INTRODUCTION
A
LA STATISTIQUE

Statistique pour l'ingénieur

Objectifs

- Prendre en compte l'aléatoire dans le processus décisionnel
- Comment prévoir en présence du hasard ?

Ambiguïté du terme

La Statistique : Ensemble de méthodes permettant d'analyser (de traiter) des ensembles d'observations (des données)

Une statistique : Donnée statistique (ex. : statistique du commerce extérieur français)

Les données

- Enquêtes socio-économiques
- Observations de phénomènes naturels
- Résultats d'expériences scientifiques
- Résultats de simulations numériques

Démarche statistique

Statistique Descriptive (Exploratoire)

- **Description synthétique des données :**
  - Représentations graphiques,
  - Tableaux,
  - Indicateurs numériques (moyenne, écart-type …).
- **Analyse de données**

  classification, analyse factorielle, …
- **Pas de modèles probabilistes dans cette étape**

Statistique Inférentielle (Décisionnelle)

- **Étendre les propriétés constatées sur un échantillon à toute une population** (inférence statistique) :
  - Estimation d'une moyenne, variance,
  - Tests d'hypothèse,
  - Proposer des modèles probabilistes pour gérer des risques d'erreurs.
- **Les probabilités jouent un rôle fondamental dans cette étape**

SOMMAIRE

Terminologie de base

Population Ω (limitée ou de très grande taille)

Individu ω : tout élément de la population

Échantillon : sous-ensemble (de taille n) de la population sur lequel sont réalisées les observations

Recensement : observation (ou interrogation) de toute la population

Enquête ou sondage : observation d'un échantillon

Variable X : $\Omega \rightarrow \Omega'$ (caractéristique définie sur la population) ;

- **Quantitative** ($\Omega'=\Re$)
  discrète *(ex : âge)* ou continue *(ex : poids)*
- **Qualitative** ($\Omega'=V$)
  nominale *(ex : sexe)* ou ordinale *(ex : mention)*

Données : ensemble des individus observés, des variables considérées et des observations de ces variables sur ces individus.

SOMMAIRE

Variable aléatoire

- **Variable aléatoire ($X : \Omega \rightarrow \Omega'$) :**
  Grandeur dépendant du résultat d'une expérience aléatoire (dont le résultat est non prévisible)

  Ex : choisir une caisse au supermarché, X = son temps d'attente

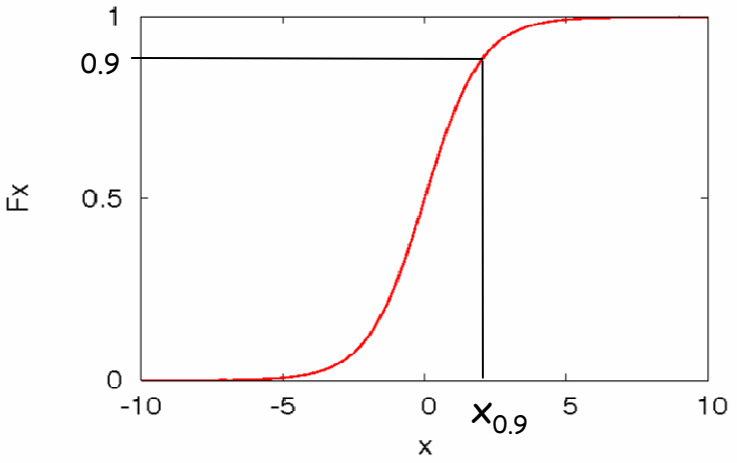
- **Réalisation :** *x* est une *réalisation* de *X* (valeur prise par *X*)

- **Fonction de répartition de *X* :**

$$F_X : \Omega' \rightarrow [0\,;1]$$

$$x \rightarrow F_X(x) = P(X \leq x)$$

$$\lim_{X \rightarrow -\infty} F_X(x) = 0 \qquad \lim_{X \rightarrow +\infty} F_X(x) = 1$$



- **Quantile (ou fractile) d'ordre q :**
  x_q tel que $P(X \leq x_q)=q \Leftrightarrow F_X(x_q)=q$

Variable aléatoire continue

- **$P(a < X \leq b) = F_X(b) - F_X(a)$**

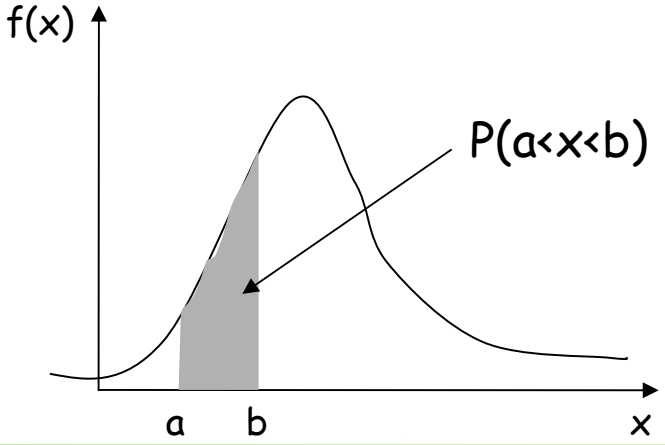
  Densité moyenne de probabilité sur [a,b] : $f_X(a,b)=[F_X(b)-F_X(a)]/(b-a)$

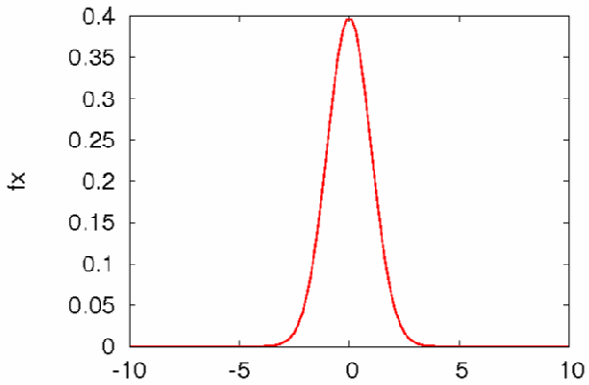
- **Densité de probabilité f_X** = dérivée de la fonction F_X

$$P(X \in I) = \int_I f_X(x)dx \quad \text{pour tout intervalle I de } \Re$$

  - f_X est une **fonction positive** telle que $\int_{-\infty}^{+\infty} f_X(x)dx = 1$ et $\lim_{x \to \pm\infty} f_X(x) = 0$

  - Sa représentation graphique met en évidence les **zones à + forte probabilité.**





Exemple : densité gaussienne

Moments des variables aléatoires

- **Espérance mathématique** d'une v.a. continue : $\mu = \mathrm{E}(X) = \int x f_X(x) dx$

Propriétés : Indicateur de tendance centrale

$\mathrm{E}[aX + b] = a\,\mathrm{E}[X] + b$

Remarque : l'existence de E(X) n'est pas garantie (ex : $f(x) = \frac{1}{\pi(x^2+1)}$)

- **Variance** d'une variable aléatoire :

$$\sigma^2 = \mathrm{var}(X) = \mathrm{E}\left[(X - \mathrm{E}(X))^2\right]$$

$$\sigma^2 = \int [x-\mu]^2 f_X(x) dx = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$$

Propriétés : Indicateur de dispersion

$\mathrm{var}(aX) = a^2\,\mathrm{var}(X)$; $\mathrm{var}(X + b) = \mathrm{var}(X)$

Remarque : variance nulle $\Leftrightarrow$ *v.a. certaine*

Moments des variables aléatoires

- **Moments d'ordre supérieur** d'une v.a. continue :

$$m_n = \mathrm{E}\left[(X - \mathrm{E}(X))^n\right]$$

De la variable centrée réduite $\tilde{m}_n = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^n\right]$

Propriétés : n = 3 ⇨ Indicateur d'**asymétrie**

n = 4 ⇨ Indicateur d'**aplatissement des extrêmes**
appelé **kurtosis**

Couple de variables aléatoires

- Variables quantitatives : $(X,Y) : \Omega \rightarrow \Re^2$

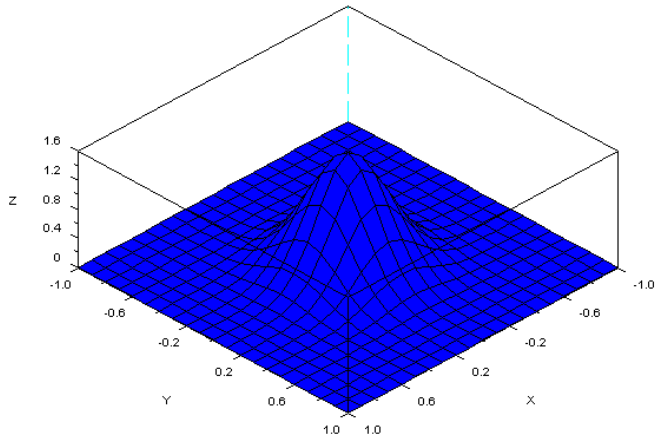
- **Fonction de répartition conjointe :**

$F_{X,Y} : \Re^2 \rightarrow [0,1]$

$(x,y) \rightarrow F_{X,Y}(x,y) = P(X \leq x, Y \leq y)$

- **Densité de probabilité conjointe $f_{X,Y}$:**

$$F_{X,Y}(x,y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u,v)\,du\,dv$$



- **Indépendance entre X et Y $\Leftrightarrow$ $f_{X,Y}(x,y) = f_X(x)\, f_Y(y)$**

Couple de variables aléatoires

- **Covariance** : $\operatorname{cov}(X,Y) = E[XY] - E[X]E[Y]$
$= E[(X - E[X])]\,E[(Y - E[Y])]$

  Propriétés : E(X+Y) = E(X)+E(Y)
  Var(X+Y) = var(X) + var(Y) + 2cov(X,Y)

- **Covariance et indépendance :**

  X et Y sont indépendantes ⇒ cov(X,Y) = 0

  cov(X,Y) = 0 ⇏ X et Y sont indépendantes

  X et Y sont décorrélées ⇔ cov(X,Y) = 0

- **Inégalité de Cauchy-Schwartz :** $\left|\mathrm{E}[XY]\right| \le \left(\mathrm{E}[X^2]\mathrm{E}[Y^2]\right)^{1/2}$

- **Coefficient de corrélation :** $\rho = \dfrac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$

  Propriétés : Inégalité C-S ⇒ $-1 \le \rho \le 1$

SOMMAIRE

Principales lois de probabilité discrètes

- **Loi uniforme** : X = {1,2,…,n} avec P(X=k) = 1/n

$$\mathrm{E}(X) = \frac{n+1}{2} \; ; \; \mathrm{var}(X) = \frac{n^2-1}{12}$$

Exemple : lancement d'un dé

- **Loi de Bernouilli $\mathcal{B}(p)$** :

$$\mathrm{E}(X) = p \; ; \; \mathrm{var}(X) = p(1-p)$$

$$X = \begin{cases} 1 \text{ avec une proba p} & \textit{(succès)} \\ 0 \text{ avec une proba 1-p} & \textit{(échec)} \end{cases}$$

- **Loi binomiale $\mathcal{B}(n,p)$** : *n* répétitions indépendantes d'une Bernouilli

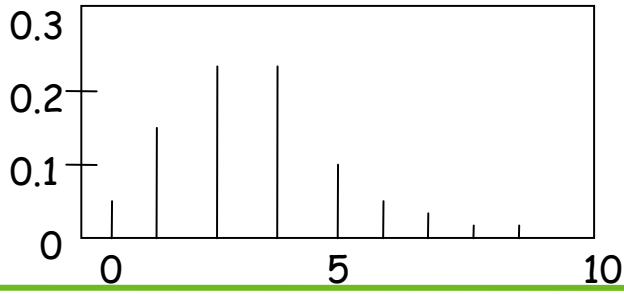
$$X = \sum_{i=1}^{n} X_i \quad \Longrightarrow \quad \mathrm{P}(X=k) = C_n^k p^k (n-p)^{n-k}$$

Exemple : sondage (OUI=1, NON=0)

p faible, n grand

- **Loi de Poisson $\mathcal{P}(\lambda)$** : loi du nombre d'occurrences d'événements « rares », sans mémoire et dans un intervalle de temps donné.

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

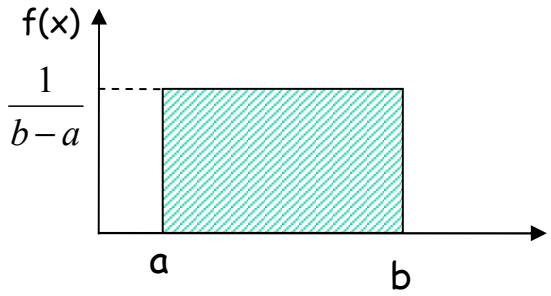


Ex : nombre de personnes dans une file, nombre d'appels à un standard

Principales lois de probabilité continues

- **Loi uniforme $\mathcal{U}$ [a,b]**

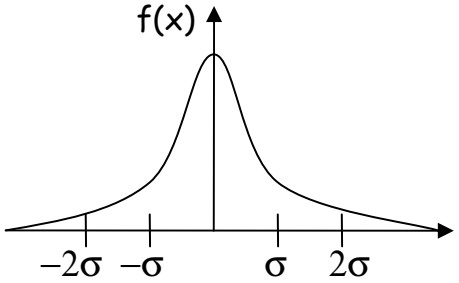
$$f(x) = \frac{1}{b-a} \text{ si } a \le x \le b \; ; \; f(x) = 0 \text{ ailleurs}$$

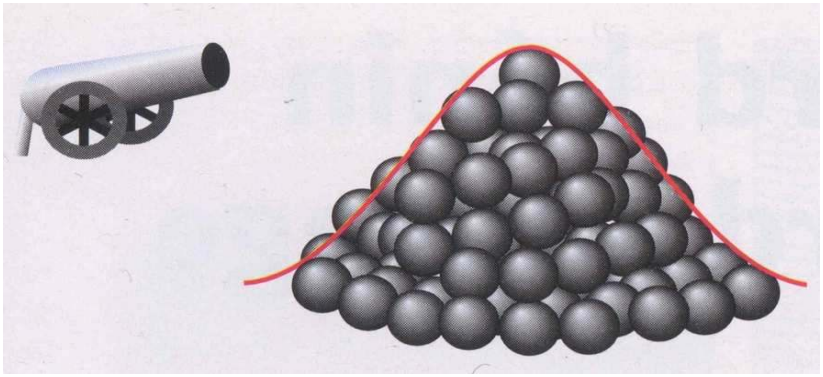


- **Loi normale $\mathcal{N}(\mu,\sigma^2)$**

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

$$\mathrm{E}(X) = \mu \; ; \; \mathrm{var}(X) = \sigma^2$$





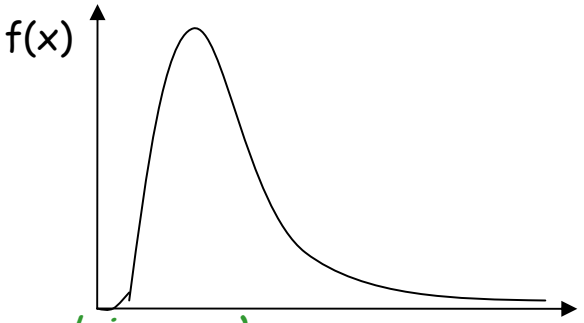
$P(\mu - \sigma < X < \mu + \sigma) = 0.68$
$P(\mu - 1.64\sigma < X < \mu + 1.64\sigma) = 0.90$
$P(\mu - 1.96\sigma < X < \mu + 1.96\sigma) = 0.95$
$P(\mu - 3.09\sigma < X < \mu + 3.09\sigma) = 0.998$

Exemples : impacts des boulets de canon (Jouffret, 1872), incertitude de mesure

Principales lois de probabilité continues

- **Loi du Chi-deux** : si $X_i \sim \mathcal{N}(0,1)$ pour *i=1,…,n* alors $\sum_{i=1}^{n} X_i^2 \sim \mathcal{X}^2(n)$

- **Loi lognormale $L\mathcal{N}(\mu,\sigma^2)$** : $\ln(X) \sim \mathcal{N}(\mu,\sigma^2)$



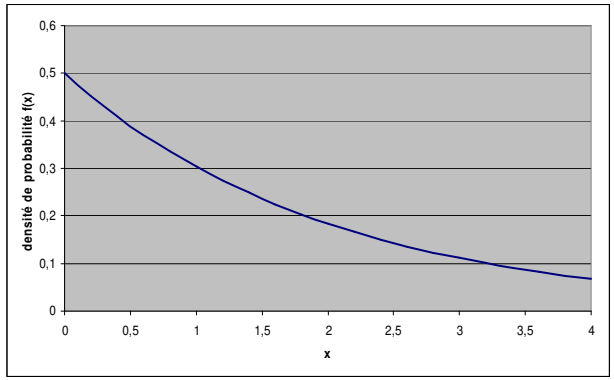
Le produit de v.a. $\xrightarrow{\mathcal{L}}$ $L\mathcal{N}$

Exemples : variables positives et asymétriques (poids, salaires, …),
résolution d'un instrument (sources d'erreur = multiplication d'un grand nombre de petits facteurs indépendants)

- **Loi exponentielle $\mathcal{E}(\lambda)$** : $f(x) = \lambda \exp(-\lambda x)$ si $x \geq 0$;

$$E(X) = \frac{1}{\lambda} \; ; \; \mathrm{var}(X) = \frac{1}{\lambda^2}$$

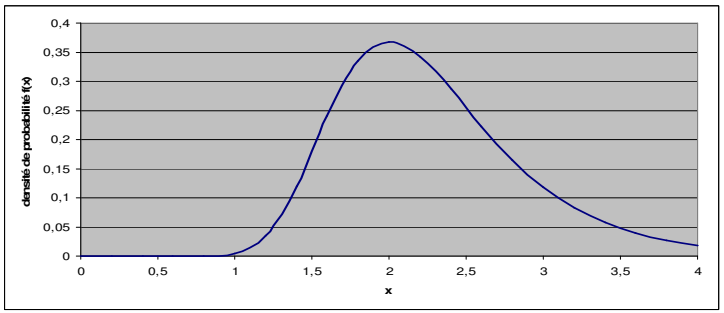
Exemples : temps d'attente,
durée de vie de systèmes sans usure
⇨ i.e. la proportion de matériels défaillants est chaque année la même.



Principales lois de probabilité continues

➢ **Loi de Gumbel G(m,s)** : $f(x) = \frac{1}{s}\exp\left(-\frac{x-\mu}{s}\right)\exp\left(-\exp\left(-\frac{x-\mu}{s}\right)\right)$

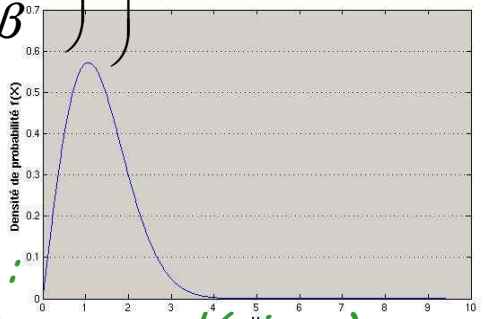
- Densité de probabilité fortement asymétrique autour du mode *m*
- les fortes valeurs restent probables



Exemple : modélisation des phénomènes climatiques extrêmes (modèle de crue, …)

➢ **Loi de Weibull W(x_0,α,β)** : $f(x) = \frac{\alpha}{\beta}\left(\frac{x-x_0}{\beta}\right)^{\alpha-1}\exp\left(-\left(\frac{x-x_0}{\beta}\right)^{\alpha}\right)$

Généralisation de la loi exponentielle



Exemple en mécanique : Durée de vie d'un matériel qui :

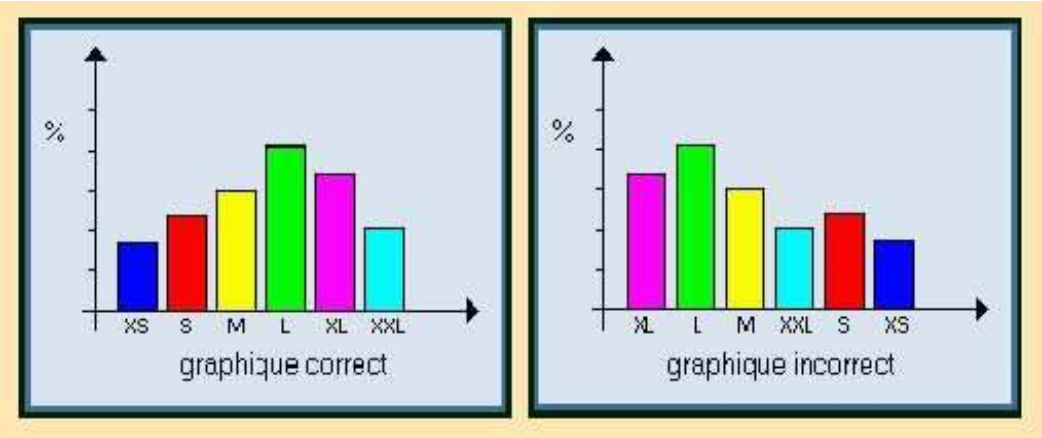
- *se dégrade pour α>1 (ténacité des cuves de réacteurs nucléaires)*
- *ou se bonifie pour α<1 (résistance du béton sans agression externe)*

SOMMAIRE

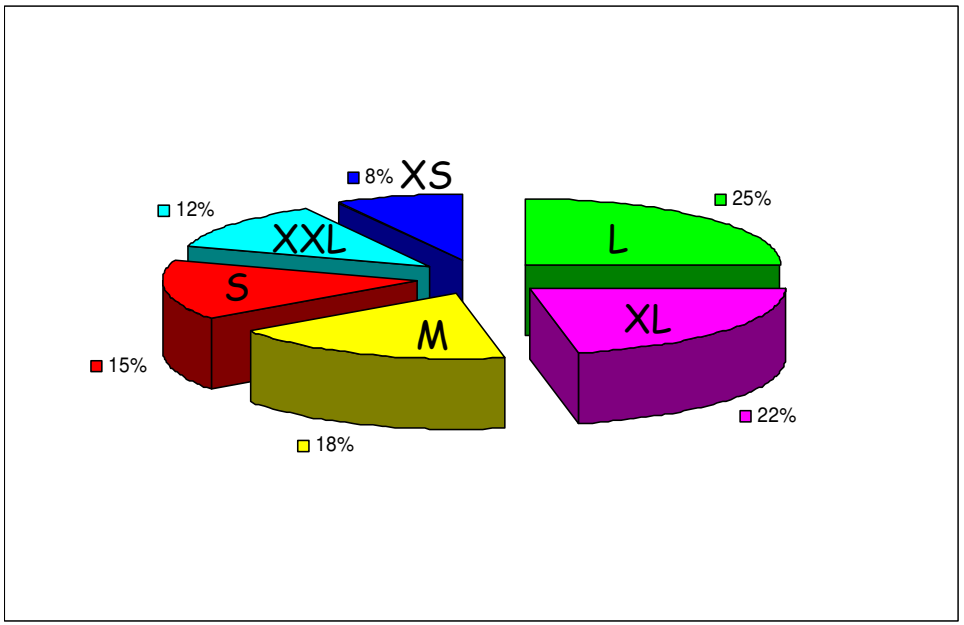
Représentations graphiques – Variables discrètes

- **Diagrammes en bâtons**

Exemple (variable ordinale) : taille des habits achetés dans un magasin



- **Diagrammes sectoriels (« camemberts »)**

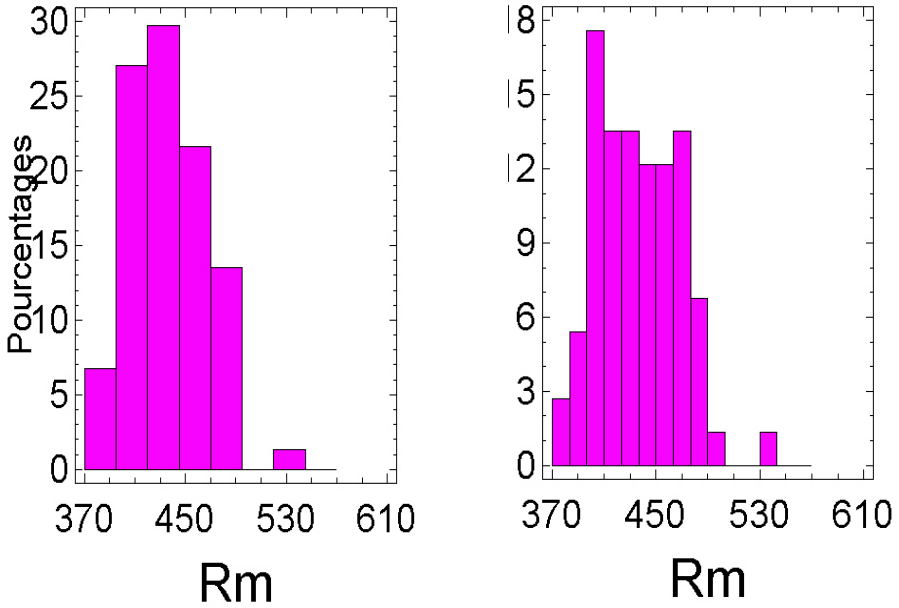


Représentations graphiques – Variables continues

Histogramme

(approximation de la densité)

Exemple : essais de traction de boulons ; limite de rupture (MPa)



Rm : résistance mécanique, valeur de la contrainte à la rupture.

Fonction de répartition empirique

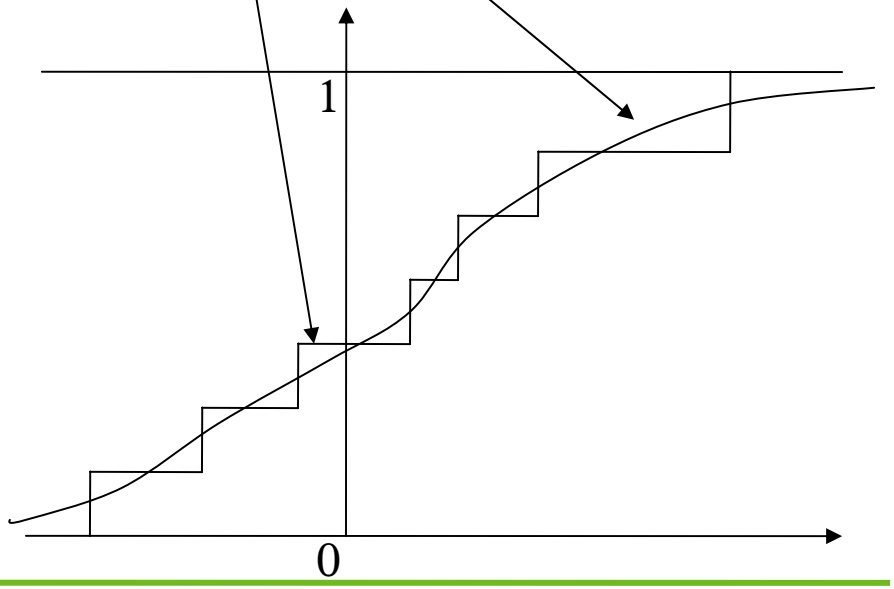
(histogramme cumulé)

$F_n : \Re \longrightarrow [0,1]$

$$x \longrightarrow F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{x_i \le x}$$

Théorème de Glivenko-Cantelli :

$$\sup_{x\in\Re} |F_n(x) - F(x)| \xrightarrow[n\to\infty]{ps} 0$$



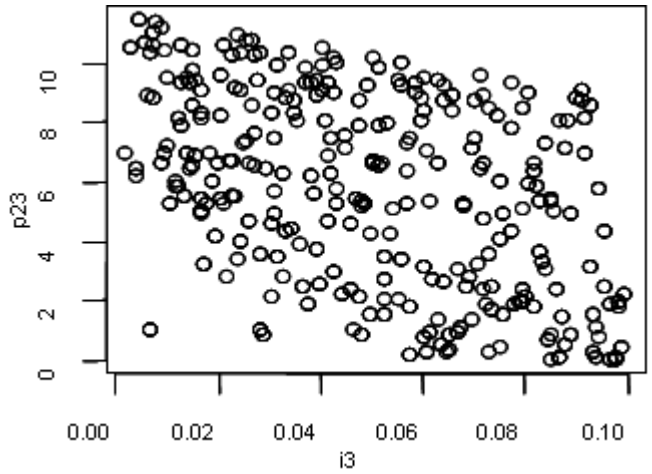
Représentations graphiques – Scatter plot

Mesure le caractère linéaire du nuage de points

- n calculs
- Graphe Sortie / chaque entrée

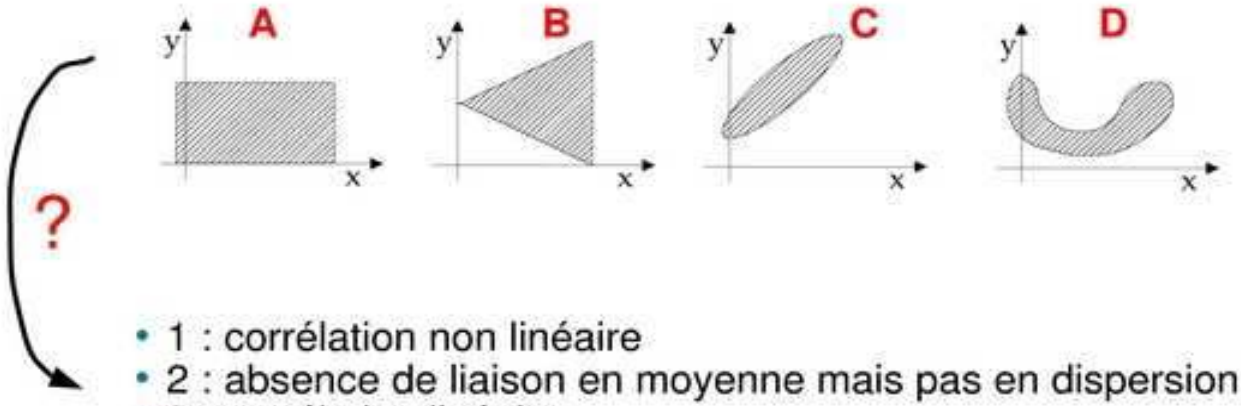
$$\rho = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Exemple : n=300



Exercice

- Nuage de points : exemples



- 1 : corrélation non linéaire
- 2 : absence de liaison en moyenne mais pas en dispersion
- 3 : corrélation linéaire
- 4 : absence de liaison

SOMMAIRE

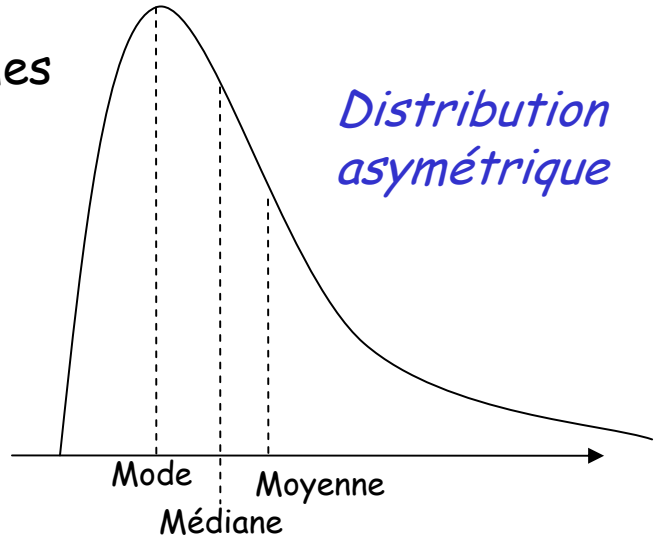
Propriétés de position

- **Moyenne μ**

  Peu robuste car sensible aux valeurs extrêmes

- **Médiane** : valeur M telle que F(M)=0.5

  Insensible aux valeurs extrêmes

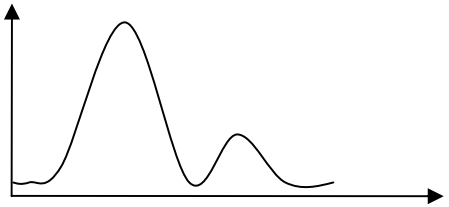


- **Mode**

  V.a. discrète : valeur la plus fréquente

  V.a. continue : pic de l'histogramme

  Plusieurs modes : distribution multimodale



- **Les valeurs minimale et maximale**

  Sensibles aux valeurs aberrantes

- **Quartiles et autres quantiles** (déciles, centiles, …) :

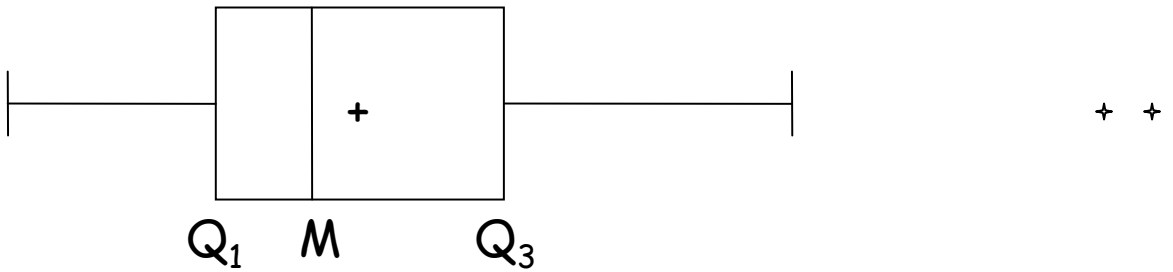
  $F(Q_1)=0.25$; $F(Q_2)=0.5$; $F(Q_3)=0.75$;

Propriétés de dispersion (1/2)

- **Étendue** (intervalle de variation) $|x_{max} - x_{min}|$
  Instable car dépendant de valeurs extrêmes

- **Intervalle interquartile** $|Q_3 - Q_1|$ où $F(Q_1)=0.25$ et $F(Q_3)=0.75$
  Mesure plus robuste que l'étendue

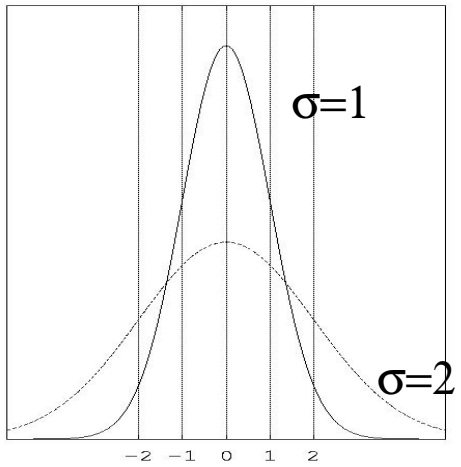
- **Diagramme en boîte – Boîte-à-moustaches** (« box plot » de Tukey) :
  résumé : min [> Q_1-1.5(Q_3-Q_1)] , Q_1, médiane, moyenne, Q_3, max [< Q_3+1.5(Q_3-Q_1)]
  + valeurs en dehors de cet intervalle



Propriétés de dispersion (2/2)

- **Variance :** $\sigma^2 = \mathrm{E}(X-\mu)^2$

  Distance moyenne des observations par rapport à la moyenne des observations



- **Écart-type** σ

  (même unité que les observations)

  Mesure cohérente avec la moyenne (distances euclidiennes)

- **Écart moyen :** $E.M. = \mathrm{E}|X-\mu|$

  Ordre de grandeur des déviations autour de la moyenne

- **Écart médian :** $E.med = \mathrm{E}|X-M|$

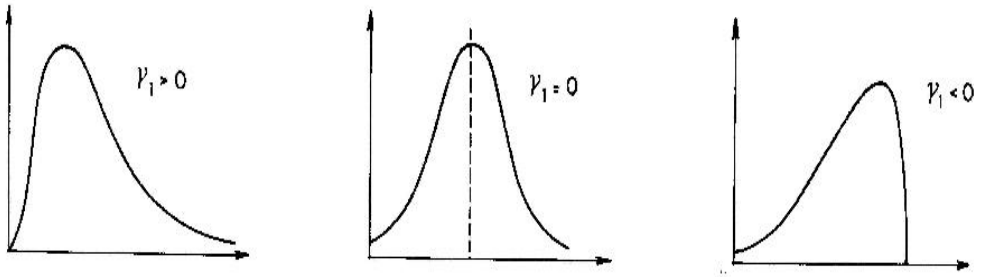
  Mesure cohérente avec la médiane

- **Coefficient de variation** $V = \dfrac{\sigma}{\mu}$ (indicateur sans dimension)

Propriétés de forme

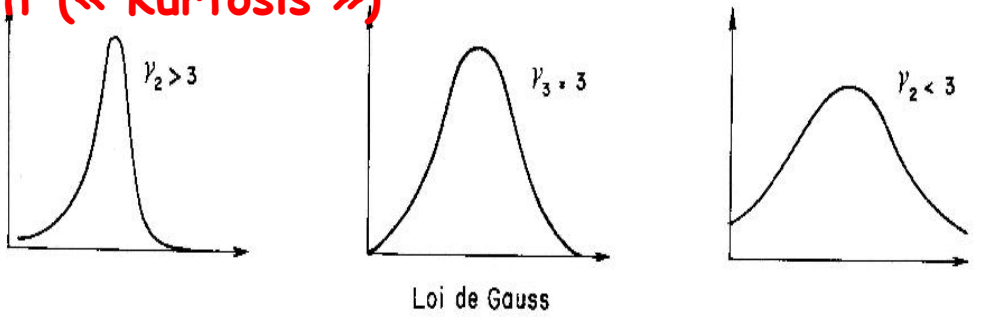
- **Coefficient d'asymétrie (« skewness »)**

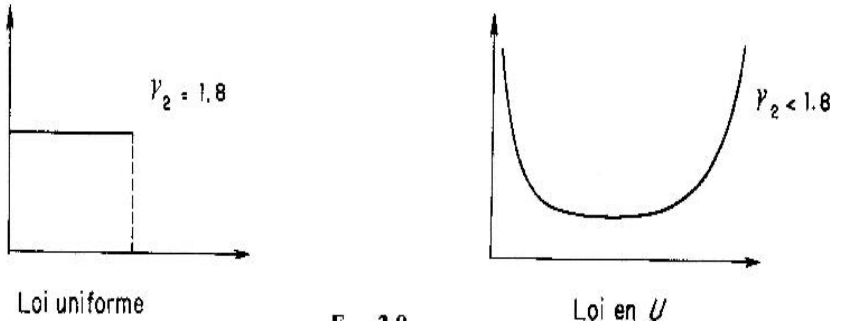
$$\gamma_1 = \frac{\mathrm{E}(X-\mu)^3}{\sigma^3}$$



- **Coefficient d'aplatissement (« Kurtosis »)**

$$\gamma_2 = \frac{\mathrm{E}(X-\mu)^4}{\sigma^4}$$



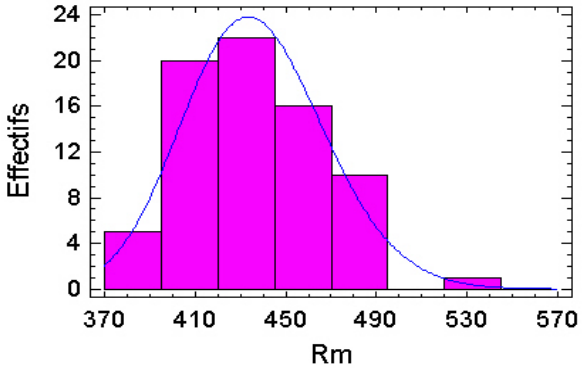


SOMMAIRE

Ajustement empirique à une densité de probabilité

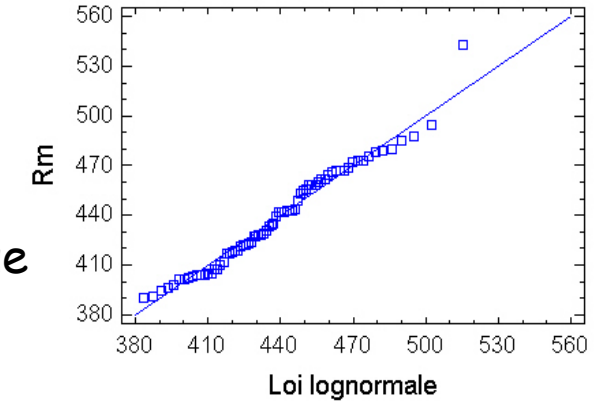
- **Forme de l'histogramme**

Exemple : essais de traction de boulons ;
limite de rupture (MPa) ;
ajustement par une loi lognormale



- **Comparaison sommaire des propriétés mathématiques**
  - Asymétrie étirée à droite (médiane=434, moyenne=437)
  - Coef. d'asymétrie et d'aplatissement du log des données
  - …

- **Ajustements graphiques**
  - QQ-plot : Graphique quantiles-quantiles
    ⇒ Quantile théorique / quantile empirique



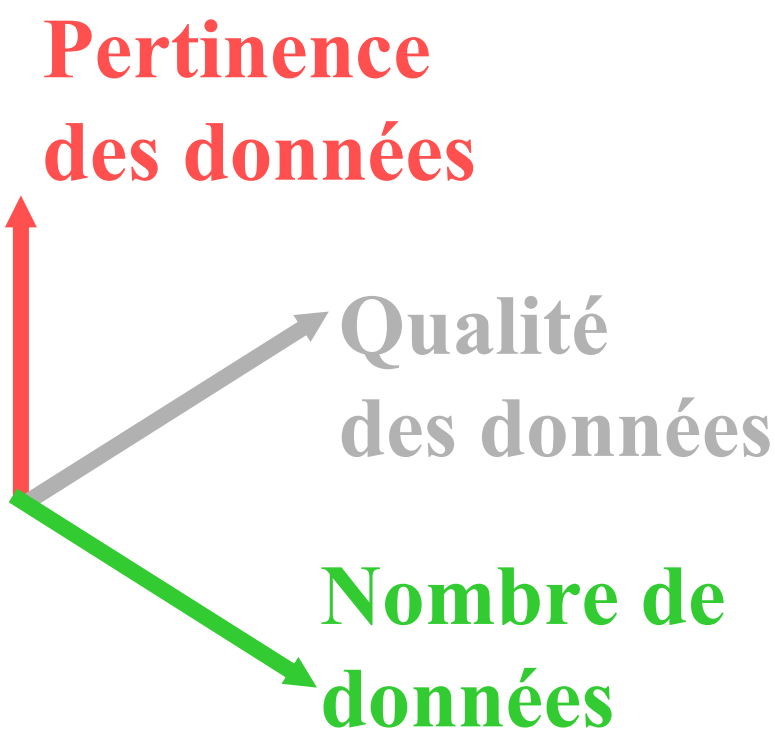
Ajustement empirique à une densité de probabilité

Sans données

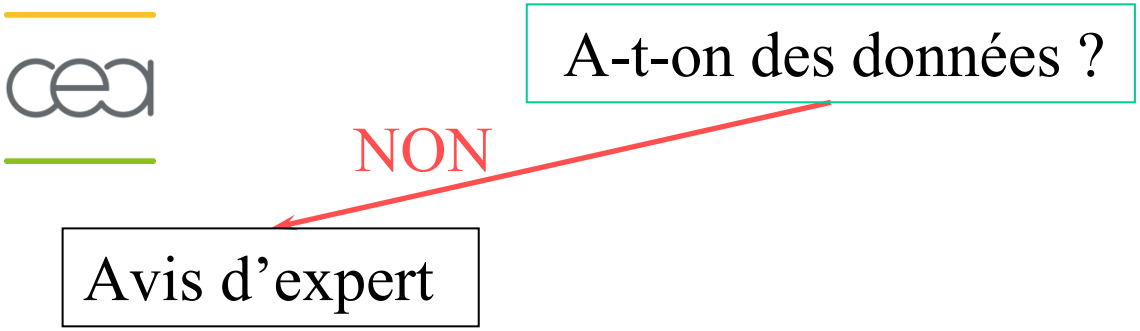
- ✓ Loi liée à la physique
- ✓ Par avis d'expert
- ✓ …

Avec données

- ✓ Minimum – maximum,
- ✓ Moyenne – écart-type,
- ✓ Distribution empirique – distribution théorique ajustée
- ✓ …



Ajustement empirique à une densité de probabilité



Ajustement empirique à une densité de probabilité

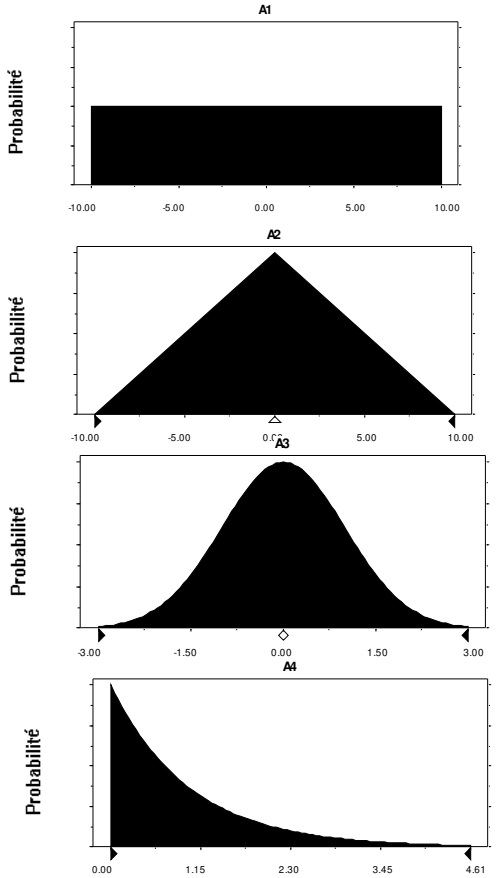
Exemples d'interprétation d'avis d'experts

Variable bornée par une valeur min et une valeur max, aucun autre a priori ⇨ loi uniforme

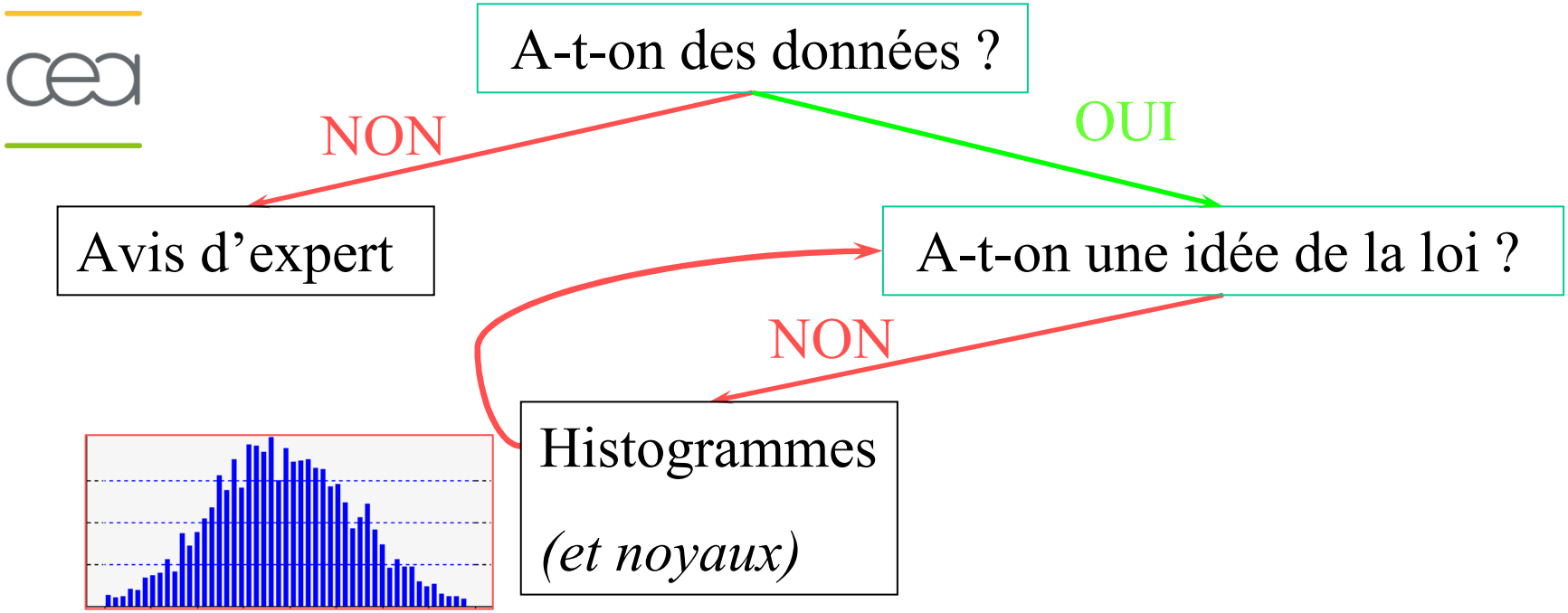
Variable bornée par une valeur min et une valeur max, une valeur plus probable que les autres ⇨ loi triangulaire

On connaît uniquement la moyenne et l'écart-type ⇨ loi normale

Variable positive, on connaît uniquement la moyenne ⇨ loi exponentielle



Ajustement empirique à une densité de probabilité



Ajustement empirique à une densité de probabilité

Histogrammes et méthode des noyaux

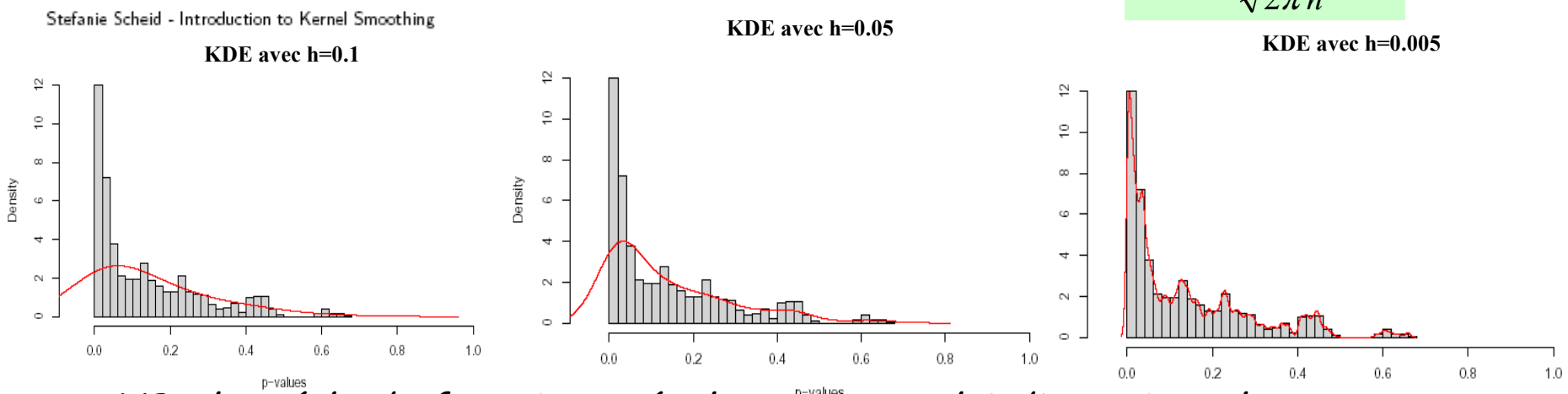
La représentation par histogrammes dépend des classes …
… on peut également représenter la densité à l'aide de noyaux
⇒ méthode non paramétrique

$$\hat{f}_h(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right)$$

Principe : "lissage" de l'histogramme

- N : taille de l'échantillon
- h : largeur de la fenêtre ⇔ paramètre de lissage
- K : noyau (kernel) ⇒ gaussien, uniforme, ..

$$K(u) = \frac{1}{\sqrt{2\pi}h} e^{-\frac{1}{2}u^2}$$



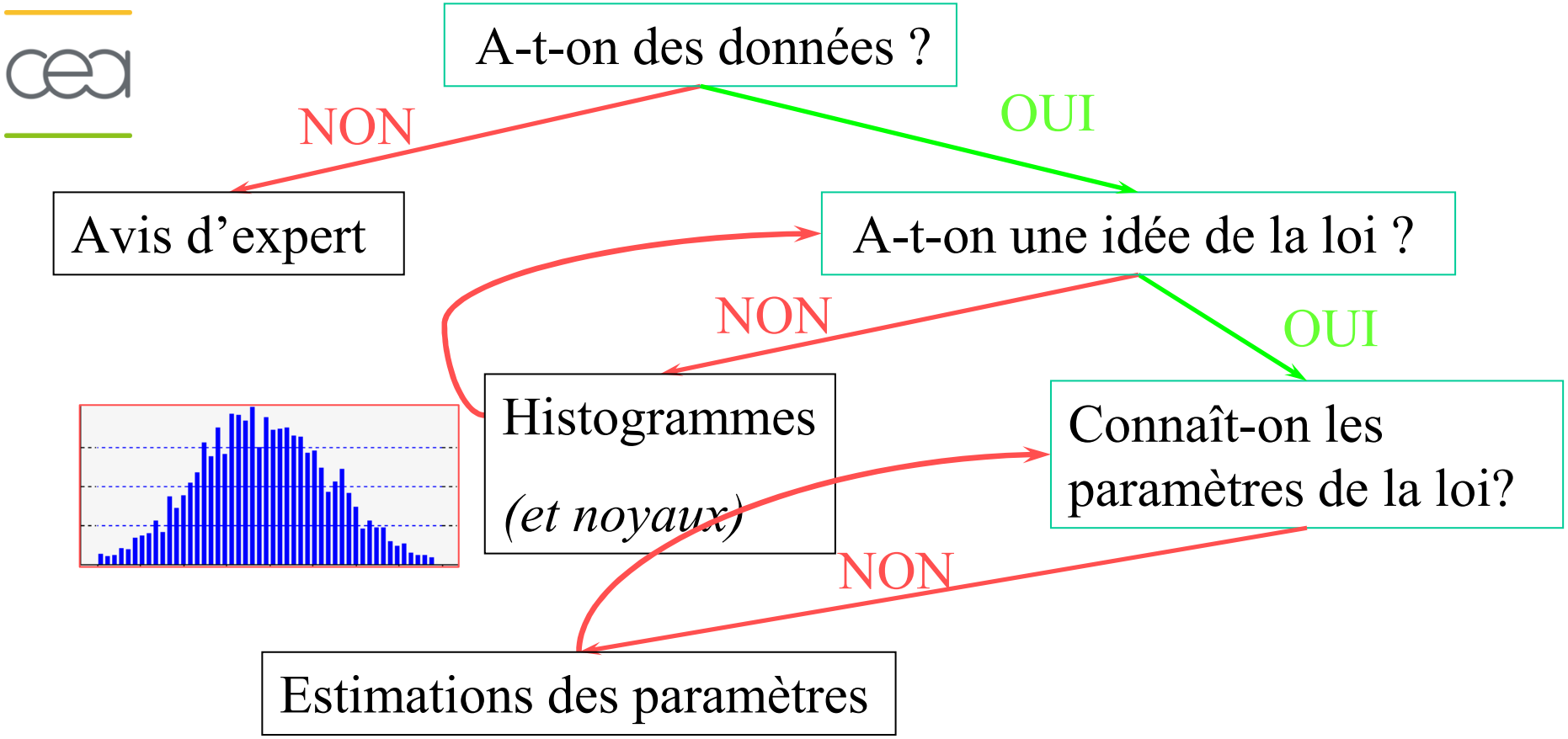
NB : la méthode fonctionne également en multi-dimensionnel

Ajustement empirique à une densité de probabilité

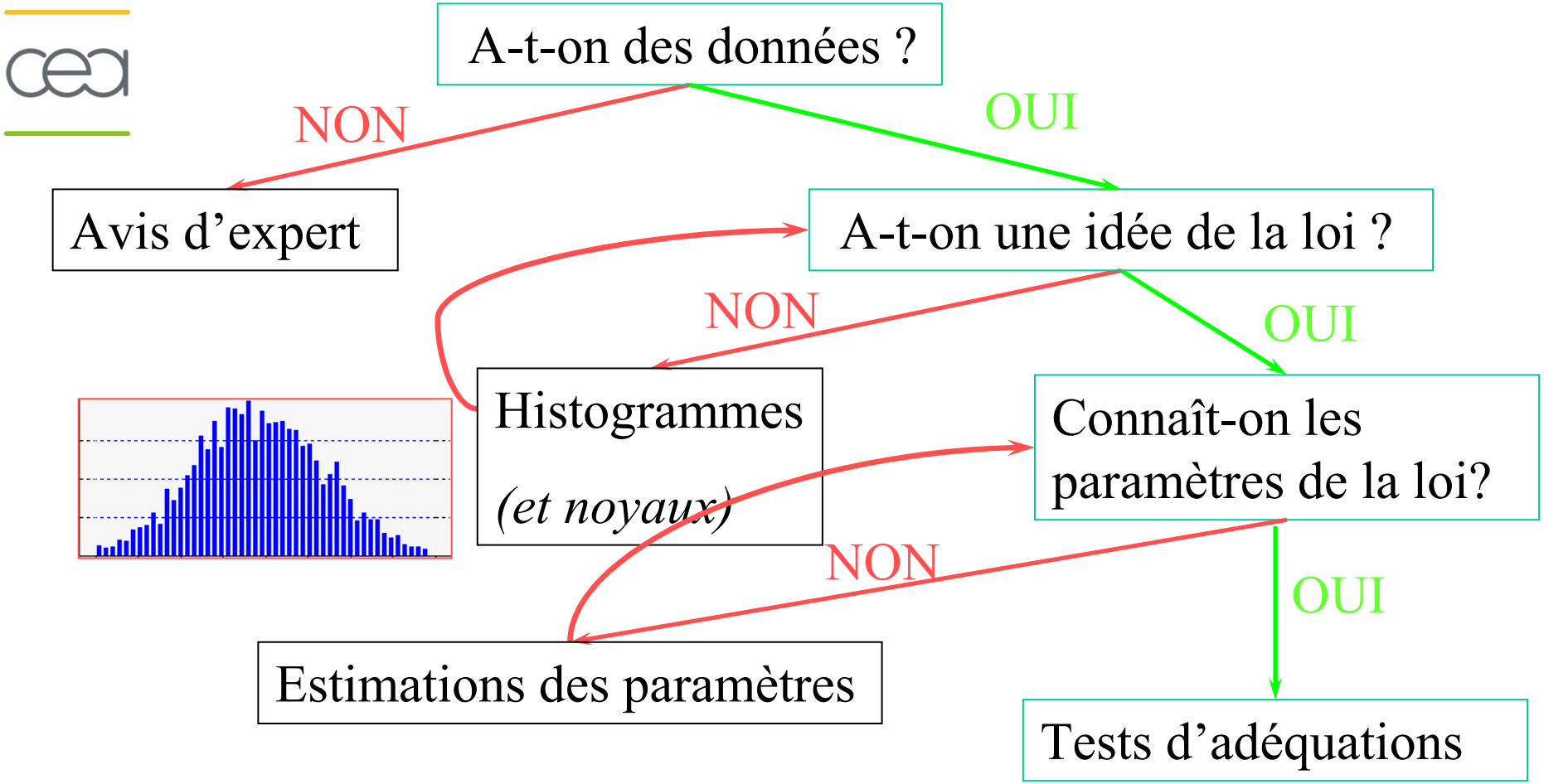
Quelques considérations sur le support des lois

Positive Continue (0,+∞)	Illimité Continue (-∞,+∞)	Limité Continue (a,b)
Exponentielle Gamma/Erlang Log normal Weibull Chi-deux F (Fisher-Snedecor) Log-Laplace Log-logistique Pareto …	Normale Cauchy Loi des Extrêmes A,B Laplace Logistique Student …	Bêta Triangulaire Uniforme …

Ajustement empirique à une densité de probabilité



Ajustement empirique à une densité de probabilité



Ajustement empirique à une densité de probabilité

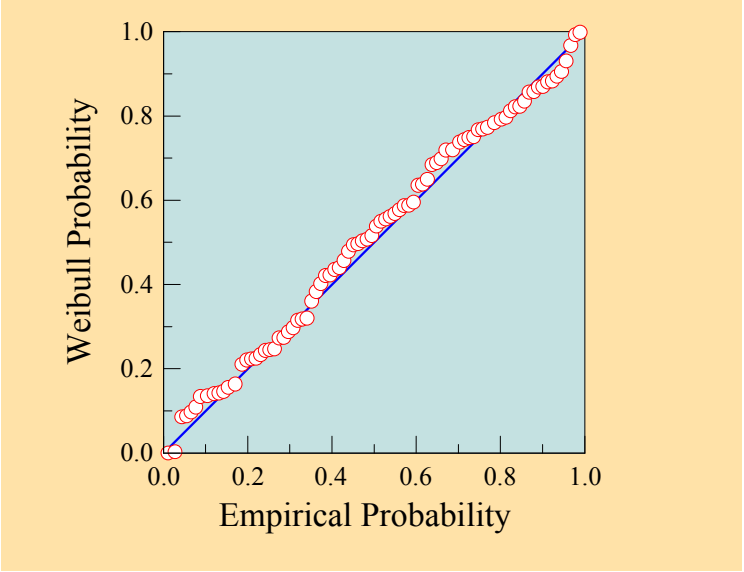
Adéquation : vérification graphique

- Comparaison des densités de probabilités théoriques et empiriques (empirique = histogramme)
- **P-P plot** : graphe des probabilités pour comparer les fonctions de répartition empiriques et théoriques

Théorique : $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$

Empirique : $F_n(x) = \frac{1}{n}\sum_{k=1}^{n} \mathrm{I}_{\{X_k \leq x\}} = \frac{\text{nb de données} \leq x}{\text{nb de données}}$

Coordonnées d'un point sur le P-P plot : $\left(\frac{k}{n}; F\left(x_{(k)}\right)\right)$



- **Q-Q plot** : graphe des quantiles empiriques et théoriques

Ajustement empirique à une densité de probabilité

Test d'adéquations

Etape 1 : Définition de deux hypothèses H_0 contre H_1

H_0 : « les données suivent une loi donnée de fonction de répartition F »

H_1 : « les données ne suivent pas cette loi »

Etape 2 : Définition de la statistique de test

Etape 3 : Définition d'un niveau de confiance α et du risque de 1ère espèce

Etape 4 : Définition de la règle de décision

Tests usuels :

- *Test de Kolmogorov Smirnov*
- *Test d'Anderson-Darling*
- *Test de Cramer Von Mises*

Plus de détails dans la partie 4 : tests d'hypothèse

SOMMAIRE

Échantillonnage

Comment assurer la « représentativité » de l'échantillon pour estimer les statistiques d'une population à partir d'observations sur un échantillon ?

- Taille souvent fixée en pratique (à cause du coût, temps, …)

  Ex : n=10 (petit échantillon) ; n=1000 (grand échantillon)

- **Types d'échantillonnage :**

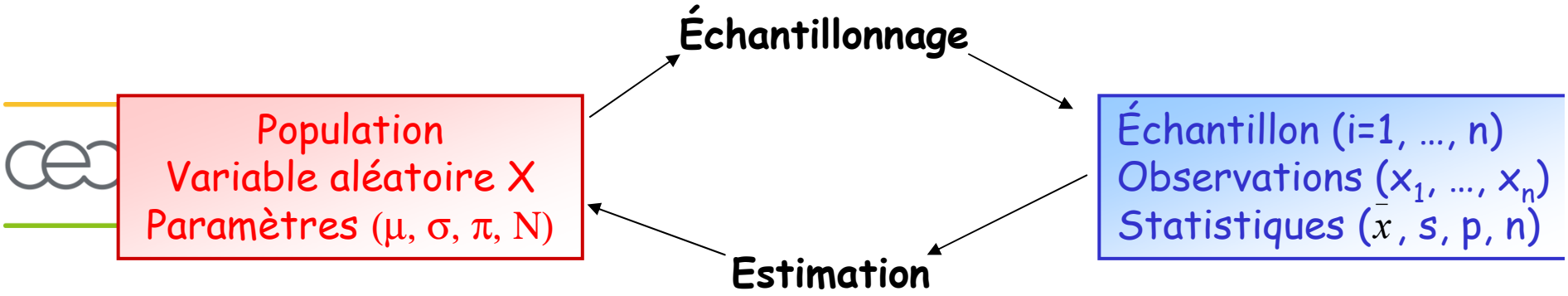
  - **Échantillonnage par choix raisonné** (méthode des quotas)
    Méthode déterministe, pas de mesure de la marge d'erreur.

  - **Échantillonnage aléatoire simple :** tirages équiprobables indépendants (i.i.d.).

  - **Échantillonnage stratifié :** découpage de la population en classes homogènes puis échantillonnage aléatoire simple dans chaque classe.

  - **Plans d'expériences :** on élabore des hypothèses sur le modèle et on cherche à extraire un maximum d'informations.

Échantillonnage et estimation



$(x_1, \ldots, x_n)$ est une réalisation de $(X_1, \ldots, X_n)$ v.a. i.i.d (de même loi « mère »)

Moyenne empirique : $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ réalisation de $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Variance empirique : $s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$ réalisation de $S_n^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X_n})^2$

$\bar{x}$ et s^2 sont des estimations ponctuelles de μ et σ^2

⇒ Construction d'un intervalle de confiance autour des estimateurs

SOMMAIRE

Convergence de variables aléatoires

- (X_n) = suite de v.a. définies sur un même espace probabilisé

Exemple : estimateur quand la taille de l'échantillon augmente

- **Convergence en loi (ou en distribution)** : $(X_n) \xrightarrow{\mathcal{L}} X$ si la suite des fonctions de répartition (F_{Xn}) converge vers F_X en tout point de continuité.

$$(X_n) \xrightarrow[n\to\infty]{\mathcal{L}} X \Leftrightarrow \lim_{n\to\infty} F_n(a) = F(a),\ \forall a \in R \text{ où F est continue}$$

- **Convergence en probabilité** : $(X_n) \xrightarrow[n\to\infty]{\mathcal{P}r} X \Leftrightarrow \forall \varepsilon > 0,\ \lim_{n\to\infty} P\left(\left|X_n - X\right| \geq \varepsilon\right) = 0$

- **Convergence presque sûre (ps)** : $(X_n) \xrightarrow[n\to\infty]{ps} X \Leftrightarrow P(\lim_{n\to\infty} X_n = X) = 1$

Propriétés :

$$(X_n) \xrightarrow{\mathcal{P}r} X \Rightarrow (X_n) \xrightarrow{\mathcal{L}} X$$

$$(X_n) \xrightarrow{ps} X \Rightarrow (X_n) \xrightarrow{\mathcal{P}r} X \Rightarrow (X_n) \xrightarrow{\mathcal{L}} X$$

Convergence de variables aléatoires

➢ Loi des grands nombres :

Soit (X_n) suite de v.a. indépendantes et de même loi d'espérance μ.

Alors la suite des v.a. $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{ps} \mu$

Remarques :

- Cela justifie l'estimation d'une espérance par une moyenne empirique.
- Cela justifie aussi l'estimation d'une probabilité par une proportion.

➢ Théorème central limite (TCL):

Soit (X_n) suite de v.a. indépendantes et de même loi (d'espérance μ et de variance σ^2 finies). Alors $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(\mu, \frac{\sigma^2}{n})$

Remarques :

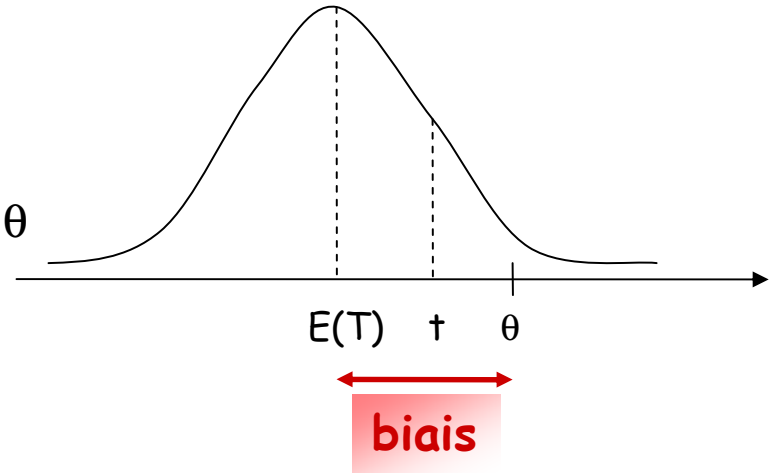
- Cela explique l'importance de la loi normale dans la nature et son usage abondant : loi de phénomènes qui résultent de l'addition de phénomènes identiques et indépendants.
- Formule asymptotique, n doit être très grand dans certains cas.

Exemple : loi des erreurs (Laplace, 1810), répartition des moyennes de séries de mesures

SOMMAIRE

Estimateurs

Soit T_n estimateur de θ



- **Estimateur sans biais :** $E(T_n) = \theta$

- **Estimateur convergent :** $T_n \xrightarrow{L^2} \theta \Leftrightarrow E(T_n - \theta)^2 \xrightarrow{n\to\infty} 0$

Si T_n est sans biais, alors $E(T_n - \theta)^2 = var(T_n)$

- **Estimateur efficace :**

Soient T_n et V_n estimateurs sans biais ;
T_n est plus efficace que V_n si $var(T_n) < var(V_n)$

Méthode des moments : moyenne empirique

$(X_1, \ldots, X_n)$ v.a. i.i.d telles que $E[X_i]=\mu$ et $var(X_i)=\sigma^2$

- **Estimateur de la moyenne :** $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- **Loi des grands nombres :** $\overline{X_n} \xrightarrow[n\to\infty]{ps} \mu$
  - **Propriétés :** $\mathrm{E}\left[\overline{X_n}\right] = \mu \quad \text{et} \quad Var\left(\overline{X_n}\right) = \frac{\sigma^2}{n}$
  - **Cas gaussien :** $\overline{X_n} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$
  - **Cas général : Théorème central limite** $\Rightarrow \frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$

La méthode des moments a pour but d'estimer les paramètres d'une loi

Exemple : loi exponentielle , $\lambda \exp(-\lambda x)$, $\hat{\lambda}_n = \frac{1}{\overline{X_n}}$

Méthode des moments : variance empirique

$(X_1, \ldots, X_n)$ v.a. i.i.d telles que $E(X_i)=\mu$ et $var(X_i)=\sigma^2$

- **Estimateur de la variance :** $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$
- **Loi des grands nombres :** $S_n^2 \xrightarrow[n\to\infty]{ps} \sigma^2$

Propriété: $\mathrm{E}(S_n^2) = \frac{n-1}{n}\sigma^2$ *(⇔ estimateur biaisé)*

- **Autre estimateur de la variance :** $S_n^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$

$\mathrm{E}(S_n^{*2}) = \sigma^2$ ⇒ estimateur non biaisé

- **Cas gaussien :** $n\frac{S_n^2}{\sigma^2} \sim \chi^2(n-1)$
- **Cas général :** **Théorème central limite**

Remarque : en théorie, cette méthode est applicable à tous moments.

Méthode de maximum de vraisemblance

$(X_1, \ldots, X_n)$ v.a. i.i.d dont la loi mère dépend d'un coefficient $\theta \in \Theta \subseteq \Re^k$

- **Fonction de vraisemblance :**
  - Si X discrète : $L(x_1,\ldots,x_n,\theta) = P(X_1 = x_1,\ldots,X_n = x_n,\theta) = \prod_{i=1}^{n} P(X_i = x_i,\theta)$
  - Si X continue : $L(x_1,\ldots,x_n,\theta) = f(x_1,\ldots,x_n,\theta) = \prod_{i=1}^{n} f(x_i,\theta)$

- **Estimateur du max de vraisemblance (EMV) :** $T_n = \underset{\theta}{\operatorname{argmax}}\, L(x_1,\ldots,x_n,\theta)$

  $\left(\Leftrightarrow \frac{\partial}{\partial\theta} L(x_1,\ldots,x_n,T_n) = 0 \right)$

  - Propriétés :
    - $T_n \xrightarrow{ps} \theta$
    - En général, l'EMV est plus efficace que celui estimé par la méthode des moments.

- *Application pour la loi normale :*

$$X \sim \mathcal{N}(\mu,\sigma^2) \Rightarrow L(x_1,\ldots,x_n,\theta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$\Rightarrow$ *Estimateurs du maximum de vraisemblance :*

$$\hat{\mu}_{EMV} = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\sigma}^2{}_{EMV} = S_n^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

Taille de l'échantillon

- **Inégalité de Bienaymé-Tchebyshev :**

  Toute variable X de moyenne μ et de variance σ^2 satisfait à :

$$P\left(|X-\mu| \geq \varepsilon\right) \leq \frac{\sigma^2}{\varepsilon^2}$$

Cette formule évalue les probabilités des écarts à la moyenne.

Application à $\overline{X_n}$:

$$P\left(|\overline{X_n}-\mu| < \varepsilon\right) > 1-\frac{\sigma^2}{n\varepsilon^2}$$

Niveau de confiance

On choisit ε et un niveau de confiance (*par ex. 0.95*)

$\Rightarrow$ On en déduit *n*.

➡ Permet de déterminer la taille nécessaire (de l'échantillon) pour avoir 95 % de chance que l'écart entre la moyenne empirique et la moyenne réelle soit faible (inférieur à ε)

Intervalle de confiance d'une estimation

$$P(T - e \leq \theta \leq T + e) = 1 - \alpha$$

- T = estimateur de θ
- e = marge d'erreur
- 1-α = niveau de confiance
- α = probabilité d'erreur

Exemple : IC d'un estimateur $\overline{X_n}$ de la moyenne d'une loi $\mathcal{N}(\mu, \sigma^2)$

$$T = \overline{X_n} \text{ et } \theta = \mu$$

❖ **σ connu :** $e = \frac{\sigma}{\sqrt{n}} u_\alpha$ tel que $\mathrm{P}(|U| > u_\alpha) = \alpha$ avec $U \sim \mathcal{N}(0,1)$

❖ **σ inconnu :** $e = \frac{S_n'}{\sqrt{n}} t_{n-1,\alpha}$ tel que $\mathrm{P}(|Y| > t_{n-1,\alpha}) = \alpha$ avec $Y \sim St(n-1)$

Méthodes paramétriques

La distribution de la variable aléatoire étudiée *X* est inconnue

Echantillon de grande taille (n>30)

➡ Ajustement d'un modèle probabiliste

Méthode des moments, Max. de vraisemblance

➡ D'après TCL, la moyenne empirique suit une loi normale de moyenne μ et d'écart-type σ^2/n telle que

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$ suit une loi N(0,1).

Calcul moyenne, écart-type, fractiles

Tables statistiques

Fonctions logicielles (Matlab, SAS, Statgraphics…)

Méthodes paramétriques

La distribution de la variable aléatoire étudiée *X* est inconnue et/ou petit échantillon

Théorème de Bienaymé-Tchebitchev :
Pour X_n de moyenne et d'écart-type finis μ et σ

$$\mathrm{P}\left(\left|\overline{X_n} - \mu\right| < \varepsilon\right) > 1 - \frac{\sigma^2}{n\varepsilon^2}$$

Commentaire :

- Cela permet d'avoir une limite de confiance « conservative » pour la moyenne de l'échantillon
- Nécessite de connaître σ

SOMMAIRE

Méthodes non paramétriques

La distribution de la variable aléatoire étudiée *X* est inconnue et/ou petit échantillon

➡ Méthodes de rééchantillonnage :
Jackknife, Bootstrap

Construction de répliques par tirage aléatoire avec ou ou sans remise dans l'échantillon disponible

Calcul d'estimateurs de paramètres et un intervalle de confiance associé à partir des répliques

Commentaire :

Méthode surtout robuste pour la moyenne et l'écart-type

Méthodes non paramétriques

La distribution de la variable aléatoire étudiée *X* est inconnue et/ou petit échantillon

➡ **Méthode de Wilks**

$(X_1, \ldots, X_N)$ v.a. i.i.d dont la loi, $X_{max} = \max\{X_1, \ldots, X_N\}$ et $X_{min} = \min\{X_1, \ldots, X_N\}$

Formule pour fractile unilatéral supérieur

$$P\left[P\left(X \le X_{\max}\right) \ge \alpha\right] \ge \beta,$$

$$N \text{ solution de } 1-\alpha^N \ge \beta$$

A.N. : pour α=β=0.95, N=59

Formule pour fractile bilatéral

$$P\left[P\left(X_{\min} \le X \le X_{\max}\right) \ge \alpha\right] \ge \beta,$$

$$N \text{ solution de } 1-\alpha^N - N(1-\alpha)\alpha^{N-1} \ge \beta$$

Méthodes non paramétriques

Commentaire :

- Méthode permettant de calculer :
  - *N*, la taille de l'échantillon minimal nécessaire
  - La Valeur du fractile
- Méthode robuste
- S'applique à tout type de distribution même multimodale ou discontinue
- Résultat conservatif

Tableau des tailles d'échantillons minimales pour un α-fractile unilatéral au niveau de confiance β

α	0.50	0.90	0.90	0.95
β	0.95	0.90	0.95	0.90
N	5	22	29	45

SOMMAIRE

1. Définitions et rappels de probabilités
2. Analyse descriptive unidimensionnelle
3. Estimations paramétriques et non paramétriques
4. **Tests d'hypothèse**
5. Plan d'expériences
6. Régréssion linéaire

Tests d'hypothèses

Mécanisme qui permet de trancher entre 2 hypothèses H_0 et H_1 (dont une seule est vraie) au vu des résultats d'un échantillon.

Vérité / Décision	Hypothèse nulle H_0	Hyp. alternative H_1
H_0	$1-\alpha$	β (erreur de 2ème espèce)
H_1	α (erreur de 1ère espèce) Région critique : $P(W \mid H_0)=\alpha$	$1-\beta$

- On fixe α ; H_0 est l'hypothèse prééminente (solide, prudente, facile, …)
  Exemple : H_0 = l'accusé est innocent ; H_1 = l'accusé est coupable
- On détermine la région critique en fonction de α
- On situe les observations par rapport à la région critique
- On rejette ou pas H_0

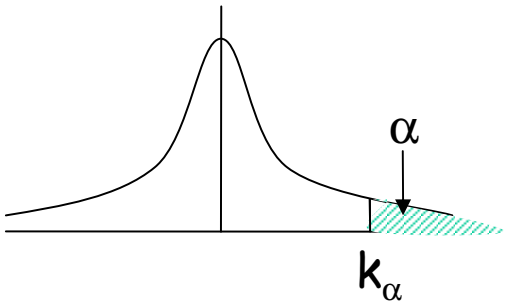
β dépend de H_1 et est le résultat d'un calcul (puissance du test)

Tests paramétriques sur un échantillon

On teste θ, paramètre de la loi de probabilité de la v.a. X

Exemple : valeur moyenne d'une loi normale (σ connu)

- **Hypothèses :** $H_0 : \mu=\mu_0$; $H_1 : \mu>\mu_0$ (test unilatéral)
- **Variable de décision :** $\overline{X_n} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$
- **Région critique :** On rejette H_0 si $\overline{X_n} > k_\alpha$



- **Décision de rejet ou pas de H_0 :**
  On trouve *k* dans les tables statistiques (ou les logiciels) par :

$$\mathrm{P}(\overline{X_n} > k_\alpha | \mu_0) = \mathrm{P}\left(U > \frac{k_\alpha - \mu_0}{\sigma/\sqrt{n}}\right) = \alpha \text{ avec } U \sim \mathcal{N}(0,1)$$

Remarques :

- *En général tous les tests paramétriques sont basés sur la loi normale.*
- *Grâce au théorème central limite, ces tests peuvent fonctionner avec d'autres lois pour de grandes tailles d'échantillon.*

Tests statistiques d'adéquation à une loi

Avec un certain niveau de confiance (par ex. 95 %), **on rejette ou on ne rejette pas l'hypothèse que l'échantillon suive une certaine loi.**

➢ **Test du Chi-deux :** $D^2 = \sum_{i=1}^{k} \left(\frac{N_i - np_i}{np_i} \right)^2 \sim \chi^2(k-1)$ où k est le nb de classes

Comparaison entre fréquences observées N_i et théoriques p_i

Test peu puissant et non robuste pour de petits échantillons (n < 50)

➢ **Tests basés sur la fonction de répartition empirique F_n :**

- **Kolmogorov-Smirnov :** $K_n = \sqrt{n} \sup_{x \in \Re} |F_n(x) - F(x)|$
- **Cramer-Von Mises :** $W_n^2 = n \int_{-\infty}^{+\infty} [F_n(x) - F(x)]^2 dF(x)$
- ...

Tests puissants, hypothèse de distribution continue.

SOMMAIRE

1. Définitions et rappels de probabilités
2. Analyse descriptive unidimensionnelle
3. Estimations paramétriques et non paramétriques
4. Tests d'hypothèse
5. **Plan d'expériences**
6. Régression linéaire

Plan d'expériences

➢ **Définir un plan d'expériences :** placer les points d'expérimentation ou de simulation dans le domaine de variation des paramètres incertains

⇒ **Optimiser l'information requise avec le moins de points possible**

➢ **Etablir les liens entre :**

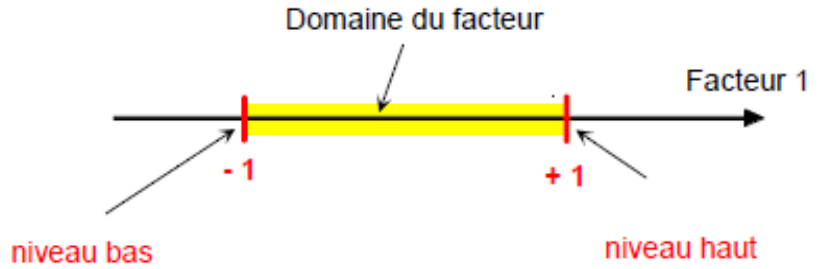
- **Réponse :** grandeur physique étudiée
- **Facteurs :** grandeurs physiques modifiables par l'expérimentateur ou le simulateur sensées influer sur les variations de la réponse
  - Différentes nature : continus, discrets ou qualitatifs
  - Domaine de variation : [borne inf ; borne sup] ⇒ discrétisation en niveaux

➢ **Différents objectifs :**

- **Recherche exploratoire :** investigation du domaine pour identifier les régions d'intérêt
- **Screening des facteurs :** identification des facteurs potentiellement influents et ceux non influents ⇒ simplification du modèle
- **Etude quantitative des facteurs :** identifier les effets des facteurs et leurs interactions
- **Optimisation**

Plan d'expériences

- **Hypothèses :**
  - K facteurs
  - 2 niveaux pour chaque facteur



- **Plan factoriel complet**
  - **Principe statistique d'orthogonalité**

    Variation de chaque facteur lorsque les autres facteurs sont fixés successivement à leurs 2 valeurs possibles.

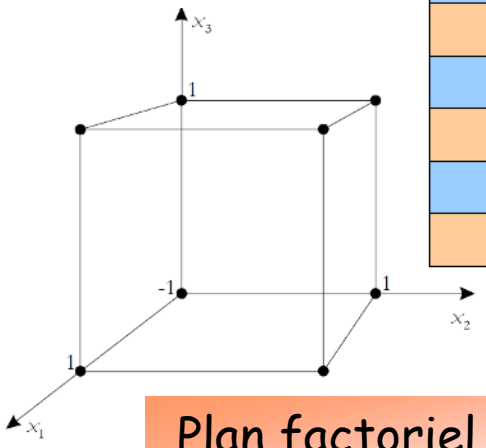
    $\Rightarrow$ *2^k expériences à réaliser.*

  - **Utilisable pour facteurs continus ou discrets.**

  - **Problème :** Nombre d'expériences trop important si K devient grand et si le nombre de niveaux augmente.

Ex : 10 facteurs $\Rightarrow$ 1024 expériences

X_1	X_2	X_3
–1	–1	–1
1	–1	–1
–1	1	–1
1	1	–1
–1	–1	1
1	–1	1
–1	1	1
1	1	1

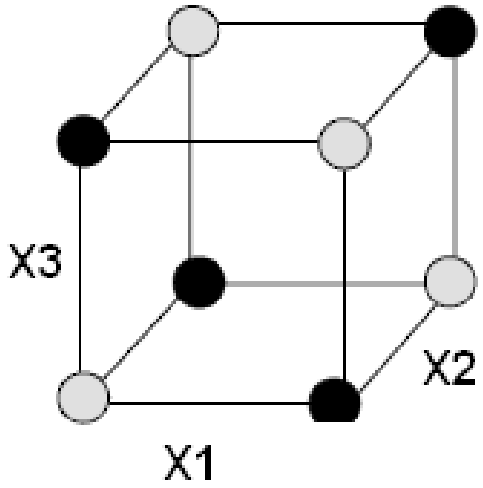


Plan factoriel à 2 niveau pour k=3 facteurs

Plan d'expériences

➢ Plan factoriel fractionnaire

- ❖ **Etude de tous les facteurs avec nombre réduit d'expériences par rapport au plans complets**
- ❖ **Fraction d'un plan complet**
  ⇒ 2^{k-q} expériences à réaliser
- ❖ **Sélection de cette fraction?**
  ⇒ Choix d'une structure d'alias
  ⇒ détermine quels effets sont confondus



Plan factoriel complet 2^3 décomposé en 2 plans factoriels fractionnaires 2^{3-1} (noir et blanc)

Plan d'expériences

Plan factoriel fractionnaire pour k=5 facteurs et q=2

❖ **Plan à 2^{5-2} = 8 expériences :**

- Plan complet à 3 facteurs pour (X_1, X_2, X_3)
- Effets de X_4 confondus avec interaction X_1X_2
- Effets de X_5 confondus avec interaction X_1X_3

  $\Rightarrow$ 3 alias de 1
  - $X_1 X_2X_4 = 1$
  - $X_1 X_3X_5 = 1$
  - $X_2X_3X_4X_5 = 1$

❖ **Résolution r :**

- **r = nombre minimal d'éléments de l'alias de 1**
  = cardinal du plus petit générateur d'alias

Exemple : ici r = III

❖ **Un plan de résolution r ne confond pas les effets d'ordre s1 et s2 avec s1 + s2 < r**

X_1	X_2	X_3	$X_4 = X_1 X_2$	$X_5 = X_1 X_3$
−1	−1	−1	1	1
1	−1	−1	−1	−1
−1	1	−1	−1	1
1	1	−1	1	−1
−1	−1	1	1	−1
1	−1	1	−1	1
−1	1	1	−1	−1
1	1	1	1	1

Plan factoriel fractionnaire 2^{5-2}

Plan d'expériences

Résolution des plans factoriels fractionnaires

- **Résolution III :** tous les effets principaux sont non confondus.
- **Résolution IV :** un effet principal ne peut être confondu avec une interaction, mais deux interactions peuvent être confondues.
- **Résolution V :** on peut poser un modèle avec toutes les interactions et effets principaux sans confusion.

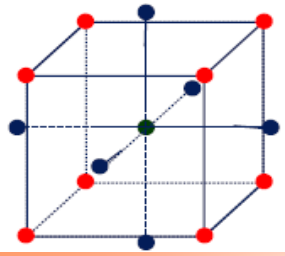
Nombre d'expériences (N) du plan fractionnaire	Nombre de facteurs (k)									
	3	4	5	6	7	8	9	10	11	
4	III									
8		IV	III	III	III					
16			V	IV	IV	IV	III	III	III	
32				VI	IV	IV	IV	IV	IV	r=7
64					VII	V	IV	IV	IV	r=6
128						VIII	VI	V	V	r=5
256							IX	VI	V	r=4
512								X	VII	r=3
1024									XI	r=2
										r=1

La résolution V est considérée comme suffisante dans toutes les situations.
La résolution III est considérée comme une propriété minimale.

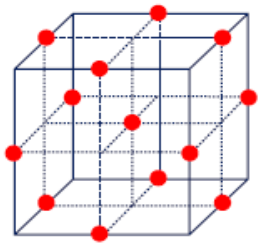
Plan d'expériences

Autre plans

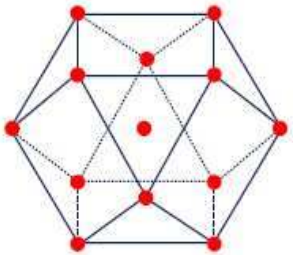
- **Plans de Packett-Burman :** matrice de Hadamard ⇒ plan de résolution III
- **Plans de Taguchi :** Plans de Plackett-Burman ou fractionnaires modifiés.
- **Plans de Koshal :** Peu connus, modèle sans interaction, utiles pour dégrossir.
- **Plans supersaturés avec effets principaux aliasés :** si beaucoup de facteurs ou très peu d'expériences possibles
- **Plans composites centrés**
- **Plan Box-Behnken**
- **Plan de Doehlert**
- **Plans de Rechtschaffner**
- **Plans D-Optimaux, …**



Plan composite centré



Plan Box-Behnken



Plan de Doehlert

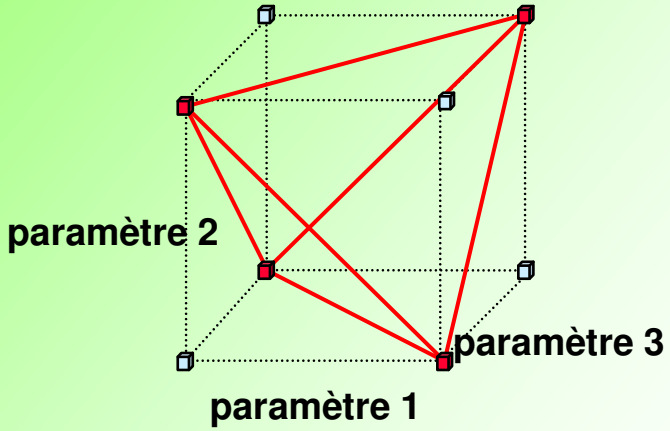
Expériences réelles / Expériences numériques

Plans pour expériences réelles

Estimer les paramètres de la régression linéaire avec le moins de calculs possible

Exemples :

- Plan factoriel complet 2^3
- Plan factoriel fractionnaire 2^{3-1}



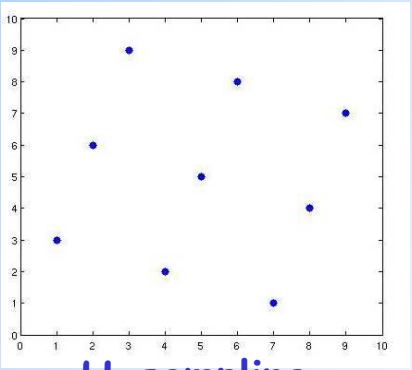
Biblio : Fisher (1917), Box et Wilson (1954), Taguschi (1960), Mitchell (1958), …

Plans pour expériences numériques

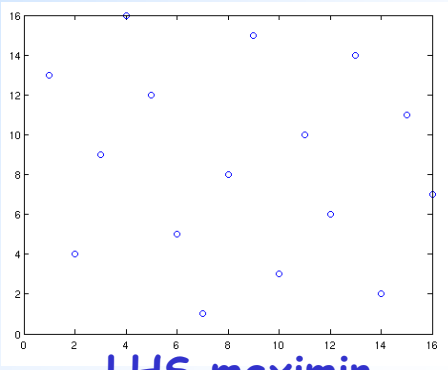
Spécificités

- expériences déterministes,
- grand nombre de variables d'entrées,
- larges domaines de variation,
- variables d'intérêt multiples,
- modèles fortement non linéaires, …

➡ **Space filling designs** : répartition uniforme dans l'espace des entrées



U-sampling



LHS maximin

Biblio : Kleijnen (1970), McKay (1979), Morris(1995), Sacks (1989), …

SOMMAIRE

1. Définitions et rappels de probabilités
2. Analyse descriptive unidimensionnelle
3. Estimations paramétriques et non paramétriques
4. Tests d'hypothèse
5. Plan d'expériences
6. **Régression linéaire**

Régression linéaire

➢ **Hypothèses :**

Y : variable à expliquer

X_1, … X_p : p variables explicatives (ou prédicteurs)

On suppose un modèle linéaire entre Y et X :

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon$$

- **Avec ε résidu aléatoire tel *que* : E[ε] = 0**
- **Avec β_j paramètres du modèle de régression**

❖ **Application à un échantillon :**

- N données : $(y_i, x_{i1}, \ldots, x_{ip})$ pour i=1,…,N

Modèle de régression linéaire :

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_j \text{ pour } i = 1,\ldots,N$$

ε_1, …, ε_N sont des variables aléatoires indépendantes et identiquement distribuées (i.i.d.) de moyenne nulle et de variance σ^2

❖ **Notation vectorielle :** $Y = X\beta + \varepsilon$ avec $X = [1\ X_1 \ldots X_p]$ et $\beta = [\beta_0\ \beta_1 \ldots \beta_p]'$

Régression linéaire

➢ Estimation des paramètres par moindres carrés :

❖ Paramètres de la régression β et σ

Estimation par moindres carrés : $\beta^* = Arg\ \min \left\| Y - X\beta \right\|_2$

$$\beta^* = (X^T X)^{-1} X^T Y$$

Rque : équivalent à l'EMV lorsque les erreurs ε sont i.i.d. gaussiennes.

Propriétés :

- Estimateur sans biais : E[β*] = β
- Estimateur de variance minimale parmi les estimateurs de la forme BY

❖ Prédicteur pour la sortie Y :

$$Y^* = X\beta^* = X(X^T X)^{-1} X^T Y = H\,Y$$
$$\text{avec } H = X(X^T X)^{-1} X^T$$

- Estimateur sans biais de Y : E[Y*] = Y

❖ Paramètres de variance σ² :

$$\sigma^{2^*} = \frac{\left\| Y - Y^* \right\|_2}{n - p - 1}$$

- Estimateur sans biais : E[σ²*] = σ²

Régression linéaire

- **Coefficient de détermination :**
  - **Sum of Squared Errors (SSE) :** $SSE = \left\| Y - Y^* \right\|_2$
  - **Total Sum of Squared (SST) :** $SST = \left\| Y - \bar{Y}1 \right\|_2 = \sum_{i=1}^{N} \left(Y_i - \frac{1}{N} \sum_{i=1}^{N} Y_i \right)^2$
  - **Regression Sum of Squared (SSR) :** $SSR = \left\| Y^* - \bar{Y}1 \right\|_2$

Propriété : SST = SSR+SSE

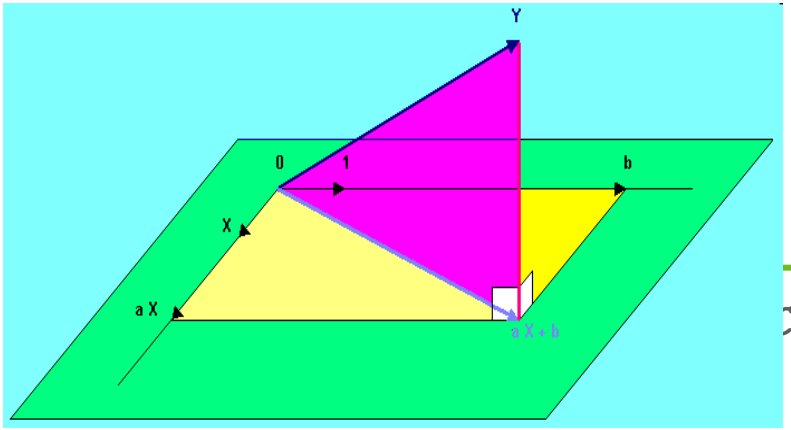
  - **Coefficient de détermination R^2 :** $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

➡ Part de variance expliquée par la régression

Rque : si N = p+1 $\Rightarrow$ R^2 =1

  - **Coefficient de détermination ajusté R^{2*} :**

$$R^{2^*} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)} = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

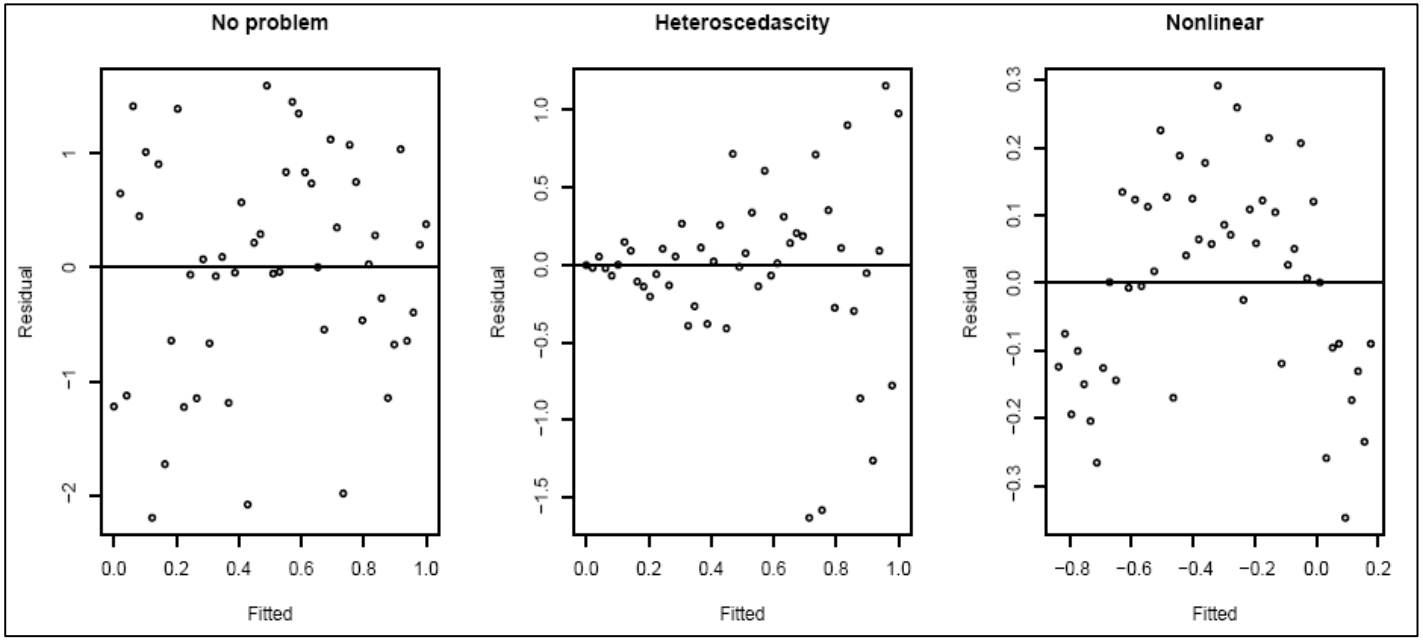


Régression linéaire

- **En pratique :**
  - **Diagnostic :**
    - Vérification des hypothèses, linéarité, normalité, données aberrantes …
  - **Transformation :**
    - Transformation de la réponse (Box-Cox)
    - Transformation des prédicteurs
    - Régression polynomiale …
  - **Sélection de variables :**
    - Stepwise procedures
    - Etude de critères AIC, BIC, Cp de Mallow

Régression linéaire

- **En pratique :**
  - **Diagnostic : étude des résidus ($Y_i - Y_i^*$)**
    - Vérification des hypothèses, linéarité, normalité, données aberrantes …



Autres méthodes : Etudes des leviers, Tests statistiques …

Régression linéaire

➢ **En pratique :**

❖ **Transformation de la réponse (Box-Cox, 1964):**

$$h_\lambda(Y) = X\beta + \varepsilon$$

$$\text{avec} \quad h_\lambda(Y) = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda} & \text{si} \quad \lambda \neq 0 \\ \ln(Y) & \text{si} \quad \lambda = 0 \end{cases}$$

⇒ **Estimation de λ**

❖ **Transformation des prédicteurs**

Visualisation graphique de Y en fonction de X_i et des résidus en fonction de X_i

Régression linéaire

- **En pratique :**
  - **Sélection de variables**

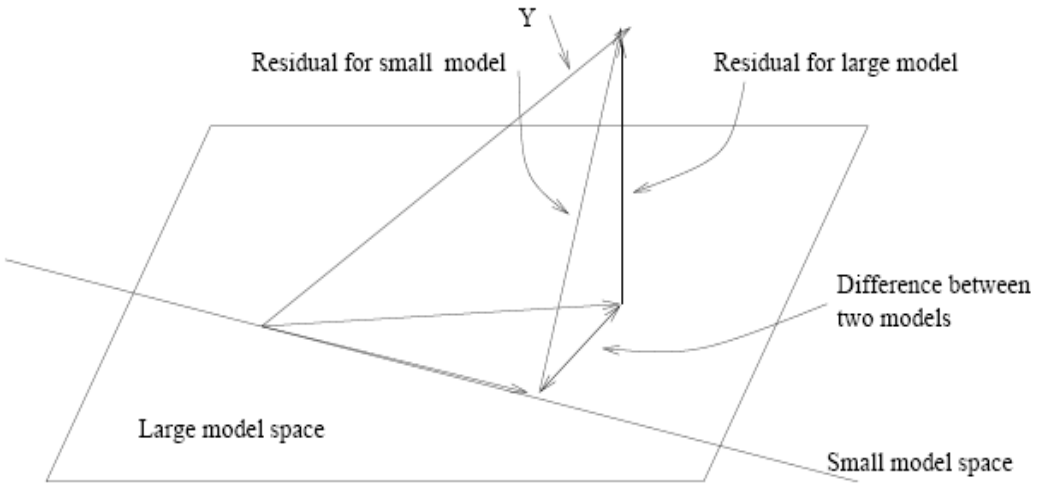
    Tests de comparaison de modèles (ANOVA) :

    - on compare **un « gros modèle » Ω**

    ex : modèle linéaire par rapport à l'ensemble des variables d'entrée

    - **à un de ses sous modèles ω**

    ex : modèle linéaire par rapport à certaines des variables d'entrée



Propriété

$$F = \frac{(SSR_\omega - SSR_\Omega)/(q-p)}{SSR_\Omega/(n-p)} \sim F_{q-p,n-q}$$

Pour la sélection de modèle, il existe des méthodes consistant à **minimiser certains critères**
ex : **AIC, BIC, *Cp de Mallow***

Régression linéaire

- **Difficultés de mise en œuvre :**
  - Choix du modèle de régression ?
  - Hypothèses d'indépendance et de bruit gaussien
    $\Rightarrow$ pas toujours possible à corriger
  - Fléau de la dimension

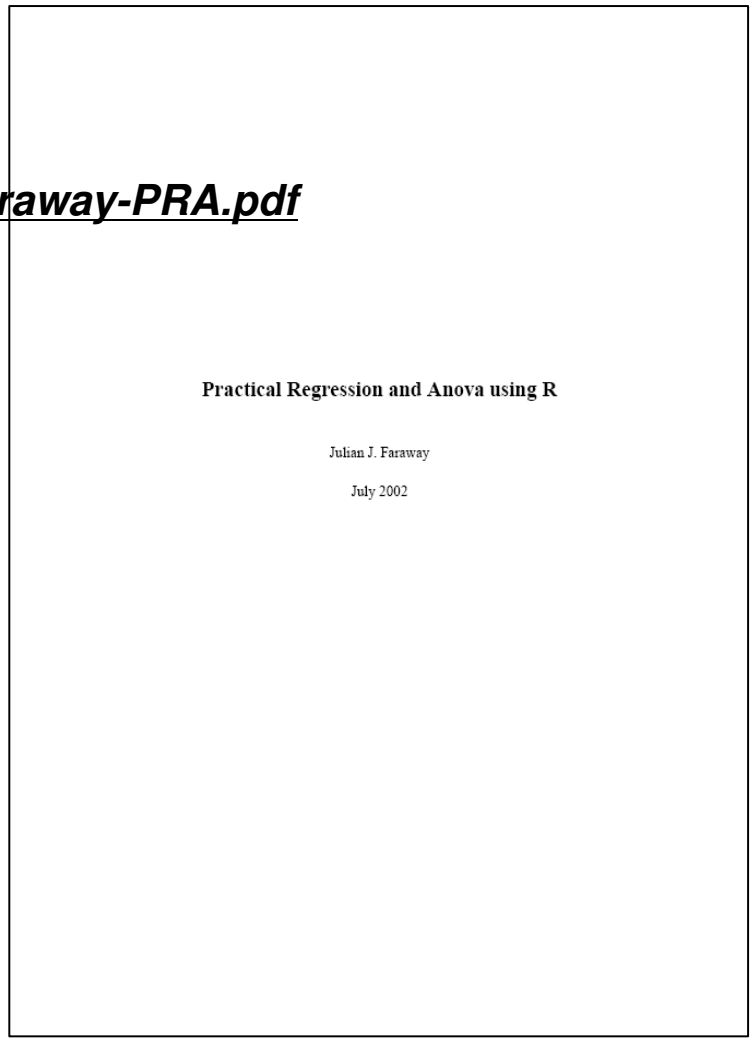
- **Avantages :**
  - Simplicité !
  - « Interprétation » du modèle obtenu
  - Techniques associées très développées : analyses statistiques, intervalle de prédiction, sélection de variables avec justification théorique …

- **Remarques générales :**
  - Méthode souvent employée de façon trop « rudimentaire » sans exploiter au mieux l'éventail des techniques associées
  - Méthode simple et fournissant des résultats satisfaisants dans la grande majorité des cas

Régression linéaire

- Référence

- ***http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf***



Autre modèle statistique : Processus Gaussien (PG)

➢ **Définition :**

Un Processus Gaussien est un processus aléatoire réel $\{Y(x)\}_{x \in D \subset R^d}$
dont toutes ses lois finies-dimensionnelles $(Y(x_1), \ldots, Y(x_n))$ *sont gaussiennes*

$$Y(x) \sim PG(\, m(x),\, C(x,x')\,)$$

où $m(x) = \mathrm{E}(x)$ et $C(x,x') = \mathrm{E}[(\, Y(x)-m(x)\,)\,(\, Y(x)-m(x'))]$

- Approche similaire : **krigeage** ⇒ conduit au même modèle

– **Différentes hypothèses de modélisation :**

Les sorties correspondent à des observations de la trajectoire d'un PG, dont la fonction de covariance vérifie : $C(x,x') = C(x - x')$ et la moyenne

- $m(x) = m$ avec m connue pour le *Krigeage Simple*
- $m(x) = m$ avec m inconnue pour le *Krigeage Oridinaire*
- $m(x) = f(x)\,\beta$ avec β inconnue pour le *Krigeage Universel*

Autre modèle statistique : Processus Gaussien (PG)

- **Hypothèse Classique :** $Y(x) = f(x)^t\beta + Z(x)$

  Avec Z, PG stationnaire tel que $E[Z(x)] = 0$ et $C(x,x') = \sigma^2 R(x-x')$

- **Estimation en x_0 à partir de n observations $(Y(x_1), \ldots, Y(x_n))$:**

  $Y^*(x_0) = E[\, Y(x_0) \mid (Y(x_1), \ldots, Y(x_n))\,]$

$$Y^*(x_0) = f(x_0)^t \beta + r(x_0)^t R^{-1} \left[Y - f(x_0)\beta \right]$$

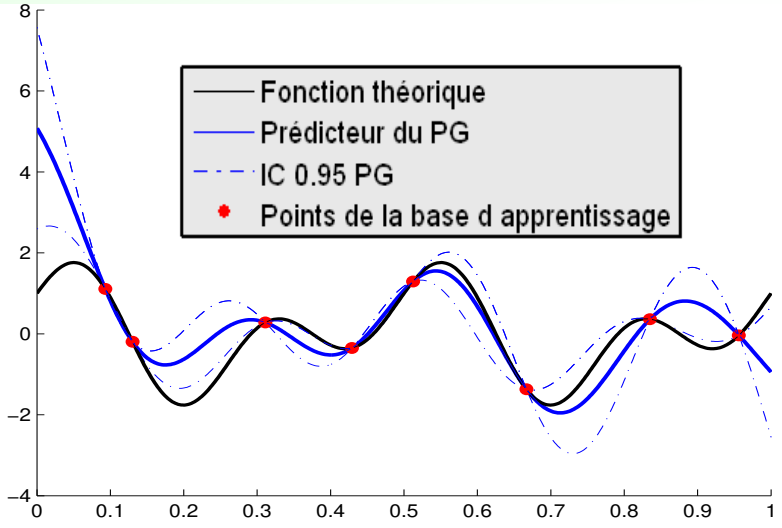
avec $r(x_0) = [\, R(x_1, x_0), \ldots, R(x_N, x_0)\,]$
et $R = (R(x_i, x_j))_{i,j}$

« Tendance déterministe »
Terme identique à celui des MC classiques

« Partie permettant l'interpolation »
Prise en compte de la configuration des données

Propriétés du prédicteur $Y^*(x_0)$:

- **Interpolateur exact des observations**
- **Sans biais et de variance minimale**



Autre modèle statistique : Processus Gaussien (PG)

- **Difficultés de mise en œuvre :**
  - Choix de la fonction de covariance ?
  - Estimations des paramètres
  - Plan d'expériences ? (Space filling Design)
  - Fléau de la dimension (d > 10)

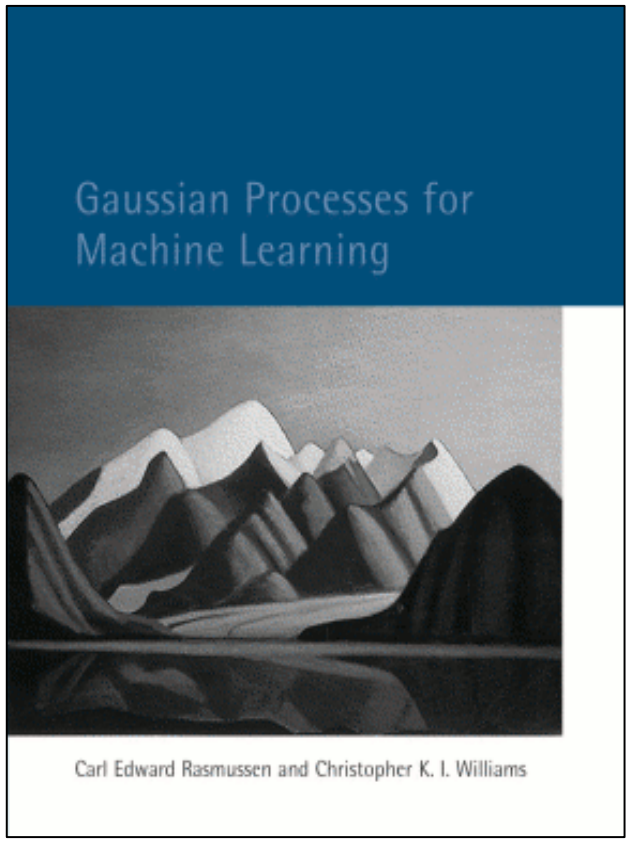
- **Avantages :**
  - Calcul des indices de Sobol possible car évaluation très rapide du prédicteur
  - Cadre statistique, expression analytique, calcul analytique (distribution de la sortie, bandes de confiance …)

- **Remarques générales :**
  - Interprétation du prédicteur délicate. La partie permettant l'interpolation « corrige » l'erreur de la partie déterministe.

Autre modèle statistique : Processus Gaussien (PG)

- Références

- *http://cran.r-project.org/web/packages/DiceKriging/index.html*



http://www.gaussianprocess.org/gpml/

Remerciements et Références

Contributeurs :

- Amandine Marrel (CEA), Nadia Perot (CEA), Marc Sancandi (CEA).
- Bertrand looss (EDF R&D).
- Vincent Feuillard (EADS).

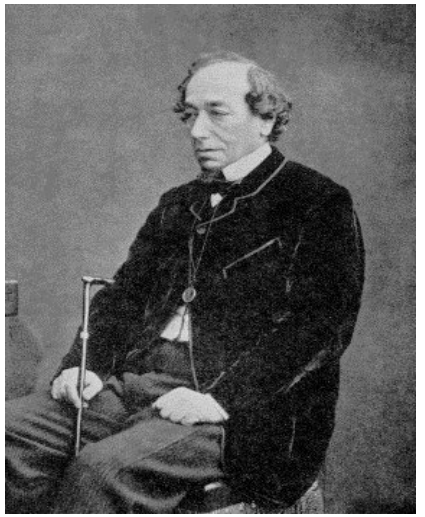
- Y. Dodge, Premiers pas en statistique, Springer, 2001
- G. Saporta, Probabilités, Analyse des données et Statistique, Ed.Technip, 1990
- M. Lejeune, Statistique : la théorie et ses applications, Springer Verlag, 2004
- Formation Incertitudes IMdR-LNE
- Cours d'O. Gaudoin : http://www-lmc.imag.fr/lmc-sms/Olivier.Gaudoin/
- Cours de P. Besse : http://www.lsp.ups-tlse.fr/Besse
- Cours de P. Leray : http://asi.insa-rouen.fr/~pleray/ftp/
- Cours de J. GOUPY : http://www-rocq.inria.fr/axis/modulad/archives/numero-34/Goupy-34/goupy-34.pdf
- Présentation de F. Campolongo, Screening methods in sensitivity analysis, SAMO Fiesole, 2010: http://sensitivity-analysis.jrc.ec.europa.eu/Events/SAMO2010_Fiesole/

Citations

Les mensonges des statistiques

"Il existe trois types de mensonges: les mensonges, les parjures et les statistiques !"

Benjamin D'Israeli



« Les statistiques sont comme les minijupes, elles cachent l'essentiel, mais donnent (parfois) de mauvaises idées »

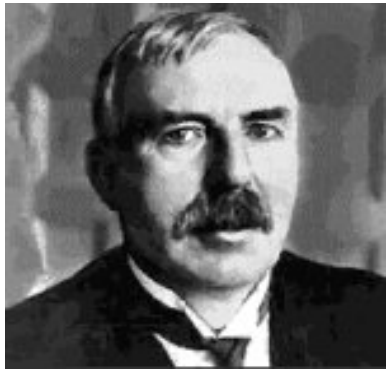
Roger Phan-Tan-Luu

Citations

Les statistiques sont indispensables ?

"Si des statistiques sont nécessaires pour interpréter une expérience, ce n'est pas une bonne expérience"

Ernest Rutherford



"Appeler un statisticien après que l'expérience soit terminée c'est comme lui demander de faire une autopsie; il pourra seulement déterminer la cause de l'échec de l'expérience"

Sir Ronald Fisher