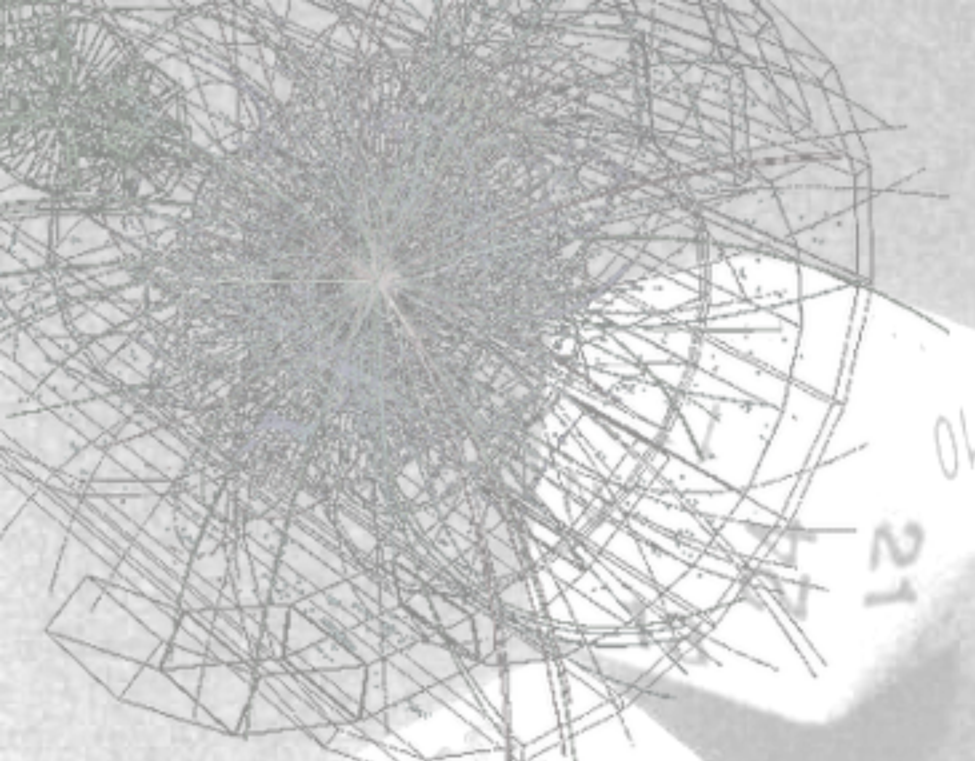
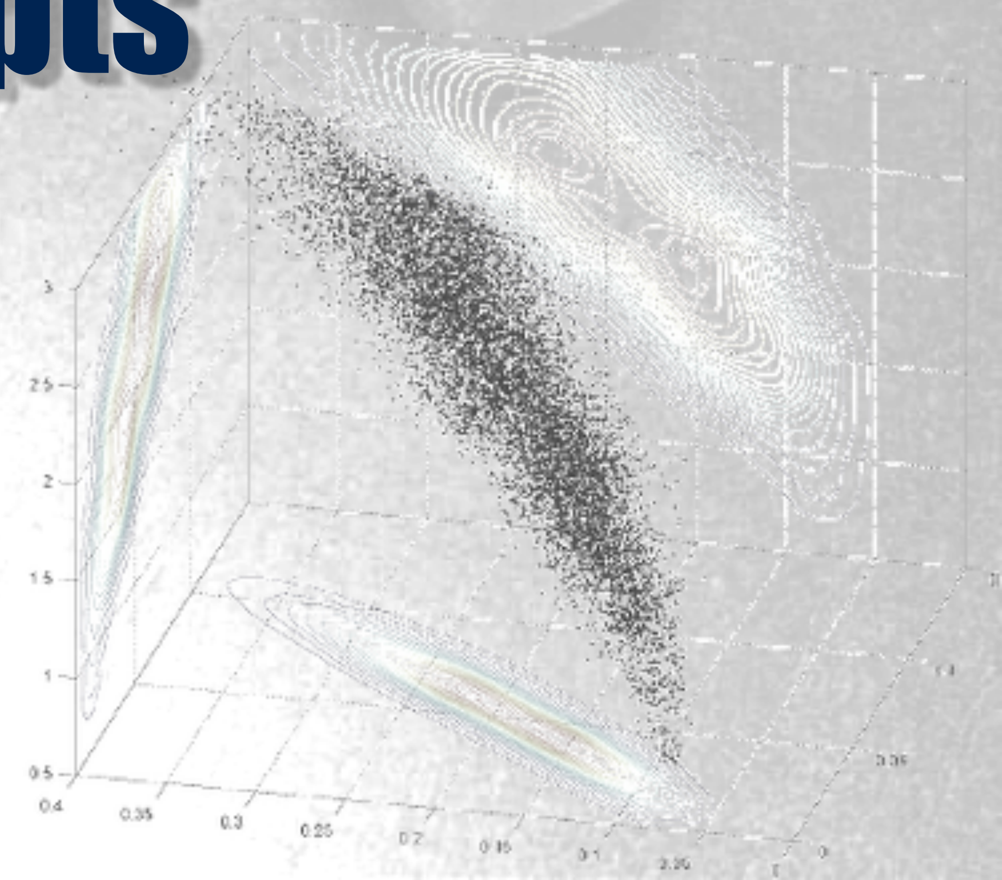


Probability and Statistics

Basic concepts

Florian RUPPIN
Université Grenoble Alpes / LPSC
ruppin@lpsc.in2p3.fr

Course content: Benoit Clément



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. Clément, Dunod

Contents

First lecture: Probability theory - Sample and population

SAMPLE

- **Finite size**
- **Selected through a random process**

eg. Result of a measurement

POPULATION

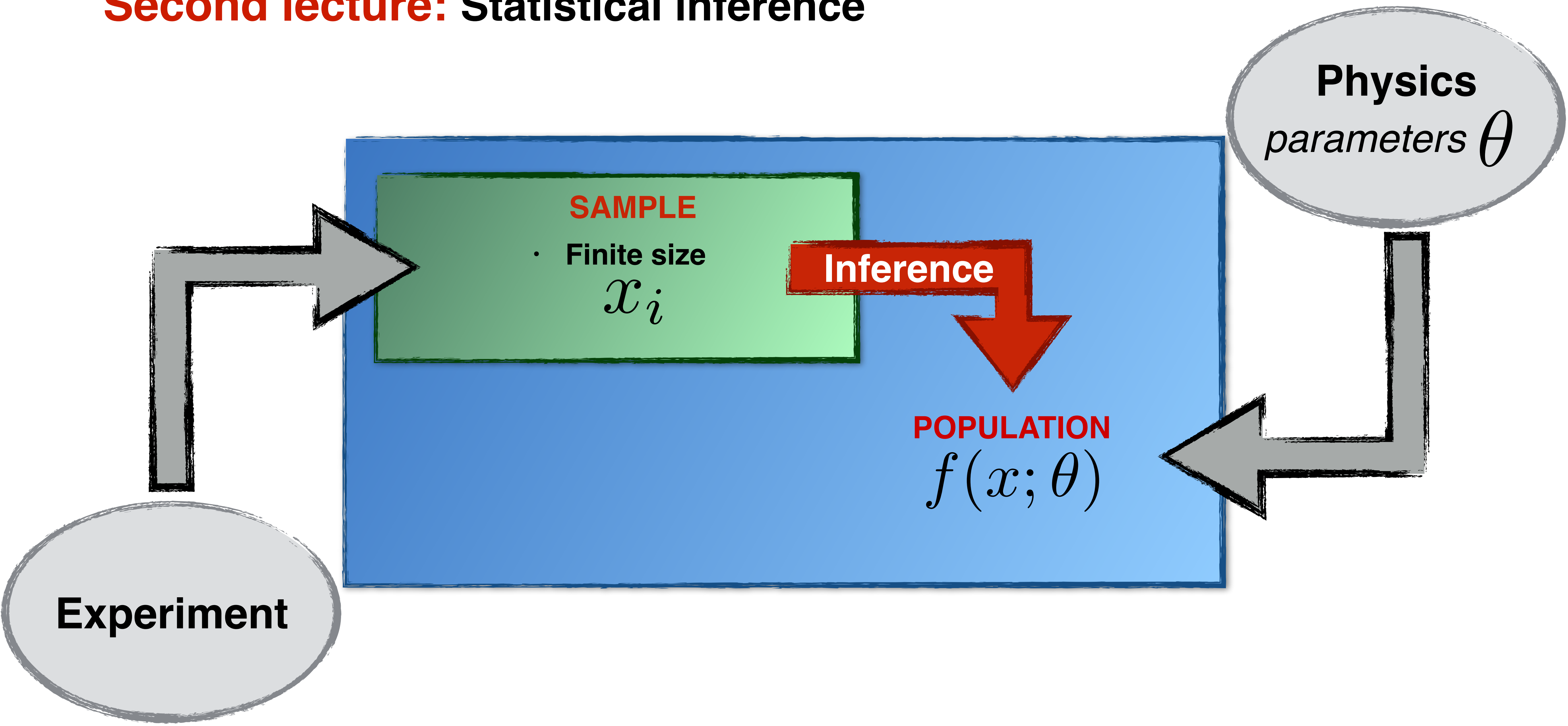
- **Potentially infinite size**

eg. All possible results

→ Characterization of the sample, the population and the sampling process

Contents

Second lecture: Statistical inference



→ Using the sample to estimate the characteristics of the population

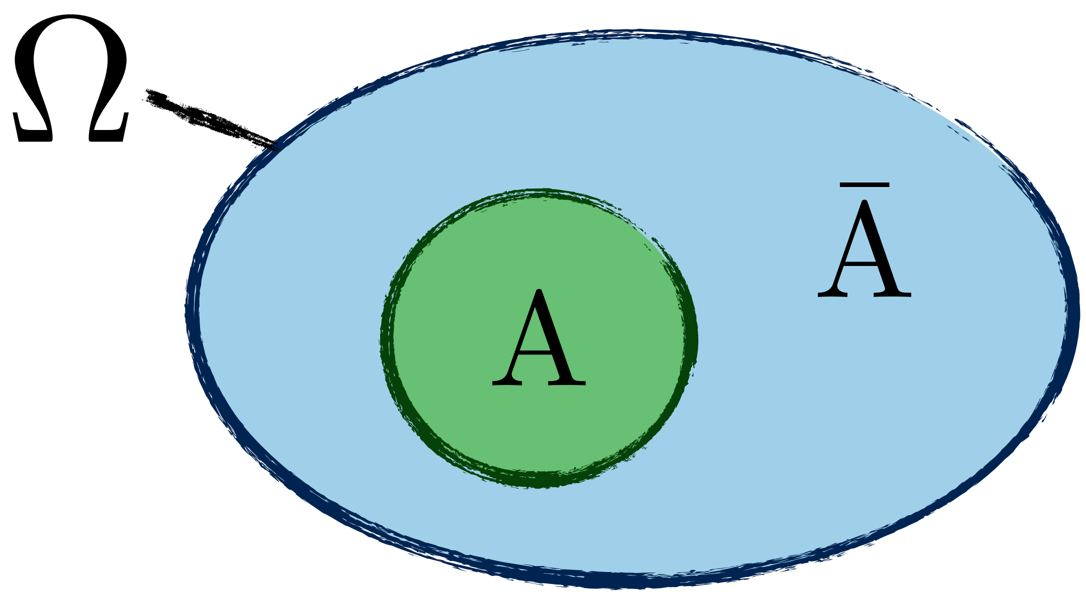
Random process

- **Random process** (“measurement” or “experiment”):
  Process whose outcome cannot be predicted with certainty.

- Described by:

  Universe: Ω = Set of all possible outcomes

  Event: Logical condition on an outcome
  Either 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**.

Probability

- **Probability function** P defined by: *(Kolmogorov, 1933)*

$$\mathrm{P} : \{\mathrm{Events}\} \longrightarrow [0:1]$$
$$\mathrm{A} \longrightarrow \mathrm{P(A)}$$

- **Properties:**

$$\mathrm{P}(\Omega) = 1$$
$$\mathrm{P(A\ or\ B)} = \mathrm{P(A)} + \mathrm{P(B)} \quad \mathrm{if\ (A\ and\ B)} = \emptyset$$

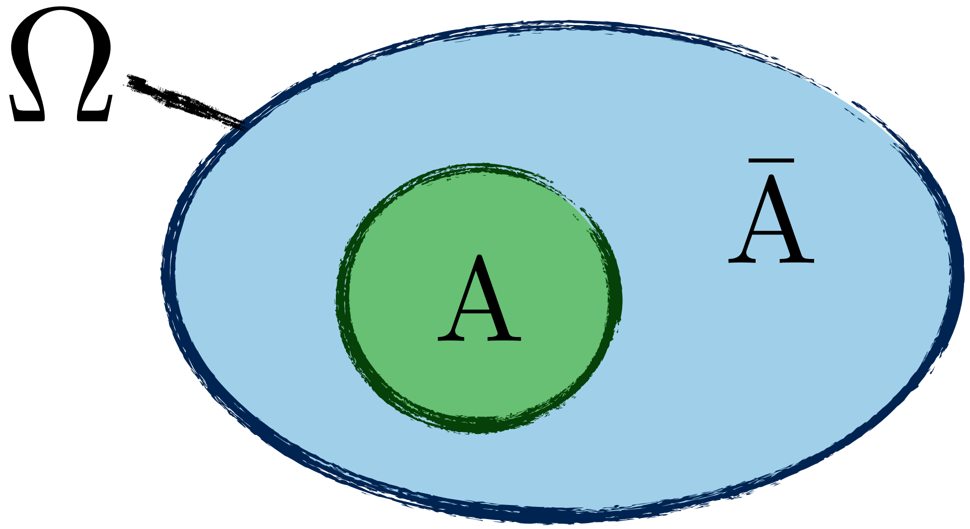
- **Interpretation:**

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

$$\lim_{n \to +\infty} \frac{n_\mathrm{A}}{n} = \mathrm{P(A)}$$ defines a probability

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

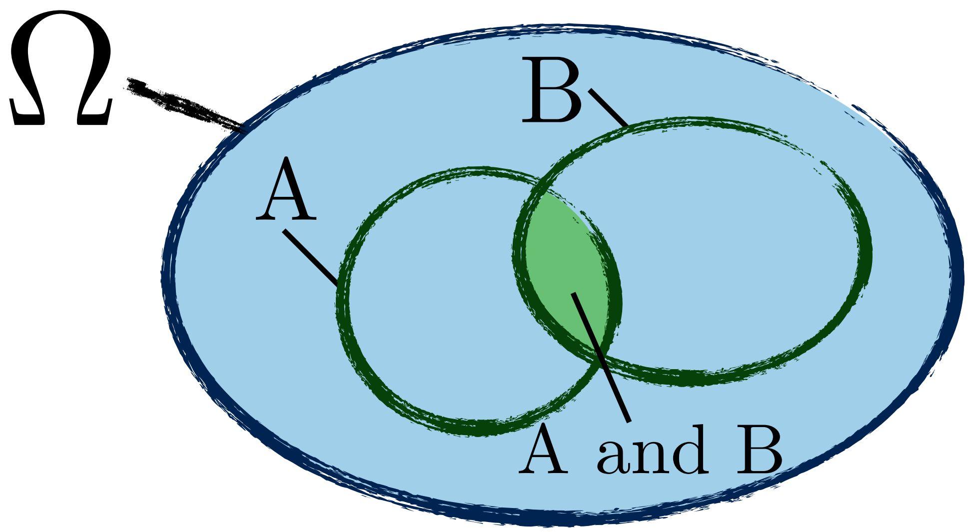
Logical relations



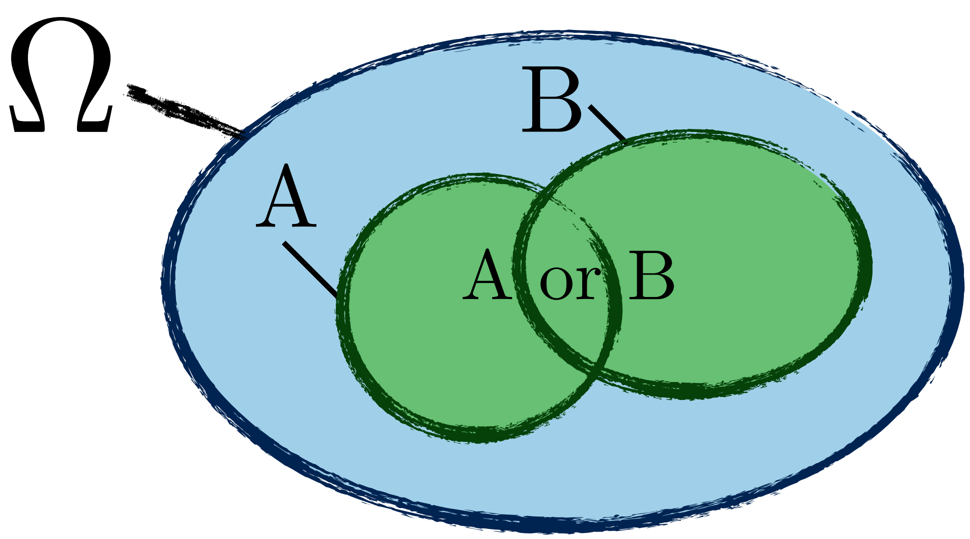
- **Event "not A"** associated with the complement of A:

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

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



- **Event "A and B"** associated with the intersection of the subsets



- **Event "A or B"** associated with the union of the subsets

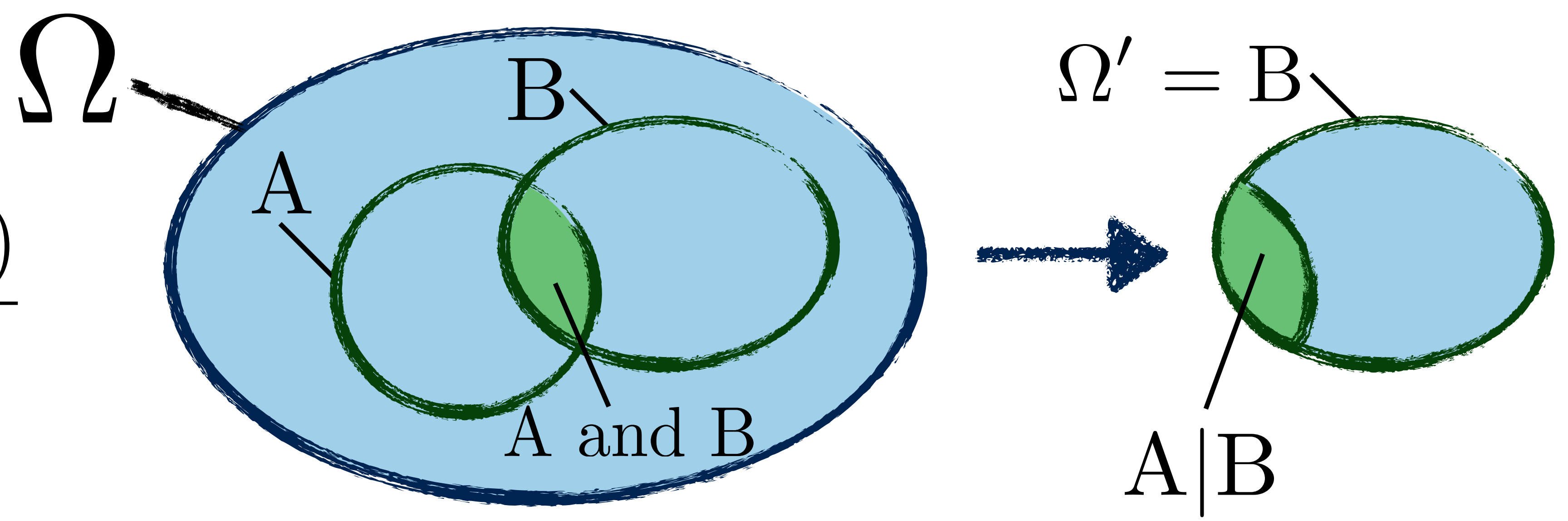
$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Conditional probability

- Event B known to be true $\longrightarrow$ restriction of the universe to $\Omega' = \mathrm{B}$
  Definition of a new probability function on this universe, the **conditional probability:**

$$\mathrm{P(A|B) = "probability\ of\ A\ given\ B"}$$

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



- The definition of the conditional probability leads to:

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

$\longrightarrow$ Relation between $\mathrm{P(A|B)}$ and $\mathrm{P(B|A)}$, the **Bayes theorem:**

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

Major role in Bayesian inference

Incompatibility and Independance

- Two **incompatible** events cannot be true simultaneously: $\mathrm{P(A\ and\ B) = 0}$

$$\longrightarrow \quad \mathrm{P(A\ or\ B) = P(A) + P(B)}$$

- Two events are **independent**, if the realization of one is not linked in any way to the realization of the other: $\mathrm{P(A|B) = P(A)}$ and $\mathrm{P(B|A) = P(B)}$

$$\longrightarrow \quad \mathrm{P(A\ and\ B) = P(A).P(B)}$$

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: $\mathrm{p}(x) = \mathrm{P}(X = x)$

- For a real variable: $\mathrm{P}(X = x) = 0$

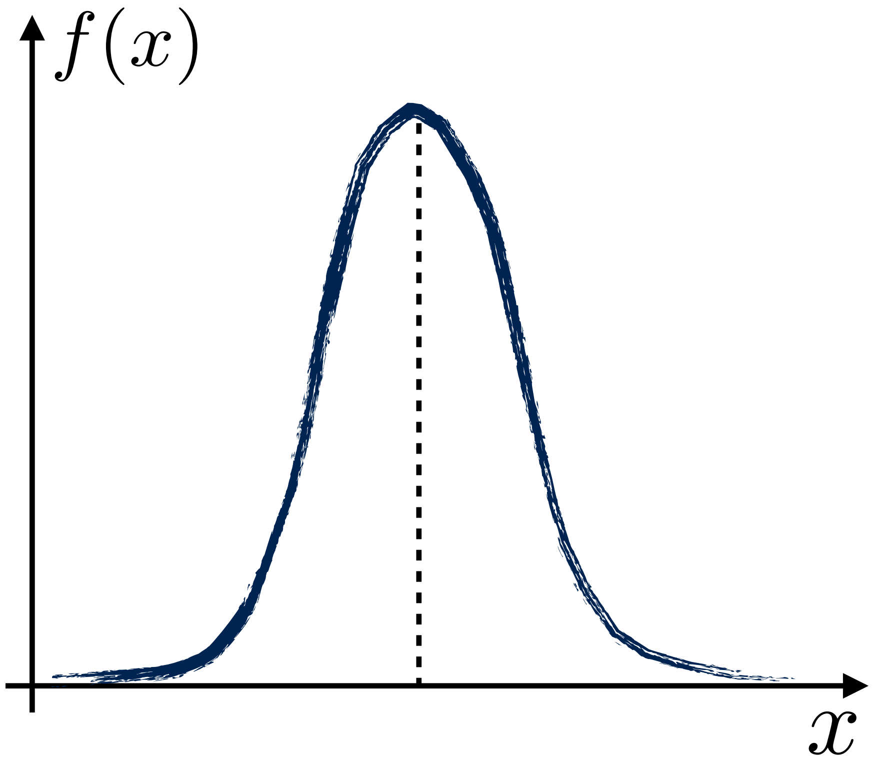
  Cumulative density function: $\mathrm{F}(x) = \mathrm{P}(X < x)$

$$
\begin{aligned}
d\mathrm{F} = \mathrm{F}(x+dx) - \mathrm{F}(x) &= \mathrm{P}(X < x+dx) - \mathrm{P}(X < x) \\
&= \mathrm{P}(X < x \text{ or } x < X < x+dx) - \mathrm{P}(X < x) \\
&= \mathrm{P}(X < x) + \mathrm{P}(x < X < x+dx) - \mathrm{P}(X < x) \\
&= \mathrm{P}(x < X < x+dx) = f(x)dx
\end{aligned}
$$

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

Density function

Probability density function:



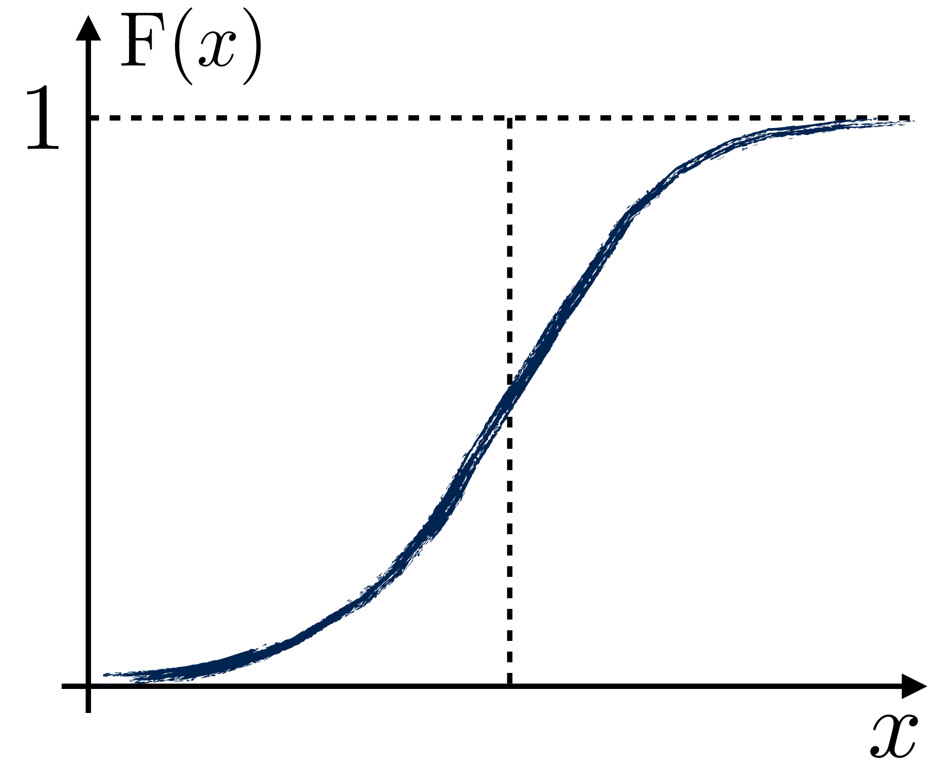
$$\int_{-\infty}^{+\infty} f(x)dx = \mathrm{P}(\Omega) = 1$$

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

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

with $\sum_i \mathrm{p}(i) = 1$

Cumulative density function:



By construction:

$$\mathrm{F}(-\infty) = \mathrm{P}(\emptyset) = 0$$

$$\mathrm{F}(+\infty) = \mathrm{P}(\Omega) = 1$$

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

$$\mathrm{P}(a < X < b) = \mathrm{F}(b) - \mathrm{F}(a) = \int_a^b f(x)dx$$

Moments

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

$$\mathrm{E}[g(X)] = \int g(x)f(x)dx \longrightarrow \text{Mean value of } g$$

- **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 called a **central variable**

2nd central moment: $\mu_2' = \sigma^2$ (variance)

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

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

$$\boxed{\mu_k = -i^k \frac{d^k\phi}{dt^k}\Big|_{t=0}}$$

Pdf entirely defined by its moments

Characteristic function: usefull tool for demonstrations

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 then associate to this process a probability density: the **empirical density** or **sample density**

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

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

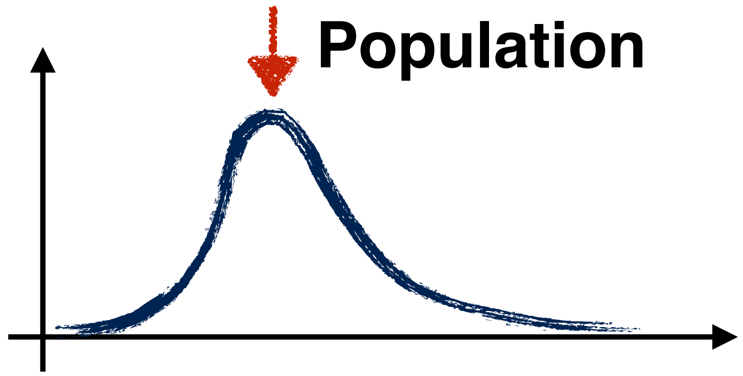
- **Measure of dispersion:**

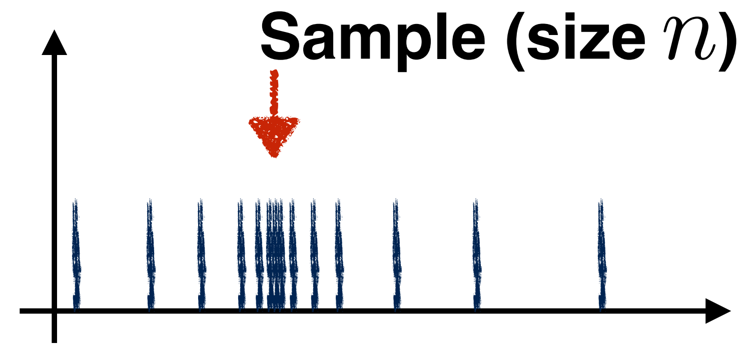
  Spread of the distribution around the central value

  → Uncertainty / Error

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

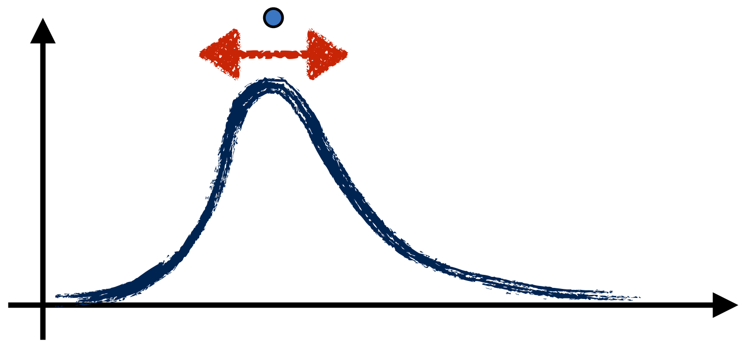
Location and dispersion

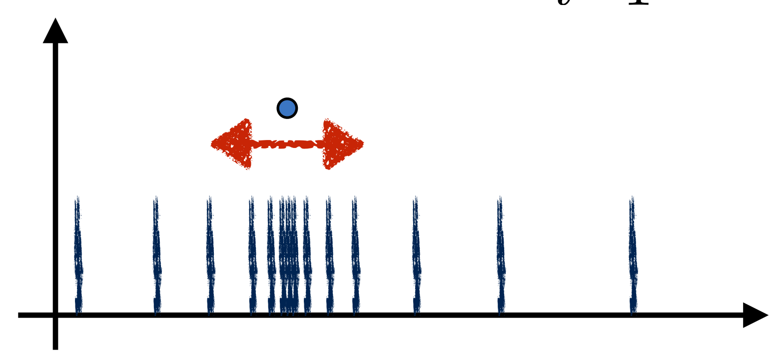




Mean value: Sum (integral) of all possible values weighted by the probability of occurrence

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





Standard deviation (σ) and variance ($v = \sigma^2$): Mean value of the squared deviation to the mean

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

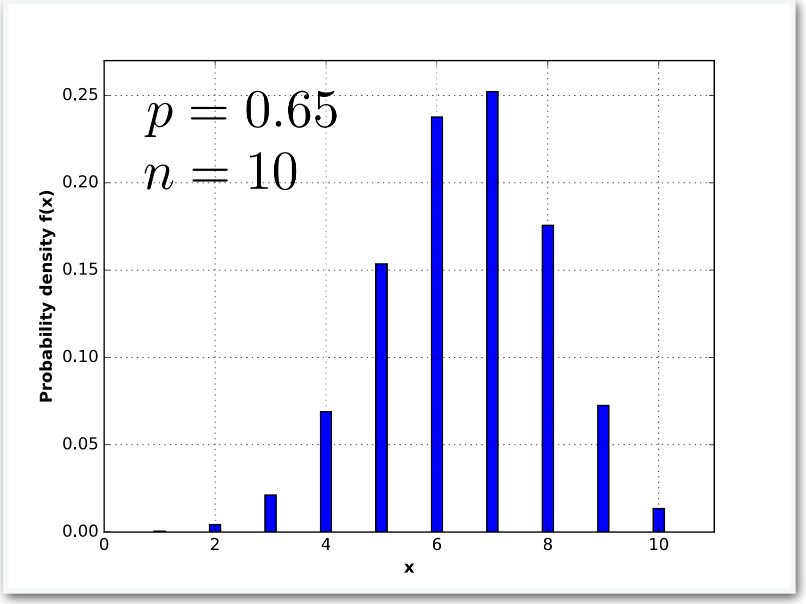
Koenig's theorem:

$$\sigma^2 = \int x^2 f(x) dx + \mu^2 \int f(x) dx - 2\mu \int x f(x) dx = \boxed{\overline{x^2} - \mu^2 = \overline{x^2} - \bar{x}^2}$$

Discrete distributions

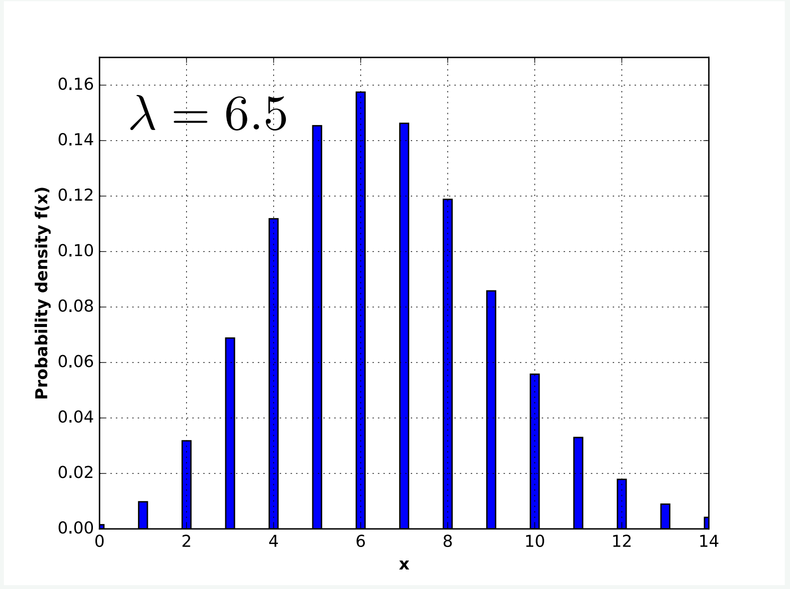
- **Binomial distribution:** randomly choosing K objects within a finite set of n, with a fixed drawing probability of p

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 $n \longrightarrow +\infty$, $p \longrightarrow 0$, $np = \lambda$
  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



Real distributions

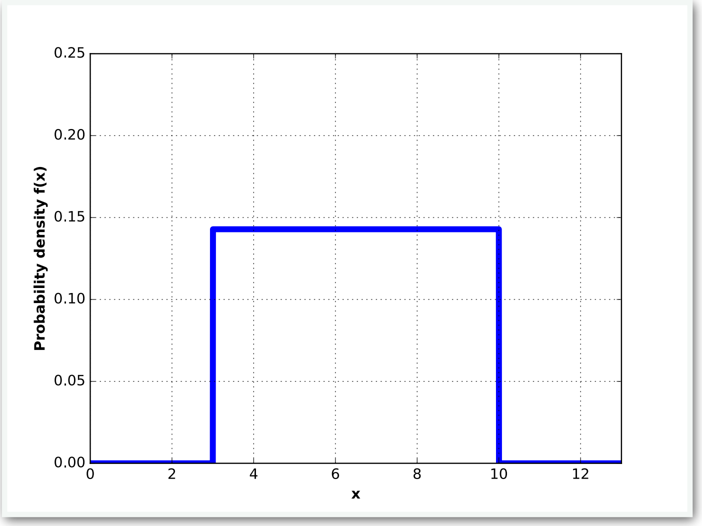
- **Uniform distribution:** equiprobability over a finite range $[a, b]$

Parameters : a, b

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

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

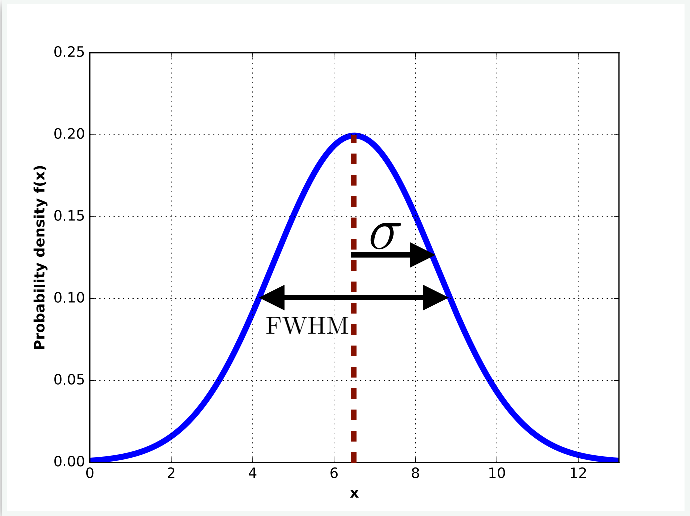
Variance : $v = \sigma^2 = (b-a)^2/12$



- **Normal distribution:** limit of many processes

Parameters : μ, σ

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



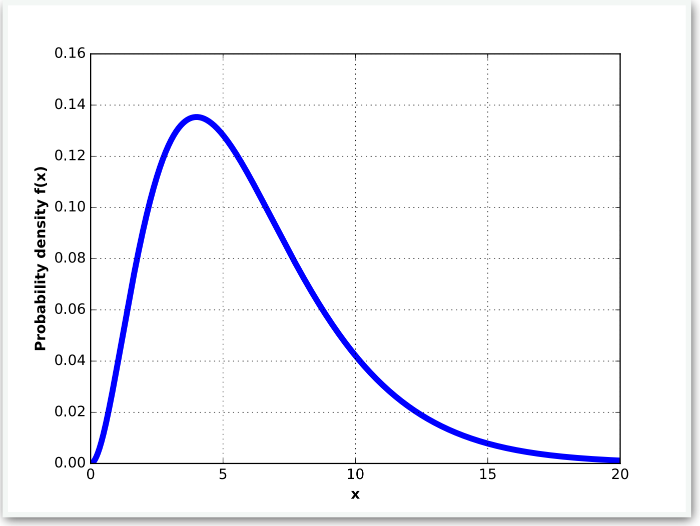
- **Chi-square distribution:** sum of the square of n normal reduced variables

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

Parameters : n

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

Mean : n Variance: $2n$



Convergence

Binomial distribution

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

$$\mu = np \qquad \sigma = \sqrt{np(1-p)}$$

p small, $k \ll n$

$np = \lambda$

Poisson distribution

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

$$\mu = \lambda \qquad \sigma = \sqrt{\lambda}$$

$\lambda \gtrsim 25$

$n \gtrsim 50$

Normal distribution

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

$$\mu = \mu \qquad \sigma = \sigma$$

Chi-square distribution

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

$$\mu = n \qquad \sigma = \sqrt{2n}$$

$n \gtrsim 30$

Multidimensional PDF

- Random variables can be generalized to random vectors:

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

- The **probability density function** becomes:

$$\begin{aligned} f(\vec{x})d\vec{x} &= f(x_1, x_2, ..., x_n)dx_1 dx_2 ... dx_n \\ &= P(x_1 < X_1 < x_1 + dx_1 \text{ and } x_2 < X_2 < x_2 + dx_2 ... \\ &\qquad ... \text{and } x_n < X_n < x_n + dx_n) \end{aligned}$$

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

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

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

Multidimensional PDF

- For a fixed value of $Y = y_0$:

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

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

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

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

- 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 $f(x,y) = f_X(x).f_Y(y)$

- For probability density functions, Bayes' theorem becomes:

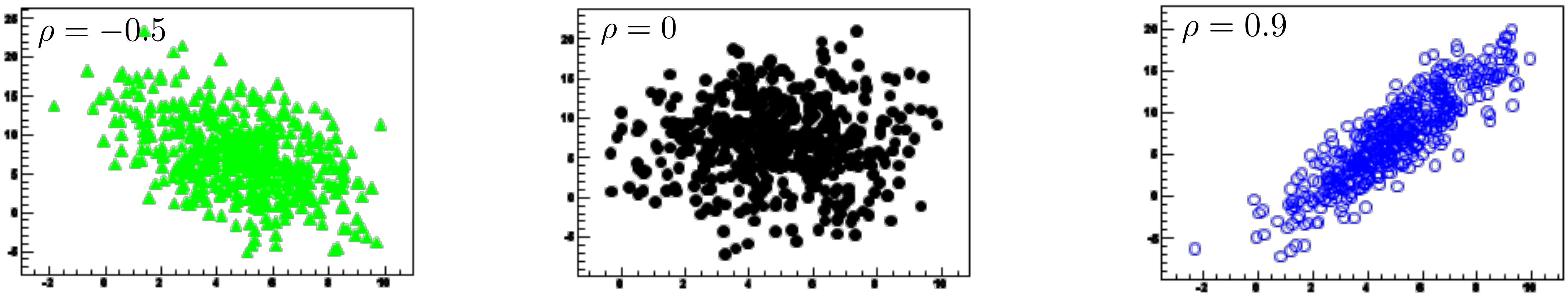
$$\boxed{f(y|x) = \frac{f(x|y)f_Y(y)}{f_X(x)} = \frac{f(x|y)f_Y(y)}{\int f(x|y)f_Y(y)dy}}$$

Covariance and correlation

- A random vector (X, Y) can be treated as 2 **separate variables** marginal densities

→ mean and standard deviation for each variable: $\mu_X, \mu_Y, \sigma_X, \sigma_Y$

- These quantities do not take into account **correlations** between the variables:



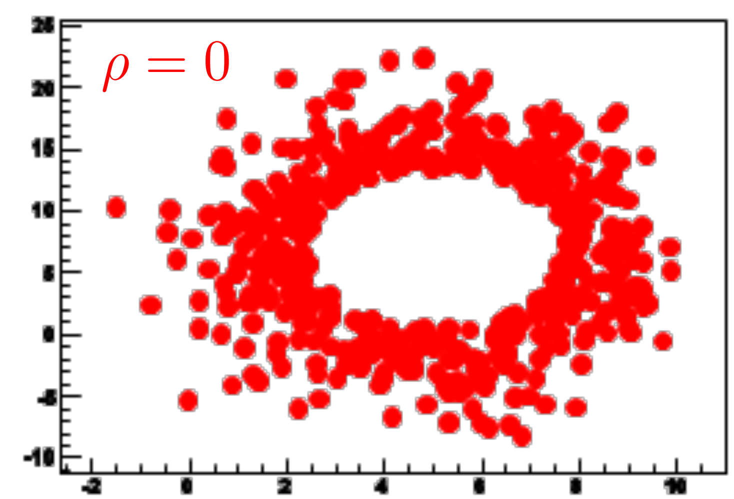
- Generalized measure of dispersion: **Covariance of X and Y**

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

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

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

Independent ⇄ **Uncorrelated**



Decorrelation

- Covariance matrix for n variables X_i :

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

- For **uncorrelated variables** $\sum$ is **diagonal**

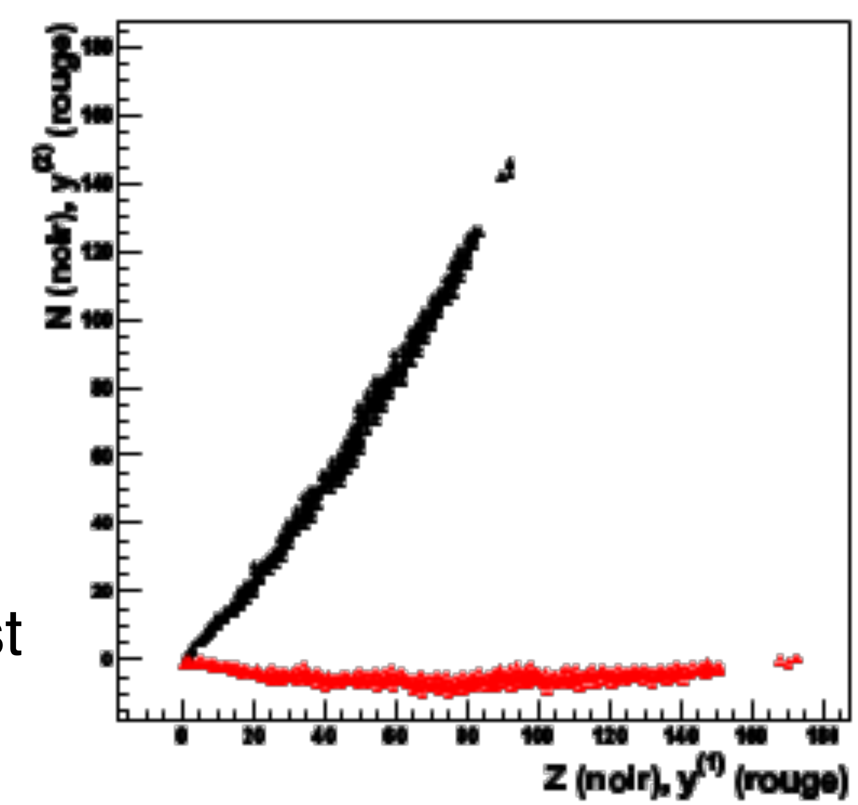
- Matrix **real** and **symmetric:** $\sum$ can be diagonalized

$\longrightarrow$ Definition of n new uncorrelated variables Y_i

$$\Sigma' = \begin{bmatrix} \sigma_1'^2 & 0 & \cdots & 0 \\ 0 & \sigma_2'^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n'^2 \end{bmatrix} = \mathrm{B}^{-1}\Sigma\mathrm{B} \quad \text{with} \quad \mathrm{Y} = \mathrm{BX}$$

$\sigma_i'^2$ are the **eigenvalues** of $\sum$

B contains the orthonormal eigenvectors



- The Y_i are the principal components. Sorted from the largest to the smallest σ', they allow dimensional reduction

Regression

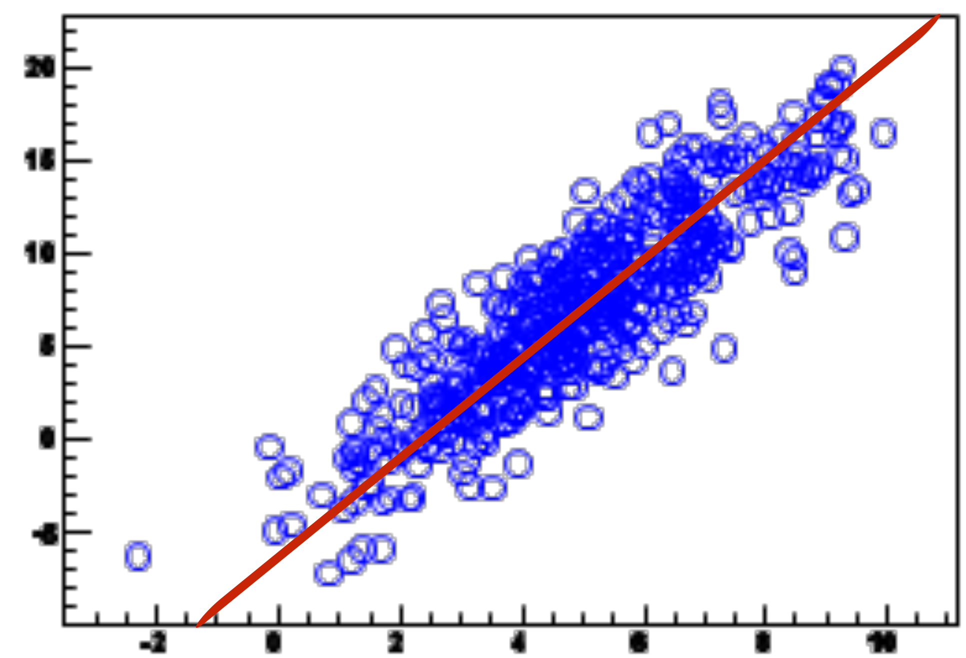
- Measure of location:
  - A point: $(\mu_\mathrm{X}, \mu_\mathrm{Y})$
  - A curve: line which is the closest to the points ⟶ **linear regression**
- Minimizing the dispersion between the curve "$y = ax + b$" and the distribution

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

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

$$\begin{cases} a(\sigma_\mathrm{X}^2 + \mu_\mathrm{X}^2) + b\mu_\mathrm{X} = \rho\sigma_\mathrm{X}\sigma_\mathrm{Y} + \mu_\mathrm{X}\mu_\mathrm{Y} \\ a\mu_\mathrm{X} + b = \mu_\mathrm{Y} \end{cases}$$

$$\boxed{\begin{cases} a = \rho\frac{\sigma_\mathrm{Y}}{\sigma_\mathrm{X}} \\ b = \mu_\mathrm{Y} - \rho\frac{\sigma_\mathrm{Y}}{\sigma_\mathrm{X}}\mu_\mathrm{X} \end{cases}}$$



Fully correlated $\rho = 1$

Fully anti-correlated $\rho = -1$

Then $\mathrm{Y} = a\mathrm{X} + b$

Multidimensional PDFs

- **Multinomial distribution:** randomly choosing $\mathrm{K}_1, \mathrm{K}_2, \ldots \mathrm{K}_\mathrm{S}$ objects within a finite set of n, with a fixed drawing probability for each category $p_1, p_2, \ldots p_\mathrm{S}$ with $\sum \mathrm{K}_i = n$ and $\sum p_i = 1$

Parameters : $n, p_1, p_2, \ldots p_\mathrm{S}$

Law : $P(\vec{k}; n, \vec{p}) = \dfrac{n!}{k_1! k_2! \ldots k_\mathrm{S}!} p_1^{k_1} p_2^{k_2} \ldots p_\mathrm{S}^{k_\mathrm{S}}$

Mean : $\mu_i = np_i$

Variance : $\sigma_i^2 = np_i(1-p_i)$ $\qquad \mathrm{Cov}(\mathrm{K}_i, \mathrm{K}_j) = -np_i p_j$

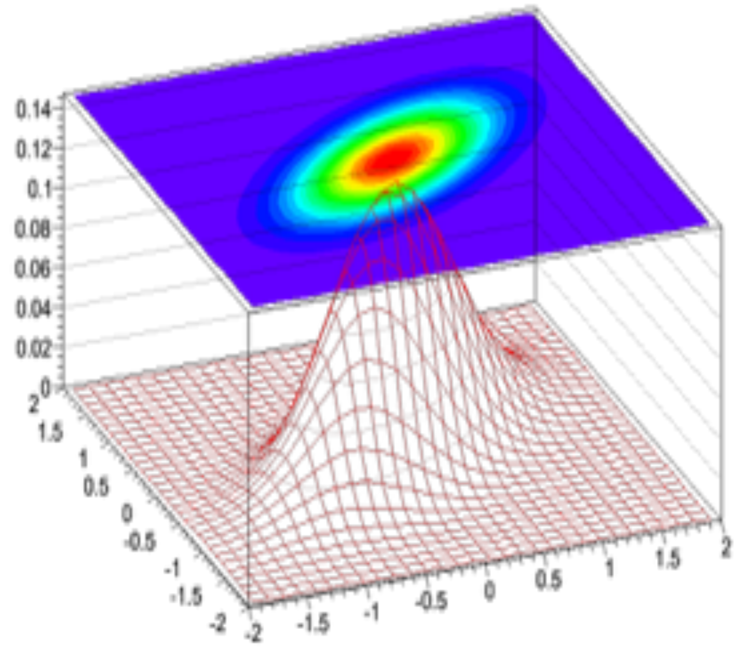
Note: Variables are not independent. The binomial corresponds to $\mathrm{S} = 2$ but has only one independent variable

- **Multinormal distribution:**

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

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

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



Independent $\rightleftarrows$ **Uncorrelated**

Sum of random variables

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

$$\mathrm{S} = \sum_{i=1}^{n} \mathrm{X}_i$$

- Assuming the mean and variance of each variable exist:
  - **Mean value** of S :

$$\mu_{\mathrm{S}} = \int \left(\sum_{i=1}^{n} x_i \right) f(x_1, ..., x_n) dx_1 ... dx_n = \sum_{i=1}^{n} \int x_i f_{\mathrm{X}_i}(x_i) dx_i \boxed{= \sum_{i=1}^{n} \mu_i}$$

The mean is an additive quantity

  - **Variance** of S :

$$\sigma_{\mathrm{S}}^2 = \int \left(\sum_{i=1}^{n} x_i - \mu_{\mathrm{X}_i} \right)^2 f(x_1, ..., x_n) dx_1 ... dx_n$$

$$= \sum_{i=1}^{n} \sigma_{\mathrm{X}_i}^2 + 2 \sum_{i} \sum_{j<i} \mathrm{Cov}(\mathrm{X}_i, \mathrm{X}_j)$$

For **uncorrelated variables**, the variance is an additive quantity

→ used for error combinations

$$\boxed{\sigma_{\mathrm{S}}^2 = \sum_{i=1}^{n} \sigma_{\mathrm{X}_i}^2}$$

Sum of random variables

- Probability density function of S: $f_{\mathrm{S}}(s)$
- Using the characteristic function:

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

For **independent variables:**

$$\phi_{\mathrm{S}}(t) = \prod \int f_{\mathrm{X}_k}(x_k) e^{itx_k} dx_k = \prod \phi_{\mathrm{X}_i}(t)$$

→ The characteristic function factorizes.

- The PDF is the **Fourier transform** of the characteristic function, therefore:

$$f_{\mathrm{S}} = f_{\mathrm{X}_1} * f_{\mathrm{X}_2} * \ldots * f_{\mathrm{X}_n}$$

$\boxed{\text{The PDF of the sum of random variables is the convolution of the individual PDFs}}$

Sum of Normal variables → Normal

Sum of Poisson variables (λ_1 and λ_2) → Poisson with $\lambda = \lambda_1 + \lambda_2$

Sum of Chi-2 variables (n_1 and n_2) → Khi-2 with $n = n_1 + n_2$

Sum of independent variables

- **Weak law of large numbers**

  Sample of size n = realization of n independent variables with the same distribution (mean μ, variance σ^2)

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

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

- **Central limit theorem**

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

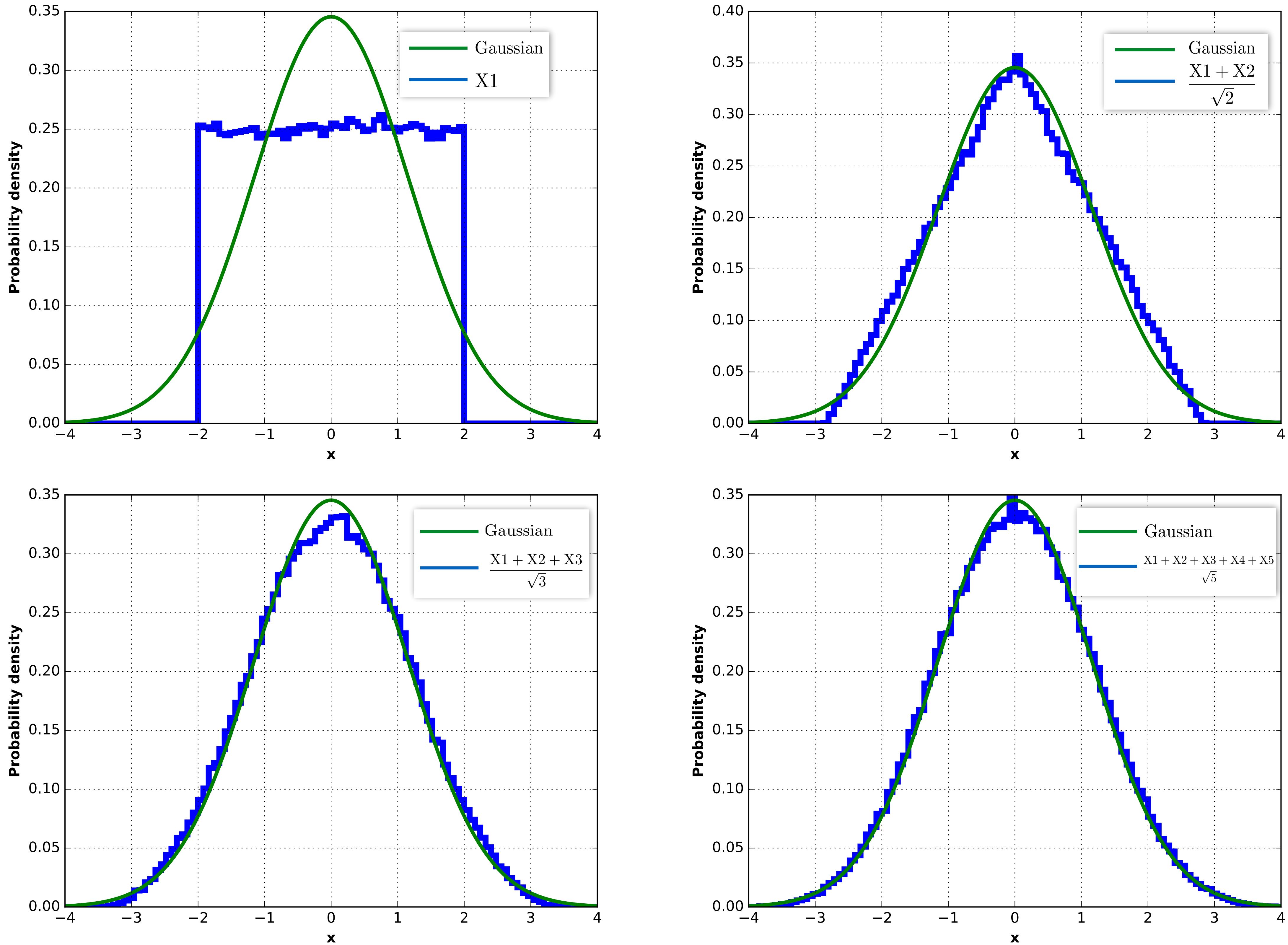
  Sum of the reduced variables: $C = \frac{1}{\sqrt{n}}\sum \frac{X_i - \mu_i}{\sigma_i}$

  The PDF of C converges to a reduced normal distribution:

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

The sum of many random fluctuations is normally distributed

Central limit theorem



Dispersion and uncertainty

- Any measure (or combination of measures) is a realization of a random variable.
  - Measured value: θ
  - True value: θ_0

- The **uncertainty** quantifies the difference between θ and θ_0 :

→ **Measure of dispersion**

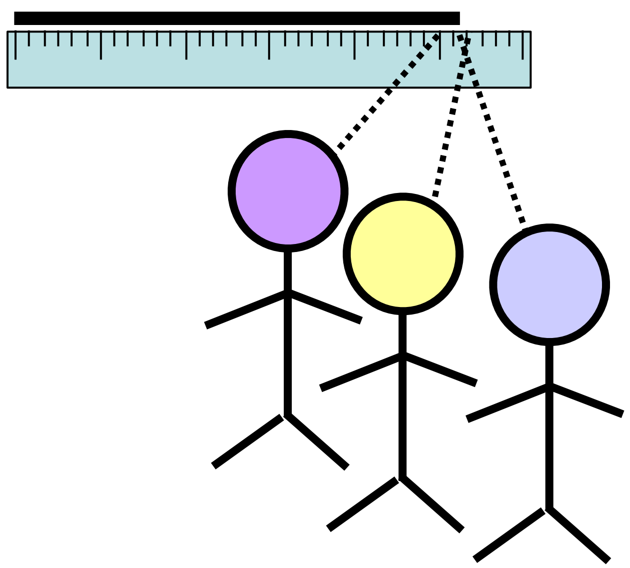
Postulate: $\Delta\theta = \alpha\sigma_\theta$ → Absolute error always positive

- Usually one differentiates:
  - **Statistical errors**: due to the measurement PDF
  - **Systematic errors** or bias: fixed but unknown deviation (equipment, assumptions, …)

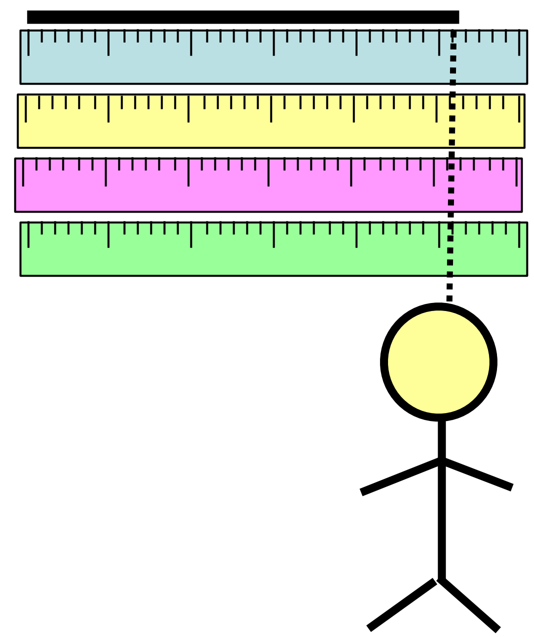
  Systematic errors can be seen as statistical error in a set of similar experiments

Error sources

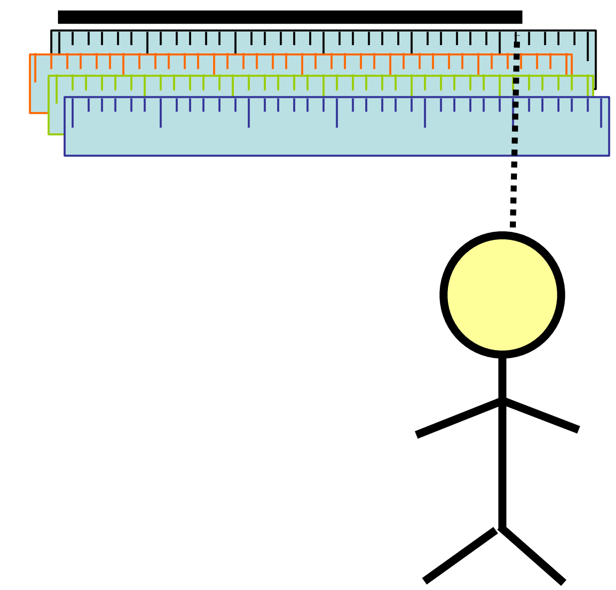
Observation error: Δ_O



Scaling error: Δ_S



Position error: Δ_P



Measured value: $\theta = \theta_0 + \delta_O + \delta_S + \delta_P$

- Each δ_i is a realization of a random variable of mean 0 and variance σ_i^2

  For **uncorrelated error sources**:

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

- **Choice for α :**

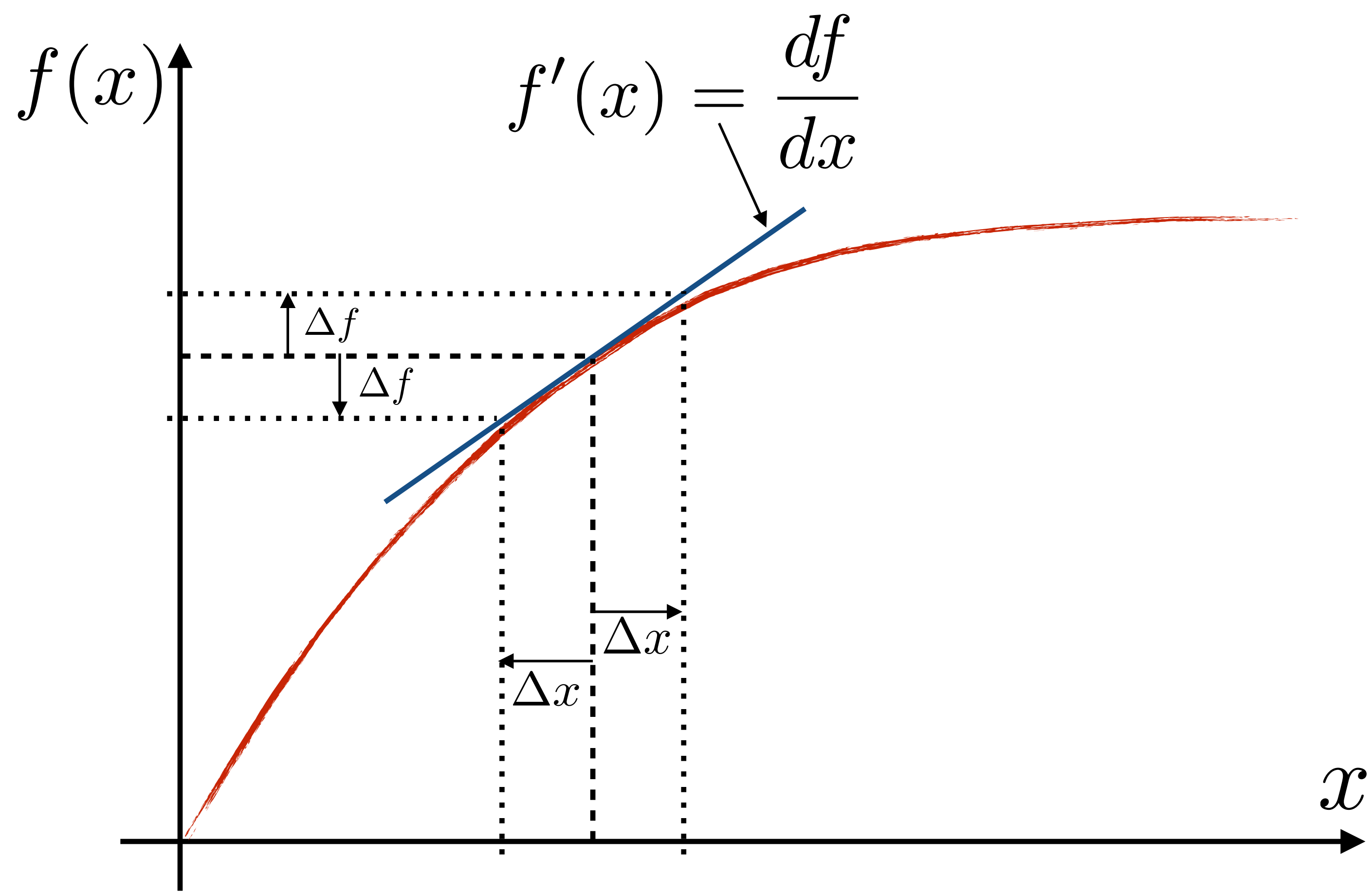
  Many sources of error $\longrightarrow$ central limit theorem $\longrightarrow$ normal distribution

  $\alpha = 1$ gives (approximately) a 68% confidence interval

  $\alpha = 2$ gives a 95% confidence interval

Error propagation

- **Measure:** $x \pm \Delta x$
- **Compute:** $f(x) \longrightarrow \Delta f$ **?**



Assuming small errors and using the Taylor expansion:

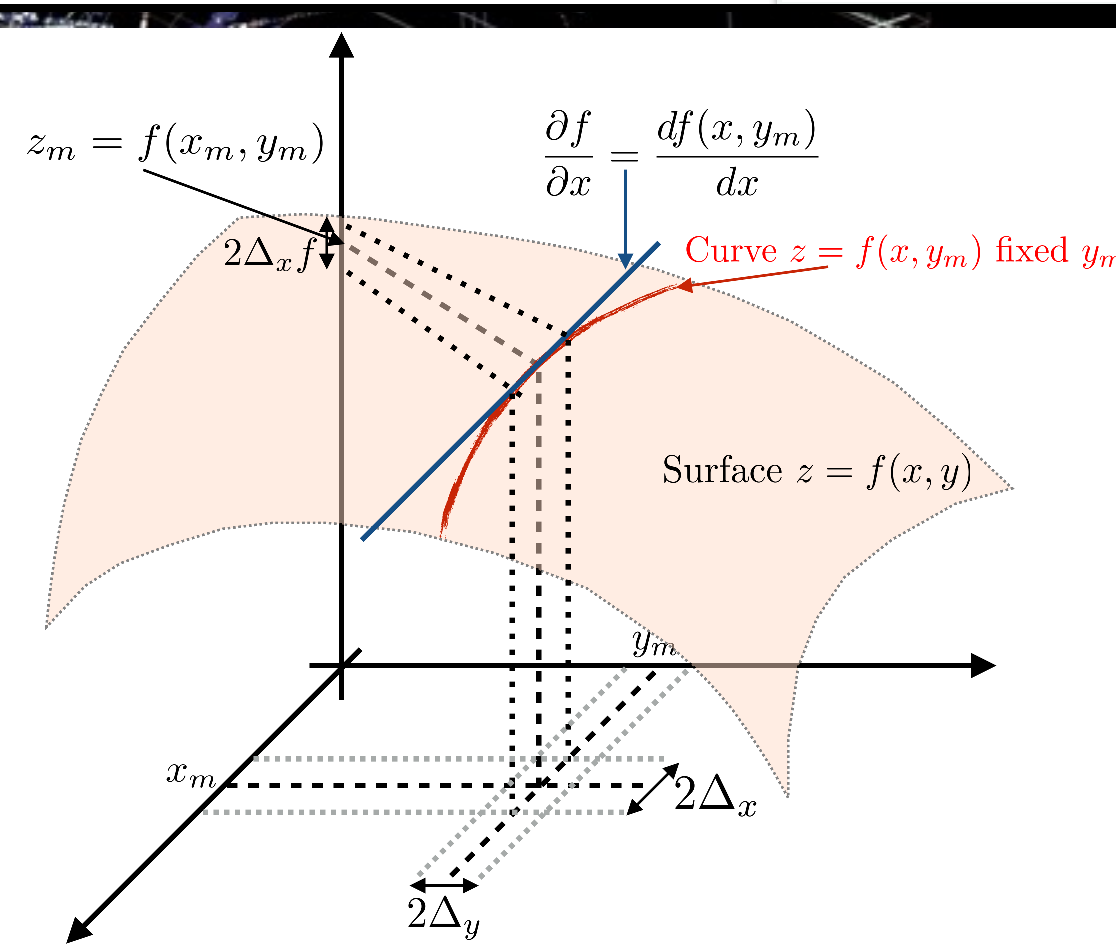
$$f(x+\Delta x) = f(x) + \frac{df}{dx}\Delta x + \frac{1}{2}\frac{d^2 f}{dx^2}\Delta x^2$$

$$f(x-\Delta x) = f(x) - \frac{df}{dx}\Delta x + \frac{1}{2}\frac{d^2 f}{dx^2}\Delta x^2$$

$$\rightarrow \quad \Delta f = \frac{1}{2}\left|f(x+\Delta x) - f(x-\Delta x)\right| \boxed{= \frac{df}{dx}\Delta x}$$

Error propagation

- **Measure:** $x \pm \Delta x\,,\ y \pm \Delta y$
- **Compute:** $f(x, y, \ldots) \longrightarrow \Delta f$ **?**



Method: Treat the effect of each variable as separate **error sources**

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

Then:

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

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

Uncorrelated | **Correlated** | **Anticorrelated**

$$\Delta f^2 = \sum_i \left(\frac{\partial f}{\partial x_i}\Delta x_i\right)^2 \qquad \Delta f = \left|\frac{\partial f}{\partial x}\right|\Delta x + \left|\frac{\partial f}{\partial y}\right|\Delta y \qquad \Delta f = \left|\left|\frac{\partial f}{\partial x}\right|\Delta x - \left|\frac{\partial f}{\partial y}\right|\Delta y\right|$$

Parametric estimation

- Estimating a parameter θ from a finite sample $\{x_i\}$
- **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 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 others

→ Need to make a choice for a given situation

Bias and consistency

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

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

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

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

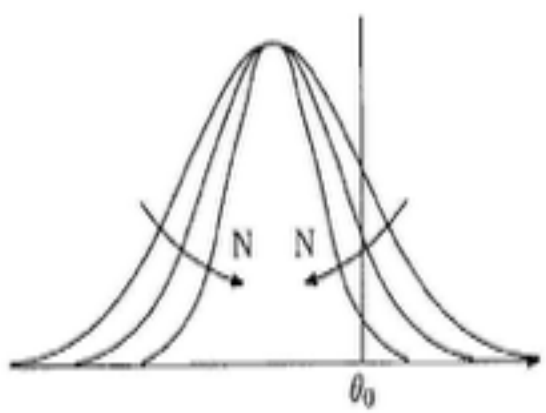
Unbiased estimator: $b(\hat{\theta}) = 0$

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

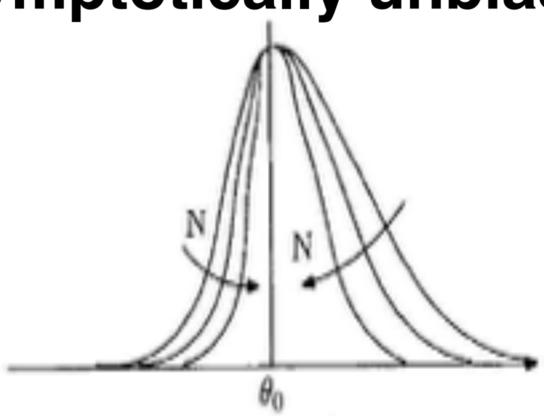
- **Consistency:** formally $\mathrm{P}(|\hat{\theta} - \theta| < \epsilon) \xrightarrow[n \to +\infty]{} 1, \ \forall \epsilon$

In practice, if the estimator is asymptotically unbiased $\boxed{\sigma_{\hat{\Theta}} \xrightarrow[n \to +\infty]{} 0}$

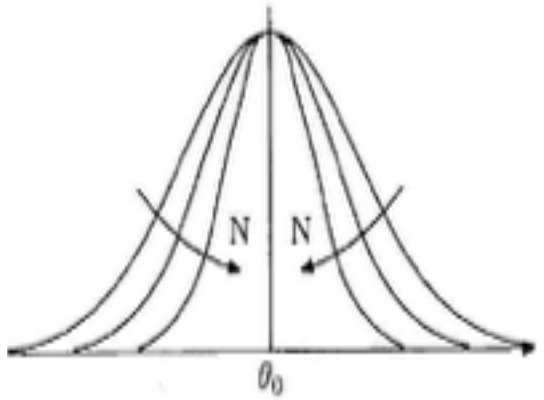
Biased



Asymptotically unbiased



Unbiased



Efficiency

- For any unbiased estimator of θ , the variance cannot exceed (Cramer-Rao bound):

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

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

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

  $\longrightarrow$ Minimal variance estimator

- The minimal variance estimator will often be biased, asymptotically unbiased

Empirical estimator

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

  $\longrightarrow$ weak law of large numbers: convergent, unbiased

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

- **Sample variance** as an estimator of the **population variance:**

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

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

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

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

Errors on estimators

Uncertainty ⟷ **Estimator standard deviation**

- Use an estimator of standard deviation: $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ (biased !)
- **Mean:** $\hat{\mu} = \frac{1}{n} \sum x_i \,,$ $\sigma^2_{\hat{\mu}} = \frac{\sigma^2}{n}$ $\longrightarrow$ $\boxed{\Delta\hat{\mu} = \sqrt{\frac{\hat{\sigma}^2}{n}}}$
- **Variance:** $\hat{\sigma}^2 = \frac{1}{n-1} \sum_i (x_i - \hat{\mu})^2,$ $\sigma^2_{\hat{\sigma}^2} \approx \frac{2\sigma^4}{n}$ $\longrightarrow$ $\boxed{\Delta\hat{\sigma}^2 = \sqrt{\frac{2}{n}}\hat{\sigma}^2}$
- Central-Limit theorem $\longrightarrow$ empirical estimators of mean and variance are **normally distributed** for large enough samples

$\hat{\mu} \pm \Delta\hat{\mu} \,, \;\; \hat{\sigma} \pm \Delta\hat{\sigma}$ define 68% confidence intervals

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

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

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

Maximum likelihood

- Let a sample of measurements: $\{x_i\}$

  The analytical form of the density is known and depends on several unknown parameters θ

  For example: Event counting follows a Poisson distribution with a parameter $\lambda_i(\theta)$ depending on the physics.

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

- An estimator of the parameters θ is given by the position of the **maximum of the likelihood function**

→ Parameter values which maximize the probability to get the observed results

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

Note: system of equations for several parameters

Note: minimizing $-\ln\mathcal{L}$ often simplify the expression

Properties of MLE

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

Asymptotically **unbiased**

Asymptotically **efficient** (reaches the Cramer-Rao bound)

Asymptotically **normally distributed**

→ Multinormal law with covariance given by a generalization of the CR bound:

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

- **Goodness of fit:** The value of $-2\ln\mathcal{L}(\hat{\theta})$ is Chi-2 distributed with

$$\mathrm{ndf} = \mathrm{sample\ size} - \mathrm{number\ of\ parameters}$$

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

Probability of getting a worse agreement

Errors on MLE

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

- Errors on the parameters given by the covariance matrix
- For **one parameter**, 68% confidence 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:

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

Confidence contours are defined by the equation:

$$\Delta\ln\mathcal{L} = \beta(n_\theta, \alpha) \text{ with } \alpha = \int_0^{2\beta} f_{\chi^2}(x; n_\theta)dx$$

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

$n_\theta \rightarrow$ / $\alpha \downarrow$	1	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

Least squares

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

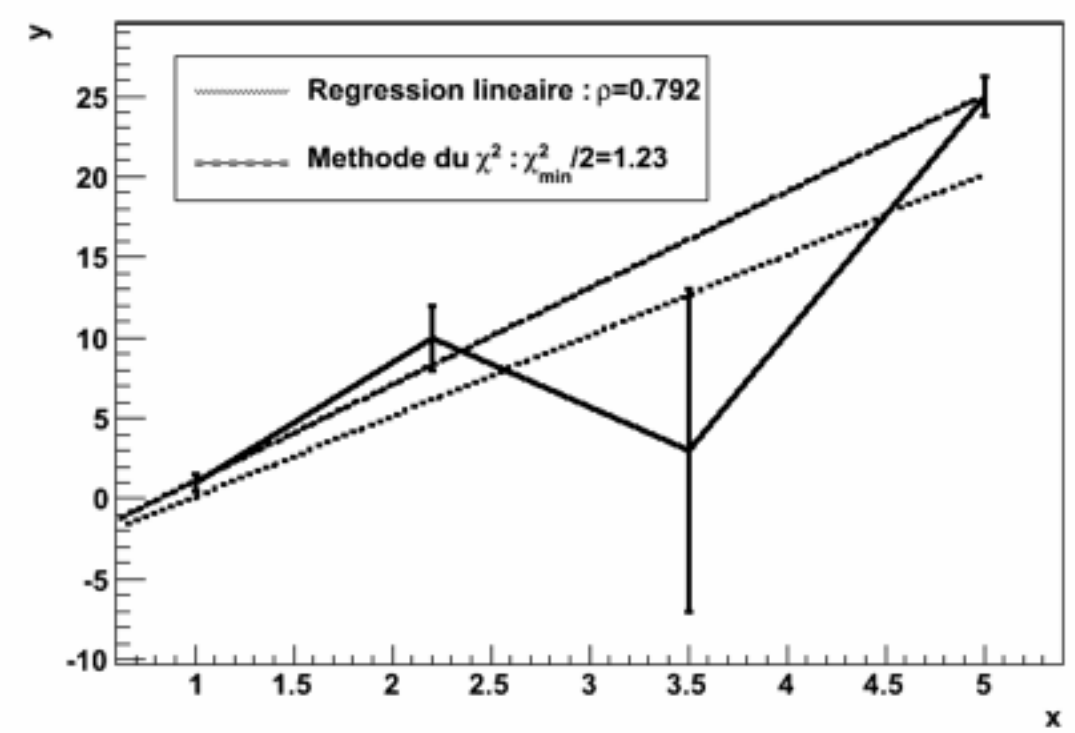
  Theoretical law given by: $y = f(x, \theta)$

- Naive 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 its associated 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 assumes that each y_i is normally distributed with a mean equal to $f(x_i, \theta)$ and a standard deviation given by Δy_i

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

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

Least squares or **Chi-2 fit** is the maximum likelihood estimator for Gaussian errors

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

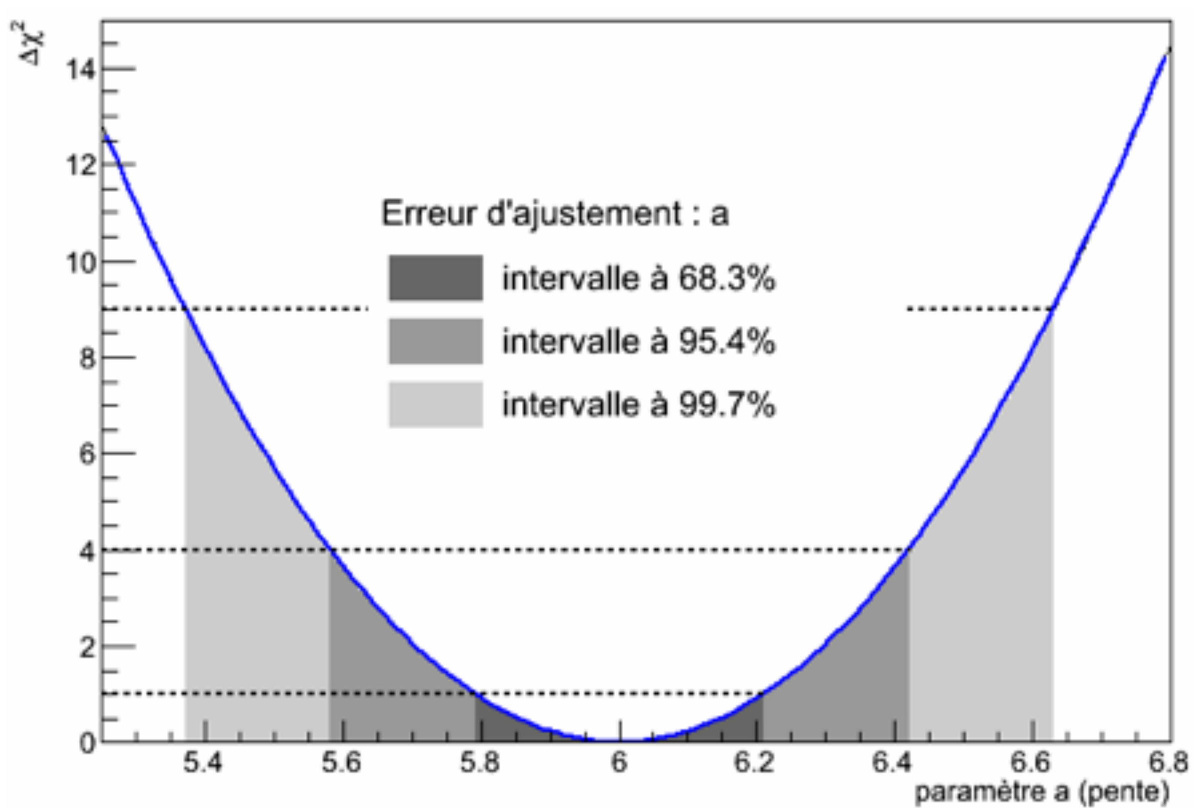
Example: fitting a line

- For $f(x) = ax$ $\quad K^2(a) = \mathrm{A}a^2 - 2\mathrm{B}a + \mathrm{C} = -2\ln\mathcal{L}$

$$\mathrm{A} = \sum_i \frac{x_i^2}{\Delta y_i^2}\,, \quad \mathrm{B} = \sum_i \frac{x_i y_i}{\Delta y_i^2}\,, \quad \mathrm{C} = \sum_i \frac{y_i^2}{\Delta y_i^2}$$

$$\frac{\partial K^2}{\partial a} = 2\mathrm{A}a - 2\mathrm{B} = 0 \quad \rightarrow \quad \boxed{\hat{a} = \frac{\mathrm{B}}{\mathrm{A}}}$$

$$\frac{\partial^2 K^2}{\partial a^2} = 2\mathrm{A} = \frac{2}{\sigma_a^2} \quad \rightarrow \quad \boxed{\Delta\hat{a} = \sigma_a = \frac{1}{\sqrt{\mathrm{A}}}}$$



Example: fitting a line

- For $f(x) = ax + b$

$$K^2(a,b) = \mathrm{A}a^2 + \mathrm{B}b^2 + 2\mathrm{C}ab - 2\mathrm{D}a - 2\mathrm{E}b + \mathrm{F} = -2\ln\mathcal{L}$$

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

$$\left.\begin{aligned} \frac{\partial K^2}{\partial a} &= 2\mathrm{A}a + 2\mathrm{C}b - 2\mathrm{D} = 0 \\ \frac{\partial K^2}{\partial b} &= 2\mathrm{C}a + 2\mathrm{B}b - 2\mathrm{E} = 0 \end{aligned}\right\}$$

$$\boxed{\hat{a} = \frac{\mathrm{BD} - \mathrm{EC}}{\mathrm{AB} - \mathrm{C}^2}\,, \qquad \hat{b} = \frac{\mathrm{AE} - \mathrm{BC}}{\mathrm{AB} - \mathrm{C}^2}}$$

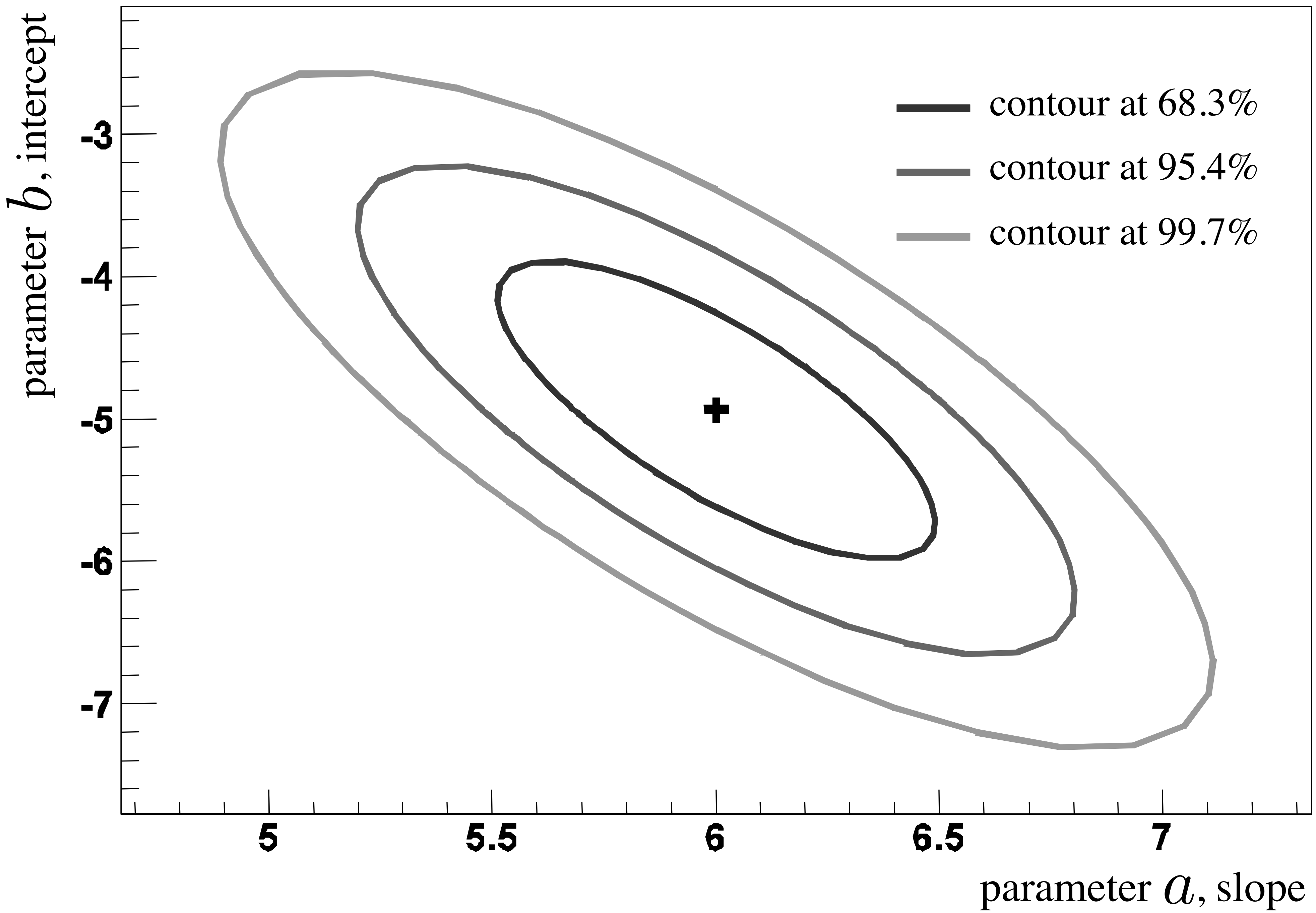
$$\left.\begin{aligned} \frac{\partial^2 K^2}{\partial a^2} &= 2\mathrm{A} = 2\Sigma^{-1}_{11} \\ \frac{\partial^2 K^2}{\partial b^2} &= 2\mathrm{B} = 2\Sigma^{-1}_{22} \\ \frac{\partial^2 K^2}{\partial a \partial b} &= 2\mathrm{C} = 2\Sigma^{-1}_{12} \end{aligned}\right\}$$

$$\Sigma^{-1} = \begin{bmatrix} \mathrm{A} & \mathrm{C} \\ \mathrm{C} & \mathrm{B} \end{bmatrix} \longrightarrow \Sigma = \frac{1}{\mathrm{AB} - \mathrm{C}^2}\begin{bmatrix} \mathrm{B} & -\mathrm{C} \\ -\mathrm{C} & \mathrm{A} \end{bmatrix}$$

$$\boxed{\Delta\hat{a} = \sigma_a = \sqrt{\frac{\mathrm{B}}{\mathrm{AB} - \mathrm{C}^2}}\,, \qquad \Delta\hat{b} = \sigma_b = \sqrt{\frac{\mathrm{A}}{\mathrm{AB} - \mathrm{C}^2}}}$$

Example: fitting a line

- Two dimensional error contours on a and b



Non-parametric estimation

- Directly estimating the probability density function
  - Likelihood ratio discriminant
  - Separating power of variables
  - Data / Monte Carlo agreement
  - …

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

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

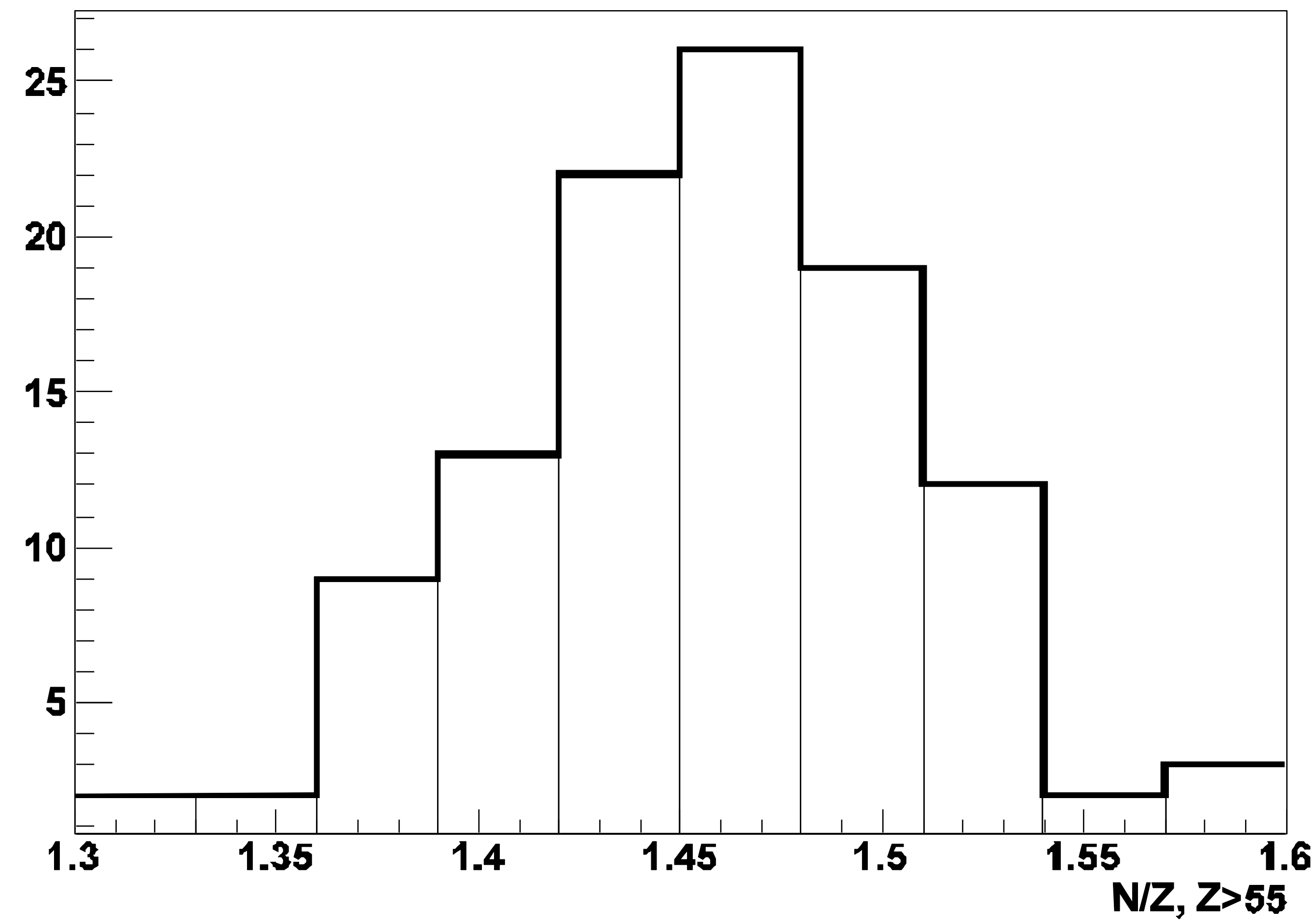
- **Histogram:** Graphical representation of the frequency table $h(x) = n_k \text{ if } x \in C_k$

Histogram

Bin	Number of N/Z	Frequency	Bin	Number of N/Z	Frequency
< 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



Histogram as PDF estimator

- Statistical description: n_k are multinomial random variables

Parameters: $n = \sum_k n_k$ $\qquad p_k = \mathrm{P}(x \in C_k) = \int_{C_k} f_{\mathrm{X}}(x)dx$

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

For a large sample:

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

For small classes (width δ):

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

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 function**

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

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

Confidence interval

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

$$\mathrm{P}(\mathrm{X} \in [a, b]) = \int_a^b f_{\mathrm{X}}(x)dx = \alpha$$

Probability of finding a realization inside the interval

- Generalization of the concept of uncertainty: interval that contains the true value with a given probability

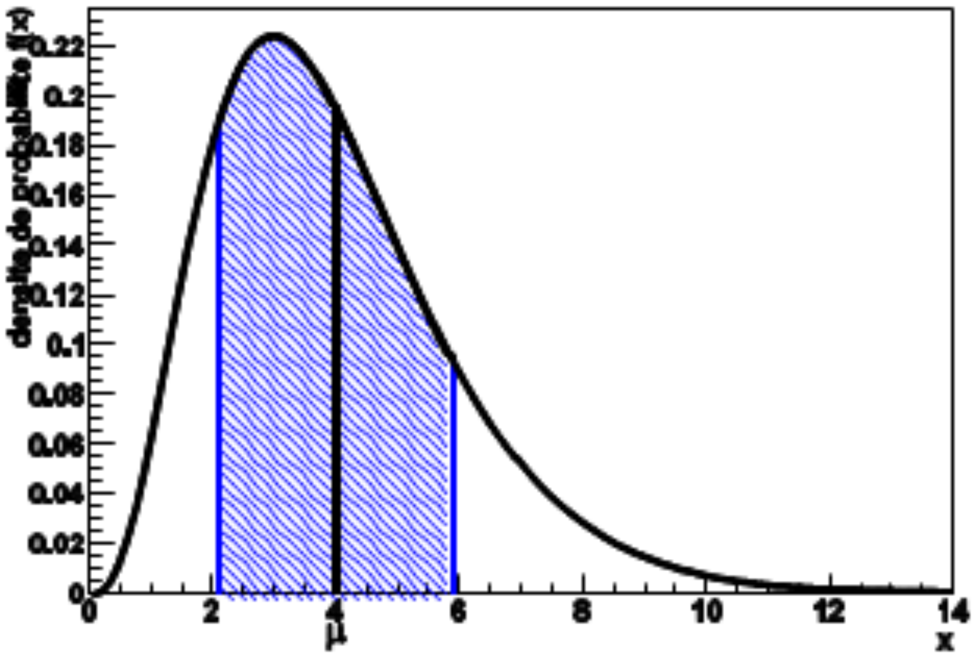
- For **Bayesians:** the posterior density is the probability density of the true value.

  ➡ It can be used to estimate an interval: $\mathrm{P}(\theta \in [a, b]) = \alpha$

- No such thing for a Frequentist: the interval itself becomes the random variable $[a, b]$ is a realization of $[\mathrm{A}, \mathrm{B}]$

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

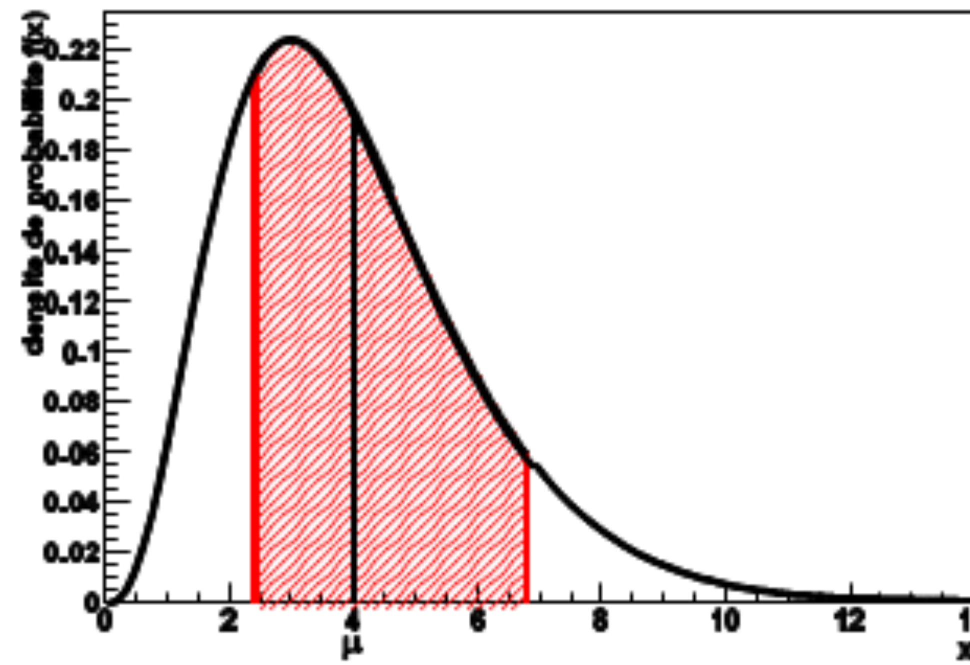
Confidence interval



- Mean centered, symetric interval:

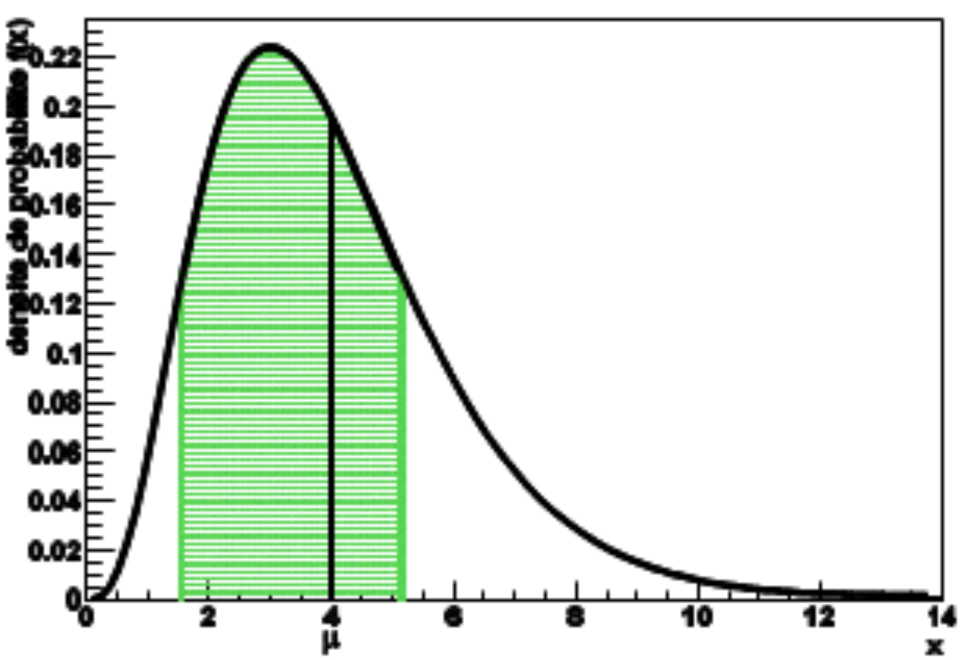
$$[\mu - a, \mu + a]$$

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



- 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) interval: $[a, b]$

$$\int_a^{b} f(x)dx = \alpha$$

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

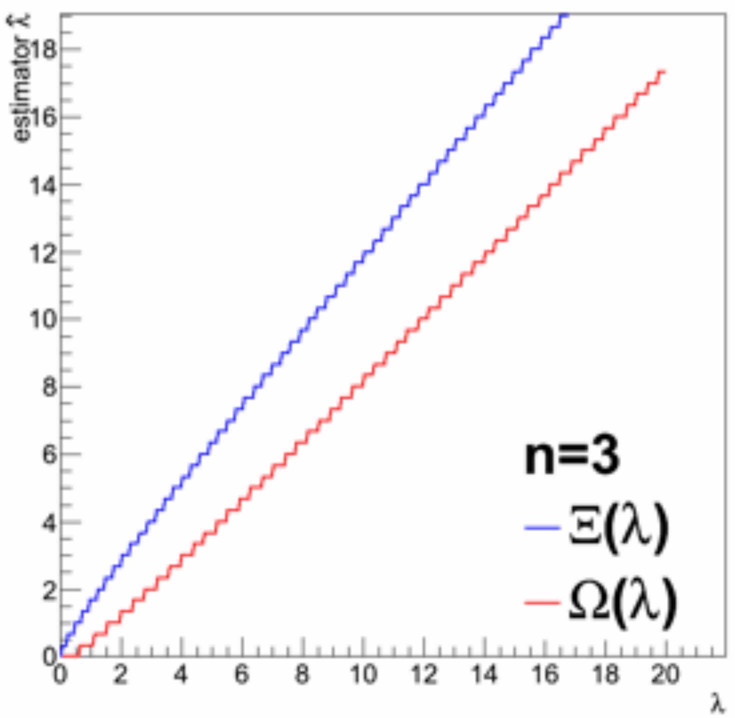
Confidence Belt

- To build a frequentist interval for an estimator $\hat{\theta}$ of θ:
  1. Make pseudo-experiments for several values of θ and compute the estimator $\hat{\theta}$ for each (Monte Carlo sampling of the estimator PDF)
  2. For each θ, determine $\Xi(\theta)$ and $\Omega(\theta)$ such that:

     $\hat{\theta} < \Xi(\theta)$ 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 **confidence level** α
  3. Inverse these functions. The interval $[\Omega^{-1}(\hat{\theta}), \Xi^{-1}(\hat{\theta})]$ satisfies:



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

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

Dealing with systematics

- The variance of the estimator only measures the statistical uncertainty.

- Often, we will have to deal with **parameters** whose **value is known with limit precision.**

  → **Systematic uncertainties**

- The likelihood function becomes:

$$\mathcal{L}(\theta, \nu) \quad \text{with} \quad \nu = \nu_0 \pm \Delta\nu \text{ or } {\nu_0}^{+\Delta\nu_+}_{-\Delta\nu_-}$$

The known parameters ν are **nuisance parameters**

Bayesian inference

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

- Usually a multinormal law is used with mean ν_0 and covariance matrix estimated from $\Delta\nu_0$ (+ correlation if needed)

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

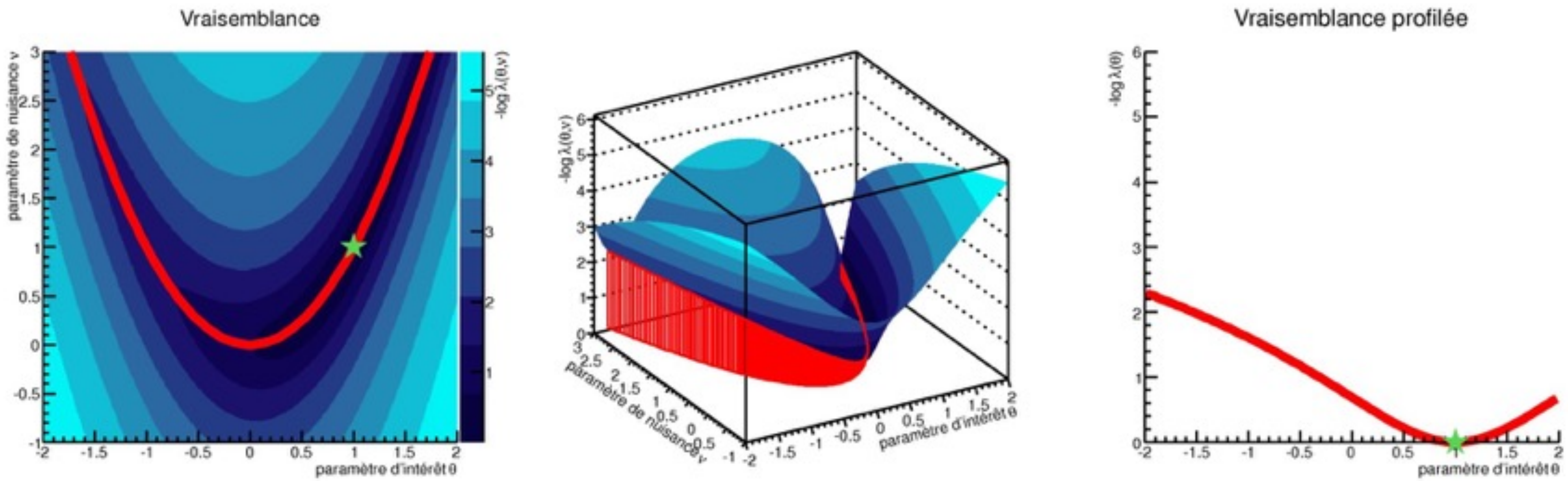
- The final posterior distribution is obtained by **marginalization** over the nuisance parameters:

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

Profile Likelihood

- No true frequentist way to add systematic effects. Popular method of the day: **profiling**
- Deal with nuisance parameters as realization of random variables:

  → extend the likelihood: $\mathcal{L}(\theta, \nu) \longrightarrow \mathcal{L}'(\theta, \nu)\mathcal{G}(\nu)$
- $\mathcal{G}(\nu)$ is the likelihood of the new parameters (identical to prior)
- For each value of θ, maximize the likelihood with respect to nuisance: profile likelihood $\mathrm{PL}(\theta)$
- $\mathrm{PL}(\theta)$ has the same statistical asymptotical properties than the regular likelihood



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 quantifies 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: $p_i = \int_{C_i} f(x)dx$
  - For each bin, the test statistics compares 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

- χ^2 follows (asymptotically) a Chi-2 law with $k-1$ degrees of freedom (one constraint $\sum n_i = n$)
- **p-value:** probability of doing worse: $\mathrm{p-value} = \int_{\chi^2}^{+\infty} f_{\chi^2}(x; k-1)dx$

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)

Kolmogorov-Smirnov test

- Test for unbinned data: compare the sample cumulative density function to the tested one
- Sample PDF (ordered sample)

$$f_{\mathrm{S}}(x) = \frac{1}{n}\sum_i \delta(x-i) \longrightarrow F_{\mathrm{S}}(x) = \begin{cases} 0 & x < x_0 \\ \frac{k}{n} & x_k \leq x < x_{k+1} \\ 1 & x > x_n \end{cases}$$

- The Kolmogorov statistic is the largest deviation:

$$D_n = \sup_x |F_{\mathrm{S}}(x) - F(x)|$$

- The test distribution has been computed by Kolmogorov:

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

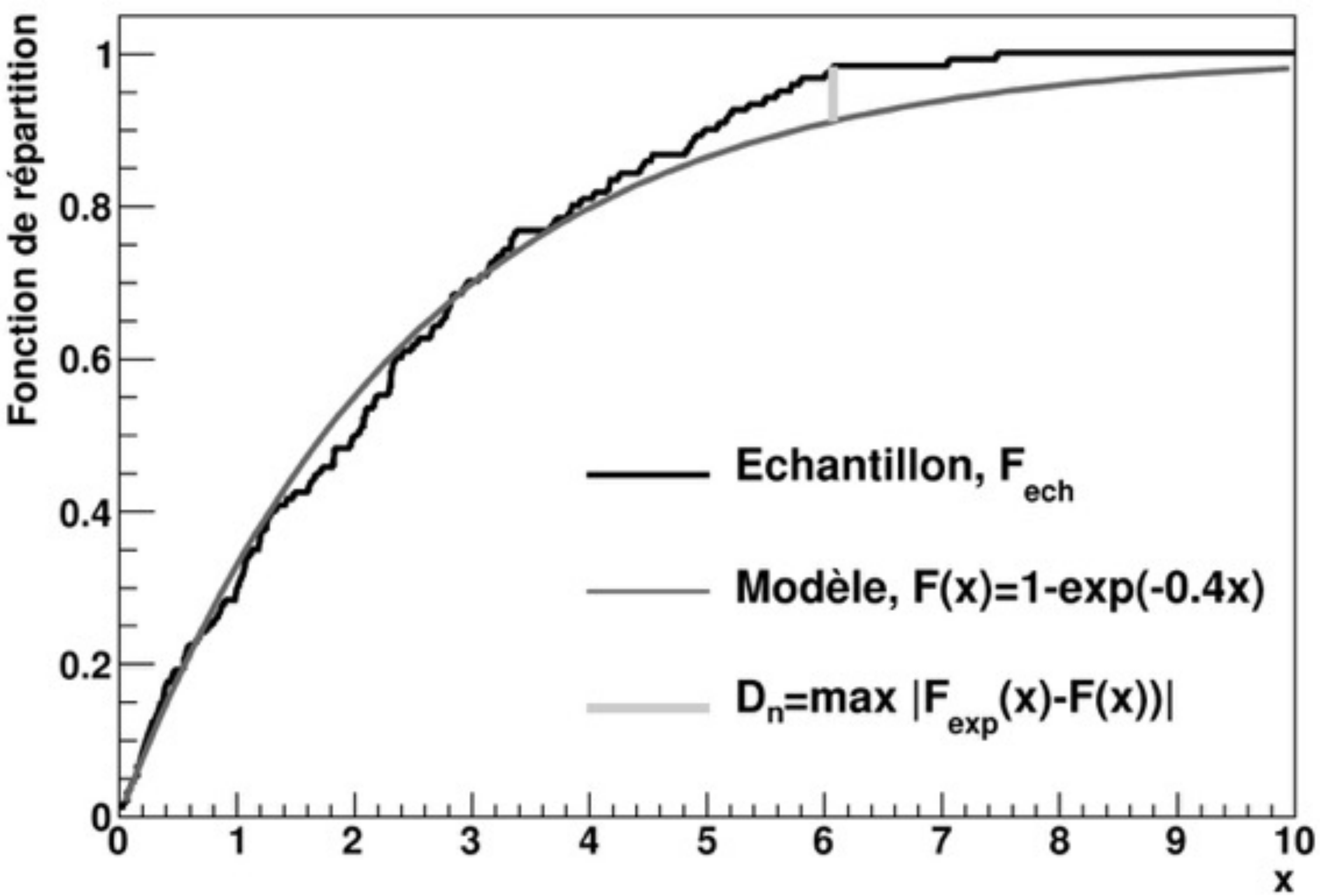
$[0, \beta]$ defines a confidence interval for D_n

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

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



$$D_n = 0.069$$

$$\mathrm{p-value} = 0.0617$$

$$1\sigma : \ [0, 0.0875]$$