[x] = \int_{-\infty}^{+\infty} x f(x) dx = \sum_i \varphi_i \cdot \int_{-\infty}^{+\infty} x f_i(x) dx = \sum_i \varphi_i E_i[x] \equiv \sum_i \varphi_i \mu_i$$

$\Leftrightarrow$ the overall expectation value is the mean of the expectation values weighted by the relative fractions in the mixture

- For the variance see next slide!

- The variance of a r.v. x for a mixture of subsamples is:

$$V[x] = E\left[(x-\mu)^2\right] = \sum_i \varphi_i \cdot E_i\left[(x-\mu)^2\right]$$ where $\mu = \sum_i \varphi_i \mu_i$ is the expectation value over the mixture

To get a more familiar expresion we must introduce the deviations $\delta_i = (\mu - \mu_i)$ and introduce in the upper expression $\mu = \delta_i + \mu_i$;

with some algebra (reported in a slide in the in-depth part) we can get:

$$V[x] = \sum_i \varphi_i \cdot \left\{ V_i[x] + \underbrace{\boxed{\left[\sum_{j \neq i} \varphi_i(\mu_j - \mu_i)\right]^2}}_{\geq 0} \right\}$$

⟺ generally, the variance is **not** just the simple mean of the variances of the sub-samples weighted by the relative fractions in the mixture, since it is always augmented because of the fact that sub-samples can have different expectation values

- On the other hand, …

($) $V[x] = \sum_i \varphi_i \cdot V_i[x]$ **IF** $\mu_i = \mu \; \forall i$ ⟺ **IF** all the distributions of r.v. x for each sub-sample in the mixture are characterized by the **same** expectation value …

… the overall variance is the mean of the variances weighted by relative fractions in the mixture

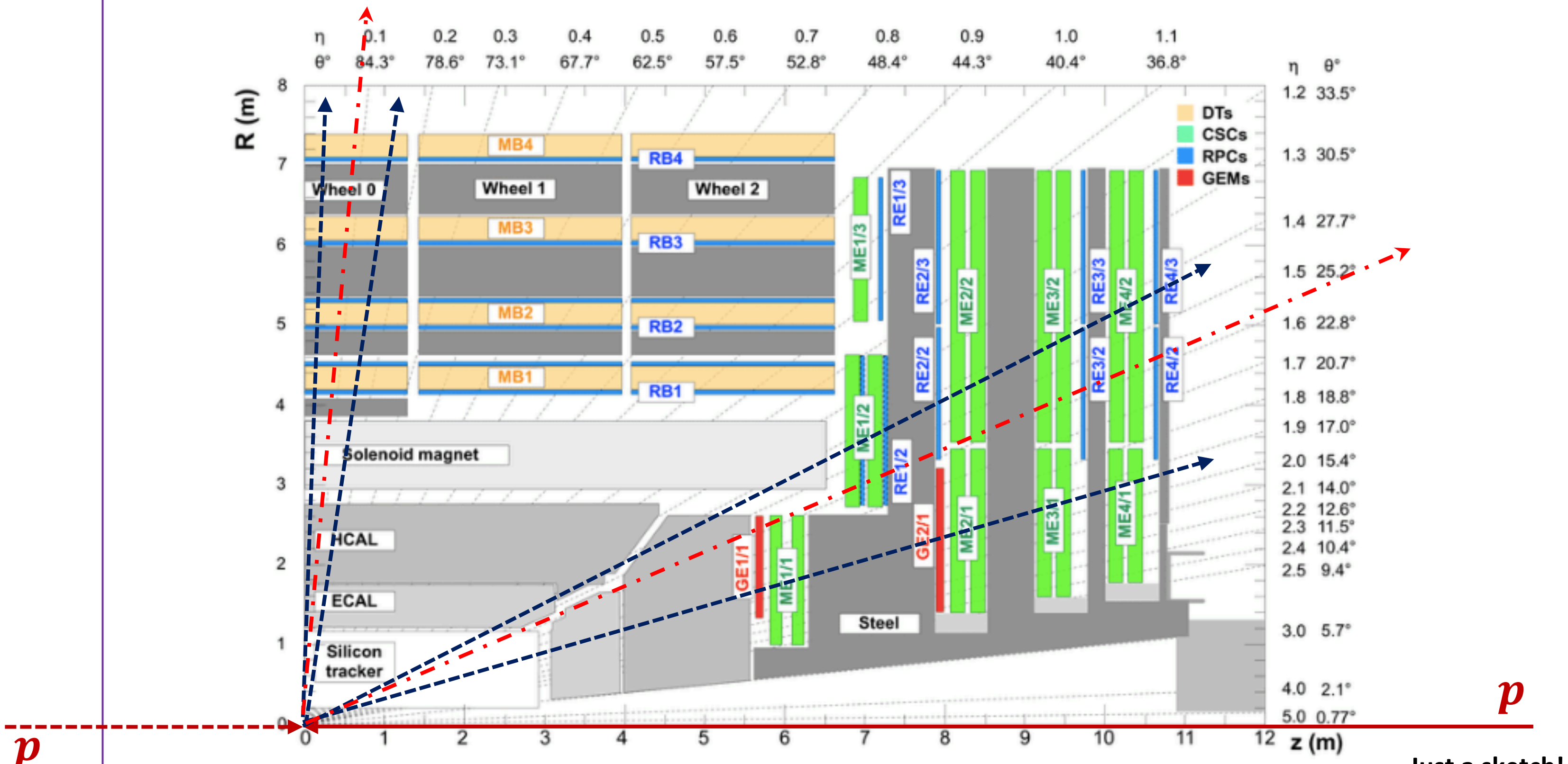
(see an example in the following slide)

Mixture of subsamples - example - I

- Example: suppose to reconstruct with the CMS detector the dimuon decays of the charmonium state $\psi(2S)$: $\psi(2S) \to \mu^+\mu^-$

This is a quadrant of the CMS detector showing the subdetectors of the muon system (including the proposed GEM detectors):

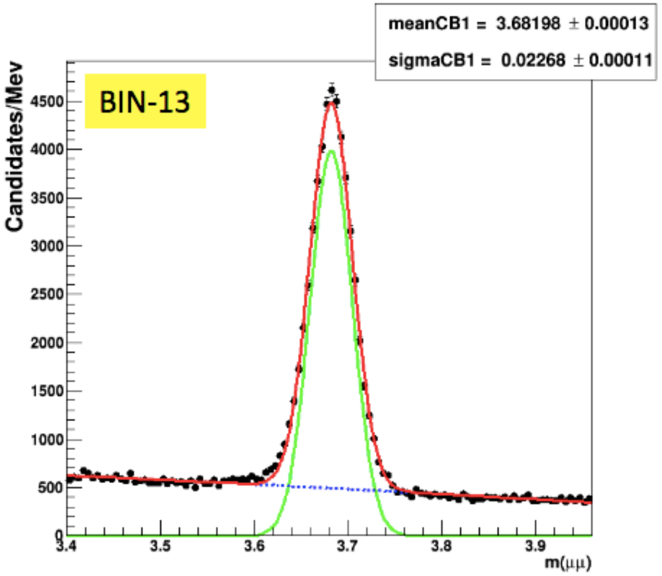
[note: pseudorapidity is the rapidity for *massless* particles: $y = \frac{1}{2}\ln\frac{1+\beta\cos\theta}{1-\beta\cos\theta} \to \frac{1}{2}\ln\frac{1+\cos\theta}{1-\cos\theta} = -\ln\tan\frac{\theta}{2} = \eta$]



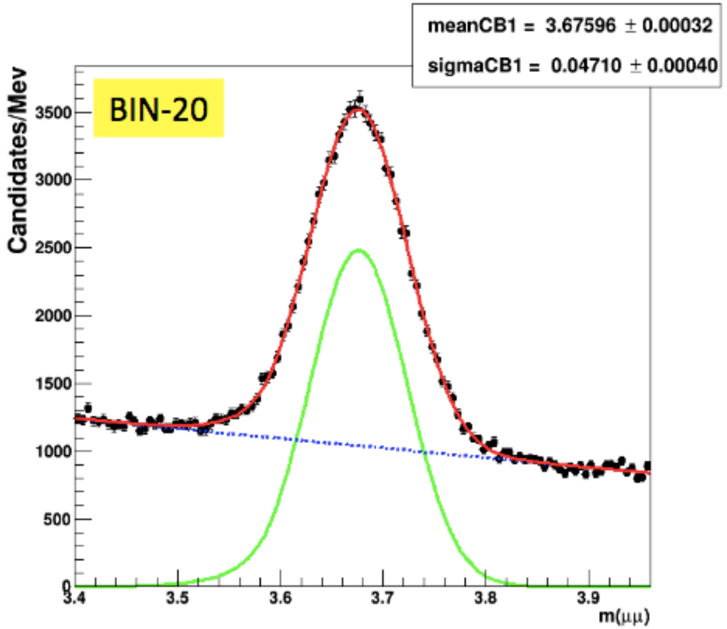
Just a sketch!

Suppose to put together two signal sub-samples of $\psi(2S)$ candidates, one with $y \in [0.,0.2]$ and the other with $y \in [1.4,1.6]$ [we neglect the combinatorial background of 2 μs (pairs by random combinations)]. The r.v. represented by the reconstructed mass, $m(\mu\mu)$, is characterized, in the two sub-samples, by the **same expectation value** [the mass of the $\psi(2S)$]; instead, the two standard deviations (square root of the two variances), that represent the mass resolution, are different for the two sub-samples since the mass resolution depends on the quality of the track reconstruction of the two μs which - in turn - depend on the detection technology of the μ-chambers: the DTs ensures a better quality w.r.t. the CSCs.

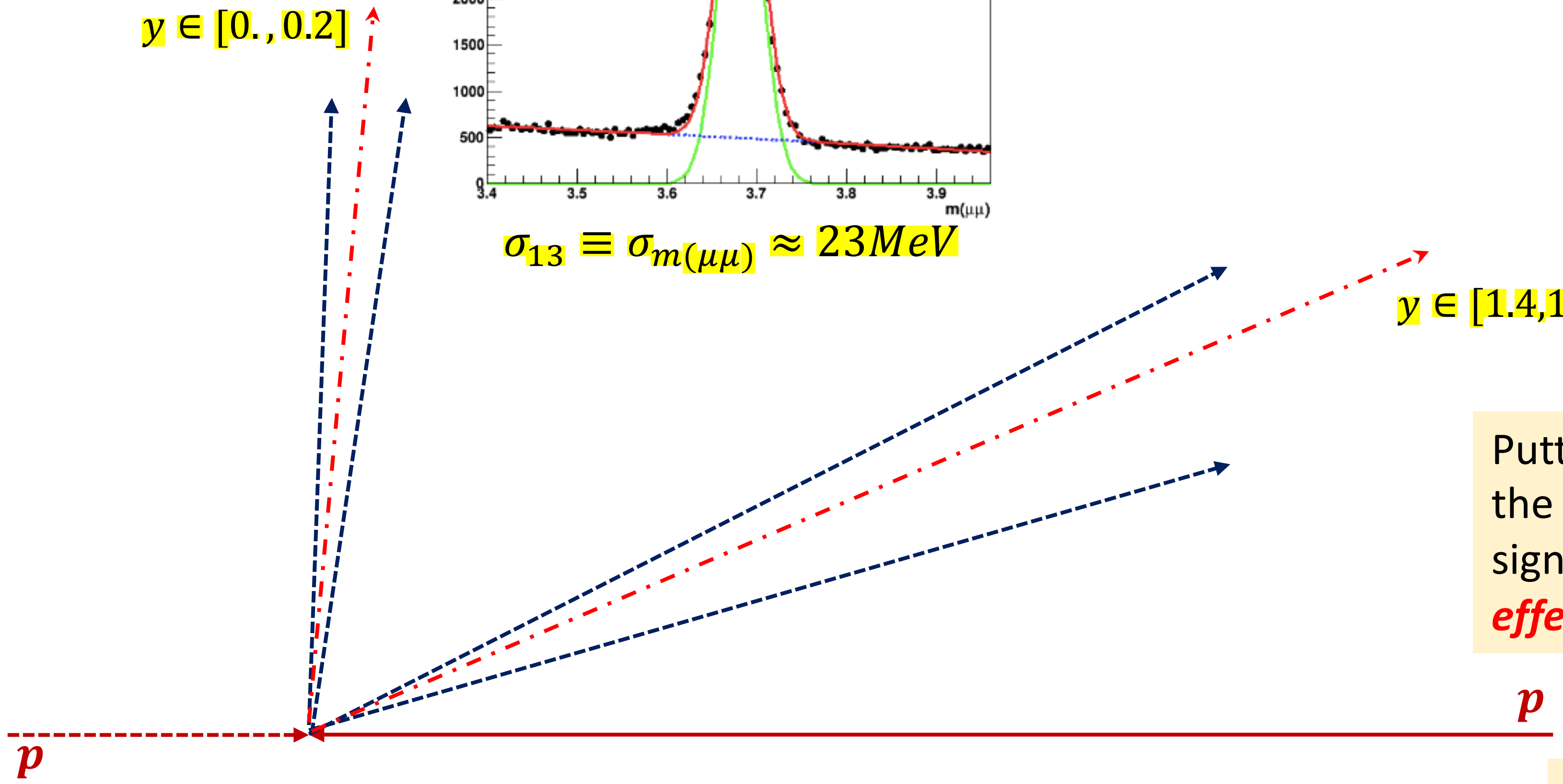
Mixture of subsamples - example - II



$\sigma_{13} \equiv \sigma_{m(\mu\mu)} \approx 23MeV$



$\sigma_{20} \equiv \sigma_{m(\mu\mu)} \approx 47MeV$



Putting together the two subsamples I would get the sum of the 2 distributions (in each one the signal can be fitted with a gaussian) and the ***effective*** *mass resolution* is expected to be:

$$\boldsymbol{\sigma_{eff\,(13+20)}} \overset{(\$)}{=} \sqrt{\boldsymbol{\varphi_{13}\,\sigma_{13}^2 + \varphi_{20}\,\sigma_{20}^2}}$$

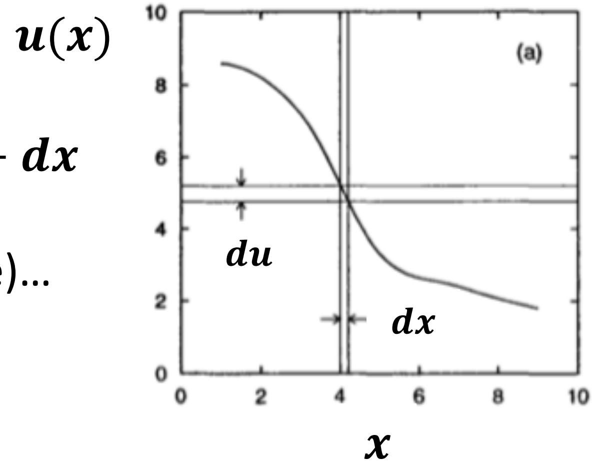
[$\boldsymbol{\varphi_{13}}$ and $\boldsymbol{\varphi_{20}}$ can be derived by the signal ***yields***]

FUNCTIONS of a R.V.

- Often experimentalists carry out **indirect measurements**, i.e the observable of interest is a function of direct measurements. For this reason we need to introduce functions of random variables! First of all have in mind that: **functions of random variables are random variables themselves** !

Suppose $u(x)$ is a continuous function of a continuous random variable x distributed according to the p.d.f. $f(x)$.

The question now is: what is the p.d.f. $\mathrm{g}(u)$ that describes the distribution of $u(x)$?



It is possible to answer requiring that the probability of x to assume values between x and $x + dx$ has to be equal to the probability for $u(x)$ to get values between u and $u + du$.
If the fuction $u(x)$ can be inverted to obtain $x(u)$ and the trasformation is 1-to-1 (i.e. bijective)…
… then we can write:

$$\boxed{\mathrm{g}(u)du = f(x)dx} \;(*) \Rightarrow \quad \mathrm{g}(u) = \frac{f(x)}{\left|\frac{du}{dx}\right|} = \frac{f(x)}{u'(x)}$$

we put the absolute value so that **g** is positive defined

(Note: du and dx may have same or odd signs!)

Since the function is a random variable:

$$E[u] = \int_{-\infty}^{+\infty} u\, \mathrm{g}(u)du \overset{(*)}{=} \int_{-\infty}^{+\infty} u(x)\, f(x)dx$$ (this will be used later!)

$$V[u(x)] = E[(u(x) - E[u(x)])^2]$$ … but … how can we calculate $E[u(x)]$?

- We can develop in series the $u(x)$ in an interval of x around μ ; thus we can substitute $u(x)$ with its development in series and for simplicity we can stop to the 2nd order:

$$u(x) \rightarrow \left.\frac{\partial u}{\partial x}\right|_{x\sim\mu} \cdot (x-\mu) + \frac{1}{2!}\left.\frac{\partial^2 u}{\partial x^2}\right|_{x\sim\mu} \cdot (x-\mu)^2 + \cancel{\ldots\ldots}$$

The substitution is applied inside the expression $E[u(x)] = \int_{-\infty}^{+\infty} u(x)\, f(x) dx$ …and after a bit of algebra one gets:

$$E[u(x)] \cong \underbrace{\boxed{E[u(\mu)]}}_{= u(\mu)} + \frac{1}{2}\left.\frac{\partial^2 u}{\partial x^2}\right|_{x\sim\mu} \cdot V[x]$$

Conclusions: 1) unless $V[x] = 0$ … the expectation value of $u(x)$ is **not** equal to the value of the function calculated with the expectation value of x, namely $u(\mu)$ ⇒ $E[u(x)] \neq u(\mu)$

2) if $u(x)$ is a linear function of x then $E[u(x)] = u(\mu)$

3) if $\left.\frac{\partial^2 u}{\partial x^2}\right|_{x\sim\mu}$ is small (slowly-varying shape) **this equality holds with a good approximation:** $E[u(x)] \approx u(\mu)$

Dealing with more than one R.V. : MARGINAL & CONDITIONAL PDFs

Case of more than 1 random variable

- If the measurement is characterized not by just one observable but instead by more than one it means …
  … we have to deal with more than 1 random variable and specifically with a vector of random variables $\vec{x} = (x_1, \dots, x_N)$;
  the associated p.d.f. would be $f(\vec{x})$. Its meaning is as follows:

  for an infinitesimal volume centered on $\vec{x}$ of sides $dx_1, \dots, dx_N$ that
  we label as $I_{\vec{x},d\vec{x}}$, the associated probability can be expressed as … $P(\vec{X} \in I_{\vec{x},d\vec{x}}) = f(\vec{x})d\vec{x}$

  We will discuss the easiest case of two r.v.s in the net slides!

- In general, this will be a complicated multi-dimentional, **unless $x_1, \dots, x_N$ are all independent among each other**
  … and in this particular case the expression of $f(\vec{x})$ is the following product:

$$f(\vec{x}) = \prod_i f_i(x_i) \quad \text{(where } f_i \text{ is the p.d.f. of } x_i \text{)}$$

We will come back to this possible *factorization* soon, in next slides!

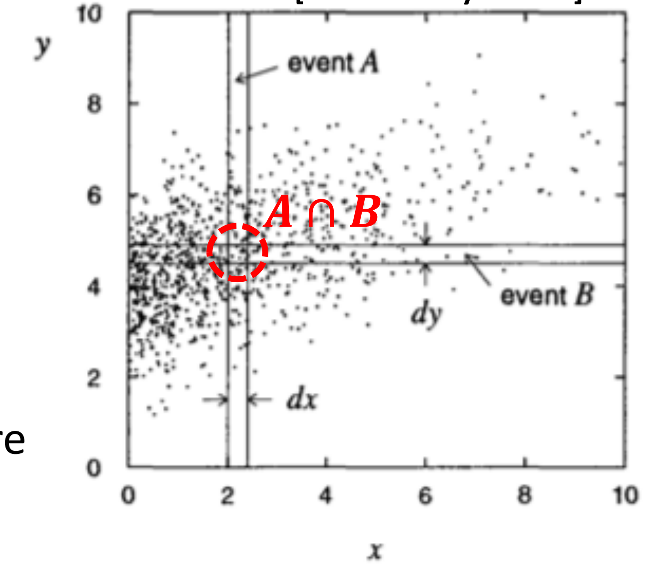
- Let's consider - in the following - to deal with **only 2** random variables: x & y !
Let's also continue to imagine to be working in the infinite sample assumption
(infinite points (x,y) in the plot): we deal with an (infinite) population, not a finite sample!

As depicted in the ***scatter plot*** in the figure, we consider :

Event **A** (vertical narrow band): observe x-values in $[x, x + dx]$ and y-values everywhere

Event **B** (horizontal narrow band): observe y-values in $[y, y + dy]$ and x-values everywhere

[borrowed by Cowan]



The event $A \cap B$ is associated to the intersection of the two bands.
Its associated probability can be expressed in terms of a joint p.d.f. (corresponding to the **density of points**) :

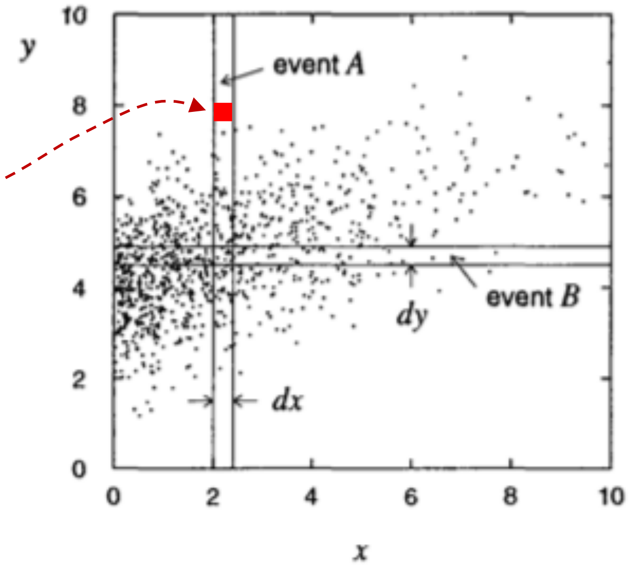
$$P(A \cap B) = P(x \in [x, x + dx], y \in [y, y + dy]) = f(x, y) dx\, dy$$

The relative normalization condition can be expressed as: $\iint_\Omega f(x,y)\, dx\, dy = 1$

- Suppose we want to know the probability for the r.v. $\boldsymbol{x}$ to get values in the interval $[\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{dx}]$ independently from the value taken by the other r.v. $\boldsymbol{y}$, i.e. we want to know the **probability of event A** (the vertical band in the scatter plot).

  The band can be considered as the set of $\boldsymbol{N}$ squares of area $\boldsymbol{dxdy_i}$ with the running index exhausting the full band:

$$P(A) = \sum_i f(x, \boldsymbol{y_i}) d\boldsymbol{y_i}\, \boldsymbol{dx} \equiv \boldsymbol{f_x(x)dx}$$



  In the limit of infinitesimal all equal intervals one gets $\boldsymbol{dy_i} = \boldsymbol{dy}$ and the sum becomes an integral ($\sum_i dy_i \rightarrow \int_{-\infty}^{+\infty} dy$)

- We can now introduce the concept of … **marginal p.d.f.** which is the **p.d.f. of 1 only random variable *once* the dependency from the other(s) is eliminated via integration of the joint p.d.f.** :

marginal p.d.f. in x : $\boldsymbol{f_x(x) = \int_{-\infty}^{+\infty} f(x,y)dy}$, **marginal p.d.f. in y** : $\boldsymbol{f_y(y) = \int_{-\infty}^{+\infty} f(x,y)dx}$

Note: the 2 marginal p.d.f.s correspond to the ***normalized* functions obtained by *projection* of the scatter plot** on the $\boldsymbol{x}, \boldsymbol{y}$ axes (again - implicitely - in the limit of infinite entries in the scatter plot) [see next slide].

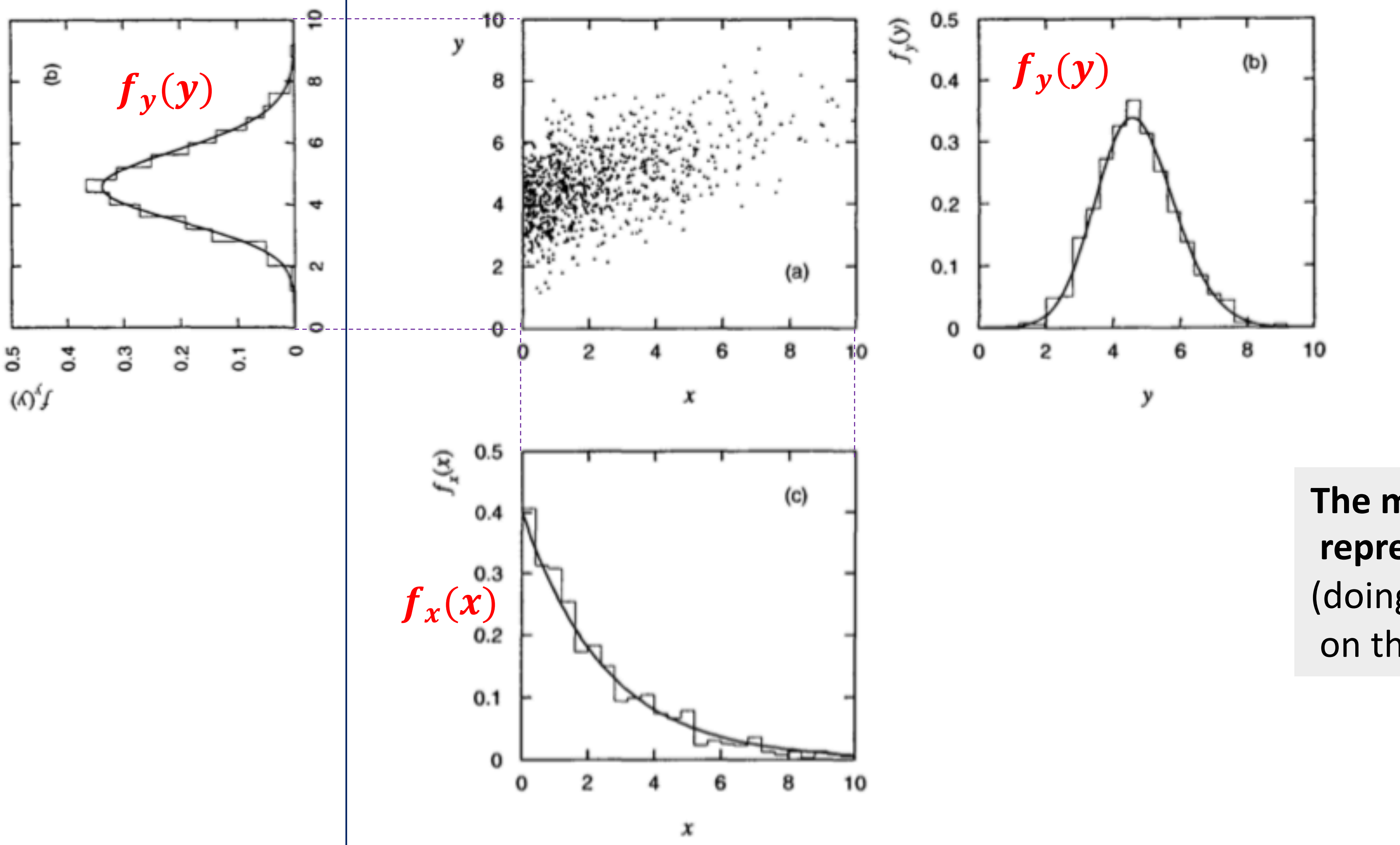


Fig. 1.5 (a) The density of points on the scatter plot is given by the joint p.d.f. $f(x, y)$. (b) Normalized histogram from projecting the points onto the y axis with the corresponding marginal p.d.f. $f_y(y)$. (c) Projection onto the x axis giving $f_x(x)$.

[borrowed by Cowan]

The marginal p.d.f.s can be easily represented as *normalized* projections
(doing a projection means integrating on the other variable)

Two random variables & Conditional p.d.f. - I

- It is now possible to introduce the concept of **conditional p.d.f.** exploiting the definition of conditional probability :

Probability for r.v. $\boldsymbol{y}$ to get values in the interval $[\boldsymbol{y}, \boldsymbol{y}+\boldsymbol{dy}]$ for any value taken by the r.v. $\boldsymbol{x}$ (event B), once it happened that $\boldsymbol{x}$ has got values in the interval $[\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{dx}]$ for any value taken by the r.v. $\boldsymbol{y}$ (event A)

is given by... $P(B|A) = \dfrac{P(A \cap B)}{P(A)} = \dfrac{\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})\boldsymbol{dx}\,\boldsymbol{dy}}{\boldsymbol{f_x}(\boldsymbol{x})\boldsymbol{dx}}$

joint p.d.f.

marginal p.d.f.

At this point it makes sense to introduce the...
conditional p.d.f. associated to the r.v. $\boldsymbol{y}$ given the r.v. $\boldsymbol{x}$
(function of the $\boldsymbol{y}$ only since $\boldsymbol{x}$ has taken a specific value) as ...

$$\boldsymbol{h}(\boldsymbol{y}|\boldsymbol{x}) = \frac{\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})}{\boldsymbol{f_x}(\boldsymbol{x})} = \frac{\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})}{\int_{-\infty}^{+\infty} f(x,y')dy'}$$

In other words: the **conditional p.d.f. of $\boldsymbol{y}$** is defined starting from the joint p.d.f. in which $\boldsymbol{x}$ has taken a specific vaue (thus, it is constant), **renormalized** so that it has unit area when integrating on $\boldsymbol{y}$ only)
(always - implicitely - in the limit of infinite entries in the scatter plot)

Similar considerations exchanging the role of $\boldsymbol{x}$ and $\boldsymbol{y}$ brings to:

$$\mathbf{g}(\boldsymbol{x}|\boldsymbol{y}) = \frac{\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})}{\boldsymbol{f_y}(\boldsymbol{y})} = \frac{\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})}{\int_{-\infty}^{+\infty} f(x',y)dx'}$$

Two random variables & Conditional p.d.f. - II

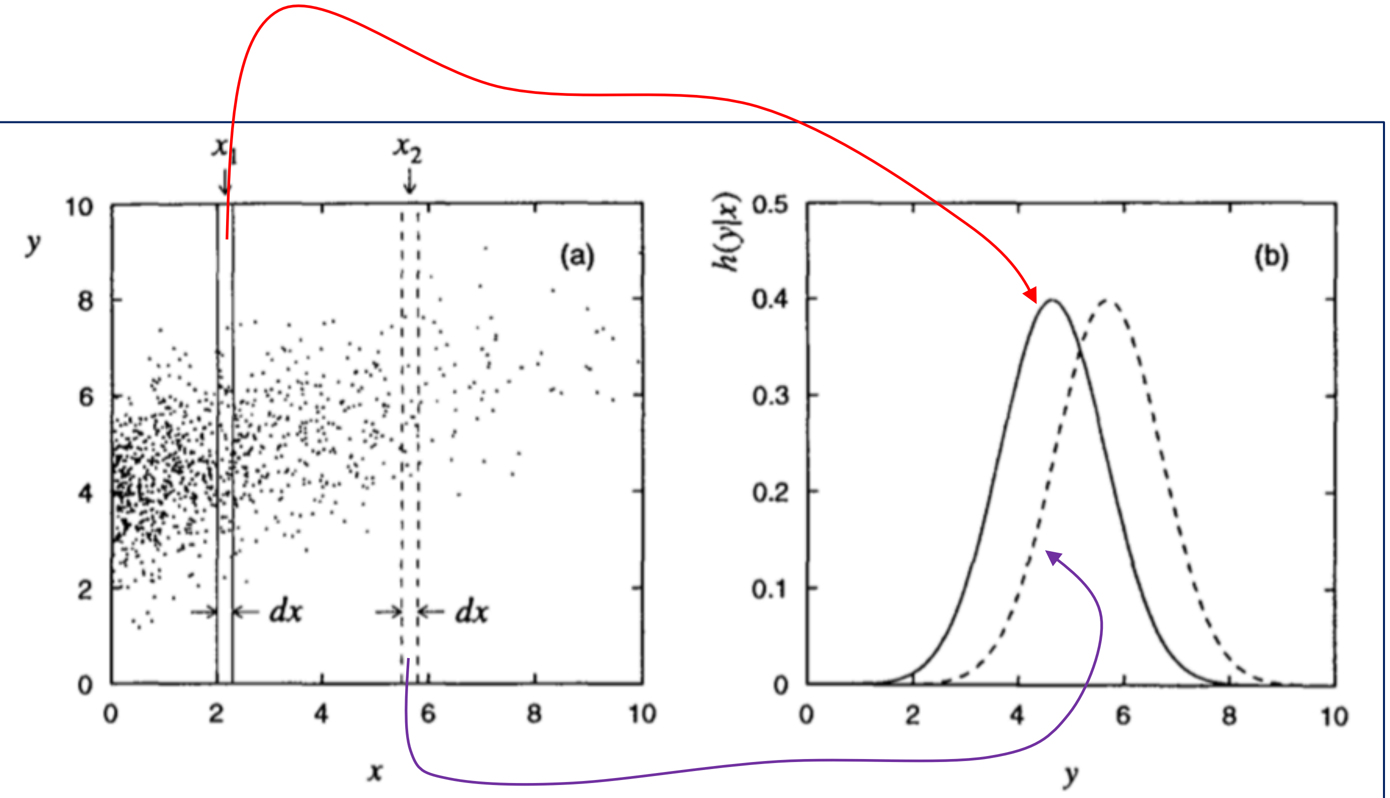


Fig. 1.6 (a) A scatter plot of random variables x and y indicating two infinitesimal bands in x of width dx at x_1 (solid band) and x_2 (dashed band). (b) The conditional p.d.f.s $h(y|x_1)$ and $h(y|x_2)$ corresponding to the projections of the bands onto the y axis.

[borrowed by Cowan]

The conditional p.d.f.s can be easily represented as *normalized* projections of narrow bands (large dx) in the conditioning variable

Two random variables & Conditional p.d.f. - III

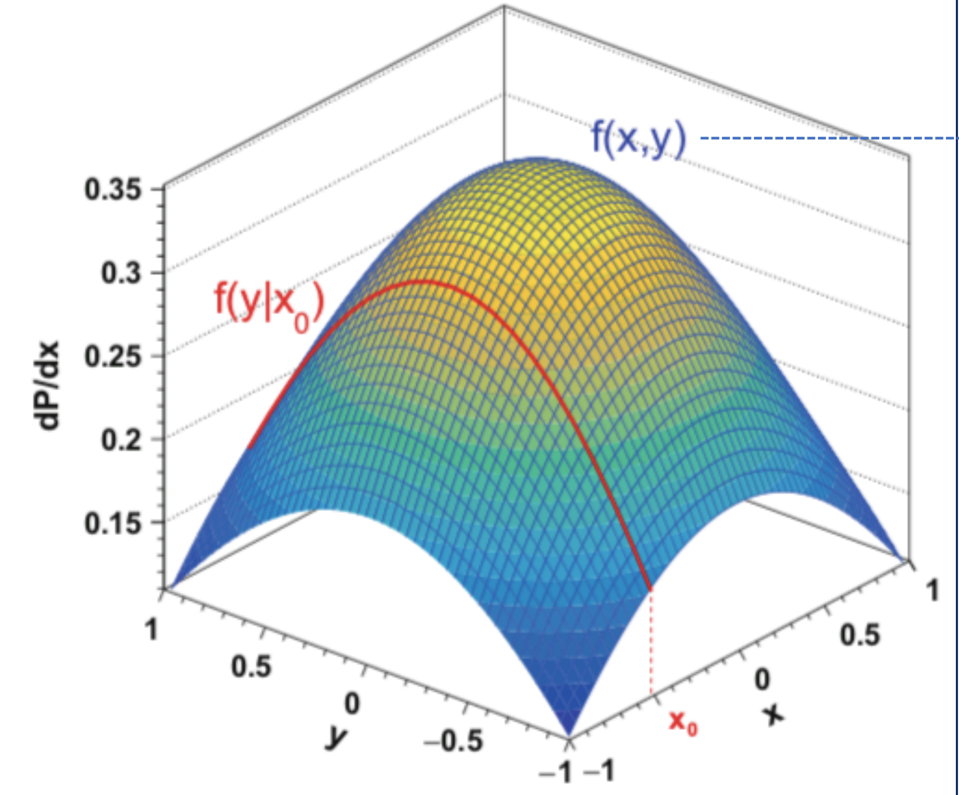


Fig. 2.19 Illustration of conditional PDF in two dimensions

joint p.d.f.

[borrowed by Lista]

Bayes theorem for random variables

- Combining together the two expressions for the conditional probability we get: $g(x|y) = g(y|x) \cdot \frac{f_x(x)}{f_y(y)}$

…which is nothing else that the **re-expression of the Bayes's theorem in the case of continuous r.v.s**!

- Rewriting the same two expressions we also get: $h(y|x) \cdot f_x(x) = f(x,y)$ $\quad g(x|y) \cdot f_y(y) = f(x,y)$

Now we can use the definition of marginal p.d.f.s to find new expressions for them:

$$f_x(x) = \int_{-\infty}^{+\infty} f(x,y)dy = \int_{-\infty}^{+\infty} g(x|y) \cdot f_y(y)dy \qquad f_y(y) = \int_{-\infty}^{+\infty} f(x,y)dy = \int_{-\infty}^{+\infty} h(y|x) \cdot f_x(x)dx$$

…which are nothing else that the **re-expression of the Law of total probability** (slide 14 part 1A)!

Independency of events expressed as factorization for joint p.d.f.

- We have discussed earlier that: $P(A) = f_x(x)dx$ (and, in the same way, $P(B) = f_y(y)dy$).
  Thus, the product of the two probabilities can be expressed as:

$$P(A) \cdot P(B) = f_x(x)\,dx \cdot f_y(y)\,dy \equiv f_x(x)\,f_y(y)\,dx\,dy \quad \text{(a)}$$

Let us remember now that … two events A and B are ***independent*** **if** $P(A \cap B) = P(A) \cdot P(B)$ [*]!

From the joint pd.f. definition $P(A \cap B) = f(x,y)dx\,dy$ we then derive from [*] that $P(A) \cdot P(B) = f(x,y)dx\,dy$ **(b)**

Expressions **(a)** & **(b)** hold <u>if and only if</u> $f(x,y)$ can be ***factorized*** into the product of the 2 marginal p.d.f.s:

$$f(x,y) = f_x(x) \cdot f_y(y)$$

From this result: x and y can be defined as independent variables if their joint p.d.f. can be written as the product of a p.d.f. of the variable x times a p.d.f. of the variable y (specifically these p.d.f.s are the 2 marginal ones)

- Additional expressions when r.v.s are independent: $h(y|x) = \dfrac{f(x,y)}{f_x(x)} = \dfrac{\cancel{f_x(x)} \cdot f_y(y)}{\cancel{f_x(x)}} \equiv f_y(y)$. Similarly: $g(x|y) = f_x(x)$

This means something obvious: the conditional pd.f. reduces simply to the marginal p.d.f. when the r.v.s. are independent.

CORRELATION between R.V.s

Let's consider 2 continuous r.v.s : (x, y) . The joint p.d.f. is written as $\boldsymbol{f(x,y)}$. We can write down the following quantities:

$$\mu_x \equiv E[x] = \int_{-\infty}^{+\infty} x f(x,y)\,dx\,dy \qquad \sigma_x^2 \equiv V[x] = E[(x-\mu_x)^2]$$

$$\mu_y \equiv E[y] = \int_{-\infty}^{+\infty} y f(x,y)\,dx\,dy \qquad \sigma_y^2 \equiv V[y] = E[(y-\mu_y)^2]$$

To take into account the possible correlations among the r.v.s, that generally are not negligible and cannot be overlooked, We need to introduce a further quantity called ***covariance***, defined as follows:

$$V_{xy} \equiv cov(x,y) = E[\,(x-\mu_x)(y-\mu_y)] = \iint_{-\infty}^{+\infty} xy \cdot f(x,y)\,dx\,dy$$

$$= E[xy - x\mu_y - y\mu_x + \mu_x\mu_y] =$$

$$= E[xy] - \mu_y E[x] - \mu_x\, E[y] + \mu_x\mu_y =$$

$$= E[xy] - \mu_y\mu_x - \cancel{\mu_x\mu_y} + \cancel{\mu_x\mu_y} =$$

$$= E[xy] - \mu_y\mu_x$$

?

Note: V_{xy} can be either positive or negative !

Note: as expected, V_{xy} gives simply the variance V_{xx} whether the r.v.s of the pair are identical (i.e. $y = x$)

Covariance for a couple of r.v.s - II

We have seen (slide 18) that $E[u(x,y)]$ can be expressed - in general - as: $E[u(x,y)] = \iint_{-\infty}^{+\infty} u(x,y)\cdot f(x,y)dxdy$

… and considering the specific case of $u(x,y) = x\cdot y$: $E[xy] = \iint_{-\infty}^{+\infty} xy\cdot f(x,y)dxdy$

Wrapping up: $V_{xy} \equiv cov(x,y) = E[\,(x-\mu_x)(y-\mu_y)] = E[xy] - \mu_x\mu_y = \iint_{-\infty}^{+\infty} xy\cdot f(x,y)dxdy - \mu_x\mu_y$ $(\equiv \sigma_{xy})$

Remember (see slide 19) that … in general $E[u] \neq u(\mu)$ and thus $E[xy] \neq \mu_x\mu_y$

In conclusion: $\boxed{V_{xy} = E[xy] - \mu_x\mu_y \neq 0}$ (i.e. one r.v. influences the other r.v. and viceversa) Note: $\boxed{V_{xy} = V_{yx}}$

- Since we can re-write: $\sigma_x^2 \equiv V[x] = E[\,(x-\mu_x)(x-\mu_x)] \equiv V_{xx}$, $\sigma_y^2 \equiv V[y] = E[\,(y-\mu_y)(y-\mu_y)] \equiv V_{yy}$

… it is possible to accommodate the 2 variances and the 2 (equal) covariances in a 2×2 **symmetric matrix**:

Covariance Matrix : $(V)_{xy} = \begin{pmatrix} V_{xx} & V_{xy} \\ V_{yx} & V_{yy} \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & cov(x,y) \\ cov(y,x) & \sigma_y^2 \end{pmatrix} = \begin{pmatrix} E[(x-\mu_x)^2] & E[xy]-\mu_x\mu_y \\ E[xy]-\mu_x\mu_y & E[(y-\mu_y)^2] \end{pmatrix}$

$\equiv \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 \end{pmatrix}$

Correlation Coefficient

- With the aim to have an adimentional measure of the "degree of correlation" between the two r.v.s and …

… it is useful to introduce the **correlation coefficient** :

$$\rho(x,y) = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y} \equiv \frac{V_{xy}}{\sqrt{V_{xx} \cdot V_{yy}}}$$

It can be demonstrated that: $\rho(x,y) \in [-1,+1]$. We get:

- maximum correlation : $\rho(x,y) = +1$
- NO correlation : $\rho(x,y) = 0$
- maximum **anti**-correlation : $\rho(x,y) = -1$

It is easy to discuss the correlation coefficient by means of these scatter plots of the r.v.s x and y :

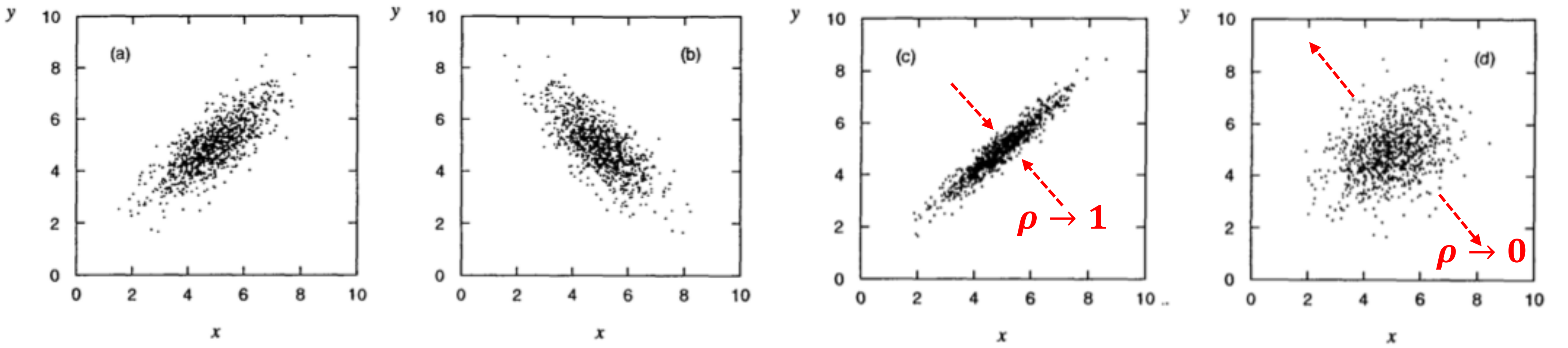


Fig. 1.9 Scatter plots of random variables x and y with (a) a positive correlation, $\rho = 0.75$, (b) a negative correlation, $\rho = -0.75$, (c) $\rho = 0.95$, and (d) $\rho = 0.25$. For all four cases the standard deviations of x and y are $\sigma_x = \sigma_y = 1$.

[borrowed by Cowan]

Independence & uncorrelation - I

- In the every-day language - often - the physicists talk about ***uncorrelated*** variables implicitely implying ***independent*** ones, although this is not correct. We will argue - instead - that strictly speaking …
  the condition of uncorrelation is *weaker* than the condition of independency !

  Indeed we will show that … **independency implies uncorrelation but the viceversa is not true**!

-

Independence & uncorrelation - I

- In the every-day language - often - the physicists talk about ***uncorrelated*** variables implicitely implying ***independent*** ones, although this is not correct. We will argue - instead - that strictly speaking …
  the condition of uncorrelation is *weaker* than the condition of independency !

  Indeed we will show that … **independency implies uncorrelation but the viceversa is not true**!

- To argue this, let me start recalling (see slide 29) that …

  … if (x, y) are (mutually) independent random variables their joint p.d.f. factorizes: $\boldsymbol{f(x,y) = f_x(x) \cdot f_y(y)}$

  and in this case: $$\boldsymbol{E[xy]} = \iint_{-\infty}^{+\infty} \boldsymbol{xy \cdot f(x,y) dxdy} = \int_{-\infty}^{+\infty} \boldsymbol{x \cdot f_x(x) dx} \cdot \int_{-\infty}^{+\infty} \boldsymbol{y \cdot f_y(y) dy} = \boldsymbol{E[x] \cdot E[y]}$$

  which implies that : $$\boldsymbol{V_{xy} = E[xy] - \mu_x\mu_y = E[x] \cdot E[y] - \mu_x\mu_y = \mu_x\mu_y - \mu_x\mu_y = 0} \; (\text{thus } \boldsymbol{\rho_{xy} = 0})$$

$\Rightarrow$) We have proved that : **INDEPENDENCY $\Rightarrow$ UNCORRELATION**

- In the every-day language - often - the physicists talk about ***uncorrelated*** variables implicitely implying ***independent*** ones, although this is not correct. We will argue - instead - that strictly speaking …
the condition of uncorrelation is *weaker* than the condition of independency !

Indeed we will show that … **independency implies uncorrelation but the viceversa is not true**!

- To argue this, let me start recalling (see slide 29) that …

… if (x, y) are (mutually) independent random variables their joint p.d.f. factorizes: $\boldsymbol{f(x,y) = f_x(x) \cdot f_y(y)}$

and in this case: $$\boldsymbol{E[xy]} = \iint_{-\infty}^{+\infty} \boldsymbol{xy \cdot f(x,y)dxdy} = \int_{-\infty}^{+\infty} \boldsymbol{x \cdot f_x(x)dx} \cdot \int_{-\infty}^{+\infty} \boldsymbol{y \cdot f_y(y)dy} = \boldsymbol{E[x] \cdot E[y]}$$

which implies that : $$\boldsymbol{V_{xy}} = \boldsymbol{E[xy] - \mu_x\mu_y = E[x] \cdot E[y] - \mu_x\mu_y = \mu_x\mu_y - \mu_x\mu_y = 0} \quad (\text{thus } \boldsymbol{\rho_{xy} = 0})$$

$\Rightarrow$) We have proved that : **INDEPENDENCY $\Rightarrow$ UNCORRELATION**

$\Leftarrow$) To prove - instead - that the viceversa does not hold, i.e. **INDEPENDENCY $\not\Leftarrow$ UNCORRELATION**

… we need to find at least one example characterized by dependency $(\boldsymbol{y = f(x)})$ in spite of existing uncorrelation $(\boldsymbol{V_{xy} = 0})$ (next slide)

- A suitably easy example is $(x, y) = (x, x^2)$ namely when $y = u(x) = x^2$!

To make easier the demonstration let's suppose that …
x is distributed symmetrically around 0, with a p.d.f. $f(x)$, i.e.: $\mu_x \equiv E[x] = \int_{-\infty}^{+\infty} x \cdot f_x(x)dx = 0$

From the definition of variance: $\sigma_x^2 \equiv \mathrm{V}[x] = \int_{-\infty}^{+\infty} (x - 0)^2 \cdot f_x(x)dx \equiv \int_{-\infty}^{+\infty} x^2 \cdot f_x(x)dx$

Independence & uncorrelation - II

- A suitably easy example is $(x, y) = (x, x^2)$ namely when $y = u(x) = x^2$!

To make easier the demonstration let's suppose that …
x is distributed symmetrically around 0, with a p.d.f. $f(x)$, i.e.: $\mu_x \equiv E[x] = \int_{-\infty}^{+\infty} x \cdot f_x(x)dx = 0$

From the definition of variance: $\sigma_x^2 \equiv \mathrm{V}[x] = \int_{-\infty}^{+\infty} (x - 0)^2 \cdot f_x(x)dx \equiv \int_{-\infty}^{+\infty} x^2 \cdot f_x(x)dx$

Let's calculate the expectation value of the r.v. y : $\mu_y \equiv E[y] = E[u(x)] = \int_{-\infty}^{+\infty} u(x) \cdot f(x)dx = \int_{-\infty}^{+\infty} x^2 \cdot f_x(x)dx = \sigma_x^2$

Note (for completeness) that: $f(x)$ is the marginal for x i.e. $f_x(x)$; analogously $g(u) = g(y) = g_y(y)$ would be the marginal for y .

Finally let's calculate the covariance: $V_{xy} = E\left[(x - \cancel{\mu_x}^{\,0})(y - \mu_y)\right] = E\left[(x)(x^2 - \sigma_x^2)\right] = E\left[x^3 - x\sigma_x^2\right] =$

$$= E\left[x^3\right] - \sigma_x^2 \cancel{E[x]}^{\,0} = E\left[x^3\right] = 0$$

(central moment of order-3 is null for a symmetric $f(x)$!)

- Visualizing the previous example:

$\rho_{xy} \approx 0$

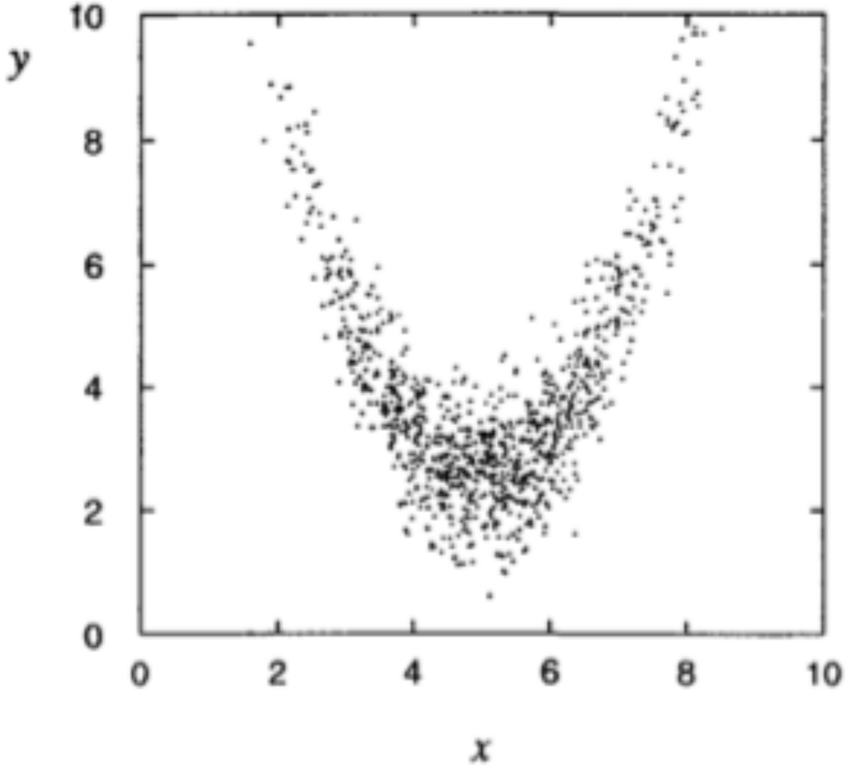


Fig. 1.10 Scatter plot of random variables x and y which are not independent (i.e. $f(x,y) \neq f_x(x) f_y(y)$) but for which $V_{xy} = 0$ because of the particular symmetry of the distribution.

borrowed by G.Cowan

Note: see in-depth slides for another example.

It is possible to remove (or introduce) a correlation by operating a change of variables, namely $(x, y) \rightarrow (x', y')$

Note that - in our 2D framework - **this change of variable corresponds to a rotation in the (x, y) plane**!

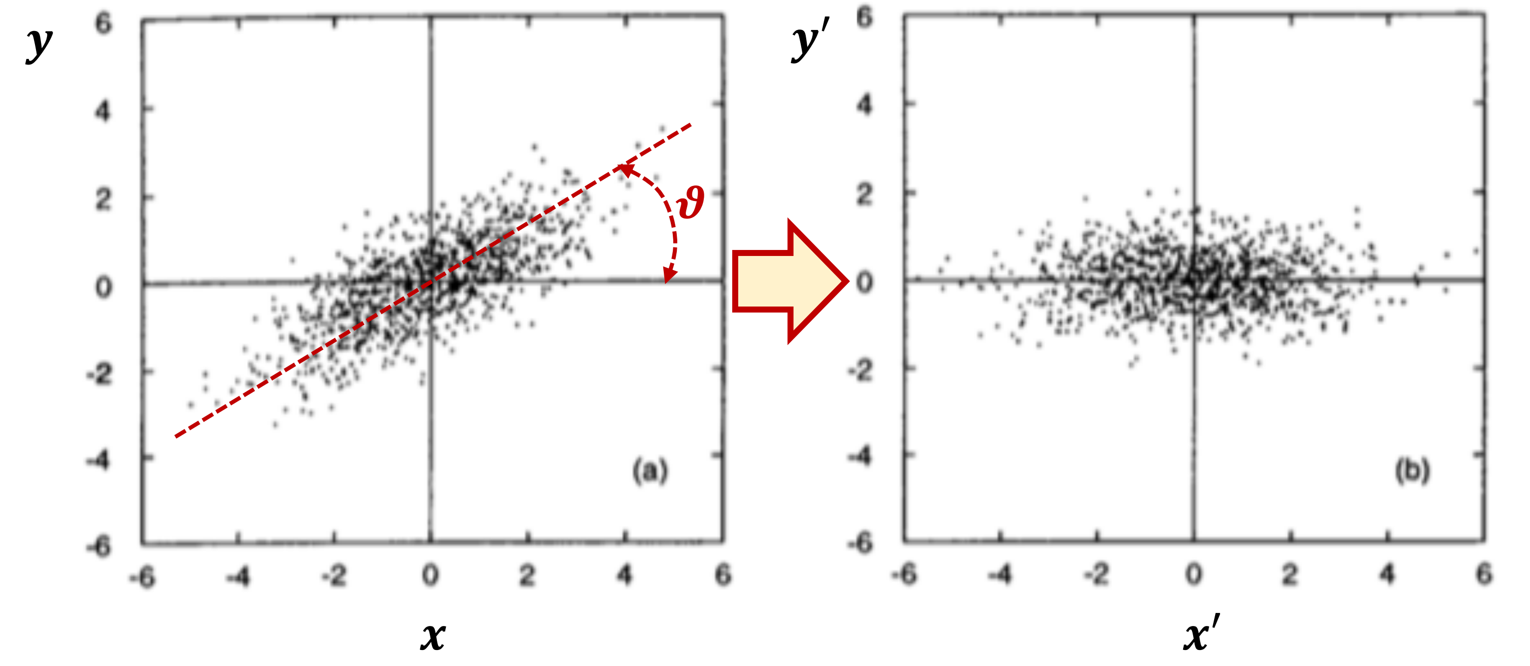


Fig. 1.11 Scatter plot of (a) two correlated random variables (x, y) and (b) the transformed variables (x', y') for which the covariance matrix is diagonal.

borrowed by G.Cowan

The rotation-in-the-plane matrix:

$$A = \begin{pmatrix} \cos\vartheta & \sin\vartheta \\ -\sin\vartheta & \cos\vartheta \end{pmatrix}$$

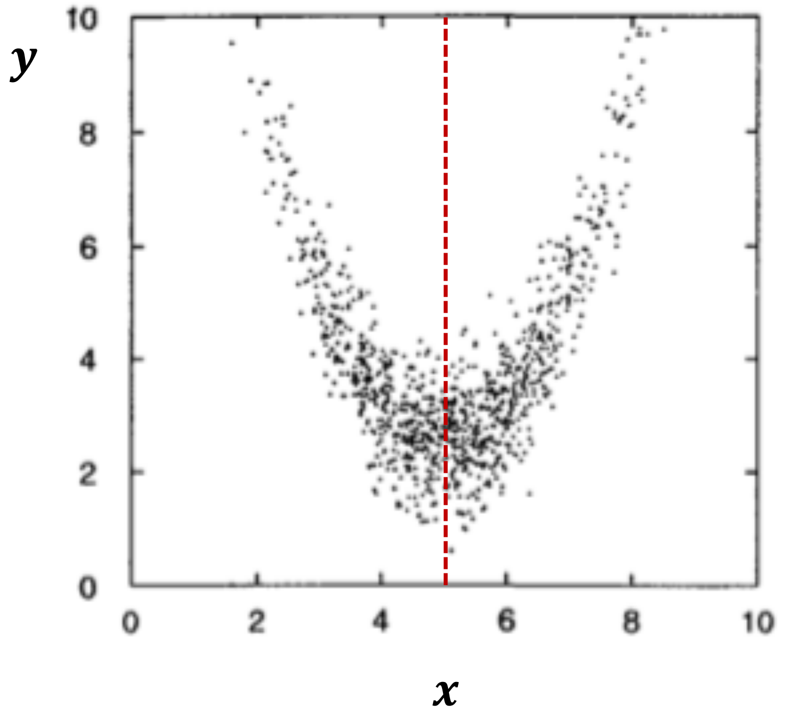
It can be calculated (G.Cowan, 1.7) that the angle is:

$$\tan(2\vartheta) = \left(\frac{2V_{xy}}{\sigma_y^2 - \sigma_x^2}\right) \equiv \left(\frac{2\rho_{xy}\sigma_x\sigma_y}{\sigma_y^2 - \sigma_x^2}\right)$$

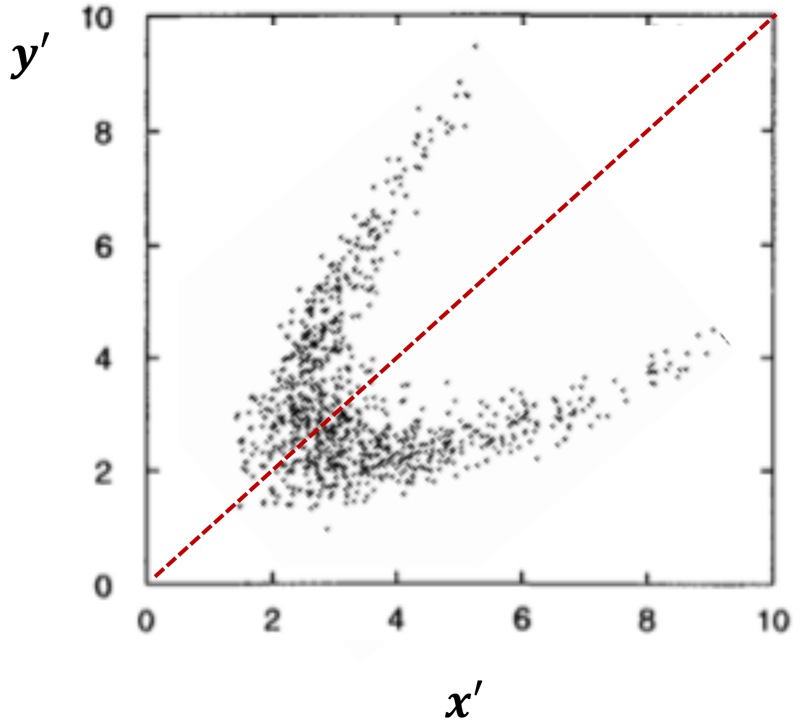
Note that **the matrix A is such that the matrix $U = A \cdot V \cdot A^T$ is diagonal** ! (I will comment further … a few slides later)

row by column products

- An example of possible introduction of some correlation between two variables is a rotation in their plane as well:



$\rho_{xy} \approx 0$



$\rho_{xy} \neq 0$

Covariance for more than 2 r.v.s

- Let's consider N r.v.s: $(x_1, \dots, x_i, \dots, x_j, \dots\ x_N)$

  The variance of the single r.v. - regardless the others - is simply defined as: $\sigma_i^2 = E[(x_i - E[x_i])^2]$

  To take into account the mutual correlations we have to introduce a coviariance for each pair (i, j):

  $$V_{ij} \equiv \sigma_{ij} = E[\,(x_i - E[x_i])(x_j - E[x_j])\,]$$

  The N variances and the $N(N-1)$ covariances (each two of them are equal by symmetry, i.e. $\sigma_{ij} = \sigma_{ji}$) can be accomodated in the **covariance matrix**, an $N \times N$ symmetric, sometimes called **error matrix**:

  $$(V)_{ij} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1N} \\ \sigma_{21} & \sigma_2^2 & & \\ \vdots & & \ddots & \vdots \\ & & \cdots & \\ \sigma_{N1} & & & \sigma_N^2 \end{pmatrix}$$

  for each pair, $i = j$ gives back the variance

  Note: If the covariance matrix is not positive defined…
  … there must be at least one linear relationship among the r.v.s.

- A ***global*** **correlation coefficient** can be introduced when $N > 2$ (see in-depth slides)

Diagonalization of the covariance matrix - I

It can be demonstrated that ….

… **it is always possible, in the framework of linear algebra, to find an *orthogonal transformation* of $N \geq 2$ variables $(x_1, \dots, x_N) \rightarrow (y_1, \dots, y_N)$ for which the "new" covariance matrix for $\vec{y}$ is diagonal while the "old" one for $\vec{x}$ was not !**

It's common to say that this transformation "diagonalizes the covariance matrix",
i. e. this transformation is able to remove any existing correlation.

Let's discuss this result:

- original variables & covariance matrix: $(x_1, \dots, x_N)$, $V_{ij} = cov(x_i, x_j)$

- transformed variables & new diagonal covariance matrix: $(y_1, \dots, y_N)$, $U_{ij} = cov(y_i, y_j)$

It can be demonstrated that it is always possible to find a **linear** transformation, namely by means of a matrix so that each y_i is a linear combination of the $(x_1, \dots, x_N)$:

$$y_i = \sum_{j=1}^{N} A_{ij} x_j \quad (\forall i) \qquad (\#)$$

In this case the transformation matrix A is such that the new matrix $U = AVA^T$ is diagonal, and has the property that the transpose matrix coincides with the inverse ($A^T = A^{-1}$) and thus $U = AVA^{-1}$.
This transformation is called *orthogonal* and it corresponds - in linear algebra - to the rotation of the vector $\vec{x}$ into the vector $\vec{y}$ so that the vector norm is kept constant. (see also next side)

We can formalize what just said using the vectorial notation and the matrix formalism:

$$\vec{x} = \begin{pmatrix} x_1 \\ \cdots \\ x_N \end{pmatrix}, \quad \vec{x}^T = (x_1 \ldots x_N), \quad \vec{y} = A\vec{x} \Leftrightarrow \begin{pmatrix} y_1 \\ \cdots \\ y_N \end{pmatrix} = \begin{pmatrix} A_{11} & \cdots & A_{1N} \\ \cdots & \cdots & \cdots \\ A_{N1} & \cdots & A_{NN} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ \cdots \\ x_N \end{pmatrix}$$

$$|\vec{y}|^2 = \vec{y}^T \cdot \vec{y} = \vec{x}^T A^T \cdot A\,\vec{x} = \vec{x}^T A^{-1} \cdot A\,\vec{x} = \vec{x}^T \,\mathrm{I}\, \vec{x} = \vec{x}^T \vec{x} = |\vec{x}|^2$$ **: vector norm is preserved**

$$U_{ij} = cov(y_i, y_j) \overset{(\#)}{=} cov\left(\sum_{k=1}^{N} A_{ik} x_k \,, \sum_{\ell=1}^{N} A_{j\ell}\, x_\ell\right) =$$

$$\begin{cases} \boldsymbol{cov(u,v) = E[uv] - \mu_u \mu_v} \,, \\ \boldsymbol{E[u(x)] = u(\mu)} \ \textbf{IF} \ \boldsymbol{u(x)} \text{ is linear in } \boldsymbol{x} \\ E[a_1 u(x) + a_2 v(x)] = a_1 E[u(x)] + a_2 E[v(x)] \end{cases}$$

$$= \sum_{k=1}^{N} \sum_{\ell=1}^{N} A_{ik}\, A_{j\ell}\, \boxed{cov(x_k \,, x_\ell)}^{V_{k\ell}} =$$

$A_{j\ell} = (A^T)_{\ell j}$ (from the def. of transpose matrix)

$$= \sum_{k=1}^{N} \sum_{\ell=1}^{N} A_{ik}\, V_{k\ell}\, A^T{}_{\ell j}$$

(here we must “saturate” on indices k and ℓ)

ERROR PROPAGATION

Propagation of the variances - I

Suppose we have N r.v.s $(x_1, \dots, x_n)$, that we can write - in a compact way - as the vector $\vec{x} \equiv (x_1, \dots, x_n)$, distributed according to the joint p.d.f. $f(\vec{x})$ that we suppose is **not** fully known since we assume we only know:

- the N expectation values, namely the vector $\vec{\mu} \equiv (\mu_1, \dots, \mu_n)$
- the $N \times N$ covariance matrix V_{ij}

Let's now consider a function $y = u(\vec{x})$ and we have seen (slides 17-18) that …
… we can determine the p.d.f. of y - say $g(u)$ - if we know the p.d.f $f(\vec{x})$ which, however, is not our case here!
Thus, we want to determine just $E[y]$ and V$[y]$.
We will see that this is possible, even if we will get **approximated** (but still **useful**) expressions!

The procedure starts from the expansion in series - truncated at 1st order - of the function $y(\vec{x})$ around the vector of the expectation values $\vec{\mu}$:

$$\boldsymbol{y(\vec{x}) \cong y(\vec{\mu}) + \sum_{i=1}^{N} \left(\frac{\partial y}{\partial xi}\right)_{\vec{x}=\vec{\mu}} \cdot (x_i - \mu_i) + \cancel{\dots}}$$

The expectation value can be easily calculated at first order:

I can apply one of the properties of the expectation value of a variable; consider that the derivatives are calculated for $\vec{x} = \vec{\mu}$ so they are just real numbers

$$\boldsymbol{E[y(\vec{x})]} \cong E[y(\vec{\mu})] + E\left[\sum_{i=1}^{N} \dots\right] =$$

$$= E[y(\vec{\mu})] + \sum_{i=1}^{N} \left(\frac{\partial y}{\partial xi}\right)_{\vec{x}=\vec{\mu}} \cdot \underbrace{\boldsymbol{E(x_i - \mu_i)}}_{0} = \boldsymbol{E[y(\vec{\mu})]}$$

(as expected: at 1st order the dependency is linear)

- Let's calculate the variance: $\sigma_y^2 = E[y^2] - (E[y])^2$ ⟶ just calculated

to be calculated here:

$$E[y^2(\vec{x})] \simeq E\left[\left(y(\vec{\mu}) + \sum_i^n \left(\frac{\partial y}{\partial x_i}\right)_{\vec{x}=\vec{\mu}}(x_i-\mu_i)\right)^2\right] =$$

$$= E\left[y^2(\vec{\mu}) + 2y(\vec{\mu})\cdot\sum_i^n \left(\frac{\partial y}{\partial x_i}\right)_{\vec{x}=\vec{\mu}}(x_i-\mu_i) + \right.$$

$$\left. + \left(\sum_i^n \left(\frac{\partial y}{\partial x_i}\right)_{\vec{x}=\vec{\mu}}(x_i-\mu_i)\right)\cdot\left(\sum_j^n \left(\frac{\partial y}{\partial x_j}\right)_{\vec{x}=\vec{\mu}}(x_j-\mu_j)\right)\right] =$$

PROPRIETÀ del VALORE di ASPETTAZIONE

$$= E[y^2(\vec{\mu})] + \cancel{2y(\vec{\mu})\sum_i^n \left(\frac{\partial y}{\partial x_i}\right)_{\vec{x}=\vec{\mu}} \underbrace{E[x_i-\mu_i]}_{=0}} +$$

$$+ E\left[\left(\quad '' \quad\right)\cdot\left(\quad '' \quad\right)\right] =$$

$$= y^2(\vec{\mu}) + \sum_i^n \sum_j^n \left(\frac{\partial y}{\partial x_i}\cdot\frac{\partial y}{\partial x_j}\right)_{\vec{x}=\vec{\mu}} \underbrace{E[(x_i-\mu_i)(x_j-\mu_j)]}_{V_{ij}}$$

Therefore:

$$\sigma_y^2 \simeq E[y^2(\vec{x})] - (E[y])^2 = \bigcirc - (y(\vec{\mu}))^2 =$$

$$= \cancel{y^2(\vec{\mu})} + \sum_i^n \sum_j^n \left(\frac{\partial y}{\partial x_i}\cdot\frac{\partial y}{\partial x_j}\right)_{\vec{x}=\vec{\mu}} V_{ij} - \cancel{y^2(\vec{\mu})}$$

Thus, we got: $\sigma_y^2 = E[y^2] - (E[y])^2 \cong \sum_{i,j=1}^{N} \left(\frac{\partial y}{\partial xi}\frac{\partial y}{\partial xj}\right)_{\vec{x}=\vec{\mu}} V_{ij}$: **equation of the error propagation**

If we conventionally define the **vector of partial derivatives** $A = \left(\frac{\partial y}{\partial x_1}, \dots, \frac{\partial y}{\partial xN}\right)$

… we can re-express this result in matrix notation:

$$\sigma_y^2 = \underbrace{\left(\frac{\partial y}{\partial x_1}, \dots, \frac{\partial y}{\partial xN}\right)}_{(1\times N)} \cdot \underbrace{\underbrace{\begin{pmatrix} V_{11} & \dots & V_{1N} \\ \dots & \dots & \dots \\ V_{N1} & \dots & V_{NN} \end{pmatrix}}_{(N\times N)} \cdot \underbrace{\overbrace{\boxed{\begin{pmatrix} \frac{\partial y}{\partial x_1} \\ \dots \\ \frac{\partial y}{\partial xN} \end{pmatrix}}}^{A^T}}_{(N\times 1)}}_{(N\times 1)}$$

… and more compactly: $\sigma_y^2 = A\,V A^T$

Do not forget that … this result is valid in the approximation in which $y(\vec{x})$ is approximated by the Taylor expansion truncated to the 1st order, namely in the linearity approximation around $\vec{\mu}$!

- In the particular case in which the $(x_1, \dots, x_n)$ are all ***uncorrelated*** among each other, i.e. $\begin{cases} V_{ii} = \sigma_i^2 \ (\forall i) \\ V_{ij} = 0 \ \ (\forall i \neq j) \end{cases}$

… then the propagation formula reduces to: $\sigma_y^2 \cong \sum_{i=1}^{N} \left(\frac{\partial y}{\partial xi}\right)^2_{\vec{x}=\vec{\mu}} \sigma_i^2$ (the well-known “error propagation formula”)

Propagation of the variances : special cases

- Usual cases are these 4:

RELAZIONE FRA x_1 e x_2	x_1, x_2 correlated $(V_{12} \neq 0)$	x_1, x_2 uncorrelated $(V_{12} = 0)$
$y = x_1 + x_2$	$\sigma_y^2 = \sigma_1^2 + \sigma_2^2 + 2V_{12}$	$\sigma_y^2 = \sigma_1^2 + \sigma_2^2$
$y = x_1 - x_2$	$\sigma_y^2 = \sigma_1^2 + \sigma_2^2 - 2V_{12}$	$\sigma_y^2 = \sigma_1^2 + \sigma_2^2$
$y = x_1 \cdot x_2$	$\frac{\sigma_y^2}{(\mu_1 \mu_2)^2} \simeq \frac{\sigma_1^2}{\mu_1^2} + \frac{\sigma_2^2}{\mu_2^2} + 2\frac{V_{12}}{\mu_1 \mu_2}$	$\frac{\sigma_y^2}{(\mu_1 \mu_2)^2} \simeq \frac{\sigma_1^2}{\mu_1^2} + \frac{\sigma_2^2}{\mu_2^2}$
$y = \frac{x_1}{x_2}$	$\frac{\sigma_y^2}{(\mu_1/\mu_2)^2} \simeq \frac{\sigma_1^2}{\mu_1^2} + \frac{\sigma_2^2}{\mu_2^2} - 2\frac{V_{12}}{\mu_1 \mu_2}$	$\frac{\sigma_y^2}{(\mu_1/\mu_2)^2} \simeq \frac{\sigma_1^2}{\mu_1^2} + \frac{\sigma_2^2}{\mu_2^2}$

dove $\mu_1 = E[x_1]$, $\mu_2 = E[x_2]$

Very usuful because we deal very often with ratios !

The *relative* standard deviations sum up in quadrature

$$\sigma^2_{y=\frac{x_1}{x_2}} \cong \frac{\sigma_1^2}{\mu_2^2} + \frac{\sigma_2^2}{\mu_2^2} \cdot \left(\frac{\mu_1^2}{\mu_2^2}\right)$$

PART 2B - IN-DEPTH SLIDES

alexis.pompili@ba.infn.it

Attribute of a p.d.f. : skewness & kurtosis

- Since all **symmetric** p.d.f.s have null odd central moments, the central moments of odd order (3, 5, …) provide a measurement of the asymmetry of a generic distribution (remember the one of 1st order is null).

  In order to have an adimentional quantity we prefer divide by $\boldsymbol{\sigma_x^3} = (\mathbf{V}[\boldsymbol{x}])^{3/2}$:

  skewness **of a p.d.f.** is defined as:

$$\boldsymbol{\gamma_1 = \frac{E[(x-\mu)^3]}{\sigma_x^3}}$$

- For a p.d.f. characterized by a central symmetric peak, its peaking “level” (or “degree”), let’s call it “**sharpness**”, can be measured through :

  kurtosis **of a p.d.f.** is defined as:

$$\boldsymbol{\gamma_2 = \frac{E[(x-\mu)^4]}{\sigma_x^4} - 3}$$

This ad hoc definition derives from the aim to have $\boldsymbol{\gamma_2 = 0}$ **for a Gaussian p.d.f.**, thus **this “sharpness” is compared to that of the Gaussian used as the reference**.

- Demonstrate the expression for $\mathrm{V}[\boldsymbol{x}]$ of a mixture of sub-samples:

$$V[x] = \sum_i \Phi_i \, E_i\left[(x_i - \mu_i - \delta_i)^2\right] =$$

$$= \sum_i \Phi_i \, E_i\left[(x-\mu_i)^2 + \delta_i^2 - 2\delta_i (x-\mu_i)\right] =$$

APPLICO LA LINEARITÀ

$$E\left[a_1 u_1(x) + a_2 u_2(x)\right] = a_1 E\left[u_1(x)\right] + a_2 E\left[u_2(x)\right]$$

$$= \sum_i \Phi_i \left\{ E_i\left[(x-\mu_i)^2\right] + E_i\left[\delta_i^2\right] - E_i\left[2\delta_i (x-\mu_i)\right] \right\} =$$

$$(E[a] = a \text{ con } a = \text{cost})$$

$$= \sum_i \Phi_i \left\{ V_i[x] + \delta_i^2 - 2\delta_i \underbrace{E_i\left[(x-\mu_i)\right]}_{=0} \right\}$$

$$= \sum_i \Phi_i \left\{ V_i[x] + \delta_i^2 \right\}$$

Now I rewrite in a useful way the deviations $\boldsymbol{\delta_i}$:

$$\delta_i = \mu - \mu_i = \sum_j \Phi_j \mu_j - \mu_i =$$

$$= \sum_{j \neq i} \Phi_j \mu_j + \Phi_i \mu_i - \mu_i =$$

$$= \sum_{j \neq i} \Phi_j \mu_j - (1 - \Phi_i)\mu_i = \quad \left(\sum_j \Phi_j = 1\right)$$

$$= \sum_{j \neq i} \Phi_j \mu_j - \sum_j \Phi_j \mu_i + \Phi_i \mu_i =$$

$$= \sum_{j \neq i} \Phi_j \mu_j - \sum_{j \neq i} \Phi_j \mu_i - \cancel{\Phi_i \mu_i} + \cancel{\Phi_i \mu_i} =$$

$$= \sum_{j \neq i} \Phi_j (\mu_j - \mu_i)$$

putting together I get the (overall) variance :

$$\boldsymbol{V[x] = \sum_i \varphi_i \cdot \left\{V_i[x] + \left[\sum_{j \neq i} \varphi_i(\mu_j - \mu_i)\right]^2\right\}}$$

Additional example of dependence with uncorrelation

Example 2.7 Uncorrelated Variables May not Be Independent
An example of PDF that describes uncorrelated variables that are not independent is given by the sum of four two-dimensional Gaussian PDFs as specified below:

$$f(x,\, y) = \frac{1}{4}\,[g(x;\, \mu,\, \sigma)\, g(y;\, 0,\, \sigma) + g(x;\, -\mu,\, \sigma)\, g(y;\, 0,\, \sigma)$$

$$g(x;\, 0,\, \sigma)\, g(y;\, \mu,\, \sigma) + g(x;\, 0,\, \sigma)\, g(y;\, -\mu,\, \sigma)]\ , \tag{2.83}$$

where g is a one-dimensional Gaussian distribution.

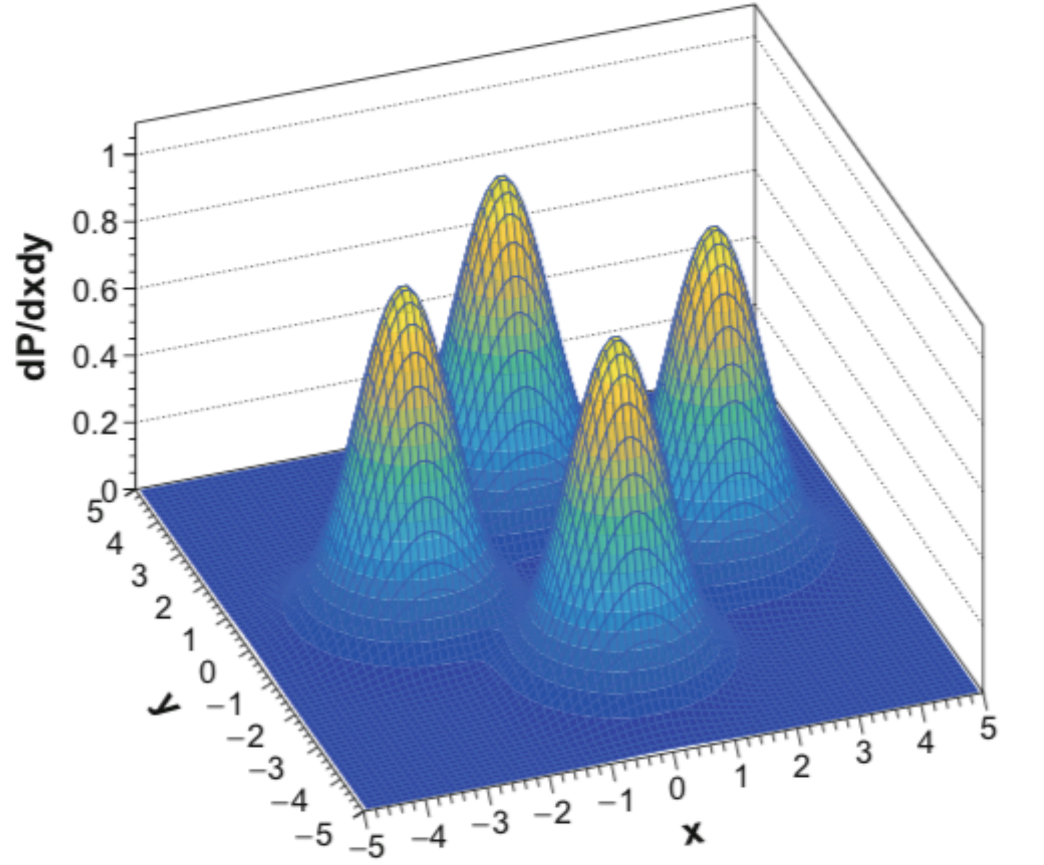


Fig. 2.18 Example of a PDF of two variables x and y that are uncorrelated but not independent

borrowed by L.Lista

Correlation coefficient for more than 2 r.v.s

- For each pair (i, j) a correlation coefficient can be defined in the standard way: $\rho(x_i, x_j) = \frac{V_{ij}}{\sigma_i \cdot \sigma_j}$

Nevertheless it can be introduced a more useful indicator, the **global correlation coefficient** :

- take a generic the r.v. x_k
- consider the correlations $\rho(x_k, y)$
- consider the linear combination y of all the other $N-1$ r.v.s $x_{i \neq k}$
- define the **global corr. coeff.** $\rho_k = max\{\rho(x_k, y)\}$

  as the quantity that measures the total amount of correlation among x_k and all the others $x_{i \neq k}$

Thus : $\rho_k = 0 \iff x_k$ is fully uncorrelated with all the others $x_{i \neq k}$

$\rho_k = 1 \iff x_k$ is fully correlated with at least one linear combination of the others $x_{i \neq k}$

- An useful result (given without demonstration) is the following: $\rho_k = \sqrt{1 - [V_{kk} \cdot (V^{-1})_{kk}]^{-1}}$

… where … $\begin{cases} (V)_{kk} : \text{diagonal element of the covariance matrix} \\ (V^{-1})_{kk} : \text{diagonal element of the inverse of the covariance matrix} \end{cases}$