Statistical Methods and Techniques for Data Analysis in High Energy Physics

MINI-COURSE @ Department of Physics - U. of Ioannina

(16 Hours = 8 theory +8 hands-on)

[24-28.04.2023]

https://web2.ba.infn.it/~pompili/teaching/data_analysis_lab/ErasmusPlus/Ioannina/mini-course-2023.html

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

Erasmus+ Teaching Mobility

Funding agency : TUCEP (thanks to EU funds)

* alexis.pompili@ba.infn.it

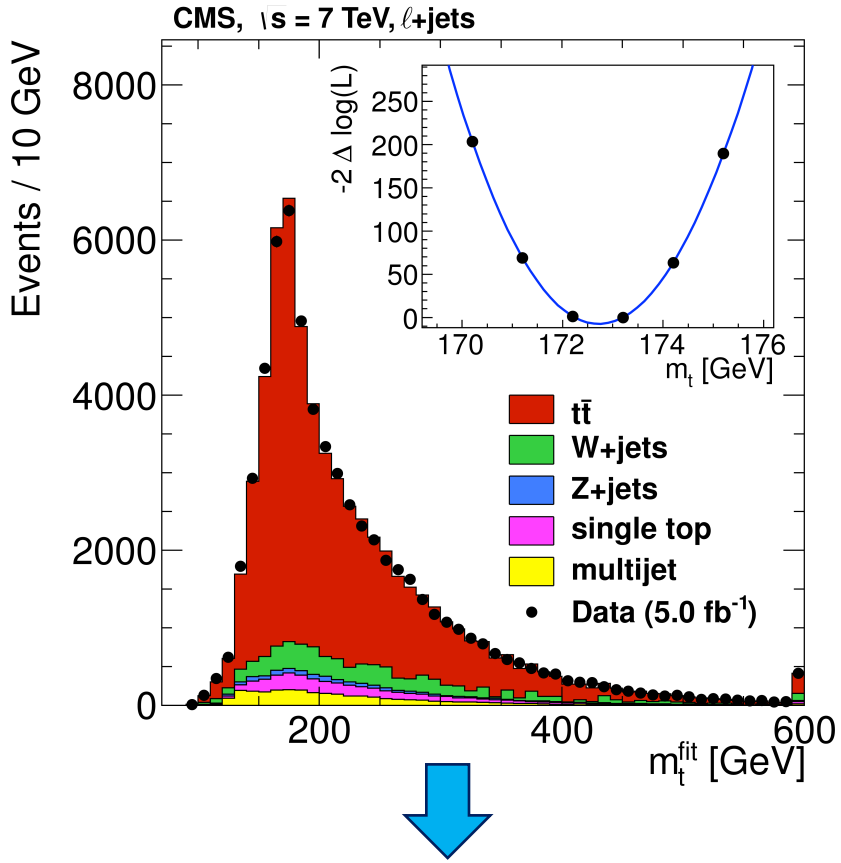
https://persone.ict.uniba.it/rubrica/alexis.pompili

Theory / Hour-1

INTRODUCTION

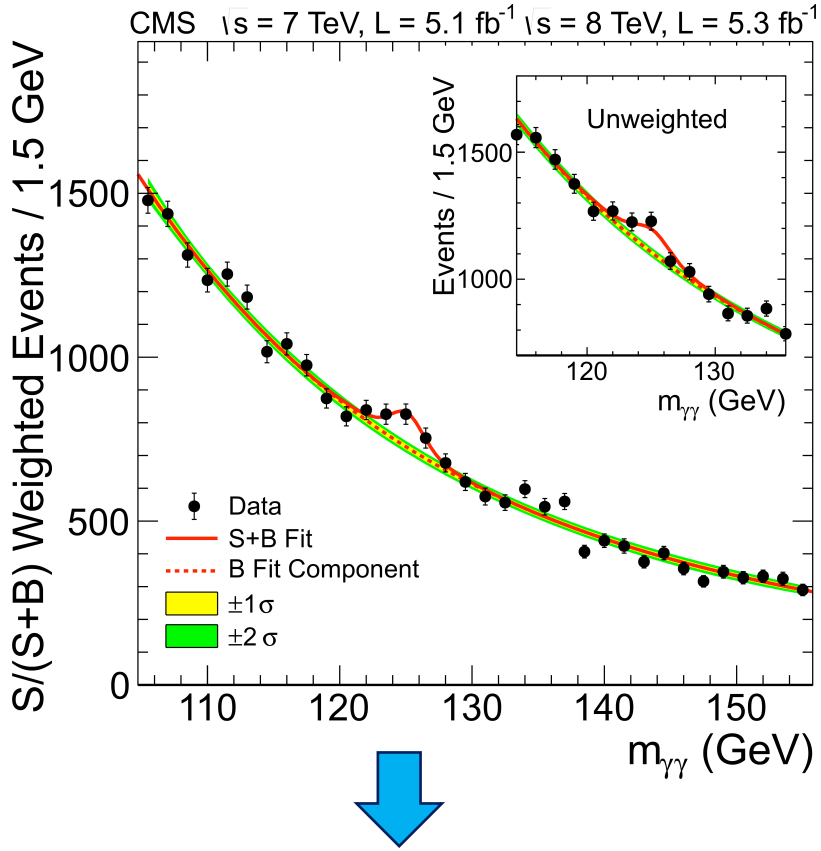
The goals of a (experimental) Particle Physicist - I

MEASUREMENTS



$m_t = 173.49 \pm 1.07$

DISCOVERIES



HIGGS BOSON

The goals of a (experimental) Particle Physicist - II

In modern particle physics experiments, **event data** are recorded by a - usually complex - system of detectors.

Measurements of particle position, particle momentum/energy, time, decay angles etc… are recorded in the **event data** and are characterized by fluctuations (due to randomness & dilution effects).

Event data are all ***different*** from each other because of:

- **Intrinsic randomness** of the physics process(es) (Quantum Mechanics: $\mathcal{P} \propto |\mathcal{A}|^2$)
- **Detector response** is somewhat random (fluctuations, resolutions, efficiencies, ….)

Tipically, a large number of events are collected by an experiment, each event usually containing large amounts of data → what we study are **distributions** of physical observables (e.g. the mass of a particle, the lifetime, etc.)

The goals of a (experimental) Particle Physicist - III

Distributions of measured quantities in data:

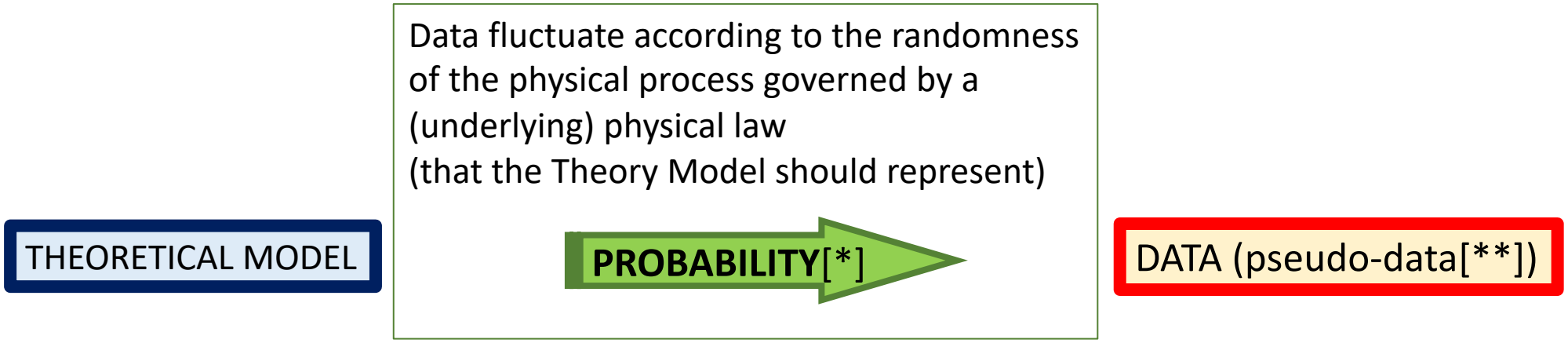
- are predicted by a theory model,
- depend on some theory parameters,
- e.g.: particle mass, cross section, etc.

Given our data sample, we want to:

- measure theory parameters,
  - e.g.: m_t= 173.49 ± 1.07 GeV, mH =125.38 GeV
- answer questions about the nature of data
  - Is there a Higgs boson? ➔ Yes! (strong evidence? Quantify!)
  - Is there a Dark Matter? ➔ No evidence, so far…
  - If not, what is the range of theory parameters compatible with the observed data? What parameter range can we exclude?

We should use probability theory on our data and our theory model in order to extract information that will address our questions → i.e.: we use statistics for data analysis

Relation between Probability & Inference - I

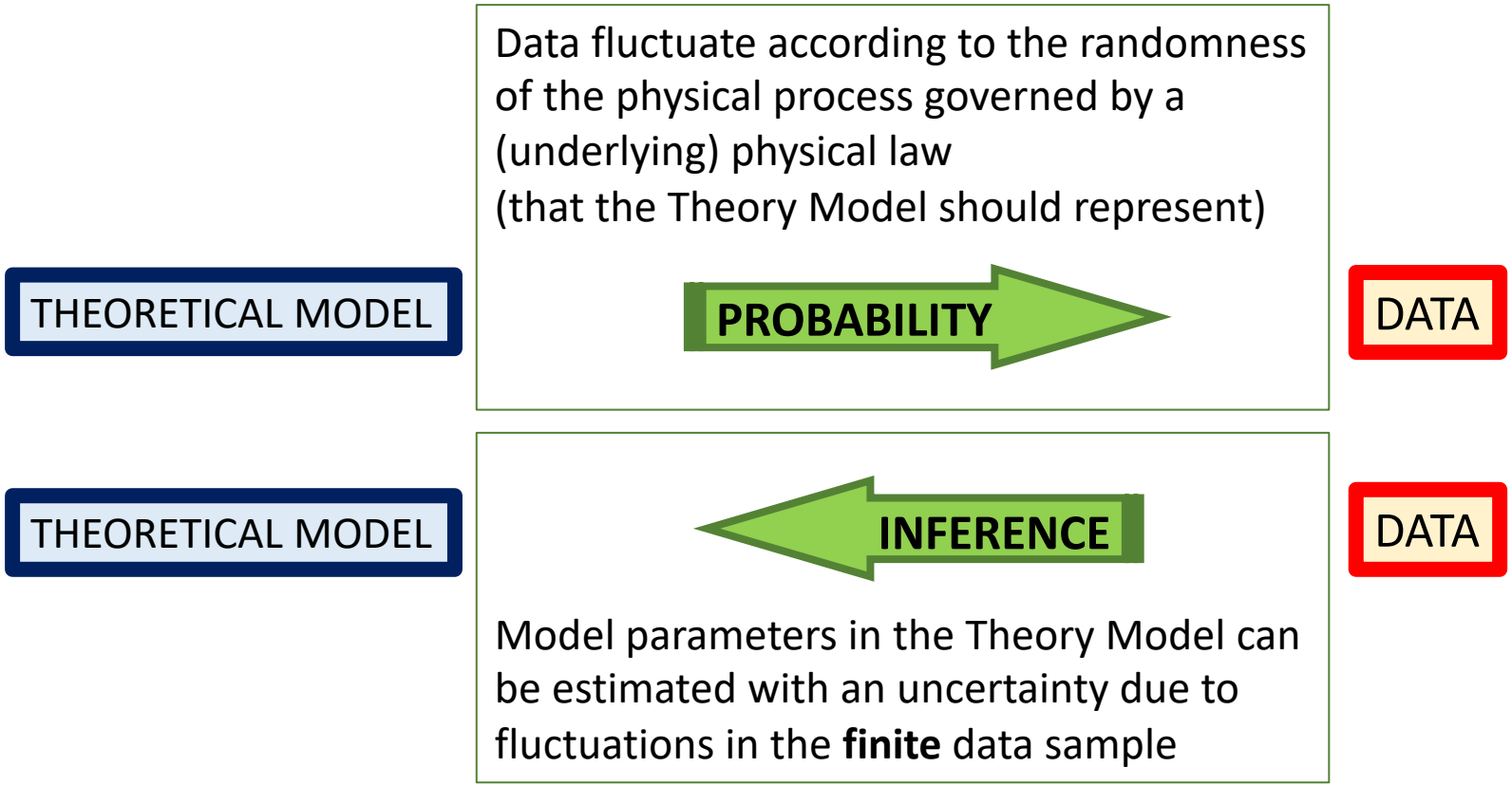


Known (or assumed correct) the physical process of generation of data (probabilistic model)
… we are able to evaluate the probability of the different outcomes of an experiment

[*] because of the randomness of the process/law ... the calculation of probabilites is involved

[**] when we generate Data according to a model (***Monte Carlo*** generators) we speak about ***pseudo-data***

Relation between Probability & Inference - I



In the **statistical inference** the approach is somehow reverted w.r.t. the **theory of probability**: the physical process or law is under investigation and the statistical methods & techiques try to induce the characteristics of the process on the basis of the (finite) experimental observations

Concept of Probability - I

Many processes in nature have uncertain outcomes (their result cannot be predicted in advance).

It is useful to introduce the concept of **random variable**: it represents the outcome of a ***repeatable*** **experiment** whose result is uncertain. Then an **event** consists of the occurrence of a certain particular condition about the value of the random variable resulting from an experiment (in simple words: it is a possible outcome of an experiment).

Note: often in physics : an ***event*** is meant as an ***elementary event***, i.e. it represents a *single outcome*;
on the countrary, in statistics : an ***event*** can represent - in general - a ***subset of possible outcomes***.

Classical probability : if N is the total number of possible outcomes ("cases") of a random variable,
if n is the number of favourable cases for which an event A is realized,
the **probability of an event A** is: $\mathbf{P}(A) = \frac{n}{N}$

(P.S.Laplace, 1749-1827)

Concept of Probability - II

Most experiments in Physics can be ***repeated*** under the same - or at least very similar - conditions.
Such experiments are examples of *random processes* in the sense that, at every repetition, a different outcome is observed.
The result of an experiment may be used to address questions about natural phenomena, …
... for instance about the knowledge of an unknown physical quantity, or the existence or not of some new phenomena.
Statements that answer those questions can be assessed by assigning them a *probability*.
Different definitions of probability apply to cases in which statements refer to repeatable experiments or not:

⇨ ***Frequentist probability*** only applies to processes that can be repeated over a reasonably long period of time:

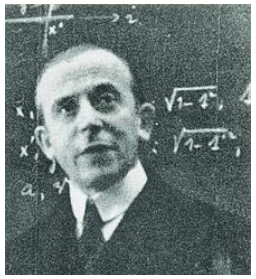
⇨

Concept of Probability - II

Most experiments in Physics can be ***repeated*** under the same - or at least very similar - conditions.
Such experiments are examples of *random processes* in the sense that, at every repetition, a different outcome is observed.
The result of an experiment may be used to address questions about natural phenomena, …
... for instance about the knowledge of an unknown physical quantity, or the existence or not of some new phenomena.
Statements that answer those questions can be assessed by assigning them a *probability*.
Different definitions of probability apply to cases in which statements refer to repeatable experiments or not:

⇨ ***Frequentist probability*** only applies to processes that can be repeated over a reasonably long period of time:

(R.Von Mises, 1883-1953)



Frequentist probability : is the fraction of the number (N_i) of possible occurrences of an event E_i over the total number of events (N) in a repeatable experiment, in the limit of a very large number of experiments: $\mathbf{P}(E_i) = \lim_{N \to \infty} \frac{N_i}{N}$

Note: - this limit must be intended in an experimental (non mathematical!) sense
- the true value of the probability would be found only repeating ∞ times the (repeatable) experiment
- in many cases, experience shows that the frequentist probability tends to the classical one (thanks to the ***Law of large numbers***) [ex.: roll a not-loaded dice & execute a large number of rolls]

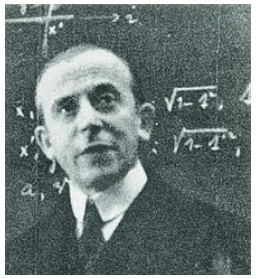
⇨

Concept of Probability - II

Most experiments in Physics can be ***repeated*** under the same - or at least very similar - conditions.
Such experiments are examples of *random processes* in the sense that, at every repetition, a different outcome is observed.
The result of an experiment may be used to address questions about natural phenomena, …
... for instance about the knowledge of an unknown physical quantity, or the existence or not of some new phenomena.
Statements that answer those questions can be assessed by assigning them a *probability*.
Different definitions of probability apply to cases in which statements refer to repeatable experiments or not:

⇨ ***Frequentist probability*** only applies to processes that can be repeated over a reasonably long period of time:

(R.Von Mises, 1883-1953)



Frequentist probability : is the fraction of the number (N_i) of possible occurrences of an event E_i over the total number of events (N) in a repeatable experiment, in the limit of a very large number of experiments: $\mathbf{P}(E_i) = \lim_{N \to \infty} \frac{N_i}{N}$

Note: - this limit must be intended in an experimental (non mathematical!) sense
- the true value of the probability would be found only repeating ∞ times the (repeatable) experiment
- in many cases, experience shows that the frequentist probability tends to the classical one (thanks to the ***Law of large numbers***) [ex.: roll a not-loaded dice & execute a large number of rolls]

⇨ ***Bayesian probability*** applies also to an hypothesis or statement that can be true (or false): the probability of a certain hypothesis (or theory) is represented by the ***degree-of-belief*** *(subjective)* that the hypothesis is true (or false).

Interpretation of Probability

We have just introduced **two different interpretations** of the probability: Frequentist & Bayesian probabilities; note that both are consistent with Kolmogorov axioms.

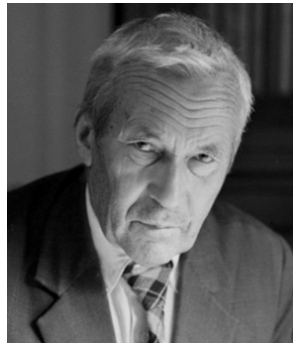
- **Frequentist probability** refers to a **relative frequency** that can be evaluated for repeatable experiments (for instance when we measure particle scatterings or radioactive decays).
  In this course we will assume/use/refer-to … this concept of probability.

- **Bayesian probability** refers to a **subjective probability** where instead of outcomes we have hypotheses (statements that can be true or false).

In particle physics the frequency interpretation is often most useful, but subjective probability can provide more natural treatment of non-repeatable phenomena (for instance the probability that Higgs boson exists, or in handling systematic uncertainties).

In most cases the two approaches give (asymptotically) similar results.

Axiomatic approach to Probability



To formalize - in a correct mathematical way - the concept probability, A.N.Kolmogorov (1903-1987) proposed (1933) an **axiomatic approach** (the ***set theory*** can help intuitively to handle axioms and theorems):

- being… Ω the set of possible outcomes, $E \in \Omega$ a certain possible outcome/result/event)

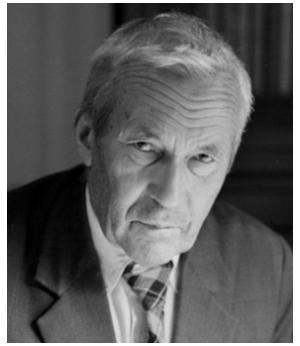
Axiom-1 : $P(\Omega) = 1$ (i.e. the experiment must have a result) [it's the ***normalization condition*** !]

Axiom-2 : $P(E \in \Omega) \geq 0$

Axiom-3: property of additivity : $P\left(\bigcup_i E_i\right) = \sum_i P(E_i)$ for ALL E_i being DISJOINT

union

Axiomatic approach to Probability



To formalize - in a correct mathematical way - the concept probability, A.N.Kolmogorov (1903-1987) proposed (1933) an **axiomatic approach** (the ***set theory*** can help intuitively to handle axioms and theorems):

- being… Ω the set of possible outcomes, $E \in \Omega$ a certain possible outcome/result/event)

Axiom-1 : $P(\Omega) = 1$ (i.e. the experiment must have a result) [it's the ***normalization condition*** !]

Axiom-2 : $P(E \in \Omega) \geq 0$

Axiom-3: property of additivity : $P\left(\bigcup_i E_i\right) = \sum_i P(E_i)$ for ALL E_i being DISJOINT

union

Every concept/definition of probability is required to be compatible with the axiomatic probabiity and with the derived …

… properties: $P(E) = 1 - P(E^*)$, $P(E \in \Omega) \leq 1$, $P(\emptyset) = 0$

… & theorems: **Additivity theorem** : $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$ with $E_1, E_2 \in \Omega$ GENERIC (→ NOT NECESSARILY DISJOINT)

(it can be easily demonstrated)

intersection

relative complement

⇨ includes Axiom-3 if E_1, E_2 are disjoint : $P(E_1 \cap E_2) = 0 \Rightarrow P(E_1 \cup E_2) = P(E_1) + P(E_2)$

Joint Probability

Joint probability : $P(A \cap B)$: probability that two events (A & B) happen concurrently

$= 0$ IF A & B DISJOINT ($A \cap B = \emptyset$)

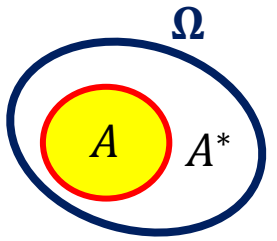
$= P(A) \cdot P(B)$ IF A & B INDEPENDENT

$= P(A) + P(B) - P(A \cup B)$ IF A & B GENERIC ⇦ from the Additivity Theorem!

To deal with **non** independent events we have to introduce the concept of ***conditional probability*** (next slide)

Conditional Probability

Suppose to *restrict* the possible outcomes of an experiment to the subset $A \subset \Omega$ and introduce the …

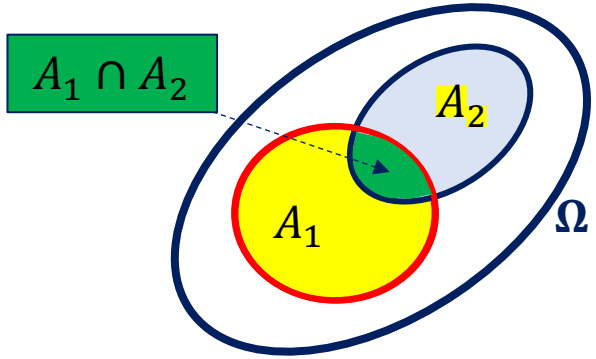


Conditional probability : $P(E|A)$: probability of event E given the restriction $A \subset \Omega$

Note: if $A^* \neq \emptyset$ it holds $P(E|A) > P(E)$; this introduces the need to "renormalize" the conditional probability: $P(A|A) \equiv 1$

The following properties hold:

1) $P(A_2|A_1) = P(A_1 \cap A_2|A_1)$ [see figure]



2) ratios of probabilities should not change with the applied restriction:

$$\frac{P(A_1 \cap A_2)}{P(A_1)} = \frac{P(A_1 \cap A_2|A_1)}{P(A_1|A_1) \to 1}$$

Putting together (1) & (2) : $\frac{P(A_1 \cap A_2)}{P(A_1)} = P(A_2|A_1)$

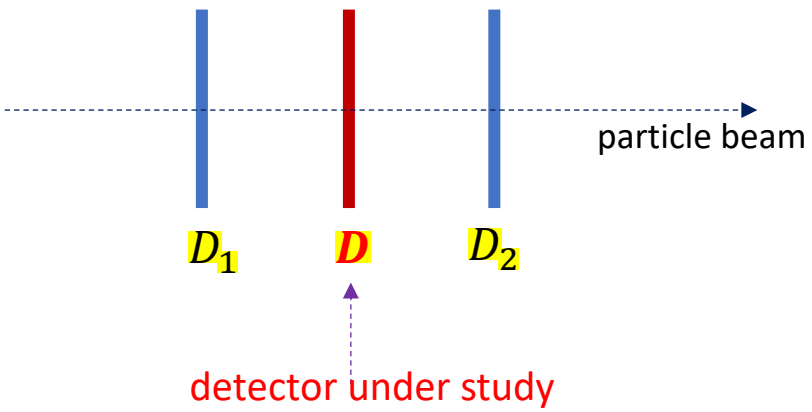
For completeness (and coherence) we define : $P(A_2|A_1) = 0 \quad IF \quad P(A_1) = 0$

We can now formally define the conditional probability: $P(B|A) = \frac{P(B \cap A)}{P(A)}$: probability of event B given the event A already happened

For ***independent*** events: $P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(B) \cdot P(A)}{P(A)} = P(B)$ (just another way to express independence)

Note: it can be demonstrated that is satisfies the axioms of Kolmogorov

Application of previous concepts - I



Detection efficiencies are probabilities !

To measure the detection efficiency of the detector under test we need to select ***all and only*** the particles that cross the system and are detected by both "telescope" detectors D_1 & D_2 (that are read in ***time coincidence***).
The intersection expresses the time coincidence in the sense that the probability to have a particle of the beam detected by both of them is given by $P(D_1 \cap D_2)$ [reminder: intersection is a logical-AND]!

Of course, $P(D_1 \cap D_2)$ is a joint probability but note that the two "telescope" detectors work independently, thus:

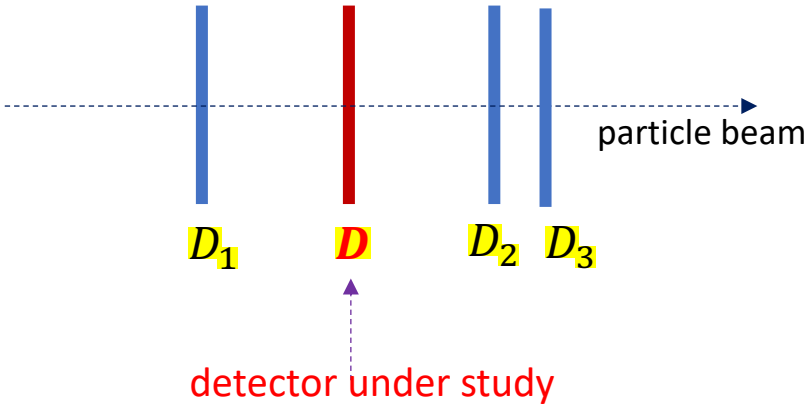
$$P(D_1 \cap D_2) = P(D_1) \cdot P(D_2)$$

As seen in previous slide, $P(D_1 \cap D_2)$ can also be expressed in terms of conditional probability as follows:

$$P(D_1 \cap D_2) = P(D_2|D_1) \cdot P(D_1)$$

and since the detectors work independently it holds $P(D_2|D_1) = P(D_2)$.

Application of previous concepts - II



Adding a third detector as in the figure implies …
to have detector D_1 in coincidence with **any** of one between D_2 & D_3 !

The involved joint probability is now: $P(D_1 \cap (D_2 \cup D_3))$

[reminder : intersection is a logical-AND, union is a logical-OR]

Now we get : $P(D_1 \cap (D_2 \cup D_3)) = P(D_1) \cdot P(D_2 \cup D_3) = P(D_1) \cdot [P(D_2) + P(D_3) - P(D_2 \cap D_3)]$

additivity theorem

Since also the detectors D_2 & D_3 work independently it holds: $P(D_2 \cap D_3) = P(D_2) \cdot P(D_3)$

Overall : $P(D_1 \cap (D_2 \cup D_3)) = P(D_1) \cdot [P(D_2) + P(D_3) - P(D_2 \cap D_3)] = P(D_1) \cdot [P(D_2) + P(D_3) - P(D_2) \cdot P(D_3)]$

In this way the ***total efficiency*** of the telescope can be calculated to know the useful particle flux to study the detector under test. It can be easily calculated that … passing from a telescope with 2 similar detectors to one with 4 similar ones increases the total efficiency by a multiplicative factor $(2 - \varepsilon_D)^2$ where ε_D is the detection efficiency of a single detector.

Bayes' theorem - I

This famous theorem by T.Bayes relates the two conditional probabilities $P(B|A)$ with $P(A|B)$ where $A, B \in \Omega$

(T.Bayes, 1702-1761)

We've already written $P(B|A) = \frac{P(B \cap A)}{P(A)}$ but we can equally write (A, B are exchangeable): $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Putting together: $P(A|B) \cdot P(B) = P(A \cap B) = P(B|A) \cdot P(A)$. Thus : $P(A|B) = P(B|A) \cdot \frac{P(A)}{P(B)}$

A generalization/extension of the theorem can be obtained by introducing the ***Law of the total probability*** as follows:

if we have sets of events $\{A_i\}_i$ that are **disjoint** and **fully cover Ω** (namely $\Omega = \bigcup_i A_i$) and $\boldsymbol{B \in \Omega}$ is a generic event, we can calculate $P(B)$ exploiting the fact that $B = B \cap \Omega = B \cap \bigcup_i A_i = \bigcup_i (B \cap A_i)$ and $(B \cap A_i)$ are disjoint, thus the total probability can be obtained by the following sum:

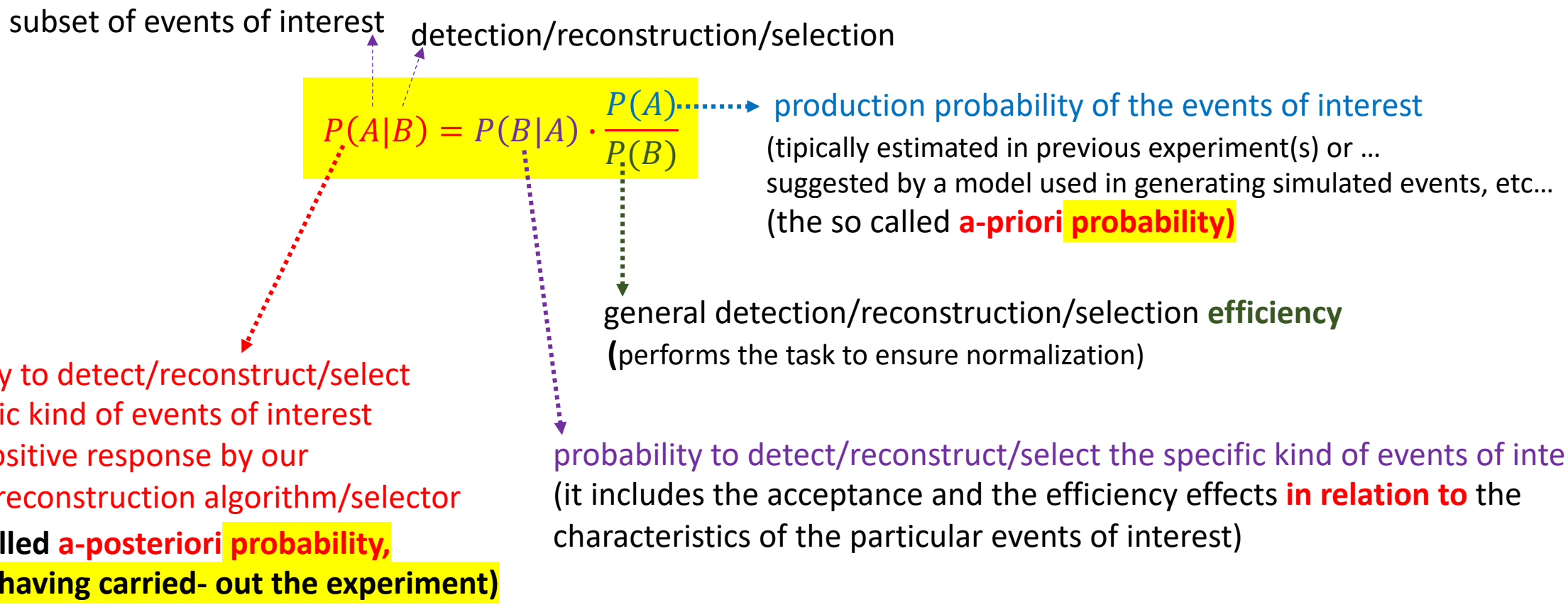
$$\boldsymbol{P(B)} = P\left(\bigcup_i (B \cap A_i)\ (B \cap A_i) \bigcup_i\right) = \sum_i P(B \cap A_i) =$$

representing the so called **Law of total probability**

and Bayes' theorem can be rewritten: $P(A|B) = \frac{P(B|A) \cdot P(A)}{\sum_i \boldsymbol{P(B|A_i) \cdot P(A_i)}}$ (nothing forbids A to be one of the A_i)

Bayes' theorem - II

This theorem can be discussed in a **frequentist context** (in which a probability cannot be associated to an hypothesis!), [and it can be helpful when designing an experiment] in the following way:



Theory / Hour-2

CHARACTERISTICS of MEASUREMENTS

STATISTICAL & SYSTEMATICS UNCERTAINTIES - I

When we carry out an experimental measurement we must separate the purely **statistical component** from those "non statistical" (called **systematics components**):

$$measure(\text{“}central\ value\text{”}) \pm statistical\ uncertainty \pm systematic\ uncertainty : m\ ^{+a}_{-b}\ ^{+c}_{-d}$$

A good measurement requires to be able to reduce as much as possible both uncertainties.

IF we have accumulated not much data (low statistics)… we can afford a conservative evaluation of the sources of systematics uncertainties (approximated by excess)

IF we have accumulated a lot of data (high statistics)… the statistical uncertrainty will be relatively small and…
…we cannot afford a conservative evaluation of systematics uncertainties: we must evaluate the systematics effect with good accuracy with the aim to bring the systematic uncertainties to the same level of the statistical uncertainty !

STATISTICAL & SYSTEMATICS UNCERTAINTIES - II

- Recap: @ “low” statistics : we can afford systematic uncertainty ≲ statistical uncertainty (relatively large)

  @ “high” statistics : we must work so that systematic uncertainty ≈ statistical uncertainty (relatively small)

- If the problem is particular difficult to require the execution - on a computing machine - of the simulation (MC) of your physical system under exam, in order to compare real and simulated data, …

  … it can happen to identify a systematic error (“**bias**”) in the real data and to correct the measurement (central value) according to a **correction (“shift”)** derived from the data-MC comparison.

  In this circumstance **the statistical uncertainty on the measurement carried out on the simulated data must be considered a systematic uncertainty for the (corrected) measurement in real data**.

  This implies the need to have enough statistics for your simulated data samples.

Example: https://arxiv.org/pdf/hep-ex/9902011.pdf (CLEO experiment’s charmed mesons lifetime measurement)[see next slide]

The systematic uncertainties for the D meson lifetimes are listed in Table I and are described below. They can be grouped into three categories:

Reconstruction of the D decay length and proper time. Errors in the measurement of the reconstructed decay length can be due to errors in the measurement of the decay vertex, the global detector scale, and the beam spot. The bias in the decay vertex position is estimated to be $(0.0 \pm 0.9\ \mu\text{m})$ from a "zero-lifetime" sample of $\gamma\gamma \to \pi^+\pi^-\pi^+\pi^-$ events. This corresponds to a measured proper-time uncertainty of ± 1.8 fs. In addition, the vertex reconstruction is checked with events with interactions in the beam pipe with a relative uncertainty of $\pm 0.2\%$. The sums of these uncertainties in quadrature yield the systematic uncertainties due to the decay vertex measurement. The global detector scale is measured to a precision of $\pm 0.1\%$ in surveys and confirmed in the study of events with interactions in the beam pipe. The changes in the lifetimes due to the variation ($\pm 2\ \mu\text{m}$) in the vertical beam spot position and height are another source of systematic error, since the interaction point is calculated from the beam spot and the reconstructed D momentum and decay vertex. Statistical uncertainties for the D masses [2] and the D momentum measurements lead to systematic errors since these quantities are used to convert the decay length into proper time.

STATISTICAL & SYSTEMATICS UNCERTAINTIES - IV

TABLE I. Systematic uncertainties for the D meson lifetimes in fs. The systematic uncertainties for the three D^0 modes are weighted with the same weights as the fitted D^0 lifetimes.

Uncertainty	D^0 $K^-\pi^+$	D^0 $K^-\pi^+\pi^0$	D^0 $K^-\pi^+\pi^-\pi^+$	D^0 combined	D^+ $K^-\pi^+\pi^+$	D_s^+ $\phi\pi^+$
Decay vertex	± 2.0	± 2.0	± 2.0	± 2.0	± 2.8	± 2.1
Global detector scale	± 0.1	± 0.1	± 0.1	± 0.1	± 0.1	± 0.1
Beam spot	$^{+0.3}_{-0.1}$	$^{+2.1}_{-0.0}$	$^{+0.3}_{-0.2}$	$^{+0.8}_{-0.1}$	$^{+1.3}_{-1.1}$	$^{+0.7}_{-1.1}$
D meson mass	± 0.1	± 0.1	± 0.1	± 0.1	± 0.3	± 0.1
D meson momentum	$^{+0.2}_{-0.0}$	$^{+0.1}_{-0.2}$	$^{+0.3}_{-0.1}$	$^{+0.2}_{-0.1}$	$^{+0.6}_{-0.0}$	± 0.1
Signal probability	$^{+0.4}_{-0.1}$	$^{+0.1}_{-0.2}$	$^{+0.1}_{-0.2}$	$^{+0.3}_{-0.1}$	$^{+1.2}_{-8.1}$	$^{+1.3}_{-1.8}$
$t - M(D)$ correlation	± 0.6	± 0.6	± 1.0	± 0.7	± 1.7	± 1.5
Large proper times	± 1.2	± 3.4	± 0.2	± 1.5	± 0.3	± 0.5
Background	± 0.5	± 2.4	± 3.0	± 1.5	± 6.3	± 2.9
MC statistics	± 0.9	± 2.3	± 2.2	± 1.6	± 6.6	± 2.4
Total	$^{+2.7}_{-2.6}$	$^{+5.6}_{-5.2}$	± 4.4	$^{+3.5}_{-3.4}$	$^{+\ 9.9}_{-12.7}$	$^{+4.9}_{-5.1}$

Checking the algorithms with simulated events. Charm meson candidate selection requirements can cause systematic biases in the lifetime measurements. We estimate these biases with simulated events and correct for the biases as described above. We include the statistical uncertainties in the measured lifetimes from the samples of simulated events as systematic uncertainties in the results.

PRECISION & ACCURACY

Precision of a measurement: term that expresses that the result of a measurement can be obtained with great detail (many significative cyphers).

Numerically, it is represented by the random (or "**statistical**") **uncertainty** !

Accuracy of a measurement: term that expresses the maximum possible deviation of the result of a measurement from the result of an ideal measurement; thus it is associated to the maximum systematic error that the experimental instrumentation can introduce in the measurement.

Numerically, it's represented by the maximum "**systematic**" **uncertainty** that the used instrumentation/method can introduce!

Wrapping up:

A precise measurement is a measurement affected by a very small statistical uncertainty;
The systematic uncertainties cannot be eliminated but enough (hopefully strongly) reduceable.

An accurate measurement is a measurement affected by a minimized systematic uncertainty (or anyway, lower than the statistical uncertainty;
The systematic uncertainties cannot be eliminated but hopefully can be *minimized*.

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 $(x_1, \dots, x_N)$ 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.

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

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

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 $f(\vec{x}) = f(x_1, \dots, x_n)$ 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: $\int_{\Omega} f(x_1, \dots, x_n) d^n x = 1$

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

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

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

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

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

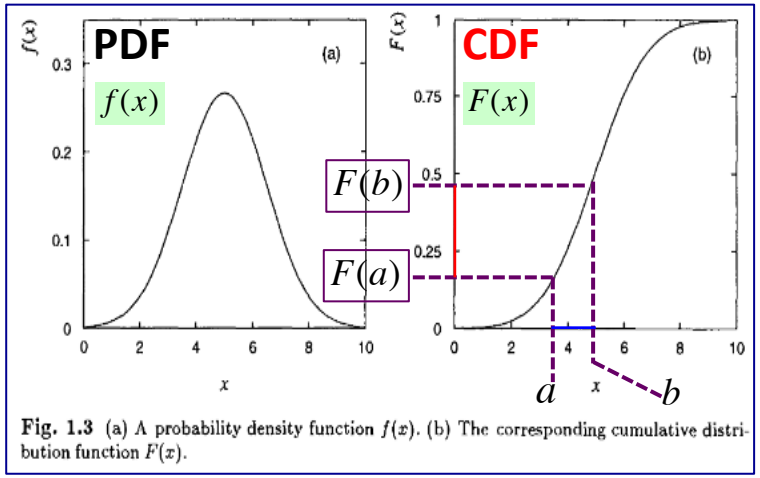


Fig. 1.3 (a) A probability density function $f(x)$. (b) The corresponding cumulative distribution function $F(x)$.

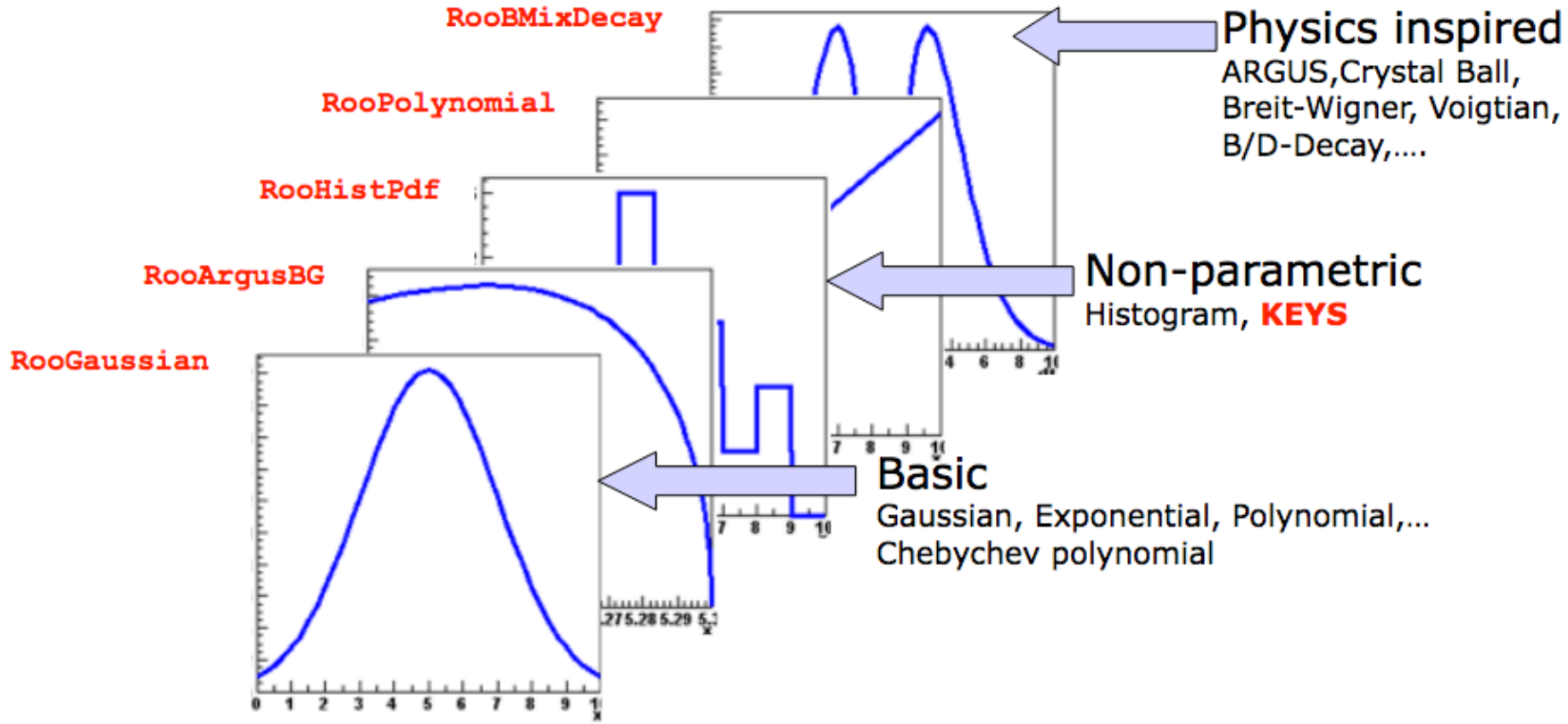
The c.d.f. can be considered to be more fundamental than the p.d.f. since the c.d.f. is an actual probability rather than a probability density. However, in applications we usually need the p.d.f. Sometimes it is easier to derive first the c.d.f. from which you get the p.d.f. by

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

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

Library of p.d.f.s in ROOT/RooFit

- RooFit provides a collection of compiled standard PDF classes

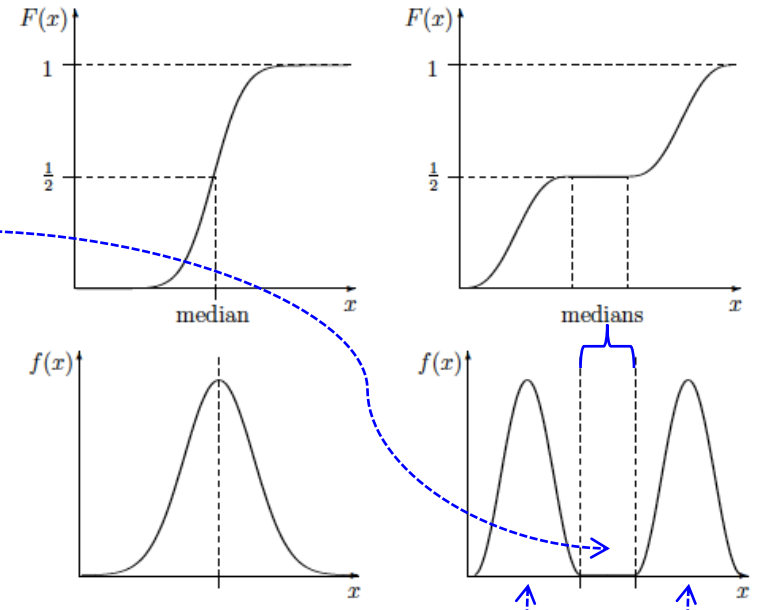


Easy to extend the library: each p.d.f. is a separate C++ class

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

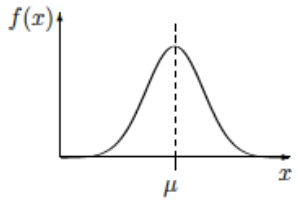
Attribute of a p.d.f. : expectation value

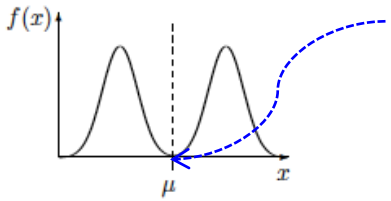
Expectation value of a p.d.f. (sometimes called "***Mean***" which is very misleading actually! Better ***population mean***): represents the **central value of a p.d.f.** and it is defined as:

$$\mu \equiv E[x] = \int_{-\infty}^{+\infty} x\, f(x)\, dx$$

Note: $E[x]$ is not a function of ***x*** (there is an integral on ***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.





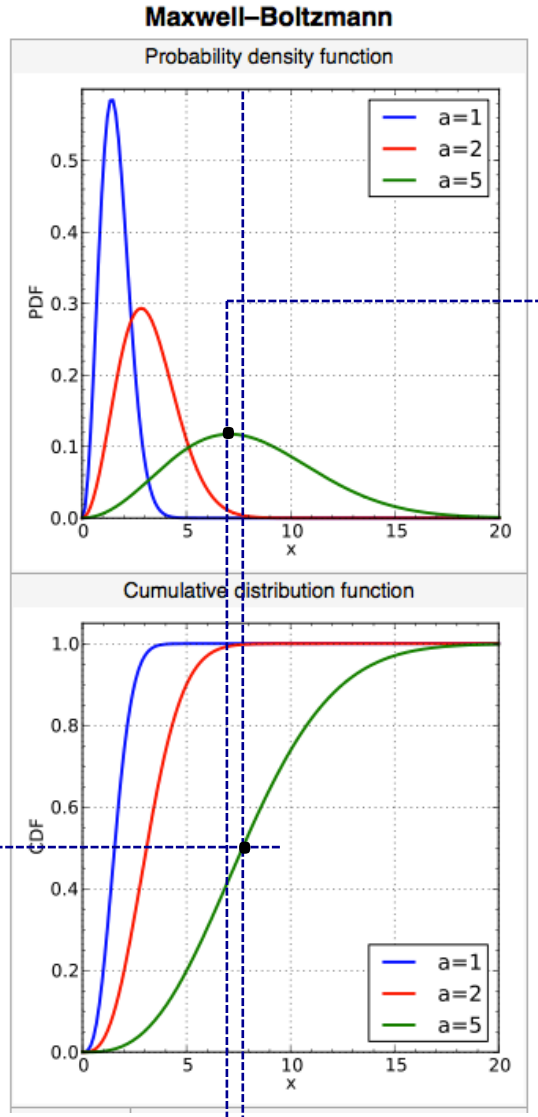
it can even happen that it is a value never taken by the x !

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

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

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



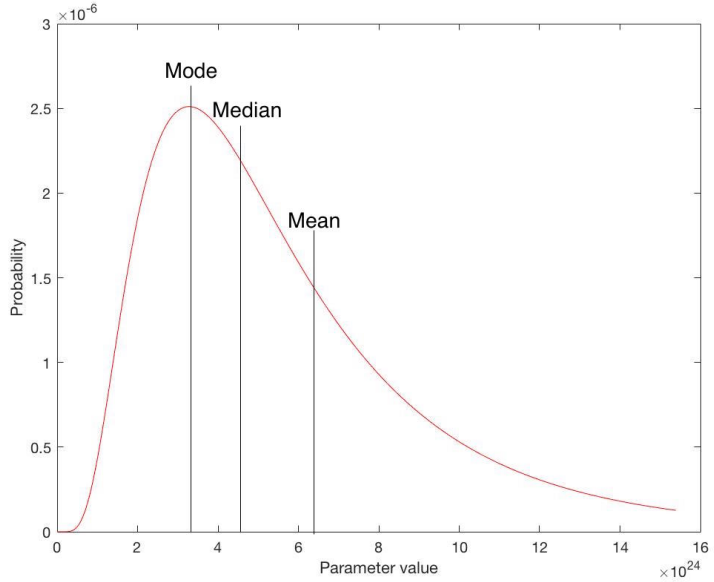
Mode

$\frac{1}{2}$

Median **> Mode**

For this distribution:

the expectation value ("Mean") > Median



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

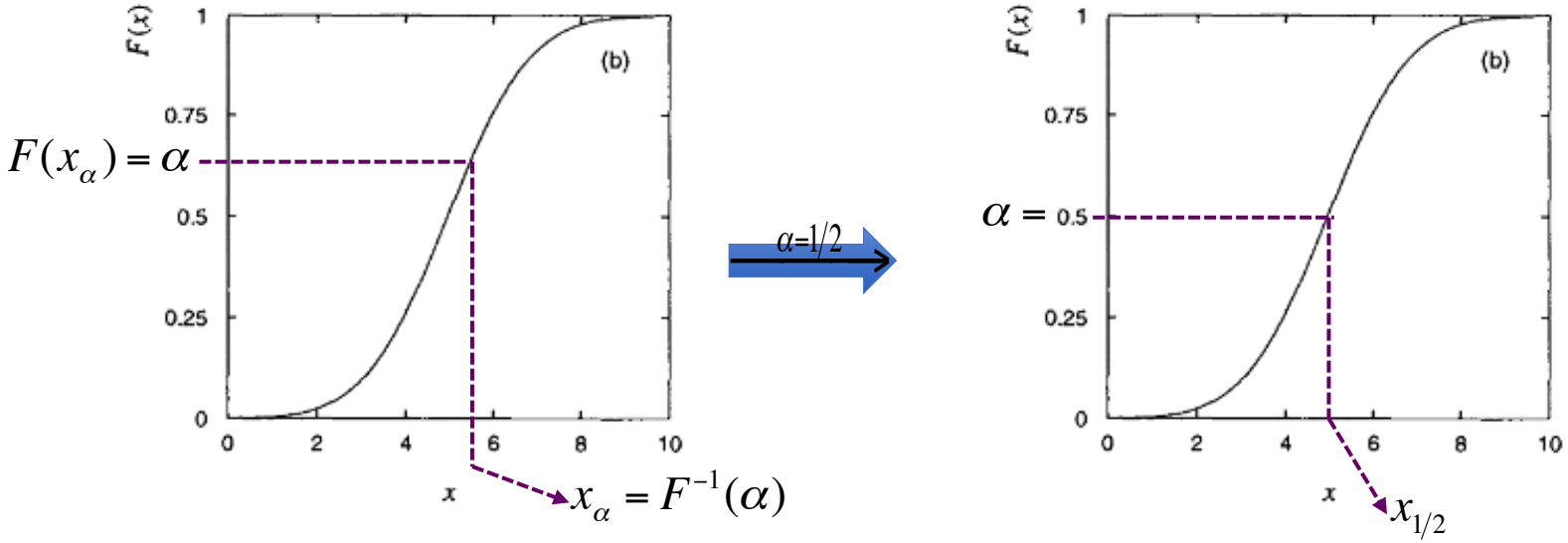
Attribute of a c.d.f. : quantile of order α

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 $x_{1/2}$.

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



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 ≡ 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 :

  =1 (normalization)

$$E[(x-\mu)^{m=1}] = \int_{-\infty}^{+\infty} (x-\mu)f(x)dx = \int_{-\infty}^{+\infty} xf(x)dx - \mu\int_{-\infty}^{+\infty} f(x)dx = \int_{-\infty}^{+\infty} xf(x)dx - \mu = E[x] - \mu = \mu - \mu = 0$$

  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!

Attribute of a p.d.f. : variance

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

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

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

The **squared root of the variance** is called **standard deviation of x** **and denoted by σ_x .**

It is often useful because **it has the same dimentional units of 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: $\mathrm{V}[ax] = E[a^2x^2] - (E[ax])^2 = a^2\, E[x^2] - (aE[x])^2 = a^2 \cdot (E[x^2] - (E[x])^2) = a^2 \cdot V[x]$