Basic concepts – part 1

SOS 2021 18-29 January, Online School

Julien Donini - UCA/LPC Clermont-Ferrand

Outline

Basics

- Sample measurements
- Error propagation
- Probabilities, Bayes Theorem
- Probability density function

Parameter estimation

- Maximum likelihood method
- Linear regression
- Least square fit

Model testings

- p-value and test statistics
- Chi2 and KS tests
- Hypothesis testing

Introductory books (non exhaustive)

Excellent book of reference

- G. Cowan, *Statistical Data Analysis* (Oxford Science Publication)

Introduction to Bayesian analysis

- D. Sivia, *Data Analysis: A Bayesian Tutorial* (Oxford Science Publication)

Classic textbook

- Louis Lyons, *Statistics for Nuclear and Particle Physicists* (Cambridge University Press)

En Français

- B. Clement, *Analyse de données en sciences expérimentales* (Dunod)

Population

- Let's consider a **sample** of values (e.g. experimental measurements)
  N measurement of a **random variable X**: $\{x_i\} = \{x_1, x_2, \ldots, x_N\}$
- There are several quantities that can be determined to **characterize this population** without any knowledge of the underlying model/theory

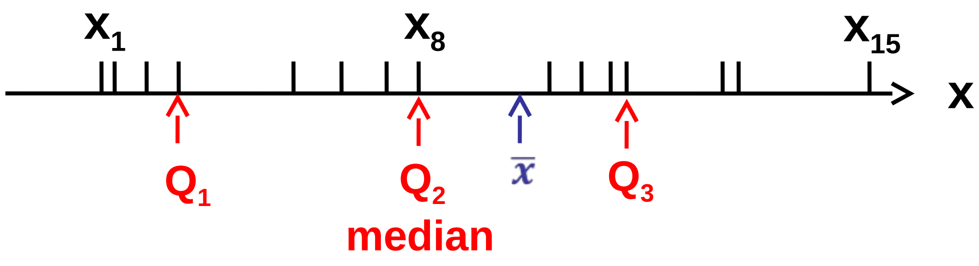
Measure of position

Arithmetic mean:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Median: value that separates sample in half

Quartiles (Q_1,Q_2,Q_3): values that separates sample in four equal-size sample



Measure of dispersion

Variance: if truth sample **mean** μ is known $$v = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{\mu})^2$$

But μ is in general not know and sample mean is used instead

- **Sample variance** (biased): $$v = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2$$

- **Estimated variance** (unbiased): $$v = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{N}{N-1}(\overline{x^2} - \bar{x}^2)$$

→ Bias is below α if N ≥ 1/α – 1 (ex for 1% bias, N≥101)

Standard deviation (is of same unit as x): $\sigma = \sqrt{v}$

Standard deviation and error

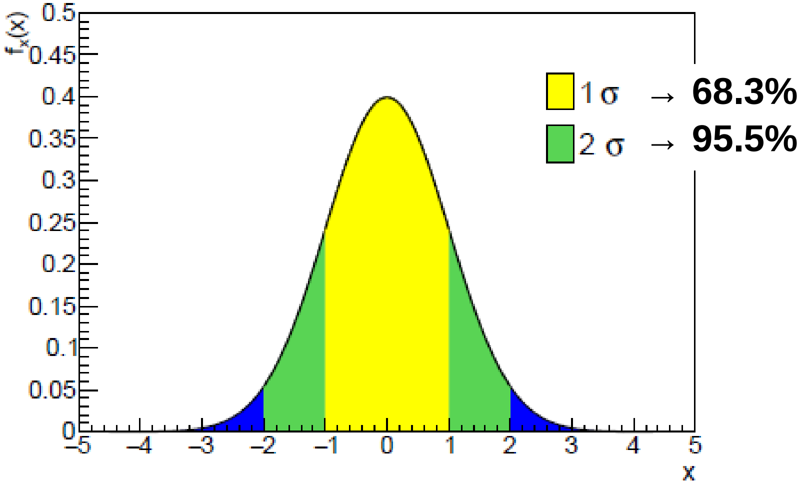
In many situations **repeating an experiment** a large amount of time produces a spread of results whose distribution is approximately **Gaussian**.

This is a consequence of the **Central Limit Theorem**.

Gaussian (a.k.a normal) distribution

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

Interval μ±σ contains 68.3% of distribution



A **measurement** = outcome of the **sum** of a large number of **effects**.

In general the distribution of this variable will be gaussian.

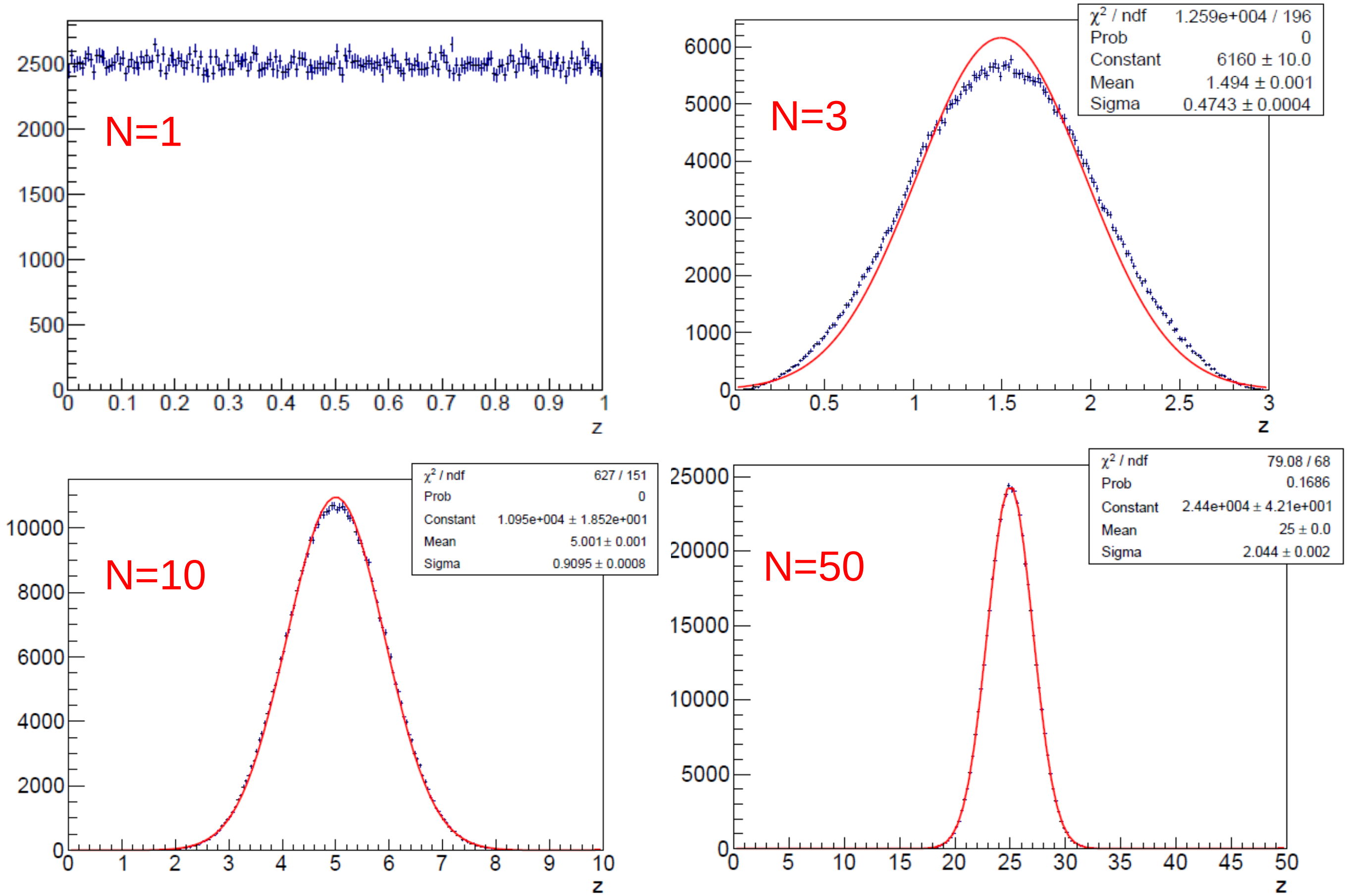
The **standard deviation** of the sample is associated to the standard deviation of the normal distribution.

The standard deviation is then interpreted as an **interval** that could contain the true value with a **68.3% confidence level**.

CLT at work

Simple illustration of CLT

- let's consider x: a random variable uniformly distributed in [0,1]
- and the distribution of the sum of N values x: $z = \sum_{i=1}^{N} x_i$



Uniform (N=1)

↓

Irwin-Hall

(see here)

↓

Gauss

(N>40)

Multidimensional samples

Case where **N measurements** are performed of **M different variables**

→ The sample then consists of **N** vectors of **M** measurements

$$\{\vec{x_i}\} = \{\vec{x_1}, \vec{x_2}, \ldots, \vec{x_N}\} \quad \text{with} \quad \begin{cases} \vec{x_1}: x_1^{(1)}, x_1^{(2)}, \ldots\ x_1^{(M)} \\ (\ldots) \\ \vec{x_N}: x_N^{(1)}, x_N^{(2)}, \ldots\ x_N^{(M)} \end{cases}$$

Mean and variance can be calculated for each variable $x_i^{(k)}$ but to quantify how of one variable behaves w.r.t another one uses the **covariance:**

For two variables x and y: $$\text{cov}(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \overline{xy} - \bar{x}\bar{y}$$

Correlation factor is defined as: $$\rho_{xy} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} \quad \text{with} \quad -1 \le \rho_{xy} \le 1$$

$\rho_{xy} = 1(-1)$ → x and y are fully (anti)correlated

$\rho_{xy} = 0$ → x and y are uncorrelated (≠ independent !)

Covariance matrix (aka error matrix) of sample $\{\vec{\boldsymbol{x}_i}\}, \mathrm{i} = 1..\mathrm{N}$

- Real, symmetric, N×N matrix of the form:

$$C = \begin{pmatrix} \mathrm{cov}(x_1, x_1) & \cdots & \mathrm{cov}(x_1, x_N) \\ \vdots & \mathrm{cov}(x_i, x_j) & \vdots \\ \mathrm{cov}(x_N, x_1) & \dots & \mathrm{cov}(x_N, x_N) \end{pmatrix} = \begin{pmatrix} {\sigma_1}^2 & \cdots & \rho_{1N}\sigma_1\sigma_N \\ \vdots & \rho_{ij}\sigma_i\sigma_j & \vdots \\ \rho_{N1}\sigma_N\sigma_1 & \dots & {\sigma_N}^2 \end{pmatrix}$$

Correlation matrix: $\rho = \begin{pmatrix} 1 & \cdots & \rho_{1N} \\ \vdots & 1 & \vdots \\ \rho_{N1} & \dots & 1 \end{pmatrix}$

Example of usage of covariance matrix:

- Transformation of input variables
- Error propagation
- Combination of correlated measurements
- …

Decorrelation: choose a **basis** $\{\vec{y_i}\}$ where **C** becomes **diagonal**.

→ transformation matrix **A** such that new covariance matrix **U** is diagonal

$$y_i = \sum_{j=1}^{N} A_{ij} x_j$$

$$Y = AX$$

$$U_{ij} = \mathrm{cov}(y_i, y_j) = \mathrm{cov}\left(\sum_{k=1}^{N} A_{ik} x_k, \sum_{l=1}^{N} A_{jl} x_l\right)$$

$$= \sum_{k,l=1}^{N} A_{ik} A_{jl} \mathrm{cov}(x_l, x_k) = \sum_{k,l=1}^{N} A_{ik} C_{kl} {A_{lj}}^T$$

$$U = ACA^T$$

(A is orthogonal $A^{-1}=A^T$)

Diagonalization of C: find orthonormal eigenvectors e_j such that $Ce_j = \lambda_j e_j$

$$A^T = \begin{pmatrix} {e_1}^{(1)} & {e_1}^{(2)} & \cdots & {e_1}^{(N)} \\ & \vdots & \vdots & \\ {e_N}^{(1)} & {e_N}^{(2)} & \cdots & {e_N}^{(N)} \end{pmatrix}$$

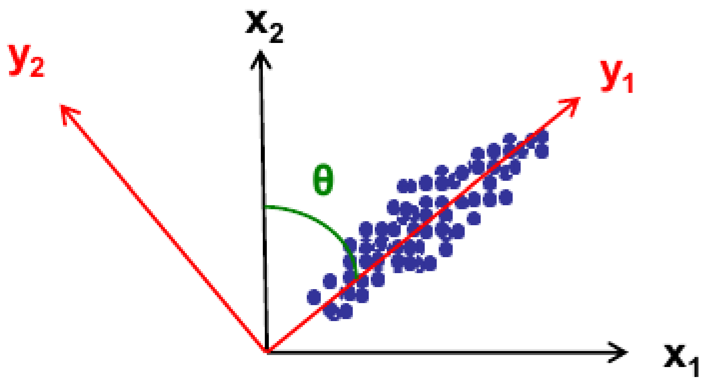
and

$$U = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{pmatrix}$$

λ_i = eigenvalues of C = σ'^2_j = variance of y_i

2D example: variables x_1 and x_2 with correlation factor ρ

$$\lambda_\pm = \frac{1}{2}\left(\sigma_1{}^2 + \sigma_2{}^2 \pm \sqrt{(\sigma_1{}^2 + \sigma_2{}^2)^2 - 4(1-\rho^2)\sigma_1{}^2\sigma_2{}^2}\right)$$



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

$$\theta = \frac{1}{2}\tan^{-1}\left(\frac{2\rho\sigma_1\sigma_2}{\sigma_1{}^2 - \sigma_2{}^2}\right)$$

Decorrelation: use cases

- Data pre-processing (for ML): remove correlation from input variables
- Reduce dimensionality of a problem: Principal Component Analysis (PCA)

Consider only the M<N dominant eigenvalues (=variance) terms in U
→ Reduced covariance matrix C: M×M

Note: the decorrelation method is able to eliminate only **linear** correlations

Function ***f*** of several variables **x**={x_1,…,x_N}

- Each variable **x_i** of mean **μ_i** and variance **σ_i^2**
- Perform **1st order Taylor expansion** of *f* around mean value

$$f(\vec{x}) \approx f(\vec{\mu}) + \sum_{i=1}^{N} \frac{\partial f}{\partial x_i}(\vec{\mu})(x_i - \mu_i)$$

$$f(\vec{x})^2 \approx f(\vec{\mu})^2 + 2f(\vec{\mu}) \sum_{i=1}^{N} \frac{\partial f}{\partial x_i}(\vec{\mu})(x_i - \mu_i) + \sum_{i,j=1}^{N} \frac{\partial f}{\partial x_i}\frac{\partial f}{\partial x_j}(\vec{\mu})(x_i - \mu_i)(x_j - \mu_j)$$

Variance of ***f(x)****:*

$$\sigma_f^2 = \overline{f(\vec{x})^2} - \left(\overline{f(\vec{x})}\right)^2 \approx \sum_{i,j=1}^{N} \frac{\partial f}{\partial x_i}\frac{\partial f}{\partial x_j}(\vec{\mu}) \times \mathrm{cov}(x_i, x_j)$$

Since $\overline{(x_i - \mu_i)} = 0$

$$\overline{(x_i - \mu_i)^2} = \sigma_i^2$$

$$\overline{(x_i - \mu_i)(x_j - \mu_j)} = \mathrm{cov}(x_i, x_j)$$

Validity: up to 2nd order, linear case, small errors

2D Example:

x and y with correlation factor ρ

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x}\sigma_x\right)^2 + \left(\frac{\partial f}{\partial y}\sigma_y\right)^2 + 2\frac{\partial f}{\partial x}\frac{\partial f}{\partial y}\mathrm{cov}(x,y)$$

$$f(x,y) = x + y \rightarrow \sigma_f^2 = \sigma_x^2 + \sigma_y^2 + 2\rho\sigma_x\sigma_y$$

$$f(x,y) = xy \rightarrow \sigma_f^2 = y\sigma_x^2 + x\sigma_y^2 + 2xy\rho\sigma_x\sigma_y$$

For a set of m function $f_1(\vec{x}), \ldots, f_m(\vec{x})$

- **C** is the covariance of variables **x**={x_i}
- We can build the covariance matrix of {$\mathbf{f_i(x)}$}: **U**

$$U_{kl} = \mathrm{cov}(f_k, f_l) = \sum_{i,j=1}^{N} \frac{\partial f_k}{\partial x_i}\frac{\partial f_l}{\partial x_j}(\vec{\mu}) \times \mathrm{cov}(x_i, x_j)$$

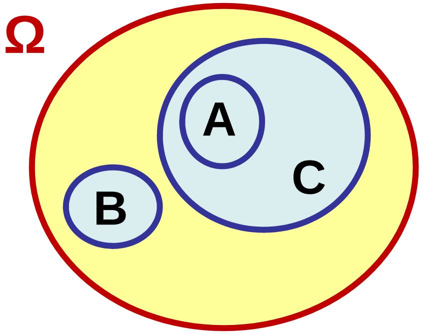
This can be expressed as $U = ACA^T$ where $A_{ij} = \frac{\partial f_i}{\partial x_j}(\vec{\mu})$ (Jacobian matrix)

Interlude

You are given a coin, you toss it and obtain “tail”.
What is the probability that both sides are “tail” ?

It depends on the **prior** that the coin is **unfair** (and on the person that gave you the coin)

Who is more likely to give a fair coin ?



Sample space: Ω

- Set of all possible results of an experiment
- Populated by events

Probability

- Frequentist: related to frequency of occurrence

$$P(A) = \frac{\text{number of time event A occurs}}{\text{number of time experience is repeated}}$$

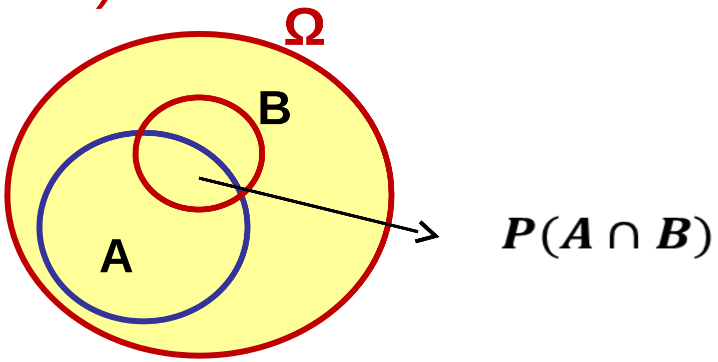
- Subjectivist (Bayesian): degree of belief that A is true
  Introduces concepts of prior and posterior probability

$$P(A|\text{data}) \propto P(\text{data}|A) \times P(A)$$

Knowledge on A increases using data

Mathematical formalization (Kolmogorov)

$$P(\Omega) = 1$$
$$0 \leq P(A) \leq 1$$
$$P(A \cap B) = P(A) + P(B) - P(A \cup B)$$



Incompatible events: $P(A \cap B) = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$

Conditional probability: $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$

Independent events: $P(A \cap B) = P(A|B)P(B) = P(A)P(B)$

Bayes theorem

Thomas Bayes (?)
c. 1701 –1761

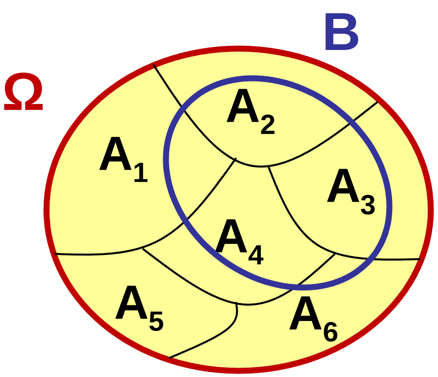
An Essay towards solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, *communicated by Mr.* Price (1763)

"If there be two subsequent events, the probability of the second b/N and the probability of both together P/N, and it being first discovered that the second event has also happened, from hence I guess that the first event has also happened, the probability I am right is P/b."

http://www.stat.ucla.edu/history/essay.pdf

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

If the sample space Ω can be divided in disjoint subsets A_i



$$A_i \cap A_j = \emptyset \ (i \neq j)$$

$$P(B) = \sum_i P(B \cap A_i) = \sum_i P(B|A_i)P(A_i)$$

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

Bayes Theorem in everyday life

Example: 10 coins, **one** of which is **unfair** (two-sided tail): You flip a random coin and obtain **tail**. What is the probability that this is the unfair coin ?

A: event where the coin is **unfair**, **B**: event where the result is **tail**

You want **P(A|B)**: $$\boldsymbol{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$

where: $$P(B) = P(B \cap A) + P(B \cap \bar{A}) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$$

$$P(B|A) = 1, P(A) = \frac{1}{10}$$

$$\boldsymbol{\Rightarrow P(A|B) = \frac{1 \times \frac{1}{10}}{1 \times \frac{1}{10} + \frac{1}{2} \times \frac{9}{10}} = \frac{2}{11}}$$

In **Bayesian** language: P(A) is the **prior** probability and P(A|B) the **posterior**

Consequences of not knowing Bayes Th.

Simple tools for understanding risks: from innumeracy to insight (2003)

G. Gigerenzer, A. Edwards, BMJ 327, 2003 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC200816/

Conditional probabilities

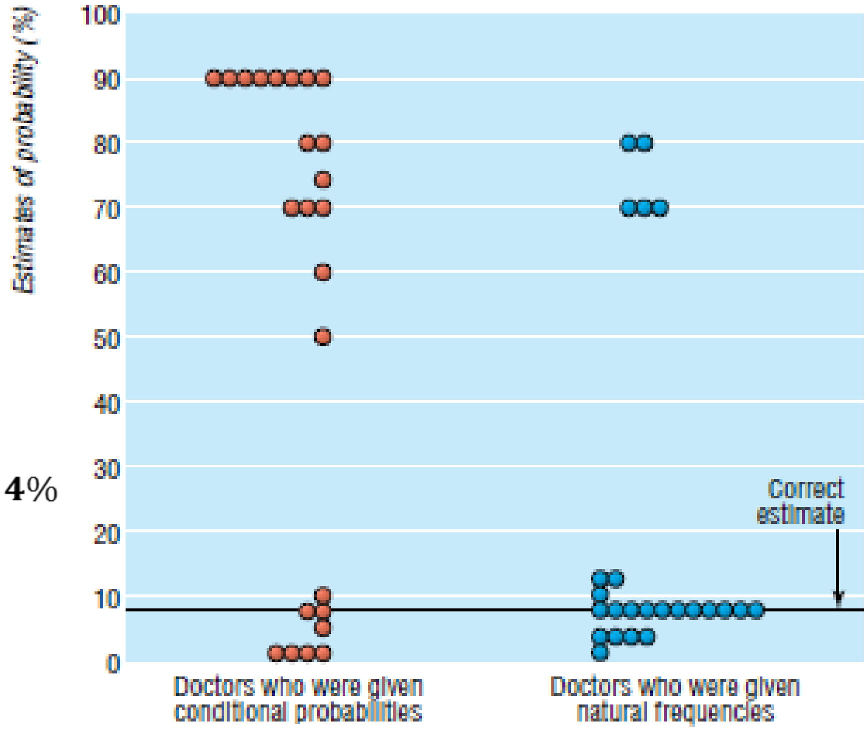
The probability that a woman has breast cancer is 0.8%. If she has breast cancer, the probability that a mammogram will show a positive result is 90%. If a woman does not have breast cancer the probability of a positive result is 7%. Take, for example, **a woman who has a positive result. What is the probability that she actually has breast cancer?**

$$P(C|+) = \frac{P(+|C)P(C)}{P(+)} = \frac{0.9 \times 0.008}{0.9 \times 0.008 + 0.07 \times 0.992} = 9.4\%$$

Natural frequencies

Eight out of every 1000 women have breast cancer. Of these eight women with breast cancer seven will have a positive result on mammography. Of the 992 women who do not have breast cancer some 70 will still have a positive mammogram. Take, for example, a sample of women who have positive mammograms. **How many of these women actually have breast cancer?**

$$\mathbf{P(C|+)} \simeq \frac{\mathbf{7}}{\mathbf{77}} = \mathbf{9.1\%}$$



"Bad presentation of medical statistics such as the risks associated with a particular intervention can lead to patients making poor decisions on treatment"

Bayes Theorem and statistical inference

Statistical inference

Estimate true parameters of a theory or a model using data

- Frequentist: perform measurement (or set limits)
- Bayesian: Improve prior knowledge using data

Going Bayesian

Likelihood of observing these data given a theory

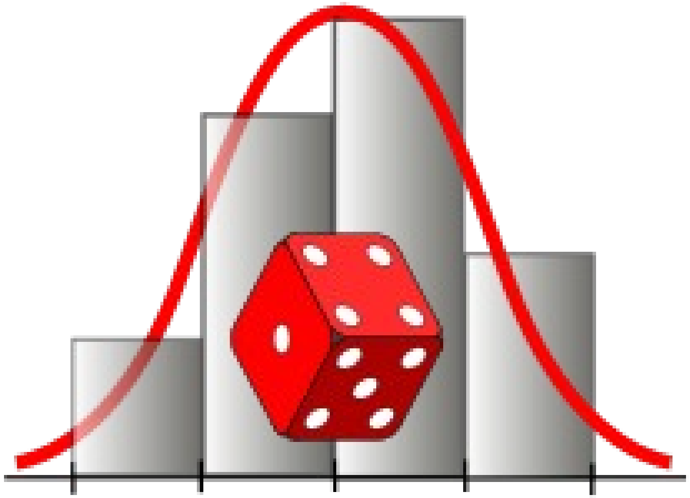
Posterior knowledge on theory

Prior knowledge on theory

$$P(\text{theory}|\text{data}) = \frac{P(\text{data}|\text{theory})\,P(\text{theory})}{P(\text{data})}$$

Usually just a **normalisation** factor

Probability distribution



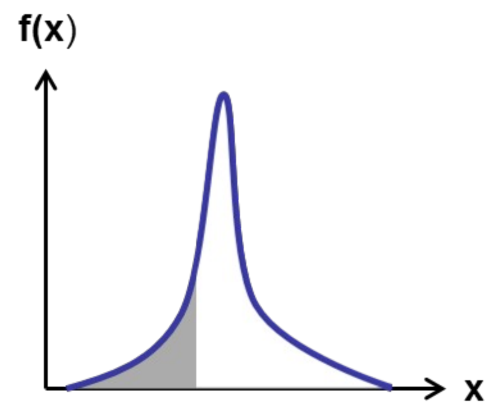
Random variable X

Discrete random variable: result (realizations) $x_i \in \Omega$ with probability $P(x_i)$

→ **P** is the **probability distribution** and $\sum_i^N P(x_i) = 1$

For continuous variable: probability of observing x in infinitesimal interval

→ Given by the **probability density function** (p.d.f) f(x)



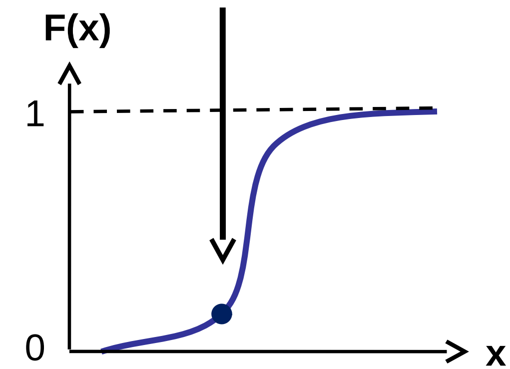
Probability of x in $[x, x+dx] = f(x)dx$

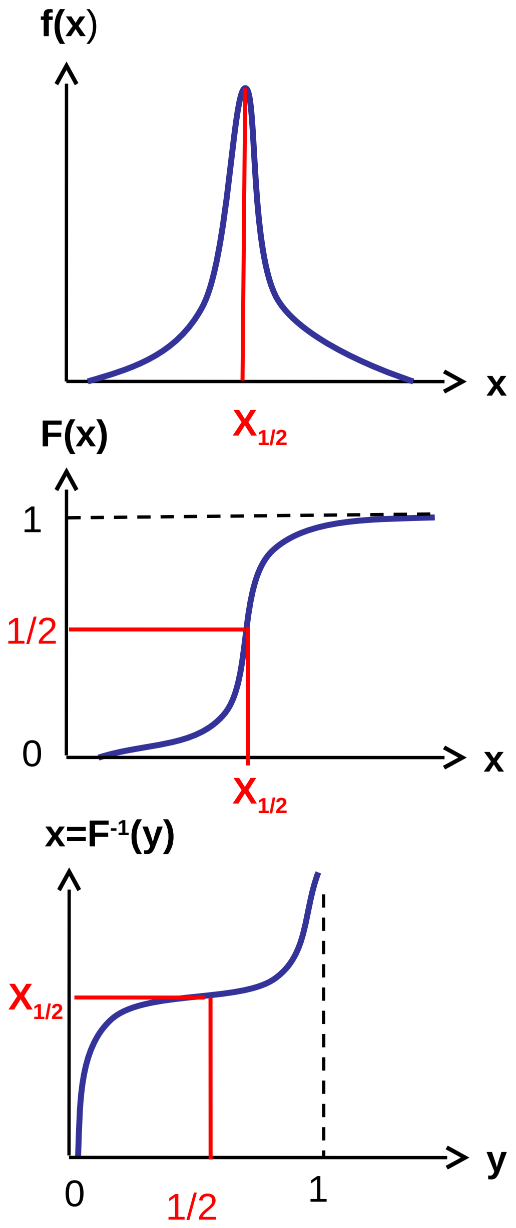
Probability of x in $[a,b] = \int_a^b f(x)dx$

with: $\int_\Omega f(x)dx = 1$

→ **Cumulative distribution** F(x): $F(x) = \int_{-\infty}^{x} f(x')dx'$

hence: $f(x) = \frac{dF}{dx}(x)$





Probability density function: $f(x)$

Cumulative distribution: $F(x)=y$

Inverse cumulative distribution: $x=F^{-1}(y)$

Median: x such that $F(x)=1/2 \rightarrow x_{1/2} = F^{-1}(1/2)$

Quantile of order α: $x_\alpha = F^{-1}(\alpha)$

- Ex: quartile, percentile, ...

Expectation value of a random variable X:

For a **function** of x, **a(x)**, the expectation value is: $E[a(x)] = \int_{-\infty}^{\infty} a(x) f(x) dx$

- **mean of X:** $$E[x] = \int_{-\infty}^{\infty} x f(x) dx = \mu$$

- **n^th^ order moment:** $E[x^n] = \int_{-\infty}^{\infty} x^n f(x) dx = \mu_n$

- **Characteristic function Φ(t):**

$$\phi(t) = E\left[e^{itx}\right] = \int e^{itx} f(x) dx = \mathrm{FT}^{-1}(f) \quad \text{where } \mu_n = (-i)^n \frac{d^n \phi}{dt^n}(0)$$

- **Variance:** $$V[x] = E\left[(x - E[x])^2\right] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx = E\left[x^2\right] - E[x]^2$$

- **Standard deviation:** $\sigma = \sqrt{V[x]}$

Some common distributions

Binomial law: efficiency, trigger rates, …

$$B(k;n,p) = C_k^n p^k (1-p)^{n-k}, \mu = np, \sigma = \sqrt{np(1-p)}$$

Poisson distribution: counting experiments, hypothesis testing

$$P(n;\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}, \mu = \lambda, \sigma = \sqrt{\lambda}$$

Gauss distribution (aka normal): many use-case (asymptotic convergence)

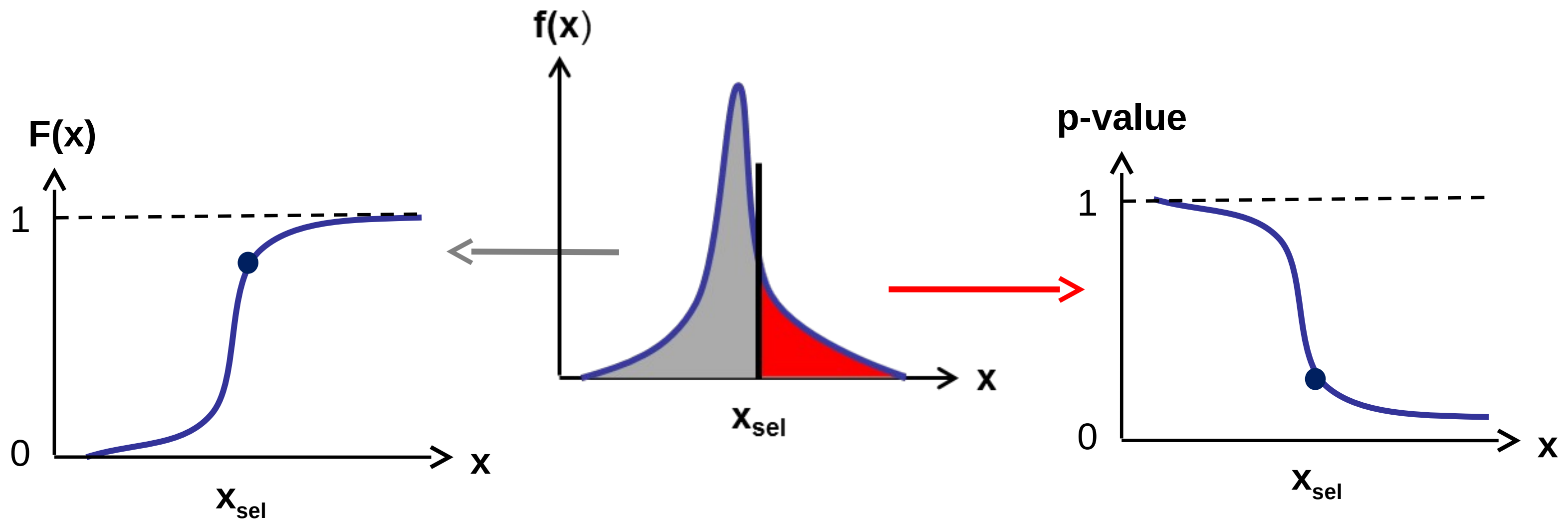
$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Cauchy distribution (aka Breit-Wigner): particle decay width, ...

$$f(x;x_0,\gamma) = \frac{1}{\pi\gamma\left[1+\left(\frac{x-x_0}{\gamma}\right)^2\right]}$$

μ and σ not defined (divergent integral)

Cumulative distribution and p-value



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

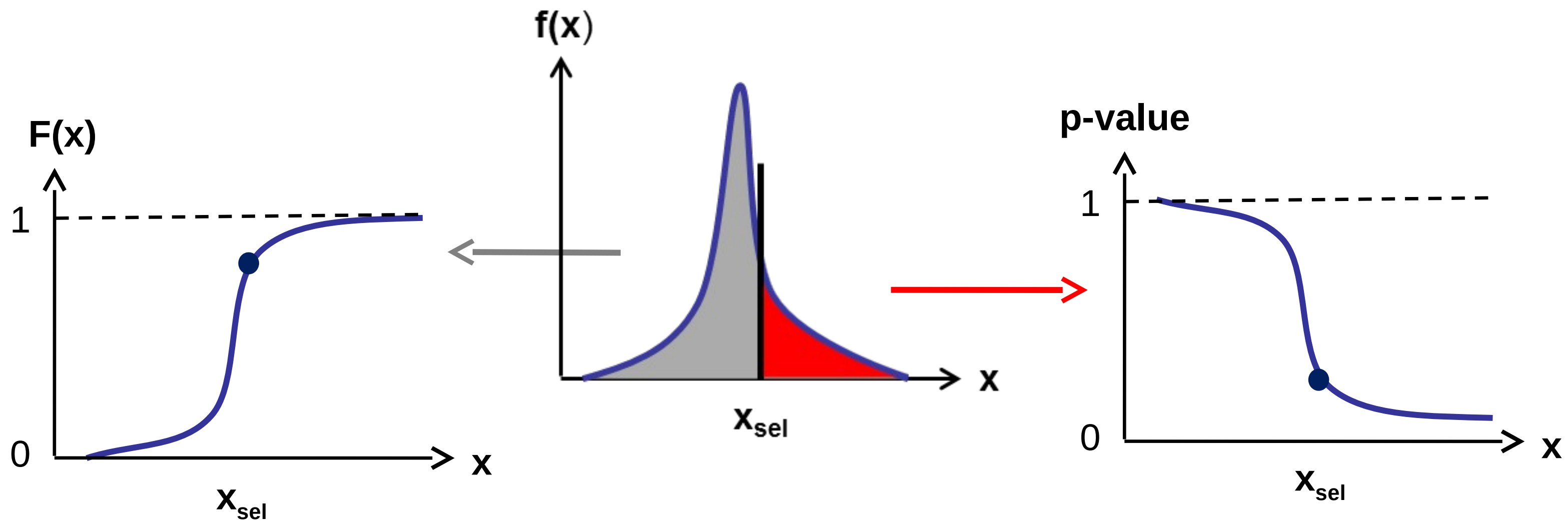
$$\text{p-value} = \int_{x}^{+\infty} f(x')dx'$$

One can choose any x_{sel} to compute F(x) or p-value, that is x_{sel} does not have a preferred value: it follows the **uniform** distribution

➔ The distributions of $F(x_{sel})$ and p-value are also **uniform** [**proof** next page]

➔ Important for MC sample generation and hypothesis testing

Cumulative distribution and p-value



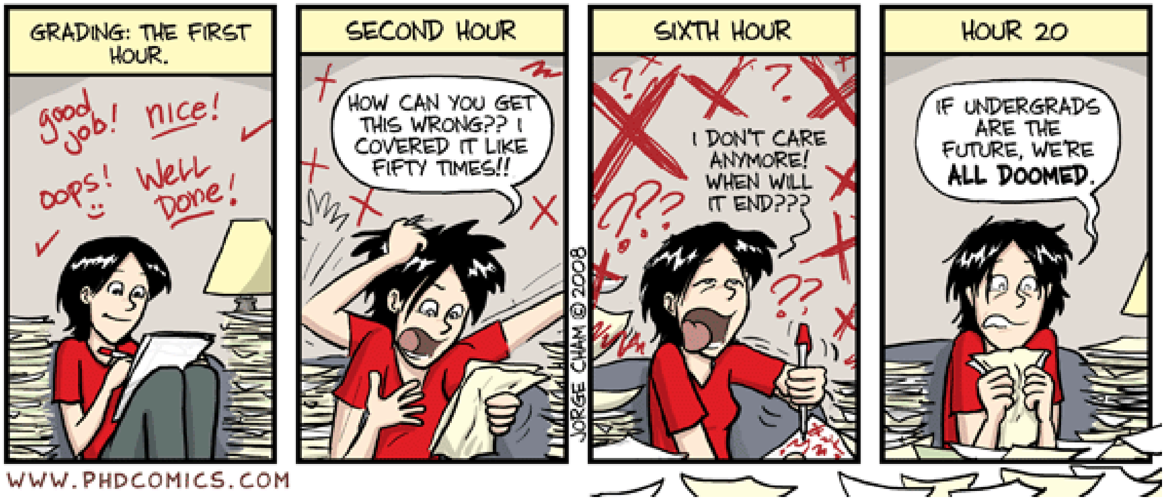
[**proof**] Given any random continuous variable X, define $Y = F_X(X)$

Then: $$\begin{aligned} F_Y(y) &= \mathrm{P}(Y \leq y) \\ &= \mathrm{P}(F_X(X) \leq y) \\ &= \mathrm{P}(X \leq F_X^{-1}(y)) \\ &= F_X(F_X^{-1}(y)) \\ &= y \end{aligned}$$

F_Y is just the cumulative distribution function of a uniform U(0,1) variable.

→ Thus, Y has a **uniform distribution** on the interval [0,1]

Grading copies:



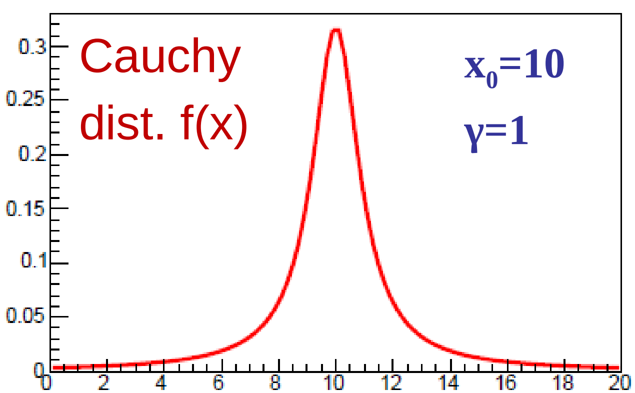
- 100 copies, grades: 0-20
- Peaked distribution at 10

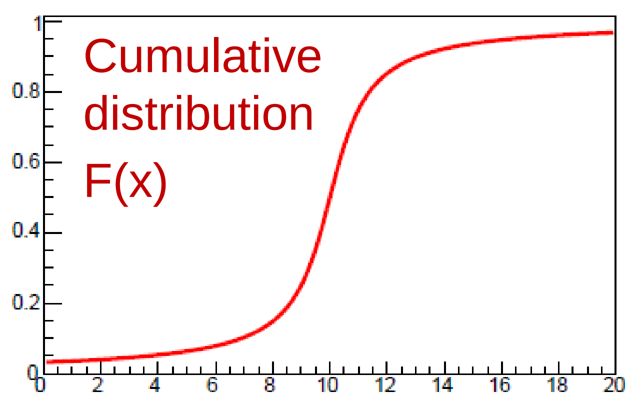
Try Cauchy distribution

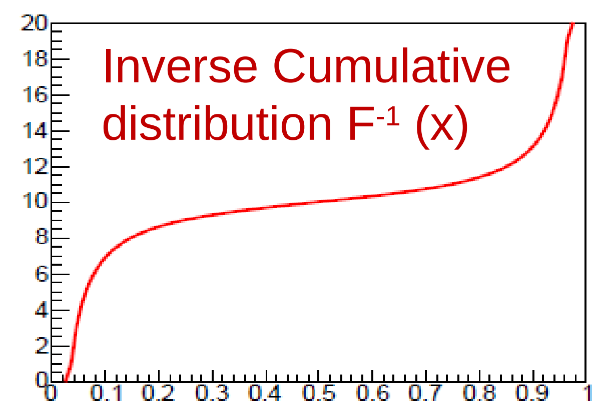
$$f(x) = \frac{1}{\pi\gamma\left[1 + \left(\frac{x - x_0}{\gamma}\right)^2\right]}$$

$$F(x) = \frac{1}{\pi}\arctan\left(\frac{x - x_0}{\gamma}\right) + \frac{1}{2}$$

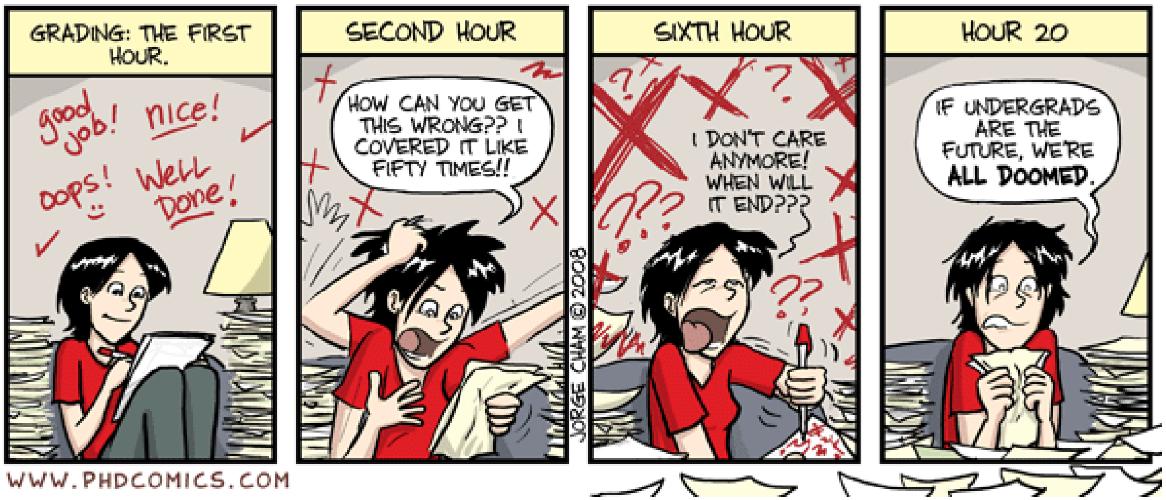
$$F^{-1}(y) = x = \gamma\tan\left(\pi\left(y - \frac{1}{2}\right)\right) + x_0$$







Grading copies:



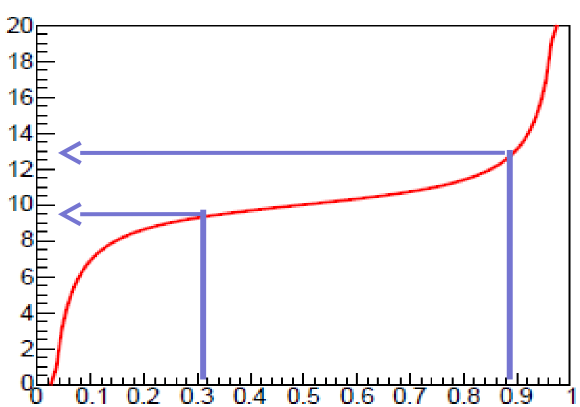
- 100 copies, grades: 0-20
- Peaked distribution at 10

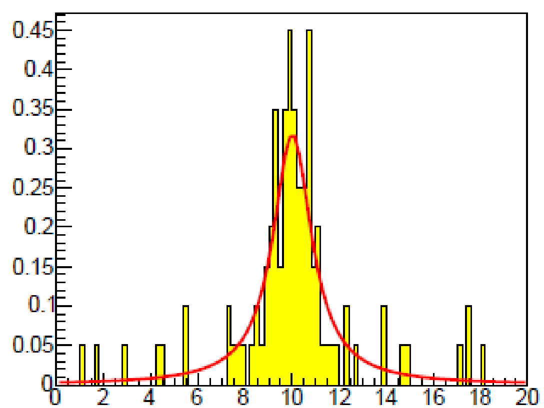
Try Cauchy distribution

$$f(x) = \frac{1}{\pi\gamma\left[1 + \left(\frac{x - x_0}{\gamma}\right)^2\right]}$$

$$F(x) = \frac{1}{\pi}\arctan\left(\frac{x - x_0}{\gamma}\right) + \frac{1}{2}$$

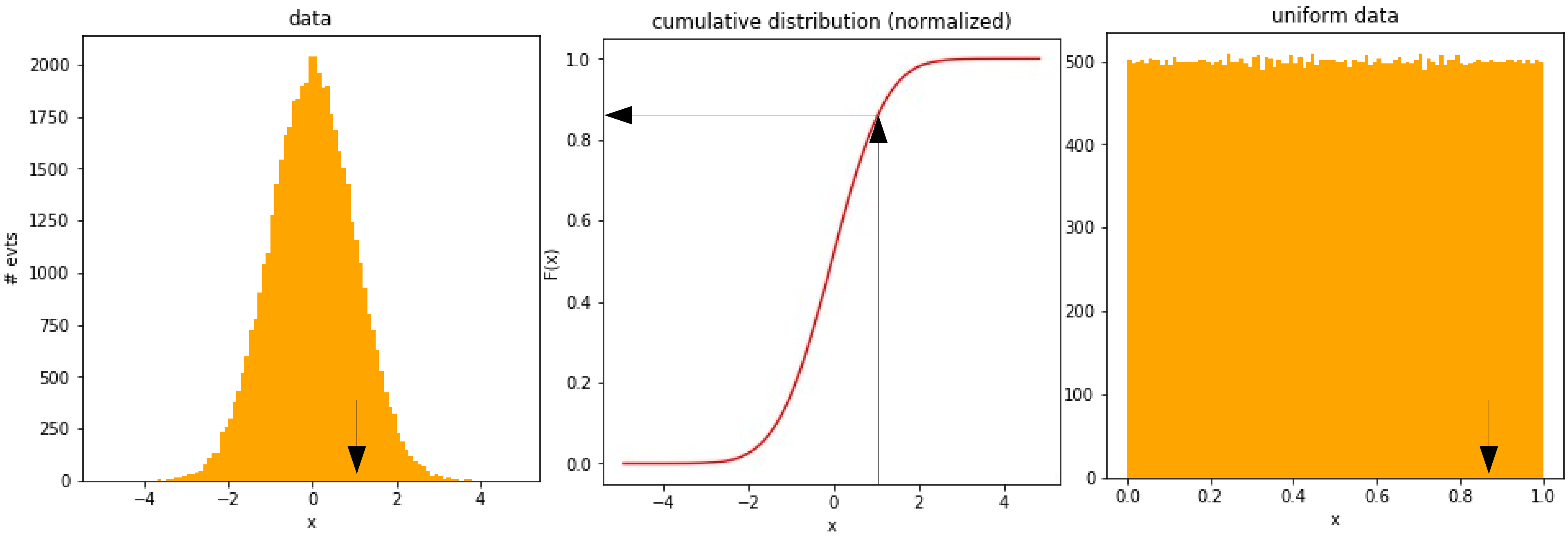
$$F^{-1}(y) = x = \gamma\tan\left(\pi\left(y - \frac{1}{2}\right)\right) + x_0$$





Data uniformization

(Inverse) cumulative distribution is naturally useful to **uniformize** data distributions



For **Machine Learning**: data preprocessing is usually the 1st step
→ Uniformization of all input variables can sometime be a good idea.

To know more about **data transformation** see for example:
https://scikit-learn.org/stable/modules/preprocessing.html

Pearson's χ^2 test: estimate global compatibility between data and a model

- The data is regrouped in an **histogram** of N bins
- A **goodness-of-fit test** K^2 is computed as follows

$$K^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i}$$

n_i: number of observed events in bin i

ν_i: expected number of events in bin i

If the data n_i are **Poisson** distributed with mean values ν_i and $\boldsymbol{n_i} > \sim\mathbf{5}$ then:
K^2 is a random variable following a **χ^2 distribution** with **N** degrees of freedom.

A variant of this test statistics is the **Neyman's χ^2**

$$K^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{n_i}$$

Easier to code (in particular for fits)

Asymptotically equivalent to Pearson's χ^2

Follows χ^2 with **N-1** degrees of freedom

χ^2 distribution

Probability density function

k degrees of freedom, x>0

$$\chi^2(x;k) = \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)}$$

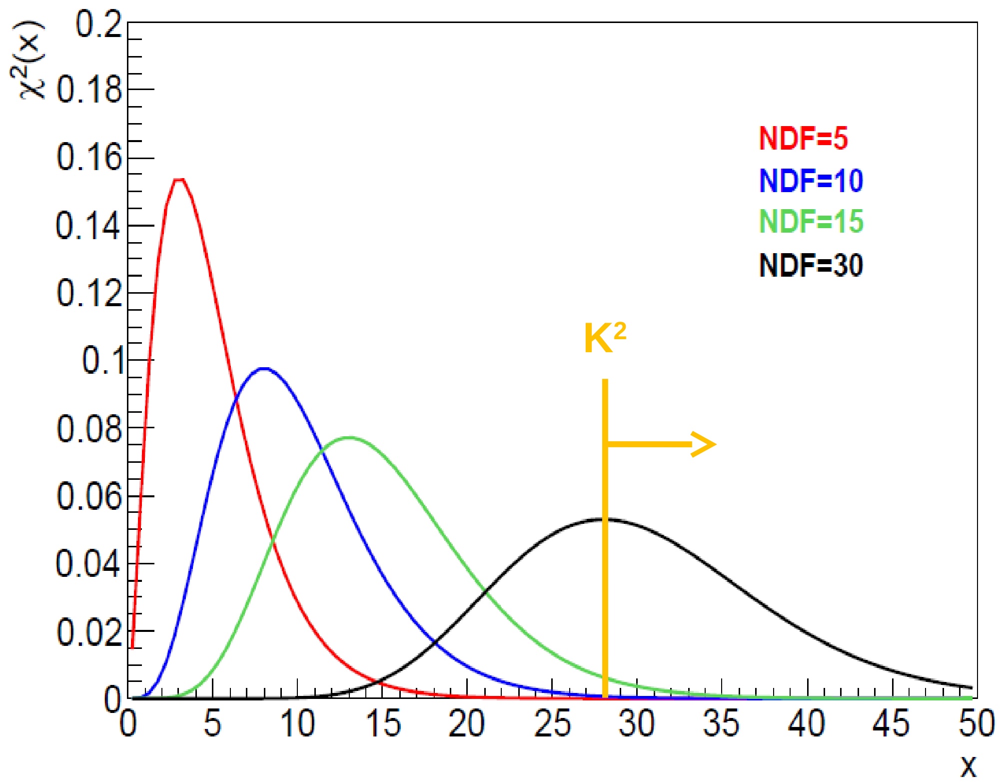
Cumulative distribution

$$F(x;k) = \frac{\gamma\left(\frac{k}{2}, \frac{x}{2}\right)}{\Gamma\left(\frac{k}{2}\right)}$$

With: $\gamma(s,x) = \int_0^x t^{s-1} e^{-t} dt$

$$\Gamma(s) = \int_0^{+\infty} t^{s-1} e^{-t} dt$$

Mean = k, variance = 2k

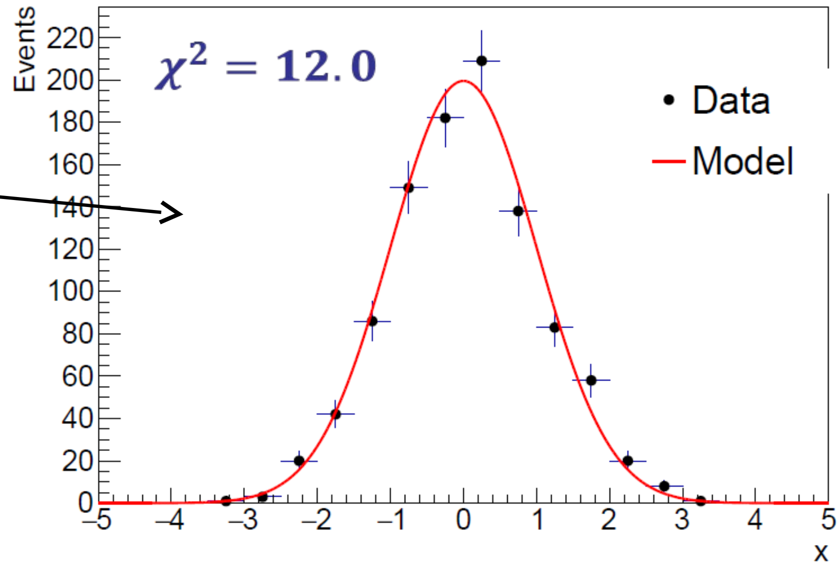


The **p-value** of a χ^2 test is obtained by integrating the χ^2 distribution **above** the measured **K^2** value.

$$\text{p-value} = \int_{K^2}^{+\infty} \chi^2(x;k)\, dx$$

Procedure

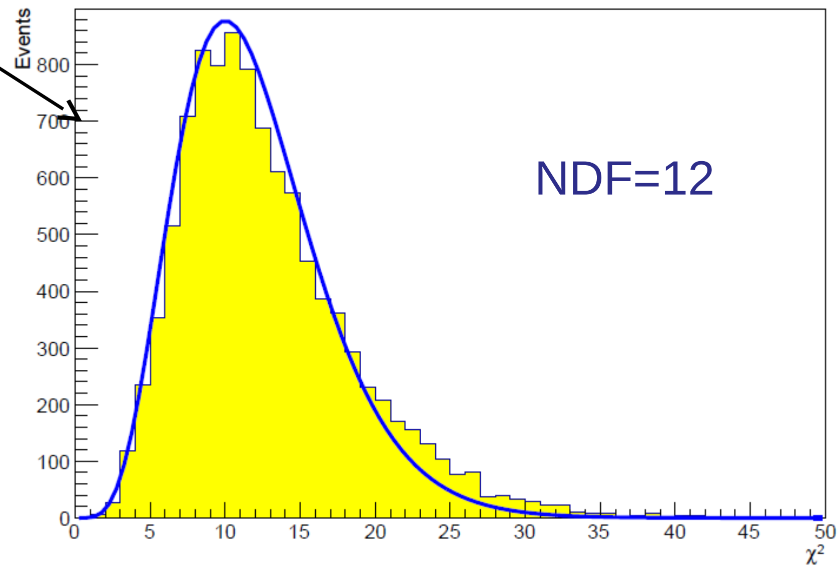
- Generate events following a Gaussian distribution
- Calculate (Neyman's) K^2
- Repeat 10k time and plot the distribution of K^2
- Compare to χ^2 distribution



Note:

K^2 is calculated only with non-empty bins

NDF is the number of non-empty bins - 1



Multi-dimensional p.d.f

An experiment can perform a set of measurement

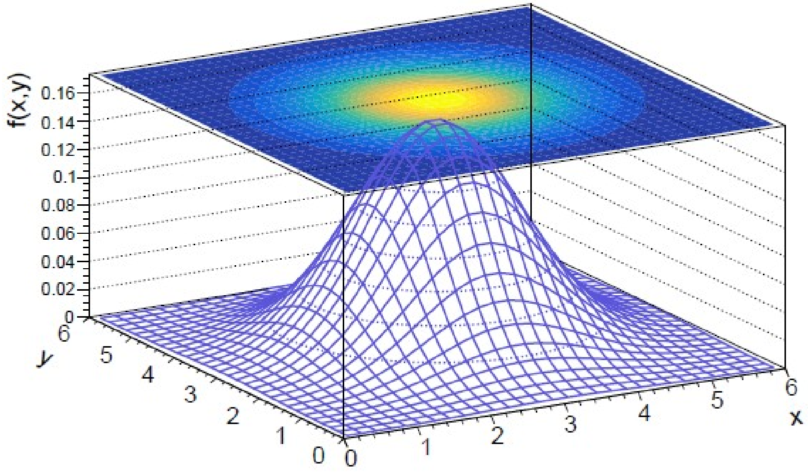
→ Vector of N measurements $\vec{x} = \{x_1, x_2, ..., x_N\}$

Probability of observing $\vec{x}$ in infinitesimal interval $\vec{x} + d\vec{x}$ given by **joint p.d.f**

$$f(\vec{x})d\vec{x} = f(x_1, \ldots, x_N)dx_1 \ldots dx_N$$

Ex: for a measurement of 2 values x and y

Probability of x in $[x, x+dx]$ and y in $[y, y+dy]$ is $f(x,y)dxdy$

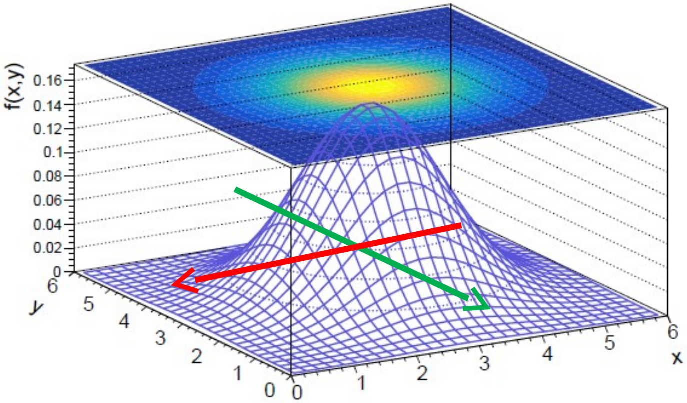


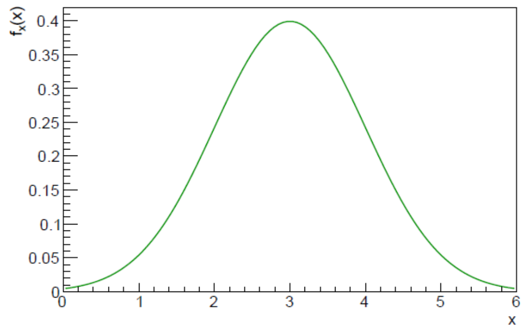
$$\iint_{\Omega} f(x,y)dxdy = 1$$

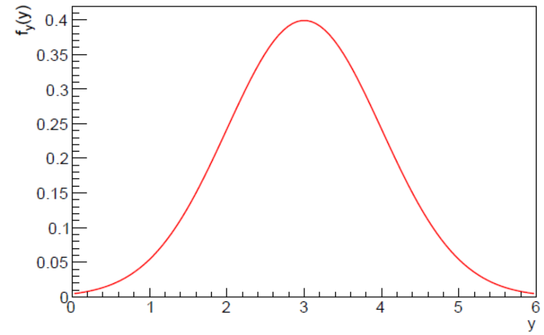
Marginal and conditional p.d.f

Marginal distribution: p.d.f of one variable regardless of the others

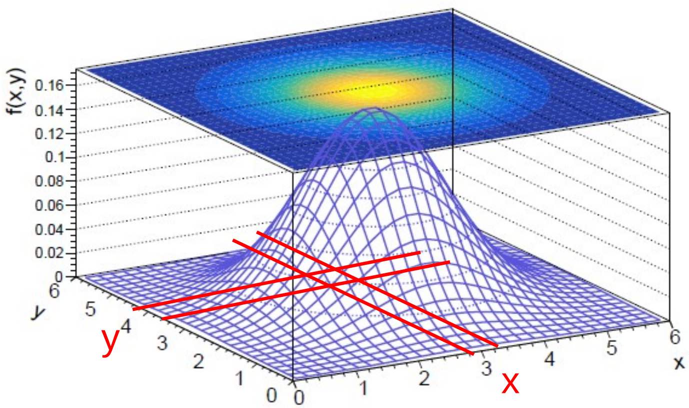
$$f_x(x) = \int_{-\infty}^{\infty} f(x,y)\,dy \qquad f_y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$







Conditional distribution: p.d.f of one variable given a constant other



$$k(y|x) = \frac{f(x,y)}{f_x(x)} = \frac{f(x,y)}{\int f(x,y')dy'}$$

$$g(x|y) = \frac{f(x,y)}{f_y(y)} = \frac{f(x,y)}{\int f(x',y)dx'}$$

Note: k and g are both functions of x and y

Marginal and conditional p.d.f

Bayes theorem for continuous variables

$$f(x,y) = g(x|y)f_y(y) = k(y|x)f_x(x) \quad \rightarrow \quad \boxed{g(x|y) = \frac{k(y|x)f_x(x)}{f_y(y)}}$$

Marginal p.d.f can also be expressed with conditional probabilities:

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

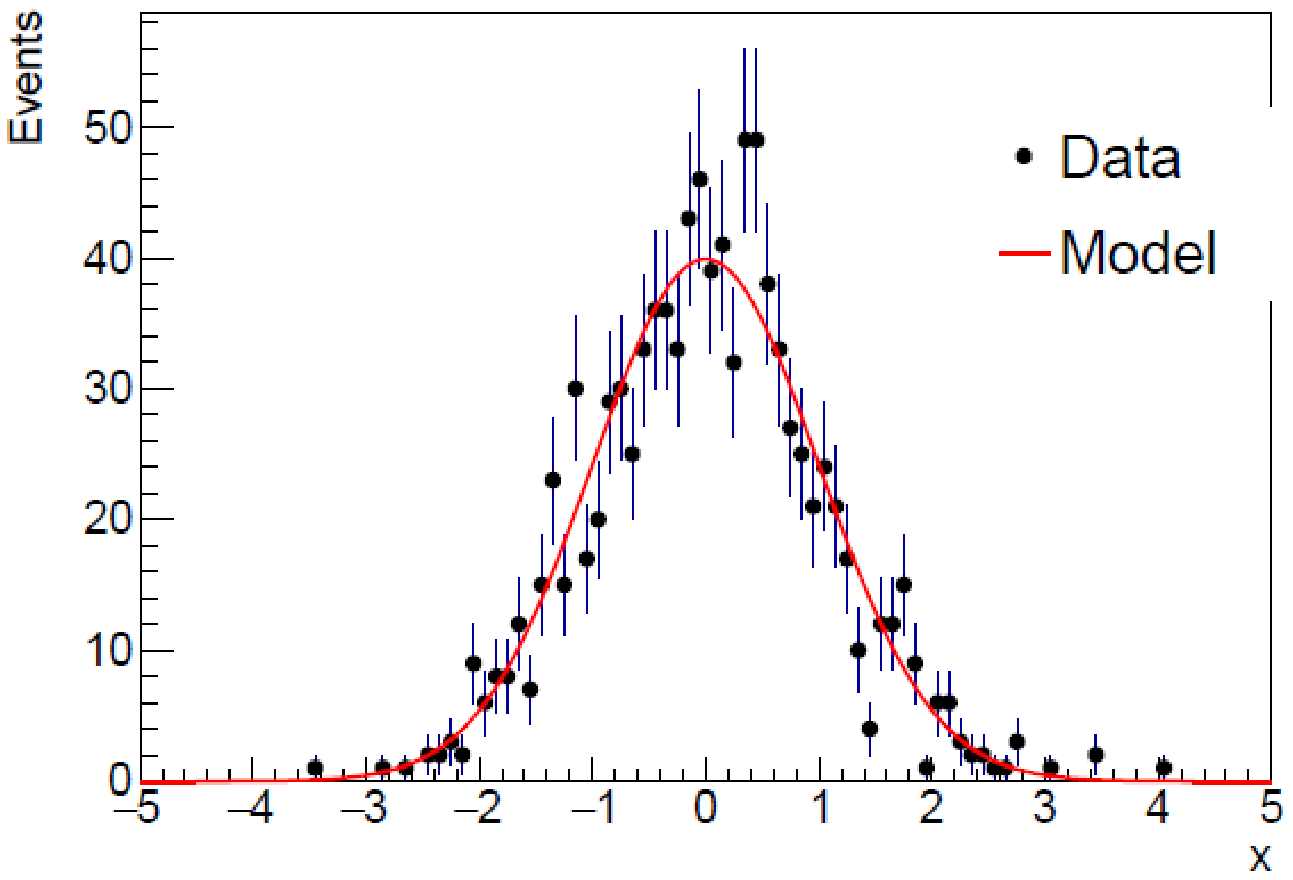
Note: this is a generalization of the relation $P(B) = \sum_i P(B|A_i)P(A_i)$ to continuous variables

Independent variables: if x and y are independent $f(x,y) = f_y(y)f_x(x)$

Ex: 2D Gaussian function with uncorrelated variables

$$\mathrm{Gaus}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left(\frac{-(x-\mu_x)^2}{2{\sigma_x}^2}\right)\exp\left(\frac{-(y-\mu_y)^2}{2{\sigma_y}^2}\right)$$

Interlude: counting experiment



What is the meaning of error bars on **observed** data ?