DE LA RECHERCHE À L'INDUSTRIE

www.cea.fr

Uncertainty quantification

A broad introduction

J-B. Blanchard, F. Gaudier
jean-baptiste.blanchard@cea.fr

HPC and Uncertainty Treatment
with Open TURNS and Uranie | 2019/05/27

Outline

Brief reminders

Descriptive statistics

- Univariate case
- Bivariate case

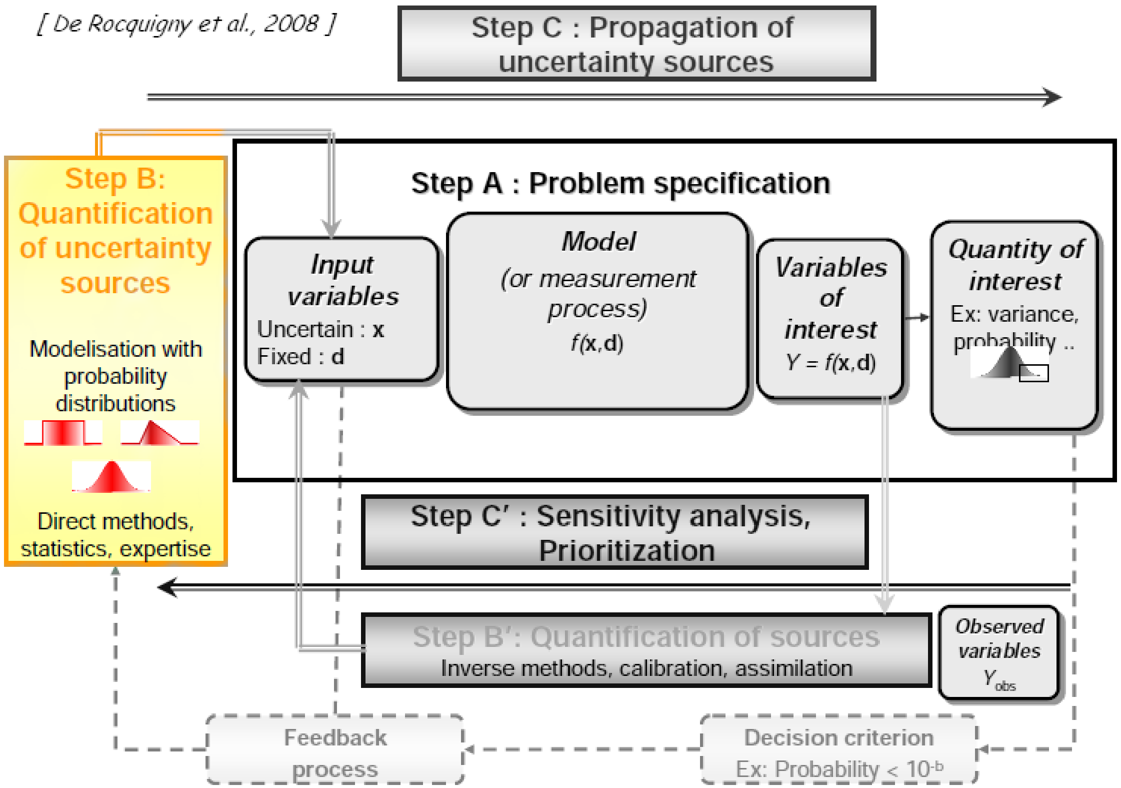
Data modelisation with PDF

- Commonly used PDFs
- Parametric PDFs
- Non-parametric PDFs

Goodness-of-fit techniques

- Graphical methods
- Statistical tests

Reminder: uncertainty quantification



Main steps:

- A: problem definition
  - Uncertain input variables
  - Variable/quantity of interest
  - Model construction
- B: uncertainty quantification
  - Choice of pdfs
  - Choice of correlations
- B': quantification of sources
  - Inverse methods using data to constrain input values and uncertainties
- C: uncertainty propagation
  - Evolution of output variability w.r.t input uncertainty
- C': sensitivity analysis
  - Uncertainty source sorting

These steps are usually model dependent, it might be useful to iterate to help converging to proper conclusions

Main distribution principle

For every random variable $X : \Omega \to \mathbb{R}$

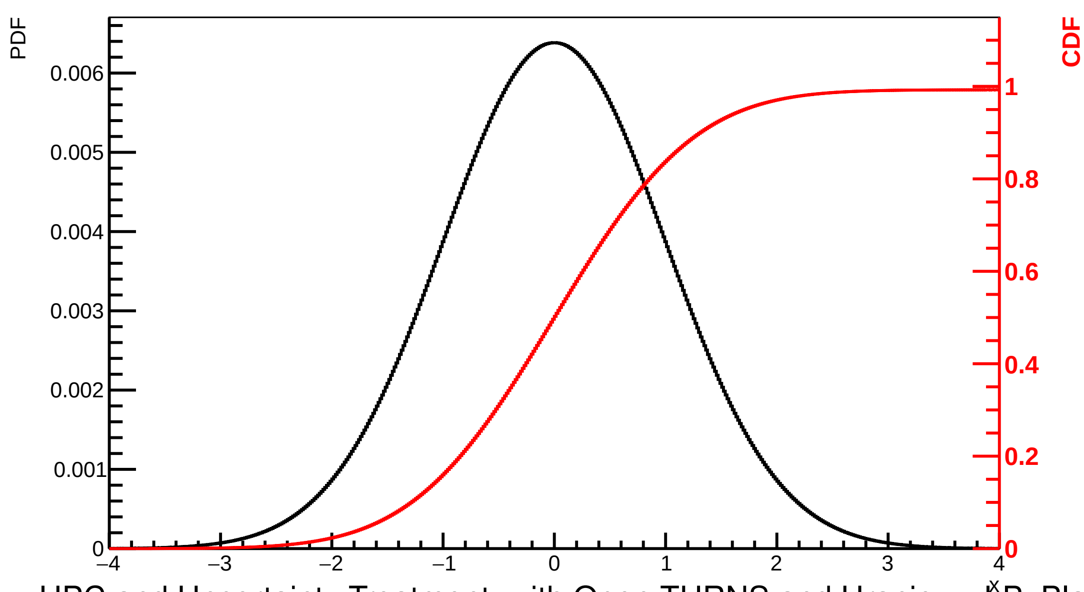
- **PDF** (Probability Density Function): if the random variable X has a density f_X, where f_X is a non-negative Lebesgue-integrable function, then

$$P\{a \leq X \leq b\} = \int_a^b f_X(s)ds$$

- **CDF** (Cumulative Distribution Function): the function $F_X : \mathbb{R} \to [0,1]$, given by

$$F_X(a) = \int_{-\infty}^{a} f_X(s)ds, \;\; a \in \mathbb{R}$$

➙ One might find **CCDF** for Complementary CDF, simply defined as $CCDF(a) = 1 - CDF(a)$



Descriptive statistics

Uni-variate case: “location” parameters

The effect of the "location" parameter is to translate the graph relative to the standard distribution

- **Mean μ**:

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

- **Mode M**: Value where the probability is the greatest value
- **α-Quantile q_{alpha} with $\alpha \in [0, 1]$**: defined as

$$\mathbb{P}[X \leq q_\alpha] = \alpha$$

- **Median $q_{0.5}$**: it is the 0.5-quantile defined as

$$\mathbb{P}[X \leq q_{0.5}] = 0.5 = \mathbb{P}[X \geq q_{0.5}]$$

- **Quartiles**: $q_{0.25}$, $q_{0.5}$, $q_{0.75}$
- **Extreme values** : Min and Max

Uni-variate case: “dispersion” parameters

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

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

$$\mathrm{Var}(X) = \frac{1}{n_S - 1} \sum_{i=1}^{n_S} (x_i - \mu)^2$$

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

$$\sigma = \sqrt{\mathrm{Var}(X)}$$

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

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

- **Range** R:

$$R = Max - Min$$

- **Interquartile interval H**:

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

Uni-variate case: “shape” parameters

Any parameter of a PDF that affect the shape of a distribution rather than simply shifting it or stretching/shrinking it.

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

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

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

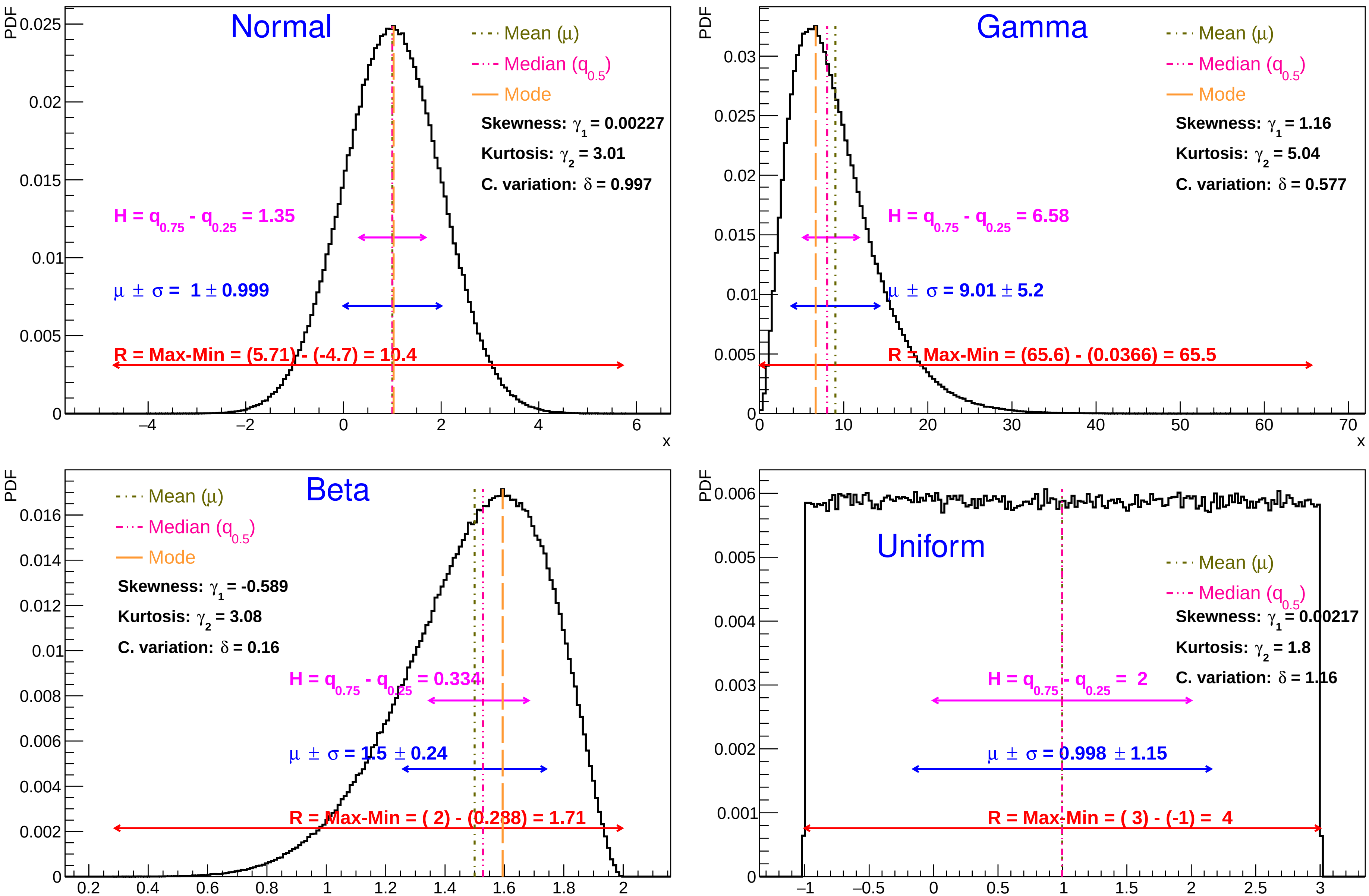
$$\gamma_1 = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\mathbb{E}(X^3) - 3\mu\sigma^2 - \mu^3}{\sigma^3}$$

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

$$\gamma_2 = \frac{\mu_4}{\sigma^4};$$

→ Normalised γ_2: sometimes -3.0 is added to it as γ_2=3.0 for $\mathcal{N}(\mu, \sigma)$

Uni-variate case: illustration of some parameters



Uni-variate case: usual graphical representation

Usual convention

- X is a random variable, whose realisation is noted x
- $\mathbf{x}$ is a vector of realisation of size n_S, x_i being its i-th element.

Many possible ways to represent data, among which:

- Histograms

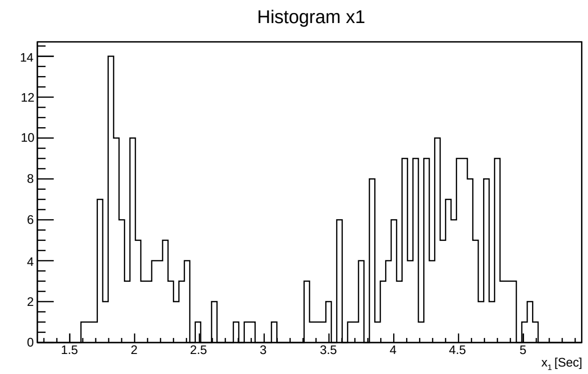
$$\forall a, b \in \mathbb{R}^2 \text{ with } a < b, H_{[a,b]}(\mathbf{x}) = \sum_{i=1}^{n_S} \mathbb{1}_{[a,b]}(x_i)$$

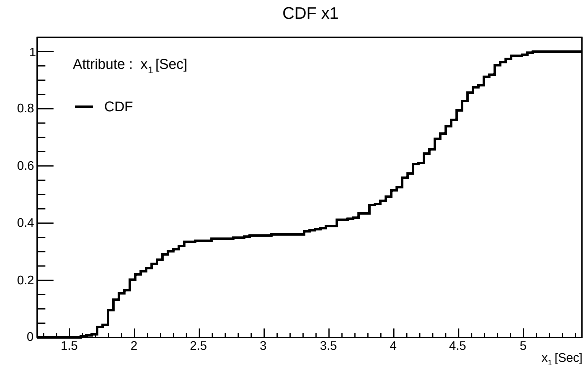
It can be normalised w.r.t total number of events, weights. . .

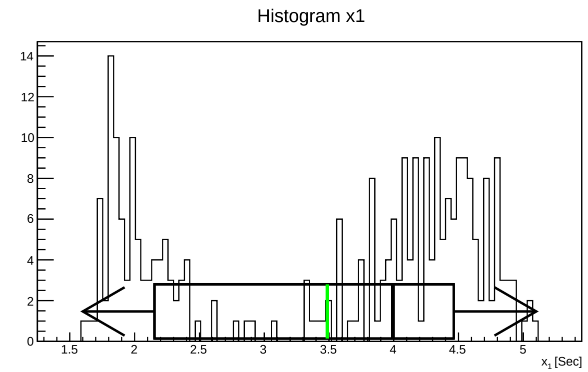
  - Sturges: $N_{\text{bin}} = \log_2(n_S) + 1$
  - Scott: $N_{\text{bin}} = (x_{\max} - x_{\min}) \times \sqrt[3]{n_S}/(3.5 \times \hat{\sigma}_x)$
  - Freedman & Diaconis: $N_{\text{bin}} = (x_{\max} - x_{\min}) \times \sqrt[3]{n_S}/2 \times (Q_x^{0.75} - Q_x^{0.25})$
- Empirical Cumulative Density Function (eCDF)

$$F_{n_S}(x) = \frac{1}{n_S} \sum_{i=1}^{n_S} \mathbb{1}(x_i \leq x)$$

- BoxPlot: Simple way to look at many information:
  - Minimum and maximum (arrows)
  - quartiles: 0.25, 0.5, 0.75 quantiles (black lines)
  - Mean: green line



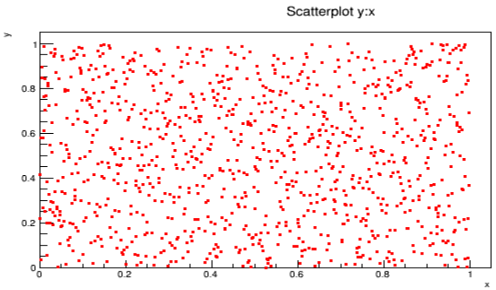




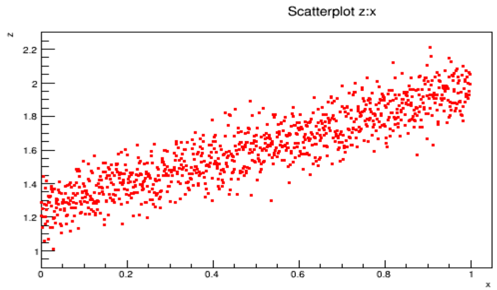
Bivariate case: graphics

Detect and describe statistical dependences between variables

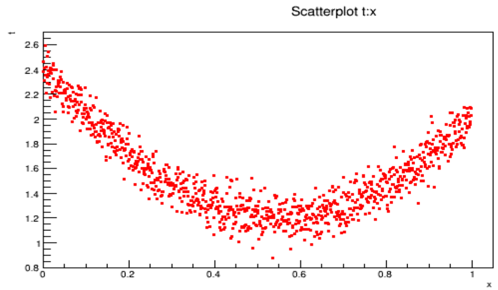
- independent variables $\Longrightarrow$ uncorrelated variables
- uncorrelated variables $\not\Longrightarrow$ independent variables



uncorrelated



linear correlation



nonlinear correlation

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

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

and the covariance estimated from a sample (x_i, y_i) is defined as

$$\hat{\mathrm{Cov}}(x, y) = \frac{1}{n_S} \sum_{i=1}^{n_S} (x_i - \bar{x})(y_i - \bar{y})$$

The sign of this coefficient is the tendency of the linear relationship between the variables, but the magnitude is not easy to interpret.

Pearson correlation coefficient

The Pearson coefficient, usually noted ρ or ρ_P, is a normalised version of the covariance: it 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,

$$\rho_P(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}.$$

It's estimation on a sample (x_i, y_i) can be written as $\hat{r}_P$:

$$\hat{r}_P = \frac{\sum_{i=1}^{n_S}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n_S}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n_S}(y_i - \bar{y})^2}}$$

Properties of this coefficient

- $\hat{r}_P \in [-1, 1]$
- $\hat{r}_P = \pm 1$ implies that a linear equation describes perfectly the relationship between X and Y, and the data points lying exactly on a positive (negative) identity line
- $\hat{r}_P = 0$, X and Y are said to be (linearly) uncorrelated (but not necessary independents !!)

Spearman correlation coefficient

The Spearman's Rank Correlation Coefficient, noted ρ_s, is a measure of the monotonic dependence between two variables X and Y.

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

$$\rho_S = \rho_P(R_X(X), R_Y(Y))$$

Rank definition

For a given sample (x_i), the n raw values x_i are converted to ranks values $x_{(i)}$:

- $x_{(i)} \in [1, 2, \ldots, n_S]$
- $x_{(1)} < x_{(2)} < \ldots < x_{(n_S)}$
- The mean of $(x_{(i)})$ is $\bar{x_{()}} = \dfrac{n_S + 1}{2}$

It's estimation on a sample (x_i, y_i) can then be written as $\hat{r}_S$:

$$\hat{r}_S = \frac{\sum_{i=1}^{n_S}(x_{(i)} - \bar{x_{()}})(y_{(i)} - \bar{y_{()}})}{\sqrt{\sum_{i=1}^{n_S}(x_{(i)} - \bar{x_{()}})^2}\sqrt{\sum_{i=1}^{n_S}(y_{(i)} - \bar{y_{()}})^2}}$$

Spearman correlation properties

- $\hat{r}_S \in [-1, 1]$
- $\hat{r}_S = \pm 1$ implies that a monotonic relationship between X and Y.
- $\hat{r}_S = 0$, X and Y are uncorrelated monotonically.

Digression: test-of-hypothesis

The aim of a test-of-hypothesis is to check the validity of a given hypothesis, providing a certain chosen confidence level $(1-\alpha)$.

Principle in few key steps

- Purpose: significance, goodness-of-fit, independence, conformity...

A factory build tubes whose lifetime $\sim \mathcal{N}(1200, 300)$. 100 tubes are produced with a new process $\bar{x} = 1265$. Is this significant ? Is the new μ_N greater than 1200 ?

- Hypothesis
  1. H_0 is the null-hypothesis (to be tested): $\mu_0 = 1200$
  2. H_1 is the alternative hypothesis: $\mu_1 > 1200$

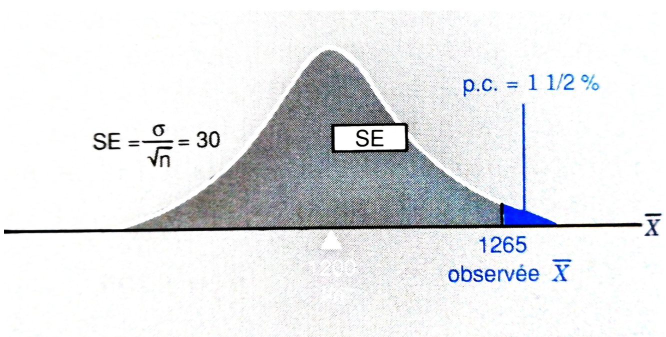
	H_0 accepted	H_0 rejected
H_0 true	Correct (1-α)	Type-I error (α)
H_0 false	Type-II error (β)	Correct $(1-\beta)$

- Confidence level: choose probability α

A usual choice is to set $\alpha = 0.05$, resulting in a 95% CL

- Statistical test to be computed

Here one can use classical test $\hat{Z} = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{N}} = \frac{1265-1200}{30} = 2.17$



- Result interpretation

Two possibilities:

1. *Look at table and see that $\hat{Z} > Z_{0.05}$ (=1.64) $\Rightarrow H_0$ rejected !*
2. *Look at table and see that $\hat{Z} = 2.17 \Leftrightarrow P_c = 0.015 \Rightarrow H_0$ rejected !*
3. *Look at table and see that to get $\hat{Z}_C = 1.64 \Leftrightarrow \bar{x}_c = 1249 \Rightarrow H_0$ rejected !*

Pearson's correlation test of independence (1/2)

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

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

- 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 (2/2)

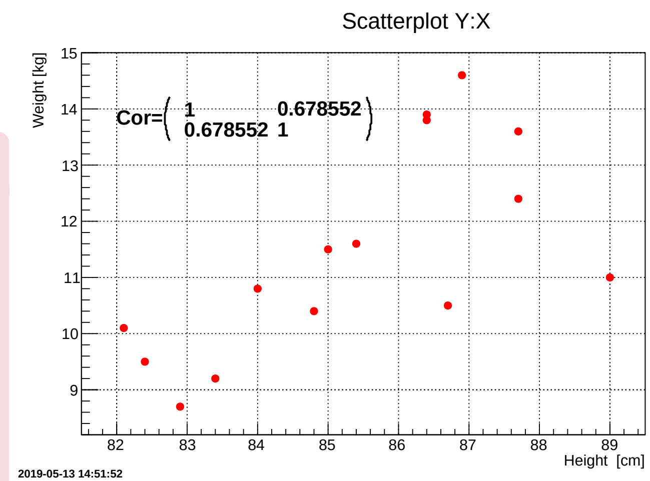
Using a 15-sample database showing weight and height for 2-year old children.

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

Setting the test

1. $n_S = 15 \Rightarrow$ degree of freedom = 13 ($n_S - 2$)
2. $\hat{r}_P = 0.6786$

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



Interpret these results

- For a chosen $\alpha = 0.05$, $t_{5\%}(13) = 2.16$
  $\Rightarrow \hat{t} > t_{5\%}(13)$
  $\Rightarrow H_0$ rejected !
  $\Rightarrow H_0$ It exists a relation between X and Y at 5% significance level
- For a chosen $\alpha = 0.01$, $t_{1\%}(13) = 3.012$
  $\Rightarrow \hat{t} > t_{1\%}(13)$
  $\Rightarrow H_0$ rejected !
  $\Rightarrow H_0$ It exists a relation between X and Y at 1% significance level
- Looking at table, $3.01 < \hat{t} < 3.37$
  $\Rightarrow$ Critical probability $0.005 < P_c < 0.01$

Spearman's correlation test of independence (1/2)

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

Spearman's correlation test of independence (2/2)

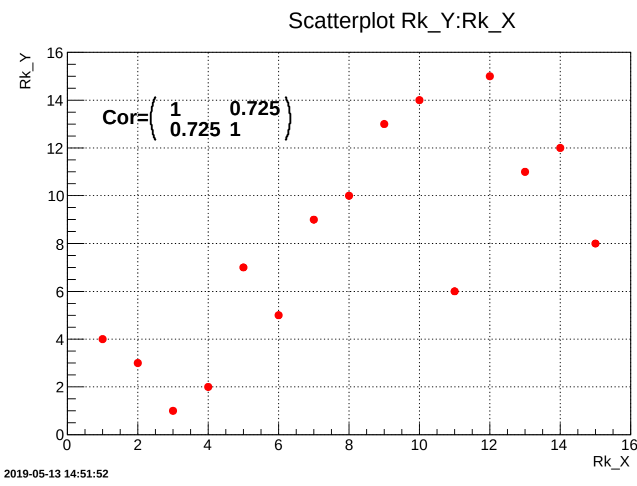
Using a 15-sample database showing weight and height for 2-year old children.

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
$X_{()}$	3	4	2	1	6	11	5	15	7	8	13.5	13.5	9.5	9.5	12
$Y_{()}$	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

Setting the test

1. $n_S = 15 \Rightarrow$ degree of freedom = 13 ($n_S - 2$)
2. $\hat{r}_S = 0.725$

$$\hat{t} = \frac{0.725 \times \sqrt{15-2}}{\sqrt{1-0.725^2}} = 3.796$$



Interpret these results

- For a chosen $\alpha = 0.05$, $t_{5\%}(13) = 2.16$
  $\Rightarrow \hat{t} > t_{5\%}(13)$
  $\Rightarrow H_0$ rejected !
  $\Rightarrow H_0$ It exists a relation between X and Y at 5% significance level
- For a chosen $\alpha = 0.01$, $t_{1\%}(13) = 3.012$
  $\Rightarrow \hat{t} > t_{1\%}(13)$
  $\Rightarrow H_0$ rejected !
  $\Rightarrow H_0$ It exists a relation between X and Y at 1% significance level
- Looking at table, $3.37 < \hat{t} < 3.85$
  $\Rightarrow$ Critical probability $0.002 < P_c < 0.005$

Data modelisation with PDF

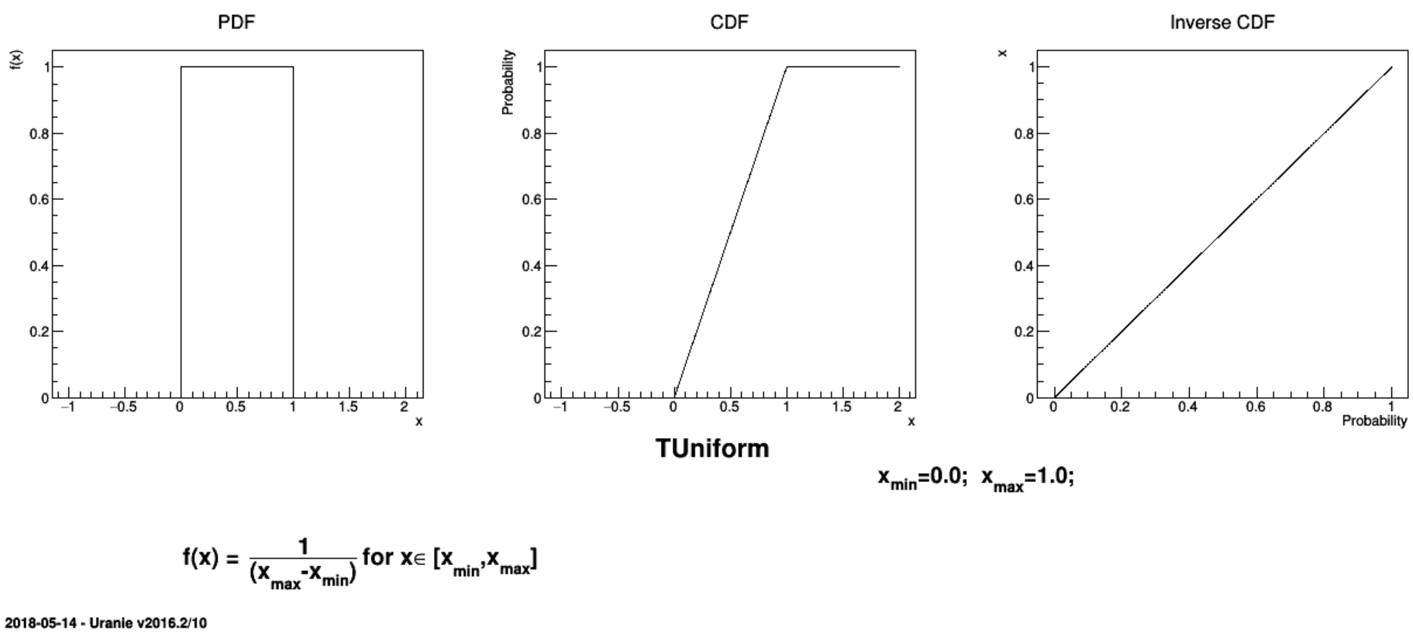
Commonly used PDF I

Uniform distribution

- Defined by 2 parameters: a (minimum) and b (maximum), $\forall a, b \in \mathbb{R}^2$ with $a < b$
- Values in $[a, b]$ are equally probable:

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

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



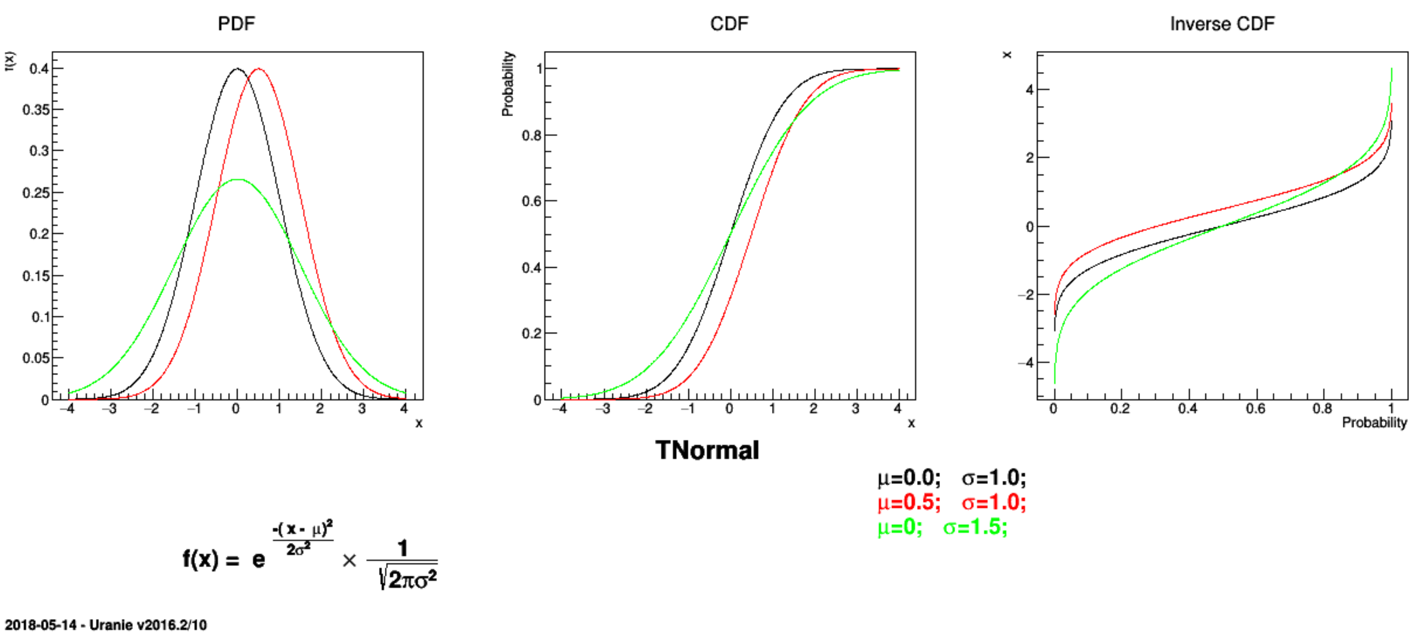
Commonly used PDF II

Normal distribution

- Defined by 2 parameters: μ (mean) and σ (standard deviation), $\forall \mu, \sigma \in \mathbb{R}^2$ with $\sigma > 0$
- Famous PDF shape:

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

- Mean, mode, median: μ



Commonly used PDF III

LogNormal distribution

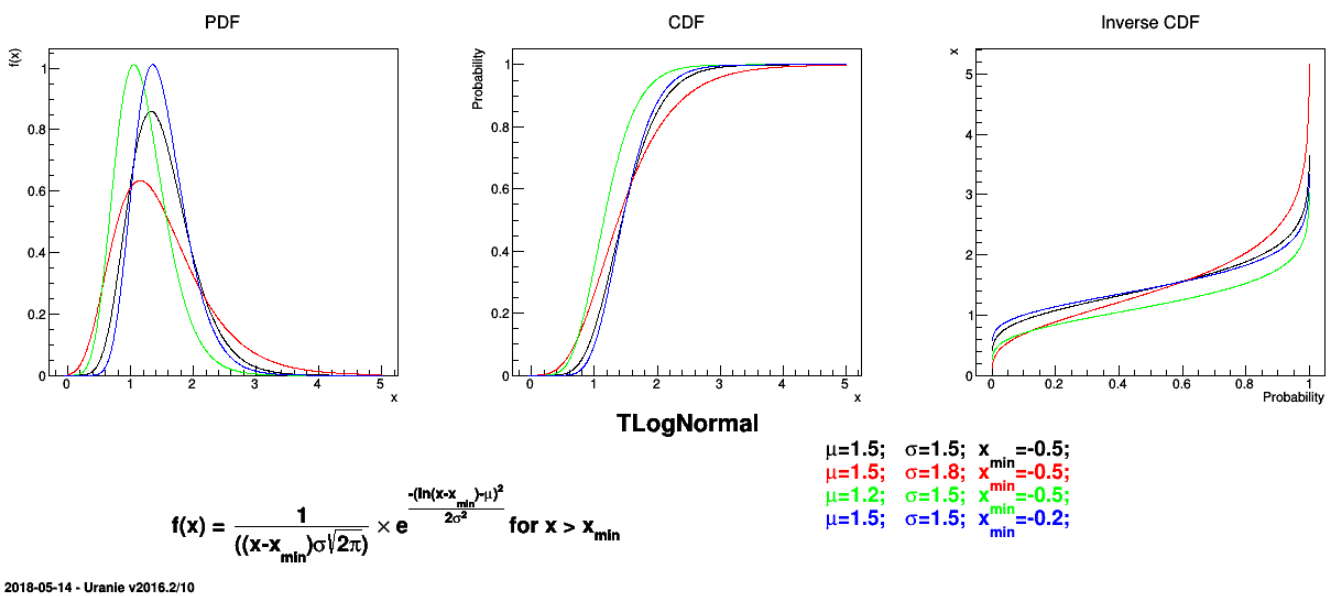
- A positive random variable x is said to follow a LogNormal law when $ln(x) \sim \mathcal{N}$
- Defined by 3 parameters: μ (mean) and σ (standard deviation) for $ln(x) \sim \mathcal{N}(\mu, \sigma)$, and x_0 (bound)

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

- Mean: $\mu_X = \exp^{(\mu+\sigma^2/2)^2}$
- Median: $\exp^{\mu}$
- Mode: $\exp^{\mu-\sigma^2}$
- Variance: $(\exp^{\sigma^2} - 1)\exp^{2\mu+\sigma^2}$

With Error Factor $Ef = \frac{q_{0.95}}{q_{0.50}}$

- $Ef = \exp^{1.645\sigma}$
- $\mu_X = \exp^{(\mu+\sigma^2/2)^2}$



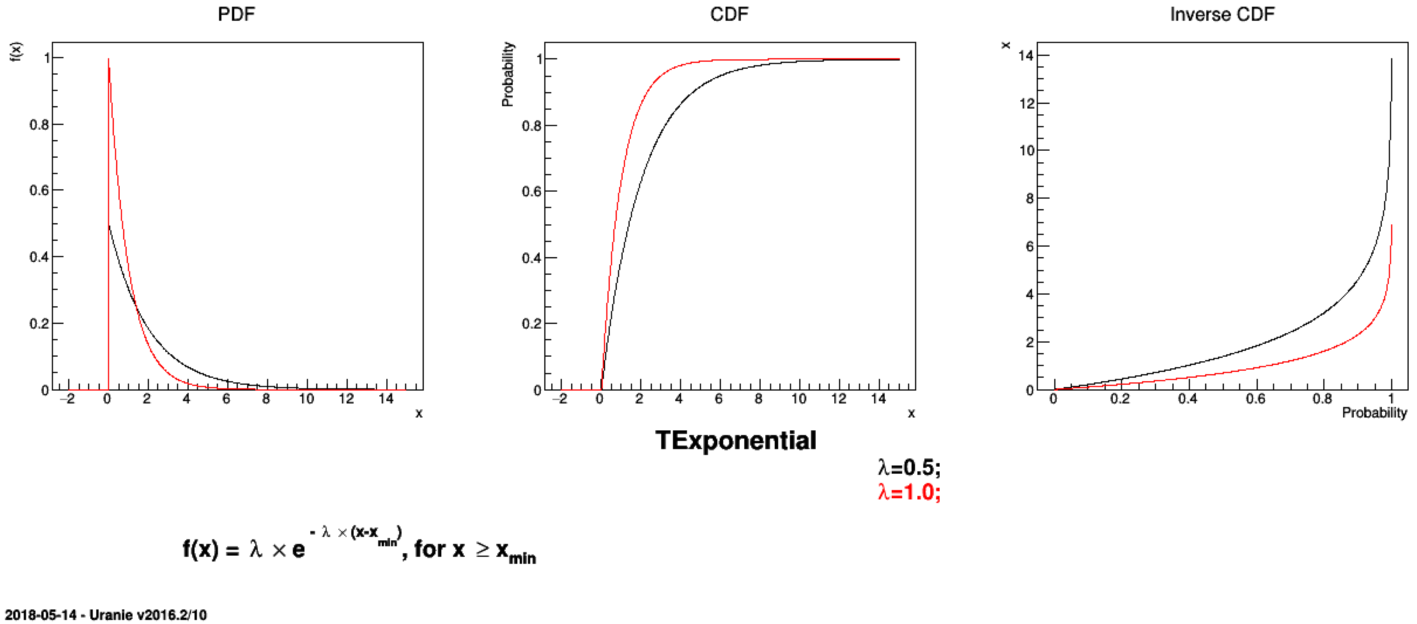
Commonly used PDF IV

Exponential distribution

- Defined by 2 parameters: λ (shape) and x_0 (bound)

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

- Mean: $x_0 + \dfrac{1}{\lambda}$
- Mode: x_0
- Variance: $\dfrac{1}{\lambda^2}$



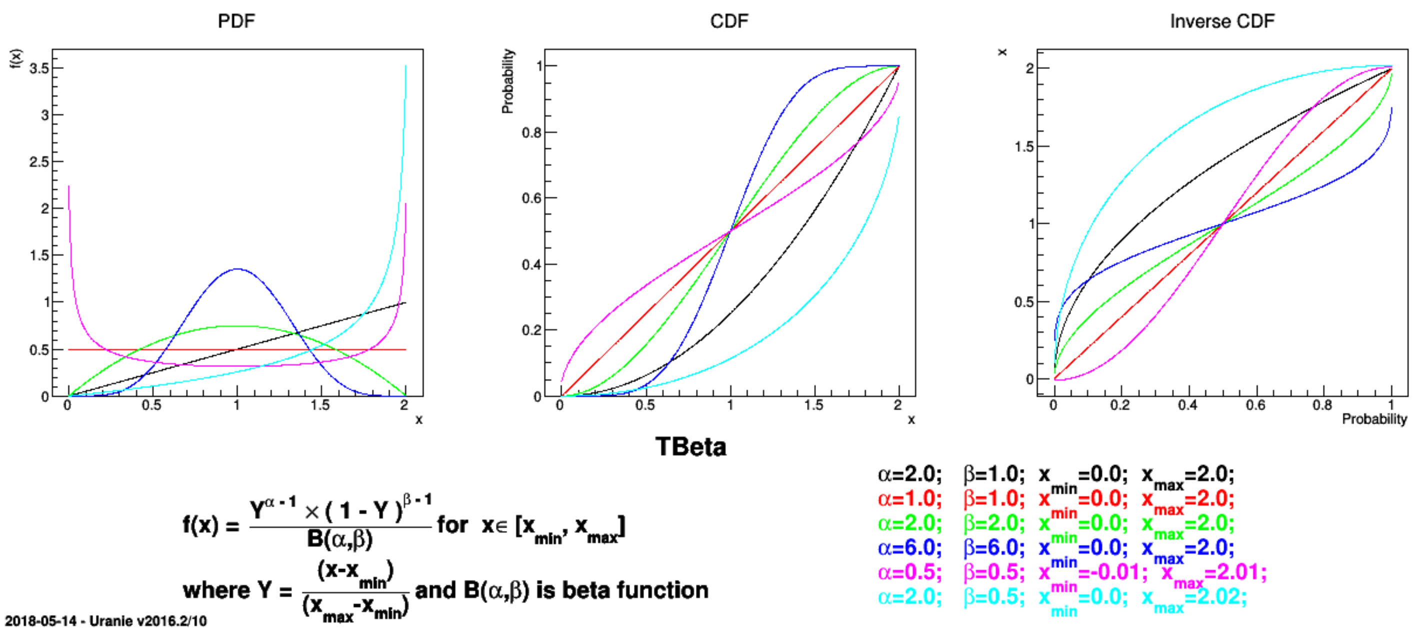
Commonly used PDF V

Beta distribution

- Defined by 4 parameters: α, β (shapes) and $x_0 < x_1$ (bounds)

$$f(x) = \frac{u^{\alpha-1} \times (1-u)^{\beta-1}}{B(\alpha,\beta)}, \forall x \in [x_0, x_1], u = \frac{x-x_0}{x_1-x_0}, B(\alpha,\beta) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$$

- Mean: $x_0 + (x_1 - x_0)\dfrac{\alpha}{\alpha+\beta}$
- Mode: depends on (α, β) values
- Variance: $(x_1 - x_0)^2 \dfrac{\alpha\beta}{\alpha+\beta+1}$



Many other possibilities

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 of parameters

Problem

- Let $(x_1, \ldots, x_{ns})$ 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 parameters θ^*, from which the data come from, is unknown
- Build an estimator $\hat{\theta}$ which would be as close to the true value θ^* as possible.

Two usual 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 $(x_i)_{1 \leq i \leq nS}$

We use the **Likelihood** function $\mathcal{L}(\theta; x_1, \ldots, x_{n_S})$:

$$\mathcal{L}(\theta; x_1, \ldots, x_{n_S}) = f(x_1, \ldots, x_{n_S}|\theta) = \prod_{i=1}^{n_S} f(x_i|\theta)$$

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

$$ln(\mathcal{L}(\theta; x_1, \ldots, x_{n_S})) = \sum_{i=1}^{n_S} ln(f(x_i|\theta))$$

or **the average log-likelihood**:

$$\hat{l}(\theta; x_1, \ldots, x_{n_S}) = \frac{1}{n_S} ln(\mathcal{L}(\theta; x_1, \ldots, x_{n_S}))$$

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

$$\hat{\theta}_{MLE} = \mathrm{argmax}_{\theta \in \Theta} \hat{l}(\theta; x_1, \ldots, x_{n_S})$$

. . . if any maximum exists

Maximum likehood, application

Let $(x_1, \ldots, x_{n_S})$ be an i.i.d sample from a normal law $\mathcal{N}(\mu, \sigma)$

If one defines $\theta = (\mu, \sigma)$ the unknown parameters, the density can be written:

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

$$\text{The Likelihood is : } \mathcal{L}(\theta; x_1, \ldots, x_{n_S}) = \prod_{i=1}^{n_S} f(x_i|\theta) = \left(\frac{1}{2\pi\sigma^2}\right)^{\dfrac{n_S}{2}} \exp^{-\dfrac{\sum_{i=1}^{n_S}(x-\mu)^2}{2\sigma^2}}$$

The average log-likelihood $\hat{l}(\theta; x_1, \ldots, x_{n_S})$ can be written as:

$$\hat{l}(\theta; x_1, \ldots, x_{n_S}) = -\frac{1}{2}\ln 2\pi - \ln\sigma - \frac{1}{2n_S\sigma^2}\sum_{i=1}^{n_S}(x_i-\mu)^2$$

- MLE for the mean parameter :

$$\frac{\partial \hat{l}}{\partial \mu} = \frac{1}{n_S\sigma^2}\sum_{i=1}^{n_S}(x_i-\mu) = 0 \Leftrightarrow \hat{\mu}_{MLE} = \bar{x} = \frac{1}{n_S}\sum_{i=1}^{n_S} x_i$$

- MLE for the variance parameter :

$$\frac{\partial \hat{l}}{\partial \sigma} = -\frac{1}{\sigma} + \frac{1}{n_S\sigma^3}\sum_{i=1}^{n_S}(x_i-\mu)^2 = 0 \Leftrightarrow \hat{\sigma}_{MLE} = \frac{1}{n_S}\sum_{i=1}^{n_S}(x_i-\hat{\mu}_{MLE})^2$$

Moments methods (MM) (1/2)

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

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 $(x_i)_{1 \leq i \leq n}$

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

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 methods (MM) (2/2)

- 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 methods, application (1/2)

- Case of the **normal distribution**

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

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

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

  Then MLE $\iff$ MM in the gaussian case

- Case of the **beta distribution**

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

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

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

Moments methods, application (2/2)

Then, the moments method gives us :

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

and

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

We obtain

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

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

Kernel methods

From the point of view of the histogramm,

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

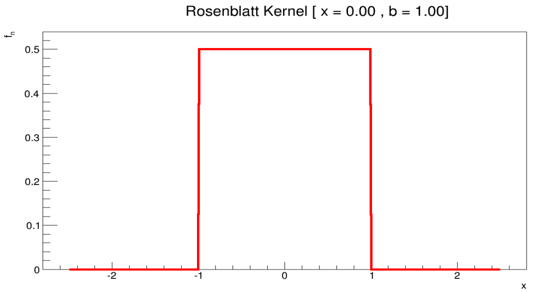
Then ***Rosenblatt*** (1956) suggests the estimator :

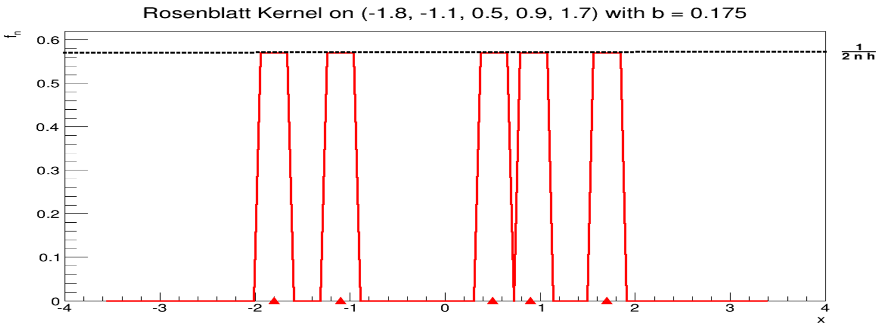
$$\hat{f}_{nS,h}(x) = \frac{\hat{F}_{nS}(x+h) - \hat{F}_{nS}(x-h)}{2 \times h}$$

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

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

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





DE LA RECHERCHE À L'INDUSTRIE

cea

PRACE

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: $K(u) > 0 \quad \forall u$

- $\forall h > 0$,

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

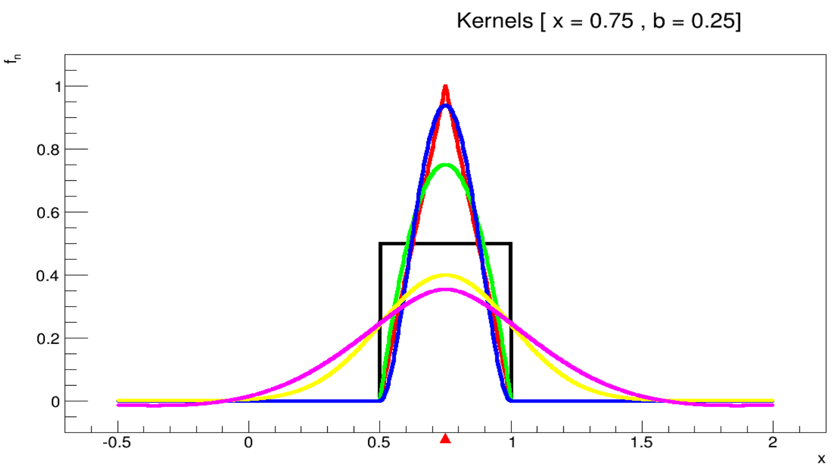
  is a ***kernel estimator*** of the density f $\quad (\int \hat{f}_{nS,h}(x)\,\mathrm{d}x = 1)$

- 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 $\hat{f}_{nS,h}$ 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 estimators - applications

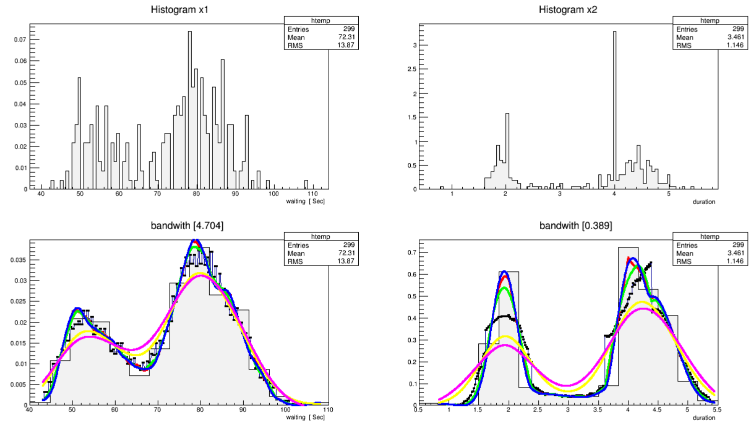
- Optimal bandwidth with the Silverman Rule (1996)

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

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	$k(x)$	σ_K
Rectangular	$1/2 \ , \ \|x\| < 1$	1.3510
Triangular	$1 - \|x\| \ , \ \|x\| < 1$	1.8882
Epanechnikov	$\frac{3}{4}(1 - x^2) \ , \ \|x\| < 1$	1.7188
Biweight	$\frac{15}{16}(1 - x^2)^2 \ , \ \|x\| < 1$	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

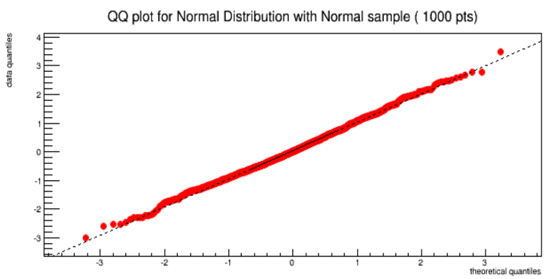
Goodness-of-fit techniques

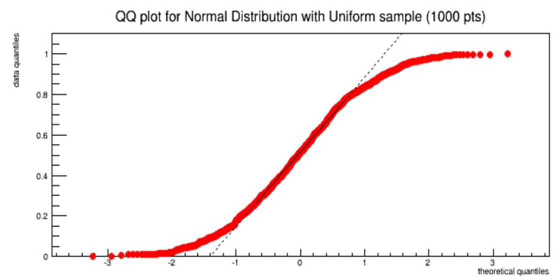
Goodness-of-fit techniques

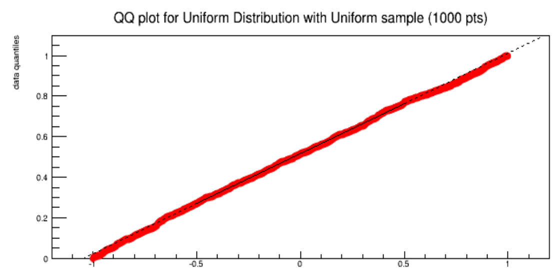
- Graphical methods
  - QQPlot
- Statistical Tests
  - Chi-Squared
  - Tests based on EDF Statistics
    - Kolmogorov-Smirnov
    - Cramer-von Misses
    - Anderson-Darling

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







Commonly used statistical tests

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

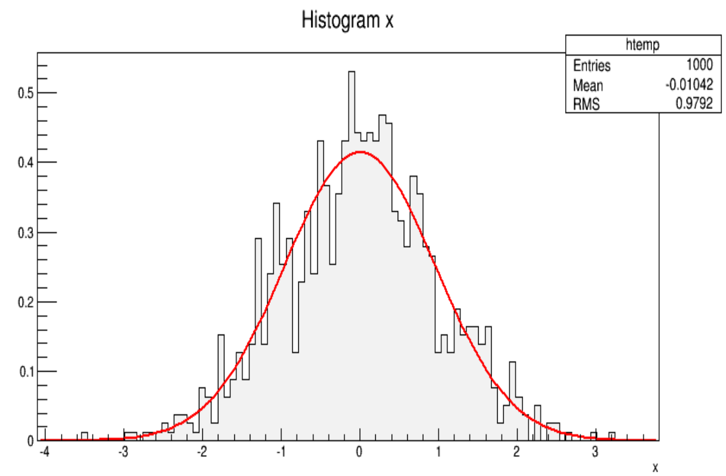
- Chi-Squared (χ^2)
- Tests based on EDF Statistics
  - Kolmogorov-Smirnov ($\mathbf{D}$)
  - Cramer-von Mises (W^2)
  - Anderson-Darling (A^2)

The chi-squared test (χ^2)

- 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

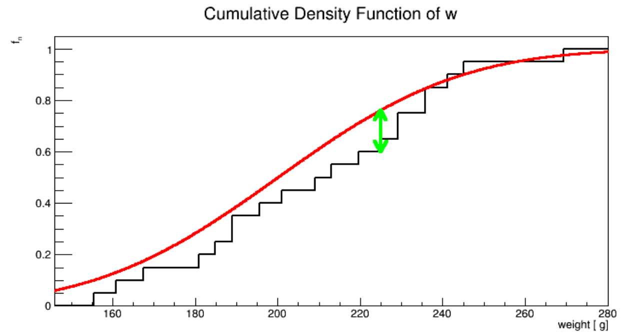


- The ratio nS/k must verify $nS/k \geq 5$
- The value of the $\chi 2$ test statistic are dependent on how the data is binned
- χ^2 test is generally less powerful than EDF tests

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 $F_{nS}(x)$ 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

$$D = \sup_x |F_{nS}(x) - F(x)|$$



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

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

where ψ is a *weight* function

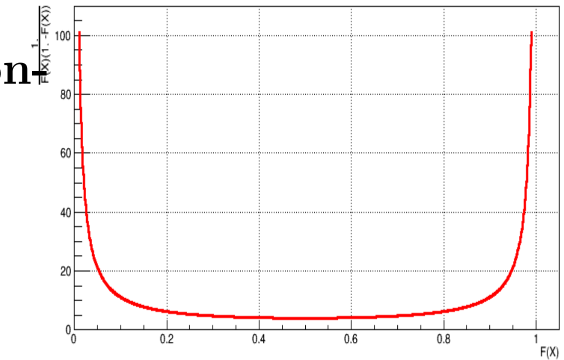
Tests based on EDF statistics (2/2)

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

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

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

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



- 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 $(x_1, x_2, \cdots, x_{nS})$, 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

Comparison 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

Commissariat à l'énergie atomique et aux énergies alternatives
Centre de Saclay | 91191 Gif-sur-Yvette Cedex
T. +33 (0)1 69 08 73 20 | F. +33 (0)1 69 08 68 86

Direction de l'énergie nucléaire
Département de modélisation des systèmes et structures
Service de Thermohydraulique et de mécanique des fluides

Etablissement public à caractère industriel et commercial | R.C.S Paris B 775 685 019

Backup outline

Backup

$\hat{t}_P$ and $\hat{t}_S$ limitation

Obvious limitation of these tests

Using a 122-points sample (not really randomly drawn).

Setting the test

1. $n_S = 122 \Rightarrow$ degree of freedom = 120 $(n_S - 2)$
2. $\hat{r}_P = 0.0668$ and $\hat{r}_S = 0.0853$

$$\hat{t}_P = \frac{0.0668 \times \sqrt{122-2}}{\sqrt{1-0.0668^2}} = 0.733 \text{ and } \hat{t}_S = \frac{0.0853 \times \sqrt{122-2}}{\sqrt{1-0.0853^2}} = 0.937$$

Interpret these results

- For a chosen $\alpha = 0.05$, $t_{5\%}(120) = 1.98$ and for $\alpha = 0.01$, $t_{1\%}(120) = 2.62$
  $\Rightarrow H_0$ accepted !
- Looking at table assuming student $\rightarrow$ normal.
  $\Rightarrow$ Critical probabilities are $0.46 < P_c^P < 0.47$ and $0.34 < P_c^s < 0.35$
  $\Leftrightarrow$ You know knothing Jon Snow !