Probability and Statistics

for experimental physicists : part I

José Ocariz
Université Paris Diderot and IN2P3

Outline

Lecture 1 (today)

Basic concepts in Probability and Statistics

Lecture 2 (Tuesday)

Maximum Likelihood theorem
Multivariate techniques

Lecture 3 (Thursday)

An analysis example from *BaBar*
Hypothesis testing, limit settings

Disclaimer

Most, if not all of you, are already familiar with many of these topics… for consistency, the scope spans from the very general concepts towards more advanced developments…

Bibliography

"The" classical reference book (912 pages) :

Stuart, K. Ord, S. Arnold, *Kendall's Advanced theory of statistics Volume 2A : Classical Inference and and the Linear Model,* John Wiley & Sons, 2009

Books on statistics, written by particle physicists, well suited for everyday's needs :

L. Lyons, *Statistics for Nuclear and Particle Physics,* Cambridge, 1986

G. Cowan, *Statistical Data Analysis*, Clarendon Press, Oxford 1998

see also http://www.p.rhul.ac.uk/~cowan/stat_course.htm

R.J. Barlow, *A Guide to the Use of Statistical Methods in the Physical Sciences* John Wiley & Sons, 1989

F. James, *Statistical Methods in Experimental Physics*, World Scientific, 2006

The *PDG* is a convenient source for quick reference :

J. Beringer *et al*. (Particle Data Group), Phys. Rev. D86, 010001 (2012) (« Mathematical Tools » section)

"Must-have" in your bookmarks, and open during most of your working time :

The ROOT users' guide

The RooFit user's guide

The TMVA user's guide

Probability

Mathematical probability

abstract axiomatic concept, developed by Kolmogorov (1933)

Probability theory : the tool to quantify our knowledge of *random processes*

A process is called random if :

- **its outcome ("an event") cannot be predicted with complete certainty**
- **if repeated under the same conditions, the outcome can be different**

In practice, the underlying sources of uncertainty can be :

- **fundamental : quantum mechanics is not a deterministic theory**
  - **particle physics is an excellent example !**
- **due to irreducible random measurement errors (i.e. thermal effects)**
- **due to reducible measurement errors (i.e. practical instrumental limitations)**

- **Let Ω be the total universe of possible outcomes (also called sample space)**
- **Let ω=A,B,... be elements of Ω**

A probability function P is defined as a map into the real numbers :

$$P : \{\Omega\} \rightarrow [0:1]$$
$$\omega \rightarrow P(\omega)$$

The mapping must satisfy the following axioms :

$$P(\Omega) = 1$$
$$if \;\; A \cap B = \varnothing \;,\; \text{then} \;\; P(A \cup B) = P(A) + P(B)$$

From which various useful properties are easily derived, i.e.

$$P(\overline{A}) = 1 - P(A)$$
$$P(A \cup \overline{A}) = 1$$
$$P(\varnothing) = 1 - P(\Omega) = 0$$
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

<u>Conditional probability :</u> by restricting the sample space Ω to a subsample B (with $P(B)\neq 0$)

$$P(A \mid B) \;=\; \text{probability of A given B}$$

<u>Independence :</u> events A and B are said to be independent (that is, their realizations are not linked in any way) if

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

If A and B are actually independent, $P(A \mid B) \;=\; P(A)$

<u>Bayes' theorem :</u> since $P(A \cap B) \;=\; P(B \cap A)$ ***one has***

$$\boxed{P(A \mid B) \;=\; \frac{P(B \mid A)P(A)}{P(B)}}$$

Useful situation: if Ω is divided into disjoint subsets A_i ("a partition"),

$$P(A \mid B) \;=\; \frac{P(B \mid A)P(A)}{\sum_i P(B \mid A_i)P(A_i)}$$

Numerical outcome of a random process (i.e. a measurement) : to each event X corresponds a number x (can be a discrete or continuous number)

<u>Probability density function (PDF) $P(x)$:</u>

- **For a discrete variable,**

$$P(X \;\; found \;\; in \;\; x_j) \; = \; p_j \, , \; with \; \sum_j p_j \; = \; 1$$

- **For a real-valued variable,**

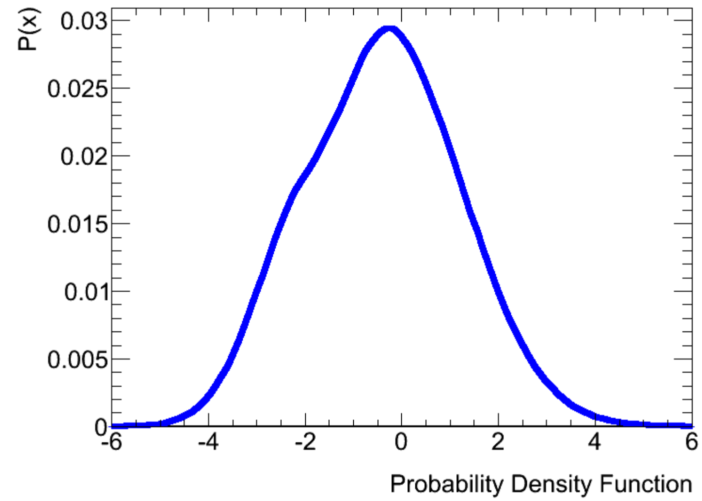
$$P(X \;\; found \;\; in \;\; [x, x+dx]) \; = \; P(x)dx$$

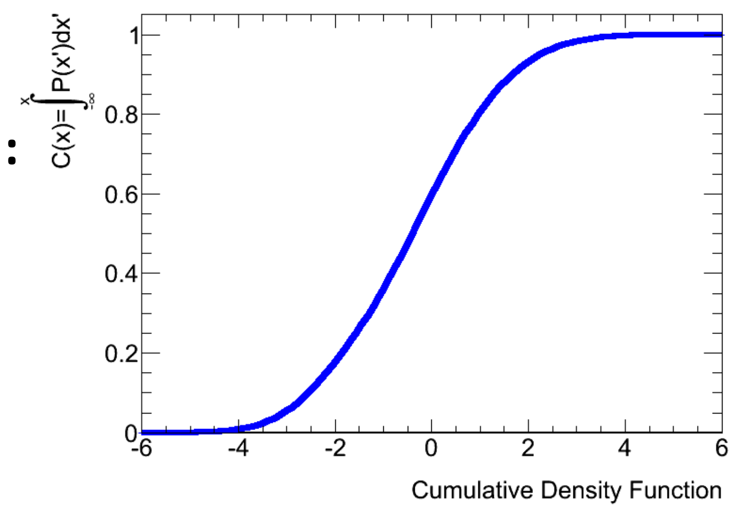
$$with \; \int_{-\infty}^{+\infty} dx' P(x') \; = \; 1$$

- **Useful definition: cumulative density (CDF) $C(x)$:**

$$C(x) \; = \; \int_{-\infty}^{x} dx' P(x')$$

$$P(a < X < b) \; = \; C(b) - C(a) \; = \; \int_a^b dx P(x)$$





Multidimensional Density Functions

Several random variables as outcome : random vectors $\vec{X} = \{X_1, X_2, \ldots, X_n\}$

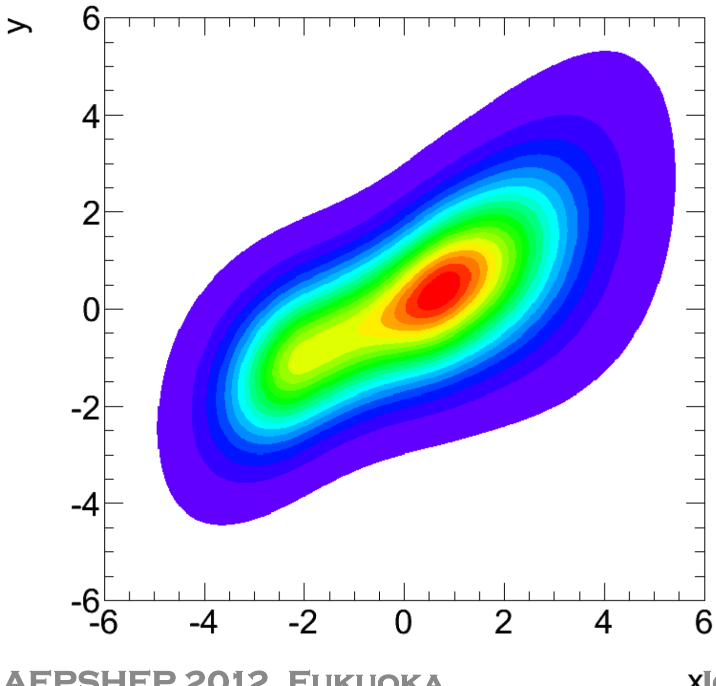
The multidimensional PDF is $P(\vec{x})d\vec{x} = P(x_1, x_2, \ldots, x_n)dx_1\, dx_2 \ldots dx_n$

Example in two dimensions : $P(a < X < b \;\; AND \;\; c < Y < d) = \int_a^b dx \int_c^d dy\, P(x,y)$

Marginal density :

$$P_X(x)dx = P(X\ in\,[x, x+dx]\, and\ Y\ in\,[-\infty, +\infty]) = dx \int_a^b dy P(x,y)$$

So that $P_X(x) = \int_a^b dy P(x,y)$



For a fixed value of Y,

$$P(x \mid y) = \frac{P(x,y)}{\int dy P(x,y)} = \frac{P(x,y)}{P_Y(y)}$$

is a *conditional* density function for X

If X,Y are independent : $P(x,y) = P_X(x) \cdot P_Y(y)$

Expectation Values

Consider a continuous random variable X with PDF $P_x(x)$. For a generic function $y(x)$, *its expectation value is defined as*

$$E[y] = \int y(x)\, P(x)\, dx$$

A few expectation values have their own name:

- **Mean value :** $\mu = E[x] = \int x\, P(x)\, dx$
- **Variance** $\sigma^2 = V[x] = E[x^2] - \mu^2 = E[(x-\mu)^2]$
- **Covariance :** $Cov[x,y] = E[xy] - \mu_x \mu_y = E[(x-\mu_x)(y-\mu_y)]$
- **The dimensionless linear correlation coefficient :** $\rho(x,y) = \dfrac{Cov[x,y]}{\sigma_x \sigma_y}$

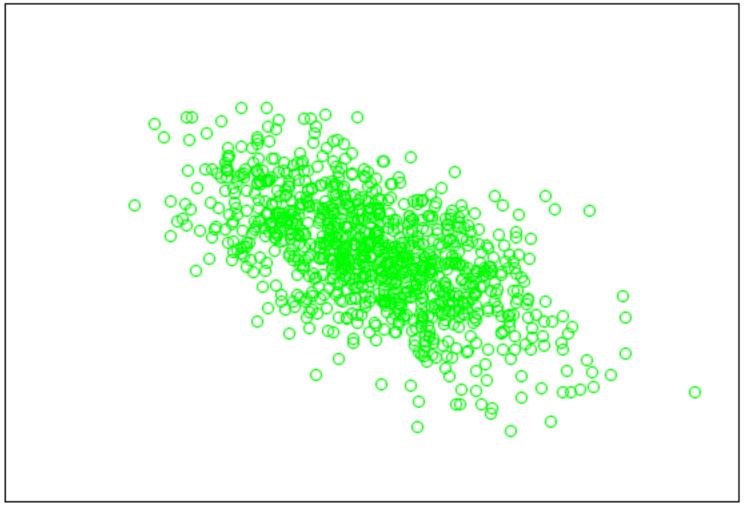
By construction, $-1 \leq \rho(x,y) \leq 1$

Note : if X,Y independent, that is $P(x,y) = P_X(x) \cdot P_Y(y)$

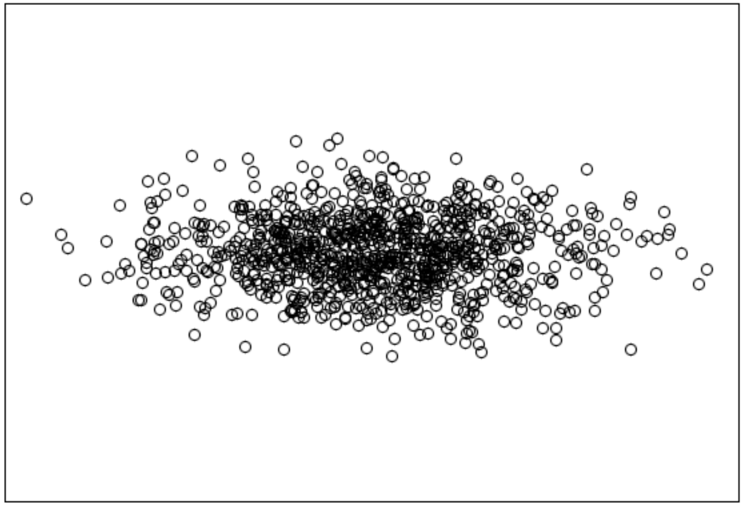
$$E[xy] = \iint xyP(x,y)\, dx\, dy = \mu_x \mu_y$$ **and thus** $\rho(x,y) = 0$

<u>(the converse needs NOT to be true!)</u>

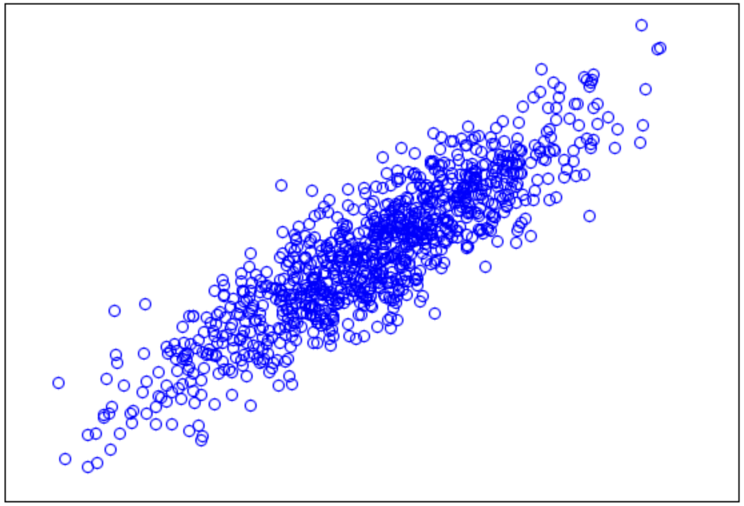
More on Correlations



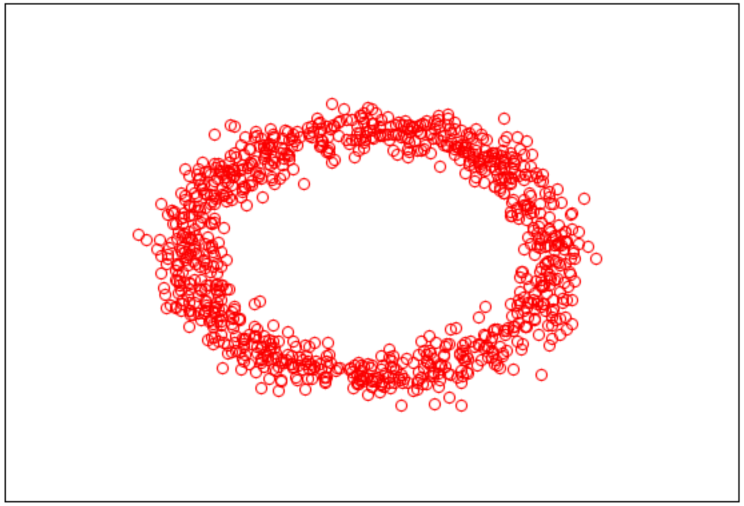
Anticorrelated variables : $\rho = -0.5$



Independent variables : $\rho = 0$



Correlated variables : $\rho = +0.9$



Correlated variables , but $\rho = 0$

Measurements : characterizing a Sample

Often, the PDF is not known, and only a finite-size sample is available (say N events)
The expectation values can be *estimated* by means of a suitable choice of *statistics*
(a *statistics* is a generic function of the reduced-size sample)

<u>Example</u> : the empirical average is an estimator of the mean value,
and characterizes the sample *location*

$$\mu = E[x] = \int x\, P(x)\, dx \quad , \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

<u>Another example :</u> the RMS (squared) is an estimator of the variance,
and characterizes the sample *dispersion*

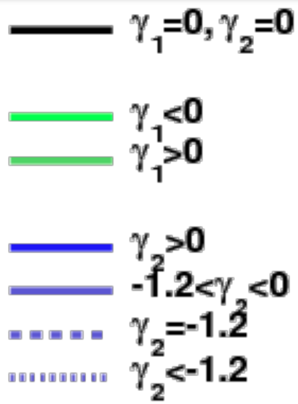
$$\sigma = \sqrt{V[x]} = \sqrt{E[x^2] - \mu^2} \quad , \quad RMS = \sqrt{\overline{x^2} - \left(\bar{x}\right)^2}$$

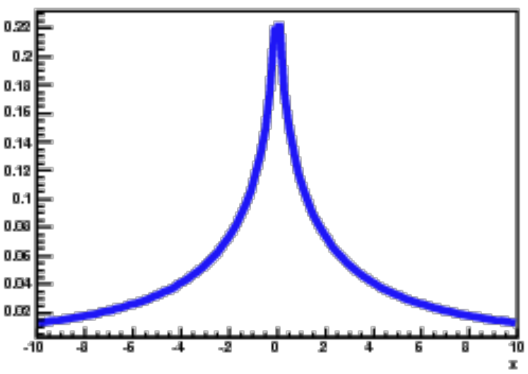
<u>Even more :</u> higher-order moments provide additional shape information :
the 3rd and 4th reduced moments estimate the *skewness* and *kurtosis* of the sample

(definition of (reduced) moments μ_k (μ''_k) follows from the Characteristic function

$$E[e^{ixt}] = \sum_k \frac{(it)^k}{k!}\mu_k \quad , \quad \mu'' = E[X''] = \frac{(X-\mu)}{\sigma}$$

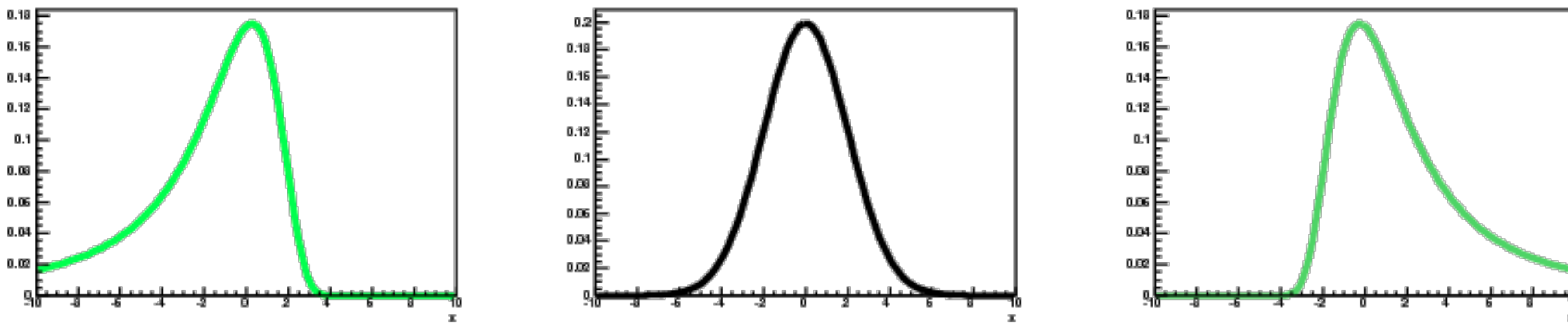
Measurements : characterizing a Sample

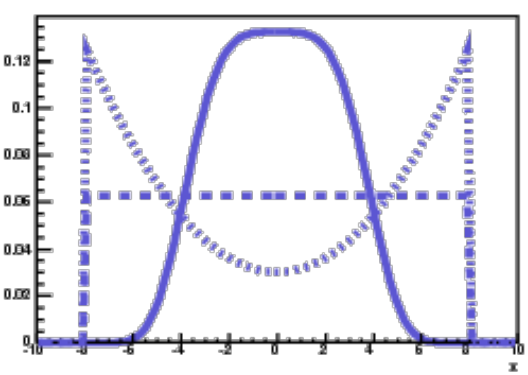




$$\gamma_1 = \mu_3''$$

$$\gamma_2 = 3 - \mu_4''$$



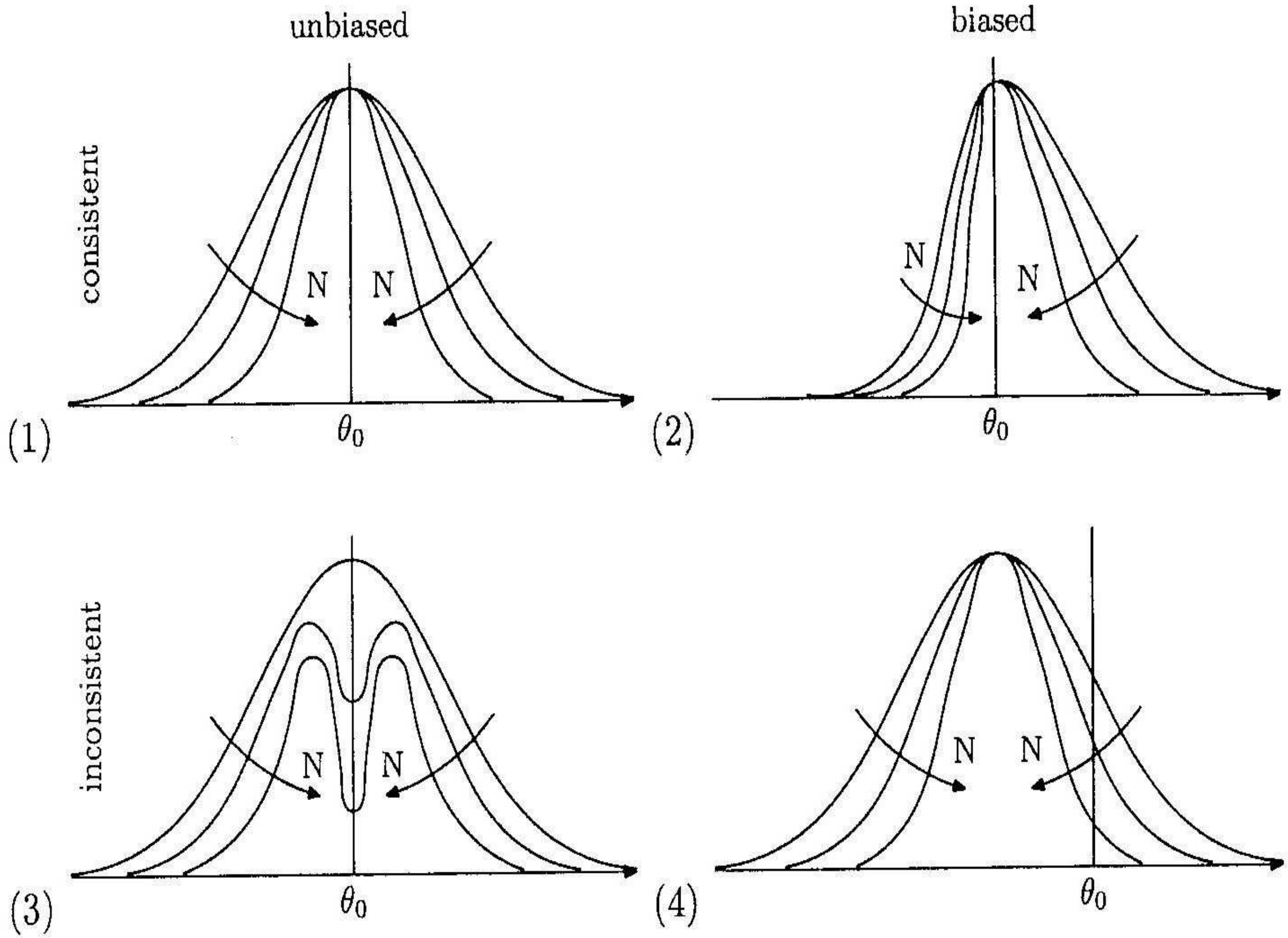


ESTIMATORS : BIAS AND ACCURACY

A ``good'' estimator should satisfy (some of) various conflicting properties :

- **be consistent,** $\lim_{n\to\infty} \overline{\theta} = E[\theta]$
- **be unbiased, or at least asymptotically unbiased**
- **Other properties :**

efficiency, robustness …



Two useful examples :

The empirical average is a convergent, unbiased estimator of the mean

$$E[\bar{x}] = \frac{1}{n}\sum_{i=1}^{n} E[x] = \mu$$

$$V[\bar{x}] = \frac{1}{n^2}\sum_{i=1}^{n} V[x] = \frac{\sigma^2}{n}$$

The RMS (squared) is a convergent, biased, asymptotically unbiased, estimator of the variance

$$RMS^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu\right)^2 - \left(\bar{x} - \mu\right)^2$$

$$E[RMS^2] = \sigma^2 - V[\bar{x}] = \frac{n-1}{n}\sigma^2$$

Estimators : Error Propagation

Consider a sample of random vectors $\vec{x} = \{x_1, x_2, \ldots, x_n\}$ **for which their covariances** $V_{ij} = \mathrm{cov}[x_i, x_j]$ **are known.**

We are interested in estimating the variance of $y(\vec{x})$ **;**
in principle it is given by $V[y] = E[y^2] - (E[y])^2$ **; in practice, one can use**

$$y(\vec{x}) = y(\vec{\mu}) + \sum_{i=1}^{n} \left[\frac{dy}{dx_i}\right]_{\vec{x}=\vec{\mu}} \left(x_i - \mu_i\right) \Rightarrow E[y(\vec{x})] \approx y(\vec{\mu})$$

$$E[y^2(\vec{x})] \approx y^2(\vec{\mu}) + 2y(\vec{\mu}) \sum_{i=1}^{n} \left[\frac{dy}{dx_i}\right]_{\vec{x}=\vec{\mu}} E[x_i - \mu_i]$$

$$+E\left[\left(\sum_{i=1}^{n} \left[\frac{dy}{dx_i}\right]_{\vec{x}=\vec{\mu}} \left(x_i - \mu_i\right)\right)\left(\sum_{j=1}^{n} \left[\frac{dy}{dx_j}\right]_{\vec{x}=\vec{\mu}} \left(x_j - \mu_j\right)\right)\right] = y^2(\vec{\mu}) + \sum_{i,j=1}^{n} \left[\frac{dy}{dx_i}\right]\left[\frac{dy}{dx_j}\right]_{\vec{x}=\vec{\mu}} V_{ij}$$

and thus

$$\sigma_y^2 \approx \sum_{i,j=1}^{n} \left[\frac{dy}{dx_i}\right]\left[\frac{dy}{dx_j}\right]_{\vec{x}=\vec{\mu}} V_{ij}$$

Estimators : Error Propagation

$$\vec{x} = \{x_1, x_2, \ldots, x_n\}$$

A few special cases :

- **if the** $\{x_i\}$ **are all uncorrelated,** $V_{ij} = \sigma_i^2 \delta_{ij}$ **and** $\sigma_y^2 \approx \sum_{i=1}^{n} \left[\frac{dy}{dx_i}\right]^2_{\vec{x}=\vec{\mu}} V_{ii}$

- **for** $y = x_1 + x_2 \, , \quad \rightarrow \quad \sigma_y^2 = \sigma_1^2 + \sigma_2^2 + 2\,\mathrm{cov}[x_1, x_2]$

  (add absolute errors in quadrature)

- **for** $y = x_1 x_2 \, , \quad \rightarrow \quad \frac{\sigma_y^2}{y^2} = \frac{\sigma_1^2}{x_1^2} + \frac{\sigma_2^2}{x_2^2} + 2\frac{\mathrm{cov}[x_1, x_2]}{x_1 x_2}$

  (add relative errors in quadrature)

- **for** $y = x_1 - x_2 \, , \; and \;\; \rho = 1 \, , \quad \rightarrow \quad \sigma_y = 0$

A Survey of useful Distributions

Distribution/PDF	Use in HEP
Binomial	Branching Ratio
Poisson	Event-counting analyses
Uniform	MonteCarlo integration
Exponential	Lifetime measurement
Gaussian	Resolution
Breit-Wigner	Mass of resonance
χ2	Goodness-of-fit

A discrete Distribution : Binomial

Consider a situation with two possible outcomes : "yes" or "no", with a fixed probability p of obtaining "yes".

If n trials are performed, $0 \leq k \leq n$ produce "yes" as outcome; only k is interesting, the sequence of trails irrelevant. This number of "yes" follows the binomial distribution,

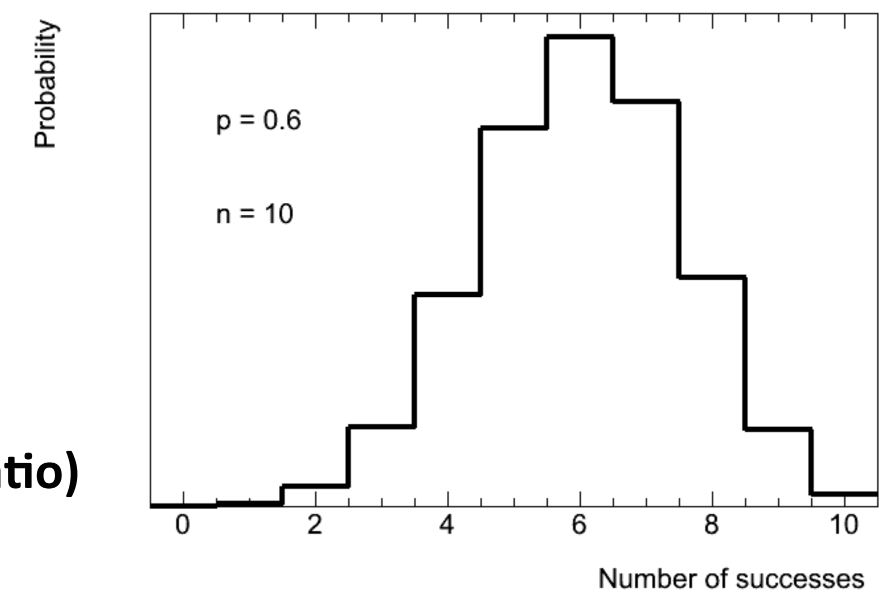
$$P_{binomial}(k; n, p) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

(k is the random variable, n and p are parameters) for which the expectation value and variance are

$$E[k] = \sum_{n=0}^{n} k \, P_{binomial}(k; n, p) = np$$

$$V[k] = E[k^2] - (E[k])^2 = np(1-p)$$

Typical example : the number of events in a specific sub-category (i.e. a branching ratio) follows a binomial distribution.



A discrete Distribution : Poisson

Consider the binomial distribution for k, in the following limit

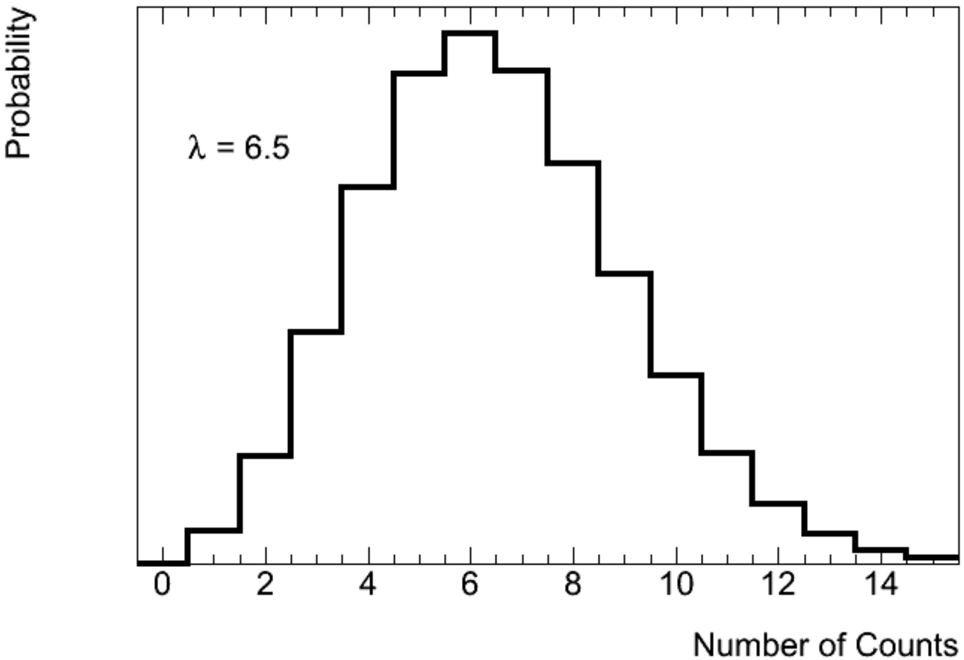
$$n \rightarrow \infty \quad , \quad p \rightarrow 0 \quad , \quad E[k] = np \rightarrow \lambda$$

The random variable k follows the Poisson distribution,

$$P_{Poisson}(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

(k is the random variable, λ is the unique parameter) for which the expectation value and variance are

$$E[k] = V[k] = \lambda$$



Typical example :

the number of expected events in one category, at a fixed number of expected events (i.e. at a given luminosity)

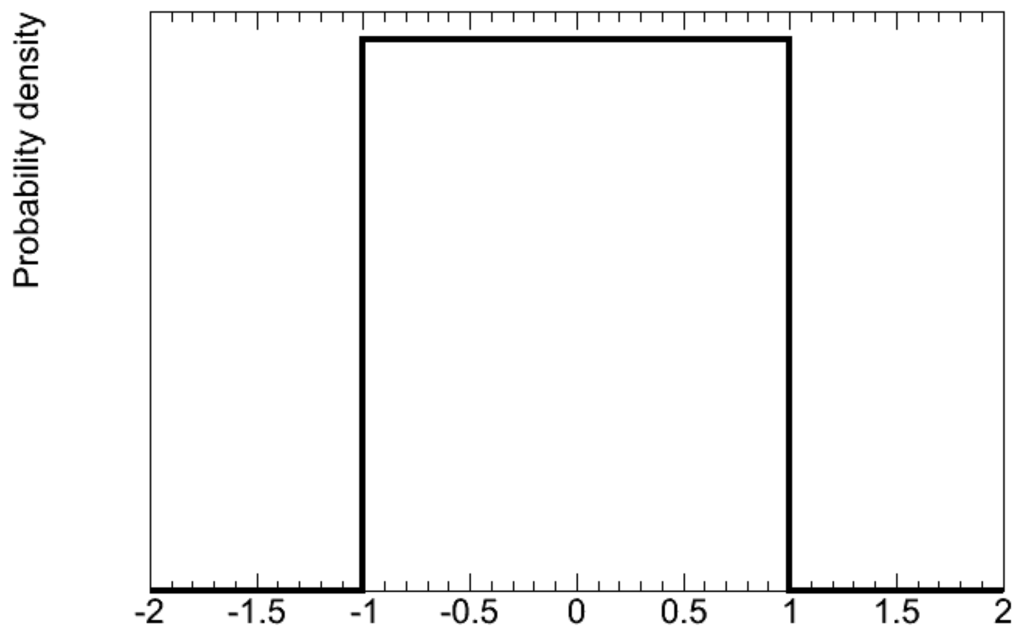
Consider a continuous random variable x, with PDF

$$P_{Uniform}(x;a,b) = \begin{cases} \dfrac{1}{b-a} & , \quad a \le x \le b \\ 0 & , \quad otherwise \end{cases}$$

for the Uniform distribution, the expectation value and variance are

$$E[x] = \frac{a+b}{2}$$

$$V[x] = \frac{(b-a)^2}{12}$$



Typical usage: accept-reject technique for MonteCarlo generation

A real-valued Distribution : Exponential

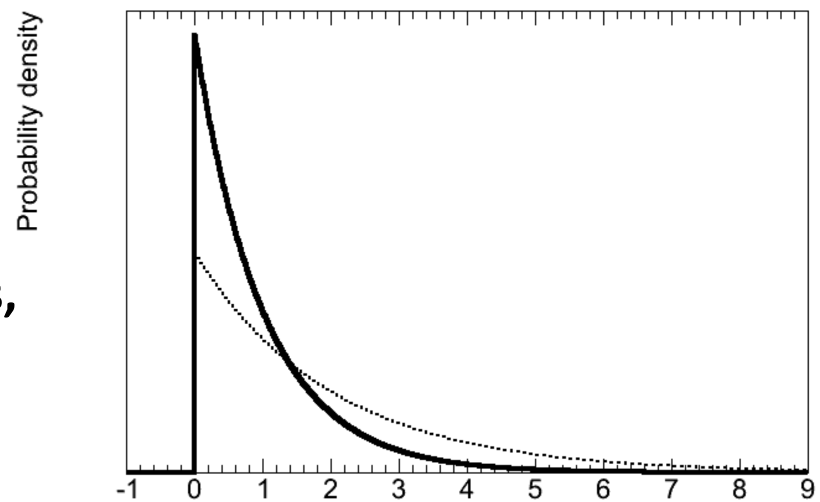
Consider a continuous random variable x, with PDF

$$P_{Exponential}(x;\xi) = \begin{cases} \frac{1}{\xi} e^{-\frac{x}{\xi}} & , \quad x \geq 0 \\ 0 & , \quad otherwise \end{cases}$$

for this exponential distribution, the expectation value and variance are

$$E[x] = \xi$$

$$V[x] = \xi^2$$



Typical examples : distribution of decay-lengths, lifetimes.

The exponential is self-similar :

$$P_{Exponential}(x - x_0 \mid x > x_0) = P_{Exponential}(x)$$

A real-valued Distribution : Gaussian

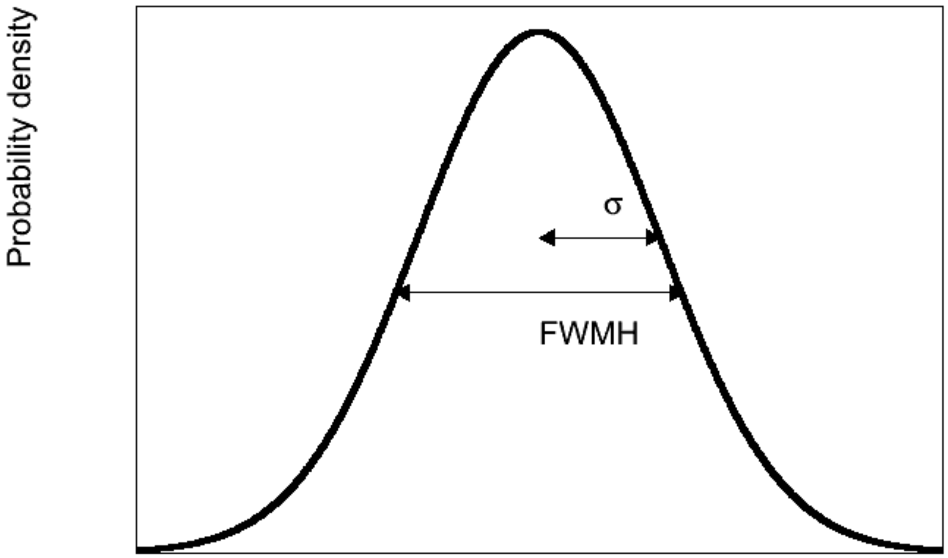
Consider a continuous random variable x, with PDF

$$P_{Gauss}(x;a,b) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

For the Gaussian (or Normal) distribution, the expectation value and variance are

$$E[x] = \mu$$

$$V[x] = \sigma^2$$



The special case $\mu = 0 \ , \sigma^2 = 1$
is often called "reduced normal".

Other parametrization often quoted:
Full Width at Half-Maximum, FWHM ~ 2.35σ

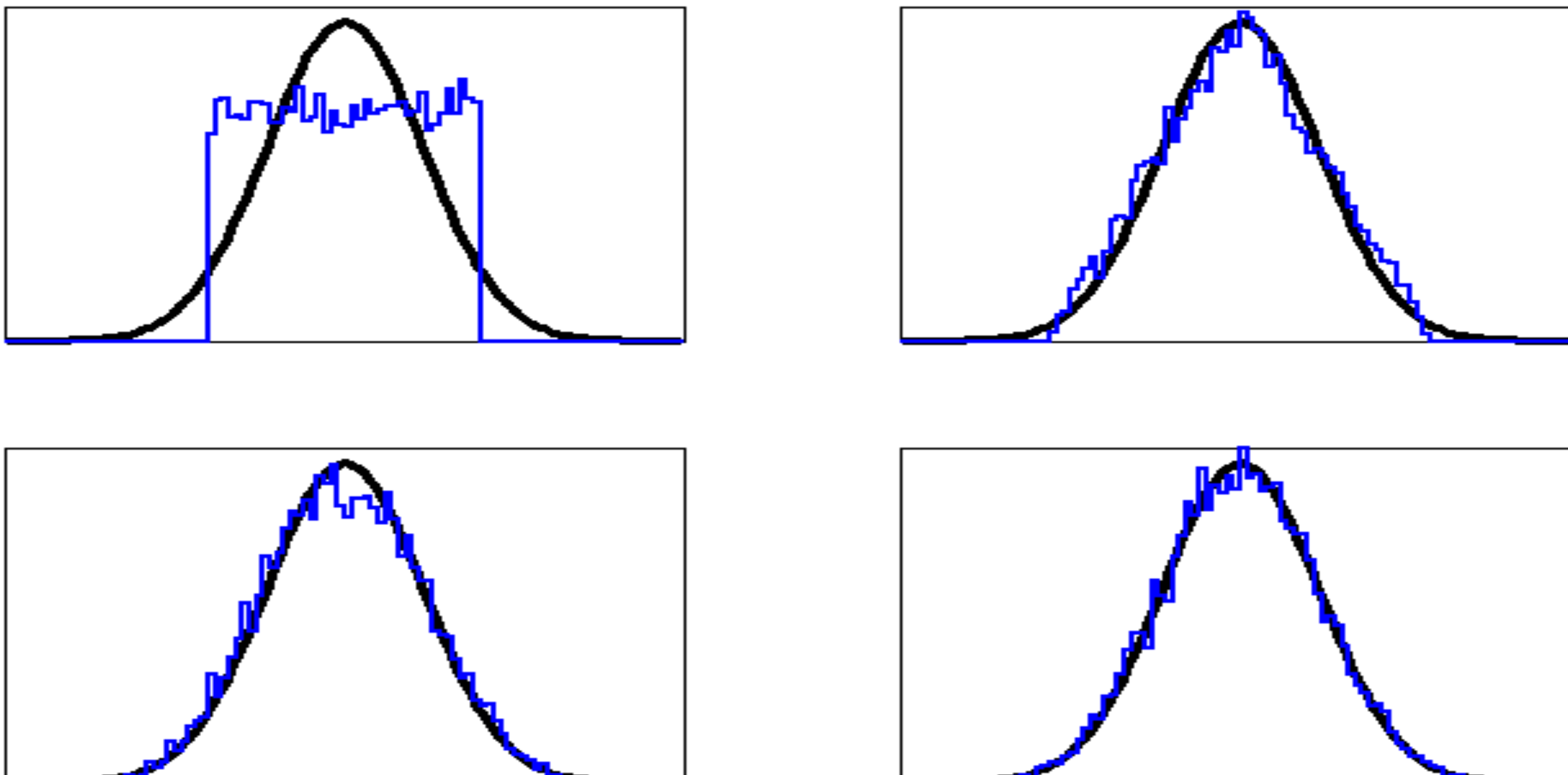
Gaussian distributions are the limit of many processes. Examples abound!

Central Limit Theorem

Consider n independent random variables $\vec{x} = \{x_1, x_2, \ldots, x_n\}$ with mean μ and variance σ^2

The sum of reduced variables $$C \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i - \mu_i}{\sigma_i}$$

converges to a reduced normal distribution, $$P(c) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{c^2}{2}}$$



A real-valued Distribution : Chi-squared (χ2)

Consider a continuous random variable x, with PDF

$$P_{\chi^2}(x;n) = \frac{x^{n/2-1}e^{-x/2}}{2^{n/2-1}\Gamma(\frac{n}{2})}$$

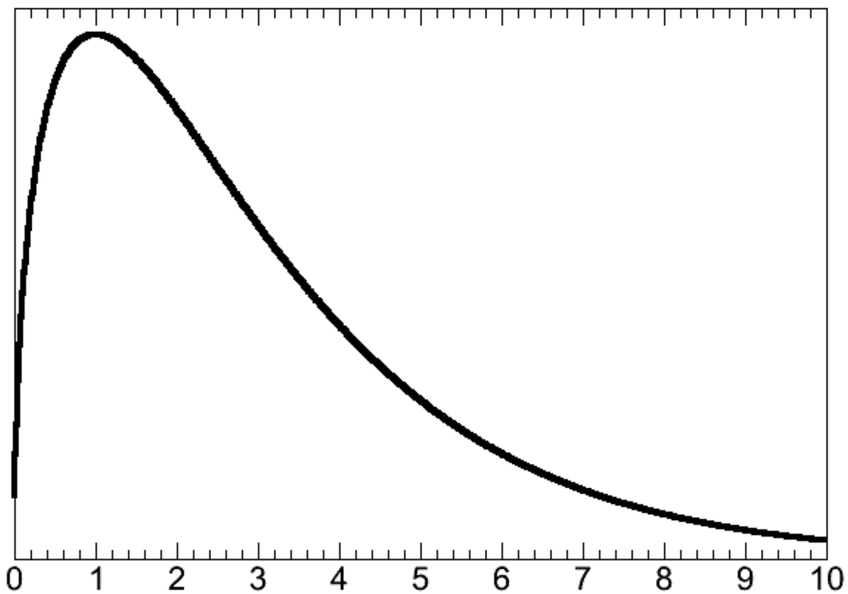
can be obtained as the sum of squares of n normal-reduced variables,

$$c = \sum_{i=1}^{n}\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2$$

the expectation value and variance are

$$E[x] = n$$

$$V[x] = 2n$$



n is called "number of degrees of freedom".
A goodness-of-fit for least-squares fits should follow a χ2 distribution.

A real-valued Distribution : Breit-Wigner

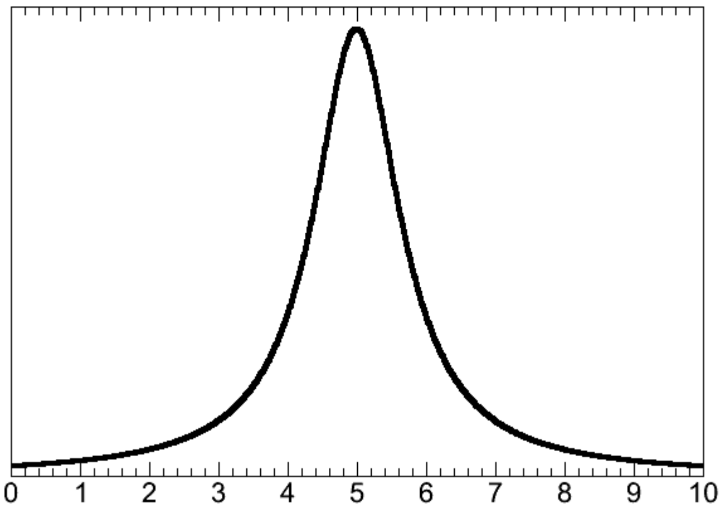
Consider a continuous random variable x, with PDF

$$P_{BW}(x;\Gamma,x_0) = \frac{1}{\pi}\frac{\left(\frac{\Gamma}{2}\right)}{(x-x_0)^2+\left(\frac{\Gamma}{2}\right)^2}$$

follows the Breit-Wigner distribution, for which neither the expectation value nor the variance are well defined. The parameters are

$$x_0 \rightarrow most\ probable\ value$$

$$\Gamma \rightarrow FWHM$$



The mass of a resonance follows a B.W. function, for which x_0 is the mass, and Γ is the decay rate