EPJ Web of Conferences **55**, 01001 (2013)
DOI: 10.1051/epjconf/20135501001

Probability and statistics: A reminder

Benoit Clément

[1] *LPSC, grenoble (IN2P3)*

Abstract. The main purpose of these lectures is to provide the reader with the tools needed to data analysis in the framework of physics experiments. Basic concepts are introduced together with examples of application in experimental physics. The lecture is divided into two parts: probability and statistics. It is build on the introduction from "data analysis in experimental sciences" given in [1]

Slides

References

[1] B. Clement, *Analyse de données en sciences expérimentale* (DUNOD, Paris, 2012).

Probability and Statistics

Basic concepts I

(from a physicist point of view)

Benoit CLEMENT – Université J. Fourier / LPSC

bclement@lpsc.in2p3.fr

School Of Statistics

Bibliography

Kendall's Advanced theory of statistics, Hodder Arnold Pub.

volume 1 : Distribution theory, A. Stuart et K. Ord

volume 2a : Classical Inference and and the Linear Model, A. Stuart, K. Ord, S. Arnold

volume 2b : Bayesian inference, A. O'Hagan, J. Forster

The Review of Particle Physics, K. Nakamura et al., J. Phys. G 37, 075021 (2010) (+Booklet)

Data Analysis: A Bayesian Tutorial, D. Sivia and J. Skilling, Oxford Science Publication

Statistical Data Analysis, Glen Cowan, Oxford Science Publication

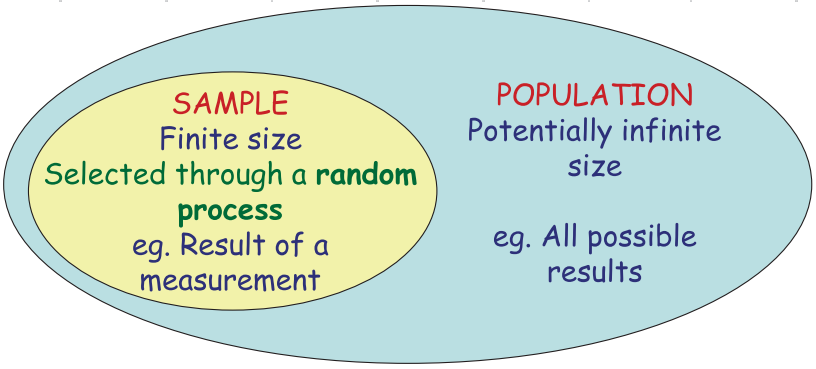
Analyse statistique des données expérimentales, K. Protassov, EDP sciences

Probabilités, analyse des données et statistiques, G. Saporta, Technip

Analyse de données en sciences expérimentales, B.C., Dunod

2

Sample and population



Characterizing the sample, the population and the drawing procedure :

-> **Probability theory** (today's lecture)

Using the sample to estimate the characteristics of the population

-> **Statistical inference** (tomorrow's lecture)

3

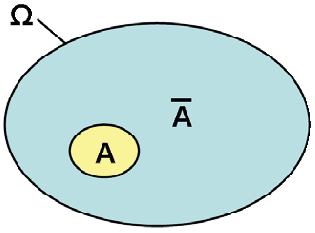
Random process

A random process (« measurement » or « experiment ») is a process whose **outcome cannot be predicted with certainty.**

It will be described by :

Universe: Ω = set of all possible **outcomes**.

Event : **logical condition on an outcome**. It can either be true or false; an event splits the universe in 2 subsets.



An event $\mathcal{A}$ will be identified by the subset **A** for which $\mathcal{A}$ is true.

4

Probability

A **probability function** P is defined by : (Kolmogorov, 1933)

$$P : \{\text{Events}\} \rightarrow [0:1]$$
$$A \rightarrow P(A)$$

satisfying :

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

Interpretation of this number :

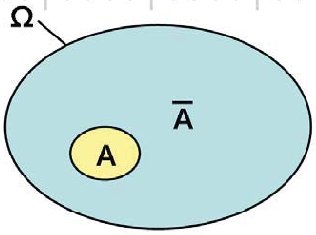
- **Frequentist approach** : if we repeat the random process a great number of times n , and count the number of times the outcome satisfy event A, n_A then the ratio :

$$\lim_{n \to +\infty} \frac{n_A}{n} = P(A) \quad \text{defines a probability}$$

- **Bayesian interpretation** : a probability is a measure of the credibility associated to the event.

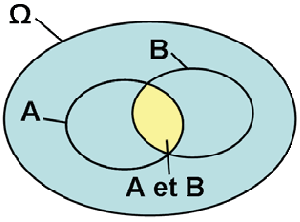
5

Simple logic

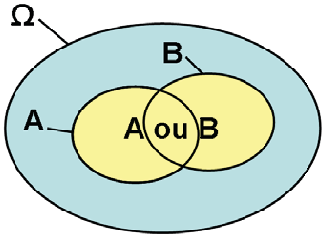


Event « **not** $\mathcal{A}$ » is associated with the complement $\bar{A}$.

$$P(\bar{A}) = 1-P(A)$$
$$P(\varnothing) = 1-P(\Omega) = 0$$



Event « $\mathcal{A}$ **and** $\mathcal{B}$ » is associated with the ensemble $A \cap B$.



Event « $\mathcal{A}$ **or** $\mathcal{B}$ » is associated with the ensemble $A \cup B$.

$$P(A \cup B) = P(A)+P(B)-P(A \cap B)$$

6

We will use and mix notations : $\mathcal{A} \leftrightarrow A$, **or** $\leftrightarrow \cup$, **and** $\leftrightarrow \cap$

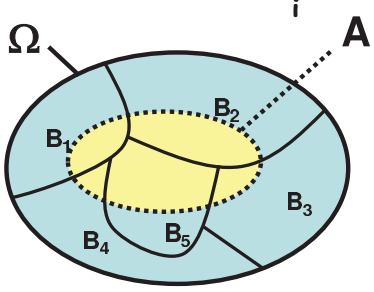
Incompatibility and partition

Two **incompatible** events cannot be true simultaneously, then :

$$P(A \text{ and } B) = 0 \quad \text{and} \quad P(A \text{ or } B) = P(A)+P(B)$$

A **partition** is a set of incompatible events that cover the full universe :

$$\Omega = \bigcup_i B_i, \qquad B_i \cap B_j = \varnothing \quad (i \neq j)$$



Then, for any event *A* :

$$P(A) = \sum_i P(A \text{ and } B_i)$$

Similar to a basis in linear algebra.

7

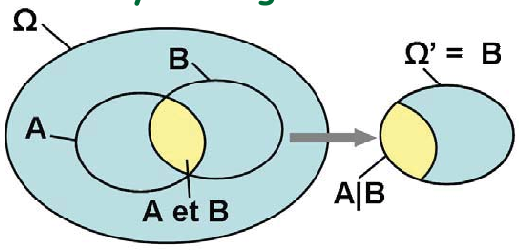
Conditional probability and independence

If an event **B** is **known to be true**, one can restrain the universe to Ω'=B and define a new probability function on this universe, the **conditional probability**.

$$P(A|B) = \text{« probability of } A \text{ given } B \text{ »}$$

From Venn diagram :

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



Two events are **independent**, if the realization of one is not linked in any way to the realization of the other :

$$P(A|B)=P(A) \quad \text{and} \quad P(B|A) = P(B)$$

From the previous relations: $P(A \text{ and } B) = P(A).P(B)$

8

Bayes theorem

The definition of conditional probability leads to :

P(A and B) = P(A|B).P(B) = P(B|A).P(A)

Hence relating P(A|B) to P(B|A) by the **Bayes theorem** :

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

Or, using a partition $\{B_i\}$:

$$P(B_i|A) = \frac{P(A|B_i).P(B_i)}{\sum_i P(A \text{ and } B_i)} = \frac{P(A|B_i).P(B_i)}{\sum_i P(A|B_i).P(B_i)}$$

This theorem will play a major role in Bayesian inference : given **data** and a **set of models**, it translates into :

$$P(\text{model}_i|\text{data}) = \frac{P(\text{data}|\text{model}_i).P(\text{model}_i)}{\sum_i P(\text{data}|\text{model}_i).P(\text{model}_i)}$$

9

Application of Bayes

100 dices in a box :

70 are equiprobable (A) 30 have a probability 1/3 to get 6 (B)

You pick one dice, throw it until you reach 6 and count the number of try. Repeating the process thrice, you get 2, 4 and 1.

What's the probability that the dice is equilibrated ?

For one throw : $P(n|A) = (1-p_6)^{n-1}p_6 = \frac{5^{n-1}}{6^n}$ $P(n|B) = \frac{2^{n-1}}{3^n}$

Combining several throw: (for one dice, throws are independents)

$$P(n_1 \text{ and } n_2 \text{ and } n_3|A) = P(n_1|A)P(n_2|A)P(n_3|A) = \frac{5^{n_1+n_2+n_3-3}}{6^{n_1+n_2+n_3}}$$

$$P(n_1 \text{ and } n_2 \text{ and } n_3|B) = \frac{2^{n_1+n_2+n_3-3}}{3^{n_1+n_2+n_3}}$$

$$P(A|n_1, n_2, n_3) = \frac{P(n_1, n_2, n_3|A)P(A)}{P(n_1, n_2, n_3|B)P(B) + P(n_1, n_2, n_3|A)P(A)}$$

$$= \frac{\frac{5^{n_1+n_2+n_3-3}}{6^{n_1+n_2+n_3}} \times 0.7}{\frac{2^{n_1+n_2+n_3-3}}{3^{n_1+n_2+n_3}} \times 0.3 + \frac{5^{n_1+n_2+n_3-3}}{6^{n_1+n_2+n_3}} \times 0.7} = \frac{\frac{5^4}{6^7} \times 0.7}{\frac{2^4}{3^7} \times 0.3 + \frac{5^4}{6^7} \times 0.7} \approx 0.42$$

10

Random variable

When the outcome of the random process is a **number** (real or integer), we associate to the random process, a **random variable X**.
Each realization of the process leads to a particular result : **X=x**.
x is a realization of **X**.

For a **discrete variable** :

Probability law : $p(x) = P(X=x)$

For a **real variable** : $P(X=x)=0$,

Cumulative density function : $F(x) = P(X<x)$

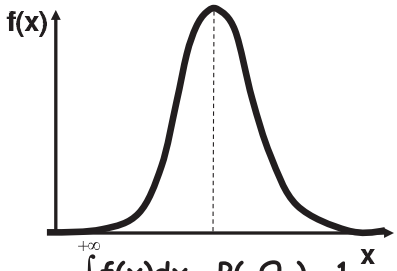
$$dF = F(x+dx)-F(x) = P(X < x+dx) - P(X < x)$$
$$= P(X < x \text{ or } x < X < x+dx) - P(X < x)$$
$$= P(X < x) + P(x < X < x+dx) - P(X < x)$$
$$= P(x < X < x+dx) = f(x)dx$$

Probability density function (pdf) : $f(x) = \dfrac{dF}{dx}$

11

Density function

Probability density function



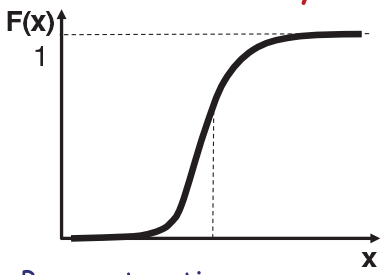
$$\int_{-\infty}^{+\infty} f(x)dx = P(\Omega) = 1$$

Note : discrete variables can also be described by a probability density function using Dirac distributions:

$$f(x) = \sum_i p(i)\delta(i-x)$$

$$\sum_i p(i) = 1$$

Cumulative density function



By construction :

$$F(-\infty) = P(\emptyset) = 0$$
$$F(+\infty) = P(\Omega) = 1$$
$$F(a) = \int_{-\infty}^{a} f(x)dx$$
$$P(a < X < b) = F(b) - F(a) = \int_a^b f(x)dx$$

12

SOS School Of Statistics

Change of variable

Probability density function of $Y = \varphi(X)$

For φ bijective

- φ increasing : $X<x \Leftrightarrow Y<y$

$$P(X<x)=F_X(x)=P(Y<y)=F_Y(Y)=F_Y(\varphi(x)) \Rightarrow f_Y(y)=\frac{dF(x)}{dy}=\frac{f(x)}{\varphi'(x)}$$

- φ decreasing : $X<x \Leftrightarrow Y>y$

$$P(X<x)=F_X(x)=P(Y>y)=1-F_Y(Y)=1-F_Y(\varphi(x)) \Rightarrow f_Y(y)=-\frac{dF(x)}{dy}=\frac{f(x)}{-\varphi'(x)}$$

in both case $$\boxed{f_Y(y)=\frac{f(x)}{|\varphi'(x)|}=\frac{f(\varphi^{-1}(y))}{|\varphi'(\varphi^{-1}(y))|}}$$

If φ not bijective : split into several bijective parts φ_i

$$f_Y(y)=\sum_i \frac{f(x)}{|\varphi_i'(x)|}=\sum_i \frac{f(\varphi_i^{-1}(y))}{|\varphi_i'(\varphi_i^{-1}(y))|}$$

Very useful for Monte-Carlo : if X is uniformly distributed between 0 and 1 then $Y=F^{-1}(X)$ has F for cumulative density

13

SOS School Of Statistics

Multidimensional PDF (1)

Random variables can be generalized to random vectors :

$$\vec{X}=(X_1,X_2,\ldots,X_n)$$

the **probability density function** becomes :

$$f(\vec{x})d\vec{x} = f(x_1,x_2,\ldots,x_n)dx_1dx_2\ldots dx_n$$
$$= P(x_1<X_1<x_1+dx_1 \text{ and } x_2<X_2<x_2+dx_2\ldots$$
$$\ldots\text{and } x_n<X_n<x_n+dx_n)$$

and $$P(a<X<b \text{ and } c<Y<d)=\int_a^b dx\int_c^d dy\ f(x,y)$$

Marginal density : probability of only one of the component

$$f_X(x)dx=P(x<X<x+dx \text{ and } -\infty<Y<+\infty)=\int(f(x,y)dx)dy$$

$$\Rightarrow f_X(x)=\int f(x,y)dy$$

14

Multidimensional PDF (2)

For a fixed value of $Y=y_0$:

$f(x|y_0)dx$ = « Probability of $x<X<x+dx$ knowing that $Y=y0$ »

is , a **conditional density** for X. It is proportional to f(x,y), so

$$f(x|y) \propto f(x,y) \qquad \int f(x|y)dx = 1$$

$$\Rightarrow f(x|y) = \frac{f(x,y)}{\int f(x,y)dx} = \frac{f(x,y)}{f_Y(y)}$$

The two random variables X and Y are **independent** if all events of the form **$x<X<x+dx$** are independent from **$y<Y<y+dy$**

$f(x|y)=f_X(x)$ and $f(y|x)=f_Y(y)$ hence **$f(x,y)= f_X(x).f_Y(y)$**

Translated in term of pdf's, Bayes' theorem becomes:

$$f(y|x) = \frac{f(x|y)f_Y(y)}{f_X(x)} = \frac{f(x|y)f_Y(y)}{\int f(x|y)f_Y(y)dy}$$

D.Sivia's lecture will detail the use of this formula for statistical inference

15

Sample PDF

A **sample** is obtained from a **random drawing** within a **population**, described by a probability density function.

We're going to discuss how to **characterize, independently from one another:**

- a **population**
- a **sample**

To this end, it is useful, to consider a sample as a finite set from which one can randomly draw elements, with equipropability

We can the associate to this process a probability density, the **empirical density** or **sample density**

$$f_{sample}(x) = \frac{1}{n}\sum_i \delta(x-i)$$

This density will be useful to translate properties of distribution to a finite sample.

16

Characterizing a distribution

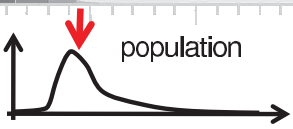
How to reduce a distribution/sample to a finite number of values ?

- **Measure of location**:
  Reducing the distribution to **one central value**
  -> **Result**
- **Measure of dispersion**:
  Spread of the distribution around the central value
  -> **Uncertainty/Error**
- **Higher order measure of shape**

- **Frequency table/histogram** (for a finite sample)

17

Measure of location



Mean value : Sum (integral) of all possible values weighted by the probability of occurrence:

$$\mu = \overline{x} = \int_{-\infty}^{+\infty} x f(x) dx \qquad \mu = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Median : Value that split the distribution in 2 equiprobable parts

$$x_1 \le x_2 \le \ldots \le x_n$$

$$\int_{-\infty}^{med(x)} f(x)dx = \int_{med(x)}^{+\infty} f(x)dx = \frac{1}{2} \quad med(x) = \begin{cases} x_{(n+1)/2} & , \text{ odd } n \\ \frac{1}{2}(x_{n/2} + x_{1+n/2}), & \text{even } n \end{cases}$$

Mode : The most probable value = maximum of pdf

$$\left.\frac{df}{dx}\right|_{x=mod(x)} = 0, \quad \left.\frac{d^2f}{dx^2}\right|_{x=mod(x)} < 0 \qquad ?$$

18

Measure of dispersion

Standard deviation (σ) and variance (v= σ²) : Mean value of the squared deviation to the mean :

$$v = \sigma^2 = \int (x-\mu)^2 f(x)dx \qquad v = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

Koenig's theorem :

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

$$\boxed{\sigma^2 = \overline{x^2} - \mu^2 = \overline{x^2} - \overline{x}^2}$$

Interquartile difference : generalize the median by splitting the distribution in 4 :

$$\int_{-\infty}^{q_1} f(x)dx = \int_{q_1}^{q_2} f(x)dx = \int_{q_2}^{q_3} f(x)dx = \int_{q_3}^{+\infty} f(x)dx = \frac{1}{4} \qquad \mathrm{med}(x) = q_2 \qquad \delta = q_3 - q_1$$

Others…

19

Bienaymé-Chebyshev

Consider the interval : $\Delta =]-\infty, \mu - a[\cup]\mu + a, +\infty[$

Then for $x \in \Delta$: $\left(\frac{x-\mu}{a}\right)^2 > 1 \Rightarrow \left(\frac{x-\mu}{a}\right)^2 f(x) > f(x)$

$$\Rightarrow \int_\Delta \left(\frac{x-\mu}{a}\right)^2 f(x)dx > \int_\Delta f(x)dx$$

$$\Rightarrow \int_{-\infty}^{+\infty} \left(\frac{x-\mu}{a}\right)^2 f(x)dx > \int_\Delta f(x)$$

$$\Rightarrow \frac{\sigma^2}{a^2} > P(|X-\mu| > a)$$

Finally **Bienaymé-Chebyshev's inequality** $\boxed{P(|X-\mu| \le a\sigma) > 1 - \frac{1}{a^2}}$

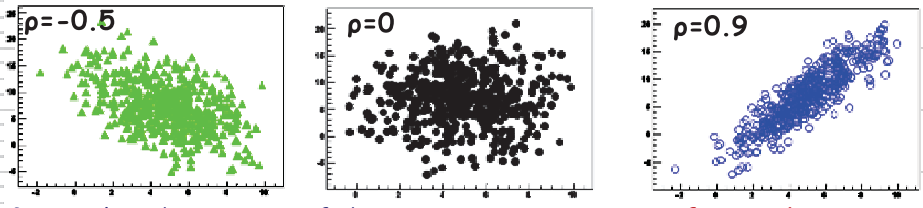
It gives a bound on the **confidence level** if the interval $\mu \pm a\sigma$

a	1	2	3	4	5
Chebyshev's bound	0	0.75	0.889	0.938	0.96
Normal distribution	0.683	0.954	0.997	0.99996	0.9999994

20

Multidimensional case

A **random vector (X,Y)** can be treated as **2 separate variables**
marginal densities : mean and variance for each variable : $\mu_X\ \mu_Y\ \sigma_X\ \sigma_Y$
Doesn't take into account correlations between the variables



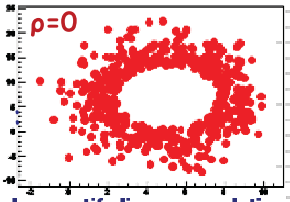
Generalized measure of dispersion : **Covariance of X and Y**

$$Cov(X,Y) = \iint (x-\mu_X)(y-\mu_Y)f(x,y)dxdy = \rho\sigma_X\sigma_Y = \mu_{XY} - \mu_X\mu_Y$$

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_X)(y_i-\mu_Y)$$

Correlation : $\rho = \dfrac{Cov(X,Y)}{\sigma_X\sigma_Y}$ **Uncorrelated** : $\rho=0$

Independent ⇄ **Uncorrelated**



only quantify linear correlation

21

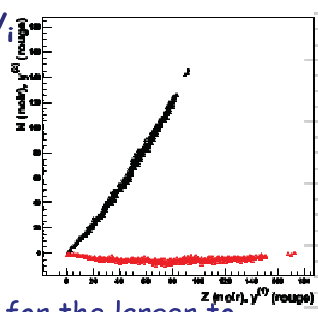
Decorrelation

Covariance matrix for **n** variables X_i:

$$\Sigma_{ij} = Cov(X_i,X_j) \Rightarrow \Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{2n}\sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1n}\sigma_1\sigma_n & \rho_{2n}\sigma_2\sigma_n & \cdots & \sigma_n^2 \end{pmatrix}$$

For **uncorrelated variables** **Σ is diagonal**
Matrix **real** and **symmetric** : can de diagonalized
On can define **n** new uncorrelated variables Y_i

$$\Sigma' = \begin{pmatrix} \sigma_1'^2 & 0 & \cdots & 0 \\ 0 & \sigma_2'^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n'^2 \end{pmatrix} = B^{-1}\Sigma B, \qquad Y = BX$$



σ'^2_i are the **eigenvalues** of **Σ**,
B contains the **orthonormal eigenvectors**.
The Y_i are the **principal components**. Sorted for the larger to the smaller **σ'** they allow **dimensional reduction**

22

Regression

Measure of location:

- a point : $(\mu_X\ ,\ \mu_Y)$
- a curve : line closest to the points -> **linear regression**

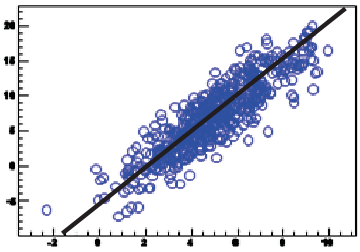
Minimizing the dispersion between the curve « y=ax+b » and the distribution :

$$w(a,b)=\iint (y-ax-b)^2 f(x,y)dxdy\left(=\frac{1}{n}\sum_i (y_i-ax_i-b)^2\right)$$

$$\begin{cases}\dfrac{\partial w}{\partial a}=0=\iint x(y-ax-b)f(x,y)dxdy\\ \dfrac{\partial w}{\partial b}=0=\iint (y-ax-b)f(x,y)dxdy\end{cases}$$

$$\Leftrightarrow\begin{cases}a(\sigma_X^2-\mu_X^2)+b\mu_X=\rho\sigma_X\sigma_Y+\mu_X\mu_Y\\ a\mu_X+b=\mu_Y\end{cases}$$

$$\Leftrightarrow\boxed{\begin{cases}a=\rho\dfrac{\sigma_Y}{\sigma_X}\\ b=\mu_Y-\rho\dfrac{\sigma_Y}{\sigma_X}\mu_X\end{cases}}$$



Fully correlated $\rho=1$

Fully anti-correlated $\rho=-1$

Then Y = aX+b

Moments

For any function **g(x)**, the **expectation** of **g** is :

$$E[g(X)]=\int g(x)f(x)dx$$ It's the mean value of g

Moments μ_k are the expectation of X^k.

0th moment : $\mu_0=1$ (pdf normalization)

1st moment : $\mu_1=\mu$ (mean)

$X' = X-\mu_1$ is a **central variable**

2nd central moment : $\mu'_2=\sigma^2$ (variance)

Characteristic function $\varphi(t)=E[e^{ixt}]=\int f(x)e^{ixt}dx=FT^{-1}[f]$

From Taylor expansion : $\varphi(t)=\int\sum_k\frac{(itx)^k}{k!}f(x)dx=\sum_k\frac{(it)^k}{k!}\mu_k$

$$\boxed{\mu_k=-i^k\left.\frac{d^k\varphi}{dt^k}\right|_{t=0}}$$

Pdf entirely defined by its moments

CF : useful tool for demonstrations

Skewness and kurtosis

Reduced variable : $X'' = (X-\mu)/\sigma = X'/\sigma$

Measure of asymmetry :

3rd reduced moment : $\mu''_3 = \sqrt{\beta_1} = \gamma_1$: **skewness**

$\gamma_1=0$ for symmetric distribution. Then **mean = median**

Measure of shape :

4th reduced moment : $\mu''_4 = \beta_2 = \gamma_2 + 3$: **kurtosis**

For the normal distribution $\beta_2 = 3$ and $\gamma_2 = 0$

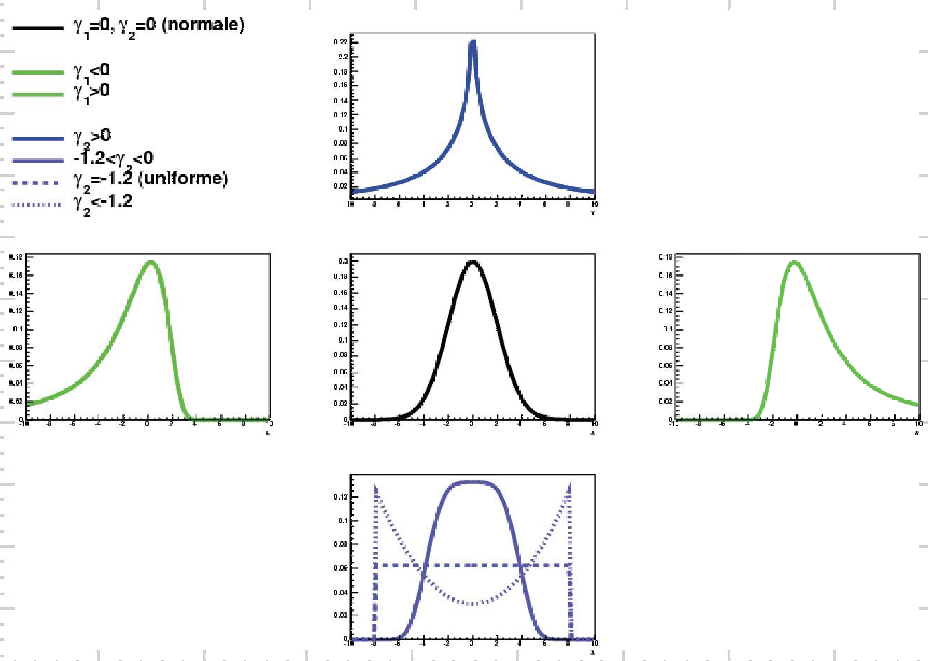
Generalized Koenig's theorem

$$\mu'_n = (-1)^n (1-n)\mu_1^n + \sum_{k=2}^{n} \frac{n!}{k!(n-k)!}(-\mu_1)^{n-k}\mu_k$$

$$\mu''_n = \left(\frac{1}{\mu'_2}\right)^{n-2} \mu'_n$$

25

Skewness and kurtosis (2)



26

Discrete distributions

Binomial distribution: randomly choosing **K** objects within a finite set of **n**, with a fixed drawing probability of **p**

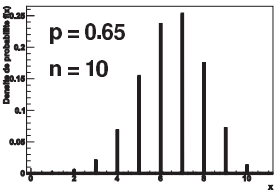
Variable : K

Parameters : n, p

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

Mean : np

Variance : np(1-p)



Poisson distribution : limit of the binomial when $n \to +\infty, p \to 0, np = \lambda$

Counting events with fixed probability per time/space unit.

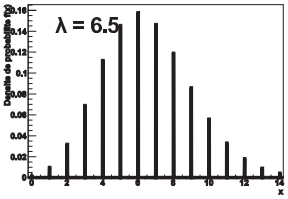
Variable : K

Parameters : λ

Law : $P(k;\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$

Mean : λ

Variance : λ



27

Real distributions

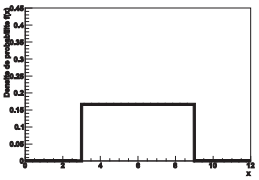
Uniform distribution : equiprobability over a finite range [a,b]

Parameters : a, b

Law : $f(x;a,b) = \frac{1}{b-a}$ if $a < x < b$

Mean : $\mu = (a+b)/2$

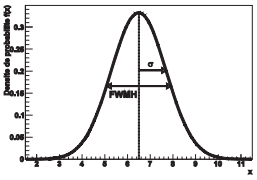
Variance : $v = \sigma^2 = (b-a)^2/12$



Normal distribution (Gaussian) : limit of many processes

Parameters : μ, σ

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

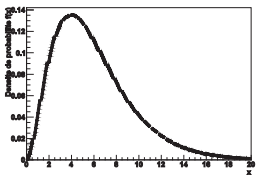


Chi-square distribution : sum of the square of **n** normal reduced variables

Variable : $C = \sum_{k=1}^{n} \left(\frac{X_k - \mu_{X_k}}{\sigma_{X_k}} \right)^2$

Parameters : n

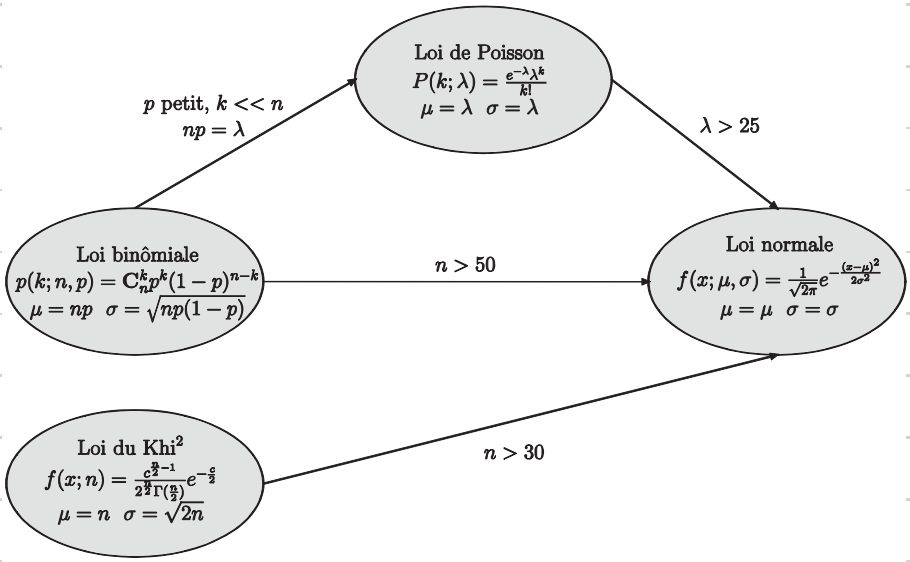
Law : $f(c;n) = c^{\frac{n}{2}-1} e^{-\frac{c}{2}} \Big/ 2^{\frac{n}{2}-1} \Gamma\left(\frac{n}{2}\right)$



Mean : n Variance : 2n

28

Convergence



29

Multidimensional Pdfs

Multinomial distribution : randomly choosing K_1 , K_2 ,... K_s objects within a finite set of n, with a fixed drawing probability for each category p_1, p_2,... p_s with $\Sigma K_i = n$ and $\Sigma p_i = 1$

Parameters : n, p_1, p_2,... p_s

Law : $P(\vec{k};n,\vec{p}) = \frac{n!}{k_1!k_2!\ldots k_s!} p_1^{k_1} p_2^{k_2} \ldots p_s^{k_s}$

Mean : $\mu_i = np_i$

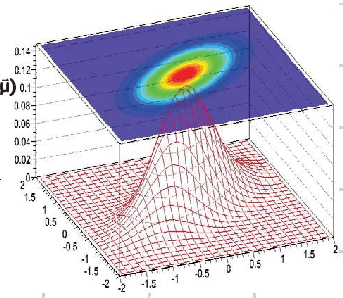
Variance : $\sigma_i^2 = np_i(1-p_i)$ $\qquad$ $Cov(K_i,K_j) = -np_ip_j$

Rem : **variables are not independent.** The binomial, correspond to $s=2$, but has only one independent variable.

Multinormal distribution :

Parameters : $\vec{\mu}, \Sigma$

Law : $f(\vec{x};\vec{\mu},\Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})}$



if **uncorrelated** $f(\vec{x};\vec{\mu},\Sigma) = \prod \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}}$

Independent $\Longleftrightarrow$ **Uncorrelated**

30

Sum of random variables

The sum of several random variable is a new random variable S

$$S = \sum_{i=1}^{n} X_i$$

Assuming the mean and variance of each variable exists,

Mean value of S :

$$\mu_S = \int \left(\sum_{i=1}^{n} x_i \right) f(x_1, \ldots, x_n) dx_1 \ldots dx_n = \sum_{i=1}^{n} \int x_i f_{X_i}(x_i) dx_i = \sum_{i=1}^{n} \mu_i$$

The mean is an additive quantity

Variance of S :

$$\sigma_S^2 = \int \left(\sum_{i=1}^{n} x_i - \mu_{X_i} \right)^2 f(x_1, \ldots, x_n) dx_1 \ldots dx_n$$

$$= \sum_{i=1}^{n} \sigma_{X_i}^2 + 2 \sum_i \sum_{j<i} Cov(X_i, X_j)$$

For uncorrelated variables,

$$\sigma_S^2 = \sum_{k=1}^{n} \sigma_{X_i}^2$$

the variance is additive -> used for error combinations

31

Sum of random variables

Probability density function of S : $f_S(s)$

Using the characteristic function :

$$\varphi_S(t) = \int f_S(s) e^{ist} ds = \int f_{\vec{X}}(\vec{x}) e^{it \sum x_i} d\vec{x}$$

For **independent variables**

$$\varphi_S(t) = \prod \int f_{X_k}(x_k) e^{itx_k} dx_k = \prod \varphi_{X_i}(t)$$

The characteristic function factorizes.

Finally the pdf is the **Fourier transform** of the cf, so :

$$f_S = f_{X_1} * f_{X_2} * \ldots * f_{X_n}$$

The pdfs of the sum is a **convolution**.

Sum of Normal variables -> Normal

Sum of Poisson variables (λ_1 and λ_2) -> Poisson, $\lambda = \lambda_1 + \lambda_2$

Sum of Khi-2 variables (n_1 and n_2) -> Khi-2, $n = n_1 + n_2$

32

Sum of independent variables

Weak law of large numbers

Sample of size **n** = realization of **n** independent variables, with the **same distribution** (mean **μ**, variance **σ²**).

The **sample mean** is a realization of $M = \frac{S}{n} = \frac{1}{n}\sum X_i$

Mean value of **M** : $\mu_M = \mu$ **Variance** of **M** : $\sigma_M^2 = \sigma^2/n$

From Bienaymé-Chebyshev : $P(|M-\mu| > a) < \frac{\sigma^2}{na^2} \xrightarrow[n\to+\infty]{} 0 \ (\forall a)$

Central-Limit theorem

n independent random variables of mean μ_i and variance σ_i^2

Sum of the reduced variables : $C = \frac{1}{\sqrt{n}}\sum \frac{X_i - \mu_i}{\sigma_i}$

The pdfs of **C** converge to a reduced normal distribution :

$$f_C(c) \xrightarrow[n\to+\infty]{} \frac{1}{\sqrt{2\pi}} e^{-\frac{c^2}{2}}$$

33

Central limit theorem

Naive demonstration:

For each X_i : X''_i has mean 0 and variance 1. So its characteristic function is :

$$\varphi_{X''_i}(t) = 1 - \frac{t^2}{2} + o(t^2)$$

Hence the characteristic function of **C** :

$$\varphi_C(t) = \varphi_{X''_i}\left(\frac{t}{\sqrt{n}}\right)^n = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n$$

For **n** large :

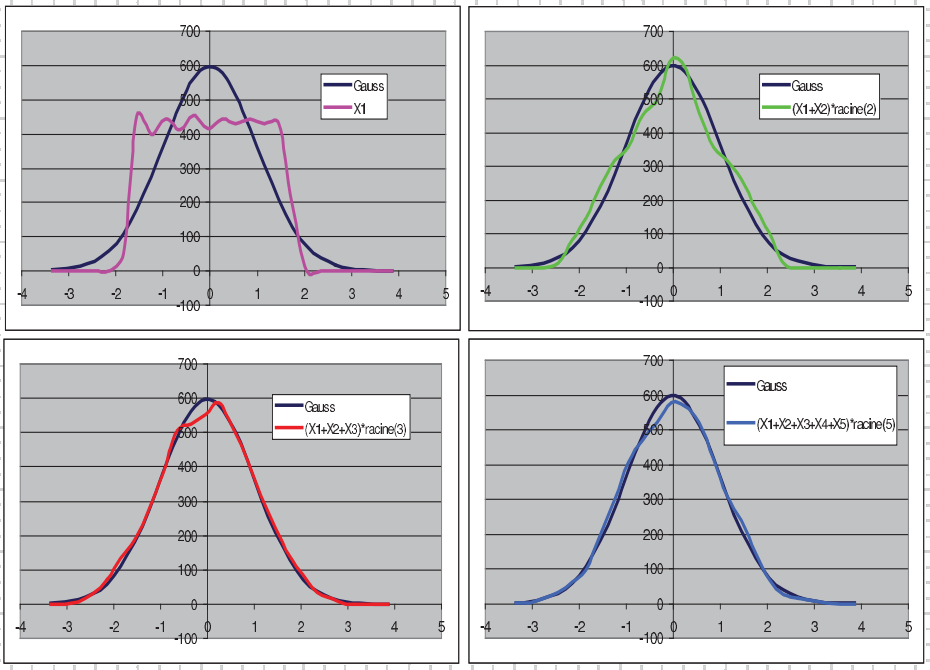
$$\lim_{n\to+\infty} \varphi_C(t) = \lim_{n\to+\infty}\left(1 - \frac{t^2}{2n}\right)^n = e^{-\frac{t^2}{2}} = FT^{-1}[f_C]$$

This is a naive demonstration, because we assumed that the moments were defined.

For CLT, only mean and variance are required (much more complex)

34

Central limit theorem



35

Dispersion and uncertainty

Any measure (or combination of measure) is a realization of a random variable.

- Measured value : θ
- True value : θ_0

Uncertainty = quantifying the difference between θ and θ_0 :

->measure of dispersion

We will postulate : $\Delta\theta = \alpha\sigma_\theta$ Absolute error, always positive

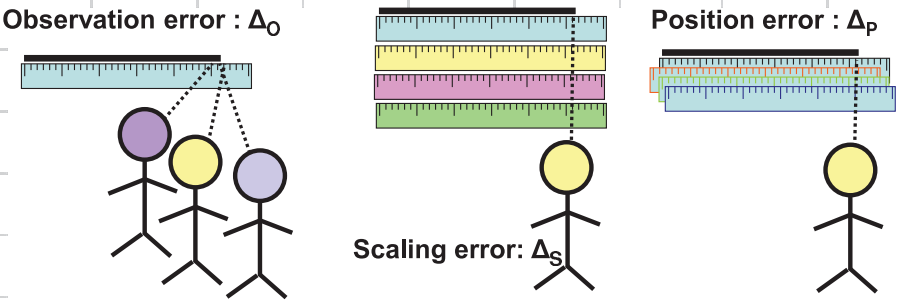
Usually one separate

- **Statistical error** : due to the measurement Pdf.
- **Systematic errors** or bias -> fixed but unknown deviation (equipment, assumptions,…)

Systematic errors can be seen as statistical error in a set a similar experiences.

36

Error sources



$$\theta = \theta_0 + \delta_O + \delta_S + \delta_P$$

Each δ_i is a realization of a random variable : mean 0 (negligible) and variance σ_i^2. For **uncorrelated error sources** :

$$\left.\begin{aligned}\Delta_O &= \alpha\sigma_O\\ \Delta_S &= \alpha\sigma_S\\ \Delta_P &= \alpha\sigma_P\end{aligned}\right\}\Delta_{tot}^2 = (\alpha\ \sigma_{tot})^2 = \alpha^2(\sigma_O{}^2 + \sigma_S{}^2 + \sigma_P{}^2) = \Delta_O^2 + \Delta_S^2 + \Delta_P^2$$

Choice of α ?

If many sources, from central-limit -> normal distribution

α=1 gives (approximately) a 68% confidence interval

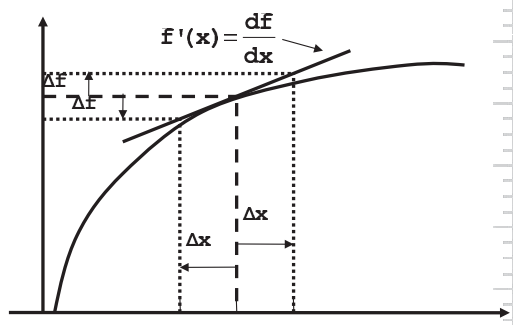
α=2 gives 95% CL (and at least 75% from Bienaymé-Chebyshev)

37

Error propagation

Measure : $x \pm \Delta x$

Compute : $f(x)$ -> Δf ?



Assuming **small errors**, using Taylor expansion :

$$f(x+\Delta x) = f(x) + \frac{df}{dx}\Delta x + \frac{1}{2}\frac{d^2f}{dx^2}\Delta x^2\left(+\frac{1}{6}\frac{d^3f}{dx^3}\Delta x^3 + \frac{1}{24}\frac{d^4f}{dx^4}\Delta x^4\right)$$

$$f(x-\Delta x) = f(x) - \frac{df}{dx}\Delta x + \frac{1}{2}\frac{d^2f}{dx^2}\Delta x^2\left(-\frac{1}{6}\frac{d^3f}{dx^3}\Delta x^3 + \frac{1}{24}\frac{d^4f}{dx^4}\Delta x^4\right)$$

$$\Rightarrow \Delta f = \frac{1}{2}\left|f(x+\Delta x) - f(x-\Delta x)\right| = \frac{df}{dx}\Delta x\left(+\frac{1}{6}\frac{d^3f}{dx^3}\Delta x^3\right)$$

38

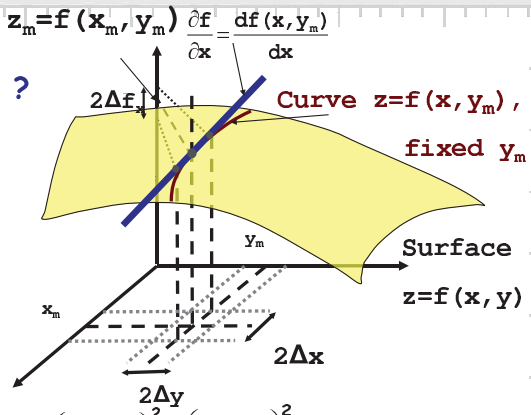
Error propagation

Measure : $x \pm \Delta x$, $y \pm \Delta y$, ...

Compute : $f(x,y,\ldots) \rightarrow \Delta f$?

Idea : treat the effect of each variable as separate **error sources**

$$\Delta_x f = \left|\frac{\partial f}{\partial x}\right| \Delta x \ , \ \Delta_y f = \left|\frac{\partial f}{\partial y}\right| \Delta y$$



Then

$$\Delta f^2 = \Delta_x f^2 + \Delta_y f^2 + \rho_{xy} \Delta_x f \Delta_y f = \left(\frac{\partial f}{\partial x}\Delta x\right)^2 + \left(\frac{\partial f}{\partial y}\Delta y\right)^2 + \rho_{xy}\frac{\partial f}{\partial x}\frac{\partial f}{\partial y}\Delta x \Delta y$$

$$\boxed{\Delta f^2 = \sum_i \left(\frac{\partial f}{\partial x_i}\Delta x_i\right)^2 + \sum_{i,j<i} \rho_{x_i x_j} \left|\frac{\partial f}{\partial x_i}\frac{\partial f}{\partial x_j}\right| \Delta x_i \Delta x_j}$$

uncorrelated

$$\Delta f^2 = \sum_i \left(\frac{\partial f}{\partial x_i}\Delta x_i\right)^2$$

correlated

$$\Delta f = \left|\frac{\partial f}{\partial x}\right| \Delta x + \left|\frac{\partial f}{\partial y}\right| \Delta y$$

anticorrelated

$$\Delta f = \left|\left|\frac{\partial f}{\partial x}\right| \Delta x - \left|\frac{\partial f}{\partial y}\right| \Delta y\right|$$

Probability and Statistics

Basic concepts II

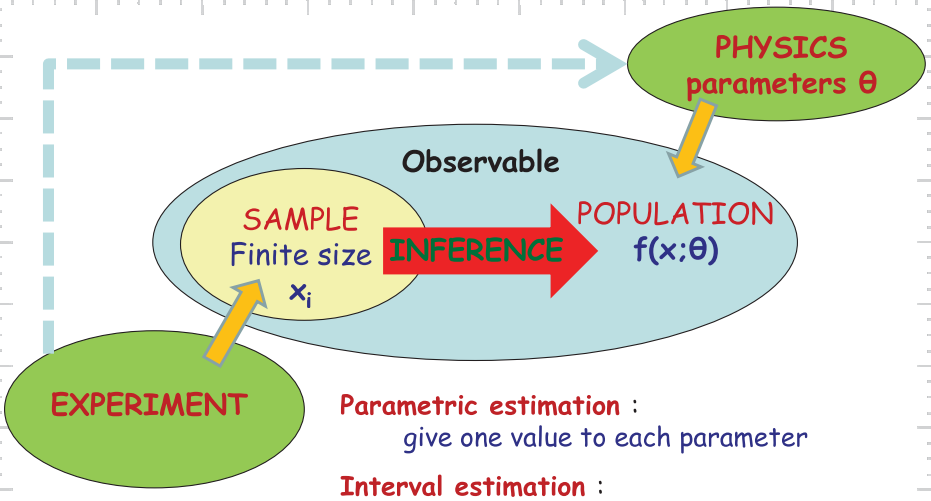
(from a physicist point of view)

Benoit CLEMENT – Université J. Fourier / LPSC

bclement@lpsc.in2p3.fr

School Of Statistics

Statistics



Parametric estimation :
give one value to each parameter

Interval estimation :
derive a interval that probably contains the true value

Non-parametric estimation :
estimate the full pdf of the population

2

Parametric estimation

From a finite sample $\{x_i\}$ -> estimating a parameter θ

Statistic = a function $S = f(\{x_i\})$

Any statistic can be considered as an **estimator** of θ

To be a good estimator it needs to satisfy :

- **Consistency** : limit of the estimator for a infinite sample.
- **Bias** : difference between the estimator and the true value
- **Efficiency** : speed of convergence
- **Robustness** : sensitivity to statistical fluctuations

A **good estimator** should at least be **consistent** and **asymptotically unbiased**

Efficient / Unbiased / Robust often contradict each other

$\Rightarrow$**different choices for different applications**

3

Bias and consistency

As the sample is a set of realization of random variables (or one vector variable), so is the estimator :

$\hat{\theta}$ **is a realization of** $\hat{\Theta}$

it has a mean, a variance,... and a probability density function

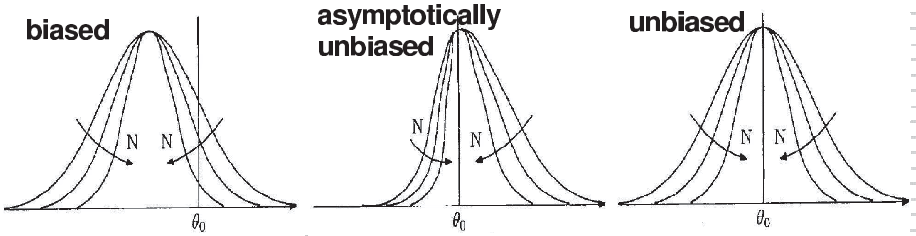
Bias : Mean value of the estimator $b(\hat{\theta}) = E[\hat{\Theta} - \theta_0] = \mu_{\hat{\Theta}} - \theta_0$

unbiased estimator : $b(\hat{\theta}) = 0$

asymptotically unbiased : $b(\hat{\theta}) \xrightarrow[n\to+\infty]{} 0$

Consistency: formally $P(|\hat{\theta} - \theta| > \varepsilon) \xrightarrow[n\to+\infty]{} 0, \forall\varepsilon$

in practice, if **asymptotically unbiased** $\boxed{\sigma_{\hat{\Theta}} \xrightarrow[n\to+\infty]{} 0}$



4

SOS School Of Statistics

Empirical estimator

Sample mean is a good estimator of the **population mean**

-> **weak law of large numbers** : convergent, unbiased

$$\hat{\mu} = \frac{1}{n}\sum x_i, \qquad \mu_{\hat{\mu}} = E[\hat{\mu}] = \mu, \qquad \sigma^2_{\hat{\mu}} = E[(\hat{\mu}-\mu)^2] = \frac{\sigma^2}{n}$$

Sample variance as an estimator of the population variance :

$$\hat{s}^2 = \frac{1}{n}\sum_i (x_i - \hat{\mu})^2 = \left(\frac{1}{n}\sum_i (x_i-\mu)^2\right) - (\mu - \hat{\mu})^2$$

$$E[\hat{s}^2] = \left(\frac{1}{n}\sum_i \sigma^2\right) - \sigma^2_{\hat{\mu}} = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$$

biased, asymptotically unbiased

unbiased variance estimator : $\boxed{\hat{\sigma}^2 = \frac{1}{n-1}\sum_i (x_i - \hat{\mu})^2}$

variance of the estimator (convergence) $\sigma^2_{\hat{\sigma}^2} = \frac{\sigma^4}{n-1}\left(\frac{n-1}{n}\gamma_2 + 2\right) \rightarrow \frac{2\sigma^4}{n}$

5

SOS School Of Statistics

Errors on these estimator

Uncertainty ⇔ Estimator standard deviation

Use an estimator of standard deviation : $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ (!!! Biased)

Mean : $\hat{\mu} = \frac{1}{n}\sum x_i, \qquad \sigma^2_{\hat{\mu}} = \frac{\sigma^2}{n} \quad \Rightarrow \quad \boxed{\Delta\hat{\mu} = \sqrt{\frac{\hat{\sigma}^2}{n}}}$

Variance : $\hat{\sigma}^2 = \frac{1}{n-1}\sum_i (x_i - \hat{\mu})^2, \ \sigma^2_{\hat{\sigma}^2} \approx \frac{2\sigma^4}{n} \quad \Rightarrow \quad \boxed{\Delta\hat{\sigma}^2 = \sqrt{\frac{2}{n}}\hat{\sigma}^2}$

Central-Limit theorem -> empirical estimators of mean and variance are **normally distributed**, for **large enough samples**

$\hat{\mu} \pm \Delta\hat{\mu} \quad ; \hat{\sigma}^2 \pm \Delta\hat{\sigma}^2$ define 68% confidence intervals

6

Likelihood function

Generic function $k(x,\theta)$

x : random variable(s)

θ : parameter(s)

fix $\theta = \theta_0$ (true value)

fix $x = u$ (one realization of the random variable)

Probability density function

$$f(x;\theta) = k(x,\theta_0)$$

$$\int f(x;\theta)\,dx = 1$$

for Bayesian $f(x|\theta) = f(x;\theta)$

Likelihood function

$$\mathcal{L}(\theta) = k(u,\theta)$$

$$\int \mathcal{L}(\theta)\,d\theta = ???$$

for Bayesian $f(\theta|x) = \mathcal{L}(\theta)/\int\mathcal{L}(\theta)d\theta$

For a **sample** : **n** independent realizations of the same variable **X**

$$\mathcal{L}(\theta) = \prod_i k(x_i,\theta) = \prod_i f(x_i;\theta)$$

7

Estimator variance

Start from the generic k function, differentiate twice, with respect to θ, the pdf normalization condition: $1 = \int k(x,\theta)dx$

$$0 = \int \frac{\partial k}{\partial\theta}dx = \int k\frac{\partial \ln k}{\partial\theta}dx = E\left[\frac{\partial \ln k}{\partial\theta}\right] \Rightarrow \boxed{(b+\theta)E\left[\frac{\partial \ln k}{\partial\theta}\right] = 0}$$

$$0 = \int \frac{\partial^2 k}{\partial\theta^2}dx = \int k\frac{\partial^2 \ln k}{\partial\theta^2}dx + \int k\left(\frac{\partial \ln k}{\partial\theta}\right)^2 dx \Rightarrow E\left[\frac{\partial^2 \ln k}{\partial\theta^2}\right] = -E\left[\left(\frac{\partial \ln k}{\partial\theta}\right)^2\right]$$

Now differentiating the estimator bias : $\theta + b = \int \hat{\theta}(x)k(x,\theta)dx$

$$1 + \frac{\partial b}{\partial\theta} = \frac{\partial}{\partial\theta}\int \hat{\theta}(x)k(x,\theta)dx = \int \hat{\theta}\frac{\partial k}{\partial\theta}dx = \int \hat{\theta}k\frac{\partial \ln k}{\partial\theta}dx = \int (\hat{\theta}\boxed{-b-\theta})k\frac{\partial \ln k}{\partial\theta}dx$$

Finally, using Cauchy-Schwartz inequality

$$\left(1+\frac{\partial b}{\partial\theta}\right)^2 \le \int (\hat{\theta}-b-\theta)^2 k dx \int k\left(\frac{\partial \ln k}{\partial\theta}\right)^2 dx \Rightarrow \boxed{\sigma^2_{\hat{\theta}} \ge \frac{(1+b')^2}{E\left[\left(\frac{\partial \ln k}{\partial\theta}\right)^2\right]}}$$

Cramer-Rao bound

8

Efficiency

For **any unbiased estimator of θ**, the variance cannot exceed :

$$\sigma^2_{\hat{\theta}} \geq \frac{1}{E\left[\left(\frac{\partial \ln \mathcal{L}}{\partial \theta}\right)^2\right]} \left(= \frac{-1}{E\left[\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right]} \right)$$

The **efficiency** of a convergent estimator, is given by its **variance**.

An **efficient estimator** reaches the Cramer-Rao bound (at least asymptotically) : Minimal variance estimator

MVE will often be biased, asymptotically unbiased

9

Maximum likelihood

For a sample of measurements, $\{x_i\}$

The **analytical form of the density is known**

It depends on **several unknown parameters θ**

eg. event counting : Follow a Poisson distribution, with a parameter that depends on the physics : $\lambda_i(\theta)$

$$\mathcal{L}(\theta) = \prod_i \frac{e^{\lambda_i(\theta)} \lambda_i(\theta)^{x_i}}{x_i!}$$

An estimator of the parameters of θ, are the ones that **maximize of observing the observed result.**

-> Maximum of the likelihood function

$$\left.\frac{\partial \mathcal{L}}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0$$

rem : system of equations for several parameters

rem : **often minimize** $-\ln\mathcal{L}$: simplify expressions

10

Properties of MLE

Mostly **asymptotic properties** : valid for large sample, often assumed in any case for lack of better information

Asymptotically **unbiased**

Asymptotically **efficient** (reaches the CR bound)

Asymptotically **normally distributed**

-> Multinormal law, with covariance given by generalization of CR Bound :

$$f(\hat{\vec{\theta}};\vec{\theta},\Sigma)=\frac{1}{\sqrt{2\pi|\Sigma|}}e^{-\frac{1}{2}(\hat{\vec{\theta}}-\vec{\theta})^T\Sigma^{-1}(\hat{\vec{\theta}}-\vec{\theta})} \qquad \Sigma_{ij}^{-1}=-E\left[\frac{\partial \ln\mathcal{L}}{\partial\theta_i}\frac{\partial \ln\mathcal{L}}{\partial\theta_j}\right]$$

Goodness of fit = The value of $-2\ln L(\hat{\theta})$ is Khi-2 distributed, with

ndf = sample size – number of parameters

$$p-\text{value}=\int_{-2\ln\mathcal{L}(\hat{\theta})}^{+\infty} f_{\chi^2}(x;ndf)dx$$

Probability of getting a worse agreement

11

Errors on MLE

$$f(\hat{\vec{\theta}};\vec{\theta},\Sigma)=\frac{1}{\sqrt{2\pi|\Sigma|}}e^{-\frac{1}{2}(\hat{\vec{\theta}}-\vec{\theta})^T\Sigma^{-1}(\hat{\vec{\theta}}-\vec{\theta})} \qquad \Sigma_{ij}^{-1}=-E\left[\frac{\partial \ln\mathcal{L}}{\partial\theta_i}\frac{\partial \ln\mathcal{L}}{\partial\theta_j}\right]$$

Errors on parameter -> from the covariance matrix

For **one parameter**, 68% interval $\Delta\theta=\hat{\sigma}_{\hat{\theta}}=\sqrt{\frac{-1}{\frac{\partial^2\ln\mathcal{L}}{\partial\theta^2}}}$

only one realization of the estimator -> empirical mean of 1 value…

More generally :

$$\Delta\ln\mathcal{L}=\ln\mathcal{L}(\hat{\theta})-\ln\mathcal{L}(\theta)=\frac{1}{2}\sum_{i,j}\Sigma_{ij}^{-1}(\theta_i-\hat{\theta}_i)(\theta_j-\hat{\theta}_j)+O(\theta^3)$$

Confidence contour are defined by the equation : $\Delta\ln\mathcal{L}=\beta(n_\theta,\alpha)$ with $\alpha=\int_0^{2\beta}f_{\chi^2}(x;n_\theta)dx$

Values of β for different **number of parameters n_θ** and **confidence levels α**

$n_\theta\rightarrow$ $\alpha\downarrow$	1 (0. 5*n_θ^2)	2	3
68.3	0.5	1.15	1.76
95.4	2	3.09	4.01
99.7	4.5	5.92	7.08

12

Least squares

Set of measurements $(x_i,\ y_i)$ with uncertainties on y_i

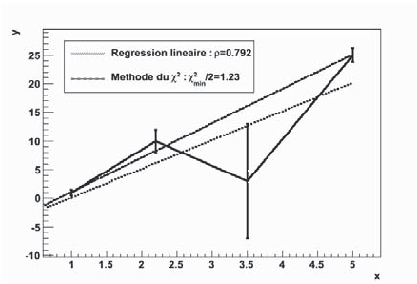
Theoretical law : $y = f(x,\theta)$

Naïve approach : use **regression**

$$w(\theta) = \sum_i (y_i - f(x_i,\theta))^2, \qquad \frac{\partial w}{\partial \theta_i} = 0$$

Reweight each term by the error

$$K^2(\theta) = \sum_i \left(\frac{y_i - f(x_i,\theta)}{\Delta y_i} \right)^2, \qquad \frac{\partial K^2}{\partial \theta_i} = 0$$



Maximum likelihood : assume each y_i is normally distributed with a mean equal to $f(x_i,\theta)$ and a variance equal to Δy_i

Then the **likelihood** is : $\mathcal{L}(\theta) = \prod_i \frac{1}{\sqrt{2\pi}\Delta y_i} e^{-\frac{1}{2}\left(\frac{y_i - f(x_i,\theta)}{\Delta_{y_i}}\right)^2}$

$$\frac{\partial \mathcal{L}}{\partial \theta} = 0 \Leftrightarrow -2\frac{\partial \ln \mathcal{L}}{\partial \theta} = \frac{\partial K^2}{\partial \theta} = 0$$

Least squares or **Khi-2 fit** is the MLE, for Gaussian errors

Generic case with correlations: $K^2(\vec{\theta}) = \frac{1}{2}(\vec{y} - \vec{f}(x,\vec{\theta})^T \Sigma^{-1} (\vec{y} - \vec{f}(x,\vec{\theta})$

13

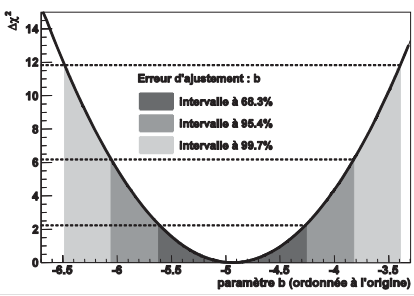
Example : fitting a line

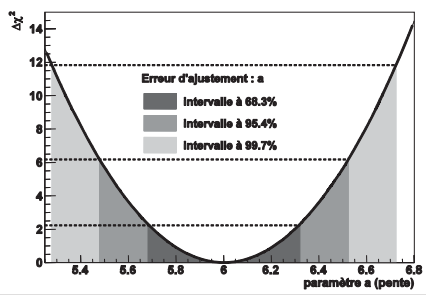
For $f(x)=ax+b$

$$A = \sum_i \frac{x_i y_i}{\Delta y_i^2},\ B = \sum_i \frac{x_i^2}{\Delta y_i^2},\ C = \sum_i \frac{x_i}{\Delta y_i^2},\ D = \sum_i \frac{y_i}{\Delta y_i^2},\ E = \sum_i \frac{1}{\Delta y_i^2}$$

$$\hat{a} = \frac{AE - DC}{BE - C^2}, \qquad \hat{b} = \frac{DB - AC}{BE - C^2}$$

$$\Delta\hat{a} = \frac{1.52}{\sqrt{B}}, \qquad \Delta\hat{b} = \frac{1.52}{\sqrt{E}}$$

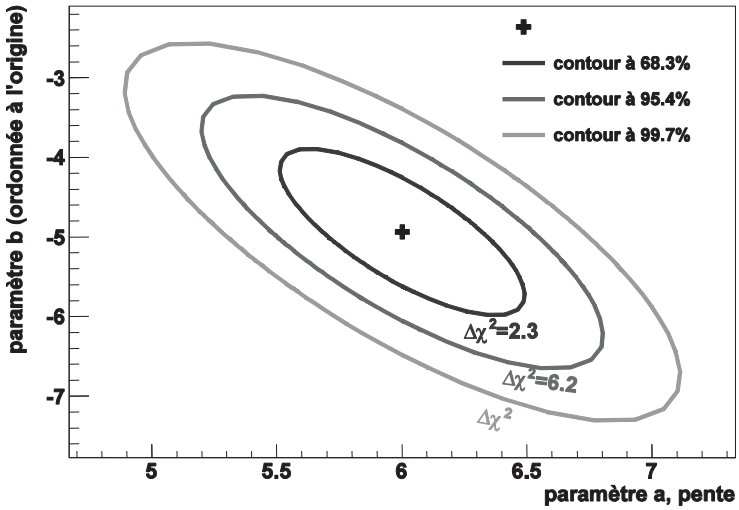




14

Example : fitting a line

2 dimensional error contours on **a** and **b**



15

Confidence interval

For a random variable, a **confidence interval** with **confidence level** **α**, is any interval **[a,b]** such as :

$$P(X \in [a,b]) = \int_a^b f_X(x)dx = \alpha$$

Probability of finding a realization inside the interval

Generalization of the concept of uncertainty:

interval that contains the **true value** with a given probability

-> slightly different concepts

For **Bayesians** : the posterior density is the probability density of the true value. It can be used to derive interval :

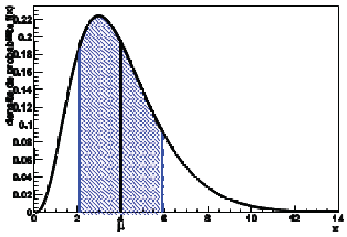
$$P(\theta \in [a,b]) = \alpha$$

No such thing for a **Frequentist** : The interval itself becomes the random variable **[a,b]** is a realization of **[A,B]**

$$P(A < \theta \text{ and } B > \theta) = \alpha$$ Independently of **θ**

16

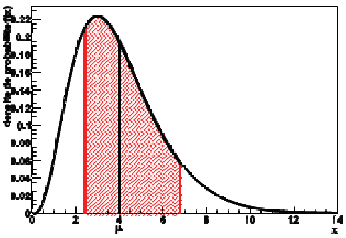
Confidence interval



Mean centered, symetric interval

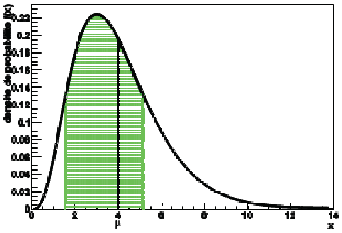
$[\mu-\alpha,\ \mu+\alpha]$

$$\int_{\mu-\alpha}^{\mu+\alpha} f(x)dx = \alpha$$



Mean centered, probability symetric interval : $[a,\ b]$

$$\int_a^{\mu} f(x)dx = \int_{\mu}^{b} f(x)dx = \frac{\alpha}{2}$$



Highest Probability Density (HDP)

$[a,\ b]\quad \int_a^b f(x)dx = \alpha$

$f(x) > f(y)$ for $x \in [a,b]$ and $y \notin [a,b]$

17

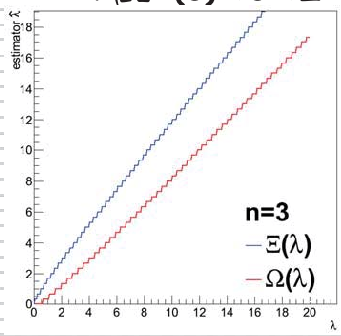
Confidence Belt

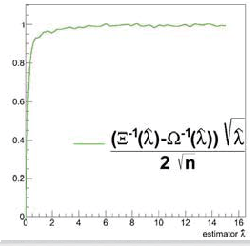
To build a frequentist interval for an estimator $\hat{\theta}$ of θ

1. Make pseudo-experiments for several values of θ and compute he estimator $\hat{\theta}$ for each (MC sampling of the estimator pdf)
2. For each θ, determine $A(\theta)$ and $B(\theta)$ such as :
   $\hat{\theta} < \Xi(\theta)$ for a fraction $(1-\alpha)/2$ of the pseudo-experiments
   $\hat{\theta} > \Omega(\theta)$ for a fraction $(1-\alpha)/2$ of the pseudo-experiments
   These 2 curves are the **confidence belt**, for a **CL α**.
3. Inverse these functions. The interval $[\Omega^{-1}(\hat{\theta}),\ \Xi^{-1}(\hat{\theta})]$ satisfy:

$$P\left(\Omega^{-1}(\hat{\theta}) < \theta < \Xi^{-1}(\hat{\theta})\right) = 1 - P\left(\Xi^{-1}(\hat{\theta}) < \theta\right) - P\left(\Omega^{-1}(\hat{\theta}) > \theta\right)$$

$$= 1 - P\left(\hat{\theta} < \Xi(\theta)\right) - P\left(\hat{\theta} > \Omega(\theta)\right) = \alpha$$





Confidence Belt for Poisson parameter λ estimated with the empirical mean of 3 realizations (68%CL)

18

Dealing with systematics

The variance of the estimator only measure the statistical uncertainty.

Often, we will have to deal with some **parameters** whose **values are known with limited precision.**

Systematic uncertainties

The likelihood function becomes :

$$\mathcal{L}(\theta, v) \qquad v = v_0 \pm \Delta v \text{ or } v_{0-\Delta v_-}^{+\Delta v_+}$$

The known parameters **v** are **nuisance parameters**

19

Bayesian inference

In **Bayesian statistics**, nuisance parameters are dealt with by assigning them a prior $\pi(v)$.

Usually a multinormal law is used with mean v_0 and covariance matrix estimated from Δv_0 (+correlation, if needed)

$$f(\theta, v \mid x) = \frac{f(x \mid \theta, v)\pi(\theta)\pi(v)}{\iint f(x \mid \theta, v)\pi(\theta)\pi(v)d\theta\, dv}$$

The final prior is obtained by **marginalization** over the nuisance parameters

$$f(\theta \mid x) = \int f(\theta, v \mid x)dv = \frac{\int f(x \mid \theta, v)\pi(\theta)\pi(v)dv}{\iint f(x \mid \theta, v)\pi(\theta)\pi(v)d\theta\, dv}$$

20

Profile Likelihood

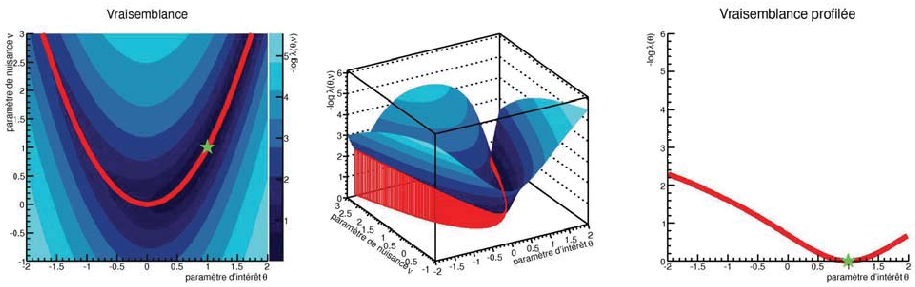
No true frequentist way to add systematic effects. Popular method of the day : **profiling**

Deal with nuisance parameters as realization if random variables : extend the likelihood : $\mathcal{L}(\theta, v) \rightarrow \mathcal{L}'(\theta, v)\mathcal{G}(v)$

$\mathcal{G}(v)$ is the **likelihood of the new parameters** (identical to prior)

For each value of θ, maximize the likelihood with respect to nuisance : **profile likelihood** $\mathcal{PL}(\theta)$.

$\mathcal{PL}(\theta)$ has the same statistical asymptotical properties than the regular likelihood



21

Non parametric estimation

Directly estimating the probability density function

- Likelihood ratio discriminant
- Separating power of variables
- Data/MC agreement
- ...

Frequency Table : For a sample $\{x_i\}$, $i=1..n$

1. Define successive intervals (bins) $C_k=[a_k, a_{k+1}[$
2. Count the number of events n_k in C_k

Histogram : Graphical representation of the frequency table

$$h(x) = n_k \text{ if } x \in C_k$$

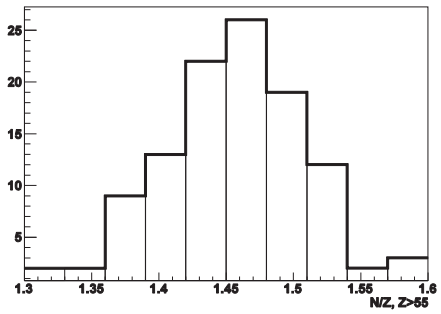
22

Histogram

Classe	Nombre de N/Z	Fréquence	Classe	Nombre de N/Z	Fréquence
< 1.30	0	0	1.45 - 1.48	26	0.2363
1.30 - 1.33	2	0.0182	1.48 - 1.51	19	0.1727
1.33 - 1.36	2	0.0182	1.51 - 1.54	12	0.1091
1.36 - 1.39	9	0.0818	1.54 - 1.57	2	0.0182
1.39 - 1.42	13	0.1182	1.57 - 1.60	3	0.0273
1.42 - 1.45	22	0.2	≥ 1.60	0	0

N/Z for stable heavy nuclei

```
1.321, 1.357, 1.392, 1.410, 1.428, 1.446,
1.464, 1.421, 1.438, 1.344, 1.379, 1.413,
1.448, 1.389, 1.366, 1.383, 1.400, 1.416,
1.433, 1.466, 1.500, 1.322, 1.370, 1.387,
1.403, 1.419, 1.451, 1.483, 1.396, 1.428,
1.375, 1.406, 1.421, 1.437, 1.453, 1.468,
1.500, 1.446, 1.363, 1.393, 1.424, 1.439,
1.454, 1.469, 1.484, 1.462, 1.382, 1.411,
1.441, 1.455, 1.470, 1.500, 1.449, 1.400,
1.428, 1.442, 1.457, 1.471, 1.485, 1.514,
1.464, 1.478, 1.416, 1.444, 1.458, 1.472,
1.486, 1.500, 1.465, 1.479, 1.432, 1.459,
1.472, 1.486, 1.513, 1.466, 1.493, 1.421,
1.447, 1.460, 1.473, 1.486, 1.500, 1.526,
1.480, 1.506, 1.435, 1.461, 1.487, 1.500,
1.512, 1.538, 1.493, 1.450, 1.475, 1.500,
1.512, 1.525, 1.550, 1.506, 1.530, 1.487,
1.512, 1.524, 1.536, 1.518, 1.577, 1.554,
1.586, 1.586
```



23

Histogram

Statistical description : n_k are **multinomial** random variables.

with parameters :

$$n = \sum_k n_k \qquad p_k = P(x \in C_k) = \int_{C_k} f_X(x)dx$$

$$\mu_{n_k} = np_k \qquad \sigma^2_{n_k} = np_k(1-p_k) \underset{p_k \ll 1}{\approx} \mu_{n_k} \qquad Cov(n_k, n_r) = -np_k p_r \underset{p_k \ll 1}{\approx} 0$$

For a large sample :

$$\lim_{n\to+\infty} \frac{n_k}{n} = \frac{\mu_k}{n} = p_k$$

For small classes (width δ):

$$p_k = \int_{C_k} f_X(x)dx \approx \delta f(x_c) \Rightarrow \lim_{\delta\to 0} \frac{p_k}{\delta} = f(x)$$

So finally :

$$\boxed{f(x) = \lim_{\substack{n\to+\infty \\ \delta\to 0}} \frac{1}{n\delta} h(x)}$$

The histogram is an estimator of the probability density

Each bin can be described by a **Poisson density**.

The 1σ error on n_k is then : $\Delta n_k = \sqrt{\hat{\sigma}^2_{n_k}} = \sqrt{\hat{\mu}_{n_k}} = \sqrt{n_k}$

24

Kernel density estimators

Histogram is a step function -> sometime need smoother estimator

On possible solution : **Kernel Density Estimator**

Attribute to each point of the sample a "kernel" function **k(u)**

$$u = \frac{x - x_i}{w}, \qquad k(u) = k(-u), \qquad \int k(u)du = 1$$

Triangle kernel : $k(u) = 1 - |u|$, for $-1 < u < 1$

Parabolic kernel : $k(u) = \frac{3}{4}(1 - u^2)$, for $-1 < u < 1$

Gaussian kernel : $k(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$

...

w = **kernel width**, similar to bin width of the histogram

The a pdf estimator is :

$$\boxed{K(x) = \frac{1}{n}\sum_i k(u_i) = \frac{1}{n}\sum_i k\left(\frac{x - x_i}{w}\right)}$$

Rem : for multidimensional pdf : $u^2 = \sum_k \left(\frac{x^{(k)} - x_i^{(k)}}{w^{(k)}}\right)^2$

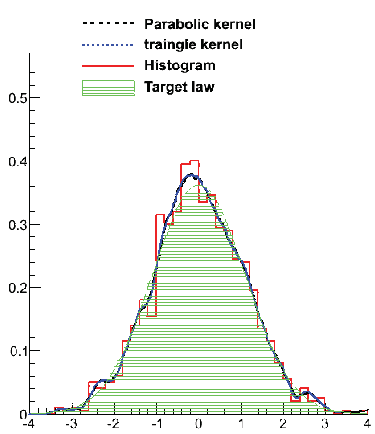
25

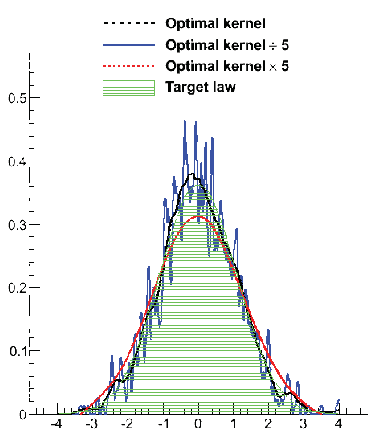
Kernel density estimators

If the estimated density is normal, the **optimal width** is :

$$w = \sigma\left(\frac{3}{(d+2)n}\right)^{\frac{1}{d+4}}$$ with **n** that sample size and **d** the dimension

As for the histogram binning, no generic result : try and see





26

Statistical Tests

Statistical tests aim at:

- Checking the **compatibility** of a dataset $\{x_i\}$ **with a given distribution**
- Checking the **compatibility of two datasets** $\{x_i\}$, $\{y_i\}$: are they issued from the same distribution.
- **Comparing different hypothesis** : background vs signal+background

In every case :

- build a statistic that quantify the agreement with the hypothesis
- **convert it into a probability** of compatibility/incompatibility : **p-value**

27

Pearson test

Test for **binned data** : use the Poisson limit of the histogram

- Sort the sample into **k** bins C_i : n_i
- Compute the probability of this class : $p_i = \int_{C_i} f(x)dx$
- The test statistics compare, for each bin **the deviation of the observation from the expected mean** to the **theoretical standard deviation.**

Data — Poisson mean

$$\chi^2 = \sum_{\text{bins } i} \frac{(n_i - np_i)^2}{np_i}$$

Poisson variance

Then χ^2 follow (asymptotically) a Khi-2 law with **k-1** degrees of freedom (1 constraint $\sum n_i = n$)

p-value : probability of doing worse, $\text{p-value} = \int_{\chi^2}^{+\infty} f_{\chi^2}(x;k-1)dx$

For a "good" agreement $\chi^2/(k-1) \sim 1$,

More precisely $\chi^2 \in (k-1) \pm \sqrt{2(k-1)}$ (1σ interval ~ 68%CL)

28

Kolmogorov-Smirnov test

Test for **unbinned data** : compare the sample cumulative density function to the tested one

Sample Pdf (ordered sample)

$$f_s(x) = \frac{1}{n}\sum_i \delta(x-i) \Rightarrow F_s(x) = \begin{cases} 0 & x < x_0 \\ \frac{k}{n} & x_k \le x < x_{k+1} \\ 1 & x > x_n \end{cases}$$

The the **Kolmogorov statistic** is the largest deviation :

$$D_n = \sup_x |F_S(x) - F(x)|$$

The test distribution has been computed by Kolmogorov:

$$P(D_n > \beta\sqrt{n}) = 2\sum_r (-1)^{r-1} e^{-2r^2z^2}$$

[0;β] define a confidence interval for D_n

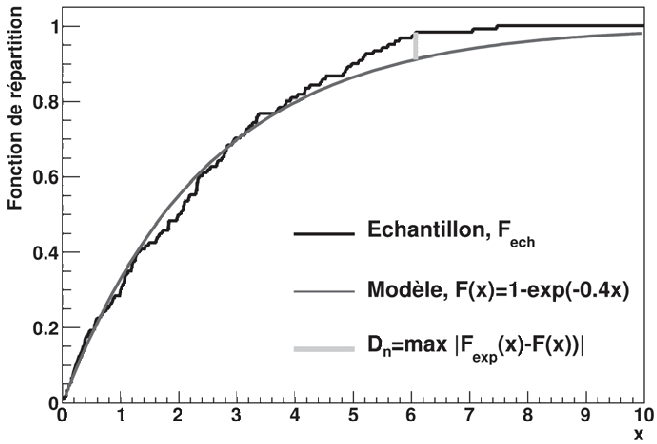
$\beta = 0.9584/\sqrt{n}$ for 68.3% CL $\beta = 1.3754/\sqrt{n}$ for 95.4% CL

29

Example

Test compatibility with an **exponential law** : $f(x) = \lambda e^{-\lambda x}$, $\lambda = 0.4$

0.008, 0.036, 0.112, 0.115, 0.133, 0.178, 0.189, 0.238, 0.274, 0.323, 0.364, 0.386, 0.406, 0.409, 0.418, 0.421, 0.423, 0.455, 0.459, 0.496, 0.519, 0.522, 0.534, 0.582, 0.606, 0.624, 0.649, 0.687, 0.689, 0.764, 0.768, 0.774, 0.825, 0.843, 0.921, 0.987, 0.992, 1.003, 1.004, 1.015, 1.034, 1.064, 1.112, 1.159, 1.163, 1.208, 1.253, 1.287, 1.317, 1.320, 1.333, 1.412, 1.421, 1.438, 1.574, 1.719, 1.769, 1.830, 1.853, 1.930, 2.041, 2.053, 2.119, 2.146, 2.167, 2.237, 2.243, 2.249, 2.318, 2.325, 2.349, 2.372, 2.465, 2.497, 2.553, 2.562, 2.616, 2.739, 2.851, 3.029, 3.327, 3.335, 3.390, 3.447, 3.473, 3.568, 3.627, 3.718, 3.720, 3.814, 3.854, 3.929, 4.038, 4.065, 4.089, 4.177, 4.357, 4.403, 4.514, 4.771, 4.809, 4.827, 5.086, 5.191, 5.928, 5.952, 5.968, 6.222, 6.556, 6.670, 7.673, 8.071, 8.165, 8.181, 8.383, 8.557, 8.606, 9.032, 10.482, 14.174



D_n = 0.069

p-value = 0.0617

1σ : [0, 0.0875]

30

Hypothesis testing

Two exclusive hypotheses H_0 and H_1

-which one is the most compatible with data

- how incompatible is the other one

P(data|H0) vs P(data|H1)

Build a **statistic**, define an interval **w**

- if the observation falls into **w** : accept H_1
- else accept H_0

Size of the test : how often did you get it right

$$\alpha = \int_w \mathcal{L}(x \mid H_0)dx$$

Power of the test : how often do you get it wrong !

$$1-\beta = \int_w \mathcal{L}(x \mid H_1)dx$$

Neyman-Pearson lemma : optimal statistic for testing hypothesis is the **Likelihood ratio**

$$\Lambda = \frac{\mathcal{L}(x \mid H_0)}{\mathcal{L}(x \mid H_1)} < k_\alpha$$

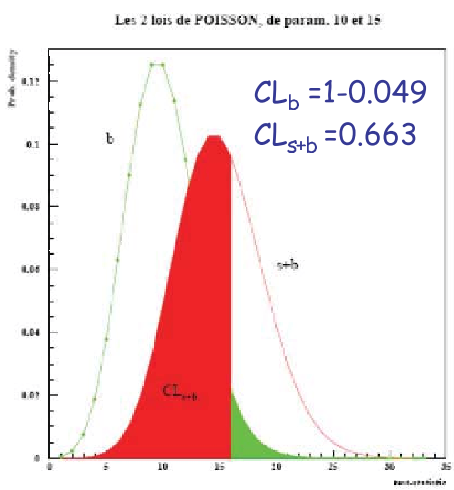
31

CL_b and CL_s

Two hypothesis, for counting experiment

- background only : expect 10 events
- signal+ background : expect 15 events

You observe 16 events



CL_b = confidence in the background hypothesis (power of the test)

Discovery : $1 - CL_b < 5.7\times10^{-7}$

CL_{s+b} = confidence in the signal+background hypothesis (size of the test)

Rejection : $CL_{s+b} < 5\times10^{-2}$

Test for signal (non standard)

$CL_s = CL_{s+b}/CL_b$

32