Chapitre 1

Introduction to probability

1 Random experiment

An experiment is said to be random if we could not predict its result in advance. A result of such experiment can be considered as un element ω of a set including all possible results. This set is called the sample space Ω.

The nature of the elements of Ω is not unique. It depends on the usage we make of the results of the experiment :
Example : 2-die game In the the 2-die game the set Ω can be that of the the different possible couples :

$$\{(1,1),(1,2)\cdots\cdots(6,6)\}$$

or that of the sum of the two dies :

$$\{(2),(3)\cdots\cdots(12)\}$$

Event :
An event is a logical proposition related to the result of on experiment

Example : Sum of the 2 dies > 7

$$\Rightarrow \quad \{(4,4),(4,5),(5,5),(4,6),(5,6),(6,6)\}$$

Kolmogorov probability

A probability is an application P from the set of events of Ω into [0,1] such that :

– $P(\Omega) = 1$

– for all sets of incompatibles events :

$$A_1, A_2, \ \ A_n, \ \text{ on a } \ P(U_i A_i) = \sum_i P(A_i)$$

Proprieties

– $P(\phi) = 0 \qquad and \qquad P(\bar{A}) = 1 - P(A)$
– $P(A) \leq P(B)$ if $A \subseteq B$
– $P(AUB) = P(A) + P(B) - P(A \cap B)$
– $P(U_i A_i) \leq \sum_i P(A_i)$
– if $\{B_i\}$ forms a complet system of events then :

$$\forall A : \qquad P(A) = \quad \sum_i \quad P(A \cap B_i)$$

Conditional Probability

$P(A/B)$: probability to have A if B is known/realized

$$P(A/B) \ = \ \frac{P(A \cap B)}{P(B)}$$

Conditional probability satisfies the probability definition requirements :

– $P(\Omega/B) = 1$
– $P(U_i A_i/B) = \sum_i P(A_i/B)$

Independance of events :

A et B are independant if

$$P(A/B) = P(A) \Rightarrow \ \frac{P(A \cap B)}{P(B)} \ = P(A)$$

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

Bayes formulae

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

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

$$P(A) = \sum_i P(A \cap B_i) \qquad \{B_i\} \qquad \text{système complet}$$

$$= \sum_i P(A/B_i)P(B_i)$$

$\Rightarrow$

$$P(B_i/A) = \frac{P(A/B_i)P(B_i)}{\sum\limits_k P(A/B_k)\, P(B_k)}$$

Example : The physics department of our University has bought 100 PC of 3 different marks. Each has a known failure rate C

mark	nombre	ϵ
m_1	30	2%
m_2	50	2%
m_3	20	3%

The PC attributed to prof. X was found corrupt. Can prof. X infer what is the mark of his PC ?

$$P(m_1/\text{corrupt}) = \frac{\frac{2}{100}\times\frac{30}{100}}{\frac{2}{100}\times\frac{30}{100}+\frac{2}{100}\times\frac{50}{100}+\frac{3}{100}\times\frac{20}{100}} = \frac{6}{22}$$

$$P(m_2/\text{corrupt}) = \frac{10}{22}$$

$$P(m_3/\text{corrupt}) = \frac{6}{22}$$

Knowing the failure rate (prior) of the different marks and the number of PCs of each mark Bayes formulae allow one to estimate the probability (posterior)of one corrupt PC to belong to one mark.

If the companies do not provide the failure rates what one can do ? We will see later how one can proceed.

Probability notions

Theoritical Concept
If one has a finite set of events and symmetry such that each elementary event has the same probability then

$\Rightarrow$ Probabilty is a counting business

$$P(A) = \frac{\text{nb of favorable cases}}{\text{nb all possibe cases}}$$

However perfect symmetries are rare and sets are not always finite. Bertrand paradox is an illustration of this :
Bertrand paradox

Let's take a circle of radius r and draw an equilateral triangle inside it. Now let's estimate the probability of any segment to be longer than the side of the triangle ?

Practical concepts :
Two notions of probability are used by physicists in practice ; one is called the objective vision and the other is called the subjective one

The objective vision or the frequentist :
this notion is based on the large numbers law. If one repeats a large number of times the same experiment, the appearance frequency of one kind of events defines the associated probability.
Critics :

- This vision could not "probabilize" the rare events. Example : What is the probability that it will snow in the Nevada desert on the 13th of August 2020 ?
- The frequentist notion is based on the law of large number which is valid only if a notion of probability exists already ! ! ! !

The subjective vision or the Bayesian one :
Since the frequentist notion is limited in its application, the subjective one

tries to enlarge its scope by using Bayes theorem :

$$P(A/B) \quad = \quad \frac{Pr(B/A)\ P(A)}{P(B)}$$

So if one has recorded the weather in Nevada during longtime and more particularly on the 13th of August of each year one may be able to predict the probablity to snow in the Nevada desert on the 13th of August 2020.
The probability of an event is subject to the information we acquire. Our knowledge (and hence our uncertainty) evolves with time.

Critics :

The Bayesian probability depends on an arbitrary choice of the prior (the failure rate in our PCs example) which is generally unkown in most of our physics application. So this probability is dependant on the observer choice. So what to choose ? This is the big question....

Random variable (X)

This is an application from a set Ω to which a probability law P is defined to an another set E. The application allows to link each element of E to one or more events of Ω. This allow indeed to a transfer of the probability law defined on Ω to E.
Example : The 2-die game (application : sum of the two dies)

$\Omega = \{(1,1),(1,2)\ \cdots\ (6,6)\}$ with the probability P defined as

$$P(\omega) = \frac{1}{36}\ \ \forall \omega \in \Omega$$

$E = \{2,3 \qquad 12\}$ sum of the two dies.
For the element $s = 7 \in E$ we can associate the probablity :

$$P_X(s=7) = P(1,6),(2,5),(3,4),(6,1),(5,2),(4,3) = \frac{6}{36}$$

The application allowed to transport the probability defined on Ω to E
$P \to P_X$
If $E \equiv \mathbf{R}$ the random variable is called a real random variable.

$$P_X(A)_{A\in\mathbf{R}} = P(\omega / X(\omega) \in A) = P(X^{-1}(A))$$

The probability density function (p.d.f)

If X is a discreet variable $\Rightarrow$ we talk about Probability

If X is a continuous one $\Rightarrow P(X = x) = 0$

$\Rightarrow$ in this case we introduce on the notion of the probability density

$$P(x_{min} < x < x_{max}) = \int_{x_{min}}^{x_{max}} f(x)dx$$

The distribution function (d.f)

$$F(x) = P(X < x)$$

F is a montonous left-continous function.

if X is a continous variable $\Rightarrow$

$$F(x) = \int_{x_{min}}^{x} f(x')dx'$$

$$F(x_{min}) = 0 \qquad , \qquad F(x_{max}) = 1$$

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

F is as important as f (if not more).

Important Application

Let's consider the following case : x is a real variable and $\phi(x)$ is a derivable function of x.
if f, F are respectively the probability density function (p.d.f) and the distribution function (d.f) associated to x, then what are the p.d.f and the d.f associated to ϕ called respectively (g, G) ?

Answer :
- Case of a bijective ϕ

In this case ϕ is monotonous.
1- ϕ is an increasing function $\rightarrow$ $\qquad \phi' > o$

$$\begin{aligned} F(x) &= G(\phi(x)) \text{ since } P(X < x) = P(\phi(X) < \phi(x)) \\ f(x) &= g(y)\, \phi'(x) \Rightarrow g(y) = \tfrac{f(x)}{\phi'(x)};\ y = \phi(x) \end{aligned}$$

2- ϕ is a decreasing function$\rightarrow$ $\qquad \phi' < o$

$$\begin{aligned} F(x) &= 1 - G(\phi(x)) \text{ since } P(X < x) = P(\phi(X) > \phi(x)) \\ f(x) &= -g(y)\, \phi'(x) \Rightarrow g(y) = \tfrac{f(x)}{-\phi'(x)};\ g = \phi(x) \end{aligned}$$

For both cases one may write :

$$g(y) = \frac{f(x)}{|\phi'(x)|}$$

• General case
If ϕ is not bijective then we divide the definition domain into intervals on which the function is either increasing or decreasing and then we apply the previous recipe.
exercice : Find the disribution function associated to : $\phi(x) = x^2$

Moments
The moments play an important role in the statistics. they allow to characterize our sample. They are defined by :

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

The expectation value :
This is a moment of the first order ($m = 1$) :

$$m = 1 \Rightarrow E(x) = \mu = \int_{-\infty}^{+\infty} x f(x) dx$$

Central moments

These are moments centered around the mean value μ

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

The variance :
This is a central moment of the second ordre ($m = 2$) :

$$m = 2 \Rightarrow V = \sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x) dx$$

In addition to the expectation value and the variance, two other moments can be sometimes of interest :

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

Called the asymmetry coefficient (skewness). The other is :

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

Called flatness coefficient (kurtosis). Skewness expresses the deformation of a pdf with respect to a gaussian $\gamma_1 > 0 (< 0)$ deformation to the right (left). The kurtosis expresses the flatness of a pdf with respect to a gaussian $\gamma_2 > 3 (< 3)$ sharp (broad) than a gaussian

Characteristic function :
This is the inverse Fourier Transform of the pdf :

$$\begin{aligned}\phi(t) &= E[e^{itx}] = \int_{-\infty}^{+\infty} e^{+itx} f(x)dx \\ &\quad \text{or} \left(\sum_{k=-\infty}^{+\infty} e_k^{itx} f(x_k) \ \text{ discreet case} \right)\end{aligned}$$

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-itx} \phi(t) dt$$

$$TF \quad \longleftrightarrow \quad TF^{-1}$$

$$\begin{aligned}\phi(t) &= E\left[\sum_{k=0}^{\infty} \frac{(itx)^k}{k!} \right] \\ &= \sum_{k=0}^{\infty} \frac{(it)^k}{K!} E[x^k] \\ E[x^k] &= \frac{1}{i^k} \frac{d^k \phi(t)}{dt^k}|_{t=0}\end{aligned}$$

Discreet case : the characteristic function is called generator function in this case and commonly noted : $G(z)$.

Application :

We can use the proprieties of the characteristic function in order to find the pdf of the the sum of two real variables $f(x+y)$:
We define $w = x + y \ \Rightarrow \ \phi(t) = E[e^{iwt}] = E(e^{itx} e^{ity})$
if x, y independant $\Rightarrow \ E(e^{iwt}) = E(e^{ixt}) E(e^{iyt}) \Rightarrow \ \phi_w(t) = \phi_x(t) \ \phi_y(t)$
now once $\phi_w(t)$is found$f(x+y)$can be determined by taking $TF^{-1}$of ϕ.

Case of multiple random real variables
Let's consider the case of two random real variables x, y to start with. The couple (x, y) is an application from (Ω) in R^2.

The adjoint density function :
$f(x, y)$ is defined in the following manner :

$$P(x < X < x + dx, \ y < Y < y + dy) = f(x, y) \ dx \ dy$$

The marginal density function :

If one is interested by the behaviour of only one of the two variables, the marginal density function can be use. It is obtained by integrating on all the values of the second variable :

$$\begin{aligned} f_1(x) &= \int f(x,y)\, dy \\ f_2(y) &= \int f(x,y)\, dx \end{aligned}$$

The conditional density function :
This it the density probability function associated to one variable when the other is known/fixed. It is given by :

$$f_c(y/x) = \frac{f(x,y)}{f_1(x)}$$

Correlation :
One of the most important aspects when analyzing data is to have a good understanding of the different variables describing those data. The correlation provide as much information as the variables themselves and should be thus quantified correctly. To do this we introduce the notion of covariance :

The covariance :
The covariance of 2 variables x, y is defined by :

$$\begin{aligned} COV(x,y) &= E[(x-\mu_x)(y-\mu_y)] \\ &= E[xy] - \mu_x\mu_y \end{aligned}$$

with

$$\mu_x = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x f(x,y)\, dx\, dy$$

$$\mu_y = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} y(f(x,y)\, dx\, dy$$

The correlation coefficient :
From the covariance we can introduce the correlation coefficient :

$$\rho_{xy} = \frac{COV(x,y)}{\sigma_x\sigma_y} \qquad -1 \leq \rho_{xy} \leq 1$$

Independance of 2 random variables
The correlation notion is different from the independence one. One can ea-

sily show that the independence of two variables leads to the absence of correlation between them. The opposite is not true. Indeed the absence of correlation does not imply their independence :

x, y independent $\Rightarrow \rho_{xy} = 0$

$\rho_{xy} = 0 \quad \not\Rightarrow x, y$ independent

Indeed,if x, y are independant we have $f(x,y) = f(x)f(y)$ which leads to : $COV(X,Y) = \int (x-\mu_x) \int (y-\mu_y) f(x,y)dxdy = \int (x-\mu_x)f(x) \int (y-\mu_y)f(y)dy = 0$
The covariance may vanish even if $f(x,y) \neq f(x)f(y)$. This occurs when the variation of one of the two variables does not affect the expectation value of the other.

Remark : If the two variables x, y are correlated ($COV(x,y) \neq 0$) one can always perform a change of variable which results in two new uncorrelated variables.

$$X, Y \Rightarrow x', y' \quad \text{such that} \quad COV(x', y') = 0$$
$$x' = x'(x,y), \;\; y' = y'(x,y) \Rightarrow g(x', y') = f(x,y)J$$

where we introduced the Jacobian J

$$J = \begin{vmatrix} \frac{\partial x}{\partial x'} & \frac{\partial y}{\partial x'} \\ \frac{\partial x}{\partial y'} & \frac{\partial y}{\partial y'} \end{vmatrix}$$

Remarks :
1- The variable change in case of a discreet variable is not accompanied by a Jacobian multiplication.

2- Idem in case of a parameter change.

Generalization to the case of many variables

$$\underline{x} = \begin{pmatrix} x_1 \\ x_2 \\ \\ x_n \end{pmatrix} \qquad \underline{x}^T = (x_1, x_2, \cdots x_n)$$

$$F(\underline{x}) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_n} f(\underline{x})\, d\underline{x}, \quad d\underline{x} = dx_1\, x_2 \cdots dx_n$$

$$f(\underline{x}) = \frac{\partial^n}{\partial x_1, \partial x_2 \cdots \partial x_n} F(\underline{x})$$

$$\mu_{\ell 1, \ell 2 \cdots \ell_n} = E(x_1^{\ell 1}, x_2^{\ell 2}, \cdots x_n^{\ell n})$$

$$\underline{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}$$

$$\underline{V} = E\left[(\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})^T\right] = \begin{pmatrix} \sigma_{11} & \sigma_{12} \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \sigma_{2n} \\ \vdots & & \\ \sigma_{n1} & & \sigma_{nn} \end{pmatrix}$$

$$\sigma_{ij} = \rho_{ij}\, \sigma_i\, \sigma_j$$

V is a symmetric matrix $\Rightarrow$ diagonalisable.

$$\Rightarrow \exists\, U \text{ such that } g = U\, x \quad \text{avec} \quad V(g) = \text{diag.}$$

Chapitre 2

Probability distributions

When analyzing data collected from experiments, we try to understand the behavior of some or all of the physical quantities which can characterize the events we are interested in. Usually we try to compare the distribution of those physical quantities to some "appropriate" distributions we use as a reference. Agreement or deviation with respect to those references play an important role to understand our data.

In high energy and nuclear physics the distributions we may meet are numerous. Here after some which are frequently used :

Bernouli

This is the simplest of the known distributions. It gives the probability of the discreet variable $X = k$ which can take only two discreet values 1 or 0 with the probability p to have the value 1 :

$$\begin{aligned} f(k,p) &= p^k(1-p)^{1-k} \\ E[k] &= p \\ V[k] &= p(1-p) \end{aligned}$$

Binomial distribution

It describes the probability of having $X = k$ favorable trials among n ones. with p being the probability of success for each trial :

$$\begin{aligned} f(n,k,p) &= B(k;n;p) = C_n^k p^k (1-p)^{n-k} \\ E[k] &= np \\ V[k] &= np(1-p) \\ \phi(t) &= (p\mathrm{e}^{it} + 1 - p)^n \end{aligned}$$

Exercice
Show that the sum of two independent variables x, y distributed respectively according to : $B(k_x; n_x, p)$ and $B(k_y, n_y, p)$ is a new variable distributed like $B(k_x + k_y, n_x + n_y, p)$.

Multinomial distribution

This is a generalization of the binomial distribution with m possible results(2 in the binomial). The following table illustrates the possible configurations of such distribution :

result	nb	prob.
1	k_1	p_1
2	k_2	p_2
$\cdots$	$\cdots$	$\cdots$
m	k_m	p_m

with $\sum_n p_m = 1, \quad \sum_m k_m = n$ and the probability associated to one configuration is given by :

$$f(k_1, k_2, \cdots \quad k_m, p_1, p_2 \cdots \quad p_m, n) = \frac{n!}{k_1! k_2! \cdots k_m!} p_1^{k_1} p_2^{k_2} \cdots \quad p_m^{k_m}$$

$$\begin{aligned} E[k_i] &= np_i \\ V[k_i] &= np_i(1 - p_i) \\ \mathrm{COV}(k_i, k_j) &= -np_i p_j \qquad i \neq j \\ \phi(t_2, t_3 \cdots t_m) &= (P_1 + P_2 e^{it2} + \cdots P_m e^{itm})^n \end{aligned}$$

Application :
1)The decay of one particle in different modes ;
2)The bins of a histogram.

Poisson distribution

This is one of the most frequent distributions. It is a discreet one. This distribution can be used to describe a phenomenon if it satisfies the following conditions :

– The number of success (called events) is known but not the number of trials ;
– The number of success in a given interval depends only on the length of this interval (there is a constant rate of events/unit of interval..) ;
– The occurrence of an event could not alter the occurrence of another event (uncorrelated events) ;
– 2 events could not occur at the same time (too scarce to have a coincidence of two events).

Those conditions allow us to derive the Poisson probability distribution starting from the binomial one.
Let Δx be an interval in which one event at most can take place. Let λ be the probability to have one event in $n\Delta X \Rightarrow$ the probability to have 1 event in ΔX will be (second condition) $p = \frac{\lambda}{n}$

Now let's estimate the probability to have k events in $n \gg k$ intervals ($n\Delta x$). It is easy to understand that this can be estimated using the binomial distribution with k success and n tries with the probability associated to one success given by : $p = \frac{\lambda}{n}$:

$$P(k,p) = \frac{n!}{(n-k)!k!} \quad p^k(1-p)^{n-k}$$

or

$$k \ll n \quad \text{(rare events)} \quad \Rightarrow \frac{n!}{(n-k)!} \simeq n^k$$

$$p = \frac{\lambda}{n} \Rightarrow \left(1 - \frac{\lambda}{n}\right)^{n-k} = e^{-\lambda}$$

$$\rightarrow \qquad f(k,\lambda) = \frac{\lambda^k}{k!}\, e^{-\lambda}$$

$$\begin{aligned} E[k] &= \lambda \\ V[k] &= \lambda \\ \gamma_1 &= \lambda^{-\frac{1}{2}} \\ \gamma_2 &= \lambda^{-1} + 3 \\ \phi(t) &= \exp(\lambda(e^{it} - 1)) \end{aligned}$$

Exercice :
Show that the sum of two random variables x, y each of them with a Poisson distribution is also a variable with a variable de Poisson according to the following scheme :

$$
\begin{aligned}
X &\rightarrow \frac{\lambda_x^{K_x}}{K_x!} e^{-i\lambda_x} \\
Y &\rightarrow \frac{\lambda_y^{K_y}}{K_y!} e^{-\lambda_y} \\
X+Y &\rightarrow \frac{(\lambda_x+\lambda_y)^k}{K!} e^{-(\lambda_x+\lambda_y)}
\end{aligned}
$$

Application :

1) Nuclear decay (big number of nuclei) ;
2) Interaction of an intense beam with a thin target

Uniform distribution This is the simplist continuous distribution :

$$
\begin{aligned}
f(x;a,b) &= \frac{1}{b-a} \qquad x \in [a,b] \\
f(x;a,b) &= 0 \qquad \text{else}
\end{aligned}
$$

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

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

Application :

1) The round-up of a number ;
2) The coordinate of a particle impact in a pixel.

Normal distribution (De Moivre, Laplace, Gauss)

This is the most known and used distribution. We will see very soon why this distribution plays such an important role

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

$$
\begin{aligned}
E[x] &= \mu \\
V[x] &= \sigma^2
\end{aligned}
$$

By introducing the following variable change :

$$
x \rightarrow z = \frac{x-\mu}{\sigma}
$$

We obtain the so-called standard normal distribution which has zero as a mean value and a variance equals to 1 : $N(x;\mu,\sigma) \rightarrow N(z;0,1)$. The characteristic function is given by :

$$\phi(t) = e^{-\frac{t^2}{2}}$$

Normal distribution with n variables

$$x \to \underline{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \qquad \mu \to \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}$$

$$-\frac{(x-\mu)^2}{2\sigma^2} \to -\frac{1}{2}(\underline{x}-\underline{\mu})^T\ A(\underline{x}-\underline{\mu})$$

To determine the matrix A we use the fact that $E\left[\underline{x}\ -\underline{\mu}\right] = 0$

$$\int (\underline{x}-\underline{\mu})\ e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T A(\underline{x}-\underline{\mu})}\ d\underline{x} = 0$$

We then differentiate with respect to $\underline{\mu}$

$$\Rightarrow \qquad E[(\underline{x}-\underline{\mu})\ (\underline{x}-\underline{\mu})^T]A = \mathbf{1}$$

(covariance matrix) $\leftarrow VA = \mathbf{1} \Rightarrow \qquad A = V^{-1}$

$$N(\underline{x};\underline{\mu};V) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}}\ \exp\ \left[-\frac{1}{2}(\underline{x}-\underline{\mu})^T\ V^{-1}(\underline{x}-\underline{\mu})\right]$$

where $|V|$ is the determinant of V

Application : Normale distribution with a 2 dimensions (x, y)

$$V = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \qquad \mathrm{COV}(x,y) = \rho\sigma_x\sigma_y$$

$$V^{-1} = \frac{1}{\sigma_x^2\sigma_y^2(1-\rho^2)} \begin{pmatrix} \sigma_y^2 & -\rho\sigma_x\sigma_y \\ \\ -\rho\sigma_x\sigma_y & \sigma_x^2 \end{pmatrix}$$

The p.d.f is given by : $f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\ \mathrm{e}^{-\frac{1}{2}G}$

$$G = \frac{1}{1-\rho^2}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right]$$

Remark : The contours of the pdf related to a given density of this distribution in the plan (x, y) can be obtained by fixing the value G. Those contours are ellipses. ρ **Cauchy distribution**

$$C(x,\mu,\alpha) \quad = \frac{1}{\pi\alpha}\ \frac{1}{1+(x-\mu)^2/\alpha^2}$$

Cette distribution is well known in nuclear and high energy physics (up to a multiplicative constant) under the name Breit-Wigner which is used to describe resonances.

$$f(m;M,\Gamma) = \frac{1}{2\pi}\ \frac{\Gamma}{(m-M)^2+\frac{\Gamma^2}{4}}$$

Remark : All the moments related to this distribution ($E[x],\cdots$) are divergent. In practice the distribution is truncated between $-L$ et $+L$ with $L \gg \alpha$.

Gamma distribution

For phenomena for which events fulfill the Poisson distribution and for which the probability to have an event per unit of interval is $\mu = \lambda/t$, the probability to have the event number k taking place at t starting from zero can be estimated from the following :

$$F(t) = P(T_k \leq t) = 1 - P(T_k > t)$$

where T_k is the occurrence time of the event number k.

$$P(T>t) \equiv \text{ (Prob. of number of decays } < k)$$

$$= \sum_{m=0}^{k-1} \frac{(\lambda t)^m\ \mathrm{e}^{-\lambda t}}{m!}$$

$$= \int_{\lambda t}^{\infty} \frac{z^{k-1}\mathrm{e}^{-z}}{(k-1)!}\, dz$$

$$\Rightarrow \qquad F(t) = 1 - \int_{\lambda t}^{\infty} \frac{z^{k-1}}{(k-1)!}\ \mathrm{e}^{-z}\, dz \ = \int_0^t \frac{\lambda^k y^{k-1}}{\Gamma(k)}\ \mathrm{e}^{-\lambda y}\, dy\ ; y = \frac{z}{\lambda}$$

$$f(t;k;\lambda) = \frac{dF}{dt} = \frac{\lambda^k t^{k-1} e^{-\lambda t}}{\Gamma(k)} \quad ; t > 0$$

This distribution is called the Gamma distribution.

$$E[t] \ = \ \tfrac{k}{\lambda}$$

$$V[t] \ = \ \tfrac{k}{\lambda^2}$$

$$\phi(x) \ = \ \tfrac{1}{(1-\frac{ix}{\lambda})^k}$$

Remarks :
1- In the case of $k = 1$ the distribution is called the exponential distribution $f(t) = \lambda e^{-\lambda t}$;
2- If $\lambda = \frac{1}{2}$ and $k = \frac{n}{2} \Rightarrow \mathrm{f}\left(t; \frac{n}{2}, \frac{1}{2}\right) \equiv \chi^2(t; n)$

χ^2 distribution

If $x_1, x_2, x_3 \cdots \ x_n$,are n independent variables normally distributed then we can define a new variable :

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

where μ_i, σ_i^2 are the mean value and the associated variances of x_i.

Case of n=1 : $\chi^2(1)$

$$\begin{aligned} \chi^2 &= \left(\tfrac{x-\mu^2}{\sigma^2}\right) = z^2 \\ \text{we define } Q &= \chi^2 = z^2 \end{aligned}$$

The p.d.f associated to z (normally distributed) is given by :

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

From this we can deduce the one associated to Q.

$$\begin{aligned} P(Q < q) &= P(-\sqrt{q} < z < \sqrt{q}) \\ G(q) &= F(\sqrt{q}) - F(-\sqrt{q}) \\ g(q) &= \tfrac{f(\sqrt{q})}{2\sqrt{q}} + \tfrac{f(-\sqrt{q})}{2\sqrt{q}} \\ g(Q) &= \tfrac{f(\sqrt{Q})}{2\sqrt{Q}} + \tfrac{f(-\sqrt{Q})}{2\sqrt{Q}} \end{aligned}$$

$$g(Q) = \frac{1}{\sqrt{2\pi Q}} e^{-\frac{Q}{2}} = \frac{1}{\sqrt{2\pi\chi^2}}\ e^{-\frac{\chi^2}{2}}$$

Unfortunately we use the same name χ^2 for the p.d.f and the variable itself :

$$\chi^2(n = 1) = \frac{1}{\sqrt{2\pi\chi^2}}\ e^{-\frac{\chi^2}{2}}$$

The previous result can be generalized to the case of n variables :

$$\chi^2(n) \to g(\chi^2, n) = \frac{(\chi^2)^{\frac{n}{2}-1}}{\Gamma(\frac{n}{2})2^{n/2}}\, e^{-\frac{\chi^2}{2}}$$

$$\begin{aligned} E[\chi^2(n)] &= n \\ V[\chi^2(n)] &= 2n \\ \phi(t) &= (1-2it)^{-n/2} \end{aligned}$$

Remarks :
1-The sum of two variables distributed according to $\chi^2(n_1)$ and $\chi^2(n_2)$ respectively is a $\chi^2(n_1+n_2)$ distributed.
2- The asymptotic limit of $\chi^2(n)$ for large n is N(n, 2n).

Student distribution

The p.d.f associated to the variable $t = \frac{z}{\sqrt{\frac{\chi^2(n)}{n}}}$ is the student distribution if z is distributed normally. The p.d.f is shown to be

$$f(t,n) = \frac{1}{\sqrt{\pi n}}\, \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}\, \frac{1}{(1+\frac{t^2}{n})^{\frac{n+1}{2}}}$$

We will see shortly the utility of this distribution when we will study the estimator of the variance. variance.

Chapitre 3

The central-limit theorem

Let $X_1, \cdots, X_i, \cdots X_n$ be n independent random variables having respectively $m_1, , \cdots m_i, \cdot m_n$ as expectation value, $\sigma_{12}, \cdots, \sigma_i^2, \cdots, \sigma_n^2$ as variance and f_i as p.d.f (not necessarily identical). If F_i is the distribution function associated to $(X_i - m_i)$ and $S_n^2 = \sum_{i=1}^{n} \sigma_i^2$ then :

$$\sum_{i=1}^{n} \frac{(X_i - m_i)}{S_n} \lim_{n\to\infty} U = N(X; 0, 1)$$

if the condition of Lindeberg-Cramer is satisfied :

$$\lim_{n\to\infty} \left[\frac{1}{S_n^2} \sum_{i=1}^{n} \int_{|x|>\epsilon\ S_n} x^2 dF_i(x) \right] = 0$$

This condition is better put under the following form : None of the variables is dominating the others and the contribution of all variables is uniformly small. c

Special case demonstration

If the X_i are all identical with a common expectation value m and a common variance σ^2 then :

$$X = \frac{1}{\sqrt{n}} \left(\frac{X_1 + X_2 + \cdots + X_n - nm}{\sigma} \right) = \frac{1}{\sqrt{n}} \sum_i \frac{(X_i - m)}{\sigma} = \sum_i^n \frac{X_i - m}{\sqrt{n}\sigma} = \sum_i^n x_i$$

is the sum of n centered variables (expectation 0) and a variance $= \frac{1}{n}$ each of them has for a characteristic function : $\phi_{xi}(t) = E(\mathrm{e}^{itx_i}) = 1 + 0 - \frac{t^2}{2n} + 0\left(\frac{1}{n^2}\right)$ The characteristic function of the sum of those independent variables is $\phi_X(t) = \prod_{1=i}^{n} \phi_{xi} = \left(1 - \frac{t^2}{2n}\right)^n$. If $n \to \infty$ then we have : $\phi_X(t) \to \exp\left(-\frac{t^2}{2}\right)$ which is the characteristic function of $N(X, 0, 1)$.

Chapitre 4

statistics

The goal of statistics is to find the characteristics of a given population (Ω) using a sample of this population. Many tools are used to attain this goal. Here we will see how one can evaluate those tools and which to choose in a given situation

Statistics :
Statistics can be defined mathematically as functions of the observations of a sample which do not depend upon the unknown characteristics of the population.

Statistics tools :

In practice, the results of a statistical study allow to evaluate some characteristic parameters. This is often cast in the following form : $\theta = A \pm B$. We will learn later what this statement means in details. However we can anticipate by saying that A is a value we think to be as close as possible to the true one. The interval $[A-B, A+B]$ can be used (following some rules to be explained later) to quantify our belief/confidene on A. In order to reach this statement we need to understand the tools we use to obtain it :

2 Estimators

These are statistics which allow to estimate one parameter of the population θ. the estimate will be noted $\hat{\theta}$.

Example :

One may ask how we can estimate the mean value μ of the population ? There are indeed many possibilities :

- $\frac{1}{n}\sum_{i=1}^{n} x_i$
- $\sqrt[n]{\prod_{i=1}^{n} x_i}$
- $\frac{max(x_i)+min(x_i)}{2}$
- "$\frac{1}{n}\sum_{i=1}^{n} x_i$" after eliminating 5% of both extremities.
- The median : the value x_0 for which $F(x_0) = 1/2$ (when possible)

The question which follows : Which one to choose ?

Estimator's proprieties

• **The bias criterium :**

$$b(\hat{\theta}) = E[\hat{\theta}] - \theta_t$$

$\hat{\theta}$ is a function of random variables so it is a random variable itself $\Rightarrow E[\hat{\theta}]$ is its expectation value. θ_t is the true value of the parameter associated to the population.
The estimator is said to be unbiased if :

$$b(\hat{\theta}) = 0$$

and asymptotically unbiased if :

$$\lim_{n-\infty} b_n(\hat{\theta}) = 0$$

where n is the number of events used in the estimation.
Example 1 :

The sample mean is an unbiased estimator of the true mean of the population :

$$E\left[\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}E\left[\sum_i x_i\right] = \frac{1}{n}\sum_i E[x_i]$$

$$= \frac{n}{n}E[x] = \mu \Rightarrow E[\hat{\mu}] = \mu$$

Example 2 :
We will study two estimators of the variance. The first is used when the true value of the mean is known while the other is used when the mean value is only estimated from the same sample :

- variance with known μ :

In this case we propose the following estimator

$$S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Let's take the expectation value of this estimator :

$$\begin{aligned} E[S_1^2] &= \tfrac{1}{n}\, E\left[\sum_{i=1}^{n}(x_i-\mu)^2\right] = \tfrac{1}{n} E\left[\sum_{i=1}^{n}(x_i^2 - 2x_i\mu + \mu^2)\right] \\ &= \tfrac{1}{n}\, E\left[\textstyle\sum_i x_i^2 - 2\mu\sum_i x_i + n\mu^2\right] \\ &= \tfrac{1}{n}\left(n\, E[x^2] - 2\mu \times nE[x] + n\mu^2\right) \\ &= \tfrac{1}{n}\left(n\, E[x^2] - 2n\mu^2 + n\,\mu^2\right) \\ &= \tfrac{1}{n}\left(n\, E[x^2] - n\,\mu^2\right) = \tfrac{n}{n}\sigma^2 = \sigma^2 \end{aligned}$$

Hence S_1^2 is an unbiased estimator of the variance σ^2 if μ is known.

- variance with unknown μ :

Since μ is unknown we replace in S_1^2, μ by its estimator :

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

We rename the new estimator S_2^2 :

$$S_2^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

$$E[S_2^2] = \frac{1}{n}\, E\left[\sum_i x_i^2\right] - E\left[\left(\frac{\sum_i x_i}{n}\right)^2\right]$$

Since $\sigma^2 = E[x^2] - \mu^2$ and $V\left[\sum x_i\right] = E\left[\left(\sum x_i\right)^2\right] - \left(E\left[\sum_i x_i\right]\right)^2$

$$E[S_2^2] = \frac{1}{n}\,(n(\sigma^2+\mu^2) - \frac{1}{n}\,\left(V\left[\sum x_i\right] + (E[\sum_i x_i])^2\right)$$

but

$$\begin{aligned} V[\sum_i x_i] &= nV[x] = n\sigma^2 \\ E[\sum_i x_i] &= n\mu \\ \Rightarrow \quad E[S_2^2] &= \tfrac{1}{n}\left\{n(\sigma^2+\mu^2) - \tfrac{1}{n}\,(n\sigma^2 + n^2\mu^2)\right\} \\ &= \tfrac{1}{n}(n-1)\sigma^2 \end{aligned}$$

This leads to the conclusion that S_2^2 is a biased estimator of the variance σ^2. One can however get rid of this bias by defining a new estimator :

$$S^2 = \frac{n}{n-1}\,S_2^2 = \frac{1}{n-1}\,\sum_i (x_i - \hat{\mu})^2$$

•Consistency

An estimator is said to be consistent if $\forall \epsilon > 0$

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta_i| \geq \epsilon) = 0$$

This means that the estimator converges to the true value with increasing n.

Remarks :
1-If the data are distributed according to a a gaussian p.d.f or in more general way according to a p.d.f for which the C.L. theorem applies then the sample mean is a consistent estimator of the true mean since in this case the estimator p.d.f is $N(\hat{x};\mu,\frac{\sigma^2}{n})$ whose width goes to zero as n increases to ∞ and then $\hat{x}$ goes to μ.
2- An estimator can be consistent without being unbiased. It should be however asymptotically unbiased.

• Efficiency
The efficiency of an estimator is a notion related to its variance. It increases when the variance decreases. As an example let's estimate the variance of the sample mean estimator and the variance estimator :

$$V\left[\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2}\sum_i V[x_i] = \frac{n}{n^2}V[x] = \frac{\sigma^2}{n}$$

$$V\left[S^2 = \frac{1}{n-1}\sum_i (x_i - \hat{\mu})^2\right] = \left(\frac{\sigma^2}{(n-1)}\right)^2 V\left[\sum_i \frac{(x_i - \hat{\mu})^2}{\sigma^2}\right]$$

If x_i is distributed according to a gaussian p.d.f or such that the C.L. theorem applies then :

$$\begin{aligned} V[S^2] &= \left(\tfrac{\sigma^2}{n-1}\right)^2 V\left[\sum_i z_i^2\right] = \left(\tfrac{\sigma^2}{n-1}\right)^2 V[\chi^2(n-1)] \\ &= \tfrac{2(\sigma^2)^2}{n-1} \end{aligned}$$

σ^2 being unknown we replace it by its unbiased estimator S^2.

$$\Rightarrow \quad V[\hat{\mu}] = \frac{S^2}{n}, \quad V[S^2] = \frac{2(S^2)^2}{n-1}$$

$$\hat{\mu} \Rightarrow \hat{\mu} \pm \sqrt{\frac{S^2}{n}} \qquad S^2 \Rightarrow S^2 \pm \sqrt{\frac{2}{n-1}}S^2$$

3 Information and likelihood

The information concept in statistics was proposed to quantify the knowledge/ignorance about the characteristics of a population. Since the variance is a quantity which expresses also the knowledge/ignorance of the searched parameters and since the efficiency of an estimator is directly related to its variance we can use the information concept to evaluate the efficiency of one estimator. The information concept should satisfy the following requirements :

• The information should increase when the number of observations increases. The precision of the estimator should be consequently improved.
• The observations not related to the studied parameters should not increase the information.
R.A. Fisher was the first to introduce the information concept. We will use here his definition :

Likelihood function

Let's consider a random variable X with a p.d.f described by $f(x,\theta)$ where θ is the studied parameter. the function of adjoint probability of n independent observations $\prod_{i=1}^{n} f(x_i,\theta) = \mathcal{L}(x_1, x_2 \cdots x_n, \theta)$ is called the likelihood function . Fisher definition of the information is based on this function :

$$\begin{aligned} I_x(\theta) &= E\left[\left(\frac{\partial \ell n(\mathcal{L}(x_i\theta))}{\partial\theta}\right)^2\right] = E\left[\left(\frac{\partial \ell}{\partial\theta}\right)^2\right] \\ &= \int_{\Omega_\theta} (\frac{\partial \ell n(\mathcal{L}(x,\theta)}{\partial\theta})^2 \, \mathcal{L}(x,\theta)\, dx \end{aligned}$$

where $\ell = \ell n\mathcal{L}$ and Ω_θ is the random variable domain which may depend on the parameter θ. The generalization to n variables and k parameters follows :

$$\left[I_{\underline{x}}(\underline{\theta})\right]_{ij} = \int_{\Omega_\theta}^{dx} \frac{\partial \ell(\underline{x};\underline{\theta})}{\partial\theta_i} \frac{\partial \ell(\underline{x},\underline{\theta})}{\partial\theta_j} \mathcal{L}(\underline{x},\underline{\theta})$$

The score

We define the score of a sample as follows :

$$\begin{aligned} S = \sum_i S_1(x_i,\theta) &= \sum_i \frac{\partial}{\partial\theta} \ell n f(x_i,\theta) \\ S &= \frac{\partial}{\partial\theta}\left(\ell(\underline{x};\theta)\right) = \frac{\partial}{\partial\theta}\ell \end{aligned}$$

We then link the information to the score :

$$I = E[S^2(\underline{x};\theta)]$$

If Ω_θ is independent of θ and $\mathcal{L}(\underline{x},\theta)$ is regular we can show easily that :

$$E[S(\underline{x},\theta)] = 0$$

$$\Rightarrow I = V[S(\underline{x};\theta)] = -E\left[\frac{\partial S}{\partial\theta}(\underline{x};\theta)\right]$$

We can also show that :

$$\begin{aligned} I(\theta) &= E[S^2(\underline{x},\theta)] = E[(\sum_i S_i(x_i,\theta))^2] \\ &= nV[S_1(x;\theta)] + n^2\left\{E[S_1(x_i\theta)]\right\}^2 \end{aligned}$$

This means that when n increases, $I(\theta)$ increases as well. We can notice also that if the observations do not depend on θ, $\mathcal{L}$ does not depend neither $\Rightarrow S_1 = 0$. This leaves I unchanged.
Therefore, I satisfies the two requirements of the information concept.

The minimal variance of an estimator

If $\hat{\theta}$ is an estimator of θ such that $E[\hat{\theta}] = \theta + b_n(\hat{\theta})$and if the associated variance is defined and the random variable domain is independent of θ then the following theorem applies :

Théorème de Rao-Cramer

$$\sigma^2(\hat{\theta}) = V[\hat{\theta}] \geq \frac{1 + b_n(\hat{\theta})}{I(\theta)}$$

if $b_n(\hat{\theta}) = 0$ then $\sigma^2(\hat{\theta}) \geq \frac{1}{I(\theta)} = \sigma^2_{min}(\theta)$

This allows us to define the efficiency of an estimator in the following way :

$$\epsilon(\hat{\theta}) = \frac{\sigma^2_{min}(\theta)}{\sigma^2(\hat{\theta})} = \frac{1}{I(\theta)\sigma^2(\hat{\theta})}$$

From the previous theorem we can deduce the following result :

If the p.d.f of the random variable of the considered population is of the form :

$$f(x;\theta) = exp\ \Big(A(\theta)\hat{\theta}\,(x) + B(\theta) + K(x)\Big)$$

then $\hat{\theta}$ is an unbiased efficient estimator of the quantity

$$E[\hat{\theta}] = -\frac{\frac{\partial B(\theta)}{\partial \theta}}{\frac{\partial A(\theta)}{\partial \theta}}$$

This can be generalized to the case of more than one parameter :

$$f(x;\underline{\theta}) = exp\left(\underline{A}(\underline{\theta})\ \underline{\hat{\theta}}(x) + B(\underline{\theta}) + K(x)\right)$$

$$E[\hat{\theta}_i] = \frac{-\frac{\partial B(\underline{\theta})}{\partial \theta_i} + \sum\limits_{j \neq i} E(\hat{\theta}_j)\ \frac{\partial A_j}{\partial \theta_i}(\underline{\theta})}{\frac{\partial A_i}{\partial \theta_i}(\underline{\theta})}$$

Application :

The normal distribution $N(x;\mu;\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\ exp\ \left(-\frac{1}{2}\ \frac{(x-\mu)^2}{\sigma^2}\right)$ can be written as :

$$N(x;\mu,\sigma^2) = exp\ \left(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{1}{2}\left(\frac{\mu^2}{\sigma^2} + \ell n(2\pi\sigma^2)\right)\right)$$

The adjoint probability of having $x_1, \cdots, x_n$ in n observation is then given by :

$$\prod_{i=1}^{n} N(x_i; \mu, \sigma^2) = \ exp\ \left(\frac{n\mu}{\sigma^2}\bar{x} - \frac{n}{2\sigma^2}\overline{x^2} - \frac{n}{2}\ \left(\frac{\mu^2}{\sigma^2} + \ell n(2\pi\sigma^2) \right) \right)$$

with the expression of $f(x, \theta)$ we can deduce :

$$\begin{array}{rcll} A_1 & = & \frac{n\mu}{\sigma^2} & \hat{\theta}_1 = \bar{x} = \frac{1}{n}\sum\limits_{i=1}^{n} x_i \\ A_2 & = & -\frac{n}{2\sigma^2} & \hat{\theta}_2 = \overline{x^2} = \frac{1}{n}\sum\limits_{i=1}^{n} x_i^2 \\ B & = & -\frac{n}{2}\left(\frac{\mu^2}{\sigma^2} + \ell n(2\pi\rho^2) \right) & \\ K & = & 0 & \end{array}$$

$\hat{\theta}_1 = \bar{x}$ is the unbiased efficient estimator of :

$$\frac{-\frac{\partial B}{\partial \mu}}{\frac{\partial A_1}{\partial \mu}} = \mu$$

and $\hat{\theta}_2 = \overline{x^2}$ is the unbiased efficient estimator of :

$$\frac{-\frac{\partial B}{\partial \sigma^2} + E(\hat{\theta}_1)\frac{\partial B}{\partial \mu}}{\frac{\partial A_2}{\partial \sigma^2}} = \mu^2 + \sigma^2$$

4 Maximum likelihood method

The liklyhood function defined by :

$$\mathcal{L}(x_1, x_2, \cdot, x_n, \underline{\theta}) = \prod_{i=1}^{n} f(x_i, \underline{\theta})$$

is the function of the adjoint probability associated to n independent measurements of $x_1, x_2 \cdots, x_n$.
It can be interpreted in two ways :

Probability interpretation : Knowing the parameter θ, $\mathcal{L}$ gives the probability that an experiment equivalent to ours produces the same measurements $x_1, x_2 \cdots$.

Statistical interpretation : Having obtened $x_1, x_2 \cdots x_n$ as measurements and believing that we know the p.d.f, one can find the parameter θ by maximizing $\mathcal{L}$.

In practice we replace $\mathcal{L}$ by its logarithm :

$$\ell = \ell n \mathcal{L}(x_1, \cdots x_n, \underline{\theta}) = \sum_i \ell n \ f(x_i, \underline{\theta}) = \sum_i \ell_i$$

maximizing ℓ leads to maximizing $\mathcal{L}$

$$\frac{\partial}{\partial \theta_i} \ell = \frac{1}{\mathcal{L}} \frac{\partial \mathcal{L}}{\partial \theta_i} = 0$$

Remark :

The previous condition leads to an extremum. One should however check that this is indeed a maximum and the maximum of the maxima.

Example : For measurements of n random variables each normally distributed according to $N(x_i; \mu_i, sigma_i^2)$. The likelihood function built with the n measurements $x_i, \cdots, x_n$ is :

$$\begin{aligned} \mathcal{L} &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \mathrm{e}^{-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}} \\ \ell &= \sum_{i=1}^{n} -\frac{1}{2} \ell n(2\pi) - \ell n \sigma_i - \frac{(x_i-\mu_i)^2}{2\sigma_i^2} \end{aligned}$$

if the μ_i are all identical $\mu_i = \mu$ while the σ_i are different but known :

$$\frac{\partial \ell}{\partial \mu} = 0 \Rightarrow \qquad \sum_i \frac{x_i}{\sigma_i^2} - \sum \frac{\mu}{\sigma_i^2} = 0$$

$$\Rightarrow \quad \hat{\mu} = \frac{\sum_i x_i/\sigma_i^2}{\sum_i 1/\sigma_i^2} \quad \text{which is a M.L. estimator of the mean value } \mu$$

It is easy to check that this estimator is unbiased : $E[\hat{\mu}] = \mu$
In addition :

$$\begin{aligned} V[\hat{\mu}] &= \frac{1}{\sum_i 1/\sigma_i^2} \\ I &= -E\left[\frac{\partial S}{\partial \mu}\right] \\ &= -E\left[\frac{\partial^2 \ell}{\partial \mu^2}\right] \\ &= \sum_i \frac{1}{\sigma_i^2} \end{aligned}$$

The previous result indicates that $V[\hat{\mu}] = 1/I$ which shows that the ML estimator $\hat{\mu}$ is an efficient one.

The asymptotic proprieties of $\mathcal{L}$

We can show that asymptotically when the number of observations becomes large :

$$S(\underline{x}, \theta) = -I(\hat{\theta})(\theta - \hat{\theta}) = \frac{\partial \ell n}{\partial \theta}$$

$$\Rightarrow \quad \ell = \ell n\, \mathcal{L} = -\frac{I(\hat{\theta})}{2}\,(\theta - \hat{\theta})^2 + \ell n\, k$$

$$k = \mathcal{L}(\hat{\theta}) = \mathcal{L}_{max}$$

$$\Rightarrow \quad \mathcal{L}(\theta) = \mathcal{L}_{max}\ \ exp\ \left(-\frac{I(\hat{\theta})}{2}\,(\theta - \hat{\theta}^2\right) \approx N(\theta; \hat{\theta}; 1/I(\hat{\theta})$$

Asymptotically $\mathcal{L}$ is proportional to a gaussian function of θ centered at $\hat{\theta}$ and having for width (variance) $\frac{1}{I(\hat{\theta})}$

Remark : $\mathcal{L}$ is not a p.d.f. of θ so any change of parameter from $\theta \to g(\theta)$ can be done in replacing in $\mathcal{L}(x;\theta)$ simply θ by $g(\theta)$:

$$\mathcal{L}'(x, \theta) = \mathcal{L}(xg(\theta))$$

Variance of the M.L. estimator

If the M.L. estimator is efficient, biased or unbiased, and if the definition domain of the random variable does not depend on θ then we can estimate the variance using the relation $V^{-1}[\hat{\theta}] = I[\theta]$. Two methods can be used to obtain I :

- $I[\theta] = E[s^2] = E\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right]$

which can be generalized to $E\left[\frac{\partial \ell}{\partial \theta_i}\ \frac{\partial \ell}{\partial \theta_j}\right] = V_{ij}^{-1}$

Using $E[S_1] = 0$ and replacing θ by $\hat{\theta}$

$$\hat{V}_{ij}^{-1}[\hat{\theta}] = \sum_{k=1}^{n} \left.\frac{\partial \ell n f(x_k, \underline{\theta})}{\partial \theta_i}\right|_{\underline{\hat{\theta}}} \left.\frac{\partial \ell n f(x_k, \underline{\theta})}{\partial \theta_j}\right|_{\underline{\hat{\theta}}}$$

• $I(\theta) = -E\left[\frac{\partial S}{\partial \theta}\right]$
We can show that :

$$\hat{V}_{ij}^{-1}[\hat{\theta}] = -\sum_{k=1}^{n} \left.\frac{\partial^2 \ell n f(x_k, \underline{\theta})}{\partial \theta_i \partial \theta_j}\right|_{\hat{\theta}}$$

Remark The two methods provide the same results in the asymptotic case.

Bayesian inference method for the estimation of the M.L estimator variance

If the M.L. estimator does not satisfy the previous conditions (domain independence of *theta...*) and if $\mathcal{L}$ could not be put under a gaussian form then we may use the bayesian method to estimate the variance :

$$f(\underline{\theta}/\underline{x}) \text{ propotional to } \; f(\underline{x}/\underline{\theta})\; f(\underline{\theta})$$

In this case $\underline{\theta}$ becomes a "random variable" $f(\underline{x}/\underline{\theta}) = \mathcal{L}(\underline{x};\underline{\theta})$ and if we take $f(\underline{\theta}) = cte$ then :

$$f(\underline{\theta}/\underline{x}) = \qquad \frac{\mathcal{L}(\underline{x};\underline{\theta})}{\int \mathcal{L}(\underline{x};\underline{\theta})\; d\underline{\theta}} \qquad \text{(normalization)}$$

$f(\underline{\theta}/\underline{x})$ being considered as p.d.f of the "variable" θ, we have :

$$V_{ij}[\underline{\hat{\theta}}] \;=\; E(\hat{\theta}_i. - \theta_i)(\hat{\theta}_j - \theta_j)]$$

$$V_{ij}[\underline{\hat{\theta}}] \;=\; \frac{\int (\hat{\theta}_i - \theta_i)(\hat{\theta}_j - \theta_j)\mathcal{L}(\underline{x},\underline{\theta}) d\underline{\theta}}{\int \mathcal{L}(\underline{x},\underline{\theta})\; d\underline{\theta}}$$

Estimation of the M.L estimator variance using a graphic method
There is a graphic method which allows to find the variance when $\mathcal{L}$ has a gaussian shape :

$\mathcal{L} = \mathcal{L}_{max}\; \mathrm{e}^{-\frac{1}{2}Q^2}$ with $Q^2 = \frac{(\hat{\theta}-\theta)^2}{\sigma^2}$

$\Rightarrow$

$$\begin{aligned} \ell = \ell n \mathcal{L} \;&=\; \ell n\; \mathcal{L}_{max} - \tfrac{1}{2}\, Q^2 \\ \ell(\theta) \;&=\; \ell n\; \mathcal{L}_{max} - \tfrac{1}{2}\frac{(\hat{\theta}-\theta)^2}{\sigma^2} \end{aligned}$$

If θ_1 is such that $\hat{\theta} - \theta_1 = \sigma$
then $\ell(\theta_1) = \ell n \mathcal{L}_{max} - \frac{1}{2}$

$$\ell(\theta_1) = \ell_{max} - \frac{1}{2}$$

This means that in order to find the variance (uncertainty of an estimation) one should find the maximum value of θ_0 associated to $\ell_max = \ell n \mathcal{L}_{max}$ and then find the value of θ_1 for which the value of is reduced by 1/2. The difference $|\theta_0 - \theta_1|$ gives then the variance value.
If $\mathcal{L}$ does not have a gaussian shape one could try to perform a "variable change" $\theta \to g(\theta)$ such that $\mathcal{L}(x; g(\theta))$ has a gaussian shape.
We then look for $g_{1,2}$ such that ; $\ell(g_{1,2}) = \ell_{max} - \frac{1}{2}$

$$\Rightarrow \quad \theta_1 \quad = \quad g^{-1}(g_1) \quad \theta_2 = g^{-1}(g_2)$$

$$\Rightarrow \quad \sigma_1 \quad = \quad |\hat{\theta} - \theta_1|, \qquad \sigma_2 = |\hat{\theta} - \theta_2|$$

σ_1, σ_2 are not necessarily identical.

Maximum likelihood with constraints

When the parameters we are trying to estimate are linked among each others or they are constrained into specific domains. We should take these constraints into consideration because they can contribute to the variance (uncertainty) reduction. Indeed those constraints add to our knowledge and hence increase the information which as we already saw is inversely proportional to the variance.

Few tricks to deal with constraints// • If there is a simple relation between the parameters such that : $\theta_1 + \theta_2 = 0$ one can replace θ_2 by $-\theta_1$.
• if $\theta_1 < \theta < \theta_2$ one can propose a new parameter ψ to replace θ such that :

$$\theta = \theta_1 + \frac{1}{2}(\sin\psi + 1)(\theta_2 - \theta_1)$$

However a more general solution does exist :

5 Lagrange multipliers

The constraints on the different parameters can be put into relations of the form :

$$\underline{g}(\underline{\theta}) = 0$$

We then extend the definition of the likelihood function to include those relations using the Lagrange multipliers

$$\ell n\mathcal{L}(\underline{x};\underline{\theta}) \rightarrow F(\underline{x};\underline{\theta},\underline{\alpha}) = \ell n\mathcal{L}(\underline{x};\underline{\theta}) + \alpha^T \underline{g}(\underline{\theta})$$

$$\alpha^T \underline{g} = \alpha_1 g_1 + \alpha_2 g_2 + \cdots$$

We then look for the maximum of the extended likelihood function with respect to $\underline{\theta}$ and $\underline{\alpha}$:

$$\left.\frac{\partial F}{\partial \theta_i}\right|_{\underline{\theta}=\hat{\theta},\underline{\alpha}=\hat{\underline{\alpha}}} = \left.\frac{\partial \ell}{\partial \theta_i}\right|_{\underline{\theta}=\hat{\underline{\theta}}} + \hat{\underline{\alpha}}^T \left.\frac{\partial g}{\partial \theta_i}\right|_{\underline{\theta}=\hat{\underline{\theta}}} = 0$$

$$\left.\frac{\partial F}{\partial \alpha_j}\right|_{\underline{\theta}=\hat{\theta},\underline{\alpha}=\hat{\underline{\alpha}}} = g(\hat{\theta}) = 0 \qquad \text{(contraintes satisfaites)}$$

From what precedes we can construct the information matrix :

$$I = -E\begin{pmatrix} \frac{\partial^2 F}{\partial\underline{\theta}\partial\underline{\theta}} & \frac{\partial^2 F}{\partial\underline{\theta}\partial\underline{\alpha}} \\ \frac{\partial^2 F}{\partial\underline{\alpha}\partial\underline{\theta}} & \frac{\partial^2 F}{\partial\underline{\alpha}\partial\underline{\alpha}} \end{pmatrix} = \begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix}$$

$$\begin{aligned} V[\hat{\theta}] &= A^{-1} - A^{-1}\, B\, V[\hat{\alpha}]\, B^T\, A^{-1} \\ V[\hat{\alpha}] &= (B^T\, A^{-1}\, B)^{-1} \end{aligned}$$

6 Least squares method

We have seen seen that the likelihood function $\mathcal{L}$ corresponding to n independent measurements distributed according to $N(x_i;\mu;\sigma_i)$ is given by :

$$\mathcal{L} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}}\, \mathrm{e}^{-\frac{(x_i-\mu)^2}{2\sigma_i^2}}$$

and that

$$\ell = \ell n\ \mathcal{L} = -\frac{n}{2}\ \ell n 2\pi + \sum_{i=1}^{n}\ \left[-\ell n\sigma_i - \frac{(x_i-\mu)^2}{2\rho_i^2}\right]$$

To maximize $\mathcal{L}$ and consequently ℓ is then equivalent to minimize "$\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma_i^2}$". The latter is clearly distributed as $\chi^2(n)$ if μ is known and according to $\chi^2(n-1)$ if μ is estimated from data("1 relation linking x_i").
This leads to the least squares method's estimator by deriving with respect to μ :

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma_i^2}$$

$$\left.\frac{\partial \chi^2}{\partial \mu}\right|_{\hat{\mu}} = -2 \sum_{i=1}^{n} \frac{(x_i - \hat{\mu})}{\sigma_i^2} = 0$$

$$\Rightarrow \quad \hat{\mu} = \frac{\sum\limits_{i=1}^{n} \frac{x_i}{\sigma_i^2}}{\sum\limits_{i=1}^{n} \frac{1}{\sigma_i^2}}$$

Remark ; We find here the same estimator that we obtained from the likelihood method. This coincidence is due to the fact that our data are distributed normally (gaussian). In gnral and for dufferent kinds of distribution the estimators obtained from the two methods are different.

Remark :

The least squares method is a special case of a category of methods based on the minimization of different kinds of distances :

$$d = \sum_i |x_i - \mu|^\alpha$$
$$d = \sum_i \left(\frac{|x_i - \mu|}{\sigma_i}\right)^\alpha$$

historically Laplace in 1792 used $|x_i - \mu|^{\alpha=1}$ but in 1805 Legendre proposed $\frac{(x_i-\mu)^2}{\sigma_i 2}$ (to determine the comets' orbits).

Least squares estimator's variance
the variance of the least square estimator of the mean value is given by l'estimateur de moindres carrés de la valeur moyenne trouvée dans le cas de distributions normales est donnée par :

$$V[\hat{\mu}] = V\left[\frac{\sum\limits_i \frac{x_i}{\sigma_i^2}}{\sum\limits_i \frac{1}{\sigma_i^2}}\right] = \left(\sum_i \frac{1}{\sigma_i^2}\right)^{-1} = \frac{1}{\sum\limits_i \frac{1}{\sigma_i^2}}$$

If we take the successive derivatives of χ^2 at $\mu = \hat{\mu}$ we have the following results :

$$\begin{aligned}
\chi^2(\mu) &= \sum_i \frac{(x_i-\mu)^2}{\sigma_i^2} \\
\left.\frac{\partial\chi^2}{\partial\mu}\right|_{\mu=\hat{\mu}} &= -2\sum_i \frac{(x_i-\hat{\mu})}{\sigma_i^2} = 0 \\
\left.\frac{\partial^2\chi^2}{\partial\mu^2}\right|_{\mu=\hat{\mu}} &= 2\sum_i \frac{1}{\sigma_i^2} = \frac{2}{V[\hat{\mu}]}
\end{aligned}$$

The derivatives of higher order are all null. This allows to rewrite the χ^2 for any value of μ in the vicinity of $\hat{\mu}$ using a Taylor development :

$$\chi^2(\mu) = \chi^2(\hat{\mu}) + \frac{(\mu-\hat{\mu})^2}{V[\hat{\mu}]}$$

Estimation of the least square estimator variance using a graphic method

The previous formula provides us with a graphic method to find the the variance of estimator. Indeed, once the the value of $\hat{\mu}$ corresponding to the minimum χ^2 value is found we look for μ_1 et μ_2 such that :

$$\mu_1 - \hat{\mu} = \hat{\mu} - \mu_2 = V[\hat{\mu}] \Rightarrow \chi^2(\mu_1) = \chi^2(\mu_2) = \chi^2(\hat{\mu}) + 1$$

which provides us with the variance.

7 Least squares linear model :

Let $y_1, y_2 \cdots y_n$ be $n$measurements of the y quantity which depends on another quantity x. For each value of x_i (that we consider exactly knwon) we associate the measurement y_i. We would like to predict the y associated to a given x. For this we propose a relation linking y to x

$$y = \theta_1 h_1(x) + \theta_2 h_2(x) + \cdots + \theta_k h_k(x)$$

where y is linear with respect to the parameters θ_i. On can then write each y_i asa function of x_i :

$$y_i = y(x_i) + \epsilon_i = \sum_{j=1}^{K} \theta_j h_j(x_i) + \epsilon_i$$

Let's suppose that theϵ_i are such that :

$E[\epsilon_i] = 0 \quad V[\epsilon_i] = \sigma_i^2 \quad$ with the $\quad \sigma_i$ are well known but not necessarily gaussian.
We try to determine the θ_i with the n measurements ($n \geq k$) using the least squares method as follows :

$$\begin{aligned} Q^2 = \sum_{i=1}^{n} \frac{\epsilon_i^2}{\sigma_i^2} &= \sum_{i=1}^{n} \frac{(y_i - y(x_i))^2}{\sigma_i^2} \\ &= \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left(y_i - \sum_{j=1}^{K} \theta_j h_j(x_i) \right)^2 \end{aligned}$$

Remark The σ_i are not necessarily gaussian $\Rightarrow$ Q^2 is not necessarily χ^2 distributed.

Let's know minimize the previous expression of Q^2 with respect to θ_ℓ. we obtain :

$$\frac{\partial Q^2}{\partial \theta_l} = 2 \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left(g_i - \sum_{j=1}^{k} \theta_j h_j(x_i) \right) h_l(x_i) = 0$$

Which can be put more in more elegant way :

$$\sum_{i=1}^{n} \frac{h_l(x_i)}{\sigma_i^2} \sum_{j=1}^{K} \hat{\theta}_j h_j(x_i) = \sum_{i=1}^{n} \frac{g_i}{\sigma_i^2} h_l(x_i)$$

and by introducing the following matrices :

$$\underline{y} = \begin{pmatrix} y_1 \\ \\ y_n \end{pmatrix}, \ \underline{\theta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_k \end{pmatrix}, \ \underline{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$\underline{H} = \begin{pmatrix} h_1(x_i) & \cdots & h_k(x_1) \\ \vdots & & \\ h_1(x_n) & & h_k(x_n) \end{pmatrix}$$

We can rewrite the previous relations as follows :

$$\begin{aligned} \underline{y} &= \underline{H}\,\underline{\theta} + \underline{\epsilon} \\ Q^2 &= \underline{\epsilon}^T \underline{V}^{-1} \underline{\epsilon} \end{aligned}$$

with

$$\underline{V} = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \cdot \\ 0 & \sigma_2^2 & 0 & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \sigma_n^2 \end{pmatrix}$$

Remark We can include the case when the measurements are not independent through correlations in the $\underline{V}$ matrix.
If we replace now ϵ_i by its expression we have :

$$\begin{aligned} Q^2 &= \left(\underline{y} - \underline{H}\,\underline{\theta}\right)^T \underline{V}^{-1}(\underline{y} - \underline{H}\,\underline{\theta}) \\ \frac{\partial Q2}{\partial \underline{\theta}} &= -2\,\underline{H}^T\underline{V}^{-1}(\underline{y} - \underline{H}\,\underline{\theta}) = 0 \\ &\Rightarrow \quad \underline{H}^T\underline{V}^{-1}\underline{H}\,\underline{\theta} = \underline{H}^T\underline{V}^{-1}\underline{y} \\ &\Rightarrow \quad \hat{\underline{\theta}} = \left(\underline{H}^T\underline{V}^{-1}\underline{H}\right)^{-1}\underline{H}^T\underline{V}^{-1}\underline{y} \end{aligned}$$

Let's evaluate now $E[\hat{\theta}]$:

$$E[\hat{\underline{\theta}}] = (\underline{H}^T\underline{V}^{-1}\underline{H})\,H^T\underline{V}^{-1}E[\underline{y}]$$

$$E[\underline{y}] = \underline{H}\,\underline{\theta} \text{ car } \; E[\underline{\in}] = 0 \;\Rightarrow\; E[\hat{\underline{\theta}}] = \underline{\theta}$$

This shows that the least square estimator $\hat{\theta}$ is unbiased.
Let's evaluate variance also :

$$\begin{aligned} \left.\frac{\partial Q^2}{\partial \underline{\theta}}\right|_{\underline{\theta}=\hat{\underline{\theta}}} &= -2\,\underline{H}^T\underline{\underline{V}}^{-1}(\underline{y} - \underline{H}\,\underline{\theta}) = 0 \\ \left.\frac{\partial^2 Q^2}{\partial \underline{\theta}^2}\right|_{\underline{\theta}=\hat{\underline{\theta}}} &= 2\,\underline{H}^T\underline{V}(\underline{y})^{-1}\underline{H} = 2\,\underline{V}^{-1}(\hat{\underline{\theta}}) \end{aligned}$$

All the following derivatives cancel and then one will write :

$$\begin{aligned} Q^2(\underline{\theta}) &= Q^2\,\hat{\underline{\theta}} + 0 + \tfrac{1}{2}(\underline{\theta} - \hat{\underline{\theta}})^T\frac{\partial^2 Q^2}{\partial \underline{\theta}^2}(\underline{\theta} - \hat{\underline{\theta}}) \\ &= Q^2(\hat{\underline{\theta}}) + (\underline{\theta} - \hat{\underline{\theta}})^T V^{-1}(\underline{\theta} - \hat{\underline{\theta}}) \end{aligned}$$

This result is identical to the one obtained with the x_i normally distributed (gaussian).

8 Gauss-Markov theorem

If $E(\epsilon_i) = 0$ and if $V(\epsilon_i)$ are well defined and independent of both y and θ) then the least squares estimator $\hat{\theta}$ is unbiased and has the lowest variance of any other estimator in the linear case whatever are the p.d.f of the ϵ_i. Consequences :

• The least squares estimator is superior to the likelihood estimator in the linear case far from the asymptotic behaviour where both are equivalent.
• Concerning samples of limited number, it becomes sometimes necessary to group data together. In this case the least square method loses its advantages with respect to the likelihood one.

Including constraints in the least squares method

Consider the existence of m constraints among the k parameters :

$$\sum_{i=1}^{K} \ell_{ij}\theta_j = R_i \qquad\qquad i = 1, \cdots m$$

written with matrices this gives :

$$\underline{L}\,\underline{\theta} = \underline{R}$$

We include the constraint in the definition of Q^2 using Lagrange multipliers :

$$Q^2 = (\underline{y} - \underline{H}\,\underline{\theta})^T\,\underline{V}^{-1}(\underline{y} - \underline{H\theta}) + 2\underline{\lambda}^T\,(\underline{L}\,\underline{\theta} - \underline{R})$$

We minimize Q^2 with respect to both $\underline{\theta}$ and $\underline{\lambda}$

$$\frac{\partial Q^2}{\partial \underline{\theta}} = 0 \;\Rightarrow\; \underline{H}^T\,\underline{V}^{-1}\underline{H}\,\underline{\hat{\theta}} + L^T\,\underline{\hat{\lambda}} = \underline{H}^T\,\underline{V}^{-1}\underline{y}$$

$$\frac{\partial Q^2}{\partial \underline{\lambda}} = 0 \;\Rightarrow\; \underline{L}\,\underline{\hat{\theta}} = \underline{R}$$

$$\Rightarrow \begin{pmatrix} \underline{c} & \underline{L}^T \\ \underline{L} & \underline{o} \end{pmatrix} \begin{pmatrix} \underline{\hat{\theta}} \\ \underline{\hat{\lambda}} \end{pmatrix} = \begin{pmatrix} \underline{s} \\ \underline{R} \end{pmatrix}$$

where

$$\underline{c} \;=\; \underline{H}^T\underline{V}^{-1}$$

$$\underline{s} \;=\; \underline{H}^T\underline{V}^{-1}\underline{y}$$

The solutions can be given by :

$$\begin{aligned}\underline{\hat{\theta}} &= \underline{F}\,\underline{S} + \underline{G}^T\,\underline{R} = \underline{F}\,\underline{H}^T\underline{V}^{-1}\,\underline{y} + \underline{G}^T\,\underline{R} \\ \underline{\hat{\lambda}} &= \underline{G}\underline{S} + \underline{E}\,\underline{R} = \underline{G}\,\underline{H}^T\,\underline{V}^{-1}\underline{\in}\end{aligned}$$

with the associated variances

$$\begin{aligned}\underline{V}\,[\underline{\hat{\theta}}] &= \underline{F} \\ \underline{V}\,[\underline{\hat{\lambda}}] &= \underline{W}\end{aligned}$$

and $COV(\underline{\hat{\theta}},\, \underline{\hat{\lambda}}) = 0$ with

$$\begin{aligned}\underline{W} &= (\underline{L}\,\underline{C}^{-1\,T}), \quad \underline{F} = \underline{C}^{-1} - \underline{C}^{-1}\,\underline{L}^T\,\underline{W}\,\underline{L}\,\underline{C}^{-1} \\ \underline{G}^{-1} &= \underline{W}\,\underline{L}\,\underline{C}^{-1}, \qquad\qquad \underline{E} = -\underline{W}\end{aligned}$$

9 Extended linear model

Let's now consider the case where x_i are not known with certainty and let's take the case of a straight line fit with two parameters to find a and b. We can write y as a function of x as :$y = ax + b$
The question we may ask ourselves is the following : Which distance we have to minimize ? Contrary to the case in which the x_i are known with certainty we should here take into account the uncertainty of x_i. The distance to minimize in this case is given by :

$$d_i^2 = \frac{(x - x_i)^2}{\sigma_{xi}^2} + \frac{(y - y_i)^2}{\sigma_{yi}^2}$$

This can be cast in more compact form :

$$d_i^2 = \frac{(y_i - y(x_i))^2}{\sigma_i^2}$$

with σ_i is the uncertainty of the quantity $y_i - ax_i - b$, given by :

$$\sigma_i^2 = \sigma_{yi}^2 + a^2\sigma_{xi}^2$$

10 Generalization of the linear model

It is useful to gather the previous results in one general frame in which uncertainties on x and y are treated coherently as well well as constraints and correlations. This frame is elegantly presented using matrices as follows :

$$Z_i = \begin{pmatrix} x_i \\ y_i \end{pmatrix} \quad Z_i^c = \begin{pmatrix} x_i^c \\ y_i^c \end{pmatrix}$$

Were Z_i^c are different from zero only when there are constraints linking x_i to y_i.

$$\underline{V_i} = \begin{pmatrix} \sigma_{xi}^2 & COV(x_i, y_i) \\ \\ COV(x_i, y_i) & \sigma_{yi}^2 \end{pmatrix}$$

$$y_i^c = \underline{H}^T(x_i^c)\underline{\theta}$$

$$Q^2 = \sum_{i=1}^{n} (\underline{Z_i^c} - \underline{Z_i})^T \ \underline{V_i}(\underline{Z_i^c} - \underline{Z_i}) + \lambda_i(y_i^c - \underline{H}^T \ (x_i^c)\underline{\theta})$$

The parameters $\underline{\theta}$ are then obtained by minimizing Q^2 with respect to $\underline{\theta}, \underline{x}^c, \underline{y}^c$ and the Lagrange multipliers$\underline{\lambda}$.

11 Binned data

To apply the least squares method we have access to x_i as well as to y_i. There are some situation where only one coordinate is given. This case is well treated by the likelihood method. Is it possible to apply the least squares method in this case also. The answer is yes. For this we have to collect the events into mutually exclusive and exhaustive classes defined with respect to the variable x. An example of this is the histogram whose bins are the mentioned classes. In this case we compare the number of events in one bin with that expected from a multinomial distribution in which the P_i is determined by the p.d.f to describe the data and which depends on the parameter we are looking for.

$$n = \sum_i n_i \qquad \sum_i P_i = 1$$

We introduce the expression of Q_1^2

$$Q_1^2 = n \sum_i \frac{(n_i/n - P_i)^2}{P_i} = Q_1^2 = \sum_i \frac{(n_i - nP_i)^2}{nP_i}$$

This expression is an approximation of the least squares formula. Indeed in the case of the multinomial distribution for each bin we have $\sigma_i^2 = nP_i(1-P_i)$ et $COV(n_i, n_j) = -nP_iP_j$ which are different from those used in χ_1^2. However , if the P_i are all small then we can admit the following approximation : $\sigma_i^2 \simeq nP_i$ and $COV(n_i, n_j) \simeq 0$. In this case χ_1^2 is a good approximation.

Consequence : to justify the use of Q_1^2 we need to have many bins so that all P_i are small.
Remark : the use of χ_1^2 is delicate because of the presence of the parameters in the denominator through χ_1^2.
this dificulty can be avoided by introducing :

$$Q_2^2 = \sum_i \frac{(n_i - nP_i)^2}{n_i}$$

In Q_2^2we have replaced the expected uncertainty related to the p.d.f from that estimated from the measurement ($\sigma_i^2 = n_i$) valid for large value of n_i.

Consequence : n_i should not be too small ($\sigma_i^2 = n_i$)and not too large neither ($COV \neq 0$).
Remark : Q_1^2 et Q_2^2 are not distributed according to $\chi^2(n-1)$.

Maximum likelihood versus least squares : The maximum likelihood can also be used in the case of binned data. The proposed p.d.f is :

$$\mathcal{L} = f = \frac{n!}{n_1!n_2! - n_k!} \qquad P_1^{n1} P_2^{n2} \qquad P_k^{nk}$$

$$\ell = \sum_i n_i \ell n P_i + cte$$

where the constant cte does not depend on P_i.
the next step is ti maximize $\mathcal{L}$.
Remark : In the case of the maximum likelihood one can eliminate the normalization coefficients if they don't include the searched parameters.

We can summerize the previous results by a kind of hierarchy

$$ML > Q_1^2 > Q_2^2$$

However minimization is easier in the case of Q_1^2 et Q_2^2.

Chapitre 5

Statistical interpretation

The choice of an estimator is based on its proprieties : Bias, consistency, efficiency..etc. Once the choice is made we can, using our sample, obtain an estimate of the searched parameter as well as the estimator variance which informs on the uncertainty of this estimate . The out put of this work s commonly written under the following form : $\theta = \hat{\theta} \pm \sigma_{\hat{\theta}}$. Let's try now to give an interpretation of this formula. To make our work even simpler we suppose that our estimator is normally distributed (gaussian) .
Two interpretations can be made :
1) The true value of θ_t has $0,683$ of probability to be in the interval $[\hat{\theta} - \sigma_{\hat{\theta}}, \hat{\theta} + \sigma_{\hat{\theta}}]$.
2)Repeating the same measurement many times and estimating each time the interval $[\hat{\theta} - \sigma_{\hat{\theta}}, \hat{\theta} + \sigma_{\hat{\theta}}]$ then, $0,683$ of those intervals will contain the true value θ_t

Those two interpretations are the main ones used by physicists when analyzing the results of one experiment. The first one is called the **Bayesian** interpretation while the second one is the **frequentist** one.
The difference between the two interpretations is not artificial as it may appear at first look. There is a thorough difference based on different understanding of the way we should apply statistics in our analyses. Here we will try to develop each interpretation and where and why it is often used :

12 The frequentist method

Neyman was the first to a give a frequentist interpretation to the result : $\theta = \hat{\theta} \pm \sigma_{\hat{\theta}}$ by elaborating the notion of confidence levels. His idea can be presented as follows :

$f(x,\theta)$ is the p.d.f associated to the random variable$x$ and it depends on the searched parameter θ. Let's consider the following probability integral :

$$\beta = \int_a^b f(x,\theta)\, dx = P(a \leq X \leq b)$$

This integral could not be estimated for any given β since the integrand contains θ which is unknown. Neyman proposed to perform a variable change replacing x by y in such a way that the p.d.f $g(y)$associated to y does not depend anymore on θ.

$$\beta = \int_A^B g(y)\, dy = P(A \leq Y \leq B)$$

The new expression can a priori be estimated. It depends on both A and B. So for a given β one can hope to fixe A and B. But since A and B depend themselves on θ this allows to determine two values of θ : $\theta_A = \theta^-$ and $\theta_B = \theta^+$.
We can then rewrite the previous integral as

$$\int_A^B g(y)\, dy = \beta = P([\theta^-, \theta^+] \text{ contains } \theta_t)$$

We should here emphasize the meaning of this relation. It says that the interval $[\theta^-, \theta^+]$has a probability β to contain θ_t and not there is a probability β that θ_t is located in the interval $[\theta^-, \theta^+]$. The parameter θ is not a random variable while θ and θ^+ are. They inherit their status from the random variable y. Now what does mean that $[\theta^-, \theta^+]$ contains θ_t ? The meaning is the following : If we repeat the same experiment 100 times and each time we estimate $[\theta^-, \theta^+]$ then we expect the true value θ_t to be in $\beta \times 100$ of the found intervals. **Remark :** We use the notion of coverage to interpret the previous result. We say there is a probability β to cover the true value θ_t

13 Construction of the confidence

The intervals : $[\theta^-, \theta^+]$ we obtain with the frequentist method are called the confidence level. We will give hereafter the method that allowed us to construct all the confidence levels associated to a given probability before to realize the experiment. We will then see how to give the confidence interval once the experiment is realized.
For this we suppose known the parameter θ and we look for an interval of the random variable x :

$$[x^-, x^+] \text{ such that } \; P(x^- \leq x \leq x^+) = \beta = \int_{x^-}^{x^+} f(x;\theta)dx$$

The choice of x^-, x^+ not being "always" unique we can select the central one defined by :

$$\int_{-\infty}^{x^-} f(x;\theta);dx = \frac{1-\beta}{2} = \int_{x^+}^{\infty} f(x;\theta)\ dx$$

We repeat the same procedure for all values of θ. This leads to two curves. The first is built from x^- . This curve will be called θ^+. The second curve is built from the $x^+$and will be called the θ^-.curve.

The experiment output will give us a certain value of $x = \hat{x}$. From this we can have an estimate of the true value θ_t. In order to quantify the uncertainty on our estimate we use the intersection of the line $x = \hat{x}$ with the two curves θ^+ and θ^+ to determine the interval $[\theta^-, \theta^+]$ This interval has a probability β to contain the true value θ_t.

14 Confidence bounds

Sometimes we are more interested by non-central confidence level than by the central ones defined in the previous paragraph. This happens when the estimated parameters are close or even beyond the physical boundary. In this case we can establish what we call superior or inferior limit which correspond to confidence intervals with one of the two boundaries is the largest or the lowest possible value.

Upper limit
The difference with respect to the confidence intervals we build before is that here for each value of θ_t we look for an interval $[x^-, x_L^+]$ where x_L^+ is the largest possible value. We then fixe the curve $\theta^+ = x^-$. The output of our experiment will provide us with $\hat{x}$ and as before we determine the intersection of the line $x = \hat{x}$ with the curve θ^+. The found value θ_{sup} is called the upper limit and we have :

$$\beta = P(\theta \leq \theta_{sup}) = \int_{\hat{x}}^{\infty} f(x;\theta_{sup})\ dx = P(x \geq \hat{x})$$

In the same way we define the lower limit θ_{inf} which is the intersection of the line $x = \hat{x}$ with with the curve $\theta^- = x^+$ obtained by the construction of confidence intervals $[x_S^-, x^+]$ where x_S^- is the lowest possible value :

$$\beta = P(\theta \leq \theta_{inf}) = \int_{-\infty}^{\hat{x}} f(x;\theta_{inf})\ dx = P(x \leq \hat{x})$$

Remark : The upper and lower limits are very important in the search of new phenomena. They allow one to "eliminate" some models which predicts too much or too few events with respect the observation.

Applications
1- Normal distribution : $N(\mu, \sigma)$
A) Estimation of the mean value in case of known variance :
The mean sample $\hat{x} = \frac{1}{n}\sum_i nx_i$ is the best estimator of the mean value μ . It is obtained using n measurements distributed normally according to $N(\mu, \sigma)$. The variance of this estimator can be given by $\sigma/\sqrt{n}$. We look for $x_{\beta/2}$ such as :

$$Pr(x^- = \mu - (\sigma/\sqrt{n})x_{\beta/2} < \hat{x} < \mu + (\sigma/\sqrt{n})x_{\beta/2} = x^+) = \beta$$

This leads to :

$$Pr(\mu^- = \hat{x} - (\sigma/\sqrt{n})x_{\beta/2} < \mu < \hat{x} + (\sigma/\sqrt{n})x_{\beta/2} = \mu^+) = \beta$$

The values of $x_{\beta/2}$ in the case of a gaussian distribution are tabulated. For instance for $\beta = 0,95$ we have $x_{\beta/2} = 1,96$.

B) Estimation of the mean value with unknown variance : In this case we replace the variance by its unbiased estimator $S_2^2 = \frac{1}{n-1}\sum_i (x_i - \hat{x})^2$. Now the variable $t = \frac{x-\mu}{\sqrt{S_2^2}}$ is distributed according to the Student distribution. As before using the tabulated values of $t_{\beta/2}$ and $-t_{\beta/2}$ so that :

$$Pr(x^- = \mu - s't_{\beta/2} \leq \hat{x} \leq \mu + s't_{\beta/2} = x^+) = \beta$$
$$Pr(\mu^- = \hat{x} - s't_{\beta/2} \leq \mu \leq \bar{x} + s't_{\beta/2} = \mu^+) = \beta$$

where $s' = \sqrt{S_2^2}$.
Remark : $t_{\beta/2}$ is determined by the following equation :

$$\int_{-\infty}^{t_{\beta/2}} t(x; n-1)\, dx = \frac{1}{2}(1+\beta)$$

Binomial distribution : $\frac{N!}{n!(N-n)!}\, P^n(1-P)^{N-n}$

We try to determine P. The best estimator is $\hat{P} = n/N$ (n number of success, N number of trials).

Since N is fixed, for each value of P we determine détermine $[n^-, n^+]$ such that :

$$\sum_{k=n^-}^{k=n^+} B(k,N,P) = \sum_{k=n^-}^{k=n^+} \frac{N!}{k!(N-k)!}\, P^k (1-P)^{N-k} \geq \beta$$

Remark : The discreet nature of the binomial distribution explains the sign $\geq$ in the previous sum rather than the equality sign since it is not always possible for a given real value of β to find two integers n^-, n^+ such that the sum is equal to β. This is called un overcoverage.

Poisson sistribution $\frac{\lambda^n}{n!} e^{-\lambda}$

The best estimator of λ of the Poisson distribution is the number of events n. As for the binomial distribution we can build the confidence level by finding $[n^-, n^+]$ for each value of λ such that :

$$\sum_{k=n^-}^{k=n^+} P(\lambda,k) = \sum_{k=n^-}^{k=n^+} \frac{\lambda^k}{k!} e^{-\lambda} \geq \beta$$

Remark : Both the Poisson and the binomial distribution are not symmetric and the choice of n^-, n^+ is not unique.
Remark : In the case of the Poisson distribution which related to rare events it is usually the upper limit which is looked for :

$$\begin{aligned} \beta = P(\lambda < \lambda^+) &= P(k > n) \\ &= \sum_{k=n+1}^{\infty} \frac{\lambda^k}{k!}\, \mathrm{e}^{-\lambda^+} \\ &= 1 - \sum_{k=0}^{n} \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda^+} \end{aligned}$$

We can show :

$$\begin{aligned} 1 - \beta = \sum_{k=0}^{n} \frac{\lambda^{+k}}{k!} \mathrm{e}^{-\lambda^+} &= P(\chi^2(2n+2) > 2\lambda^+) \\ &= \int_{2\lambda^+}^{\infty} \chi^2(2n+2)\, d\chi^2 \end{aligned}$$

For the lower limit we can use the formula :

$$\beta = \int_{2\lambda^-}^{\infty} \chi^2(2n)\, d\chi^2$$

15 Background consideration :

The search of new phenomena is one of the most interesting topics of the high energy physics . When we search for supersymmetric particles for instance we select events which respect the characteristics given by the theory of the supersymmetry. Unfortunately, phenomena of the standard physics can also show up sometimes when their characteristics are similar to the supersymmetric related events.We should then take into account this contribution when we analyze our sample and before to announce a discovery or to set a limit.
In the case of rare events the distribution of both signal and background is rather Poissonian. We have then

$$\lambda = \lambda_s + \lambda_{bg}$$

Since the estimator of λ is N and since we are looking rather for λ_s. Then we can put a limit on *lambda* and then we deduce the limit on lambda λ_s

$$\lambda_s^{lim} = \lambda^{lim} - \lambda_{bg}$$

RemarK if N is small (a negative fluctuation) resulting in $\lambda_{lim} \leq \lambda_{bg}$ for certain values of β

$$\text{alors} \quad \lambda_s^{lim} < 0$$

This looks absurd but it is not from the statistics point of view. However one can try to remedy by adopting a more robust method.

16 Unified frequentist method

In order to avoid the problem of finding a negative value of an upper limit of a positive quantity as mentioned before. There is a method proposed by Cousins and Field which try to construct the confidence intervals as well as the upper and lower limits without having those intervals in non physical zones. The application of this method in the case of Poisson distribution is the following.

$$P(n \mid \lambda_s, \lambda_{bg}) = \frac{(\lambda_s + \lambda_{bg})^n}{n!} \, \mathrm{e}^{-(\lambda_s + \lambda_{bg})}$$

To build the intervals we fix λ_s and then we order the different n according to the value of the following ratio : $R(n, \lambda_s, \lambda_{bg}) = \frac{P(n\ \lambda_s, \lambda_{bg})}{P(n\ \lambda_s^{best}, \lambda_{bg})}$ with $\lambda_s^{best} = max[0, n - \lambda_{bg}]$.

The n are selected with decreasing$R$ and we stop when $\sum_n P(n) = \beta$

17 Bayesian method

The Bayesian method is based on the use of Bayes formula :

$$f_{post}(\theta/x) = \quad f(x/\theta)\ f(\theta) = \quad \mathcal{L}(x;\theta)\ f_{prior}(\theta)$$

It is interesting to analyze the previous formula in order to understand the difference between the Bayesian method and the frequentist one. Indeed in this formula $f_{prior}(\theta), f_{post}(\theta\ x)$ are probability density functions of θ. This means that θ becomes a variable and no more a parameter. In addition we need to find $f(\theta)_{prior}$ and none could be really justified although the uniform distribution is often used.

As in the frequentist method we can define intervals we call them credit intervals which gives the probability of the true θ_t to belong to :

$$\beta = P(a \leq \theta \leq b) = \int_a^b f(\theta/x)d\theta$$

By choosing a prior with a uniform distribution we can rewrite the previous equation in the following way :

$$\beta = P(a \leq \theta \leq b) = \frac{\int_a^b \mathcal{L}(x;\theta)\ d\theta}{\int_{\theta_{inf}}^{\theta_{sup}} \mathcal{L}(x;\theta)\ d\theta}$$

Where the denominator represents is the integral of the likelihood function between the two extreme values of θ and was introduced to ensure probability normalization. the two values θ_{inf} and θ_{sup} can be fixed from physics consideration and introduced as the boundary values of the uniform distribution of the prior.
One can easily see here from what preceds how the Bayesian method can handle the boundary and the physics constraint. This is an important advantage of this method.

18 Application

The upper limit on the parameter λ_s of a Poisson distributed variable can be obtained with the Bayesian method including the fact that the number of the observed events could not be less than the background contribution :

$$\beta = 1 - \frac{\sum_{k=0}^{N} \frac{(\lambda_s^{sup}+\lambda_{bg})^k}{k!} \, \mathrm{e}^{-(\lambda_s^{sup}+\lambda_{bg})}}{\sum_{k=0}^{N} \frac{\lambda_b^k}{k!} \, \mathrm{e}^{-\lambda_b}}$$

19 Using the simulation

It happens that the p.d.f of both the signal and the background are of complicated form and different from those we mentioned in the previous chapters. In this case (but also in the case of known p.d.f) we can use the simulation to build the confidence/credit intervals or to set upper and lower limits.

For instance, let us take the case of Poisson distribution variables for both the signal and the background and let's see how we can proceed to set an upper limit on λ_s , we can proceeds as following : We select a range of λ_s of interest. For each value of λ_s we generate signal events according a Poisson distribution with λ_s as a parameter. We do the same for the background (fixed λ_{bg}). The output of each trial (we have to make many) is $N_{tot} = N_s + N_{bg}$ then we count the cases for which $N_{tot} \leq N_{obs}$

$$1 - \beta = \frac{\text{Cases with} N_{tot} \leq N_{obs}}{\text{all cases}}$$

if we consider the upper limit at 95% this means that we have to take $\beta = 95\%$ then the upper limit will be the λ_s which gives :

$$\frac{\text{Cases with} N_{tot} \leq N_{obs}}{\text{all cases}} = 0.05$$

Remark : We can also include the constraint that the observed number of events could not be less than the background contributionN_{tot} easily by excluding all the cases for which $N_{bg} < N_{obs}$

Chapitre 6

Statistical Tests

A statistical test is a test which allows one to choose one of two hypotheses :
Examples :
• We are searching for the Higgs boson in the $H \to \gamma\gamma$. We observed an excess in this canal with respect to expected background. Is this a Higgs or a positive fluctuation of the background ?
• We are using two different gas mixtures for our gaseous detector. Which mixture is the more appropriate ?
• The same kind of crystals is produced by two different companies. How to be sure that the two productions have the same characteristics ?

There are two kinds of statistical tests :
1- Parametric tests :
Those tests are used when we compare two distributions defined by some parameters (μ, σ for the normal distributions).

2- Non parametric tests :
Those tests compare hypotheses without reference to the parameters .

Statistical tests may fail and we can distinguish two kinds of failures :

20 First and second kind of error

In order to explain in practice the two kinds of error that one may make when dealing with statical tests let's take the following example :

Two fertilizers A, B were used by a farmer for his tomatoes. The results obtained with the fertilizer A in number of kilos of tomatoes can be given by a normal distribution $N(x; \mu_A = 600, \sigma_A = 30)$. The results of the fertilizer B can be also given by a normal distribution but with different parameters $N(x; \mu_B = 660, \sigma_B = 40)$.

We choose one fertilizer and we take off the mark on its box. We ask the farmer to use it for his tomatoes production. At the end of the season we ask him if he can tell us what of the two fertilizers we provided him with ?
The farmer who followed in his youth lectures on statistical treatment of data will proceed as follows :
1-The farmer will decide to consider the hypothesis with the product A as his hypothesis of reference and call that H_0 and the other one as the hypothesis H_1.
2- The farmer will fix an upper limit beyond based on the normal distribution associated to the hypothesis H_0. The upper limit will be defined at a level of confidence $(1 - \alpha)$ in the following way :

$$x_{lim} = \mu_A + a(\alpha)\sigma_A$$

a being a function of α. It is defined such that :

$$\alpha = \int_{\mu_A + a\sigma_A}^{\infty} N(x; \mu_A, \sigma_A)dx$$

For instance if we take $\alpha = .05$ from the normal distribution A will be equal to 1,64.
3-The farmer will compare the production result $x_0$with x_{lim}. If $x_0 \leq x_{lim}$ the farmer will conclude that the product used is the product A. Otherwise it is the product B associated to the hypothesis H_1.

Indeed the quantity of tomatoes produced this year was 630 kg which is less than $x_{lim} = 600 + 1,64 \times 30 = 649,2$. So according to the farmer's rule the product used was A.
Was the farmer right in choosing the hypothesis H_0. There are indeed 4 possible cases : 4 cas de figures sont possibles

$x_0 \leq x_{lim}$	and the used product is	A
$x_0 \leq x_{lim}$	and the used product is	B
$x_0 \geq x_{lim}$	and the used product is	A
$x_0 \geq x_{lim}$	and the used product is	B

If H_0 is true and we select $H_1 \rightarrow$ **error of the first kind** .

if H_1 is true and we select $H_0 \rightarrow$ **error of the second kind** .

Remark : H_0 et H_1 are not treated on the same footing since it is H_0 which determines x_{lim}. The decision may change if the reference was chosen to the product B in our previous example with(a lower limit is used in this case)

α being fixed we try to reduce the error of the second kind by.

21 Statistical tests : simples and composites

There are two kinds of tests : simple and composite
A simple test is of the kind : HH is true if $\theta = \theta_0$) (one value)
A test is said composite if HH is true if : $\theta \in A$ (set)

22 Tests proprieties

Definitions We call the critical region W the variable domain which allow to eliminate the hypothesis H_0 and keep H_1 $P(W/H_0) = \alpha, P(W/H_1) = 1 - \beta$
$\bar{W}$ is the complementary set $P(\bar{W}/H_0) = 1 - \alpha$
α is called the amplitude of the test
$1 - \beta$ is called the power of the test $= P(\theta)$.
With those definitions we can introduce the proprieties of the statistical tests
a) The power
Let's consider two hypotheses :

$$H_0 : \theta = \theta_0 \qquad\qquad H_1 : \theta = \theta_1$$

A test comparing the two H_0 et H_1 is the more powerful the bigger $1 - \beta$ for $\theta = \theta_1$ at fixed α .
b) La consistency
A test is said to be consistent if $\lim_{N\rightarrow\infty} P(\underline{\theta} \in W \ H_1) = 1$ where N is the number of trials.
c) Le biais
A test is said to be biased if we can find $\theta_1 \neq \theta_0$ (valeur spécifiée par H_1 et H_0 respectivement) telle que $P(\theta_1) = 1 - \beta(\theta_1) < \alpha$.

I Test choice

The choice of the hypothesis based on the previous proprieties should be made once we optimize both α and β In the case of two simple hypotheses we can use the method of Neyman and Pearson to determine the critical region W_α which reduces β.

1 Neyman-Pearson test

Let's call $f(x, \theta_0)$ the p.d.f associated to the hypothesis H_0 and $g(x, \theta_1)$the p.d.f of the hypothesis H_1 then we can write :

$$P(x \in W_\alpha / H_0) = \int_{W_\alpha} f(x; \theta_0)\, dx = \alpha$$

$$P(x \in W_\alpha / H_1) = \int\limits_{W_\alpha} g(x; \theta_1)\, dx = 1 - \beta$$

Remember that the aim is to find the critical region associated to the valueα and which maximizes $1 - \beta$. The latter can be written as :

$$\begin{aligned} 1 - \beta &= \int\limits_{W\alpha} \frac{g(x;\theta_1)}{f(x;\theta_0)}\, f(x; \theta_0)\, dx \\ &= E_{W\alpha}\left[\frac{g(x;\theta_1)}{f(x;\theta_0)}\, H_0\right] \end{aligned}$$

Neyman and Pearson propose to order the x according to the decreasing $\mathbf{R} = \frac{g(x;\theta_1}{f(x;\theta_0)}$. We include the values of x associated to the higest values of R until we have $\int_{W_\alpha}, f(x, \theta_0)dx = \alpha$. This determines the $C_\alpha = 1/R_{lim}$.
The test of Neyman-Pearson is then based on the inverse ratio of R called λ estimated for the value x_0 of the experiment output.

$$\lambda = \frac{f(x_0; \theta_0)}{g(x_0; \theta_1)}$$

this can be extended to include other measurements using the likelihood function. In this case we have :

$$\lambda = \frac{\mathcal{L}(x, \theta_0)}{\mathcal{L}(x; \theta_1)}$$

We choose H_0 if $\lambda > c_\alpha$
We reject H_0 if $\lambda \leq c_\alpha$

Bayesian tests

The Neyman-Pearson method is applied on the p.d.f associated to the random variables. It inherits the same problem of the frequentist method when it concerns the boundary conditions. Here again the Bayes formula can be used in order to incorporate the boundary limits in a coherent manner. However we know for now the price to pay for...

$$P(H_i/\underline{x}) = \frac{P(\underline{x}/H_i)}{P(\underline{x}/H_0) + P(\underline{x}/H_1)} \; P_P(H_i) \qquad \text{where } i = 0, 1$$

where H_0, H_1 are simple hyposises.
If we compare at present $P(H_0/\underline{x}$ et $P(H_1/\underline{x})$ we have

$$\frac{P(H_0/\underline{x})}{P(H_1/\underline{x})} = \frac{(P\underline{x}/H_0)}{P(\underline{x}/H_1)} \; \frac{P(H_0)}{P(H_1)}$$

we can as for the Neyman-Pearson method define λ :

$$\frac{P(\underline{x}/H_0)}{P(\underline{x}/H_1)} = \frac{\mathcal{L}(\underline{x}/H_0)}{\mathcal{L}(\underline{x}/H_1)} = \lambda$$

and if we decide to choose arbitrarily a uniform distribution for both H_0 and H_1

$$\Rightarrow \quad \frac{P(\underline{x}/H_0)}{P(\underline{x}/H_1)} = \lambda' = c\,\lambda$$

where c is the ratio :$P_P(H_0)/P_P(H_1)$

Remark : The Neyman-Pearson as well as the Bayesian methods can be extended (not without a certain amount of complexity) for comparisons with composite tests.

Adjustment test

When H_1 is not specified then the previous tests could not be used. This is the case when we would like to compare our data to a p.d.f associated to a given H_0. In this case we would like to know if the the proposed H_0 is good and how good it is.
Few adjustment tests exist here are after some of the most used :

χ^2 test

$$\chi^2 = \sum_{i=1}^{N} \frac{(g_i - f_i)^2}{\sigma_i^2}$$

f_i are given by the p.d.f associated to H_0
if the σ_i are normally distributed then X^2 is a $\chi^2(N)$ distributed.

We can introduce here what we call the confidence level for this test defined as the probability to have a result of the fit as bad or even worse to the one we found :

$$CL = \int_{x^2}^{\infty} \chi^2(y; N) dy$$

The meaning of the CL is clear. It measures the deviation of our hypothesis from the data.
Remark : If one uses the least squares method to determine $f_i(Q^2_{min})$, then χ^2 still follows a χ^2 distributed but with a $N - K$ degree where K is the number of fixed parameters using the least squares method. In addition of some constraints are used the degree of the χ^2 is then $N-K+M$ (M number of constraints).

Remark : We can use the likelihood method in the same way as for the χ^2. We need however to specify the $g(\ell)$ associated to the likelihood function. In this case we define the confidence level as CL=$\int_{-\infty}^{\ell_{max}} g(\ell) d\ell$.

Run test

The χ^2 test could be misleading sometimes. It can be completed by a test called RUN defined as follows. Let r be the number of de sequences of results with the same sign with respect to the chi^2 fit and :
K_A number of data with a positive sign with respect to the fit
K_B number of data with a negative sign with respect to the fit
$K = K_A + K_B$

we can show that

$$E[r] = 1 + \frac{2K_AK_B}{4K}, V[r] = \frac{2K_AK_B(2K_AK_B - K)}{K^2(K-1)}$$

For $r > 10$ we can use the gaussian approximation in order to deduce the confidence level and hence we can eliminate H_0 even if the output of the χ^2 test is ok.

Kolmogorov test

We use a d.p.f function such $F(x)$ and we compare it to a function $S_n(x)$ built from the data in the following way :

$$S_n(x) = \begin{cases} 0 & x < x(1) \\ \frac{r}{n} & x(r) \leq x \leq x(r+1) \\ 1 & x(n) \leq x \end{cases}$$

with $x_{(1)} \leq x_{(2)} \leq \qquad x_{(r)} \leq \qquad x(n)$

$$S_n(x) \to F(x) \text{when } n \to \text{if } quad H_0 \qquad \text{is true}$$

We then estimate :

$$D_n = max\{|S_n(x) - F(x)|\} \qquad \text{for all} \qquad x$$

if $n \geq 80$

α	we reject H_0 if $\sqrt{n}D_n >$
$0,01$	$1,63$
$0,05$	$1,36$
$0,1$	$1,22$
$0,15$	$1,07$

Non parametric tests

These are tests where the parameters are not used. We have three kinds of them. They allow to :

- Verify that 2 variables are independent

- Verify the random nature of one variable
- Verify that 2 samples have the same p.d.f

Tests of variables independence

If we have a smaple of events with the 2 variables x, y distributed according to $f(x, y)$. We want to check taht :

$$H_0 : f(x, y) = g(x)h(y) \qquad g, h \quad \text{are the marginal p.d.f l}$$

For this we consider the estimated correlation coefficient :

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i + y_i \; - \bar{x}\bar{y}}{S_x \; S_y} \; = \; \frac{\overline{xy} - \overline{xy}}{S_x \; S_y}$$

if x, y are independent $\Rightarrow E[r] = 0$ and $v(r) = 1/(n-1)$

We can show that $t = r\sqrt{\frac{n-2}{1-r^2}}$ is distributed according to a student distribution of the $n-2$ order.
This allows to estimate the confidence level (and hence we can reject H_0 if $|t|$ is large).

Test of randomness

If we want to study the distribution of a variable as a function of time or other quantity we build a sample (x_i, y_i) where x_i is the studied variable and $y_i = t_i$ then we try to show that the two variables are independent.

Verification of two samples

• Kolmogorov : We remplace

$$\begin{array}{rcl} D_n & = & sup\,\{|S_n(x) - F(x)|\} \quad \text{by} \\ D_{n1n2} & = & sup\,\{|S_{n1}(x) - S_{n2}(x)|\} \quad \forall x \end{array}$$

$\sqrt{n}D_n$ is then replaced by $\sqrt{\frac{n_1 n_2}{n_1+n_2}}\, D_{n1n_2}$ We apply then the same method as for the Kolmogorov test

• Run : We combine the members of the two samples but we keep track of their sample origin. The members of the first become the members of A of the run test and those of the second members of B. We then use the run test as before.

• χ^2 : If we compare 2 histograms with the same binning :

$$\Rightarrow \quad X^2 = \sum_{j=1}^{2}\sum_{i=1}^{k} \frac{(n_{ij} - N_j P_i)^2}{N_j P_i}$$

Pi is then replaced by :

$$\hat{P}_i = \frac{n_{1i} + n_{2i}}{N_1 + N_2}$$

The result is distributed as

$$\chi^2(k-1) \qquad k \text{ number of bins}$$

Références

• Probabilités, analyse des données et statistique, Gilbert SAPORTA, Ed. Technip

• Statistics for nuclear and particle physics, Louis LYONS, Ed. Cambridge

• Statistical methods in data analysis, W.J. METZGER, Nijmegen

• Workshop on conference limits, CERN, Geneva, Switzerland, CERN-2000-005, 30 May 2000

• The advanced theory of statistics, Kendall and Stuart, Ed. Griffin