Probability and statistics: A reminder

Benoit Clément

1LPSC, 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

Probability and Statistics

Basic concepts I
(from a physicist point of view)

Benoît CLEMENT – Université J. Fourier / LPSC
bclement@lpsc.in2p3.fr

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
Sample and population

SAMPLE
Finite size
Selected through a random process
eg. Result of a measurement

POPULATION
Potentially infinite size
eg. All possible results

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)

Random process

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

It will be described by:

Universe: \( \Omega \) = 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 \( \mathcal{A} \) for which \( \mathcal{A} \) is true.
**Probability**

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

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

$$A \rightarrow P(A)$$

satisfying:

$$P(\Omega)=1$$

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

**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:

$$\lim_{n \to \infty} \frac{n_A}{n} = P(A)$$

defines a probability

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

---

**Simple logic**

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

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

$$P(\emptyset) = 1 - P(\Omega) = 0$$

Event « $A$ and $B$ » is associated with the ensemble $A \cap B$.

Event « $A$ or $B$ » is associated with the ensemble $A \cup B$.

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

We will use and mix notations: $\leftrightarrow A$, 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:

\[ \Omega = \bigcup_i B_i \quad B_i \cap B_j = \emptyset \quad (i \neq j) \]

Then, for any event \( A \):

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

Similar to a basis in linear algebra.

Conditional probability and independence

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

\[ P(A|B) = \langle \text{ probability of } A \text{ given } B \rangle \]

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:

\[ P(A \text{ and } B) = P(A) \cdot P(B) \]
Bayes theorem

The definition of conditional probability leads to:
\[ P(A \text{ and } B) = P(A | B) \cdot P(B) = P(B | A) \cdot P(A) \]

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

\[
P(B | A) = \frac{P(A | B) \cdot P(B)}{P(A)}
\]

Or, using a partition \( \{B_i\} \):
\[
P(B_i | A) = \frac{P(A | B_i) \cdot P(B_i)}{\sum_i P(A \text{ and } B_i)} = \frac{P(A | B_i) \cdot P(B_i)}{\sum_i P(A | B_i) \cdot P(B_i)}
\]

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

\[
P(\text{model}_1 | \text{data}) = \frac{P(\text{data} | \text{model}_1) \cdot P(\text{model}_1)}{\sum_i P(\text{data} | \text{model}_i) \cdot P(\text{model}_i)}
\]

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

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

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

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

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

\[
= \frac{\frac{5^{n_1-n_2-n_3-3}}{6^{n_1+n_2+n_3}} \times 0.7}{\frac{2^{n_1-n_2-n_3-3}}{3^{n_1+n_2+n_3}} \times 0.3 + \frac{5^{n_1-n_2-n_3-3}}{6^{n_1+n_2+n_3}} \times 0.7} = \frac{\frac{5^4}{6^7} \times 0.7}{\frac{2^4}{3^7} \times 0.3 + \frac{5^4}{6^7} \times 0.7} \approx 0.42
\]
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): $f(x) = \frac{dF}{dx}$

Density function

Probability density function

Cumulative density function

By construction:

$F(-\infty) = P(\emptyset) = 0$

$F(+\infty) = P(\Omega) = 1$

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

$$P(a < X < b) = F(b) - F(a) = \int_{a}^{b} f(x)dx$$
**Change of variable**

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

For \( \varphi \) bijective

- **\( \varphi \) increasing :** \( X < x \Leftrightarrow Y < y \)
  
  \[
  P(X < x) = F_X(x) = P(Y < y) = F_Y(y) = F_Y(\varphi(x)) \Rightarrow f_Y(y) = \frac{dF(x)}{dy} = \frac{f(x)}{\varphi'(x)}
  \]

- **\( \varphi \) 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 \( f_Y(y) = \frac{f(x)}{\varphi'(x)} \)

If \( \varphi \) not bijective : split into several bijective parts \( \varphi_i \)

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

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

---

**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:

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

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

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

\[
\Rightarrow f_X(x) = \int f(x, y)dy
\]
Multidimensional PDF (2)

For a fixed value of $Y=y_0$:

$$f(x|y_0)dx = \text{Probability of } x < X < x + dx \text{ knowing that } Y=y_0 \text{ is a conditional density for } X.$$ It is proportional to $f(x,y)$, so

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

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

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

$$f(x|y) = f_X(x) \text{ and } f(y|x) = f_Y(y) \quad \text{hence } f(x,y) = f_X(x).f_Y(y)$$

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.

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 equiprobability.

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.
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)

---

Measure of location

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

\[ \mu = \bar{x} = \int_{-\infty}^{+\infty} xf(x) dx \]

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

Median: Value that split the distribution in 2 equiprobable parts

\[ \int_{-\infty}^{\text{med}(x)} f(x) dx = \int_{\text{med}(x)}^{+\infty} f(x) dx \]

\[ \begin{align*}
\text{med}(x) &= \frac{x_1 + x_2 + \ldots + x_n}{n} \\
\text{odd } n &= \frac{x_{(n+1)/2}}{2} \\
\text{even } n &= \frac{x_{n/2} + x_{(n/2)+1}}{2}
\end{align*} \]

Mode: The most probable value = maximum of pdf

\[ \frac{df}{dx} \bigg|_{x=\text{mod}(x)} = 0, \quad \frac{d^2f}{dx^2} \bigg|_{x=\text{mod}(x)} < 0 \]
Measure of dispersion

Standard deviation ($\sigma$) and variance ($\nu = \sigma^2$) : Mean value of the squared deviation to the mean :
$$\nu = \sigma^2 = \int (x - \mu)^2 f(x) dx$$
$$\nu = \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

Koenig's theorem :
$$\sigma^2 = \int x^2 f(x) dx + \mu^2 \int f(x) dx - 2\mu \int xf(x) dx$$
$$\sigma^2 = \bar{x}^2 - \mu^2 = \bar{x}^2 - \bar{x}^2$$

Interquartile difference : generalize the median by splitting the distribution in 4 :
$$\int_{q_1}^{q_2} f(x) dx = \int_{q_2}^{q_3} f(x) dx = \int_{q_3}^{q_4} f(x) dx = \int_{q_4}^{+\infty} f(x) dx = \frac{1}{4} \quad \text{med}(x) = q_2$$
$$\delta = q_3 - q_1$$

Others...

Bienaymé-Chebyshev

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

Then for $x \in \Delta$ :
$$\left(\frac{x - \mu}{a}\right)^2 > 1 \Rightarrow \left(\frac{x - \mu}{a}\right)^2 f(x) > f(x)$$
$$\Rightarrow \int_{\Delta} \left(\frac{x - \mu}{a}\right)^2 f(x) dx > \int_{\Delta} f(x) dx$$
$$\Rightarrow \int_{-\infty}^{+\infty} \left(\frac{x - \mu}{a}\right)^2 f(x) dx > \int_{\Delta} f(x) dx$$
$$\Rightarrow \frac{\sigma^2}{a^2} > P(|X - \mu| > a)$$

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

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

<table>
<thead>
<tr>
<th>$a$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chebyshev's bound</td>
<td>0</td>
<td>0.75</td>
<td>0.889</td>
<td>0.938</td>
<td>0.96</td>
</tr>
<tr>
<td>Normal distribution</td>
<td>0.683</td>
<td>0.954</td>
<td>0.997</td>
<td>0.99996</td>
<td>0.9999994</td>
</tr>
</tbody>
</table>

01001-p.11
**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 \(Y\)

\[
\text{Cov}(X, Y) = \int \int (x - \mu_X)(y - \mu_Y)f(x, y)dx\,dy = \rho \sigma_X \sigma_Y = \mu_{XY} - \mu_X \mu_Y
\]

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

Correlation: \(\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}\)

Independent \(\iff\) Uncorrelated: \(\rho = 0\)

**Decorrelation**

Covariance matrix for \(n\) variables \(X_i\):

\[
\Sigma_{ij} = \text{Cov}(X_i, X_j) \Rightarrow \Sigma = \begin{pmatrix}
\sigma_1^2 & \rho_{12} \sigma_1 \sigma_2 & \cdots & \rho_{1n} \sigma_1 \sigma_n \\
\rho_{21} \sigma_1 \sigma_2 & \sigma_2^2 & \cdots & \rho_{2n} \sigma_2 \sigma_n \\
\vdots & \vdots & \ddots & \vdots \\
\rho_{n1} \sigma_1 \sigma_n & \rho_{n2} \sigma_2 \sigma_n & \cdots & \sigma_n^2
\end{pmatrix}
\]

For uncorrelated variables, \(\Sigma\) is diagonal.

Matrix real and symmetric: can be diagonalized.

On can define \(n\) new uncorrelated variables \(Y_i\):

\[
\Sigma' = \begin{pmatrix}
\sigma_1^2 & 0 & \cdots & 0 \\
0 & \sigma_2^2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \sigma_n^2
\end{pmatrix}
\]

\[
= \mathbf{B}^{-1} \Sigma \mathbf{B}, \quad Y = BX
\]

\(\sigma_i^2\) are the eigenvalues of \(\Sigma\).

\(\mathbf{B}\) contains the orthonormal eigenvectors.

The \(Y_i\) are the principal components. Sorted for the larger to the smaller \(\sigma_i\) they allow dimensional reduction.
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:

\[
\begin{align*}
\int \int (y - ax - b)^2 f(x, y) dx dy &= \frac{1}{n} \sum_{i=1}^{n} (y_i - ax_i - b)^2 \\
\frac{\partial w}{\partial a} &= 0 = \int x(y - ax - b)f(x, y) dx dy \\
\frac{\partial w}{\partial b} &= 0 = \int (y - ax - b)f(x, y) dx dy \\
\end{align*}
\]

\[
\begin{align*}
a (\sigma_y^2 - \mu_y^2) + b \mu_x &= \rho \sigma_x \sigma_y + \mu_x \mu_y \\
a \mu_x + b &= \mu_y
\end{align*}
\]

\[
\begin{align*}
a &= \frac{\rho \sigma_y}{\sigma_x} \\
b &= \mu_y - \rho \frac{\sigma_y}{\sigma_x} \mu_x
\end{align*}
\]

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 \(\mu_k\) are the expectation of \(X^k\).

1st moment: \(\mu_0 = 1\) (pdf normalization)
2nd moment: \(\mu_1 = \mu\) (mean)

\(X' = X - \mu\) is a central variable
2nd central moment: \(\mu'_2 = \sigma^2\) (variance)

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

From Taylor expansion:

\[
\phi(t) = \sum_{k} \frac{(it)^k}{k!} f(x) dx = \sum_{k} \frac{(it)^k}{k!} \mu_k
\]

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:
- \( 3^{rd} \) reduced moment: \( \mu''_3 = \beta_1 = \gamma_1 \): skewness
- \( \gamma_1 = 0 \) for symmetric distribution. Then mean = median

Measure of shape:
- \( 4^{th} \) 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

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

\[
\mu''_n = \left( \frac{1}{\mu_2'} \right)^{n-2} \mu'_n
\]
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 $n \to +\infty, p \to 0, np = \Lambda$

- **Variable**: $K$
- **Parameters**: $\Lambda$
- **Law**: $P(K; \Lambda) = \frac{\Lambda^k}{k!} e^{-\Lambda}$
- **Mean**: $\Lambda$
- **Variance**: $\Lambda$

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

**Normal distribution (Gaussian)**: limit of many processes

- **Parameters**: $\mu, \sigma$
- **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**: $C = \sum_{k=1}^{n} \left( \frac{X_k - \mu_k}{\sigma_k} \right)^2$
- **Parameters**: $n$
- **Law**: $f(c; n) = c^{\frac{n}{2} - 1} e^{-\frac{c}{2}} 2^{\frac{n}{2}} \Gamma \left( \frac{n}{2} \right)$
- **Mean**: $n$
- **Variance**: $2n$
**Convergence**

Loi de Poisson

\[ P(k; \lambda) = \frac{e^{-\lambda} \lambda^k}{k!} \]

\[ \mu = \lambda \quad \sigma = \lambda \]

\( p \) petit, \( k \ll n \)

\[ np = \lambda \]

\( \lambda > 25 \)

Loi binomiale

\[ p(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k} \]

\[ \mu = np \quad \sigma = \sqrt{np(1-p)} \]

\( n > 50 \)

Loi normale

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

\[ \mu = \mu \quad \sigma = \sigma \]

\( n > 30 \)

**Multidimensional Pdfs**

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

**Parameters**: \( n, p_1, p_2, \ldots, p_s \)

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

**Mean**: \( \mu = np_i \)

**Variance**: \( \sigma_i^2 = np_i (1-p_i) \)

\( \text{Cov}(K_i, K_j) = -np_i p_j \)

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

**Multinormal distribution**:

**Parameters**: \( \hat{m}, \Sigma \)

**Law**: \[ f(\hat{x}; \hat{m}, \Sigma) = \frac{1}{(2\pi)^{s/2} |\Sigma|^{1/2}} e^{-\frac{1}{2} (\hat{x}-\hat{m})^T \Sigma^{-1} (\hat{x}-\hat{m})} \]

**if uncorrelated** \[ f(\hat{x}; \hat{m}, \Sigma) = \prod_{i=1}^{s} \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{1}{2 \sigma_i^2} (x_i - m_i)^2} \]

**Independent** \( \iff \) **Uncorrelated**
**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 $S$:**

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

The mean is an additive quantity.

**Variance of $S$:**

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

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

For uncorrelated variables, the variance is additive -> used for error combinations.

---

**Sum of random variables**

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

Using the characteristic function:

$$\varphi_S(t) = \int f_S(s)e^{ist} ds = \int f_X(x) e^{it \sum x_i} dx$$

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 ($\lambda_1$ and $\lambda_2$) -> Poisson, $\lambda = \lambda_1 + \lambda_2$

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

Weak law of large numbers
Sample of size \( n \) = realization of \( n \) independent variables, with the same distribution (mean \( \mu \), variance \( \sigma^2 \)).
The sample mean is a realization of \( M = \frac{S}{n} = \frac{1}{n} \sum X_i \)
Mean value of \( M : \mu_M = \mu \quad \text{Variance of } M : \sigma_M^2 = \frac{\sigma^2}{n} \)

From Bienaymé-Chebyshev: \( P(\lvert M - \mu \rvert > a) \xrightarrow{n \to \infty} 0 \) (\( \forall a \))

Central-Limit theorem
\( n \) independent random variables of mean \( \mu_i \) and variance \( \sigma_i^2 \)
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:
\[
f_C(c) \xrightarrow{n \to \infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{c^2}{2}}
\]

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:
\[
\lim_{n \to \infty} \varphi_C(t) = \lim_{n \to \infty} \left(1 - \frac{t^2}{2n}\right)^n = e^{-\frac{t^2}{2}} = FT^{-1}[\xi]
\]
This is a naive demonstration, because we assumed that the moments were defined.
For CLT, only mean and variance are required (much more complex)
Central limit theorem

Dispersion and uncertainty

Any measure (or combination of measure) is a realization of a random variable.
- Measured value: $\theta$
- True value: $\theta_0$

Uncertainty = quantifying the difference between $\theta$ and $\theta_0$:
$\rightarrow$ measure of dispersion

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

Usually one separate
- Statistical error: due to the measurement Pdf.
- Systematic errors or bias $\rightarrow$ fixed but unknown deviation (equipment, assumptions, ...)

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

Observation error: $\Delta_0$

Position error: $\Delta_p$

Scaling error: $\Delta_S$

Each $\delta_i$ is a realization of a random variable: mean 0 (negligible) and variance $\sigma_i^2$. For uncorrelated error sources:

\[
\begin{align*}
\Delta_0 &= a_0 \sigma_0 \\
\Delta_S &= a \sigma_S \\
\Delta_p &= a \sigma_p \\
\end{align*}
\]

\[
\Delta_{tot}^2 = (a \sigma_{tot})^2 = a^2 (\sigma_0^2 + \sigma_S^2 + \sigma_p^2) = \Delta_0^2 + \Delta_S^2 + \Delta_p^2
\]

**Choice of $a$?**

If many sources, from central-limit $\rightarrow$ normal distribution:

- $a=1$ gives (approximately) a 68% confidence interval
- $a=2$ gives 95% CL (and at least 75% from Bienaymé-Chebshev)

---

**Error propagation**

Measure: $x \pm \Delta x$

Compute: $f(x) \rightarrow \Delta f$?

Assuming small errors, using Taylor expansion:

\[
\begin{align*}
f(x + \Delta x) &= f(x) + \frac{df}{dx} \Delta x + \frac{1}{2} \frac{d^2f}{dx^2} \Delta x^2 + \frac{1}{6} \frac{d^3f}{dx^3} \Delta x^3 + \frac{1}{24} \frac{d^4f}{dx^4} \Delta x^4 \\
f(x - \Delta x) &= f(x) - \frac{df}{dx} \Delta x + \frac{1}{2} \frac{d^2f}{dx^2} \Delta x^2 - \frac{1}{6} \frac{d^3f}{dx^3} \Delta x^3 + \frac{1}{24} \frac{d^4f}{dx^4} \Delta x^4 \\
\Rightarrow \Delta f &= \frac{1}{2} |f(x + \Delta x) - f(x - \Delta x)| = \left| \frac{df}{dx} \Delta x + \frac{1}{6} \frac{d^3f}{dx^3} \Delta x^3 \right|
\end{align*}
\]
Error propagation

Measure: $x \pm \Delta x$, $y \pm \Delta y$ ...
Compute: $f(x, y, ...) \rightarrow \Delta f$ ?
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

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

- uncorrelated
- correlated
- anticorrelated

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

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

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

Basic concepts II
(from a physicist point of view)

Benoit CLEMENT - Université J. Fourier / LPSC
bclement@lpsc.in2p3.fr

Statistics

- **PARAMETRIC ESTIMATION**: give one value to each parameter
- **INTERVAL ESTIMATION**: derive an interval that probably contains the true value
- **NON-PARAMETRIC ESTIMATION**: estimate the full pdf of the population
**Parametric estimation**

From a finite sample \( \{x_i\} \) \( \rightarrow \) estimating a parameter \( \theta \)

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

Any statistic can be considered as an **estimator** of \( \theta \)

To be a good estimator it needs to satisfy:

- **Consistency**: limit of the estimator for an 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

---

**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 \( \Theta \)

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

**Bias**: Mean value of the estimator

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

- **unbiased estimator**
  \( b(\hat{\theta}) = 0 \)
- **asymptotically unbiased**
  \( b(\hat{\theta}) \xrightarrow{n \to \infty} 0 \)

**Consistency**: formally

\[
P(\left|\hat{\theta} - \theta\right| > \varepsilon) \xrightarrow{n \to \infty} 0, \quad \forall \varepsilon
\]

In practice, if asymptotically unbiased

\( \sigma_{\hat{\theta}} \xrightarrow{n \to \infty} 0 \)
**Empirical estimator**

Sample mean is a good estimator of the population mean

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

Sample variance as an estimator of the population variance:

\[ s^2 = \frac{1}{n} \sum (x_i - \hat{\mu})^2 = \left( \frac{1}{n} \sum (x_i - \mu)^2 \right) - (\mu - \hat{\mu})^2 \]

\[ E[s^2] = \left( \frac{1}{n} \sum \sigma_i^2 \right) - \sigma_{\hat{\mu}}^2 = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n} \sigma^2 \]

biased, asymptotically unbiased

unbiased variance estimator:

\[ \hat{\sigma}^2 = \frac{1}{n-1} \sum (x_i - \hat{\mu})^2 \]

variance of the estimator (convergence)

\[ \sigma_{\hat{\sigma}^2}^2 = \frac{\sigma^4}{n-1} \left( \frac{n-1}{n} \gamma_2 + 2 \right) \rightarrow \frac{2\sigma^4}{n} \]

---

**Errors on these estimator**

Uncertainty \(\Leftrightarrow\) Estimator standard deviation

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

Mean:

\[ \hat{\mu} = \frac{1}{n} \sum x_i, \quad \sigma_{\hat{\mu}}^2 = \frac{\sigma^2}{n} \Rightarrow \Delta \hat{\mu} = \sqrt{\frac{\sigma^2}{n}} \]

Variance:

\[ \hat{\sigma}^2 = \frac{1}{n-1} \sum (x_i - \hat{\mu})^2, \quad \sigma_{\hat{\sigma}^2}^2 \approx \frac{2\sigma^4}{n} \Rightarrow \Delta \hat{\sigma}^2 = \sqrt{\frac{2\sigma^2}{n}} \]

Central-Limit theorem \(\rightarrow\) empirical estimators of mean and variance are normally distributed, for large enough samples

\[ \hat{\mu} \pm \Delta \hat{\mu} ; \hat{\sigma}^2 \pm \Delta \hat{\sigma}^2 \text{ define } 68\% \text{ confidence intervals} \]
**Likelihood function**

**Generic function** \( k(x, \theta) \)
- \( x \): random variable(s)
- \( \theta \): 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 \( f(\theta|x) = \frac{\mathcal{L}(\theta)}{\int \mathcal{L}(\theta) d\theta} \)

For a sample: \( n \) independent realizations of the same variable \( X \)
\[
\mathcal{L}(\theta) = \prod_{i} k(x_i, \theta) = \prod_{i} f(x_i; \theta)
\]

**Estimator variance**

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

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

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

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

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

Finally, using Cauchy-Schwartz inequality
\[
\left(1 + \frac{\partial b}{\partial \theta}\right)^2 \leq \int (\hat{\theta} - b - \theta)^2 k \left( \frac{\partial \ln k}{\partial \theta} \right)^2 \, dx
\]

**Cramer-Rao bound**
Efficiency

For any unbiased estimator of $\theta$, the variance cannot exceed:

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

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

Maximum likelihood

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

The analytical form of the density is known

It depends on several unknown parameters $\theta$

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

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

An estimator of the parameters of $\theta$, are the ones that maximize of observing the observed result.

$\rightarrow$ Maximum of the likelihood function

$$\frac{\partial L}{\partial \theta}\bigg|_{\theta=\hat{\theta}} = 0$$

rem: system of equations for several parameters
rem: often minimize $-\ln L$ : simplify expressions
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

CR Bound:
\[
f(\hat{\theta}; \hat{\theta}, \Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}} e^{-\frac{1}{2}(\hat{\theta} - \theta)^T \Sigma^{-1}(\hat{\theta} - \theta)}
\]

Covariance matrix of \( \Sigma^{-1} = -E \left[ \frac{\partial \ln L}{\partial \theta_i} \frac{\partial \ln L}{\partial \theta_j} \right] \)

Goodness of fit = The value of \(-2\ln L(\hat{\theta})\) is Khi-2 distributed, with ndf = sample size - number of parameters

\[
p - value = \int_{-\infty}^{+\infty} f_{\chi^2}(x; \text{ndf})dx \quad \text{Probability of getting a worse agreement}
\]

Errors on MLE

Errors on parameter \( \theta \) from the covariance matrix

For one parameter, 68% interval
\[
\Delta \theta = \sigma_{\theta} = \sqrt{-\frac{1}{\partial^2 \ln L / \partial \theta^2}}
\]

Only one realization of the estimator \( \theta \) \( \Rightarrow \) empirical mean of 1 value...

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

Confidence contour are defined by the equation:
\[
\Delta \ln L = \beta(n_\theta, a) \quad \text{with} \quad a = \int_0^{2\beta} f_{\chi^2}(x; n_\theta)dx
\]

Values of \( \beta \) for different number of parameters \( n_\theta \) and confidence levels \( a \)

<table>
<thead>
<tr>
<th>( n_\theta \rightarrow )</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a \rightarrow )</td>
<td>(0, 5n_\theta^2)</td>
<td>1.15</td>
<td>1.76</td>
</tr>
<tr>
<td>68.3</td>
<td>0.5</td>
<td>1.15</td>
<td>1.76</td>
</tr>
<tr>
<td>95.4</td>
<td>2</td>
<td>3.09</td>
<td>4.01</td>
</tr>
<tr>
<td>99.7</td>
<td>4.5</td>
<td>5.92</td>
<td>7.08</td>
</tr>
</tbody>
</table>
**Least squares**

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

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

Naïve approach: use regression

\[
\frac{\partial w}{\partial \theta_i} = 0
\]

Reweight each term by the error

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

Maximum likelihood: assume each \(y_i\) is normally distributed with a mean equal to \(f(x_i, \theta)\) and a variance equal to \(\Delta y_i^2\)

Then the likelihood is:

\[
\mathcal{L}(\theta) = \prod_i \frac{1}{\sqrt{2\pi \Delta y_i}} e^{-\frac{1}{2} \left( \frac{y_i - f(x_i, \theta)}{\Delta y_i} \right)^2}
\]

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

Generic case with correlations:

\[
K^2(\tilde{\theta}) = \frac{1}{2} (\tilde{y} - \tilde{f}(x, \tilde{\theta}))^T \Sigma^{-1} (\tilde{y} - \tilde{f}(x, \tilde{\theta}))
\]

---

**Example: fitting a line**

For \(f(x) = ax + b\)

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

\[
\hat{a} = \frac{AE - DC}{BE - C^2}, \quad \hat{b} = \frac{DB - AC}{BE - C^2}
\]

\[
\Delta \hat{a} = \frac{1.52}{\sqrt{B}}, \quad \Delta \hat{b} = \frac{1.52}{\sqrt{E}}
\]
Example: fitting a line

2 dimensional error contours on a and b

Confidence interval

For a random variable, a **confidence interval** with confidence level $\alpha$, 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
- $\rightarrow$ 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 $\theta$
Confidence interval

Mean centered, symmetric interval:

\[ [\mu - a, \mu + a] \]

\[ \int_{\mu - a}^{\mu + a} f(x)\,dx = a \]

Mean centered, probability symmetric interval:

\[ [a, b] \]

\[ \int_{a}^{b} f(x)\,dx = \frac{a}{2} \]

Highest Probability Density (HDP):

\[ [a, b] \]

\[ \int_{a}^{b} f(x)\,dx = a \]

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

Confidence Belt

To build a frequentist interval for an estimator \( \hat{\theta} \) of \( \theta \):

1. Make pseudo-experiments for several values of \( \theta \) and compute the estimator \( \hat{\theta} \) for each (MC sampling of the estimator pdf)

2. For each \( \theta \), 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 \( \alpha \).

3. Inverse these functions. The interval \([\Omega^{-1}(\hat{\theta}), \Xi^{-1}(\hat{\theta})]\) satisfy:
   \[
   P(\Omega^{-1}(\hat{\theta}) < \theta < \Xi^{-1}(\hat{\theta})) = 1 - P(\Xi^{-1}(\hat{\theta}) < \theta) - P(\Omega^{-1}(\hat{\theta}) > \theta)
   = 1 - P(\hat{\theta} < \Xi(\theta)) - P(\hat{\theta} > \Omega(\theta)) = \alpha
   \]

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

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

The known parameters $v$ are nuisance parameters.

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 $\Delta v_0$ (+correlation, if needed)

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

The final prior is obtained by marginalization over the nuisance parameters

$$f(\theta | x) = \int f(\theta, v | x)dv = \frac{\int f(x | \theta, v)\pi(\theta)\pi(v)dv}{\iint f(x | \theta, v)\pi(\theta)\pi(v)d\theta \, dv}$$
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: \( \mathcal{L}(\theta, \nu) \rightarrow \mathcal{L}^*(\theta, \nu) \mathcal{G}(\nu) \)

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

For each value of \( \theta \), 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

---

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

\[ h(x) = n_k \text{ if } x \in C_k \]
**Histogram**

<table>
<thead>
<tr>
<th>Classe</th>
<th>Nombre de N/Z</th>
<th>Fréquence</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 1.30</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1.30 - 1.33</td>
<td>2</td>
<td>0.0182</td>
</tr>
<tr>
<td>1.33 - 1.36</td>
<td>2</td>
<td>0.0182</td>
</tr>
<tr>
<td>1.36 - 1.39</td>
<td>9</td>
<td>0.0818</td>
</tr>
<tr>
<td>1.39 - 1.42</td>
<td>13</td>
<td>0.1182</td>
</tr>
<tr>
<td>1.42 - 1.45</td>
<td>22</td>
<td>0.2</td>
</tr>
<tr>
<td>1.45 - 1.48</td>
<td>26</td>
<td>0.2363</td>
</tr>
<tr>
<td>1.48 - 1.51</td>
<td>19</td>
<td>0.1727</td>
</tr>
<tr>
<td>1.51 - 1.54</td>
<td>12</td>
<td>0.1091</td>
</tr>
<tr>
<td>1.54 - 1.57</td>
<td>2</td>
<td>0.0182</td>
</tr>
<tr>
<td>1.57 - 1.60</td>
<td>3</td>
<td>0.0273</td>
</tr>
<tr>
<td>&gt; 1.60</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**N/Z for stable heavy nuclei**

1.321, 1.357, 1.392, 1.410, 1.428, 1.446, 1.464, 1.481, 1.438, 1.444, 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, 1.435, 1.437, 1.437, 1.453, 1.468, 1.500, 1.446, 1.363, 1.393, 1.423, 1.493, 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, 1.476, 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, 1.486, 1.493, 1.421, 1.447, 1.460, 1.473, 1.486, 1.500, 1.526, 1.480, 1.506, 1.439, 1.461, 1.487, 1.500, 1.512, 1.530, 1.493, 1.450, 1.475, 1.500, 1.512, 1.523, 1.550, 1.506, 1.530, 1.487, 1.512, 1.524, 1.536, 1.518, 1.577, 1.554, 1.586, 1.586

---

**Histogram**

Statistical description: $n_k$ are multinomial random variables.

With parameters:

$$n = \sum_k n_k \quad p_k = P(x \in C_k) = \frac{\int f_x(x)dx}{C_k}$$

$$\mu_k = np_k \quad \sigma^2_k = np_k(1 - p_k) \approx \mu_k \quad \text{Cov}(n_k, n_r) = -np_k p_r \approx 0$$

For a large sample:

$$\lim_{n \to \infty} \frac{n_k}{n} = \frac{\mu_k}{n} = p_k \quad p_k = \frac{\int f_x(x)dx}{C_k} \approx \delta f(x) \Rightarrow \lim_{\delta \to 0} \frac{p_k}{\delta} = f(x)$$

So finally:

The histogram is an estimator of the probability density.

Each bin can be described by a Poisson density.

The $1\sigma$ error on $n_k$ is then:

$$\Delta n_k = \sqrt{\sigma^2_k} = \sqrt{\mu_k} = \sqrt{n_k}$$
**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 \( k(u) \)

\[
\frac{u = \frac{x - x_i}{w}}{k(u) = k(-u), \quad \int k(u) du = 1}
\]

**Triangle kernel**: \( k(u) = 1 - |u|, \text{ for } -1 < u < 1 \)

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

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

\[w = \text{kernel width}, \text{ similar to bin width of the histogram}\]

The a pdf estimator is:

\[
K(x) = \frac{1}{n} \sum_{i} k(u_i) = \frac{1}{n} \sum_{i} k\left(\frac{x - x_i}{w}\right)
\]

Rem: for multidimensional pdf: \( u^2 = \sum x^2 - \frac{x_i^2}{w^2} \)

---

**Kernel density estimators**

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

\[
w = \sigma \left( \frac{3}{(d+2)n} \right)^{1/d+4}
\]

with \( n \) that sample size and \( d \) the dimension

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

---

**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.

\[
\chi^2 = \sum_{i} \frac{(n_i - np_i)^2}{np_i}
\]

Then \( \chi^2 \) follow (asymptotically) a Khi-2 law with \( k-1 \) degrees of freedom (1 constraint \( \Sigma n_i = n \))

\( p\)-value: probability of doing worse,

\[
P - \text{value} = \int_{\chi^2}^{+\infty} f_{\chi^2}(x; k-1) dx
\]

For a “good” agreement \( \chi^2 / (k-1) \sim 1 \).

More precisely \( \chi^2 \in (k-1) \pm \sqrt{2(k-1)} \) (1\(\sigma\) interval \~ 68\%CL)
Kolmogorov-Smirnov test

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

Sample Pdf (ordered sample)
\[ f_s(x) = \frac{1}{n} \sum_i \delta(x - i) \Rightarrow F_s(x) = \begin{cases} 0 & x < x_0 \\ \frac{k}{n} & x_k \leq x < x_{k+1} \\ 1 & x > x_n \end{cases} \]

The Kolmogorov statistic is the largest deviation:
\[ D_n = \sup_x |F_s(x) - F(x)| \]

The test distribution has been computed by Kolmogorov:
\[ P(D_n > \beta \sqrt{n}) = 2 \sum_{r=0}^{n} (-1)^{r-1} e^{-2r^2} \]

[0; \beta] define a confidence interval for \( D_n \)
\[ \beta = 0.9584/\sqrt{n} \text{ for 68.3\% CL} \quad \beta = 1.3754/\sqrt{n} \text{ for 95.4\% CL} \]

Example

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

Fonction de répartition

\( D_n = 0.069 \)
\( p\text{-value} = 0.0617 \)
\( \sigma : [0, 0.0875] \)
Hypothesis testing

Two exclusive hypotheses $H_0$ and $H_1$
- which one is the most compatible with data
- how incompatible is the other one

$P(\text{data}|H_0) \quad \text{vs} \quad P(\text{data}|H_1)$

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

$$a = \int_{w} L(x|H_0) dx$$

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

$$1 - \beta = \int_{w} L(x|H_1) dx$$

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

$$\Lambda = \frac{L(x|H_0)}{L(x|H_1)} < k_a$$

CL$_b$ and CL$_s$

Two hypothesis, for counting experiment
- background only: expect 10 events
- signal + background: expect 15 events

You observe 16 events

\[ \text{CL}_b = 1 - 0.049 \]
\[ \text{CL}_{s+b} = 0.663 \]

CL$_b$ = confidence in the background hypothesis (power of the test)
Discovery: \[ 1 - \text{CL}_b < 5.7 \times 10^{-7} \]

CL$_{s+b}$ = confidence in the signal + background hypothesis (size of the test)
Rejection: \[ \text{CL}_{s+b} < 5 \times 10^{-2} \]

Test for signal (non standard)

\[ CL_s = \frac{\text{CL}_{s+b}}{\text{CL}_b} \]