Università degli Studi di Milano Bicocca
Probability and statistics

Corso di Laurea triennale in Fisica

Probability and statistics

Lecture notes

Candidato:
Giorgio Cividini, g.cividini1@campus.unimib.it
Matricola 896836

Docente:
Prof. Pietro Govoni

Anno Accademico 2023 - 2024

Contents

Preface

The following pages are an elaborated and quite complete version of the lecture notes of the "Laboratorio di calcolo e statistica" course held by professor Pietro Govoni in Università degli studi di Milano Bicocca during the first semester of the 2023-2024 academic year.
These pages have been checked and revised, but I do not guarantee the absence of errors. I assume responsibility for each of them and I thank in advance anyone who will report them to me.
I thank Daniele and Giulia for the help in the revision of the work.

Giorgio Cividini

First version, january 2024

Chapter 1

Probability - An introduction

Physics is made of two realities: theory and data collection. How can these two things be related? Probability is something that can predict data from a certain theory, statistics tells us how good a set of data fits a mathematical model.
Let's focus our attention on probability.

1.1 What is probability?

Let's take a simple experiment: the tossing of a coin. We can identify:

- The set of possible outcomes is $\{H, T\}$ and it is called the *sample space* [1] identified with Ω.
- The real outcome of the experiment, repeating the tossing a certain number of times, is called *sample*; for example $\{H, H, T, T, T, H, T, H\}$.
- The ideal outcome, which is an infinite amount of outcomes and is an idealized concept; is called *population*; $\{H, ..., T, ..., H, ..., T, ...\}$.

1.2 Kolmogorov axiomatic definition

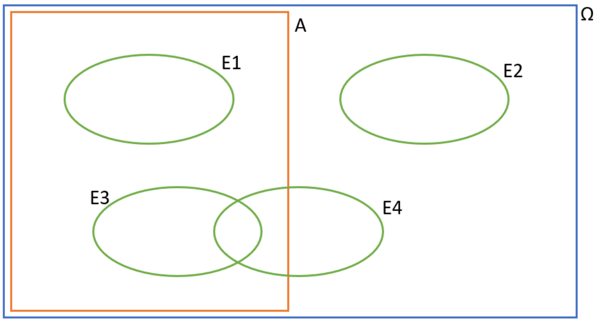
Given an event E contained in the sample space Ω, the following three conditions, called *Kolmogorov axioms*, must be respected:

$$\begin{cases} p(E) \geq 0 \\ p(\Omega) = 1 \\ p(E_1 \cup E_2) = p(E_1) + p(E_2) \text{ if } E_1 \cap E_2 = \varnothing \end{cases}$$

1.3 Conditional probability - Bayes theorem

The conditional probability of an event is the probability of an event given a preexisting condition or another event that has already happened. For example, we can discuss about the probability of the event E1 given the fact that we want to consider it in the A set.

[1] data collection taken out from the population



The conditional probability of the event E1 given A is

$$p(E1|A) = \frac{p(E \cap A)}{p(A)}$$

Let's make some math: $p(E \cap A) = p(E|A)p(A)$ and $p(A \cap E) = p(A|E)p(E)$. But $p(E \cap A) = p(A \cap E)$ so $p(E|A)p(A) = p(A|E)p(E)$ which leads to the *Bayes theorem*.

BAYES THEOREM

$$p(E|A) = \frac{p(A|E)p(E)}{p(A)}$$

The Bayes theorem is an useful tool because it can be used to calculate *inverted probabilities*:

$$p(A|E) = \frac{p(E|A)p(A)}{p(E)} = \frac{p(E|A)p(A)}{p(E|A)p(A) + p(E|A^*)p(A^*)}$$

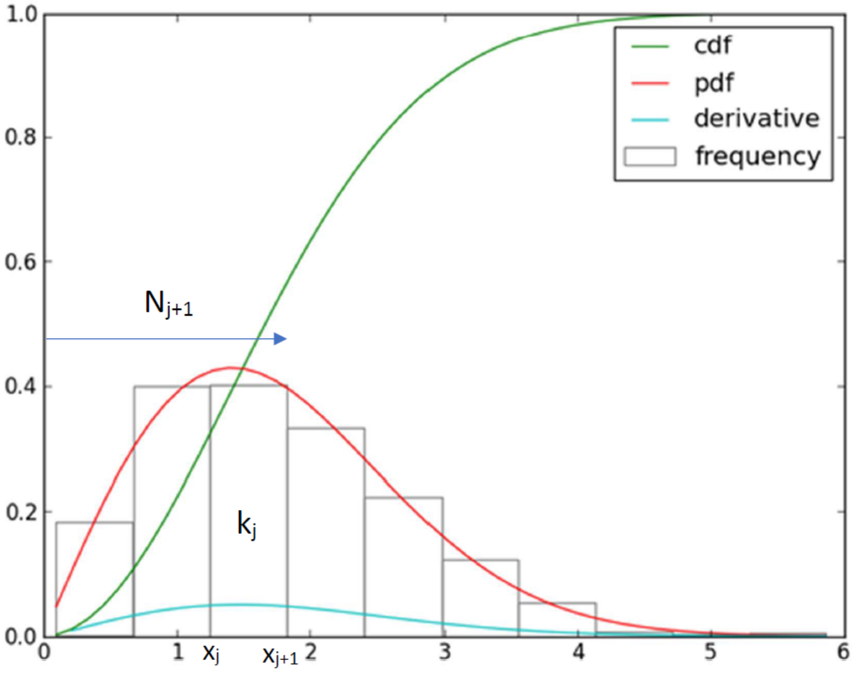
note The Bayes theorem is the fundament of *bayesian statistics* which allows to express a degree of belief in an event given the theory behind it.

1.4 Von Mises frequentist definition

Given a sample $E = \{x_i\}_{i=1,\dots,N} \quad x_i \in \Omega \forall i$, the probability of the event $E1 = \{x_j\}_{j=1,\dots,M} \subseteq E$ is

$$p(E1) = \lim_{N \to \infty} \frac{M}{N}$$

How can we represent this frequentist probability? Basically with histograms.
With discrete values, life is easy: each bin corresponds to a value. For example, the tossing of a dice: there will be six bins, each one corresponding to a value from 1 to 6.
With continuous values, bins correspond to ranges that have to be chosen. The secret is to choose the right width of the bins: not too large, not too narrow. It's also important to have a significant number of measurements.



$$p_j = \frac{k_j}{N} = \frac{N_{j+1} - N_j}{N} = F_{j+1} - F_j = \frac{F_{j+1} - F_j}{\Delta x}\,\Delta x = \frac{F(x+\Delta x) - F(x)}{dx}\,dx$$

$$p(x_j) = \lim_{\Delta x \to \infty} \frac{F(x+\Delta x) - F(x)}{dx}\,dx = F'(x)\,dx$$

Defining $f(x) = F'(x)dx$ we obtain the *probability density function* $f(x)$.

Because $f(x)$ is a density function, to obtain the probability itself the extrapolation of subtended area is needed. Because of this, the probability of a single point is zero: an interval is always needed.

$$p(x \in (x_j, x_{j+1})) = \int_{x_j}^{x_{j+1}} f(x)\,dx$$

This integer has some characteristics: its value from $-\infty$ to $+\infty$ is 1, its value in a non null interval must be ≥ 0 and it is linear for the sum (given $a < b < c$, the integer from a to c equals to the sum of the integer from a to b and the integer from b to c).

Function $F(x)$ is called *cumulative distribution function*

note Probability density functions tell us that there are no wrong measurements, but only measurements with different probabilities to manifest in a set.

Chapter 2

Distributions

As said in the previous chapter, continuous values are represented by functions called probability density functions.
Each *pdf* is different and it is possible to identify some values that characterize each distribution.
Before starting, let's recall the formula of the *expectation value*;

$$E[\mu(x)] = \int \mu(x) f(x)\, dx$$

This operator is linear.
The expectation value over a sample is the *average value* $\overline{x} = \frac{\sum x_i}{N}$, over a probability density function is the mean (or central value) $\mu = \int_{-\infty}^{+\infty} x f(x)\, dx$.

2.1 Momenta of a probability density function

We call *moment* of a probability density function the study of the expectation value elevated to an m power [1].

$E(x^m)$

- $m = 0 \qquad E(x^0) = \int x^0 f(x)\, dx = \int f(x)\, dx = 1 \longrightarrow$ this parameter checks the Kolmogorov axioms
- $m = 1 \qquad E(x^1) = \int x^1 f(x)\, dx = \int x f(x)\, dx = \mu \longrightarrow$ this gives us the mean value of the *pdf* , so it tells where the distribution is located

If we translate the function in order to have the mean value on the origin, it is possible to calculate the *central momenta* which give us other useful information about the shape of the *pdf.*

$E\left((x-\mu)^m\right)$

- $m = 1 \qquad E\left((x-\mu)^1\right) = \int (x-\mu) f(x)\, dx = \int x f(x)\, dx - \mu \int x f(x)\, dx = 0 \longrightarrow$ this checks what said before
- $m = 2 \qquad E\left((x-\mu)^2\right) = \int (x-\mu)^2 f(x)\, dx = v \longrightarrow$ this value is called *variance* and gives us information about the ***width of the distribution***

  The value $\sigma = \sqrt{v}$ is called ***standard deviation***
- $m = 3 \qquad E\left((x-\mu)^3\right) = \int (x-\mu)^3 f(x)\, dx = s \longrightarrow$ this value is called *skewness* and gives us information about the ***symmetry of the distribution around*** μ

  The value $\gamma_1 = \frac{E\left[(x-\mu)^3\right]}{\sigma^3}$ is $\begin{cases} \gamma_1 < 0 & \text{if there is asymmetry to the left} \\ \gamma_1 = 0 & \text{if there is symmetry} \\ \gamma_1 > 0 & \text{if there is asymmetry to the right} \end{cases}$

[1] the integrals in this paragraph are all calculated from $-\infty$ to $+\infty$. These extremes are not written on each integral only because of a graphical issue

- $m=4 \qquad E\left((x-\mu)^4\right) = \int (x-\mu)^4 f(x)\,dx = k \longrightarrow$ this value is called *kurtosis* and gives us information about the ***"importance" of the tails***

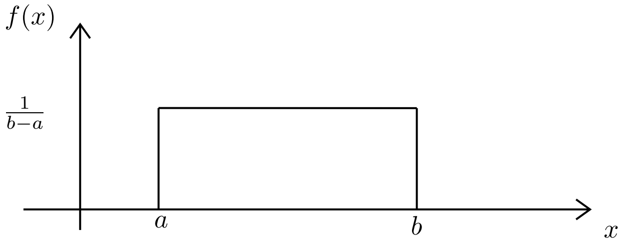
  The higher the value $\beta_2 = \frac{E\left[(x-\mu)^4\right]}{\sigma^4}$, the greater the extremity of deviations

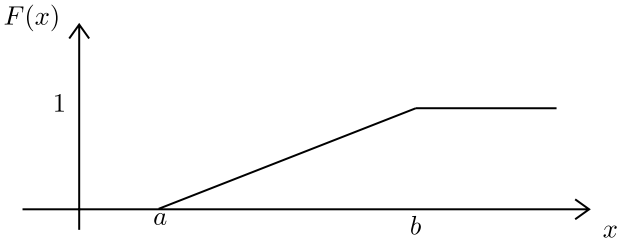
  note. $\beta_2^{Gaussian} = 3 \Longrightarrow \gamma_2 = \beta_2 - 3$

Other ways to analyze the symmetry of a function is to calculate mode [2], median [3] and expectation value: if there is symmetry, these values should be the same

2.2 Uniform distribution

$$f(x) = \begin{cases} 0 & x < a \text{ or } x > b \\ \frac{1}{b-a} & x \in (a,b) \end{cases}$$





- $\mu =$ mean $= \frac{b+a}{2}$
- $v = \frac{(b-a)^2}{12}$
- $\sigma = \frac{|b-a|}{\sqrt{12}}$

$v = \int (x-\mu)^2 f(x)\,dx = \int x^2 f(x)\,dx + \mu^2 \int f(x)\,dx - 2\mu \int x f(x)\,dx = \int x^2 f(x)\,dx - \mu^2$

$\int x^2 f(x)\,dx = \int_a^b x^2 \frac{1}{b-a}\,dx = \frac{1}{b-a}\int_a^b x^2\,dx = \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b = \frac{1}{b-a}\frac{1}{3}(b^3-a^3) = \frac{1}{b-a}\frac{1}{3}(b-a)(b^2+a^2+ab) = \frac{1}{3}(b^2+a^2+ab)$

$\mu^2 = \left(\int_a^b x f(x)\,dx\right)^2 = \left(\frac{1}{b-a}\left(\frac{b^2}{2}-\frac{a^2}{2}\right)\right)^2 = \left(\frac{(b-a)^2}{2}\frac{1}{b-a}\right)^2 = \frac{b^2+a^2-2ab}{4}$

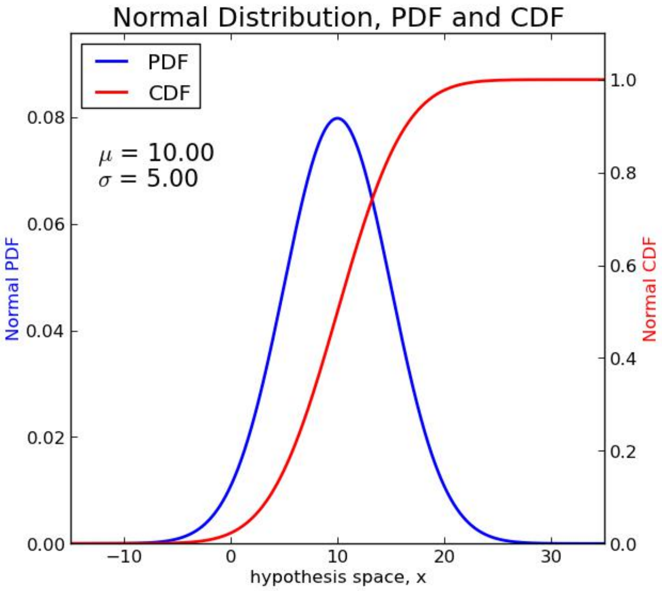
$$v = \frac{b^2+a^2+ab}{3} - \frac{b^2+a^2-2ab}{4} = \frac{b^2+a^2+2ab}{12} = \frac{(b-a)^2}{12}$$

[2] most repeated value

[3] central value of the ordered data

2.3 Gaussian distribution

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



- $E[x] = \mu$
- $v = \sigma^2$
- $\gamma_1 = 0$
- $\gamma_2 = 0$
- $1\sigma = 0.68 \quad | \quad 2\sigma = 0.95 \quad | \quad 3\sigma = 0.997 \quad | \quad 4\sigma = 1 - 3 \cdot 10^-5 \quad | \quad 5\sigma = 1 - 6 \cdot 10^-7$
- $F(x) = \frac{1}{2}\left(1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)$
- $erf = \frac{2}{\sqrt{\pi}} \int e^{-t^2}\, dt$

The gaussian distribution has the ***reproducibility*** property: if x, y have the same ***pdf***, then $z = x + y$ is described with the same probability density function.
It is possible to normalize the gaussian with the transformation $y = \frac{x-\mu}{\sigma} \rightarrow f(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$

2.4 Log-normal distribution

For values that cannot be less than zero, it is possible to use a modified gaussian called log-normal distribution

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

2.5 Central limit theorem

Given a set of **random**, **indipendent** values $\{x_i\}_{i=1,\ldots,N}$, each defined by $f_i(x_i)$ with **mean and variance**,
the limit of the sum of the values, for $\lim_{N\rightarrow\infty}$, is a gaussian distribution.

$$s = \sum x_i$$

$$\lim_{N\to\infty} s = G\left(s; \sum \mu_i, \sqrt{\sum v_i}\right)$$

If the values have all the same *pdf* $f(\mu, \sigma)$, then

$$\overline{x} = \frac{\sum x_i}{N}$$

$$\lim_{N\to\infty} \overline{x} = G\left(\overline{x}; \mu, \sqrt{\sum \frac{\sigma^2}{N}}\right)$$

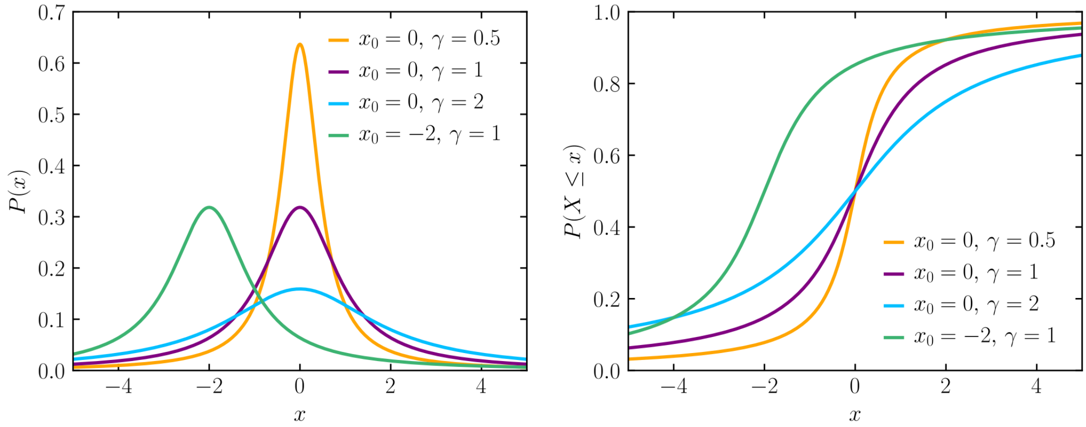
This theorem is used to interpret

- the distribution of repeated measurements of something;
- the standard deviation of a probability density function in association to the uncertainty of its values.

In fact, σ is the error on the single measurement and μ is itself a random number with it's own *pdf*. Calling $\hat{\mu}$ the central value of μ's pdf, the gaussian of $\hat{\mu}$ is narrower for incrementing number of values because $\hat{\sigma} = \frac{\sigma}{\sqrt{N}}$. This value is called ***error of the mean***.
The correct way to represent the mean value is $\mu = (\hat{\mu} \pm \hat{\sigma})$.

2.6 Cauchy distribution

$$f(x, \alpha, \mu) = \frac{1}{\pi\alpha}\frac{1}{1 + \frac{(x-\mu)^2}{\alpha^2}} \to \alpha = \frac{2}{\gamma} \to f(x, M, \gamma) = \frac{1}{2\pi}\frac{\gamma}{(x-M)^2 + \left(\frac{\gamma}{2}\right)^2}$$



This distribution is similar to the Gaussian: simmetric around the maximum value with decreasing tails. Yet, it is different because it is not possible to calculate momenta for this distribution because they're not defined: the central limit theory does not hold for this distribution.

It's possible to use a trick: shrinking the interval of integration to $(M - \delta, M + \delta)$ with δ quite large

$$\mu = \int_{-\infty}^{\infty} x f(x, M, \gamma)\, dx \to \mu_\delta = \int_{M-\delta}^{M+\delta} x f(x, M, \gamma)\, dx \longrightarrow_{\delta\to+\infty} \mu$$

The maximum is in $\left(M, \frac{2}{\pi\gamma}\right)$ and the *full width half maximum* is γ.
The Cauchy distribution is typical for resonance phenomena.
The Cauchy distribution has the *reproducibility* property.

2.7 Bernoulli distribution

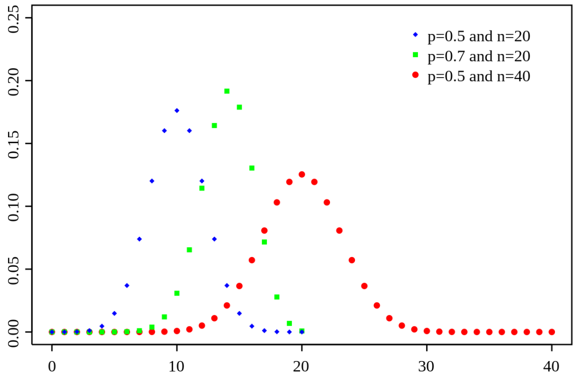
$$f(x) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \end{cases} \qquad x \in \Omega = \{1 \text{ (success)}, 0 \text{ (failure)}\}$$

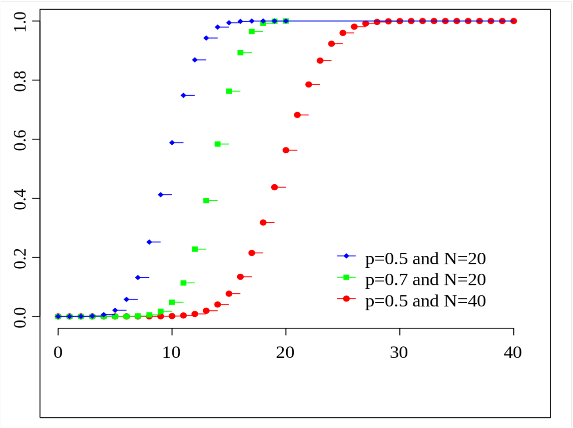
Given an event with two possible outcomes (success or failure), the Bernoulli distribution describes the probability p of a certain outcome.

$E(x) = \sum x f(x) = p$
$V(x) = p(1-p)$

2.8 Binomial distribution

$$B(k,p,N) = \frac{N!}{k!\,(N-k)!} p^k (1-p)^{N-k} = \binom{N}{k} p^k (1-p)^{N-k}$$





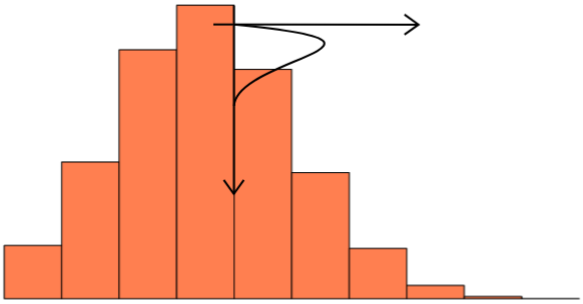
This distribution describes the probability, over N attempts, of having k successes of an event with a probability p to happen.

- $\mu = N \cdot p$
- $v = N \cdot p(1-p)$
- $\gamma_1 = \frac{1-2p}{\sqrt{N \cdot p(1-p)}}$
- $\gamma_2 = \frac{1-6p(1-p)}{N \cdot p(1-p)}$

For $N \to \infty$ the binomial, which is a discrete *pdf*, tends to a Gaussian.
The binomial distribution has the *reproducibility* property.

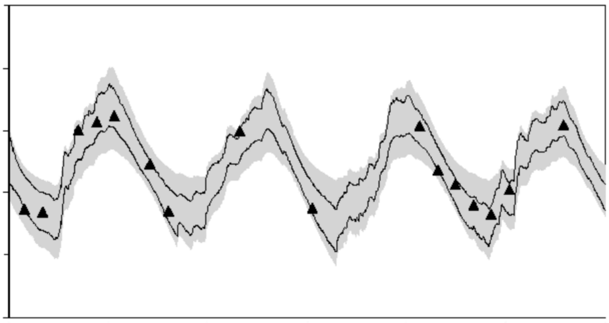
Why is the binomial distribution so important?

Let's take an histogram and the corresponding probability density function of the measurements of a continuous variable. The height of each bin is related to the probability of a single measurement to fall into the range of the bin.
Focusing on a single bin, *the number of events expected in a bin follow a binomial distribution.* The same *pdf* can describe different histogram (the so called "toy experiment", i.e. repeating an experiment several times) without errors: this is because there is an intrinsic possibility of variation of the probability of each bin.



The histogram's variation of the same probabiliy density function is called ***binomial fluctuation*** and goes as $\sqrt{v} = \sigma$.

In graphic representation of a probability density function, each point has an expected uncertainty that can be represented with *uncertainty bands*, which statistically include 68% of the points.



2.9 Multinomial distribution

The generalization of a binomial distribution is a multinomial distribution. Let's take a binomial distribution: $B(k,p,N) = \frac{N!}{k!(N-k)!}p^k(1-p)^{N-k}$.

It is possible now to write this distribution as the probability of an event 1 and an event 2, rather than the probability of one event to happen or not: $k \to N_1$, $(N-k) \to N_2$, $p \to p_1$, $1-p \to p_2$.

The distribution becomes $B = \frac{N!}{N_1!N_2!}p_1^{N_1}p_2^{N_2}$.

Extending to multiple events, we obtain

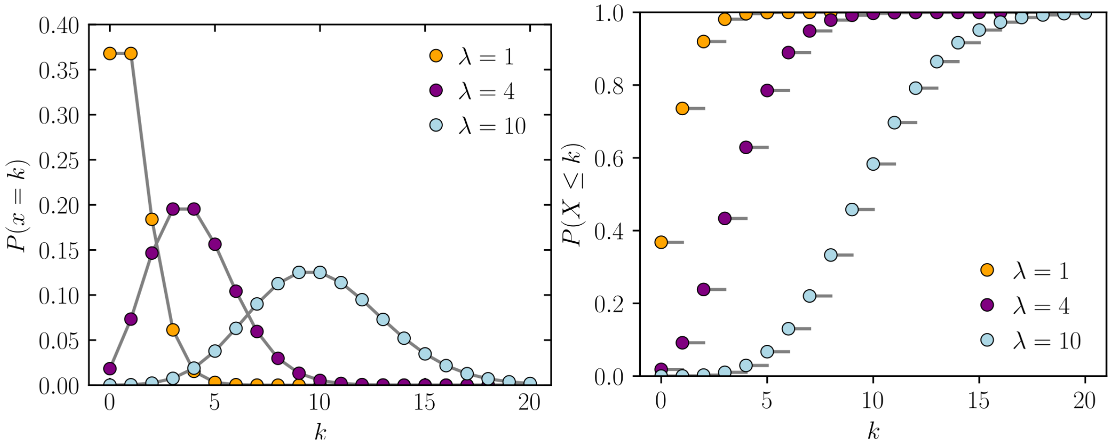
$$M(\underline{N_i}, \underline{p_i}, N) = N! \prod \frac{p_i^{N_i}}{N_i!}$$

note the bins are correlated and not independent: reducing one means increasing other bins.

- $\mu_i = E[k_i] = Np_i$
- $\sigma_i^2 = V[k_i] = Np_i(1-p_i)$

2.10 Poisson distribution

$$P(k, \lambda) = e^{-\lambda} \frac{\lambda^k}{k!}$$

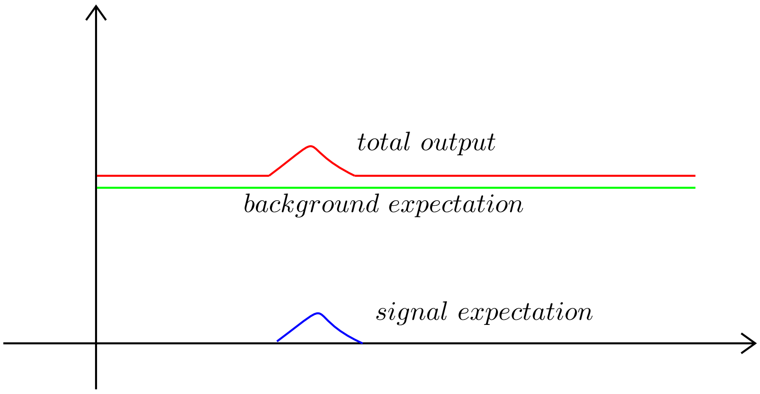


This distribution describes counting experiment of independent event with a constant expected rate λ. The number of events is proportional to the observation window.

- $\mu = \lambda$
- $v = \lambda$
- $\gamma_1 = \frac{1}{\sqrt{\lambda}}$
- $\gamma_2 = \frac{1}{\lambda}$

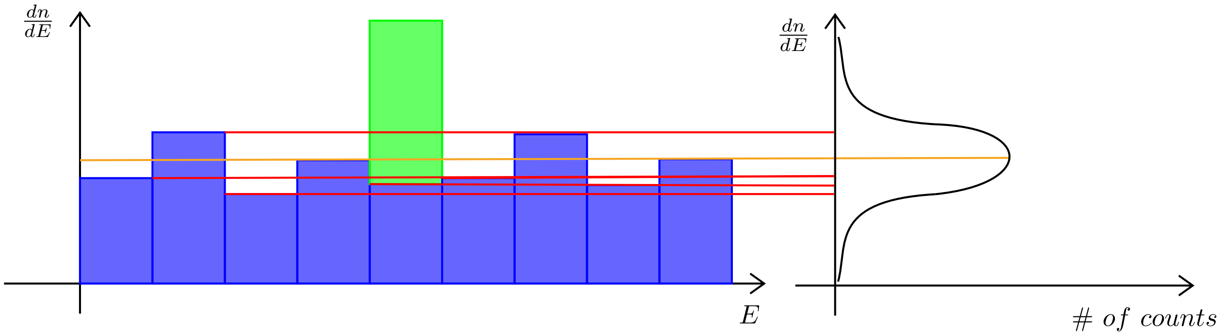
The distribution becomes more and more symmetric for increasing λ.
The variance of this distributions tells us that there is an *intrinsic fluctuation* in the outcome of the experiment which is directly correlated to the expected rate: the expected events are $\lambda \pm \sqrt{\lambda}$.
The Poisson distribution has the *reproducibility* property.

Fluctuations: signal + background experiment



Observing the output of a measuring instrument, the shown curve is the red one. The real output, however, is the total output minus the "background noise". Let's suppose that the background expectation can be calculated: λ_{bkg}. The actual background is described by a Poisson distribution and the expected value is $\lambda_{bkg} \pm \sqrt{\lambda_{bkg}}$. A characteristic value is the *significance*: $\frac{\lambda_{signal}}{\sqrt{\lambda_{bkg}}}$.

How do I decide if there's a signal or not? The signal is significative if its value is much larger than the fluctuation of the background value. The background, as said before, is distributed as a Poissonian so there is signal if, given a $\lambda \pm \sqrt{\lambda}$ background, the signal $S \gg \sqrt{\lambda}$



2.11 Binomial vs Poisson distribution

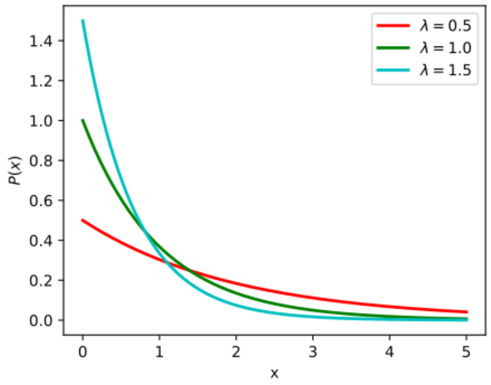
The binomial and Poisson distribution are related: when the number of observation in a binomial becomes very big, the probability correlated to each bin is close to zero. This means that it tends to a Poisson distribution. In addition, for the central limit theorem both distributions tends to a Gaussian distribution: the binomial for $N \to +\infty$, the Poisson distribution for $\lambda \to +\infty$.

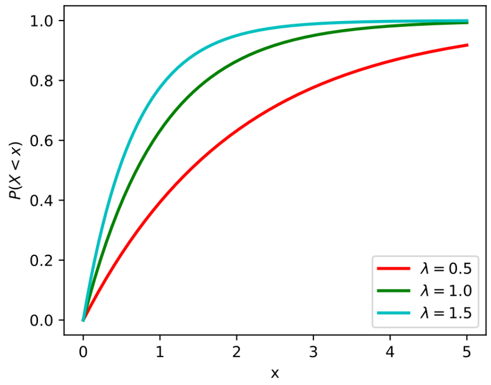
$$\begin{tikzcd}
Binomial \arrow[rr, "p \to 0 \quad N \to +\infty", "pN = \lambda"'] \arrow[dr, "N \to +\infty"'] & & Poisson \arrow[dl, "\lambda \to +\infty"] \\
& Gauss &
\end{tikzcd}$$

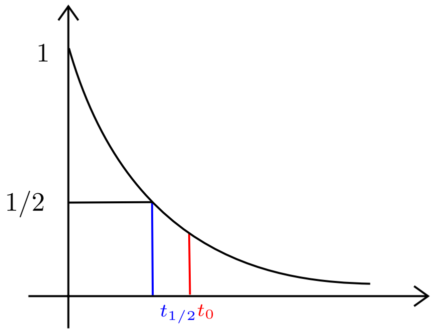
The Poisson distribution is used for observation with an expected rate; the binomial to describe a distribution with the total number of observation known.
Moreover can be said thinking about an experiment: it is possible to use the Poisson distribution before the acquisition of the data, to predict the outcome; the binomial to describe the actual results after the acquisition, observing their division into bins and the fluctuations of the bins themselves.

2.12 Exponential distribution

$$f(t, t_0) = \frac{1}{t_0} e^{-\frac{t}{t_0}}$$







- $\mu = t_0$
- $v = \sigma^2 = t_0^2$
- $\gamma_1 = 2$
- $\gamma_2 = 6$

Mean and variance are defined: the central limit theorem holds.
This means that the exponential distribution has the *reproducibility* property.
The central value is $\sum t_0$

2.12.1 Radioactive decays

Instable radioactive materials decay with an *half-life time* (time occurred for half of the original quantity to decay).
The decay of an atom is a probabilistic event, with an associate probability p to happen.

Let's suppose that we want to calculate the number of radioactive material left afetr a certain amount of time:
$N(t+dt) = N(t) - N_{decay}(dt) = N(t) - N(t) \cdot p \cdot dt$
$N(t+dt) - N(t) = -N(t)p\,dt \quad \Rightarrow \quad \frac{N(t+dt)-N(t)}{dt} = -N(t) \cdot p$
$\frac{dN(t)}{dt} = -pN(t) \qquad \text{for } dt \to 0 \quad N(t) = N_0 \cdot e^{-pt}$

It's now possible to write a function that describes the number of atoms decayed:
$N_{decay}(t) = N_0 \int_0^t f'(t)\,dt' = N_0 - N(t) = N_0(1-e^{-pt})$ with $f(t)$ corresponding to the function that describes the decay probability for a single atom.

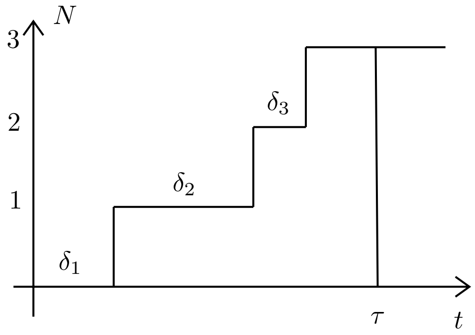
To get to this function, it's necessary to derive: $f(t) = pe^{-pt}$. Changing the variable $p \to \frac{1}{t_0}$ we obtain the exponential distribution described before.

However, the decay of a material is a physical event: how is it possible to identify the parameter t_0?
$\Delta N = \int_0^{\Delta t} \frac{1}{t_0} e^{-\frac{t}{t_0}}\,dt = N_0(1-e^{-\frac{\Delta t}{t_0}}) \simeq -N_0 \frac{\Delta t}{t_0}$
$|\Delta N| = N_0 \frac{\Delta t}{t_0}$

$$t_0 = N_0 \frac{\Delta t}{\Delta N}$$

ΔN is the result of a counting experiment, therefore it has its own intrinsic uncertainty, which can be described with a binomial distribution $B = N_0 p(1-p)$ or with a Poisson distribution $P = N_0 p$.

If the parameter t_0 is much larger than the typical measurement interval, $f(t, t_0) = \frac{1}{t_0}$ because $e^{-\frac{t}{t_0}} \simeq 1$.
The function becomes $f(t, t_0) = \frac{1}{t_0}$ and, for an interval τ, the number of observed event is a Poisson distribution with an expected constant rate $\lambda = \frac{\tau}{t_0}$.



The interval of time between two decays obeys to an exponential *pdf* $f(\delta)$.

2.12.2 Survival probability

Given the $f(t, t_0)$ probability of non-survival, it is possible to obtain the probability of survival $q(t) = 1 - f(t, t_0)$.

Let's calculate $q(t+\tau) = e^{-\frac{t+\tau}{t_0}} = e^{-\frac{t}{t_0}}e^{-\frac{\tau}{t_0}} = q(t)q(\tau)$.
Because t is a non defined time during the experiment, this means that the past is not important to describe the probability at a certain time: the exponential distribution has the *memoryless property*.

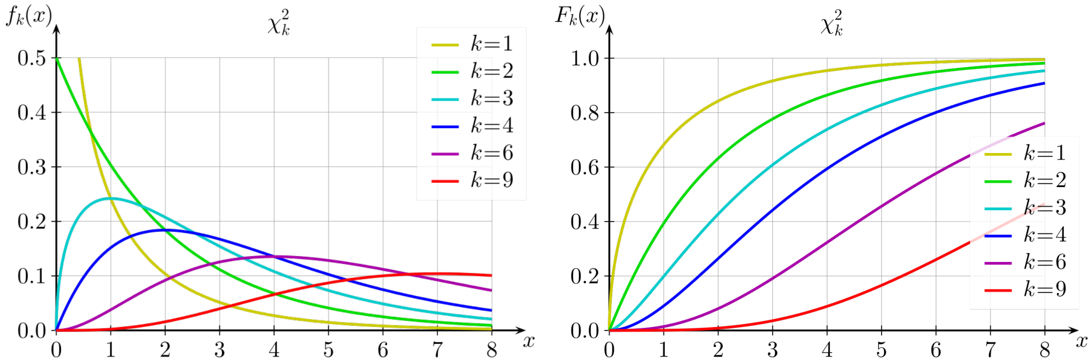
- *pdf* with $q(t)$ without memory $\Longleftrightarrow$ *pdf* is exponential;
- a counting experiment following a Poisson distribution $\Longleftrightarrow$ interval time between two events obey to an exponential distribution.

2.12.3 Monte Carlo simulations

The connection between the exponential and the Poisson distribution shown before can be used in algorithmic simulations called *Monte Carlo simulations*. A random number is generated following an exponential distribution, which represents the interval time between two events. Numbers generated this way are summed until the value becomes greater than a fixed value we want to study. At this point, the number of values generated is inserted into an histogram. The algorithm is repeated various time.
The final histogram follows a Poisson distribution.
This simulation are used to describe the number of events during a certain time, with an expected value between an event and the following.

2.13 χ^2 distribution

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



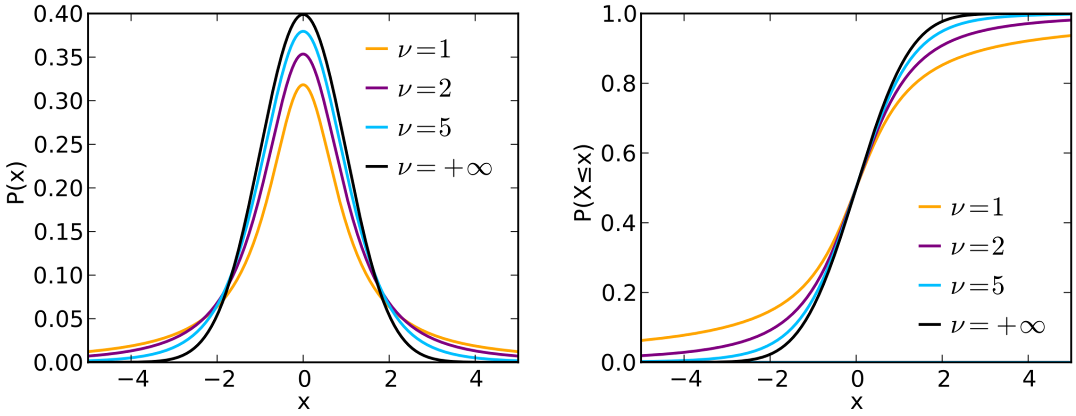
The χ^2 distribution is asymptotical to a Gaussian distribution for $n \to +\infty$ and has the *reproducibility property*.
The n value is called *degrees of freedom of the function*.

- $E(\chi^2) = n$
- $V(\chi^2) = 2n$
- $\gamma_1 = 2\sqrt{\frac{2}{n}}$
- $\gamma_2 = \frac{12}{n}$

Note that Γ is a function that generalizes the factorial operator in the complex field: $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt$ with $\Re(z) > 0$.

2.14 Student's t distribution

$$f(t,\nu) = \frac{1}{\sqrt{\pi\nu}}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

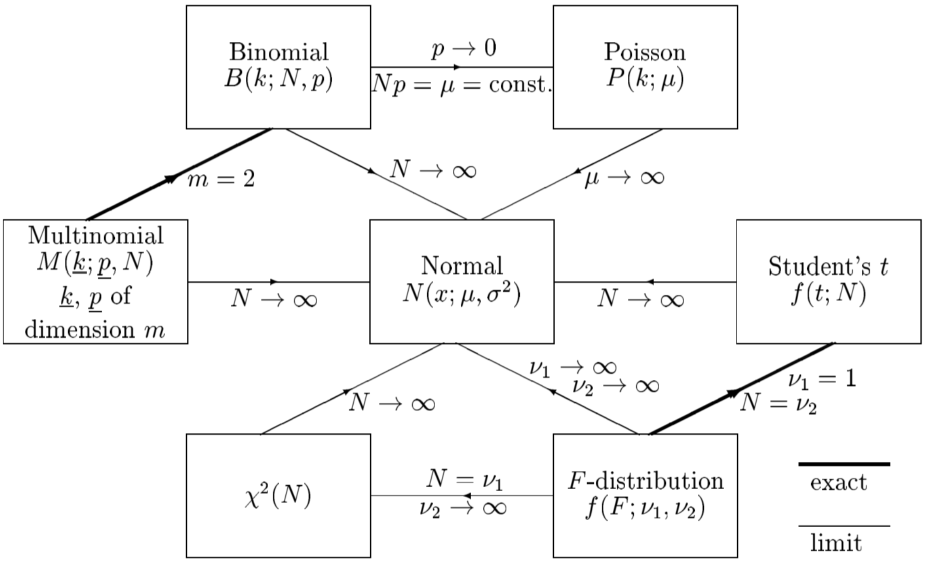


The *student's t* distribution is asymptotical to a Gaussian distribution for $\nu \to +\infty$ and to a Cauchy distribution for $\nu = 1$.

The ν value is called *degrees of freedom of the function.*

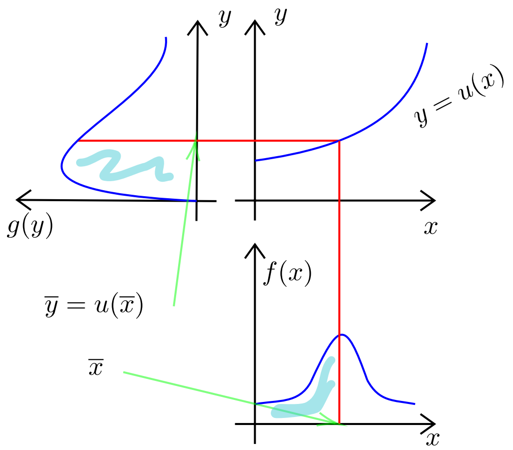
- $E(t) = 0 \qquad \nu > 1$
- $V(t) = \frac{\nu}{\nu-2} \qquad \nu > 2$
- $\gamma_1 = 0$
- $\gamma_2 = \frac{6}{\nu-4} \qquad \nu > 4$

2.15 Relations between distributions



2.16 Composite *pdf*'s

Let's suppose a distribution $x : f(x)$, and a composite function $y = u(x)$. What is the distribution of the composed function $y : g(y)$?



The function $y = u(x)$ can be seen as a change of variables, but $g(y) \neq f\left(u(x)\right)$ because there is a change in density! Let's proceed with ordere.

Fixing a value $\overline{x}$ we can determine two values: $f(\overline{x})$ and $F(\overline{x}) = \int_{-\infty}^{\overline{x}} f(t)\,dt$. The blue areas are identical. Assuming $v(\overline{y}) = u^{-1}(\overline{y})$

$$\int_{-\infty}^{v(\overline{y})} f(t)\,dt = \int_{-\infty}^{\overline{x}} f(t)\,dt = \int_{-\infty}^{\overline{y}} g(t)\,dt$$

$$g(\overline{y}) = \frac{d}{dy}F(\overline{y}) = \frac{d}{dx}\int_{-\infty}^{v(\overline{y})} f(t)\,dt = \frac{d}{dx}\Phi\left(v(\overline{y})\right) = \Phi'(\overline{y})\frac{d}{d\overline{y}}v(\overline{y}) = f\left(v(\overline{y})\right)\cdot|v'(\overline{y})|$$

2.17 Propagation of uncertainties

Given a *pdf* $\underline{x} \to f(\underline{x})$ with mean and variance known, how does the errors propagate to a function $y = u(\underline{x})$ with $u : \mathbb{R} \to \mathbb{R}$ know but mean and variance unknown?

Mean

$$y = u(\underline{x}) \simeq u(\mu_x) + \sum_{i=1}^{N} \frac{du(\underline{x})}{dx_i}(x_i - \mu_i)$$

$$\mu_y = E(y) = E(u(\underline{x})) \simeq E\left(u(\mu_x) + \sum_{i=1}^{N} \frac{du(\underline{x})}{dx_i}(x_i - \mu_i)\right) = E(u(\mu_x)) + \sum_{i=1}^{N} \frac{d}{dx_i} u(\mu_x) E(x_i - \mu_i) =$$

$$= E(u(\mu_x)) = u(\mu_x)$$

$$\sigma_y^2 = E\left((y - \mu_y)^2\right) = E(y^2) - E(y)^2 = E(u(\mu_x)^2) + E\left(\sum_{i,j} \frac{du}{dx_i}\frac{du}{dx_j}(x_i - \mu_i)(x_j - \mu_j)\right) - \mu_y^2 =$$

$$= \mu_y^2 + \sum_{i,j} \frac{du}{dx_i}(\mu_x)\frac{du}{dx_j}(\mu_x)\sigma_{ij} - \mu_y^2 = \sum_{i,j} \frac{du}{dx_i}(\mu_x)\frac{du}{dx_j}(\mu_x)\sigma_{ij} = \sum_{i=1}^{N} \left(\frac{du}{dx_i}\mu_x\right)^2 \sigma_i^2 + \sum_{i \neq j} \frac{du}{dx_i}(\mu_x)\frac{du}{dx_j}(\mu_x)\sigma_{ij}$$

If variables are not linearly correlated, the second term is equal to zero. To better understand this, see the next chapteer.

Chapter 3

Multidimensional *pdf*'s

The probability density functions and distributions seen so far are monodimensional, but *multidimensional* functions are important to describe complex systems.

3.1 Joint, marginal and conditional distributions

A multidimensional distribution is called *joint* distribution

$$\underline{x} \in \Omega \subseteq \mathbb{R}^n \quad f(\underline{x}) \to \mathbb{R} \quad \underline{x} = (x_1, ..., x_n)$$

$$pdf(\underline{x} \in E) = \int_E f(\underline{x})\, d\underline{x}$$

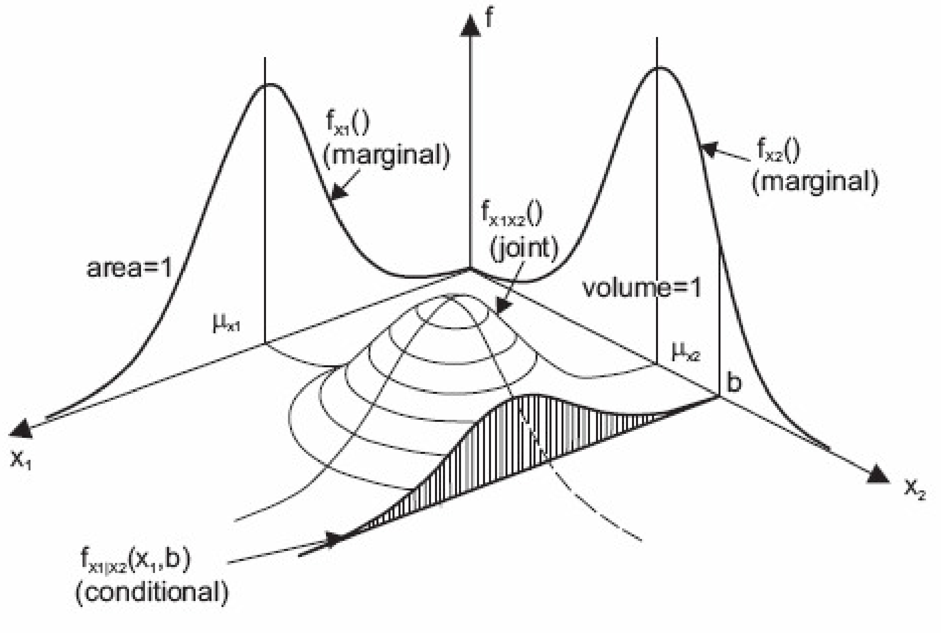
The *marginal* probability is the probability of one of the variables, regardless the others

$$pdf(x_1) = \int_\Omega f(\underline{x})\, d\underline{y} \quad \underline{y} = (x_2, ..., x_n)$$

The *conditional* probability is the probability of one variable given a fixed value of the others

$$pdf(x_1|\underline{y_k}) = \frac{pdf(x, \underline{y_k})}{pdf(\underline{y_k})}$$

oss. $pdf(x, \underline{y_k}) = pdf(x|\underline{y_k}) pdf(\underline{y_k})$



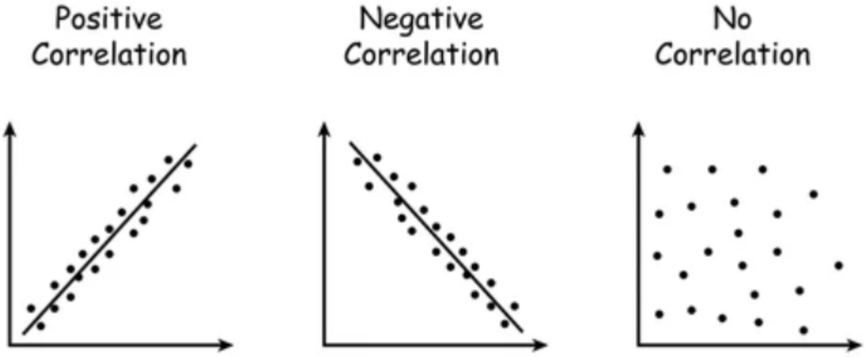
3.2 Multidimensional momenta

For simplicity, the following momenta are calculated for a *bidimensional pdf*; they can be extended to multidimensional without any problem.

- $\begin{cases} \mu_x = \int x f(x,y)\, dx\, dy = E(x) \\ \mu_y = \int y f(x,y)\, dx\, dy = E(y) \end{cases}$
- $\begin{cases} \sigma_x^2 = E\left((x-\mu_x)^2\right) \\ \sigma_y^2 = E\left((y-\mu_y)^2\right) \end{cases}$
- There's a new parameter called *covariance* which is a measure of the linear correlation of two variables:
  $cov(x,y) = \sigma_{xy} = E\left((x-\mu_x)(y-\mu_y)\right) = E\left(xy - \mu_x y - \mu_y x + \mu_x\mu_y\right) =$
  $= E(xy) - \mu_x E(y) - \mu_y E(x) + \mu_x\mu_y = E(xy) - \mu_x\mu_y - \mu_y\mu_x + \mu_x\mu_y = E(xy) - \mu_x\mu_y$
- It is also used a parameter called *linear correlation coefficient* $\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$ with $|\rho| \leq 1$

  For a total correlation $\rho = 1$, for independent variables $\rho = 0$.

If $\sigma_{xy} \neq 0$, x and y can be *correlated* ($\rho > 0$) or *anti-correlated* ($\rho < 0$). If the values are evenly distributed, it is not possible to make considerations about their correlation; $\sigma_{xy} = 0$.



ex. given a set of measurements $\{x_i\}_{i=1,\dots,N}$ not necessarily correlated, $z = \sum a_i x_i$.
$\mu_z = E(\sum a_i x_i) = \sum a_i E(x_i) = \sum a_i \mu_i$
$\sigma_z^2 = E\left((z-\mu_z)^2\right) = E\left((\sum a_i x_i - \sum a_i\mu_i)^2\right) = E\left((\sum a_i x_i - a_i\mu_i)^2\right) =$
$= \sum \left(E\left(a_i x_i - a_i\mu_i\right)\right)^2 + \sum_{i\neq j} E\left((a_i x_i - a_i\mu_i)(a_j x_j - a_j\mu_j)\right)$

The first term is the variance of the measurements, the second term is the covariance: if measurements are not linearly independent, this value is different from zero. It is important to clarify this: covariance tells us about linear correlation: if the correlation is not linear, covariance will be zero even if the coefficient are correlated [1].

The covariance is a correction factor in case of duplication of results: for a set of measurements, the greater the number of measurements the smaller the variance gets. In case of duplicate measurements, the covariance factor increases the variance, so there isn't an understimation of its value.

[1] it is possible to visualize this thinking about a donut shape in a cartesian plane, sliced perpendicularly to the hole; each slice has a "mean value" zero even if for each piece x and y are correlated

3.3 Independent variables

The conditional distribution tells us that $f(x,y) = f(x|y) \cdot f_y(y) = f(y|x) \cdot f_x(x)$.
Because of marginal distribution, $f_y(y) = \int f(x,y)\,dx = \int f(y|x) f_x(x)\,dx$.
If $f(y|x)$ does not depend on x, $\int f_x(x)\,dx = 1 \longrightarrow f_y(y) = f(y|x)$.
In conclusion, $f(x,y) = f_x(x) \cdot f_y(y)$.

Two variables are *independent* if $f(y|x)$ does not depend on x (it is enough to define independence because of the Bayes theorem).

Also, if x, y are independent, $cov(x,y) = E[(x-\mu_x)(y-\mu_y)] = \left(\int (x-\mu_x) f_x(x)\,dx\right)\left((y-\mu_y) f_y(y)\,dy\right) = \left(\int x f_x(x)\,dx - \mu_x\right)\left(\int y f_y(y)\,dy - \mu_y\right) = (\mu_x - \mu_x)(\mu_y - \mu_y) = 0$

$$(x,y) \text{ independent } \begin{cases} \Leftrightarrow f(x,y) = f_x(x) \cdot f_y(y) \\ \Rightarrow \sigma_{xy} = 0 \end{cases}$$

3.4 Generalization to N variables/dimensions

Given $\underline{x} \in \Omega \subseteq \mathbb{R}^N$, $f : \mathbb{R}^N \to \mathbb{R}$ and $A \subseteq \Omega$:

- $p(\underline{x} \in A) = \int_A f(\underline{x})\,d\underline{x}$
- $\underline{\mu} = E(\underline{x}) = \int_\Omega \underline{x} f(\underline{x})\,d\underline{x}$
- $\sigma_i^2 = \int_\Omega (x_i - \mu_i)^2 f(\underline{x})\,d\underline{x}$
- $\sigma_{ij} = \int_\Omega (x_i - \mu_i)(x_j - \mu_j)\,d\underline{x}$

It is possible to introduce a matrix called ***covariance matrix***:

$$V = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1N} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2N} \\ \dots & \dots & \dots & \dots \\ \sigma_{N1} & \sigma_{N2} & \dots & \sigma_N^2 \end{pmatrix}$$

Because $\sigma_{ij} = \sigma_{ji}$ the matrix is symmetric, it can be diagonalized. It is always possible to make a change of basis and obtain a rotation, therefore is always possible to identify a change of base matrix that make the variables not linearly correlated.

$$V' = U^{-1} V U$$

3.5 Multidimensional Gaussian distribution

$$f(\underline{x}) = \frac{1}{(2\pi)^{\frac{N}{2}} \, (\det V)^{\frac{1}{2}}} \cdot e^{\left(-\frac{1}{2}(\underline{x}-\underline{\mu})^t V^{-1} (\underline{x}-\underline{\mu})\right)}$$

In 2 dimensions: $f(x,y|\mu_x \mu_y \sigma_x \sigma_y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \cdot e^{\left(-\frac{1}{2}\frac{1}{(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} + \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right)\right)}$

If $\rho = 0$, $G = \frac{1}{2\pi\sigma_x\sigma_y} \cdot e^{-\frac{1}{2}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]}$
which is the product of two monodimensional gaussian: in a gaussian distribution in two dimensions, if the two variables are not correlated then the variables are independent.

Chapter 4

Statistics

Let's suppose we have a set of data $\{x_i\}_{i=1,\ldots,N}$ called *experiment*. This values are independent and identically distributed *(IID)* as a probability density function $x : f(x, \underline{\theta})$ which depend on a parameter $\underline{\theta}$. The aim is to find out the true value of the unknown parameter $\underline{\theta}$.

Statistic is a function of random variables (data set) which does not depend on the parameter; it is a random variable itself with its own *pdf*.

An ***estimator*** is a statistic designed to find an estimate of a parameter.

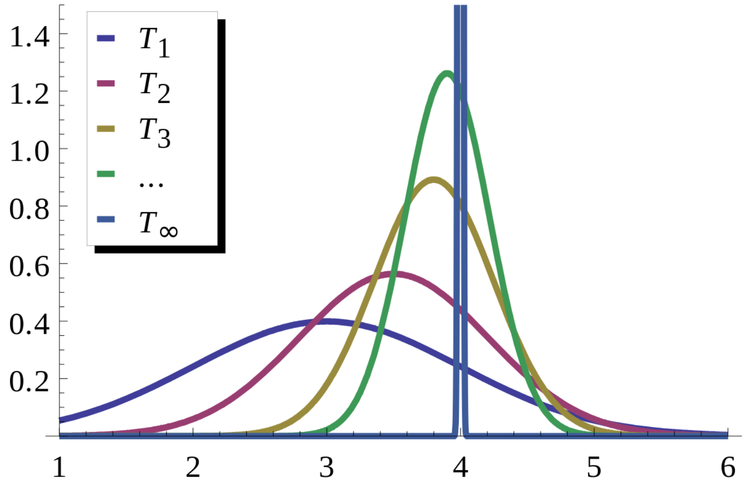
4.1 Properties of an estimator

4.1.1 Consistency

An estimator is consistent if, given N experiments each with its own mean value, we have θ_n,

$$\forall \epsilon, \delta \quad \exists N \text{ such that } \forall n > N \quad p\left(\left|\hat{\theta_n} - \theta_t\right| > \delta\right) < \epsilon$$

This means that for increasing number of experiments, the mean value of the pdf tends to the true value; the tails of the distribution become negligible.



obs. convergence has different meanings: it can be referred to distributions (for n repeated experiments, f_n distributions, $\lim_{n\to+\infty} f_n(x) = f(x)\ \forall x$); it can be referred to the consistency of an estimator; it can be *almost certain* or *in R-average* [1].

[1] convergence in R-th mean, also known as R-average, measures convergence using a specific metric, usually involving the R-th power of the differences between the random variables in the sequence and their limit

Large numbers law

Given a set of data $\{x_i\}_{i=1,\dots,N}$ *independent and identically distributed*, known the mean μ and the variance σ^2 from the probability density function, the **sample average** is $\overline{x_N} = \frac{1}{N}\sum_{i=1}^{N} x_i$.
As $\lim_{N\to+\infty} \overline{x_N} = \mu$, the sample average is a consistent estimator of the true value.
In general, given a function $g(x)$, $\frac{1}{N}\sum_{i=1}^{N} g(x_i) \to E(g(x))$ for $N \to +\infty$.

4.1.2 Bias

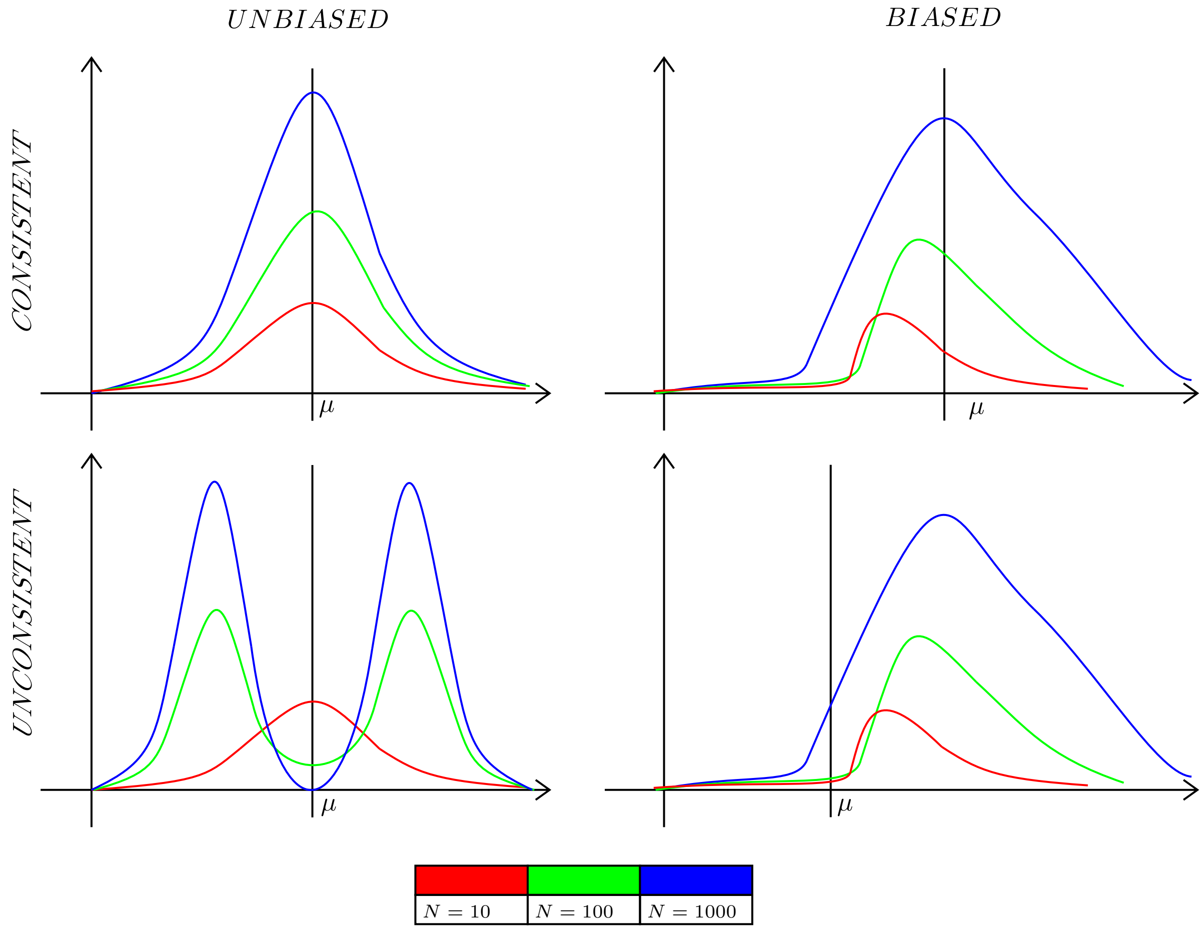
An estimator is unbiased if, given N experiments each with its own $E\left(\hat{\theta_n}\right)$,

$$E(\hat{\theta_n} - \theta_t) = E(\hat{\theta_n}) - \theta_t = b_n(\hat{\theta}) = 0 \quad \forall n \in N$$

If $b_n(\hat{\theta}) \to 0$ for $N \to +\infty$, the estimator is *asymptotically unbiased.*
If the previous condition isn't verified, the estimator is biased (distorted).

Simple mean is, in general, unbiased:
$E(\overline{x}) = E\left(\frac{\sum x_i}{N}\right) = \frac{1}{N}\sum_{i=1}^{N} E(x_i) = \frac{1}{N}\sum_{i=1}^{N} \int x_i f(x_i)\, dx_i = \frac{1}{N}\sum_{i=1}^{N} \mu = \frac{1}{N} N\mu = \mu$

4.1.3 Relation between consistency and bias



If an estimator is unbiased, that does not mean that a function $g(\hat{\theta})$ is unbiased, since $E()$ is a linear operator.

4.2 Expectation value of an estimator

Given an estimator $\hat{\theta}(\underline{x})$, with $\{x_i\}_{i=1,\ldots,N}$ independent and identically distributed,

$$E(\hat{\theta}) = \int_\Omega \hat{\theta}(\underline{x}) f(\underline{x})\, d\underline{x} = \int_\Omega \hat{\theta}(\underline{x}) \prod_{i=1}^{N} f(x_i, \theta_t)\, dx_i$$

because $f(\underline{x})$ is the *joint pdf* of all measurements.

4.3 Variance of an estimator

Variance is defined as $\sigma^2_\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$ so:

$E(\hat{\sigma^2_\mu}) = E\left(\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2\right) = \frac{1}{N}\sum_{i=1}^{N} E(x_i^2+\mu^2-2\mu x_i) = \frac{1}{N}\sum_{i=1}^{N}\left[E(x_i^2)+\mu^2-2\mu E(x_i)\right] =$
$= \frac{1}{N}\sum_{i=1}^{N}\left(\sigma_t^2+\mu^2-\mu^2\right) = \frac{1}{N}\sum_{i=1}^{N}\sigma_t^2 = \frac{1}{N}N\sigma_t^2 = \sigma_t^2$

Note that $E(x_i^2) - E(x_i)^2 = \sigma^2 \to E(x_i^2) = \sigma^2 + E(x_i)^2 = \sigma^2 + \mu^2$

It has been used μ: what if is it unknown? μ and $\overline{x}$ for a set of data can be "distant" and become equal only for $N \to +\infty$.

If $\overline{x}$ is used, $\sigma^2_{\overline{x}} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2 = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)^2$

$E\left(\hat{\sigma^2_{\overline{x}}}\right) = \frac{1}{N}E\left(\sum_{i=1}^{N}x_i^2\right) - \frac{1}{N^2}E\left(\left(\sum_{i=1}^{N}x_i\right)^2\right) = \frac{N}{N}(\mu^2+\sigma_t^2) - \frac{1}{N^2}(N\sigma_t^2+N^2\mu^2) = \mu^2+\sigma_t^2-\frac{\sigma_t^2}{N}-\mu^2 =$
$= \sigma_t^2 - \frac{\sigma_t^2}{N} = \sigma_t^2\left(\frac{N-1}{N}\right)$

Note that $V(\sum x_i) = E((\sum x_i)^2) - E(\sum x_i)^2 \to E((\sum x_i)^2) = V(\sum x_i) + E(\sum x_i)^2 =$
$= V(\sum x_i) + (N\mu)^2 = \sum V(x_i) + (N\mu)^2 = N\sigma_t^2 + (N\mu)^2$

$\hat{\sigma^2_{\overline{x}}}$ is asymptotically unbiased to σ_t^2: it is possible to define $\hat{\sigma^{2\prime}_{\overline{x}}} = \frac{N}{N-1}\hat{\sigma^2_{\overline{x}}}$ which is unbiased. The correction coefficient $\frac{N}{N-1}$ is called ***Bessel correction***.

4.4 About the *pdf*'s of an estimator

The mean value $\overline{x}$, because of the *central limit theorem* hypotesis, is distributed as a gaussian $G\left(\overline{x};\overline{x},\hat{\sigma^{2\prime}_{\overline{x}}}\right)$.
If x_i are distributed accordingly to a gaussian, the quantity $z_i = \frac{x_i-\mu}{\sigma_i}$ called *residual* (normalized deviation for the mean) is normally distributed $G(0,1)$ and the sum of the squared residuals

$$\sum z_i^2 = \frac{1}{\sigma_t^2}\sum_{i=1}^{N}(x_i-\mu)^2$$

is distributed as a χ^2 pdf with N degrees of freedom.
If μ is unknown, then it is possible to use the sample mean: doing so, 1 degree of freedom is lost because of the dipendence between the sample mean and the values of the sample themselves ($x_i - \overline{x}$ has "less power").
1 degree of freedom is lost for each parameter used.

4.5 Variance of the variance

The value of the variance itself, being related to an estimator, has its own variance, even if it has the Bessel correction.
$V(\sigma^2) = V\left(\sum_{i=1}^{N}\frac{(x_i-\overline{x})^2}{N-1}\right) = V\left(\sum_{i=1}^{N}\sigma^2\frac{1}{\sigma^2}\frac{(x_i-\overline{x})^2}{N-1}\right) = \frac{\sigma^4}{(N-1)^2}V\left(\frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i-\overline{x})^2\right).$

The value $A = \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i - \overline{x})^2$ is distributed as a χ^2 with $N-1$ degrees of freedom, so its variance is $V(A) = 2(N-1)$.
Therefore, $V(\sigma^2) = \frac{\sigma^4}{(N-1)^2}2(N-1) = 2\frac{(\sigma^2)^2}{N-1}$.

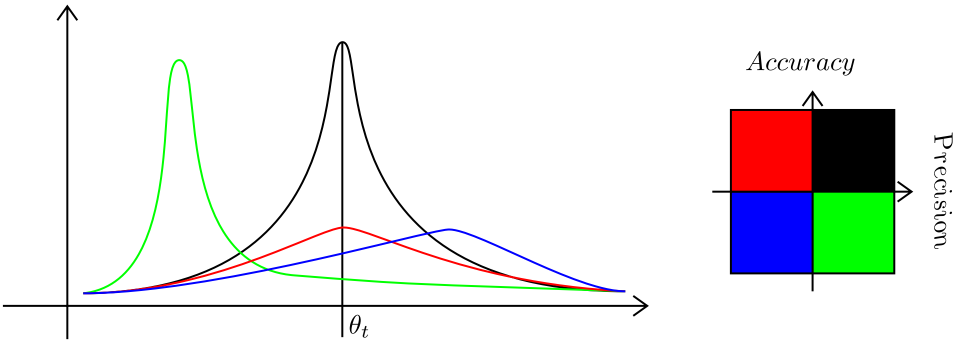
In conclusion,

$$\mu \to \hat{\mu} \pm \frac{\hat{\sigma_{\overline{x}}}'}{\sqrt{N}}$$

$$\sigma \to \hat{\sigma_{\overline{x}}}' \pm \sqrt{\frac{2\left(\hat{\sigma_{\overline{x}}}'\right)^4}{N-1}}$$

4.6 About measurements

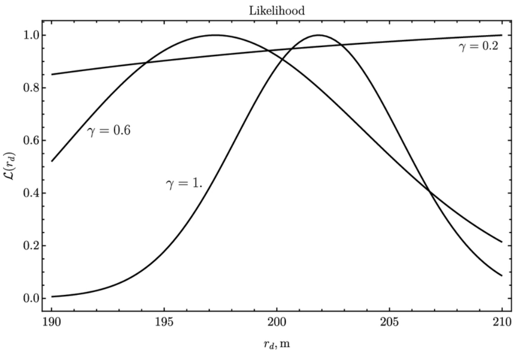
In statistics, accuracy and precision are not synonyms: the first one is connected with the bias of the estimator, the second one with its variance. It is important to balance these two things while choosing the estimator, to obtain the best one possible.
A value that combine this two things is the ***mean square error*** $MSE = b_n^2(\hat{\theta}) + V(\hat{\theta})$



4.7 Likelihood

The *likelihood* is a function that combines information from data (e.g. $\{x_i\}_{i=1,\dots,N}$ independent and identically distributed) and knowledge from the model (e.g. $x : f(x, \underline{\theta})$).

$$\mathcal{L}(\underline{x}, \underline{\theta}) = \prod_{i=1}^{N} f(x_i, \underline{\theta})$$



The function $\mathcal{L}$ depends on both the measurements x_i and on the parameters $\underline{\theta}$. However, once the measurements have been taken, the x_i are fixed. Then the likelihood can be regarded as a function of $\underline{\theta}$ only.
It is useful to define the logarithm of the likelihood as $l = \sum_{i=1}^{N} \ln f(x_i, \underline{\theta})$ because of the properties of the logarithms: the product becomes a sum, which is easier to operate.

The likelihood function changes accordingly to the sensitivity of the apparatus for the parameter: it is constant for no sensitivity and it becomes more and more shrinked for increasing sensitivity. For increasing *d.o.f.* it tends to a gaussian.

4.8 Cramer-Rao Theorem

It turns out that there is a lower limit to the variance of an estimator under certain general conditions. For Ω_x independent of θ and $\int dx$ that commutes with ∂_θ,

$$V(\hat{\sigma}) \geq \frac{\left(1 + \partial_\theta b_n(\hat{\theta})\right)^2}{E\left[\left(\frac{\partial l}{\partial \theta}\right)^2\right]} = \sigma^2_{\min}\left(\hat{\theta}\right)$$

The numerator is linked to the change of bias accordingly to the parameter itself (if the bias changes, variance is affected); the denominator is called *Fischer information*.

$\sigma^2_{\min}\left(\hat{\theta}\right)$ is called *minimum variance bound*; the closer variance is to this value, the more efficient the estimator is. Efficiency is measured between 0 and 1.

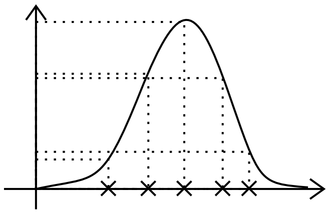
Chapter 5

Estimator Building

There are different and various techniques to build estimators:

- Substitution ($\hat{\mu} = \overline{x}$)
- Calculating momenta
- Least squares method (based on the minimum distance)
- Maximum likelihood method
- Machine learning-based methods

5.1 Maximum likelihood



Given a set of measure (see the x's on the graph) the maximum value of the *likelihood function* is reached for the true value of the parameter. Assuming that the true distribution is unknown, it is possible to find the best estimator by maximising the likelihood function (in fact, for values that are not the true one, the value is always lower; the closer the estimator is to the true value, the greater the likelihood function is).

Given the following hypothesis

- $\mathcal{L}, l$ dependent on θ
- $\mathcal{L}, (l)$ known and numerically differentiable twice
- Ω_x not dependent on θ
- $\int\, dx$ and ∂_θ commute

$$\frac{\partial \mathcal{L}(x, \underline{\hat{\theta}})}{\partial \underline{\hat{\theta}}} = 0 \qquad \frac{\partial l(x, \underline{\hat{\theta}})}{\partial \underline{\hat{\theta}}} = 0$$

5.1.1 Properties of maximum likelihood estimators

The finded estimators are

- consistent
- asymptotically unbiased
- asymptotically efficient

In frequestistic probability, the invariance principle holds too: if $\hat{\theta}$ is a maximum likelihood estimator for θ_t, $g(\hat{\theta})$ is a maximum likelihood estimator for $g(\theta_t)$.

Example - Exponential distribution

$\underline{t} = \{t_i\}_{i=1,...,N} \qquad f(t_i) = \frac{1}{\tau}e^{-\frac{t_i}{\tau}}$
$\mathcal{L}(\underline{t},\tau) = \prod_{i=1}^N \frac{1}{\tau}e^{-\frac{t_i}{\tau}} \to l(\underline{t},\tau) = \sum_{i=1}^N(-\ln\tau - \frac{t_i}{\tau}) \to \frac{\partial l(\tau)}{\partial\tau} = \sum_{i=1}^N(-\frac{1}{\tau}+\frac{1}{\tau^2}t_i) = 0$
$\frac{N}{\tau} = \sum_{i=1}^N \frac{t_i}{\tau^2} = \frac{\sum t_i}{\tau^2} \to \hat{\tau} = \frac{\sum t_i}{N}$

$f(t,\lambda) = \lambda e^{-\lambda t} \to l(\lambda) = \sum_{i=1}^N(\ln\lambda - \lambda t_i) \to \sum_{i=1}^N \frac{1}{\lambda} = \sum_{i=1}^N t_i \to \frac{N}{\lambda}\sum t_i \to \hat{\lambda} = \frac{N}{\sum t_i} = \frac{1}{\hat{\tau}}$
the invariance principle exposed before is, therefore, demonstrated.

Example - Gaussian distribution

$f(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \to \quad l(\mu,\sigma) = \sum_{i=1}^N\left(\ln\frac{1}{\sqrt{2\pi}} - \ln\sigma - \frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right)$
$\frac{\partial l(\mu,\sigma)}{\partial\mu} = \sum_{i=1}^N -\frac{1}{2}\frac{1}{\sigma^2}2(x_i-\mu)(-1) = 0 \to \sum_{i=1}^N x_i = \sum_{i=1}^N \mu$

Skipping the calculations, $\hat{\mu} = \frac{\sum x_i}{N} \qquad \hat{\sigma}^2 = \sum_{i=1}^N \frac{(x_i-\hat{\mu})^2}{N}$

If σ_i have different values for each measurement, $\hat{\mu} = \frac{\sum_{i=1}^N \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^N \frac{1}{\sigma_i^2}} \qquad V(\hat{\mu}) = \frac{1}{\sum_{i=1}^N \frac{1}{\sigma_i^2}}$

5.1.2 Variance of maximum likelihood estimators

Hessian method

Given an estimator obtained with maximum likelihood method, it is asymptotically unbiased and efficient. The Cramer-Rao theorem therefore is, asymptotically

$$V(\hat{\theta}) = \frac{1}{-E\left(\frac{\partial l}{\partial\theta}\right)^2} \simeq \frac{1}{-\frac{\partial^2 l(\hat{\theta})}{\partial\hat{\theta}^2}}$$

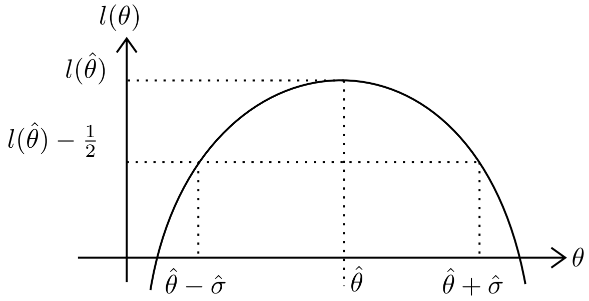
which, generalized for the multi-dimensional case, becomes

$$V(\hat{\underline{\theta}}) \simeq -\left(\frac{\partial^2 l(\hat{\underline{\theta}})}{\partial\theta_i\partial\theta_j}\right)^{-1}$$

Graphical method

Expanding with Taylor up to the 2nd order, $l(\theta) = l(\hat{\theta}) + \frac{\partial l(\hat{\theta})}{\partial\theta}(\theta-\hat{\theta}) + \frac{1}{2}\frac{\partial^2 l(\hat{\theta})}{\partial\theta^2}(\theta-\hat{\theta})^2 = l(\hat{\theta}) - \frac{1}{2}\frac{(\theta-\hat{\theta})^2}{\hat{\sigma}^2}$

Making the calculations, $l(\hat{\theta} \pm \hat{\sigma}) = l(\hat{\theta}) - \frac{1}{2}$

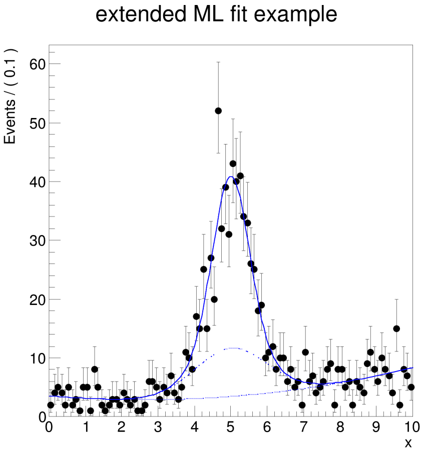

$l(\theta)$
$l(\hat{\theta})$
$l(\hat{\theta}) - \frac{1}{2}$
θ
$\hat{\theta} - \hat{\sigma}$
$\hat{\theta}$
$\hat{\theta} + \hat{\sigma}$

5.1.3 Extended likelihood

The likelihood function can be extended, with a "trick", for a complex case with more than one set of events going on at the same time, each with a probability to happen but with unknown number of events:

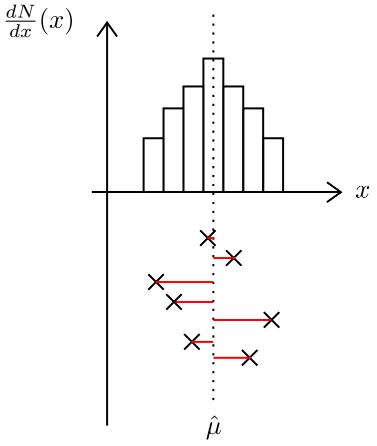
$$\mathcal{L} = P(n, N) \prod_{i=1}^{N} f(x_i, \theta)$$

This can be the case of an experiment with background and signal, all collected in the same way by the apparatus.



5.2 Least squares

Another method to calculate estimators is the *least squares method*, which is built upon the minimization of a *contrast function*. The objective is to search an estimator that is the closest to all the values.



Given $\{x_i\}_{i=1,\dots,N}$ values independent and identically distributed,

$$d(\{x_i\}, \hat{\mu}) = \sum_{i=1}^{N} \frac{(x_i - \hat{\mu})^\alpha}{\sigma_i^\alpha}$$

$\hat{\mu}$ is found when the value above is minimal.

The least squares method uses $\alpha = 2$ because of two reasons: first of all, it doesn't cancel out values (e.g. for $\hat{\mu} = 0$, -100 and 100 are very far and opposite but, if the power is not even, they cancel each other out); then, given the hypothesis of x_i distributed as a Gaussian, Q^2 is distributes as a χ^2.

$$Q^2 = \sum_{i=1}^{N} \frac{(x_i - \hat{\mu})^2}{\sigma_i^2} \to \hat{\mu} \text{ is such that } \frac{\partial Q^2(\hat{\mu})}{\partial \mu} = 0$$

The estimator for the value $\hat{\mu}$ is $-2\sum_{i=1}^{N} \frac{x_i - \hat{\mu}}{\sigma_i^2} = 0 \to \sum_{i=1}^{N} \frac{x_i - \hat{\mu}}{\sigma_i^2} = 0 \to \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} = \hat{\mu} \sum_{i=1}^{N} \frac{1}{\sigma_i^2}$

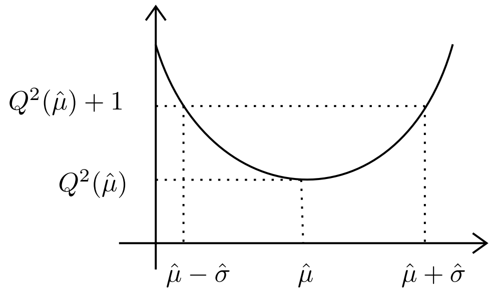
$$\hat{\mu} = \frac{\sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}}$$

$$V(\hat{\mu}) = \frac{1}{\left(\sum_{i=1}^{N} \frac{1}{\sigma_i^2}\right)^2} \sum_{i=1}^{N} \left(\frac{1}{\sigma_i^2}\right)^2 \cdot V(x_i) = \frac{1}{\left(\sum_{i=1}^{N} \frac{1}{\sigma_i^2}\right)^2} \sum_{i=1}^{N} \frac{\sigma_i^2}{\sigma_i^4} = \frac{1}{\left(\sum_{i=1}^{N} \frac{1}{\sigma_i^2}\right)^2} \left(\sum_{i=1}^{N} \frac{1}{\sigma_i^2}\right) = \frac{1}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}}$$

Uncertainty of $\hat{\mu}$

Defining $Q^2 = \sum_{i=1}^N \frac{(x_i-\mu)^2}{\sigma_i^2}$ and its derivatives $\partial_\mu Q^2 = -2\sum_{i=1}^N \frac{x_i-\mu}{\sigma_i^2} = 0$ and $\partial_\mu^2 Q^2 = \frac{2}{V(\hat{\mu})}$, it is possible to expand the least squares as
$Q^2(\mu) = Q^2(\hat{\mu}) + (\mu-\hat{\mu})\partial_\mu Q^2(\hat{\mu}) + \frac{1}{2}\partial_\mu^2 Q^2(\hat{\mu})(\mu-\hat{\mu})^2 + o(\hat{\mu}) = Q^2(\hat{\mu}) + \frac{1}{2}\frac{2}{V(\hat{\mu})}(\mu-\hat{\mu})^2 + o(\hat{\mu}) = Q^2(\hat{\mu}) + \frac{(\mu-\hat{\mu})^2}{V(\hat{\mu})}$

So, $Q^2(\hat{\mu} \pm \sqrt{V(\hat{\mu})}) = Q^2(\hat{\mu}) + \frac{\left(\hat{\mu}-\hat{\mu}\pm\sqrt{V(\hat{\mu})}\right)^2}{V(\hat{\mu})} = Q^2(\hat{\mu}) + 1$



5.2.1 Check

If the values x_i are distributed accordingly to a Gaussian around μ, the value of the Q^2 is distributed accordingly to a χ^2 distribution with $(N-k)$ degrees of fredoom, with k number of parameters.

$$\begin{cases} Exp.1 \to Q^2_{(1)} \\ ... \\ Exp.M \to Q^2_{(M)} \end{cases} \longrightarrow \text{ the union of alla data is the final experiment.}$$

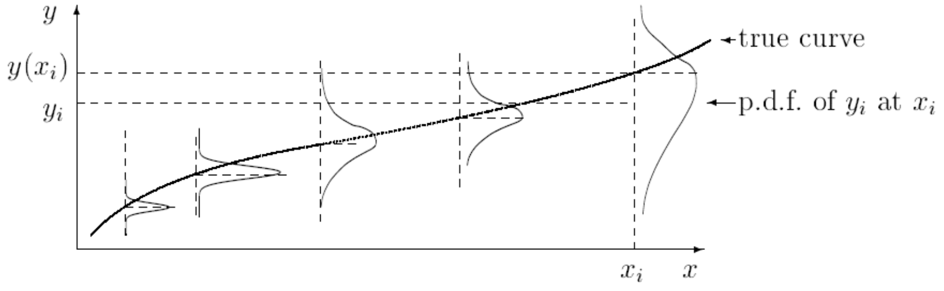
Viceversa, it is possible to divide an experiment into smaller sets and check the distribution of the least squares. This technique is called *bootstrapping.*

5.2.2 The linear model

Let's suppose to have $\{x_i, y_i\}_{i=1,...,N}$ independent measures that follows $y = \varphi(x, \underline{\theta})$. x_i and y_i are random numbers and have their own pdf, but they're not identically distributed, with a joint pdf of (x, y).
It is possible to built a linear model such that $\varphi(x, \underline{\theta}) = \sum_{j=1}^k h_j(x)\theta_j$ with $\underline{\theta}$, which is the parameter of interest, and k, which is the number of parameters.

Each y_i is distributed as $y_i = \varphi(x_i, \underline{\theta}) + \epsilon_i$. It is reasonable to assume the following statements:

- x_i is known without error ($\sigma_{x_i} = 0$);
- the expectation value $E(\epsilon_i)$ is zero such that $y_i = \varphi(x_i, \underline{\theta})$;
- the uncertainty associated with y_i is the one associated with ϵ_i, $\sigma^2_{y_i} = \sigma^2_{\epsilon_i}$;
- any deviation of a point y_i from this curve is due to measurement error or some other unbiased effects beyond our control, but whose distribution is known from previous study of the measuring process.



$$Q^2 = \sum_{i=1}^{N} \frac{[y_i - \varphi(x_i, \underline{\theta})]^2}{\sigma_{y_i}^2} = \sum_{i=1}^{N} \frac{\epsilon_i^2}{\sigma_i^2} = \sum_{i=1}^{N} \frac{\left[y_i - \sum_{j=1}^{k} h_j(x_i)\theta_j\right]^2}{\sigma_i^2}$$

It is possible to express the relation above with matrix notation: Assuming

$$\{x_i\} \to \underline{x} \quad \{y_i\} \to \underline{y} \quad \{h_j(x_i)\} \to H = \sum_{ij} H_{ij} \quad \{\sigma_i\} \to V = \begin{pmatrix} \sigma_1^2 & ... & 0 \\ 0 & ... & 0 \\ 0 & ... & \sigma_N^2 \end{pmatrix} \quad \{\theta_j\} \to \underline{\theta}$$

$$\underline{y} = H \cdot \underline{\theta} + \underline{\epsilon}$$

Skipping all the calculations (see Metzger "Statistical Methods in Data Analysis" § 8.5.2) the estimator and its variance are

$$\begin{cases} \underline{\hat{\theta}} = \left({}^tHV^{-1}H\right)^{-} 1^tHV^{-1}\underline{y} \\ V\left(\underline{\hat{\theta}}\right) = \left({}^tHV^{-1}H\right)^{-1} \end{cases}$$

Properties

- $E(\underline{\hat{\theta}}) = E(\underline{\theta}) = \underline{\theta_t}$: the estimator is unbiased even for a small set of data;
- the parameter does not depend directly on the variance, but only on its relative size (if $V = kW$, k cancels out);
- if ϵ_i are distributed as a gaussian, Q^2 follows a χ^2 distribution with $N - k$ degrees of freedom and $E(Q^2) = N - k$.

It is possible to use the last point above to work on uncertainties: let's assume that there is a misestimation on the variance such as $W = \alpha V \to V = \frac{1}{\alpha}W$.

$$Q^2 = {}^T\left(\underline{y} - H\underline{\hat{\theta}}\right)W^{-1}(\underline{y} - H\underline{\hat{\theta}}) = {}^T\left(\underline{y} - H\underline{\hat{\theta}}\right)\frac{1}{\alpha}V^{-1}(\underline{y} - H\underline{\hat{\theta}}) = \frac{1}{\alpha}{}^T\left(\underline{y} - H\underline{\hat{\theta}}\right)V^{-1}(\underline{y} - H\underline{\hat{\theta}}) = \frac{1}{\alpha}E(Q^2)$$

It is possible to conclude that $Q^2(\hat{\theta}) = \frac{1}{\alpha} \cdot (N - k) \to \alpha = \left(\frac{Q^2(\hat{\theta})}{N-k}\right)^{-1}$ which is the approximation factor of the true variance. The objective is to reduce this value near to 1 as much as possible.

5.2.3 Gauss-Markov Theorem

If $E(\epsilon_i) = 0$ and the covariance matrix of the ϵ_i $V(\underline{\epsilon})$ is finite and fixed, i.e. independent of $\underline{\theta}$ and $\underline{y}$ (it does not have to be diagonal), then the least squares estimate $\underline{\hat{\theta}}$ is *b.l.u.e.*, best linear unbiased estimator. This means that, regardless of the p.d.f. for the ϵ_i, it is the most efficient among the linear estimators. It is important to remember that there could be more efficient estimators, but not linear: therefore, the unbiased behaviour cannot be guaranteed.

The hypothesis condition is called *homoscedasticity* and it means that all the random variables have the same finite variance (so, in this case, a constant variance of $\underline{\epsilon}$ or constant uncertainty of $\underline{y}$).

5.2.4 Linear fit

The least squares technique can be used to perform a linear [1] fit: for example, a function $\varphi(x) = y = A + Bx$ with A, B parameters and x, y measurements taken from $\{x_i, y_i\}_{i=1,\dots,N}$ independent measurements. Let's start from a simple case, assuming that $V(x_i) = 0$, $y_i = \varphi(x_i) + \epsilon_i$ and $E(\epsilon_i) = 0$. If σ_{ϵ_i} are not dependent on A, B or x the hypotheses of the Gauss-Markov theorem hold.
It is now possible to proceed doing the calculations:

$$\underline{y} = \begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix} \quad \underline{\theta} = \begin{pmatrix} A \\ B \end{pmatrix} \quad H = \sum h_i(x_j) = \begin{pmatrix} 1 & x_1 \\ \dots & \dots \\ 1 & x_N \end{pmatrix} \quad V = \begin{pmatrix} \sigma_{y_1}^2 & \dots & 0 \\ 0 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \sigma_{y_N}^2 \end{pmatrix}$$

$$\hat{\theta} = \begin{cases} \hat{A} = \frac{1}{\sum_{i=1}^N \frac{1}{\sigma_i^2}} \sum_{i=1}^N \frac{y_i - \hat{B x_i}}{\sigma_i^2} = \overline{y} - \hat{B}\overline{x} \\ \hat{B} = \frac{1}{\sum_{i=1}^N \frac{1}{\sigma_i^2}} \sum_{i=1}^N \frac{x_i y_i - \hat{A x_i}}{\sigma_i^2} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} \end{cases}$$

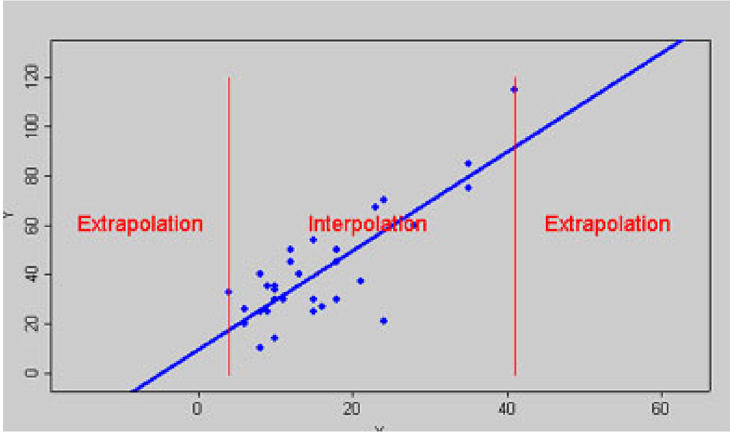
$$V(\hat{\theta}) = \begin{pmatrix} V(\hat{A}) & \text{cov}(\hat{A}, \hat{B}) \\ \text{cov}(\hat{A}, \hat{B}) & V(\hat{B}) \end{pmatrix} = \frac{\sigma^2}{N(\overline{x^2} - \overline{x}^2)} \begin{pmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}$$

Some observations can be made:

- If $\overline{x} = 0$ there is no linear correlation between A and B;
- In normal conditions, $\hat{A}$ and $\hat{B}$ are linearly *anti*correlated;
- If the σ_i are different (Gauss-Markov does not hold), it is necessary to substitute all the means with the weighted ones and replace σ^2 with $\overline{\sigma^2} = \frac{N}{\sum_{i=1}^N \frac{1}{\sigma_i^2}}$;
- A larger $V(x) = \overline{x^2} - \overline{x}^2$ means that the x are spread over a bigger interval; therefore the linear fit can be more precise as $V(\hat{\theta})$ decreases.

5.2.5 Inter&extra-polation

Assuming that a model is uniform all over the possible values, an extrapolation can be done. Let's take the linear fit model explained before: assuming that this model is valid for values greater than the biggest x_i or smaller than the smallest x_i, it is possible to estimate the value of y for unknown x's.



[1] the elements that have to be linear are the parameters, not the x and y values: $\varphi(x) = A + Bx + Cx^2$ is in the linear model case

Nov $\hat{A}$ and $\hat{B}$ have become random variables with their own uncertainties:

$$y_0 = \varphi(x_0) = \hat{A} + \hat{B}x_0$$

$$\sigma^2_{y_0} = \left(\frac{\partial\varphi(x_0)}{\partial\hat{A}}\right)^2 \sigma^2_{\hat{A}} + \left(\frac{\partial\varphi(x_0)}{\partial\hat{B}}\right)^2 \sigma^2_{\hat{B}} + 2\left(\frac{\partial\varphi(x_0)}{\partial\hat{A}} \cdot \frac{\partial\varphi(x_0)}{\partial\hat{B}}\right)\sigma_{\hat{A}\hat{B}} = V(\hat{A}) + x_0^2 V(\hat{B}) + 2x_0 \operatorname{cov}(\hat{A}\hat{B})$$

$$V(y_0) = \frac{\sigma^2}{N}\left(1 + \frac{(x_0 - \overline{x})^2}{\overline{x^2} - \overline{x}^2}\right)$$

The larger N gets and the more spread x_i are, the better y_0 is.
The uncertainty over y_0 is very small if x_0 is close to the mean value $\overline{x}$, and it becomes larger and larger for values that go far from it.

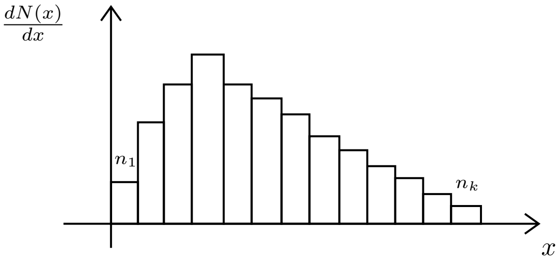
5.2.6 Uncertainties over x_i

What if the x's are known with uncertainties? It is possible to translate the uncertainties of the x-values to the y-values through the slope of the function and use the model explained before with the new uncertainties.

$$\sigma'_y = \sigma_x \cdot \varphi'(x) = \hat{B}\sigma_x \quad \rightarrow \quad Q^2 = \sum_{i=1}^{N} \frac{(y_i - \hat{A} - \hat{B}x_i)^2}{\sigma^2_{y_i} + \hat{B}^2\sigma^2_{x_i}}$$

5.3 Histogram case

The estimator building techniques can be used in the case of histograms, too. Given a set of events $\{x_i\}_{i=1,\ldots,N}$ independent and identically distributed accordingly to $f(x_i, \underline{\theta})$, its associated histogram has k bins w_i, each with n_i elements.
$\Omega_x = \cup_{i=1}^k w_i$, $w_i \cap w_j = \varnothing$ for $i \neq j$, $\sum_{i=1}^k n_i = N$



5.3.1 With least squares

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

π_i is the probability associated to each bin.

It is possible to do some semplifications:

- in the case of an elevated number of bins, $\pi_i \ll 1$ and $(1 - \pi_i) = 1$: $Q^2 = \sum_{i=1}^k \frac{(N\pi_i - n_i)^2}{N\pi_i}$ which is called *Pearson formulation*;
- if the number of event n_i is elevated, it gets closer and closer to $N\pi_i$ going into the poissonian case: $Q^2 = \sum_{i=1}^k \frac{(N\pi_i - n_i)^2}{n_i}$ which is called *Neyman formulation.*

As said before, in the poissonian case each n_i has different uncertaintiy: the Gauss-Markov theorem does not hold (there isn't homoscedasticity). However, the estimator is *b.a.n.*, best asymptotically normal.

5.3.2 With maximum likelihood

The histogram distribution is multinomial:

$$\mathcal{L}(n_i, \underline{\theta}) = \frac{N!}{\prod_{i=1}^{k} n_i!} \prod_{i=1}^{k} \pi_i^{n_i}$$

The minimization is

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = 0 \rightarrow \frac{\partial \prod \pi_i^{n_i}}{\partial \theta_i} = 0 \qquad \frac{\partial l}{\partial \theta_i} = 0 \rightarrow \frac{\partial \sum n_i \ln \pi_i}{\partial \theta_i} = \sum \frac{n_i}{\pi_i} \frac{\partial \pi_i}{\partial \theta_i} = 0$$

This method generates the best asymptotically normal estimator, too. It is preferrable to use the maximum likelihood method for few bins because of the assumptions made in order to simplify the calculations in the least squares case and because of the not-homoscedasticity of the estimator obtained.

Chapter 6

Tests, compatibility and confidence

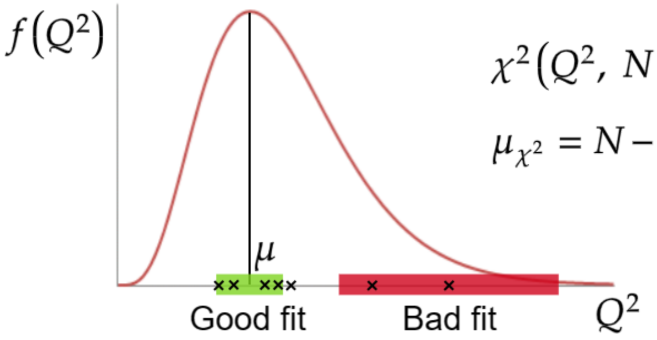
In the previous chapter, methods to determine models and estimate parameters were shown. It is also important, however, to develop tools that are able to determine if there is compatibility between the data and the model built. These tests can be divided into two categories: *goodness-of-fit* tests and *comparing values* tests.

6.1 χ^2 test

Given a set of $\{x_i\}_{i=1,\ldots,N}$ independent measurements and a $\varphi(x)$ model that can describe the data, it is possible to perform a χ^2 test to make some considerations about the compatibility between the model and the data.
Defining

$$\hat{Q}^2 = \sum_{i=1}^{N} z_i = \sum_{i=1}^{N} \frac{(y_i - \varphi(x_i))^2}{\sigma_i^2}$$

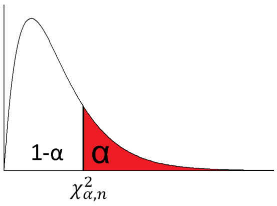
if the residuals $\epsilon_i = y_i - \varphi(x_i)$ are distributed accordingly to a gaussian *pdf*, then Q^2 follows a χ^2 probability density function, with $N - k$ degrees of freedom (N is the number of measurements, k is the number of parameters of the model).



Q^2 can be defined, informally, as a sort of distance between the expected value from the model and the outcome of the experiment.
A wrong model presents values shifted to the right: this means that the integral

$$p = \int_{\hat{Q}^2}^{+\infty} \chi^2(x; N - k)\, dx$$

tells how worse one could do with respect to the experiment that produced $\hat{Q}^2$.
The lower the $p - value$ is, the worse the model describes data, and viceversa.



The χ^2 test leads to the acceptance or the rejection of the model associatd with the measurements. To do so, it is compulsory to define a *Q-trasher*, a limit level that determines if I declare compatibility or not: greater p means accepting the compatibility, lower p the opposite.
The *p-value* depends on data and, therefore, is a random variable with its own probability density function.

$$\begin{cases} y = u(x) \\ x = v(y) \end{cases} \qquad v = u^{-1} \qquad \begin{cases} x = f(x) \\ y = g(y) \end{cases}$$

$$x \to Q^2 \qquad f(x) \to \chi^2$$

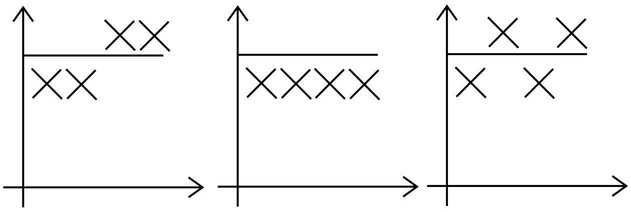
$$y \to p - value \qquad u(x) = 1 - F(x)$$

$$g(y) = f(v(y))\,|v'(y)| = f(v(y))\,\left|[(1 - F(x))^{-1}]'\right| = f(v(y))\frac{1}{f(v(y))} = 1$$

The probability density function of the *p-value* is the uniform one.

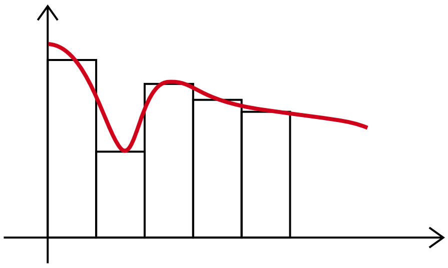
6.1.1 Errors in the test

It is important to understand that the χ^2 test is not universal and error-less.
First of all, the same $\hat{Q}^2$ can be associated to different models that can fit the data better or worse (same Q^2, different goodness).
Then, a probability greater than zero to discard a good fit always exists. There are two types of errors: the false negative (a good fit is rejected) and the false positive (a model is accepted even if it doesn't fit the data properly).
Another aspect to consider is a Q^2 too low: this could be caused by an overestimation of the uncertainties or an overfitting of the data.



6.1.2 Histogram case

This test can be used, for example, with histograms, but only with a decent number of events because of the asymptotical tendence of a binomial distribution to a poissonian and therefore to a gaussian *pdf.*



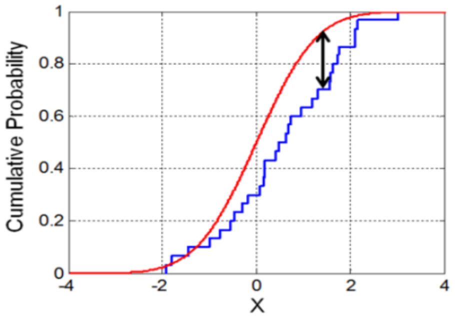
$$\hat{Q}^2 = \sum_{i=1}^{M} \frac{(\pi_i - n_i)^2}{n_i}$$

$$Q^2 = \chi^2(n - k - 1)$$

One degree of freedom is lost because the total number of events is known.

6.2 Kolmogorov-Smirnov test

Another test to study hypotheses is the Kolmogorov-Smirnov. Unlike the χ^2 test, which is built on a probability density function, this test is built on probability itself because it compares a cumulative density function with a step function which is the empirical *cdf*.



$$F(x) = \int_{-\infty}^{x} f(t,\ \theta)dt$$

$$S_n(x) = \frac{1}{n}\sum \theta(x_i)$$

The test takes the largest distance between the two functions and uses it as a parameter called

$$D_n = \sup_n |S_n(x) - F(x)|$$

If we multiply this value with the square root of n we obtain a value that follows a Kolmogorov distribution:

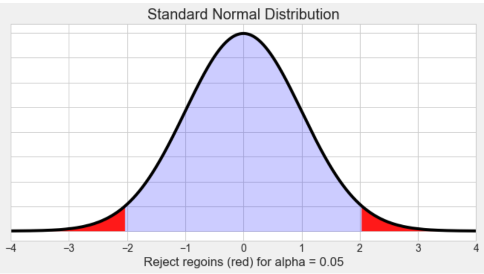
$$d_n = \sqrt{n} \cdot D_n = \sqrt{n} \cdot \sup_n |S_n(x) - F(x)|$$

d_n	1,07	1,22	1,36	1,48	1,65	1,73	1,95
$p - value$	0,20	0,10	0,05	0,021	0,01	0,005	0,001

It is possible to draw the Kolmogorov distribution with *toy-experiment*.
The Kolmogorov-Smirnov test holds even for experiment with few data.

6.3 z-test

The *z-test* is used to compare values with known uncertainty.
Given a set of N measurements $\{x_i\}_{i=1,\dots,N}$ independent and identically distributed, with σ_i of x_i known, the expected value is the mean $\overline{x} = \frac{\sum x_i}{N}$.



Because of the *central limit theorem*, $\overline{x}$ is distributed accordingly to a gaussian distribution

$$\overline{x} = G\left(\overline{x}, \mu, \frac{V}{N}\right)$$

To study the compatibility between $\overline{x}$ and the expected value μ it is possible to define a *p-value*

$$p = \int_{-\infty}^{-|\overline{x}-\mu|} G\left(t, 0, \frac{V}{N}\right) dt + \int_{|\overline{x}-\mu|}^{+\infty} G\left(t, 0, \frac{V}{N}\right) dt$$

and treat is similarly to the previous case.

6.4 t-test

The *t-test* is a test similar to the previous one, but is used when the uncertainty is unknown and needs, therefore, to be calculated from data.

The value

$$t = \frac{\overline{x} - \mu}{\hat{\sigma_{\overline{x}}}} = \frac{\overline{x} - \mu}{\sqrt{\frac{\hat{\sigma}^2}{N}}}$$

is distributed as a gaussian when σ is known, but when the uncertainty is unknown this value is distributed accordingly to the *student* distribution with $\nu = N - 1$ degrees of freedom.

6.5 Comparing two data

In the previous paragraphs, *z-test* and *t-test* were shown to compare a random variable with an exact value. The same tests can be used to compare two random variables, simply by comparing their difference with zero.

6.6 Confidence

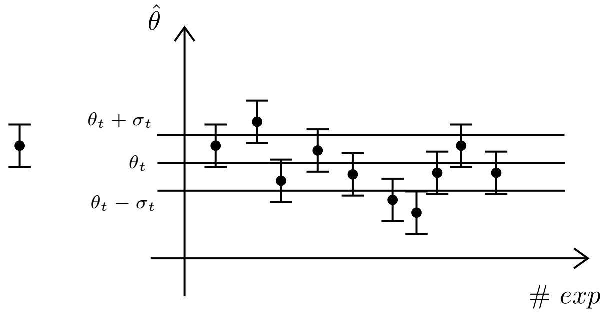
It is possible to determine some statistical relations between the true value θ_t (which is a fixed number) and the estimator $\hat{\theta} \pm \hat{\sigma}_\theta$(which is a random variable with its own *pdf*).

6.6.1 Gaussian case

The first and most common case is the gaussian one. This means that $\hat{\theta}$ is distributed as $G(\hat{\theta}; \theta_t, \sigma_t)$. Two opposed yet correct interpretations can be done:

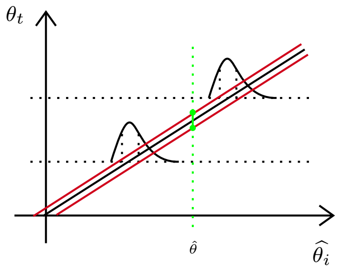
- $\hat{\theta} \in (\theta_t - \sigma_t, \theta_t + \sigma_t)$ 68% of the time: this means that 68% of the points are between the two bands;
- $\theta_t \in \left(\hat{\theta} - \sigma_t, \hat{\theta} + \sigma_t\right)$ 68% of the time: this means that the error bar intercepts θ_t for 68% of the points.

The first case describes a probability; the second case set a *confidence interval* with a related *coverage* (probability of an interval to contain θ_t).

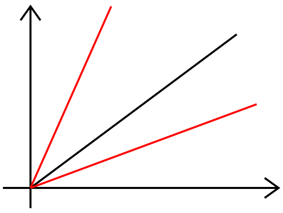


6.6.2 General case

If the probability density function is generic, the same reasoning can be adapted to this case. There is no "right and only" way to find the 68% range: the important thing is to be consistent through the whole process.
In this case, there will be a *confidence belt*, which can be built performing toy experiments.



It is important to observe that, if the probability density function isn't a gaussian with constant sigma, the shape of the belt could become quite complicated (i.e. could diverge).



6.6.3 Discovery significance

Considerations about significance are important to determine the significance of a discovery.
Supposing a counting experiment, this includes two poissonian distributions which are going to mix together (one for the background f_b, one for the signalf_s). In case of signal presence, $E(s+b) = \nu_s + \nu_b$.

Supposing N_0 events, it is possible to compare this value with the distribution of the background, running a sort of *p-value* called *confidence level*:

$$P = \sum_{N_0}^{+\infty} f_b(N_0, \nu_b) = 1 - \sum_{0}^{N_0} f_b(N_0, \nu_b)$$

The smaller this value is, the least probable it becomes for the event to be from the background; an high confidence level means that the presence of the signal is very probable.