Uncertainty Quantification

F. Gaudier

fabrice.gaudier@cea.fr

HPC and Uncertainty Treatment
Examples with OPEN TURNS and URANIE

EDF R&D - Phimeca - IMACS - Airbus Group - CEA

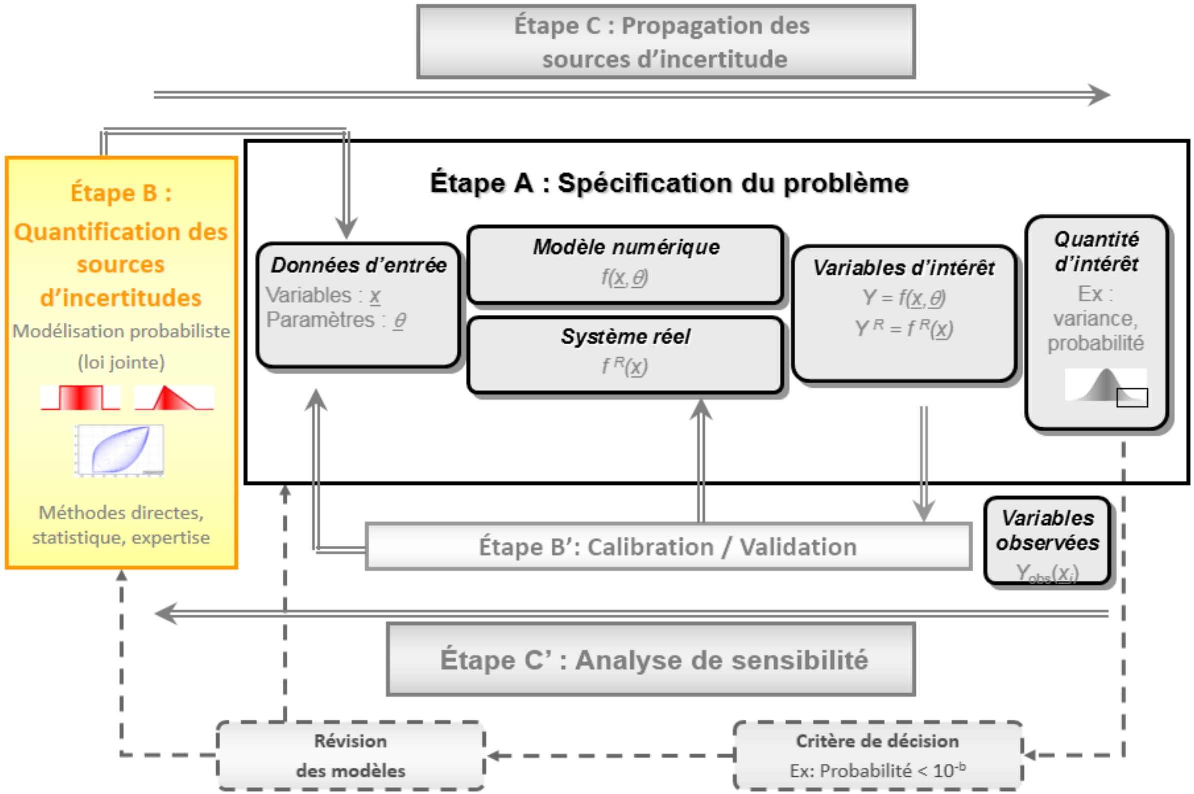
PRACE ADVANCED TRAINING CENTER

Maison de la Simulation, France

2016/05/17-19

www.cea.fr

General Flowchart



Outline

- Descriptive Statistics
  - Univariate case
  - Bivariate case
- Data Modelisation with Prabability Density Function (***PDF***)
  - Commonly used PDF
  - Parametric Probability Density Estimation
  - Nonparametric Probability Density Estimation
- Goodness-of-Fit Techniques
  - Graphical Method
  - Statistical Tests Methods

Univariate Case : "Location" parameters

The effect of the "location" parameter is to translate the graph relative to the standard distribution (nS is the size of the sample)

- **Mean μ** :

$$\mu \quad = \quad \frac{1}{nS} \sum_{i=1}^{nS} x_i$$

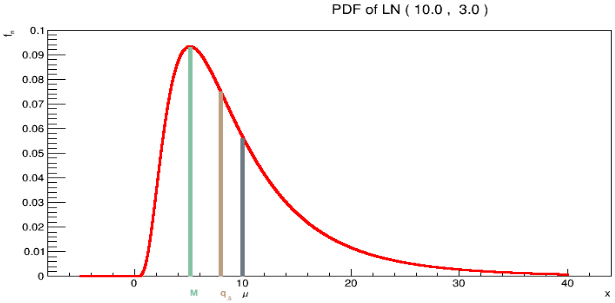
- **Mode M** : Value where the probability is the greatest value
- **Mediane $q_{0.5}$** : it is the 0.5-quantile

$$q_{0.5} \qquad as \qquad I\!P\left[X \leq q_{0.5}\right] \;=\; 0.5 \;=\; I\!P\left[X \geq q_{0.5}\right]$$

- **α-Quantile q_α with $\alpha \in [0, 1]$** :

$$q_\alpha \qquad as \qquad I\!P\left[X \leq q_\alpha\right] \;=\; \alpha$$

- **Quartiles $q_{0.25}, q_{0.50}, q_{0.75}$**
- **Extremes values min, max**



Univariate Case : "Dispersion" parameters

The effect of a "dispersion" parameter is to stretch|shrink the standard distribution

- **Variance** $Var[X]$: measure of spread in the data about the mean $Var[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$, and can be estimated by :

$$Var[X] = \frac{1}{nS - 1} \sum_{i=1}^{nS} (x_i - \mu)^2$$

- **Standard Deviation** σ : to have an information in the same unit as the variable

$$\sigma = \sqrt{Var[X]}$$

- **Coefficient of Variation** δ : σ does not indicate **the degree (%)** of dispersion around the mean value μ, a **nondimensional** term can be introduced :

$$\delta = \frac{\sigma}{\mu}$$

- **Range** R :

$$R = Max - Min$$

- **InterQuartile interval** H :

$$H = q_{0.75} - q_{0.25}$$

Univariate Case : "Shape" parameters

A "shape" parameter is any parameter of a PDF that is neither a location parameter nor a scale parameter. Such a parameter must affect the shape of a distribution rather than simply shifting it (location parameter) or stretching/shrinking it (dispersion parameter).

- **Moment order p : μ_p** $\quad := \quad \mathbb{E}[(X - \mathbb{E}[X])^p]$

$$\mu_p \quad = \quad \frac{1}{nS} \sum_{i=1}^{nS} (x_i - \mu)^p$$

- **Skewness : γ_1** is a measure of the asymmetry of the PDF

$$\gamma_1 \; := \; \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3 \right] \quad = \quad \frac{\mu_3}{\sigma^3} \quad = \quad \frac{\mathbb{E}[X^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3}$$

- **Kurtosis : γ_2** is a measure of the "peakedness" of the PDF

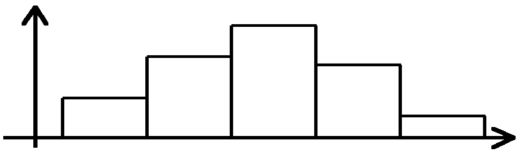
$$\gamma_2 \quad = \quad \frac{\mu_4}{\sigma^4} - 3.0$$

Univariate Case : Graphical Representations

- **Histogram**

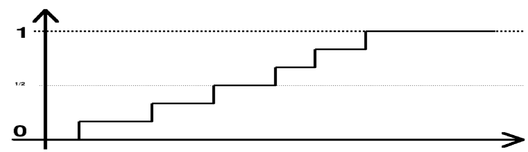
$$H(x) = \frac{\sum_{i=1}^{nS} \mathbb{1}_{[t_i,t_{i+1}]}(x_i)}{nS(t_{i+1} - t_i)} \quad \text{when} \quad x \in [t_i, t_{i+1}]$$

where $[a,b] = \bigcup_i [t_i, t_{i+1}]$

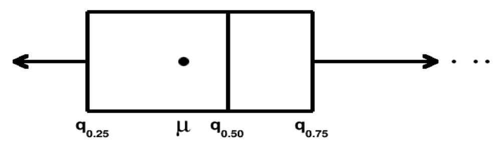


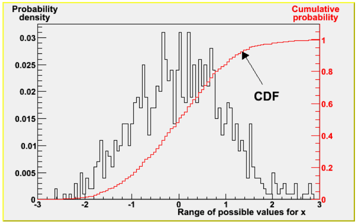
- **Empirical Cumulative Density Function (*eCDF*)**

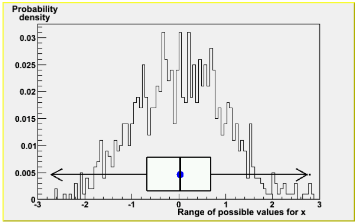
$$F_n(x) \quad = \quad \frac{1}{nS}\sum_{i=1}^{nS} \mathbb{1}(X_i \leq x)$$



- **Boxplot (Tukey)**



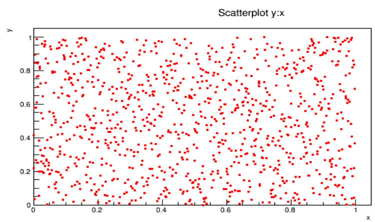




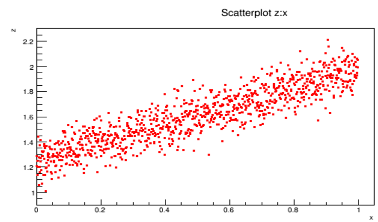
Bivariate Case : Graphics

Detect and describe statistical dependences between variables

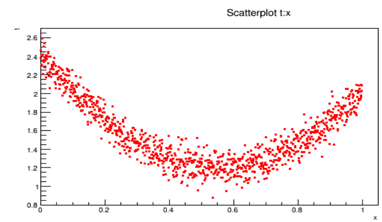
- independent variables $\Rightarrow$ uncorrelated variables ($\rho = 0$)
- but uncorrelated variables $\not\Rightarrow$ independant variables



uncorrelated



linear correlation



nonlinear correlation

The *covariance* is a measure of how much two random variables change together

$$Cov(X,Y) \quad = \quad \mathbb{E}[X - \mathbb{E}[X]] \times \mathbb{E}[Y - \mathbb{E}[Y]]$$

and the covariance estimated on a sample $(x_i\ ,\ y_i)$ is:

$$\widehat{Cov}(x,y) \quad = \quad \frac{1}{nS-1}\sum_{i=1}^{nS}(x_i - \bar{x})(y_i - \bar{y})$$

The sign the tendency in the linear relationship between the variables, but the magnitude is not easy to interpret ($\Rightarrow$ found a normalized version)

Pearson's Correlation Coefficien ("PCC")

The Pearson's Correlation Coefficien (*"PCC"*) is the normalized version of the covariance (the covariance is divided by the product of the two standard deviations)

It is a measure of the linear correlation (dependence) between two variables X and Y

$$\rho_{\mathrm{P}}(X,Y) \quad = \quad \frac{Cov(X,Y)}{\sigma_{\mathrm{X}}\,\sigma_{\mathrm{Y}}}$$

and the estimation on a sample $(x_i\ ,\ y_i)$, noted $\widehat{r_p}$, is given by :

$$\widehat{r_p} \quad = \quad \frac{\sum_{i=1}^{nS}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{nS}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{nS}(y_i-\bar{y})^2}}$$

- $\widehat{r_p} \in [-1., 1.]$
- if $\widehat{\mathrm{r_p}} = -1$ or $+1$, implies that a linear equation describes the relationship between X and Y perfectly, and the data points lying exactly on a line
- $\widehat{r_p} = 0.$ implies that X and Y are uncorrelated, but they can be dependents :

  Example: if $X \sim \mathcal{N}(0,1)$, then X and X^2 are uncorrelated (i.e. $\widehat{r_p} = 0.$) but they are dependents

Spearman's Rank Correlation Coefficient

The Spearman's Rank Correlation Coefficientit ("SRCC" noted ρ_s) is a measure of the **monotone** dependence between two variables X and Y

It is defined as the Pearson correlation coefficient between the ranked variables F:

$$\rho_S \quad = \quad \rho_P\left(F_X(X), F_Y(Y)\right)$$

with F_X the CDF of the distribution X
With a sample (x_i, y_i), the n raw values x_i, y_i are converted to ranks values $x_{(i)}, y_{(i)}$:

- $x_{(i)} \in [1, 2, \cdots, nS]$
- $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(nS)}$
- The mean of $(x_{(i)})$ is $\overline{x_{()}} = \frac{nS+1}{2}$

Then r_s is computed from the PCC formula with the ranks values:

$$r_s \quad = \quad \frac{\sum_i (x_{(i)} - \overline{x_{()}})(y_{(i)} - \overline{y_{()}})}{\sqrt{\sum_i (x_{(i)} - \overline{x_{()}})^2 \sum_i (y_{(i)} - \overline{y_{()}})^2}}$$

- $\widehat{r_s} \in [-1., 1.]$
- if $\widehat{\mathrm{r_s}} = -1$ or $+1$, implies that exists a monotone relationship between X and Y
- $\widehat{r_s} = 0.$ implies that X and Y are uncorrelated monotonically

Kendall Rank Correlation Coefficient (τ)

the Kendall Rank Correlation Coefficient (τ) is a measure of the *"association"* between two variables X and Y

$$\tau(X,Y) = \frac{(\text{Number of concordant pairs}) - (\text{Number of discordant pairs})}{\frac{1}{2}nS(nS-1)}.$$

where, any pairs (x_i, y_i) and (x_j, y_j) of sample (x_i, y_i) are said to be :

- **concordant** if the ranks for both elements agree
  if both $x_i > x_j$ and $y_i > y_j$ or
  if both $x_i < x_j$ and $y_i < y_j$
- **discordant** if the ranks for both elements disagree
  if $x_i > x_j$ and $y_i < y_j$ or
  if $x_i < x_j$ and $y_i > y_j$.

The denominator is the total number pair combinations, so $-1 \leq \tau \leq 1$

- If the agreement between the two rankings is perfect, $\tau = 1$
- If the disagreement between the two rankings is perfect, $\tau = -1$
- If X and Y are independent, then we would expect $\tau \simeq 0$

Copulas

- the notion of *"Copula"* was introduced to separate the effect of dependence from the effect of marginal distributions in a multivariate distribution
- Copulas are functions that join or *"couple"* multivariate distributions to their one-dimensional marginal distributions
- Alternatively, Copulas are multivariate distributions whose one-dimensional margins are uniform on $[0, 1]$
- Definition: ***Sklar (1959)***

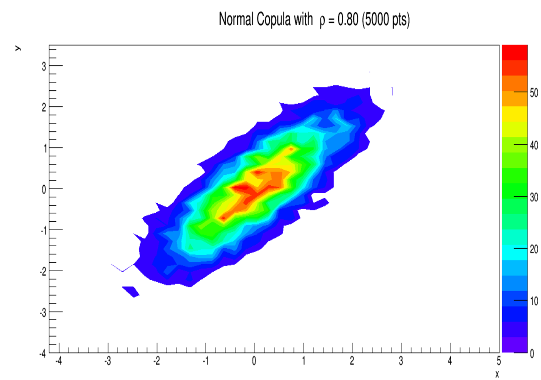
  Distributions X with marginal F_X and Y with marginal F_Y are joined by copula C if their joint distribution can be written by

  $$F_{XY}(x, y) = C(F_X(x), F_Y(y))$$

- Every **continous bivariate** distrubution can be represented in terms of a **unique copula**
- Exemple: The normal copula with correlation ρ

  $$C_\rho(u, v) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v)) \qquad u, v \in [0, 1]^2$$

  with Φ^{-1} the inverse of the standard univariate normal PDF and Φ_ρ is the bivariate normal CDF with correlation ρ



Copulas (2/2)

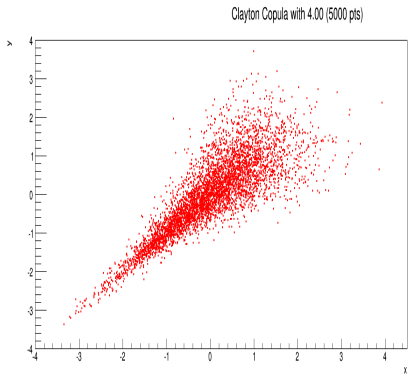
- two mainly family copulas:
  1. **Elliptical copula**

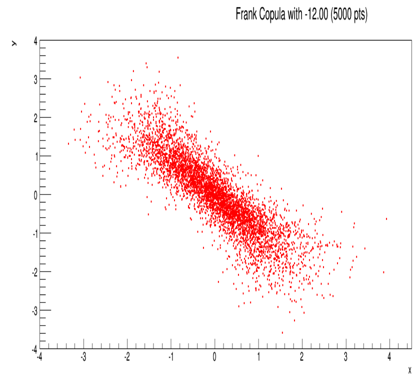
     the Elliptical copula is absolutely continous and can realize any correlation values $\rho \in [-1, 1]$

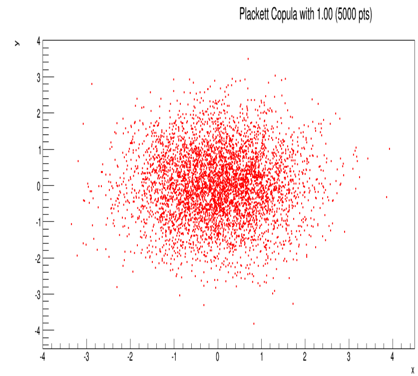
     Normal Copulas are Elliptical copula
  2. **Archimedian copula**

     Copulas are generated from a function $\varphi : (0, 1] \to [0, +\infty)$, called **generator**, which is convex, strictly decreasing with a positive derivate such as $\varphi(0) = 1$

$$C(u, v) \quad = \quad \varphi^{-1}(\varphi(u) + \varphi(v)) \qquad u, v \in [0, 1]^2$$







Pearson's Correlation Test of Independence

Assume that $(X, Y) \sim \mathcal{N}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ follows a bivariate normal distribution

- Hypothesis H_0 : test *"X and Y independent"*, $H_0 \ : \ \rho = 0$.
- Hypothesis H_1 : against *"it exists relation between X and Y"*, $H_1 \ : \ \rho \neq 0$.
- Test statistic t : we compare the test statistic

$$t = \frac{r\sqrt{(nS-2)}}{\sqrt{1-r^2}}$$

to the **Student** t distribution with $(nS-2)$ degrees of freedom with

$$r = \frac{S_{XY}}{S_X \, S_Y} = \frac{\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}}{\sqrt{(\sum X_i^2 - \frac{(\sum X_i)^2}{n})(\sum Y_i^2 - \frac{(\sum Y_i)^2}{n})}}$$

- Choose the risk α : Compute or look for in a table the quantile q_α for $\mathbf{t}$ $(nS-2)$
- Rule of the test :
  - if $|\hat{t}| > q_\alpha$ reject the hypothesis H_0 (***then*** *it exists a relation between X and Y*)
  - else accept H_0 (***then*** *X and Y are independents*)

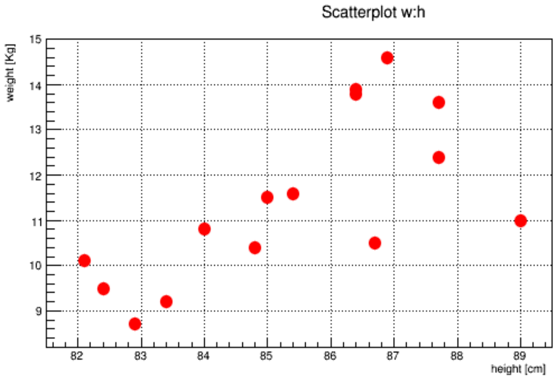
Pearson's Correlation Test of Independence - Example

15-sample (X_i, Y_i) for the height (cm) and the weight (kg) for children two years old:

X : Height (cm)	82.9	83.4	82.4	82.1	84.8	86.7	84.	89.	85.	85.4	87.7	87.7	86.4	86.4	86.9
Y : Weight (kg)	8.7	9.2	9.5	10.1	10.4	10.5	10.8	11.	11.5	11.6	12.4	13.6	13.8	13.9	14.6

- $nS = 15$
- $\hat{r} = 0.6786$

  Then $\hat{t} = \frac{0.6786 * \sqrt{15-2}}{\sqrt{1-0.6786^2}} = 3.33067$



- For $\alpha = 5\%$, $\quad t_{5\%}(13) = 2.16$

  so $\hat{t} > t_{5\%}(13)$ then we reject the hypotheses H_0 : it exists a relation between X and Y at the significance level 5%

- Even for $\alpha = 1\%$, $\quad t_{1\%}(13) = 3.012$

  so $\hat{t} > t_{1\%}(13)$ then we reject the hypotheses H_0 : it exists a relation between X and Y at the significance level 1%

Rank Spearman Correlation Test of Independence

No hypothesis about the bivariate distribution of (X, Y)

- Hypothesis H_0 : test "*X and Y independent*", $H_0 \; : \; r_s = 0$.
- Hypothesis H_1 : against "*it exists relation between X and Y*", $H_1 \; : \; r_s \neq 0$.
- Test statistic t : we compare the test statistic with the order statistic $X_{(i)}$

$$t \quad = \quad \frac{r_s \sqrt{(nS-2)}}{\sqrt{1-r_s^2}}$$

to the **Student** t distribution with $(nS-2)$ degrees of freedom with

$$r_s \quad = \quad 1. \; - \; \frac{6 \sum (x_{(i)} - y_{(i)})^2}{nS(nS^2-1)}$$

as $\sum_{i=1}^{n} X_{(i)} = \frac{nS(nS+1)}{2}$

- Choose the risk α : Compute or look for in a table the quantile q_α for **t** $(nS-2)$
- Rule of the test :
  - if $|\hat{t}| > q_\alpha$ reject the hypothesis H_0 (***then*** *it exists a relation between X and Y*)
  - else accept H_0 (***then*** *X and Y are independents*)

Rank Spearman Test of Independence - Example

15-sample (X_i, Y_i) for the height (cm) and the weight (kg) for children two years old:

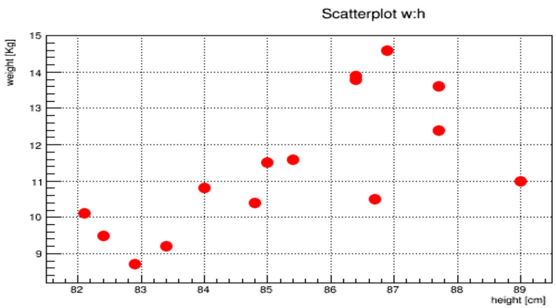
X : Height (cm)	82.9	83.4	82.4	82.1	84.8	86.7	84.	89.	85.	85.4	87.7	87.7	86.4	86.4	86.9
$X_{(i)}$	3	4	2	1	6	11	5	15	7	8	13.5	13.5	9.5	9.5	12
$Y_{(i)}$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Y : Weight (kg)	8.7	9.2	9.5	10.1	10.4	10.5	10.8	11.	11.5	11.6	12.4	13.6	13.8	13.9	14.6

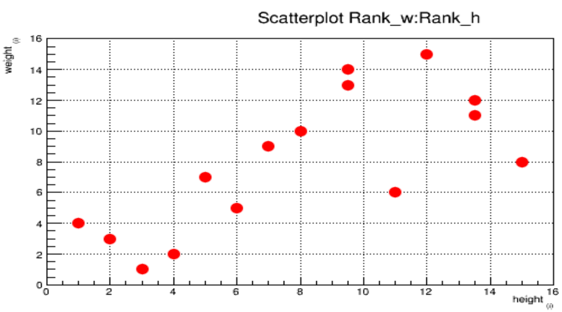
- $nS = 15$
- $\hat{r_s} = 0.72$

  Then $\hat{t} = \frac{0.72*\sqrt{15-2}}{\sqrt{1-0.72^2}} = 3.79$
- For $\alpha = 5\%$, $t_{5\%}(13) = 2.16$

  so $\hat{t} > t_{5\%}(13)$ then we reject the hypotheses H_0 : it exists a relation between X and Y at the significance level 5%
- Even for $\alpha = 1\%$, $t_{1\%}(13) = 3.012$

  so $\hat{t} > t_{1\%}(13)$ then we reject the hypotheses H_0 : it exists a relation between X and Y at the significance level 1%
- Results similary with the Pearson Test





Data Modelisation with PDF

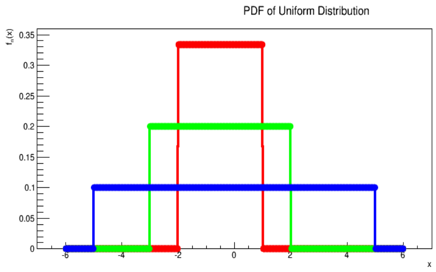
- With big dataset : Fitting the degree of freedom ("parameter")
  - Parametric methods when the family of the PDF is known
  - Else Nonparametric methods
- With small dataset : (not treated in the training session)
  - Expert judgement
  - Bayesian methods
  - Bootstrap methods (resampling)
- Inverses methods : (not treated in the training session)

1. Uniform Distribution

- The values in the interval $[a,b]$ are equally probable
- 2 parameters a (*"Minimum"*) and b (*"Maximum"*)

$$f(x) = \frac{1}{b-a}\,\mathbb{1}_{[a,b]}(x)$$

- Mean : $\mu = \frac{b-a}{2}$ (*Median*)
- Mode : any value in $[a,b]$
- Variance : $\sigma^2 = \frac{(b-a)^2}{12}$

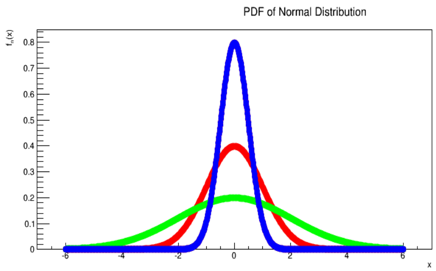


2. Normal Distribution

- 2 parameters μ (*"Mean"*) and σ (*"Standard-Deviation"*)

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

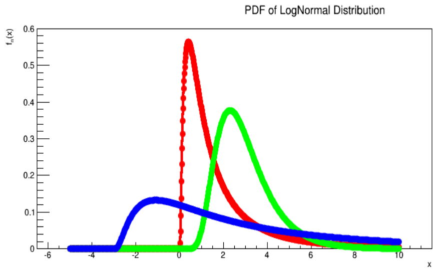
- Mean : μ (*Mode, Median*)
- Variance : σ^2



Commonly Used PDF

3. **LogNormal Distribution**

- A **positive** random variable x is said to follow a *LogNormal* law when $\ln x \sim \mathcal{N}$
- 3 parameters x_0 (lower bound) and (μ, σ) when $ln(X) \sim \mathcal{N}(\mu, \sigma)$



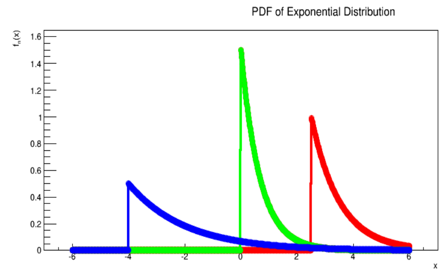
$$f(x) = \frac{1}{(x - x_0)\sigma\sqrt{2\pi}} \exp^{\frac{-(\ln(x - x_0) - \mu)^2}{2\sigma^2}} \quad \forall x > x_0$$

- Mean : $\mu_X = \exp^{(\mu + \frac{\sigma^2}{2})}$
- Median : $\exp^{(\mu)}$
- Mode : $\exp^{(\mu - \sigma^2)}$
- Variance : $\mu^2 \times (\exp^{\sigma^2} - 1.)$
- Another representation without (μ, σ) : μ_X and *"Error Factor"* with $Ef = \frac{q_{0.95}}{q_{0.50}}$
  - ⋆ $\sigma = \frac{ln(Ef)}{1.645} \iff Ef = \exp(1.645\sigma)$
  - ⋆ $\mu = ln(\mu_X) - \frac{\sigma^2}{2} \iff \mu_X = \exp^{(\mu + \frac{\sigma^2}{2})}$

4. **Exponential Distribution**
   - 2 parameters λ (shape) and x_0 (bound)

$$f(x) = \lambda \exp^{-\lambda(x-x_0)} \quad \forall x > x_0$$

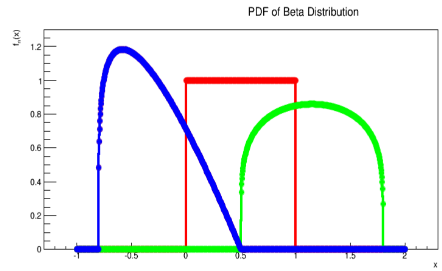
   - Mean : $x_0 + \frac{1}{\lambda}$
   - Mode : x_0
   - Variance : $1/\lambda^2$
   - $\widehat{\lambda_{\mathrm{MLE}}} = 1.0/(\bar{x} - x_0)$



5. **Beta Distribution**
   - 4 parameters α, β (shapes) & $x_0 < x_1$ (bounds)

$$f(x) = \frac{u^{\alpha-1} * (1-u)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} \quad \forall x \in [x_0, x_1]$$

   with $u = \frac{x-x_0}{x_1-x_0}$
   - Another notation (r,t) : $r = \alpha$, $t = \alpha + \beta$
   - Mean : $x_0 + (x_1 - x_0)\frac{\alpha}{\alpha+\beta}$
   - Mode : depends on (α, β)
   - Variance : $(x_1 - x_0)^2 \frac{\alpha\beta}{\alpha+\beta+1}$



The main PDF

Continuous			Discrete
Bounded	**positive**	**Umbounded**	Binomial
Uniform	Exponential	Normal	Multinomial
Beta	LogNormal	Cauchy	Poisson
Triangular	Weibull	Gumbel	...
Trapezium	Gamma	...	
Uniform by parts	Khi-two		
LogUniform	Pareto		
LogTriangular	...		
...			

Parametric Estimation

- Let $(x_1, x_2, \cdots, x_n)$ an *i.i.d* sample of a PDF $f(x, \theta)$ where $\theta \in \Theta$ is a vector of parameters for this family
- The true value of the parameter $\theta^\star$, which the data come from, is unknown
- Build an estimator $\hat{\theta}$ which would be as close to the true value $\theta^\star$ as possible

The two mainly methods are:

1. Maximum Likelihood (*MLE*)

   The method of maximum likelihood selects the set of values of the model parameters that maximizes the *likelihood* function. This function measures the *"agreement"* of the selected model with the observed data.

2. Moments Method (*MM*)

   - One starts with deriving equations that relate the population moments to the parameters θ
   - The moments are estimated from the given sample
   - The equations are then solved for the parameters θ, using the sample moments in place of the (unknown) population moments

Maximum Likelihood (MLE)

Build an estimator $\hat{\theta}$ for the model's parameters of the $f(x,\theta)$ from the data $(x_i)_{1\leq i\leq n}$

We use the ***Likelihood*** function $\mathcal{L}(\theta; x_1,\cdots,x_n)$:

$$\mathcal{L}(\theta; x_1,\cdots,x_n) = f(x_1,\cdots,x_n|\theta) \quad = \quad \prod_{i=1}^{n} f(x_i|\theta)$$

In practice it is often more convenient to work with the average of the logarithm of the likelihood function, called the ***average log-likelihood***:

$$\ln(\mathcal{L}(\theta; x_1,\cdots,x_n)) \quad = \quad \sum_{i=1}^{n} \ln(f(x_i|\theta))$$

or the average log-likelihood:

$$\hat{l}(\theta; x_1,\cdots,x_n) \quad = \quad \frac{1}{n}\ln(\mathcal{L}(\theta; x_1,\cdots,x_n))$$

MLE estimtaes $\hat{\theta}_{\mathrm{MLE}}$ by finding the value of θ that maximizes the $\hat{l}$ function

$$\hat{\theta}_{MLE} \quad = \quad \arg\max_{\theta\in\Theta} \quad \hat{l}(\theta; x_1,\cdots,x_n) \quad \text{... if any maximum exists}$$

Maximum Likelihood - Application

We have an *i.i.d* sample $(x_1, \cdots, x_n)$ from a normal law $\mathcal{N}(\mu, \sigma)$ where $\theta = (\mu, \sigma)$ unknown. The density is :

$$f(x|\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The Likelihood is

$$\mathcal{L}(\theta; x_1, \cdots, x_n) = \prod_{i=1}^{n} f(x_i|\theta) = (\frac{1}{2\pi\sigma^2})^{n/2} \exp^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}$$

The $\hat{l}$ function (*Average Log-Likelihood*) is:

$$\hat{l}(\theta; x_1, \cdots, x_n) = -\frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{1}{2n\sigma^2}\sum(x_i - \bar{x})^2 - \frac{1}{2\sigma^2}(\bar{x} - \mu)^2$$

- MLE for the mean μ parameter : $\frac{\partial \hat{l}}{\partial \mu} = -(\bar{x} - \mu)/\sigma^2 \rightarrow = 0$

$$\widehat{\mu}_{\text{MLE}} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- MLE for the variance σ^2 parameter : $\frac{\partial \hat{l}}{\partial \sigma} = -\frac{1}{\sigma} + \frac{\sum(x_i-\mu)^2}{n\sigma^3} \rightarrow = 0$

$$\widehat{\sigma^2}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_{\text{MLE}})^2$$

Moments Method (MM)

Build an estimator $\hat{\theta}$ for the model's parameters of the $f(x,\theta)$ from the data $(x_i)_{1\leq i\leq n}$

Suppose the first k moments of the true PDF can be expressed as functions of θ:

$$\begin{aligned}
\mu_1 = \mathbb{E}[X] \; &= g_1(\theta_1, \theta_2, \cdots, \theta_k) \\
\mu_2 = \mathbb{E}[X]^2 &= g_2(\theta_1, \theta_2, \cdots, \theta_k)... \\
\mu_k = \mathbb{E}[X]^k &= g_k(\theta_1, \theta_2, \cdots, \theta_k)
\end{aligned}$$

We compute the same first k moments from the sample $(x_i)_{1\leq i\leq n}$

$$\widehat{\mu_j} \quad = \quad \frac{1}{n}\sum_{i=1}^{n} x_i^j$$

The moments method estimator for (θ_j) denoted by $\hat{\theta}_{\mathrm{MM}}$ is defined as the solution (if there is one) to the system of equations:

$$\begin{aligned}
\widehat{\mu_1} &= g_1(\widehat{\theta_1}, \widehat{\theta_2}, \cdots, \widehat{\theta_k}) \\
\widehat{\mu_2} &= g_2(\widehat{\theta_1}, \widehat{\theta_2}, \cdots, \widehat{\theta_k})... \\
\widehat{\mu_k} &= g_k(\widehat{\theta_1}, \widehat{\theta_2}, \cdots, \widehat{\theta_k})
\end{aligned}$$

Moments Method (MM)

- The moments method is fairly simple and yields consistent estimators (under very weak assumptions), though these estimators are often biased

- Estimates by the moments method may be used as the first approximation to the solutions of the likelihood equations, and successive improved approximations may then be found by the Newton Raphson method. In this way the moments method and the method of maximum likelihood are symbiotic

- In some cases, as in the example of the gamma distribution, the likelihood equations may be intractable without computers, whereas the moments method estimators can be quickly and easily calculated by hand

Moments Method - Applications

- Case of the **normal distribution**

  We have an *i.i.d* sample $(x_1, \cdots, x_n)$ from a normal law $\mathcal{N}(\mu, \sigma)$ where $\theta = (\mu, \sigma)$ unknown.

  - $\mu = \mu_1 = 1/N \sum x_i$
  - $\mathbb{E}[X^2] = \mu_2 = Var[X] + \mathbb{E}[X]^2 = \sigma^2 + \mu_1^2$

$$\hat{\sigma^2} = 1/n \sum (x_i - \mu_1)^2$$

  Then MLE $\Longleftrightarrow$ MM in the gaussian case

- Case of the **beta distribution**

$$\mathbb{E}[X] = \frac{\alpha}{\alpha + \beta}$$

$$\mathbb{E}[X^2] = \frac{\alpha + 1}{\alpha + \beta + 1} \mathbb{E}[X]$$

$$\widehat{\mu}_1 = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{and} \quad \widehat{\mu}_2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2$$

Then, the moments method gives us :

$$\mathbb{E}[X] = \frac{\hat{\alpha}}{\hat{\alpha} + \hat{\beta}} = \widehat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and

$$\mathbb{E}[X^2] = \frac{\hat{\alpha} + 1}{\hat{\alpha} + \hat{\beta} + 1}\mathbb{E}[X] = \widehat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

We obtain

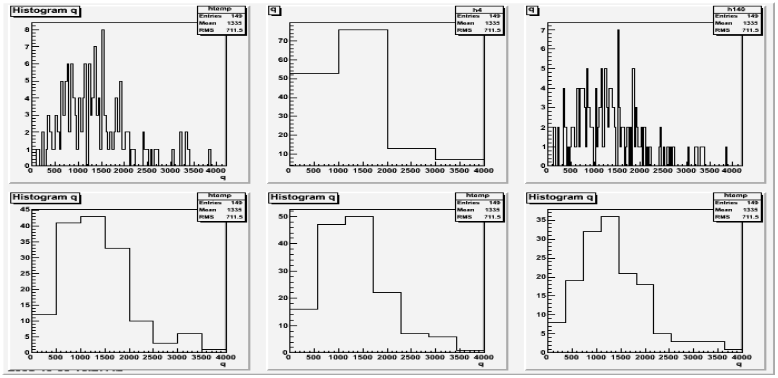
$$\hat{\alpha} = \widehat{\mu}_1 \frac{\widehat{\mu}_2 - \widehat{\mu}_1}{\widehat{\mu_1}^2 - \widehat{\mu}_2}$$

$$\hat{\beta} = \hat{\alpha}\,\frac{1 - \widehat{\mu}_1}{\widehat{\mu}_1} = (1 - \widehat{\mu}_1)\frac{\widehat{\mu}_2 - \widehat{\mu}_1}{\widehat{\mu_1}^2 - \widehat{\mu}_2}$$

Histogram

- The histograms are classical density estimation
- The followings steps are needed to build the histogram:
  - Arrange the sample in increasing order;
  - Subdivide the range of the sample into several equal intervals, and count the number of observations in each intervals;
  - plot the number of observations in each interval versus the random variable
- but the form depends on the number of bins

1. **Sturges** $N_{bin} = log_2(n) + 1$
2. **Scott** $N_{bin} = (x_{max} - x_{min}) * \sqrt[3]{n}/3.5\hat{\sigma}_x$
3. **Freedman & Diaconis** $N_{bin} = (x_{max} - x_{min}) * \sqrt[3]{n}/2 * (Q_x^{0.75} - Q_x^{0.25})$



Kernel Methods

From the point of view of the histogramm,

$$f(x) = F'(x) \simeq \frac{F(x+h) - F(x-h)}{2 \times h} \quad \forall h > 0 \,,\ h \text{ "small"}$$

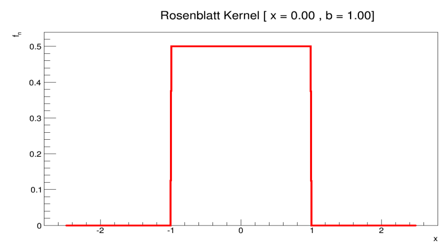
Then ***Rosenblatt*** (1956) suggests the estimator :

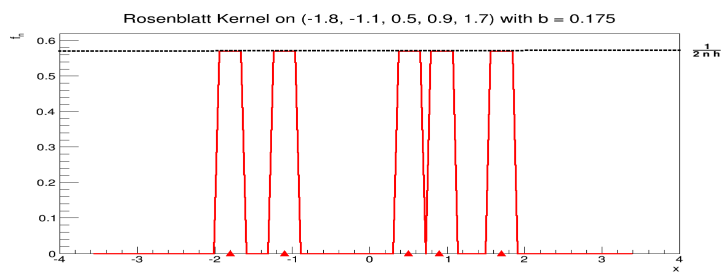
$$\hat{f}_{n,h}(x) = \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2 \times h}$$

which has another representation ***Parzen*** (1962)

$$\hat{f}_{n,h}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - x_i}{h}\right)$$

$$\text{with} \quad K(u) = \frac{1}{2} \times \mathbb{1}_{[-1.,1.]}(u)$$





Kernel Estimators - Definitions

- A function $K : \mathbb{R} \to \mathbb{R}$ is said a **Kernel** if

$$\int K(u)\ \mathrm{du} \quad = \quad 1.$$

- Often, but not necessarily,
  - K is symmetric around the origin: $K(-u) = K(u) \quad \forall u$
  - K is positive: $K(u) > 0 \quad \forall u$

- $\forall h > 0$,

$$\hat{f}_{n,h}(x) \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\frac{X_i - x}{h})$$

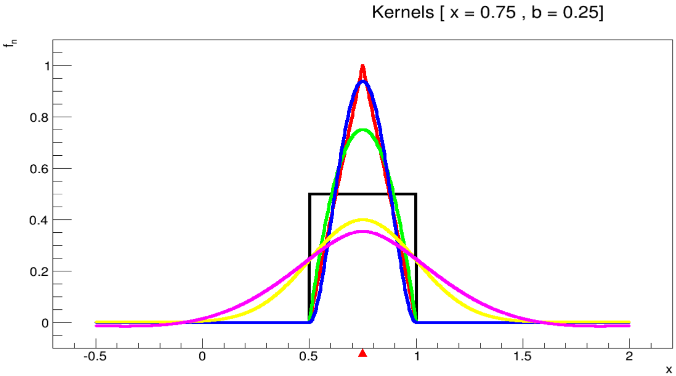
  is a ***kernel estimator*** of the density f $\quad (\int \hat{f}_{n,h}(x)\ \mathrm{d}x = 1)$

- Kernel approach is a histogram which, for estimating the density of $f(x)$, has been shifted so that x, say, lies at the center of a mesh interval. And For evaluating the density at another point, say y, the mesh is shifted again, so that y is at the center of a mesh interval.

- The parameter h is a *smoothing* parameter called ***bandwidth***; More greater h is, more the estimation $\hat{f}_{n,h}$ is smooth.

Kernel Estimators - Exemples

- *Rectangular* (***Rosenblatt***) (*black*) $$K(u) = \frac{1}{2} \times 1\!\!1_{[-1.,1.]}(u)$$
- *Triangular* (*red*) $$K(u) = (1 - |u|) \times 1\!\!1_{[-1.,1.]}(u)$$
- ***Epanechnikov*** (*blue*) $$K(u) = \frac{3}{4}(1 - x^2) \times 1\!\!1_{[-1.,1.]}(u)$$
- *Biweight* (*green*) $$K(u) = \frac{15}{16}(1 - x^2)^2 \times 1\!\!1_{[-1.,1.]}(u)$$
- *Gaussian* (*yellow*) $$K(u) = \frac{\exp^{-x^2/2}}{\sqrt{2\pi}}$$
- ***Silverman*** (*magenta*) $$K(u) = \frac{1}{2} \exp^{-|u|/\sqrt{2}} \sin\left(|u|/\sqrt{2} + \pi/4\right)$$



Kernel Methods - Application

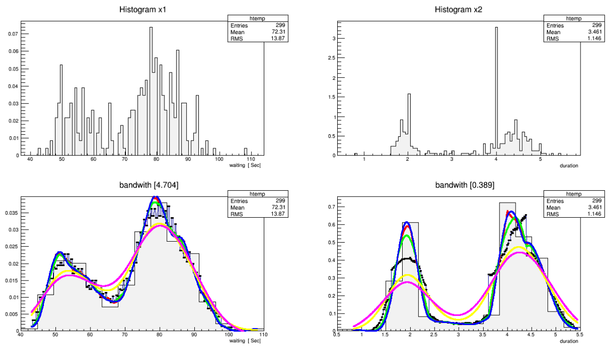
- Optimal bandwidth with the Silverman Rule (1996)

$$h_n = 1.364 \times \alpha_K \times \textsc{Min}\{\hat{\sigma}, \frac{\text{IQR}}{1.349}\} \times \mathrm{n}^{-1/5}$$

with

1. $\hat{\sigma}$ is the sample standard deviation
2. IQR is the "*I*nter*Q*uartile *R*ange" ($\text{IQR} = q_{0.75} - q_{0.25}$)
3. α_K is a constant that only depends on the used kernel

Kernel	$K(x)$	α_K
Rectangular	$1/2 \;,\; \lvert x\rvert < 1$	1.3510
Triangular	$1 - \lvert x\rvert \;,\; \lvert x\rvert < 1$	1.8882
Epanechnikov	$\frac{3}{4}(1 - x^2) \;,\; \lvert x\rvert < 1$	1.7188
Biweight	$\frac{15}{16}(1 - x^2)^2 \;,\; \lvert x\rvert < 1$	2.0362
Gaussian	$\frac{\exp^{-x^2/2}}{\sqrt{2\pi}}$	0.7764



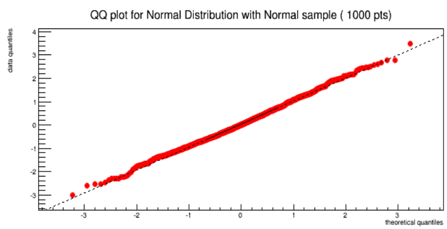
Geyser database for Gaussian Kernel (*left*) waiting b = 4.70, (*right*) duration b = 0.39

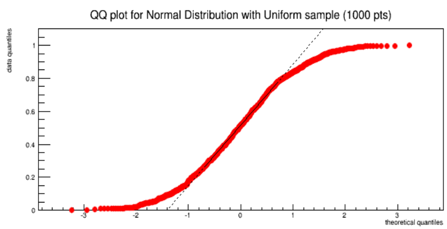
Goodness-of-Fits techniques

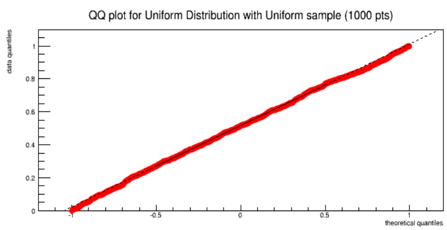
- Graphical methods
  - QQPlot
- Statistical Tests
  - Chi-Squared
  - Tests based on EDF Statistics
    - Kolmogorov-Smirnov
    - Cramer-von Misses
    - Anderson-Darling

Graphical method : QQ-plot

- a **QQ-plot** (*"Q" stands for quantile*) is a probability plot to compare two probability distributions by plotting their quantiles against each other
- A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate).
- If the two distributions being compared are similar, the points in the QQ-plot will approximately lie on the line $y = x$
- If the distributions are linearly related, the points in the QQ-plot will approximately lie on a line, but not necessarily on the line y = x.
- Select one axe for the theoretical distribution for Goodness-of-Fit test







Statistical Tests

Hypothesis testing is the use of statistics to determine the probability that a given hypothesis is true. The usual process of hypothesis testing consists of four steps :

1. Formulate the **null hypothesis** H_0 (*e.g.* two population means are equal) and the **alternative hypothesis** H_1 (*e.g.* two population means are not equal)

2. Identify a **test statistic** that can be used to assess the truth of the null hypothesis

3. Select a **Significance Level** α which defines the sensitivity of the test *(type I error)*
   In practice, the common values of α are 0.1, 0.05 or 0.01

4. Compute the **P-value**
   which is the probability that a test statistic at least as significant as the one observed would be obtained assuming that the null hypothesis were true.
   And compare the P-value to α
   If $p \leq \alpha$, that the observed effect is statistically significant, the null hypothesis is ruled out, and the alternative hypothesis is valid

   The smaller the P-value, the stronger the evidence against the null hypothesis.

	Reject H_0	Don't reject H_0
reality H_0	α	$1-\alpha$
reality H_1	$1-\beta$	β

Commonly Used Statistical Tests

In Goodness-of-Fit work, the commonly used statistical tests are:

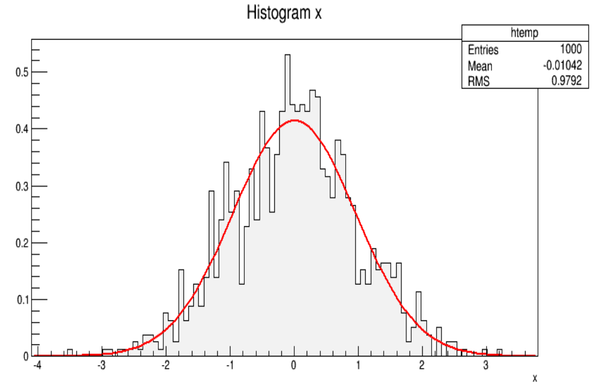
- Chi-Squared (χ^2)
- Tests based on EDF Statistics
  - Kolmogorov-Smirnov ($\mathbf{D}$)
  - Cramer-von Mises (W^2)
  - Anderson-Darling (A^2)

The Chi-Squared χ^2 Test

- The χ^2 test is used to test if a sample (x_i) came from a specific distribution
- Useful when data are discrete, and applied to continous distribution with a large number of observations
- The basic idea is to partitioned the range of the sample into k cells, and compare the observed frequency O_i with the expected frequency E_i in each cell i
- The statistic test is:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

which follows a χ^2 distribution with $(k - 1 - t)$ degrees of freedom, where t is the number of parameters of the distribution to estimate

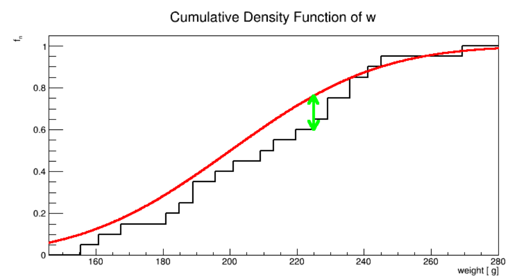


- The ratio n/k must verify $n/k \geq 5$
- The value of the $\chi 2$ test statistic are dependent on how the data is binned
- χ^2 test is generally less powerful than *EDF* tests

The tests based on EDF Statistics (1/2)

- Graphical methods have a wide appeal in deciding if a random sample appears to come from a given PDF
- We consider now tests of fit based on the *Empirical Distribution Function* ("*EDF*")
- *EDF* statistics are measures of the discrepancy between the empirical CDF and the theorical CDF of the PDF
- They are based on the vertical differences between $F_n(x)$ and $F(x)$, and divided into two classes :
  1. **the supremum statistics** : select the largest vertical difference between the two CDF; it is the **Kolmogorov-Smirnov** test D

$$D = \sup_x |F_n(x) - F(x)|$$



  2. **the quadratic statistics** : measure of discrepancy given by the Cramer-von Mises family

$$Q = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 \psi(x) \mathrm{d}x$$

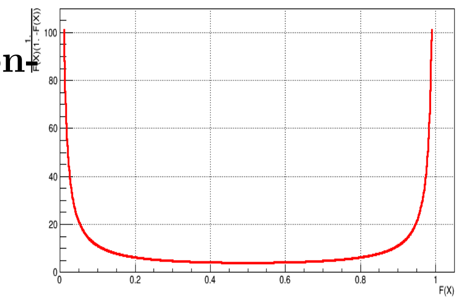
where ψ is a *weight* function

– For $\psi(x) = 1$ we obtain the **Cramer-von Mises** Tests, denoted as W^2:

$$W^2 = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 \mathrm{d}x$$

– For $\psi(x) = \frac{1.}{F(x)(1.0-F(x))}$ we obtain the **Anderson-Darling** test, denoted A^2:

$$A^2 = n \int_{-\infty}^{+\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1.0 - F(x))} \mathrm{d}x$$



- To compute these statistics, we use the *Probability Integral Transformation* (*"PIT"*)
  – Let $X \sim F$ with F is the true CDF
  – If $Z = F(X)$, then $Z \sim \mathcal{U}[0., 1.]$
  – For The sample $(x_1, x_2, \cdots, x_n)$, compute $z_i = F(x_i)$ and compare the empirical CDF of the z_i with the CDF of the uniform distribution

$$F^\star(z) = z \quad , \quad 0 \leq z \leq 1$$

  – EDF statistics computed from the EDF of the z_i compared with the uniform distribution will take the same values as if they were computed from the EDF of the x_i compared with F

Comparaisons of the Goodness-of-Fit tests

- The χ^2 statistic is the lower powerfull for continous PDF
- EDF statistics are usually much more powerfull than the χ^2 statistic (where data must be grouped, then loss of informations)
- the D statistic is the most well-known of the EDF statistics, but it is often much less powerfull than the quadratic statistics W^2 and A^2
- A^2 and W^2 give often similarly values, but A^2 is on the whole more powerfull when the distribution F departs from the true distribution in the tails (weight function)
- In Goodness-of-Fit work, departure in the tails is often important to detect, so A^2 is the recommanded statistic

Conclusions

- Review of descriptive statistics and dependence between variables
- Present several methods to estimate the Probability Density Function (Parametric and Nonparametric)
- Verify, or sort out, the selected distribution(s) by Goodness-of-Fit methods

Bibliography

D'Agostini, *R.* and Stephens, *M.* (1986). *Goodness-of-Fit techniques*. Dekker, Statistics monographs, vol 68.

De Rocquigny, *E.*, Devictor, *N* and Tarantola, *S.* (2008). *Uncertainty in Industrial Practice: A Guide to Quantitative Uncertainty Management*. John Wiley and Sons.

Haldar, *E.* and Mahadevan, *S.* (2000). *Probability, Reliability and Statistical Methods in Engineering Design*. John Wiley and Sons.

Hörmann, *W.*, Leydold, *J.* and Derflinger, *G.* (2000). *Automatic Nonuniform Random Variate Generation*. Springer, Statistics and Computing.

Kurowicka, *D.* and Cooke, *R.* (2006). *Uncertainty Analysis with High Dimensional Dependence Modeling*. John Wiley and Sons.

Nelsen, *R.*, (2006). *An Introduction to Copulas*. Springer Series in Statistics.

Tapia, *R.* and Thompson, *J.* (1978). *Nonparametric Probability Density Function*. Johns Hopkins.