

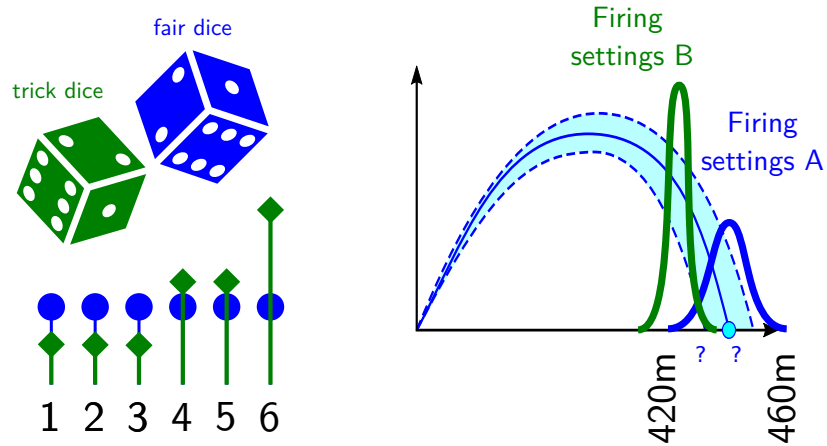
(outils pour la) **Modélisation de l'incertitude**

Formation - Quantification des incertitudes dans les simulations

numériques — Novembre 2024

William Fauriat (ce cours) + Équipe statistiques et incertitudes (formation)

CEA / DAM



Quelle incertitude ?

Quels outils pour la décrire ?

Quels issues possibles ?

Sommaire

1 Outils de la “théorie” des probabilités

2 Science statistique et inférence

3 Choix et tests statistiques





1.

Outils de la “théorie” des probabilités

Définition et propriétés mathématiques de la mesure de probabilité

Soient :

- Ω l'ensemble des **issues possibles** ou Univers
- E_i des **événements** dans un espace d'événements possibles noté \mathcal{F} (des sous-parties de Ω)

$p : E_i \in \mathcal{F} \mapsto p_i \in [0, 1]$ est la **mesure de probabilité**

On note : $p_i = \Pr(E_i)$ la **probabilité d'occurrence de l'événement** E_i

On dira aussi : la probabilité que E_i se réalise - ou - la probabilité que E_i soit "vrai" - ou - la probabilité que E_i soit observé

Le cadre formel (axiomatique de **Kolmogorov**) donne les **propriétés fondamentales** suivantes :

$$0 \leq p_i \leq 1, \text{ pour tout } E_i \in \mathcal{F} \quad (\text{i})$$

$$\Pr(\Omega) = 1 \quad (\text{ii})$$

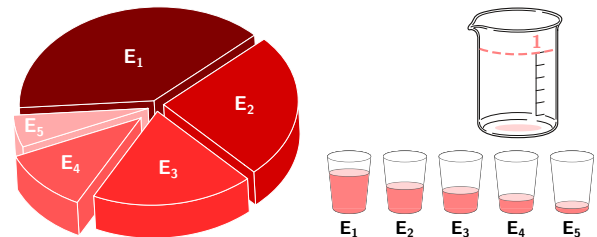
$$\Pr(\cup_{i=1}^k E_i) = \sum_{i=1}^k \Pr(E_i), \quad (\text{iii})$$

pour tout $(\cup_{i=1}^k E_i)$, un ensemble d'événements *mutuellement exclusifs*

Ceci revient à **distribuer** de la "masse" (en totalité) entre des états (ou valeurs, ou issues) possibles (*un seul à la fois*) :

comme par exemple :

- en partageant un gâteau
- en servant un jus de fruits

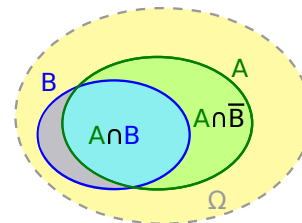


Concept de probabilité conditionnelle

Définition de la probabilité conditionnelle

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)}$$

$\Pr(B|A)$ se lit : **probabilité de B sachant A**
où A et B sont deux événements, ou **propositions**
(pouvant être vraies ou fausses)



$\Pr(B|A)$ est un **ratio** de masses de probabilité, comparant :

- les cas dans lesquels B est vraie en même temps que A (l'aire en **bleu clair** : $A \cap B$)
- parmi tous les cas dans lesquels A est vraie (l'aire en **vert** : A , dont celle en dessous de la **bleu clair**)

Ainsi :

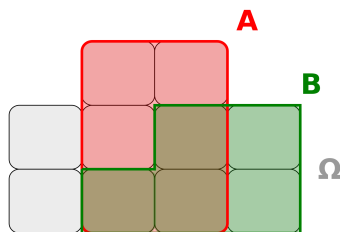
- plus B “**tend à être vraie**” *parmi les cas* dans lesquels A est vraie, plus le **ratio augmente** (jusqu'à $\Pr(B|A) = 1$, c-à-d si j'ai A alors j'ai aussi B)
- et, plus $A \rightarrow B$ “**tend à être**” une **relation valide** (sans “trop d'exceptions”)

Remarques à méditer ! (fondements du “Bayésianisme” ou du “probabilisme”)

- $\Pr(B|A)$ caractérise **ce que l'on peut conclure** sur B à partir de la **connaissance** de A
- On peut **extraire de l'information** sur une proposition B à partir d'une proposition (connue) A
- On peut **ajuster une estimation** initiale : depuis $\Pr(B)$ vers $\Pr(B|A)$ si l'on sait que A est vraie

Conditionnement et indépendance

Considérons l'exemple ci-dessous (ici on a une connaissance parfaite de l'univers Ω , ce qui est rare en pratique) :



Ici : $\Pr(A) = 6/10$, $\Pr(B) = 5/10$ et $\Pr(A \cap B) = 3/10$

Par définition : $\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{3/10}{6/10} = \frac{1}{2}$ et donc on constate que $\Pr(B|A) = \Pr(B) = \frac{1}{2}$

L'indépendance entre deux propositions se matérialise par la relation suivante :

$$\Pr(B|A) = \Pr(B)$$

ce qui donne naturellement, pour deux proposition indépendantes :

$$\Pr(A \cap B) = \Pr(A) \times \Pr(B)$$

Savoir que A est réalisé, n'apporte aucune information quant à la conclusion vis-à-vis de B

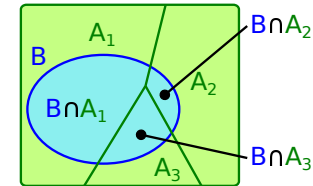


Relation des probabilités totales

$$\Pr(B) = \sum_{i=1}^k \Pr(B \cap A_i) = \sum_{i=1}^k \Pr(B|A_i)\Pr(A_i)$$

avec :

- $(A_i)_{i \in [1,k]}$ un ensemble **complet** d'événements **mutuellement exclusifs**
- c-à-d : $\cup_{i=1}^k A_i = \Omega$ et $A_i \cap A_j = \emptyset, \forall i \neq j$
(on parle d'une **partition** de Ω)



Ici on a soit A_1 , soit A_2 , soit A_3

On somme les différentes sections en **bleu** : $B \cap A_i$ (ainsi on “reconstitue” B au complet), puis on utilise la définition de la probabilité conditionnelle

Schématiquement : on prend en compte toutes les **options** A_i **possibles**, avec leur “poids” respectif $\Pr(A_i)$, et **ce qu’elles nous apprennent** sur B , via $\Pr(B|A_i)$:

on somme le tout afin de pouvoir conclure quant à B

Probabilité et valeurs possibles d'une grandeur

On peut considérer :

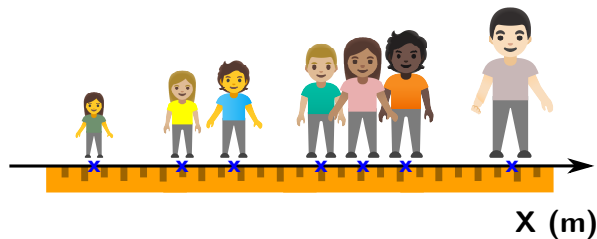
- **Différents** événements, issues, états, propositions \Rightarrow cas **discret**
 - **Différentes valeurs** d'une grandeur d'intérêt (température, distance, date ...) \Rightarrow cas **continu**
- dont la réalisation est **incertaine** (qu'elle qu'en soit la raison, idée d'"aléa" ou non)

Outil fondamental du corpus de la **théorie des probabilités**

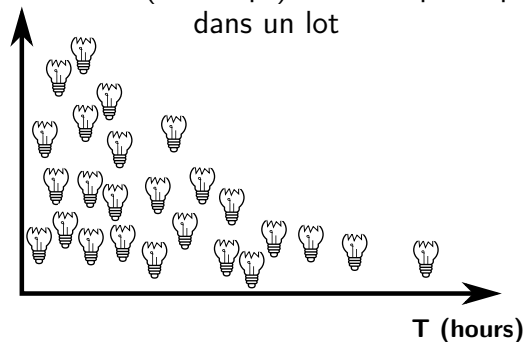
On note X une **variable aléatoire** (on pourrait aussi parler de **variable incertaine**)

- X peut prendre un ensemble fini ou un ensemble continu de valeurs (par exemple, dans \mathbb{R}^+ pour une température, ou dans $\{1, 2, 3, 4, 5, 6\}$ pour un jet de dé)
- On note $X = x$, où x est une **réalisation (possible)** de la variable aléatoire X
- La proposition " $X = x$ " peut être vue comme une *issue* ou un *événement* ou un *état*

La taille (une longueur) d'un individu pioché dans une population



La durée de vie (un temps) d'une ampoule piochée dans un lot



Distribution (des valeurs possibles) d'une variable aléatoire

Considérons le cas d'une variable X continue (dans \mathbb{R})

Soit $\Pr(x \leq X \leq x + dx)$: la probabilité que la **valeur prise par** X soit dans un (petit) intervalle $[x, x + dx]$

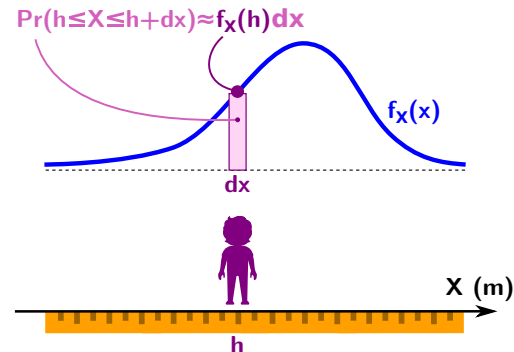
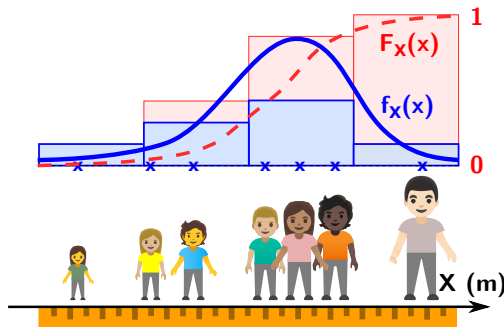
On définit la **fonction de densité de probabilité** $f_X(x)$ (ou fonction de distribution associée à X)
ou **PDF** Probability Density Function :

$$f_X(x)dx \approx \Pr(x \leq X \leq x + dx)$$

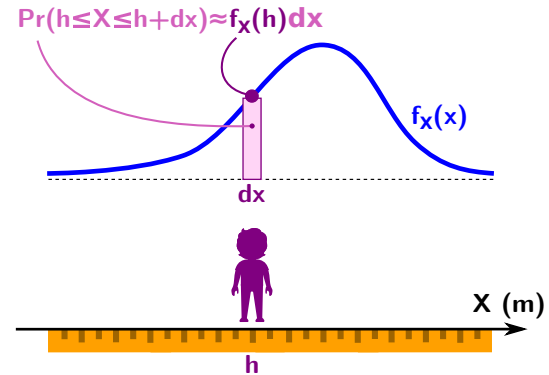
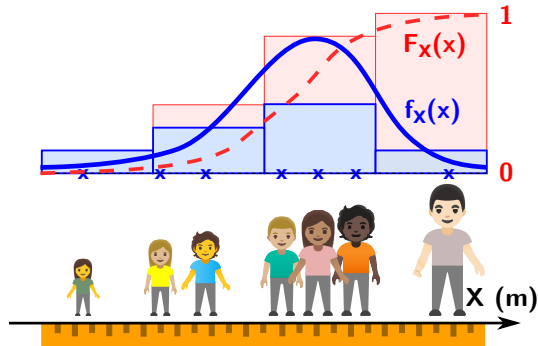
En "sommant" les **contributions possibles** (pour X) jusqu'à une valeur particulière x_0 on définit la fonction de **répartition** $F_X(x)$ (ou de densité cumulée, **CDF** Cumulative Density Function)

$$F_X(x_0) = \Pr(X \leq x_0) = \int_{-\infty}^{x_0} f_X(x)dx$$

- $X \leq x_0$ correspond à la proposition : " X prend une valeur inférieure à x_0 "
- Plus x_0 est grand, plus $\Pr(X \leq x_0)$ est grand et $\lim_{x_0 \rightarrow +\infty} F_X(x_0) = 1$



Pourquoi une “densité” ?



On parle de **densité** de probabilité

En effet, la probabilité d'observer **exactement** une valeur particulière x_u pour $X \in \mathbb{R}$ est **nulle** : on a $\Pr(X = x_u) = 0$, ce qui n'est pas une description très utile en pratique

Pour faire simple : (voir aussi figures ci-dessus)

- En resserrant l'intervalle :
 $\Pr(x \leq X \leq x + dx)$ devient très petit, tout comme dx ,
et $f_X(x_u)$ reste une valeur “stable” avec : $f_X(x_u) = \lim_{\Delta x \rightarrow 0} \frac{\Pr(x_u \leq X \leq x_u + \Delta x)}{\Delta x}$
- Si on s’“attend à observer” **beaucoup de réalisations dans un intervalle donné**, alors la **densité est élevée** : beaucoup de masse de probabilité (par unité dx) est distribuée à cet endroit
- On peut voir f_X comme le lissage d'un histogramme (**rectangles en bleu** ci-dessus) pour lequel la largeur des classes dx deviendrait très fine (des classes trop fines ne contiendraient souvent aucun point)

Densité et “somme de masse”

Remarque : f_X est la dérivée de F_X (c-à-d $F_X' = f_X$) avec :

$$F_X(x) = \Pr(X \leq x) = \int_{-\infty}^x f_X(u) du$$

On peut obtenir la masse dans un intervalle $[x_a, x_b]$ avec :

$$\Pr(x_a \leq X \leq x_b) = \Pr(X \leq x_b) - \Pr(X \leq x_a) = F(x_b) - F(x_a) = \int_{x_a}^{x_b} f_X(u) du$$

Schématiquement :

- on “somme” toute la masse (de probabilité) contenue entre x_a et x_b , en soustrayant de la masse comprise avant x_b , celle comprise avant x_a (pour ne retenir que la zone entre les deux)
- on “somme” chaque “constituant élémentaire” de largeur du et de densité $f_X(u)$

Remarque : pour le cas discret on considère chaque $x_i \in \Omega$ et on a $\sum_{i=1}^{\text{Card}(\Omega)} \Pr(X = x_i) = \sum_{i=1}^{\text{Card}(\Omega)} p_i = 1$



L'**espérance** d'une variable aléatoire $X \in \mathbb{R}$ de PDF ou **distribution** f_X est définie par :

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x \cdot f_X(x) dx$$

Remarque : c'est la "somme pondérée" (par leur poids respectifs $f_X(x)$) des valeurs x possibles pour X

Remarque : cette quantité est parfois aussi appelée **moyenne**

(on préférera, autant que possible, réserver cette dénomination à son usage empirique, c-à-d en relation avec des données plutôt qu'avec les variables aléatoires associées)

On peut calculer l'**espérance** d'une **fonction** g d'une **variable aléatoire**, par exemple $\mathbb{E}[(X - c)^2]$ avec ici $g(X) = (X - c)^2$, et notamment lorsque $c = \mathbb{E}[X]$:

La **variance** d'une variable aléatoire $X \in \mathbb{R}$ est définie par :

$$\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \int_{-\infty}^{+\infty} (x - \mu)^2 \cdot f_X(x) dx = \sigma^2$$

où on note $\mu = \mathbb{E}[X]$ et $\sigma = \sqrt{\text{Var}[X]}$ aussi appelé couramment écart-type

La variance est l'"*espérance des écarts (au carré) à la moyenne*" et donc assez naturellement une mesure de "**dispersion**" des valeurs possibles



Distribuer en pratique

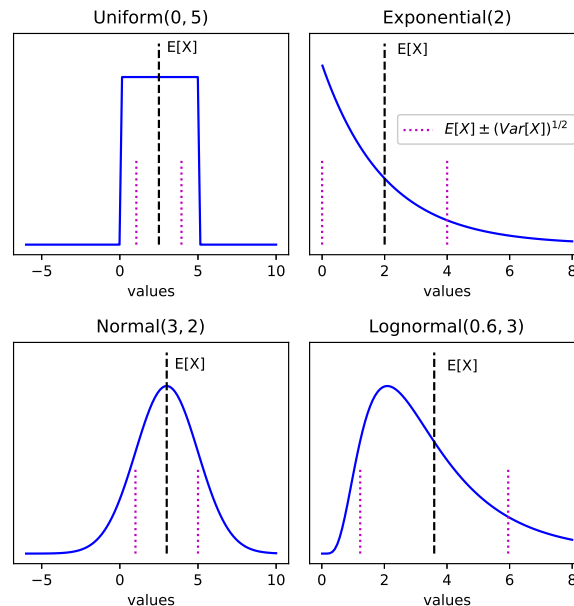
- Il faut traduire notre connaissance quant aux valeurs possibles de X sous forme d'un **modèle**
- On peut considérer un **modèle** de distribution **paramétrique**

On peut citer (parmi d'autres) les distributions, ou "lois" ou formes paramétriques (continues) :

- uniforme
- normale
- exponentielle
- lognormale

f_X est une fonction de la forme $f(x, \theta)$ et θ représente un (ou plusieurs) paramètres à définir (lesquels déterminent la position, la forme, l'"étalement", etc., de la distribution)

Par exemple, pour la loi **normale** (loi de Gauss) : $f_X(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$, et on note $X \sim \mathcal{N}(\mu, \sigma)$





$\mathbb{E}[X]$ et $\text{Var}[X]$ sont des valeurs scalaires **calculées à partir de la connaissance complète de la distribution** f_X associée à X

A ce titre :

- Elles sont “moins informatives” que la connaissance de la description complète f_X
- Deux variables X_1 et X_2 avec la même moyenne et/ou la même variance, peuvent être associées à des **distributions** très **différentes**,
c'est le cas lorsque la différence réside dans la **richesse en valeurs “extrêmes”**
(qui sont généralement rares et dont le poids pèse peu dans l'estimation de μ ou σ)

Remarque : avec la définition $\mathbb{E}[X] = \int_{\mathbb{R}} x \cdot f_X(x) dx$:

l'espérance est “**attirée**” par les **valeurs** x **associées à une densité** $f_X(x)$ **élevée**

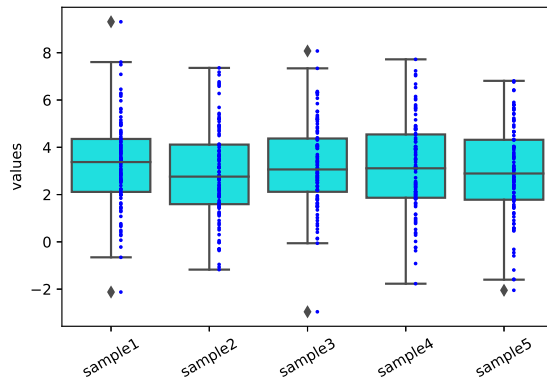
Elle représente une sorte de “**barycentre**” des valeurs possibles, en fonction de leur poids respectifs

Remarque : l'espérance est dite : moment centré d'ordre 1, la variance : moment centré d'ordre 2.

Les moments $\mathbb{E}[(X - \mu)/\sigma]^k$ d'ordre 3 et 4 sont dit respectivement asymétrie (ou skewness) et aplatissement (ou kurtosis)

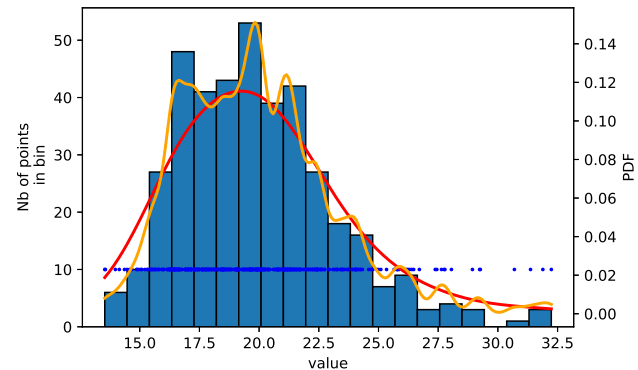
Remarque : ci-dessous, d'autres éléments pouvant être utiles à la description de la distribution de X

Boxplot (moyenne, écart interquartile, "outliers") :
pour information, échantillons générés à partir d'une loi
 $X \sim \mathcal{N}(3, 2)$



Histogramme et estimation non paramétrique de densité dite à noyaux, ou kernel density estimate (KDE) :

⇒ on somme des contributions locales : de la masse est distribuée autour de chaque observation, selon la forme du noyau choisi





On peut étendre les outils en considérant simultanément plusieurs grandeurs incertaines

Vecteur aléatoire

Un **vecteur aléatoire** \mathbf{X} (dans \mathbb{R}^d) est le regroupement de d variables aléatoires tel que :

$$\mathbf{X} = (X_1, X_2, \dots, X_d)$$

Sa distribution est caractérisée au moyen de sa **fonction de répartition conjointe** (joint-CDF) :

$$F_{\mathbf{X}}(x_1, x_2, \dots, x_d) = \Pr(X_1 \leq x_1 \cap X_2 \leq x_2 \cap \dots \cap X_d \leq x_d)$$

ou de façon analogue par sa densité conjointe (joint-PDF)

On définit la **distribution marginale** (par exemple selon X) par :

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x, y) dy$$

où $f_{X,Y}(x, y)$ est la densité conjointe du vecteur aléatoire (X, Y)

Remarque : on retrouve ici la relation des **probabilités totales** \Rightarrow on “somme” sur tous les y possibles



Définition de l'indépendance

On peut exprimer l'indépendance entre deux variables aléatoires X et Y par :

$$F_{X,Y}(x, y) = \Pr(X \leq x \cap Y \leq y) = \Pr(X \leq x)\Pr(Y \leq y) = F_X(x)F_Y(y)$$

ou de façon équivalente par :

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

Ceci signifie aussi que $\Pr(Y \leq y | X = x) = \Pr(Y \leq y)$,

ainsi “connaître” x ou “savoir que $X = x$ ” n'a pas d'effet sur la valeur de Y



Covariance de deux variables aléatoires

Pour deux variables aléatoires X et Y , la covariance s'écrit :

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

Remarque : $\text{Cov}[X, X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \text{Var}[X]$

On définit la **matrice de covariance** (pour plusieurs variables) :

$$\mathbf{R} = \begin{bmatrix} \text{Cov}[X, X] & \text{Cov}[X, Y] \\ \text{Cov}[Y, X] & \text{Cov}[Y, Y] \end{bmatrix}$$

et le **coefficient de corrélation** (entre deux variables) :

$$\rho = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X]\text{Var}[Y]}} \in [0, 1]$$

On peut passer de \mathbf{R} à $\boldsymbol{\rho}$ (matrice de corrélation) en utilisant : $\boldsymbol{\rho} = \mathbf{D}\mathbf{R}\mathbf{D}$ où $\mathbf{D} = \text{diag}(\frac{1}{\sqrt{\text{Var}[X_i]}})$

Cas multidimensionnel : vecteurs aléatoires



Ci-dessous, deux exemples de distributions où les marginales sont affichées

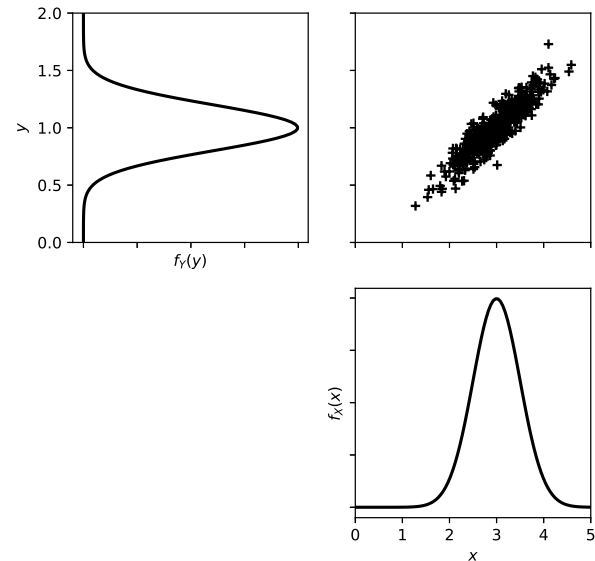
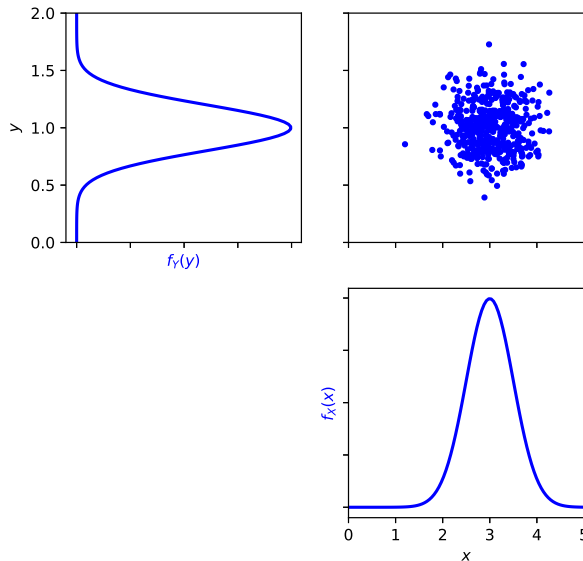
(moyennes et variances sont identiques, mais la corrélation varie de très faible (**gauche**) à forte (**droite**), notez que cela ne se voit pas sur les marginales seules)

$$(X, Y) \sim \mathcal{N}\left(\boldsymbol{\mu} = [3, 1], \mathbf{R} = \begin{bmatrix} 0.5^2 & 0.001 \\ 0.001 & 0.2^2 \end{bmatrix}\right)$$

ici $\rho = 0.01$

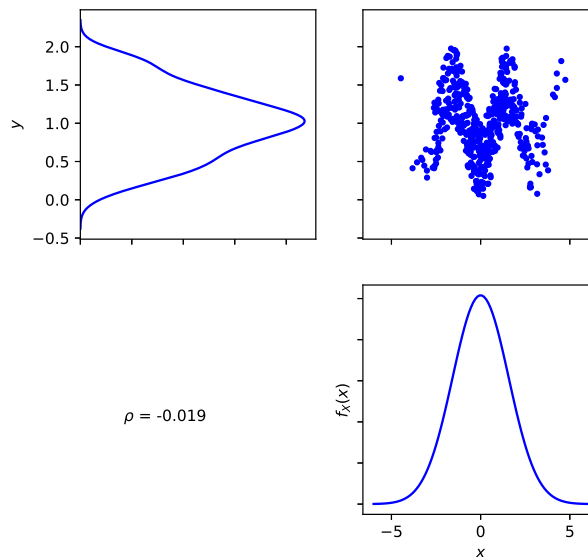
$$(X, Y) \sim \mathcal{N}\left(\boldsymbol{\mu} = [3, 1], \mathbf{R} = \begin{bmatrix} 0.5^2 & 0.09 \\ 0.09 & 0.2^2 \end{bmatrix}\right)$$

ici $\rho = 0.9$



Cas multidimensionnel : indépendance et covariance nulle

Attention : $\text{Cov}[X, Y] = 0$ ou $\rho = 0$ n'implique pas que X et Y sont indépendantes (voir exemple ci-dessous)



Par contre, **si** X et Y sont **indépendantes**, alors **nécessairement** $\text{Cov}[X, Y] = 0$

$$\begin{aligned} \text{(preuve : } \text{Cov}[X, Y] &= \mathbb{E}_{X,Y}[(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])] \\ &= \mathbb{E}_X[(X - \mathbb{E}[X])] \cdot \mathbb{E}_Y[(Y - \mathbb{E}[Y])] \\ &= (\mathbb{E}[X] - \mathbb{E}[X]) \cdot (\mathbb{E}[Y] - \mathbb{E}[Y]) = 0 \end{aligned}$$

Structure de dépendance multidimensionnelle

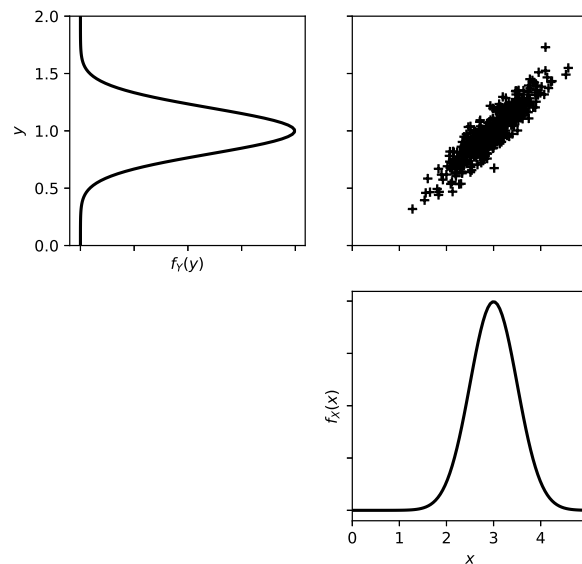
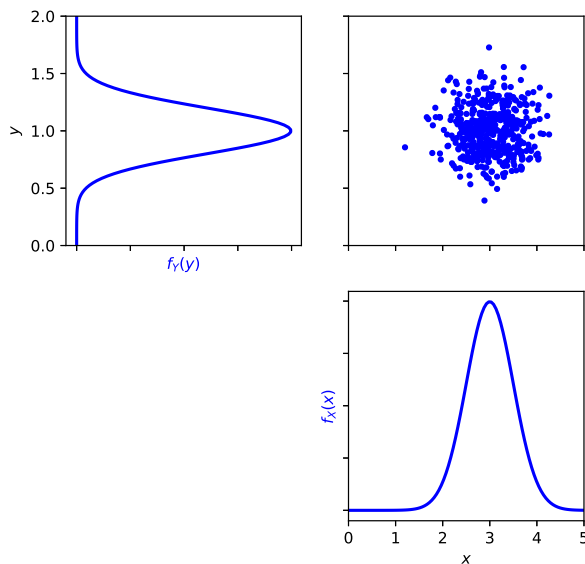
La description (probabiliste) d'une densité jointe $F_{X_1, \dots, X_d}(x_1, \dots, x_d)$ peut être une tâche complexe. On fera souvent usage de représentation paramétriques, notamment en découplant l'effet des densités marginales à l'aide de copules, telles que :

$$F_{X_1, \dots, X_d}(x_1, \dots, x_d) = C(F_{X_1}(x_1), \dots, F_{X_d}(x_d)) = \Pr(X_1 \leq x_1 \cap X_2 \leq x_2 \cap \dots \cap X_d \leq x_d)$$

où $C(\cdot, \dots, \cdot)$ est la copule choisie et les $F_{X_i}(x_i)$ sont les CDF marginales

(ici) On utiliserait plutôt une copule dite indépendante

(ici) On utiliserait plutôt une copule Gaussienne





On peut étendre la notion d'une variable aléatoire, vers un vecteur aléatoire, jusqu'à un processus aléatoire :

Processus aléatoire ou processus stochastique

On appelle processus stochastique, une **collection** continue (infinie) de variables aléatoires indexées par un paramètre :


$$\{X(t)\} = \{X(t = t_0), X(t = t_1), \dots, X(t = t_k), \dots\}$$

où par exemple $t \in \mathbb{R}$, représente un index de temps, ou d'espace $t \in \mathbb{R}^3$ (on parlera alors de champ aléatoire)

Lorsque t est fixé, $X(t)$ est une variable aléatoire

La définition complète d'un tel objet implique de spécifier la **densité conjointe du processus** (celle d'une collection infinie de variables aléatoires corrélées)

- En général, on va faire des hypothèses sur la **structure de dépendance** et/ou sur les **marginale**s (à t fixé, par exemple, choisir des distributions Gaussiennes : ceci définissant alors un *processus Gaussien*), pour permettre une spécification plus aisée de l'objet mathématique
- Pour simplifier, la structure de dépendance est souvent spécifiée via la covariance entre deux points $\text{Cov}[X(u), X(v)]$ pour tout u, v (ce qui se traduit généralement par une hypothèse de normalité, la loi normale étant pleinement définie par ses moments d'ordre 2 : espérance et variance)
- Souvent on ajoute l'hypothèse que le processus considéré est **stationnaire**, ce qui se traduit par une covariance ne dépendant que de la **distance** $|u - v|$ et pas des valeurs particulières de u et de v
- Il est généralement pertinent de faire des hypothèses simples, à défaut d'information qui attesterait d'un besoin de plus de complexité / de spécificité (principe du rasoir d'Occam)



2. Science statistique et inférence



- Comment **quantifier l'incertitude** ? avec quels **outils** ? \Rightarrow avec les **outils probabilistes** (variables aléatoires, distribution, espérance, probabilités conditionnelles, ...)
- Comment **“utiliser”** les outils probabilistes **en pratique** ? comment quantifier la **variabilité** (ou la méconnaissance) des grandeurs d'intérêt ?

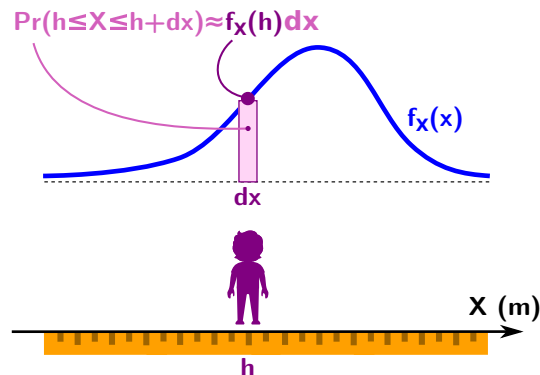
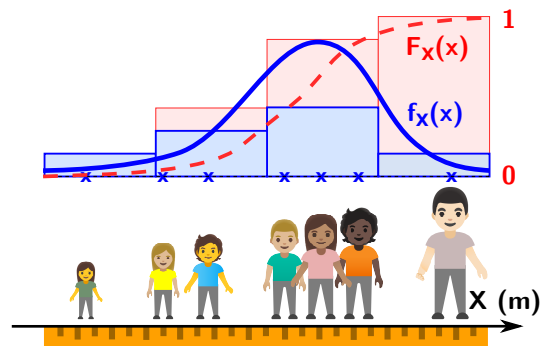
Par le biais de la science statistique ...

Qu'est ce que la (science) statistique ?

- La science de l'**empirisme** : des **observations (données)** et de leur **exploitation**
- Historiquement, la science de l'“**Etat**” (*statista* \leftrightarrow “homme d'état”),
 \Rightarrow la science de l'administration des “choses complexes” (variables, connues imparfaitement, étendues, diverses, recensées partiellement, ...)
- L'**interface** entre **problèmes parfaits** (mathématiques) et **questions concrètes** (sciences physiques, ingénierie, science de la décision, ...)

Qu'est ce que la (science) statistique ?

- La science de l'empirisme : des **observations** (**données**) et de leur **exploitation**
- L'**interface** entre **problèmes parfaits** (mathématiques) et **questions concrètes** (sciences physiques, ingénierie, science de la décision, ...)
- Des **observations** (réelles, constatées) : les tailles des individus mesurés (un **échantillon**) (**croix bleues**)
- Un **modèle** (idéal, une construction mathématique) : la **distribution** de probabilité (**courbe bleu**)
- Une **question concrète** : la taille d'un *nouvel individu* pioché dans la *même population* (prévision vis-à-vis d'un **individu (incertain)** en **violet**)





On va chercher à utiliser les **données** et **observations** disponibles \Rightarrow pour **identifier** un **modèle**

Ce modèle pourra être utilisé ensuite pour faire des prédictions ou inférences

On considérera :

- L'identification d'un **modèle de relation fonctionnelle** $y = g(x)$
- L'identification d'un **modèle probabiliste** : symboliquement, via $\Pr(X)$ ou $\Pr(Y|X)$

Stratégies (ou principes) considérés

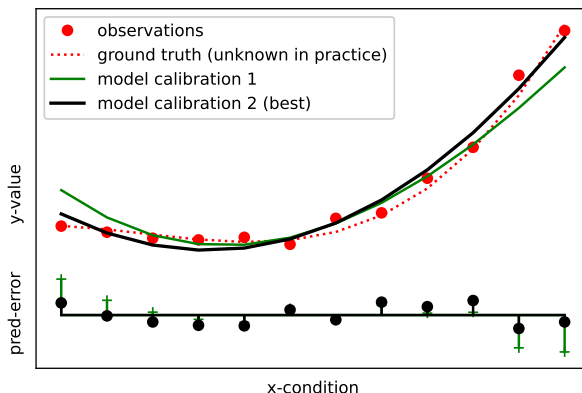
- I - Minimisation de l'erreur de prédiction
- II - Maximisation de la vraisemblance (des données)
- III - Mise à jour Bayésienne

Choisir une **stratégie** (un principe) plutôt qu'une autre, c'est faire des hypothèses spécifiques, c'est **généraliser** ou **extrapoler** à partir de ce qu'on a observé

Principe I - Minimisation de l'erreur de prédiction

On cherche à **identifier** un **modèle de relation fonctionnelle** entre \mathbf{x} et y , noté $g : \mathbf{x} \mapsto y = g(\mathbf{x})$ (dans le champ de l'apprentissage machine (ML), on parlerait de problème de régression supervisée)

On va choisir pour g (parmi tous les g qu'on pourrait proposer) une **forme paramétrique**, telle que :
 $g : \mathbf{x} \mapsto y = g(\mathbf{x}, \beta)$

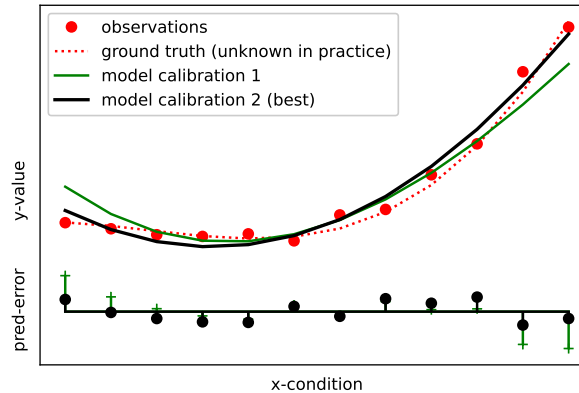


Une estimation du “**meilleur modèle**” (une fois g choisi) **compte tenu des données** est :

Minimisation de l'erreur de prédiction empirique (constatée) :

$$\hat{\beta} = \arg \min_{\beta} \left[\sum_{i=1}^N |y_{\text{mes}}(\mathbf{x}_i) - g(\mathbf{x}_i, \beta)|^2 \right]$$

où le “**coût**” de l'erreur de prédiction est (ici) évalué en fonction de l'écart quadratique $L(u, v) = |u - v|^2$



$$\hat{\beta} = \arg \min_{\beta} \left[\sum_{i=1}^N |y_{\text{mes}}(\mathbf{x}_i) - g(\mathbf{x}_i, \beta)|^2 \right]$$

Avec cette stratégie, on ajuste β (par optimisation)

Le “meilleur modèle” est celui qui donne les “meilleures prédictions” $g(\mathbf{x}_i, \hat{\beta})$ dans les conditions \mathbf{x}_i où l’on dispose d’une “référence” $y_{\text{mes}}(\mathbf{x}_i)$ à laquelle se comparer (d’où apprentissage **supervisé**)

Un tel modèle “reproduit” bien ce que l’on a observé empiriquement. On l’utilisera ensuite pour réaliser de nouvelles prédictions



Remarques :

- le **choix** de g influence largement la capacité du modèle à “bien” **généraliser**
- le choix de g s'appuie souvent sur une tentative de minimisation de **l'erreur de généralisation** (souvent de façon empirique : par validation croisée)

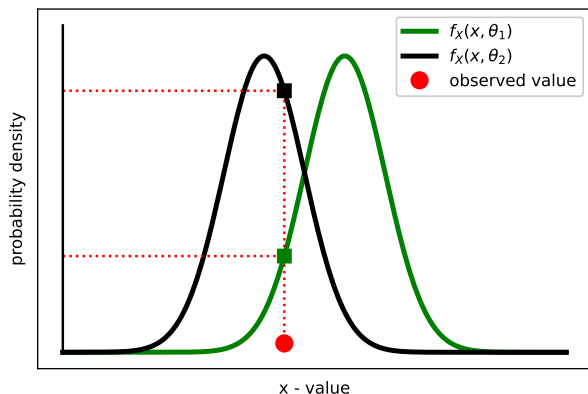
$$\mathbb{E}_{\mathbf{X},Y} [L(Y, g(\mathbf{X}, \beta))]$$

où $\mathbb{E}_{\mathbf{X},Y}[\cdot]$ désigne l'espérance sur tout les couples (y, \mathbf{x}) que l'on **pourrait observer** (par exemple en augmentant le budget expérimental) \Rightarrow cependant, on ne peut que **supposer** cette relation “réelle” (“sous-jacente”) entre \mathbf{X} et Y , car c'est **précisément ce que l'on cherche à déterminer**

- minimiser l'erreur de prédiction empirique seulement, pour une forme g “trop flexible”, peut amener à ne faire que “reproduire” les données : phénomène de **sur-apprentissage**
- au delà de l'identification de g , on peut **étudier (au sens probabiliste) l'erreur de prédiction**
On cherche alors à “adjoindre” à une prédiction ponctuelle, un modèle pour l'**incertitude de prédiction**, noté $\varepsilon_{\text{mod}}(\mathbf{x})$:
ainsi on prédit via : $y_{\text{pred}}(\mathbf{x}) = g(\mathbf{x}, \beta) + \varepsilon_{\text{mod}}(\mathbf{x})$

Principe II - Maximisation de la vraisemblance des données

- On a observé N réalisations (indépendantes) x_i d'une variable aléatoire notée X
- On cherche à **identifier** (ajuster) la **densité** f_X qui semble **“correspondre” le mieux aux observations**
- On se limite aux cas où f_X adopte une forme paramétrique $f_X(x, \theta)$ et on cherche à déterminer les **“meilleures” valeurs** de ces paramètres θ



Ici, par exemple, $f_X(x, \theta_2) > f_X(x, \theta_1)$, il est donc plus “vraisemblable” d’observer x quand $\theta = \theta_2$ que quand $\theta = \theta_1$. Or x a été observé, donc θ_2 “correspond mieux” à cette observation

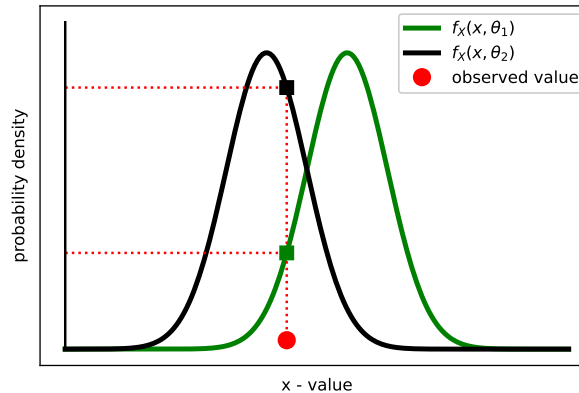
Maximisation de la **vraisemblance** des données

$$\hat{\theta} = \arg \max_{\theta} \left[\prod_{i=1}^N f_X(x_i, \theta) \right]$$

où la valeur la plus pertinente pour θ , notée $\hat{\theta}$, est celle qui **rend maximale la probabilité d’observer les données qu’on a effectivement observées (vraisemblance)**

Etant donné qu’on les a effectivement observé, c’est le “mieux” que l’on puisse dire de θ

Maximiser la vraisemblance (jointe) de données (indépendantes), c’est maximiser le produit des densités

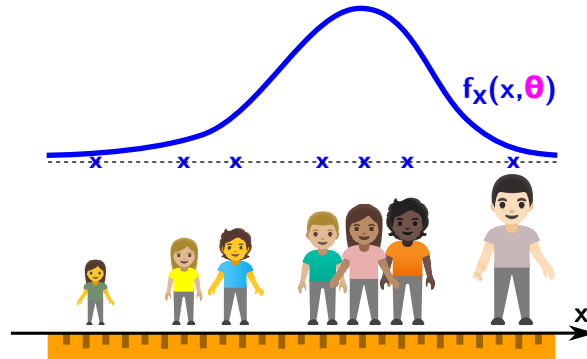


Avec cette stratégie, une fois qu'une forme paramétrique $f_X(x, \theta)$ a été choisie, on ajuste θ jusqu'à ce que la **probabilité d'observer ce que l'on a effectivement observé soit maximale**

Tout autre candidat θ , qui rendrait la vraisemblance des données plus faible, "correspond moins bien" aux observations disponibles

On dispose alors d'un modèle probabiliste "**ajusté**", pour la quantité d'intérêt $X \sim f_X(x, \hat{\theta})$

Principe II - Maximisation de la vraisemblance des données



Cette approche peut donner des résultats (nécessairement ponctuels, $\hat{\theta}$) **très variables** (si l'on venait à répéter l'estimation) quand le **nombre d'observations disponibles** est faible \Rightarrow on va ajuster pleinement et uniquement par rapport ce que l'on observe (qui est alors limité)

Le **choix** de la **forme paramétrique** pour f_X en dit long sur ce que l'on s' "**attend à observer**" (consciemment ou non) en considérant des réalisations de X

Autrement dit, ce choix détermine largement comment on va **généraliser** au-delà des données observées, (pour produire une **inférence**)

Vers le principe III : Relation de Bayes

Par définition de la probabilité conditionnelle : $\Pr(B|A) = \frac{\Pr(B \cap A)}{\Pr(A)}$

et naturellement $\Pr(A \cap B) = \Pr(B \cap A)$,

d'où $\Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A)$ et on en déduit l'expression suivante :

Relation de Bayes (on parle parfois d'“inversion probabiliste” : depuis les données vers les hypothèses)

$$\Pr(\text{Hypothèse} = h | \text{Observations}) = \frac{\Pr(\text{Observations} | \text{Hypothèse} = h) \cdot \Pr(\text{Hypothèse} = h)}{\Pr(\text{Observations})}$$

en spécifiant :

- A : une **hypothèse**, susceptible de générer (d'expliquer, de conduire à), selon $A \rightarrow B$
- B : une ou plusieurs **observations** réalisées

On schématise généralement en écrivant : **posterior** \propto **vraisemblance** \times **prior**

Ces éléments, connectés par la **relation de Bayes**, peuvent s'interpréter comme suit :

- $\Pr(\text{Hypothèse} = h)$ traduit notre “**jugement initial**” vis-à-vis d'une proposition h sur laquelle on souhaite se prononcer
- $\Pr(\text{Observations} | \text{Hypothèse} = h)$ relie notre hypothèse h à ce que l'on s'“**attend à observer**” si h est vraie (caractérisé de façon symbolique par $h \rightarrow \text{Observations}$)
- $\Pr(\text{Hypothèse} = h | \text{Observations})$ traduit notre “**jugement mis à jour**” (vis-à-vis de h) **compte tenu des observations exploitées**

Relation de Bayes (un exemple chiffré pour mieux comprendre)

Réarrangeons les termes et explicitons :

$$\Pr(\text{Hypothèse} = h | \text{Données}) = \frac{\Pr(\text{Données} | \text{Hypothèse} = h)}{\Pr(\text{Données})} \cdot \Pr(\text{Hypothèse} = h)$$

- Par exemple, si $\Pr(\text{Données} | \text{Hypothèse} = h) / \Pr(\text{Données}) > 1$, il est plus probable d'observer ces données quand h est vraie (en sachant h) que sans cette information
- D'après la formule, observer ces données renforce ainsi la crédibilité de h , car :
 $\Pr(\text{Hypothèse} = h | \text{Données}) > \Pr(\text{Hypothèse} = h)$
- et inversement...

Ci-dessous, un exemple avec :

H une variable binaire, un jugement initial non-informatif, c'est-à-dire $\Pr(H = h) = \Pr(H = \bar{h}) = 0.5$ et $\Pr(D | H = h) \gg \Pr(D | H = \bar{h})$ des données plus vraisemblables sous h que sous \bar{h}

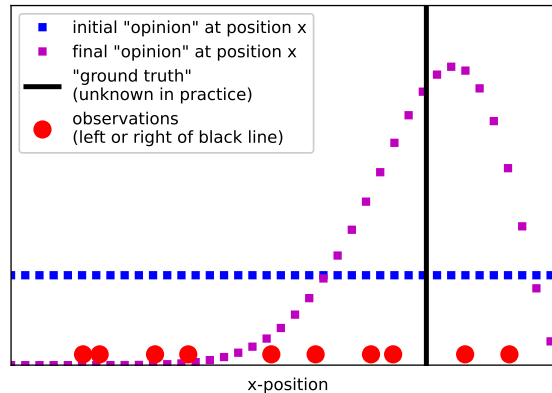
$$\begin{aligned} \Pr(H = h | D) &= \frac{\Pr(D | H = h) \Pr(H = h)}{\Pr(D | H = h) \Pr(H = h) + \Pr(D | H = \bar{h}) \Pr(H = \bar{h})} \\ &= \frac{0.9 \times 0.5}{0.9 \times 0.5 + 0.2 \times 0.5} = 0.81 \text{ à comparer à } \Pr(H = h) = 0.5 \end{aligned}$$

où le dénominateur est calculé par la formule des probabilités totales, c'est-à-dire en sommant sur les deux états possibles (h ou \bar{h})

Principe III - Mise à jour Bayésienne

Considérons l'exemple ci dessous :

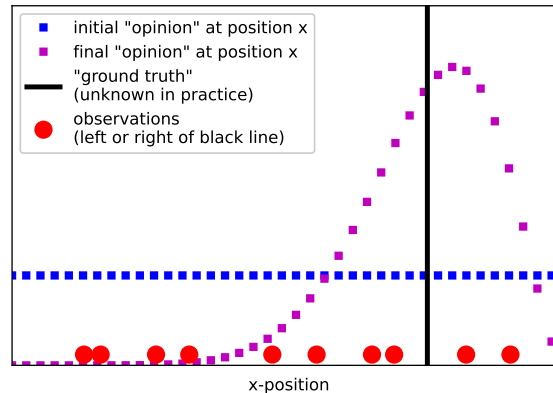
- on cherche la **position z d'une ligne verticale** (inconnue) :
c'est la quantité d'intérêt sur laquelle on cherche à se prononcer (à partir des données disponibles)
- on dispose de 10 lancers pour lesquels l'information obtenue est : on se situe à gauche ou à droite de la ligne



$$\Pr(\text{Hypothèse} = z | \text{Observations}) = \frac{\Pr(\text{Observations} | \text{Hypothèse} = z) \cdot \Pr(\text{Hypothèse} = z)}{\Pr(\text{Observations})}$$

La conclusion *a posteriori* (pour chaque position candidate z ou hypothèse) après l'observation des données est une combinaison de :

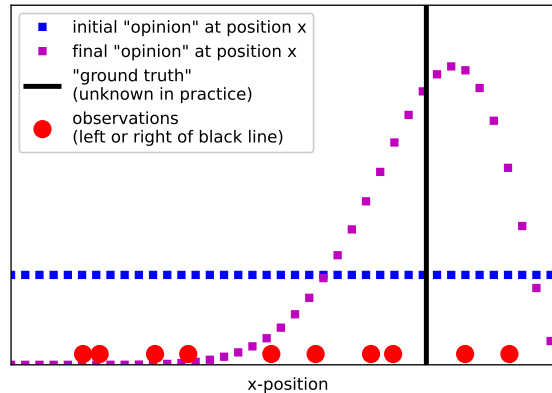
- l'opinion *a priori* vis-à-vis de l'hypothèse considérée
- la probabilité d'observer effectivement les données observés (*vraisemblance des données*) si cette hypothèse est valide



Avec cette stratégie, il s'agit de spécifier un jugement initial, noté $f_{\Theta}(\theta)$ dans le cas continu où Θ est une variable aléatoire (la valeur inconnue du paramètre d'un modèle à identifier) qui représente ce que l'on cherche à caractériser, puis de le **"mettre à jour"**, à la vue des observations disponibles \mathbb{X} et à partir de la relation de Bayes :

$$f_{\Theta}(\theta|\mathbb{X}) \propto f_{\mathbf{X}}(\mathbb{X}|\theta) \cdot f_{\Theta}(\theta)$$

où $\mathbf{X} = \mathbb{X}$ est l'ensemble des observations disponibles, $f_{\mathbf{X}}(\mathbb{X}|\theta)$ la vraisemblance des données pour une valeur particulière $\Theta = \theta$ et $f_{\Theta}(\theta|\mathbb{X})$ le jugement "mis à jour" (*a posteriori*)



Dans ce cadre (principe III), la conclusion obtenue $f_{\Theta}(\theta|\mathbb{X})$ est exprimée **sous la forme d'une distribution probabiliste** (et non une valeur ponctuelle, comme l'estimation du "meilleur" $\hat{\theta}$, issue du principe II)

La manipulation de l'expression :

$$f_{\Theta}(\theta|\mathbb{X}) = \frac{f_{\mathbf{X}}(\mathbb{X}|\theta) \cdot f_{\Theta}(\theta)}{f_{\mathbf{X}}(\mathbb{X})}$$

est généralement une opération non-triviale. Elle implique alors le recours à l'estimation par échantillonnage (approches Monte-Carlo),

notamment à cause du terme dit de vraisemblance marginale $f_{\mathbf{X}}(\mathbb{X})$ au dénominateur qui souvent, ne peut être calculé (le terme qui "somme" sur tous les θ possibles)



On peut ensuite mettre à profit le **modèle probabiliste du paramètre** Θ , c-à-d : $f_{\Theta}(\theta|\mathbb{X})$, afin de pouvoir exploiter le **modèle de la variable d'intérêt** X , c-à-d : $f_X(x, \theta)$, en “sommant” sur l’“état de connaissance” (mis à jour) relatif à Θ , soit en écrivant :

$$f_X(x) = \int_{\Theta} f_X(x, \theta) f_{\Theta}(\theta) d\theta$$

ou en explicitant la prise en compte des données \mathbb{X} :

Prédiction (inférence) basée sur l’“ensemble” de la connaissance disponible (données + a priori)

$$f_X(x|\mathbb{X}) = \int_{\Theta} f_X(x, \theta) f_{\Theta}(\theta|\mathbb{X}) d\theta$$

on “marginalise” par rapport à Θ

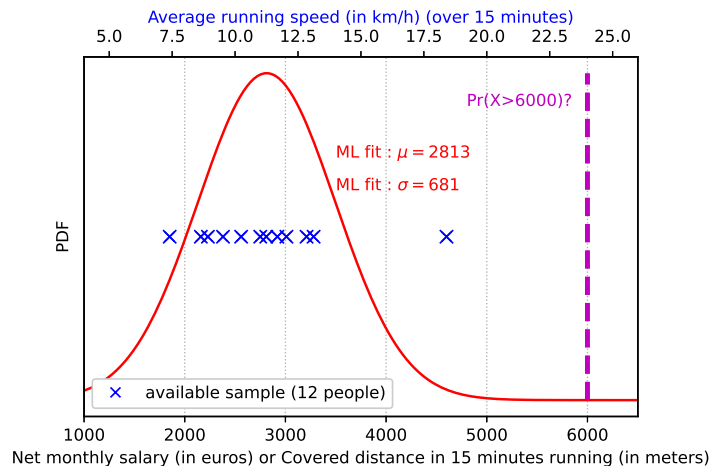
Un exemple et un enseignement : portée limitée de l'outil mathématique



Deux exemples :


Quel ordre de grandeur pour $\Pr(X > 6000)$: 1 pour 100, 1 pour 1.000.000 ?

- 1 Échantillon des salaires net mensuels (12 personnes) d'un petit groupe au sein d'une entreprise
- 2 Échantillon de distances parcourues (12 personnes) en course en 15 minutes



Probablement pas la même réponse pour $\Pr(X > 6000)$:

- le **problème mathématique est identique dans les deux cas**, les données sont les mêmes, l'estimation du maximum de vraisemblance donne le même résultat
- **que doit répondre le statisticien interrogé ?**



3. Choix et tests statistiques



Estimateur statistique

$$t : \mathbb{X} \mapsto \hat{\theta} = t(\mathbb{X})$$

c'est la fonction (ou la procédure) qui fournit une **estimation** :

- pour un paramètre θ (inconnu) dont on recherche la valeur,
- ou pour un descripteur associé à une distribution (ex : sa moyenne, un quantile, ...)
- à partir d'**observations** disponibles \mathbb{X}

Minimisation de l'“erreur espérée” ou du “risque” (Bayésien) \Leftrightarrow critère de choix “optimal”

$$\min_a \mathbb{E}_{\Theta}[L(\theta, a)]$$

où :

- Θ représente la quantité, supposée unique mais **inconnue**, que l'on cherche à **prédire** ou **identifier**
- a représente une **proposition** que l'on peut faire (généralement à partir d'observations \mathbb{X}) à propos de cette grandeur θ inconnue et donc représentée par une variable aléatoire Θ
- $L : (\theta, a) \mapsto L(\theta, a) \in \mathbb{R}$ est une fonction qui caractérise le “**coût**” que l'on subit en choisissant a alors que la “vraie” valeur (inconnue en pratique) est $\Theta = \theta$

On voudrait que $L(\theta, t(\mathbb{X}))$ soit **minimal**, c-à-d que $t(\mathbb{X}) = \hat{\theta}$ soit “proche” de la valeur vraie θ

Mais voici le point névralgique du problème, **on ne connaît pas** θ : ainsi, **comment évaluer** $L(\theta, t(\mathbb{X}))$ en pratique afin de pouvoir choisir un bon estimateur t ?

\Rightarrow Il faudrait “moyenner L ” sur les valeurs de Θ que l'on juge possible (*a priori* \Leftrightarrow **risque bayésien**)



Estimateur statistique

$$t : \mathbb{X} \mapsto \hat{\theta} = t(\mathbb{X})$$

c'est la fonction (ou la procédure) qui fournit une **estimation** :

- pour un paramètre θ (inconnu) dont on recherche la valeur,
- ou pour un descripteur associé à une distribution (ex : sa moyenne, un quantile, ...)
- à partir d'**observations** disponibles \mathbb{X}

Minimisation de l'“**erreur espérée**” ou du “**risque**” (Bayésien) \Leftrightarrow critère de choix “optimal”

$$\min_a \mathbb{E}_{\Theta}[L(\theta, a)]$$

Remarque : le meilleur estimateur est $t : \mathbb{X} \mapsto \theta, \forall \mathbb{X}$.

Mais voici le point névralgique du problème, **on ne connaît pas θ**

\Rightarrow Il faudrait “moyenner L ” sur les valeurs de Θ que l'on juge possible (*a priori* \Leftrightarrow **risque bayésien**)

- C'est satisfaisant en théorie
- C'est difficile à faire en pratique
- Il faudrait proposer un *a priori* pour la distribution de Θ (que pourrait être le θ “réel” que l'on cherche à estimer?)
- Il faudrait aussi chercher l'estimateur t qui donne de bons résultats quel que soit l'échantillon \mathbb{X} utilisé (donc “moyenner” ici aussi sur les réalisations) et répéter l'opération pour différentes tailles d'échantillon



En statistique **classique** (ou **fréquentiste**), par opposition au formalisme de risque bayésien, on va procéder différemment

- On va faire des **hypothèses** : par exemple X suit une loi normale (noté $X \sim \mathcal{N}(\mu, \sigma)$)
- On va construire un **estimateur**, via le choix d'une **stratégie particulière** : par exemple un estimateur du **maximum de vraisemblance**
- On va étudier les **propriétés de cet estimateur** sous ces hypothèses : on parle souvent d'analyse de la distribution d'échantillonnage ou des propriétés **asymptotiques** de l'estimateur (quand la taille de l'échantillon \mathbb{X} augmente) \Rightarrow voir cours (C2) propagation



Pour la loi normale $f_X(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$ et pour un échantillon d'observations indépendantes $\mathbb{X}^N = (x_1, x_2, \dots, x_N)$,

Les **estimateurs du maximum de vraisemblance** de la moyenne et de l'écart-type sont :

$$(\hat{\mu}, \hat{\sigma}) = \arg \max_{u, v} \prod_{i=1}^N f_X(x_i, u, v)$$

après calcul (en prenant le log puis en dérivant) :

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N x_i \text{ et } \hat{\sigma} = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{\mu})^2$$

on pourrait montrer (avec le calcul des espérances associées) que :

$$\mathbb{E}_{\mathbb{X}^N}[\hat{\mu}] = \frac{1}{N} \sum_{i=1}^N \mathbb{E}[X_i] = \mu \text{ (non-biaisé) et que } \mathbb{E}_{\mathbb{X}^N}[\hat{\sigma}] = \frac{(N-1)\sigma}{N} \text{ (biaisé)}$$

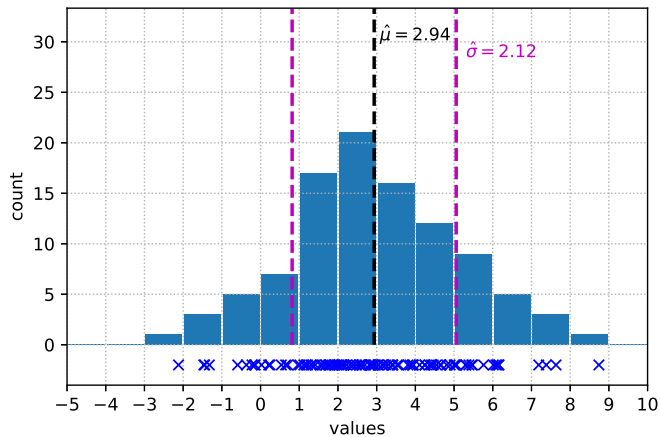
conditionnellement à l'hypothèse que chaque $X_i \sim f_X(\cdot, \mu, \sigma)$

⇒ En clair, on ne connaît pas μ ni σ mais on observe des réalisations de X via x_i à partir desquelles on va tenter de les identifier (on a supposé que $X \sim \mathcal{N}(\mu, \sigma)$)

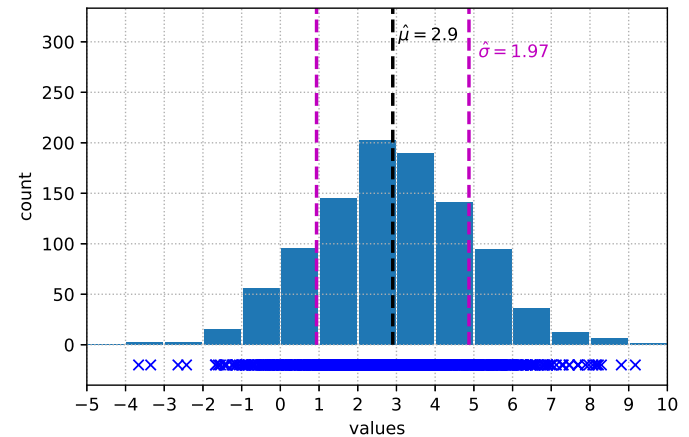
Qualité de l'estimateur et "quantité d'information"

On expérimente avec $X \sim \mathcal{N}(\mu = 3, \sigma = 2)$

(ici on connaît la valeur "vraie" $\theta = (\mu, \sigma)$, puisqu'on la fixe pour étudier la qualité des estimateurs)

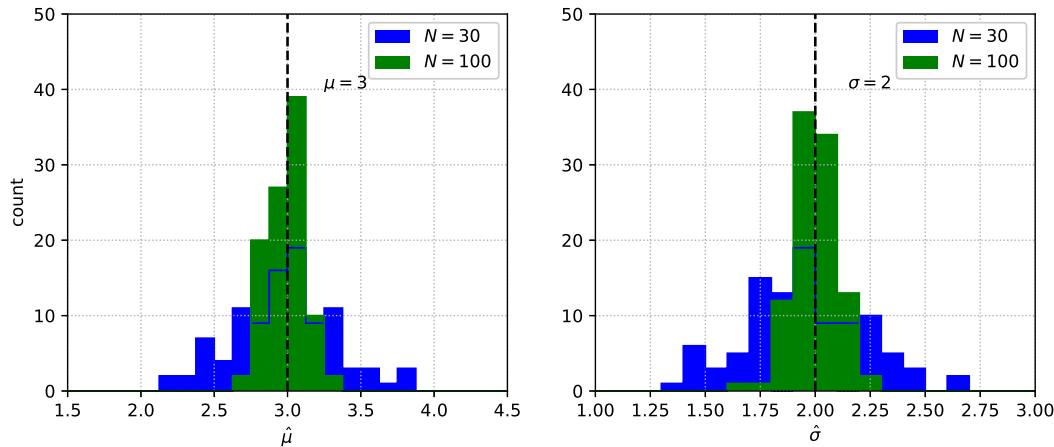


pour $N = 100$ (points dans l'échantillon)



pour $N = 1000$ (points dans l'échantillon)

Qualité de l'estimateur et "quantité d'information"



on retrouve visuellement
(avec 100 répétitions de l'opération d'estimation, respectivement pour $N = 30$ et $N = 100$) :

$$\mathbb{E}_{\mathbb{X}^N}[\hat{\mu}] = \mu \text{ (non-biaisé)}$$

$$\mathbb{E}_{\mathbb{X}^N}[\hat{\sigma}] = \frac{(N-1)\sigma}{N} \text{ (biaisé, mais l'influence du biais diminue avec } N)$$

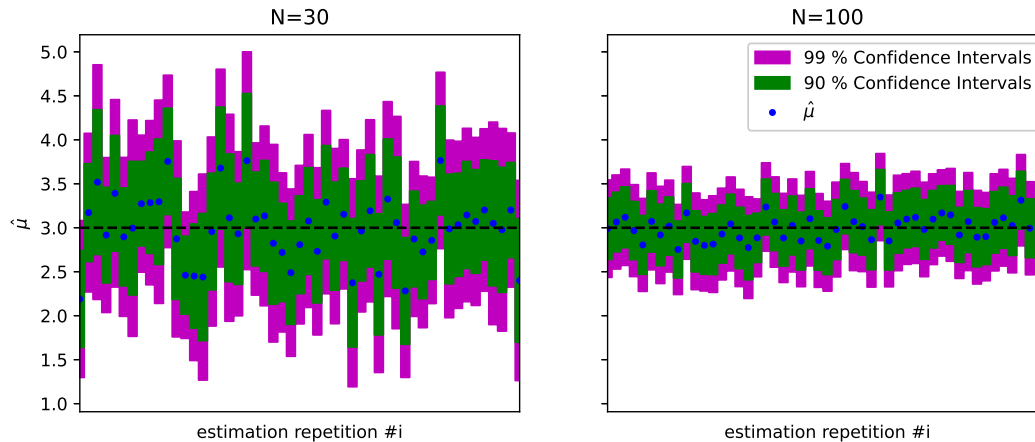
et $\text{Var}_{\mathbb{X}^N}[\hat{\theta}] \rightarrow 0$ quand $N \rightarrow \infty$ dans les deux cas ($\hat{\mu}$ et $\hat{\sigma}$)
(convergence : plus on a de points, "plus fine" est l'estimation)

Intervalles de confiance (basés sur l'estimateur $\hat{\mu}$)

Si $X \sim \mathcal{N}(\mu, \sigma)$ avec μ et σ inconnus (et on observe X), on pourrait montrer que :

$$\frac{\sqrt{N}}{\hat{\sigma}}(\hat{\mu} - \mu) \sim T = Student(N - 1)$$

On s'intéresse à : $\Pr(q_{inf} \leq \frac{\sqrt{N}}{\hat{\sigma}}(\hat{\mu} - \mu) \leq q_{sup}) = 1 - \alpha$, où q_{inf} et q_{sup} sont les valeurs qui “encadrent” $(1 - \alpha)\%$ de la masse totale de la distribution (celle de T) (q_{inf} et q_{sup} sont des valeurs tabulées pour $T(N - 1)$)
en réarrangeant : $\Pr(\hat{\mu} + q_{inf} \frac{\hat{\sigma}}{\sqrt{N}} \leq \mu \leq \hat{\mu} + q_{sup} \frac{\hat{\sigma}}{\sqrt{N}}) = 1 - \alpha$, (les bornes du CI pour μ , à \mathbb{X} donné)



Interprétation : il y a 10% de probabilité que l'intervalle de confiance (CI) à $\alpha = 10\%$, centré en $\hat{\mu}$, **ne contienne pas la valeur “vraie”**, ici $\mu = 3$ (étant donné un échantillon composé de N observations)



La logique des tests statistiques peut être délicate à saisir. Il faut être attentif vis-à-vis de **quelle expérience de pensée on réalise**

Soit H_0 une hypothèse à tester. On la **suppose vraie**, puis on en étudie ses **implications**, en relation avec les données disponibles

Risque de première espèce et "p-value"

$$\Pr(t(\mathbb{X}) \geq u | H_0) = \alpha(u)$$

où :

- α est dit risque de première espèce : probabilité de rejeter H_0 alors qu'elle est vraie,
- u définit la borne de rejet : à partir de laquelle on rejette H_0 (associée à l'estimateur que l'on teste t)
- \mathbb{X} est l'échantillon de données disponible pour statuer

$p = \Pr(t(\mathbb{X}) \geq u | H_0)$ est généralement appelée "**p-value**" et elle est comparée à un niveau de risque jugé acceptable :

⇒ par exemple, on rejette H_0 si $p < 5\%$ (au regard des données ayant servi au test \mathbb{X})



On peut tester différentes questions :

- Le fait qu'un échantillon est issue d'une distribution de paramètre donné, par exemple μ_0 (test paramétrique)
- Le fait qu'un échantillon soit bien décrit par une forme paramétrique particulière de distribution (test d'adéquation)
- Le fait que deux variables sont indépendantes (test d'indépendance)
- Le fait qu'une action ait un effet sur un état et sa distribution (par ex sa moyenne) (ex : prendre un médicament, sur le traitement d'une pathologie)

chaque fois en supposant l'hypothèse vraie, puis en évaluant la **probabilité d'observer ce que l'on a effectivement observé si l'hypothèse est vraie**

si cela est peu vraisemblable, ou moins vraisemblable qu'une alternative H_1 , alors on rejette H_0 ou on lui préfère H_1

Remarque : Le choix de H_0 , voire de H_0 et H_1 , et de t , est une question délicate

Remarque : chaque règle de décision s'accompagne généralement d'un **risque** (à évaluer et à maîtriser) : Que se passe t-il si je "conserve" H_0 alors qu'elle est fausse (en réalité) ?



Illustrons via l'exemple précédent (loi Gaussienne) où l'estimateur du maximum de vraisemblance est : $t = \hat{\mu}$ et on pourrait montrer que $\hat{\mu} \sim \mathcal{N}(\mu, \frac{\sigma^2}{N})$

On peut réarranger $\Pr(t(\mathbb{X}) \geq u | H_0) = \alpha$ pour obtenir $u = \mu + \frac{\sigma}{\sqrt{N}} \Phi^{-1}(\alpha)$ où $\Phi^{-1}(\cdot)$ est l'inverse de la CDF centrée réduite $\mathcal{N}(0, 1)$

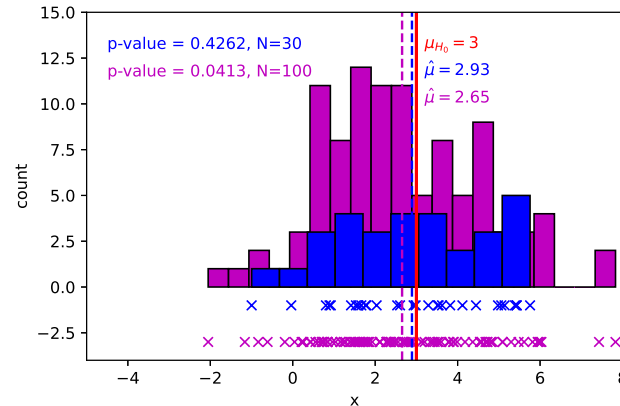
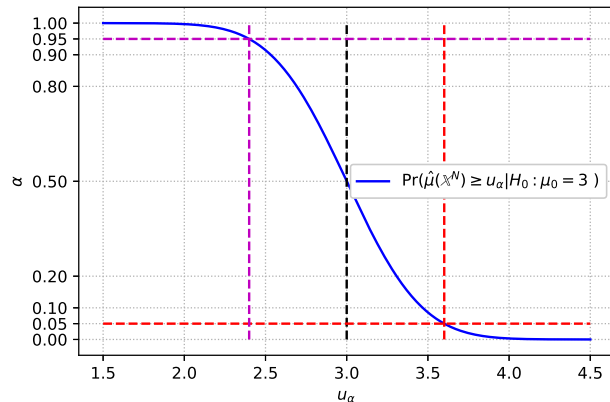
Test sur la moyenne : un exemple

Prenons l'hypothèse $H_0 : \mu_0 = 3$ et supposons $\sigma = 2$ connu, alors (pour $N = 30$) on a : $u \approx 3.601$

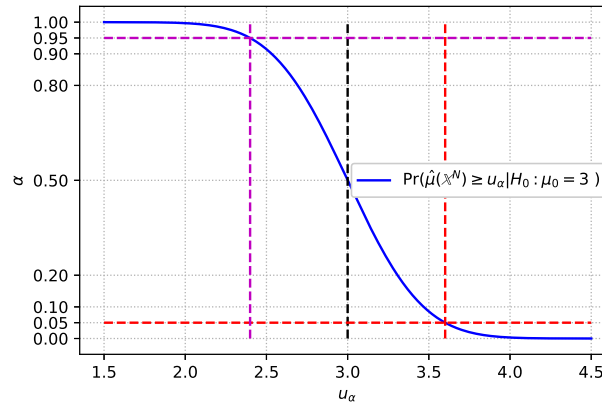
En utilisant la borne u calculée sous hypothèse H_0 , on a : $\Pr(\hat{\mu} > 3.601 | \mu = \mu_0 = 3) \approx 5\%$

Autrement dit, si en calculant $\hat{\mu}$ à partir de \mathbb{X}^N pour $N = 30$, on trouve une valeur **supérieure** à u (ce qui n'a qu'une probabilité de $\alpha \approx 5\%$ de se produire quand $\mu = \mu_0$), alors cela "semble indiquer" que $H_0 : \mu = \mu_0$ est fausse

Néanmoins, en statuant ainsi, on a 5% de probabilité de se tromper (en rejetant H_0 à tort), car une valeur aussi grande de $\hat{\mu}$ peut quand même se produire dans environ 5% des cas lorsque $\mu = \mu_0$



Exemple : ici on a généré deux échantillons \mathbb{X} selon $X \sim \mathcal{N}(2.5, 2)$ et on teste $H_0 : X \sim \mathcal{N}(3, 2)$
 \Rightarrow dans le premier cas ($N = 30$) on ne peut pas rejeter H_0 car $p = 0.42 > 0.05$, dans le second ($N = 100$) on peut rejeter H_0 car $p = 0.04 < 0.05$



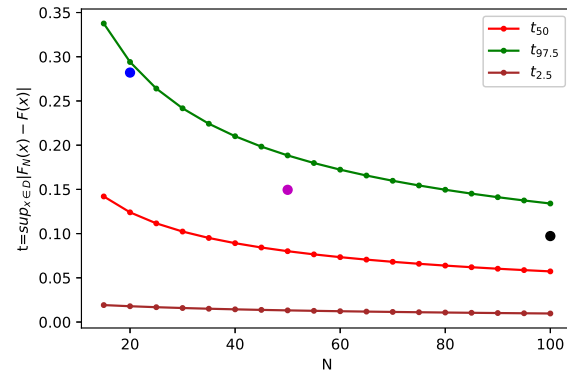
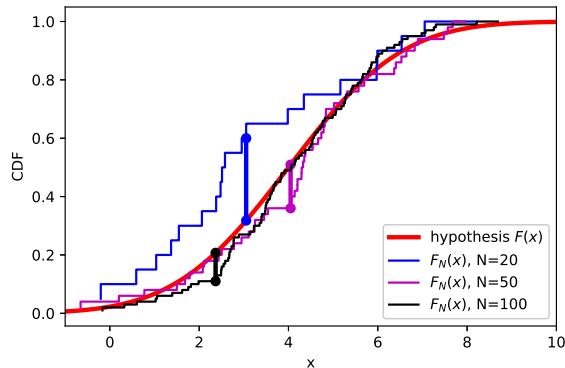
On pourrait diminuer α et donc ici, augmenter (vers la droite) la **borne de décision** u pour éviter de **rejeter à tort** H_0 , mais alors on augmenterait d'autant le risque de **garder** H_0 **alors qu'elle est fausse** (pour poursuivre l'analyse, il faudrait alors spécifier une hypothèse alternative H_1 et étudier le risque de rejeter H_1 à tort ou risque de seconde espèce)

Test de Kolmogorov-Smirnov

Test non-paramétrique qui compare une distribution théorique de CDF (supposée) F et un échantillon, via la Statistique :

$$t(\mathbb{X}) = \max_{x_i \in \mathbb{X}} |F_N(x_i) - F(x_i)|$$

où F_N est la CDF empirique (construite à partir de l'échantillon)



On peut obtenir l'expression qui donne $\Pr(t \geq u_N | H_0 : X \sim F)$ en fonction de la taille N de l'échantillon (voir figure de droite)

Ici (par exemple), on a généré trois échantillons selon F et on réalise trois tests :

$p_{N=20} = 0.06$, $p_{N=50} = 0.10$, $p_{N=100} = 0.18$. On ne peut rejeter H_0 dans aucun des trois cas (pour $\alpha = 5\%$)



A considérer...

- L'**outil de probabilité** est construit et dispose des **propriétés** nécessaires au partage d'une quantité totale (masse de probabilité) entre un ensemble d'**issues** ou de **valeurs possibles** d'une grandeur d'intérêt
- Cette **distribution** s'effectue sur la base de l'**exploitation** de **données** et/ou d'**hypothèses** : elle y est conditionnée et en résulte.
- En pratique ceci s'effectue au travers des règles de **manipulation** algébrique des **probabilités** (conditionnement, probabilités totales, relation de Bayes) et de diverses **stratégies** employées dans le **champ statistique** (minimisation de l'erreur empirique, maximisation de la vraisemblance des données, mise à jour Bayésienne) permettant l'identification de **modèles** pouvant ensuite être interrogés pour prédire (ou **inférer**)
- Il est possible de **tester les choix** associés à différentes hypothèses : souvent en étudiant la probabilité (résultante) d'observer les données **effectivement observées**, **conditionnellement** à de telles hypothèses



4. Références



L. J. Savage.

The foundations of statistics (2nd ed).

Dover Publications, 1972.



J. O. Berger.

Statistical decision theory and bayesian analysis (2nd ed).

Springer Verlag, 1985.



T. Bedford and R. Cooke.

Probabilistic risk analysis : foundation and methods.

Cambridge University Press, 2001.



G. Saporta.

Probabilités, analyse des données et statistique (2nd ed).

Edition technip, 2006.



T. Hastie, R. Tibshirani, and J. Friedman.

The elements of statistical learning : data mining, inference and prediction (2nd ed).

Springer, 2009.



J. Garnier.

Gestion des incertitudes et analyse de risque (support de cours).

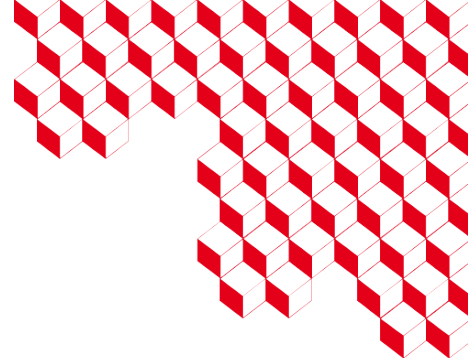
MAP 568 - Département de mathématiques appliquées, école polytechnique, 2017.



J. Garnier, S. Méléard, and N. Touzi.

Aléatoire (support de cours).

Département de mathématiques appliquées, école polytechnique, 2021.



MERCI

William Fauriat
CEA DAM Île-de-France
Bruyères-le-Châtel
91297 Arpajon cedex