

Statistiques et probabilités

Cours n°5

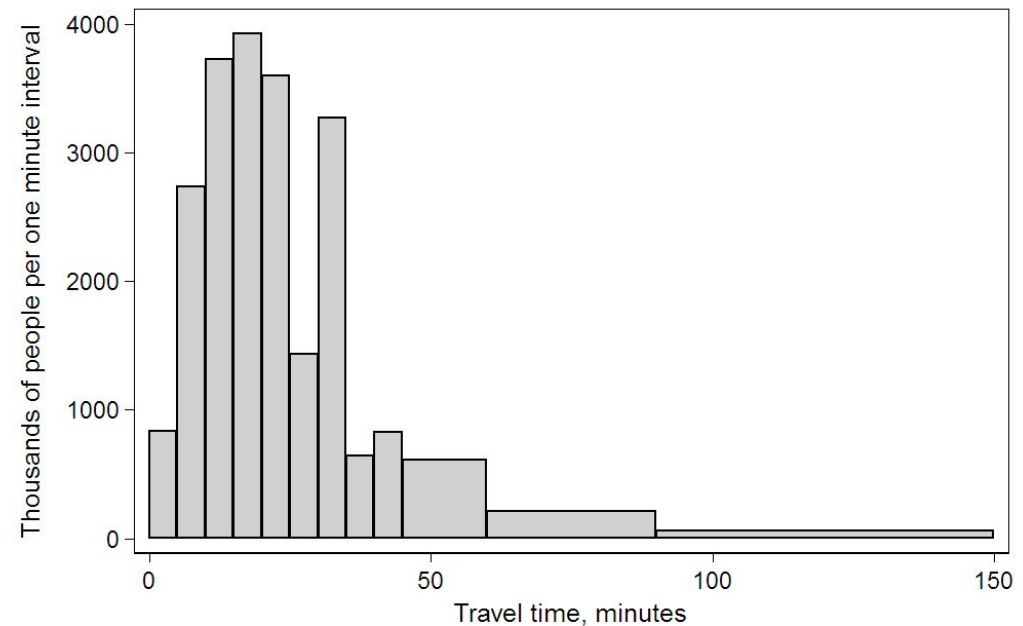
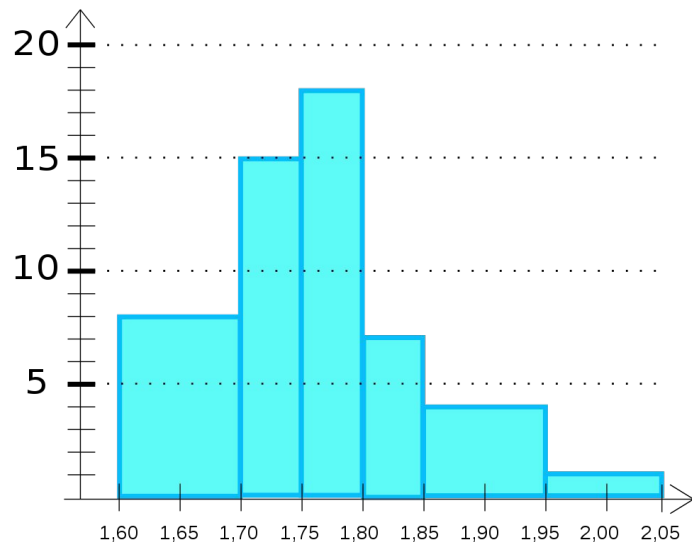
Guillaume Postic

Université Paris-Saclay, Univ. Evry
Département informatique

Master 1 MIAGE - 2023/2024

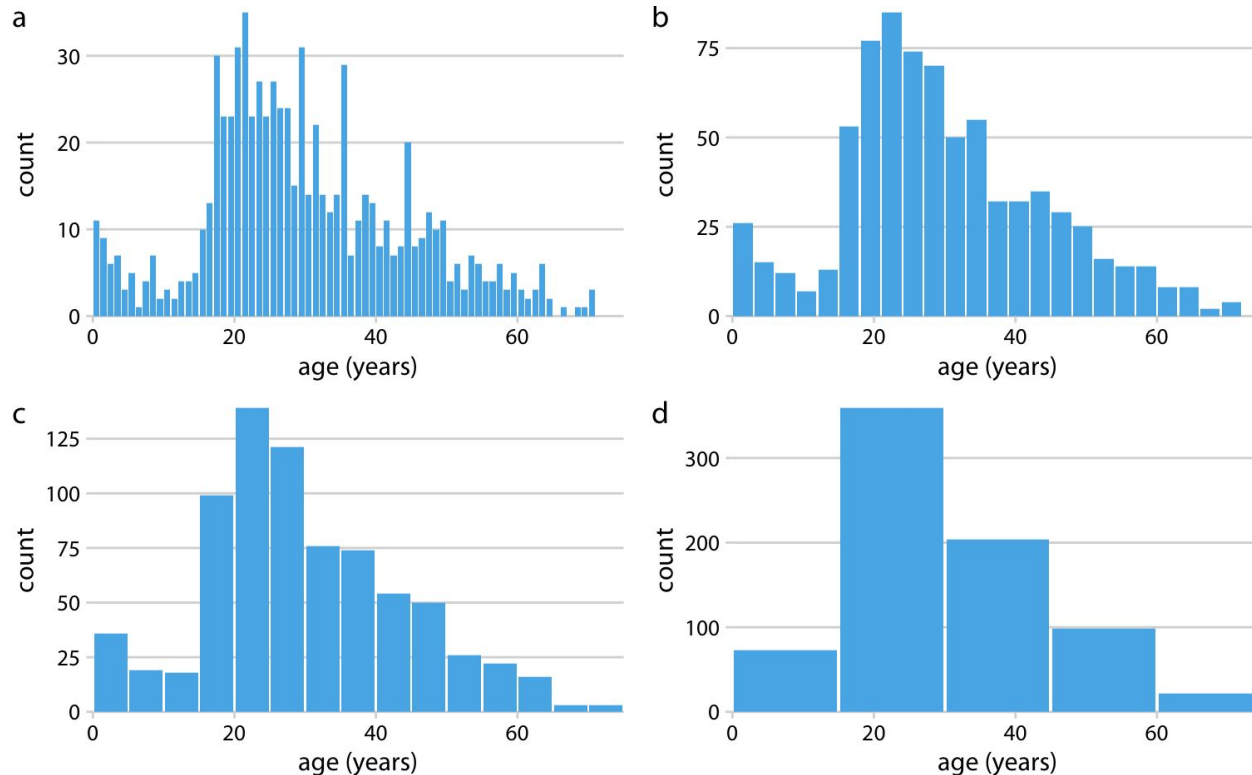
Estimation de la densité : histogramme

Représentation graphique approximant la distribution d'une variable aléatoire par groupement des données en classes représentées par des colonnes **contiguës**.



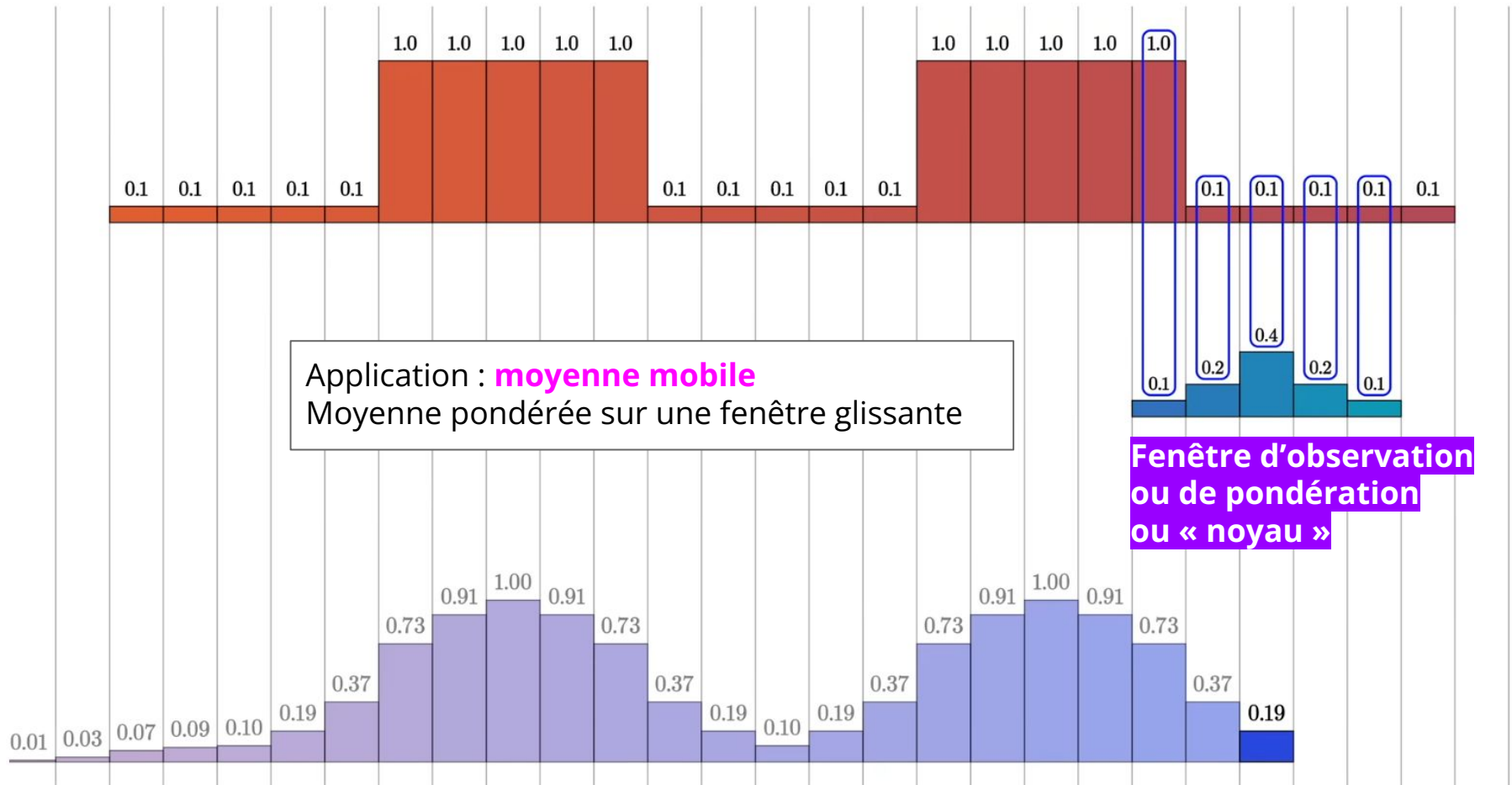
Estimation de la densité : histogramme

Représentation d'un jeu de données avec différentes largeurs d'intervalles :



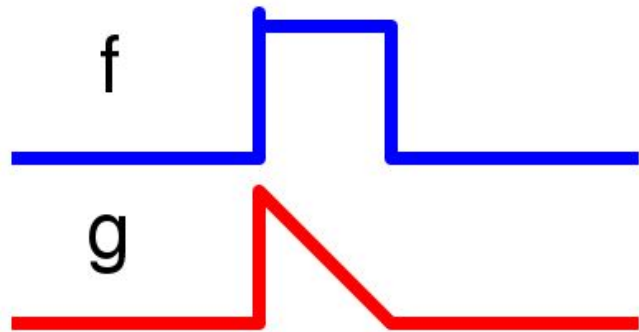
💡 Quelle que soit la largeur des intervalles, un histogramme est un estimateur de la densité de probabilité sous-jacente par **maximum de vraisemblance** (cf. définition d'un modèle ci-après)

Rappel : convolution discrète



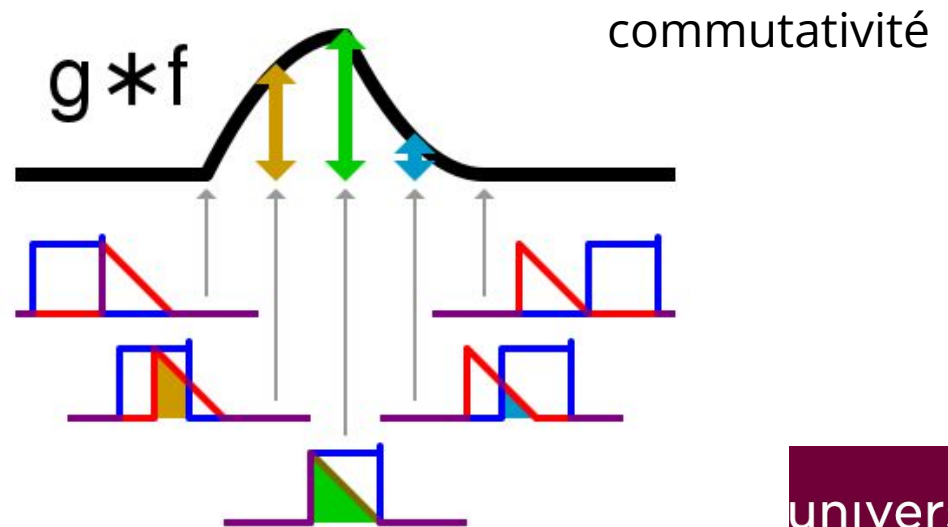
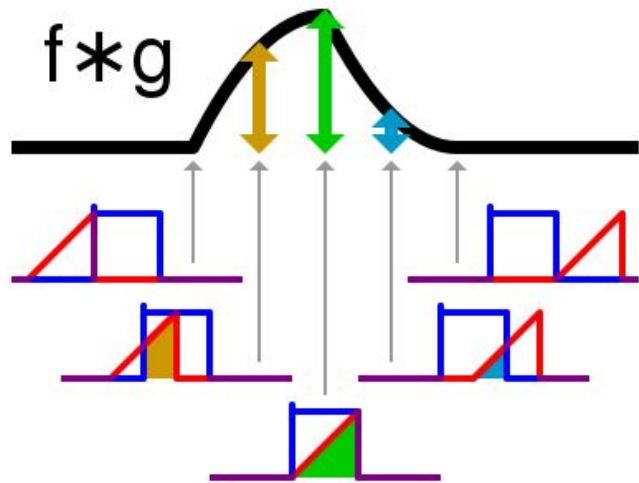
Effet de **lissage**
Exemple : flou gaussien en analyse d'images

Convolution continue (1)



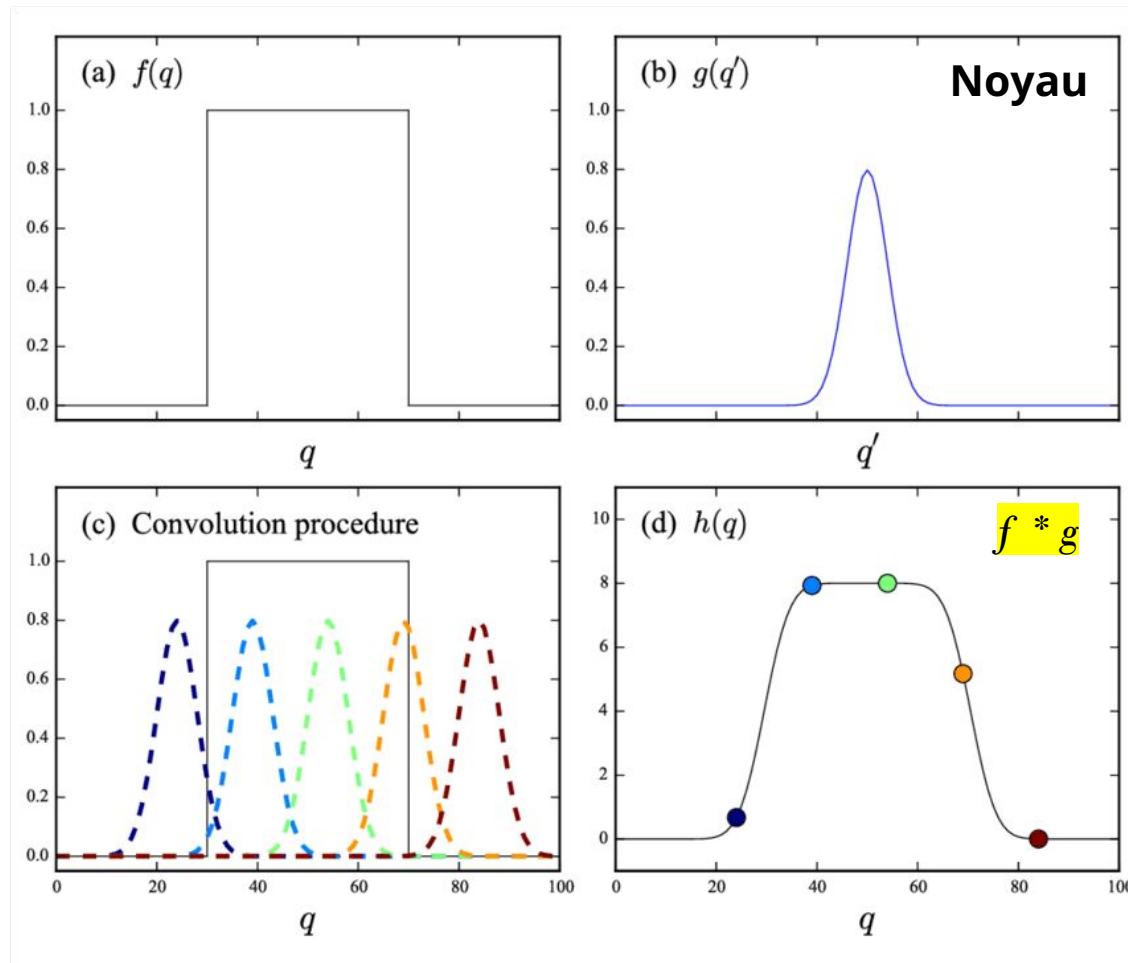
$$(f * g)(t) := \int_{-\infty}^{\infty} f(\tau)g(t - \tau) d\tau$$

$$:= \int_{-\infty}^{\infty} f(t - \tau)g(\tau) d\tau$$



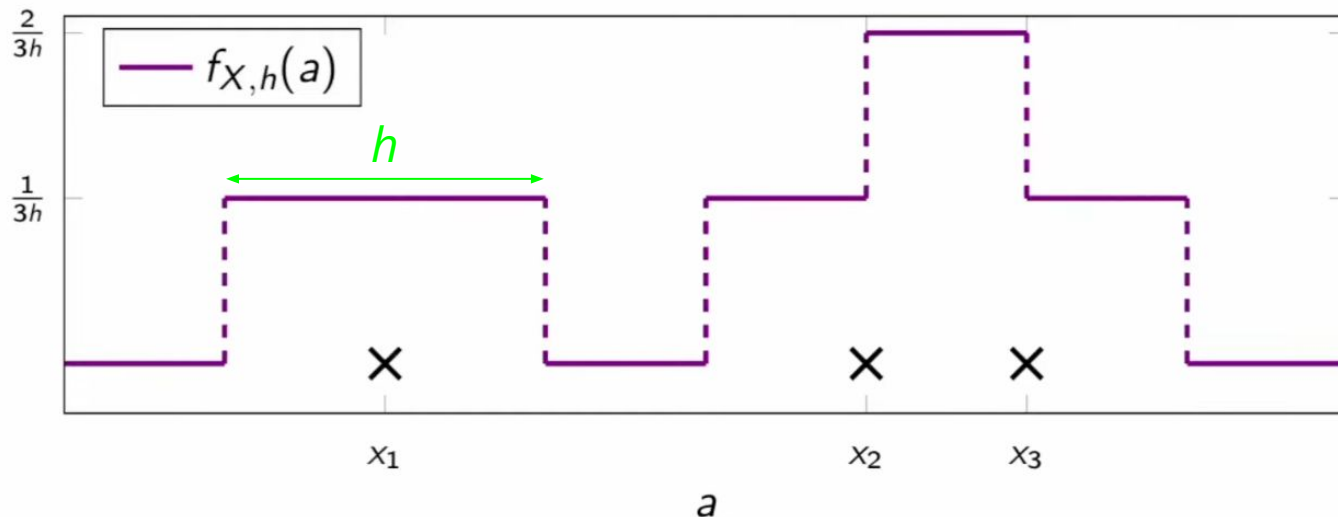
Convolution continue (2)

Moyenne mobile et lissage



Estimation de la densité par noyau

- Aussi appelée méthode de Parzen-Rosenblatt
- **Généralisation de l'estimation par histogramme**
- Deux étapes principales :
 - Sur chaque point du jeu de données, un noyau est centré
 - On calcule la **somme des fonctions noyaux** pour obtenir l'estimation de la densité f
- Exemple avec trois points et un noyau uniforme (ou rectangle) :

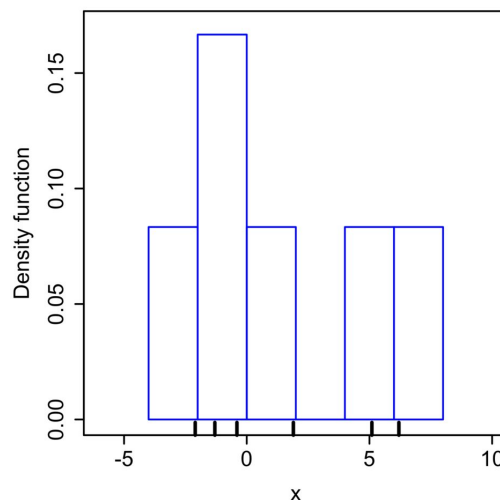


Estimation de la densité par noyau

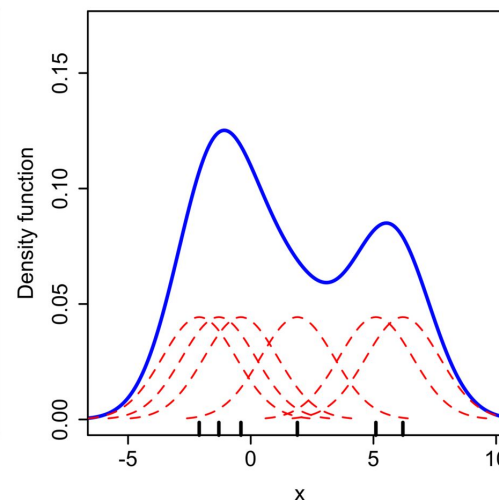
$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - x_i) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)$$

- Différents types de noyaux K peuvent être utilisés → **influe peu**
- Différentes méthodes pour choisir le paramètre de lissage $h > 0$, ou *fenêtre* (c.-à-d., la « largeur du noyau ») → **influe beaucoup**
 - Pour un noyau gaussien, la *fenêtre* $h > 0$ est égale à l'écart type

Histogramme



KDE



Variable qualitative quantitative

- Variable aléatoire quantitative
 - Quantitative continue
 - Mesure de longueur, de temps...
 - Quantitative discrète
 - Résultat d'un dé, longueur discrétisée...
- Variable aléatoire qualitative
 - Qualitative catégorielle (ou nominale)
 - Couleurs (bleu, vert, rouge)...
 - Qualitative ordinale
 - Lettres (A, B, C), mois de l'année...

Variable qualitative ordinale

- ⚠ Catégories ordonnées, mais les distances qui les sépareraient sur une échelle ne sont pas définies

Nul ! OK Bien Génial !



Nul ! OK Bien Génial !



Nul ! OK Bien Génial !

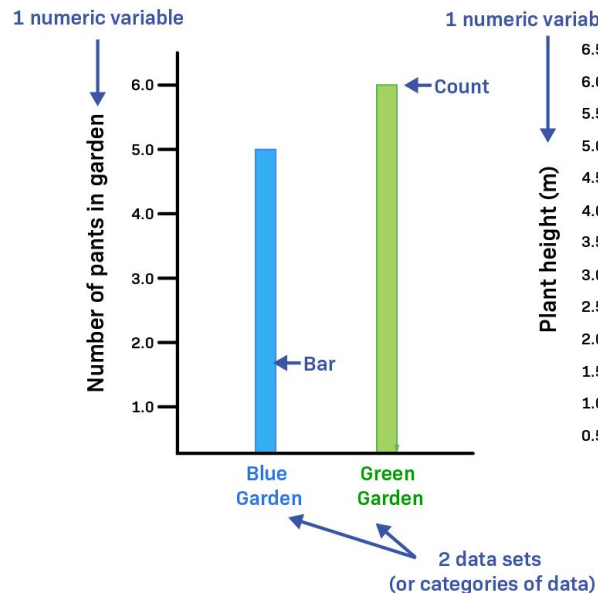


Diagramme en bâtons (*bar plot/chart*)

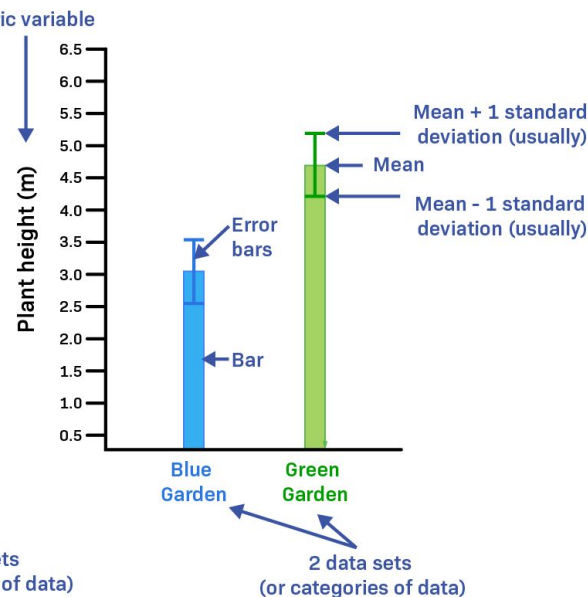
- Graphique qui présente des **variables qualitatives** avec des barres rectangulaires avec des hauteurs ou des longueurs proportionnelles aux valeurs qu'elles représentent.
- Largeur des barres arbitraire et identique pour toutes

⚠ **Ne pas confondre avec histogramme**

(A) Barplot representing amounts




(B) Barplot summarizing data sets




Variables qualitatives

- **Indicateurs de tendance centrale**

-  La moyenne et la médiane ne sont pas calculables
- Le **mode** peut être calculé : catégorie la plus représentée
 - Aussi calculable pour les variables quantitatives
 - **Distributions multimodales**

- **Indicateurs de dispersion**

-  La variance et l'écart-type ne sont pas calculables
- **Étendue** (*range*)
 - = valeur la plus haute – valeur la moins haute
- IQR
- Entropie (Shannon)

Analyses multivariées (1)

- S'intéressent à des **lois de probabilité à plusieurs variables (dites jointes)**
- Rappel : si X_1, X_2, \dots, X_n n tirages aléatoires indépendants, chacun avec $k = 2$ résultats possibles (Bernoulli), alors

$$N = \sum_k X_k \sim B(n, p)$$

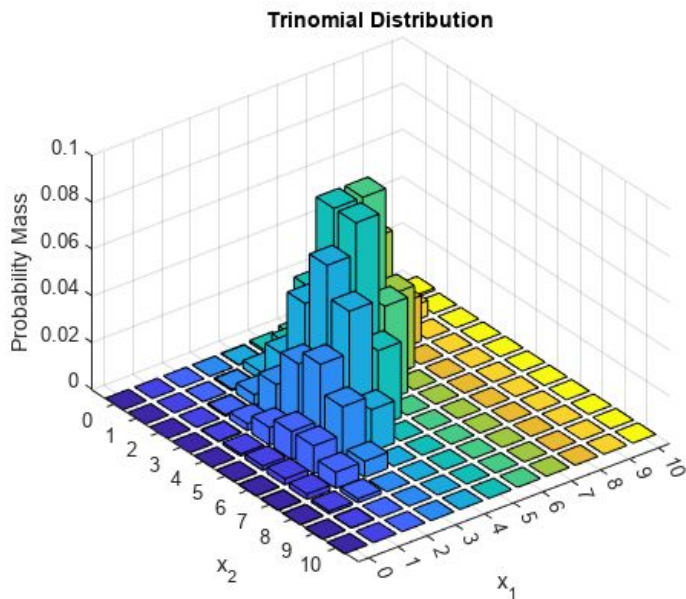
- La loi binomiale modélise la probabilité du nombre de succès : une seule variable N

Analyses multivariées (2)

- Soit un dé équilibré à six faces, que l'on lance n fois
- On définit $c = 3$ trois évènements (catégories ou classes) :
 - e_1 : valeur ≥ 4
 - e_2 : valeur $\in \{2, 3\}$
 - e_3 : valeur $= 1$
- Soit le vecteur à $c = 3$ dimensions \mathbf{X} , tq $\mathbf{X} = (X_1, X_2, X_3)$
 - avec X_i le nombre de réalisations de l'évènement e_i (succès pour la catégorie i)
- On dit que \mathbf{X} est une « variable multinomiale »
 - Elle suit la loi multinomiale, qui donne la probabilité du nombre de succès pour chacune des c catégories
 - Par exemple, pour $n = 10$ lancers, quelle est la probabilité de faire $x_1 = 7, x_2 = 2$ et $x_3 = 1$?
- Note : pour $n = 1$, \mathbf{X} est appelée « variable catégorielle »

Analyses multivariées (3)

- Les probabilités des événements sont : $p(e_1) = \frac{1}{2}$, $p(e_2) = \frac{1}{3}$ et $p(e_3) = \frac{1}{6}$
- Pour $n = 10$ lancers, on a donc $E[\mathbf{X}] = (5 ; 3,3 ; 1,7)$



Note : X_3 est déterminé par la contrainte
 $X_1 + X_2 + X_3 = n$

Exemple d'un tirage \mathbf{x} de \mathbf{X}

val. ≥ 4	val. $\in \{2, 3\}$	val. $= 1$	} $n = 10$
0	1	0	
1	0	0	
1	0	0	
1	0	0	
0	1	0	
1	0	0	
1	0	0	
0	0	1	
1	0	0	
7	2	1	

Analyses multivariées (4)

- **Loi multinomiale**

- Fonction de masse :

$$f(x_1, \dots, x_k; n, p_1, \dots, p_k) = \Pr(X_1 = x_1 \text{ et } \dots \text{ et } X_k = x_k)$$

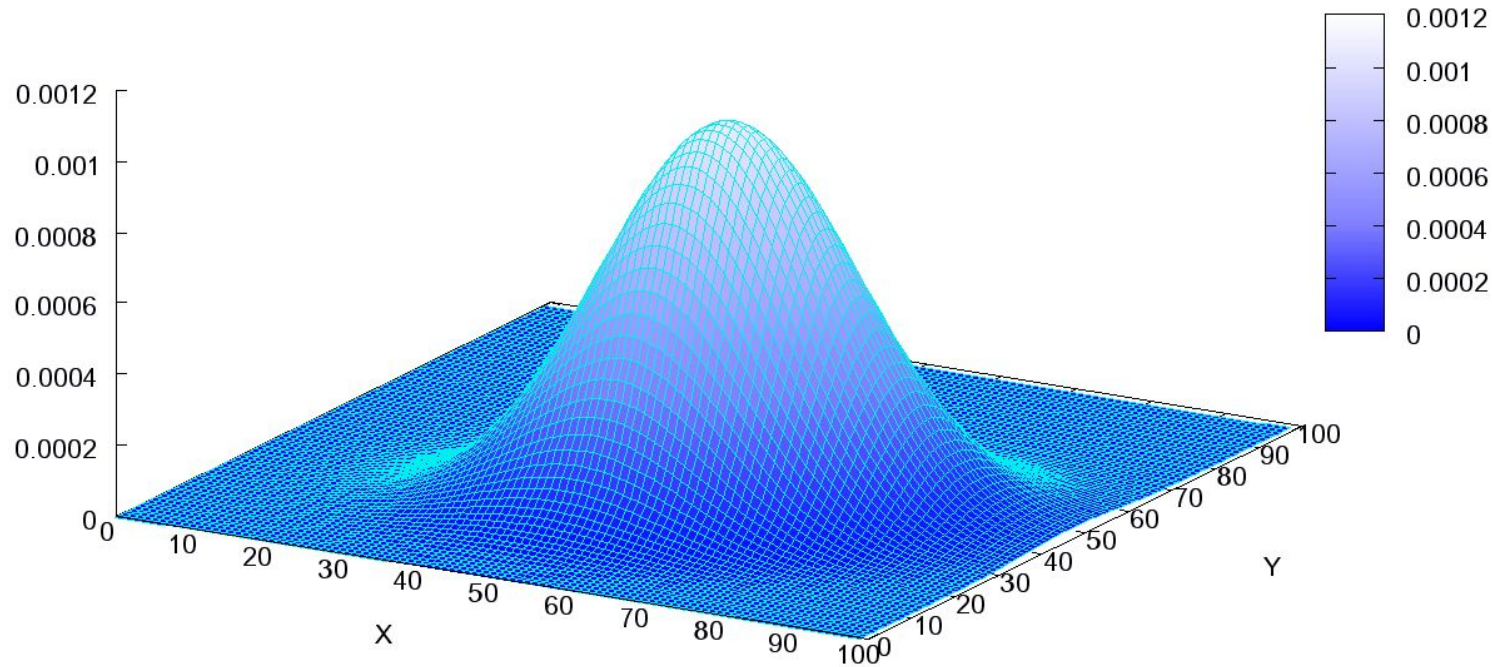
$$= \begin{cases} \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \times \dots \times p_k^{x_k}, & \text{quand } \sum_{i=1}^k x_i = n \\ 0 & \text{autrement,} \end{cases}$$

Coefficient multinomial

- Pas de représentation graphique au-delà de 3 (ou 4) catégories
- $E[X_i] = np_i$
- $\text{Var}(X_i) = np_i(1 - p_i)$
- $\text{Cov}(X_i, X_j) = -np_i p_j \ (i \neq j)$

Autres lois de probabilité jointes

Multivariate Normal Distribution



Modèle bivarié (1)

Exemple 1

Jet de deux dés à 6 faces : premier X , second Y

Deux lois uniformes

Table des probabilités jointes

$X \backslash Y$	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

Modèle bivarié (2)

Exemple 2

Jet de deux dés à 6 faces : premier X , total (somme) T

Une loi uniforme et une loi triangle

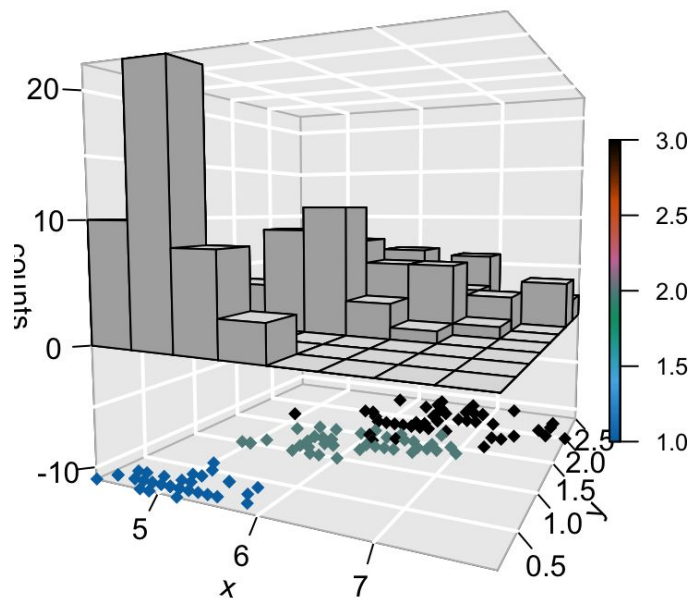
Table des probabilités jointes

$X \backslash T$	2	3	4	5	6	7	8	9	10	11	12
1	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0	0
2	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0
3	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0
4	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0
5	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0
6	0	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36

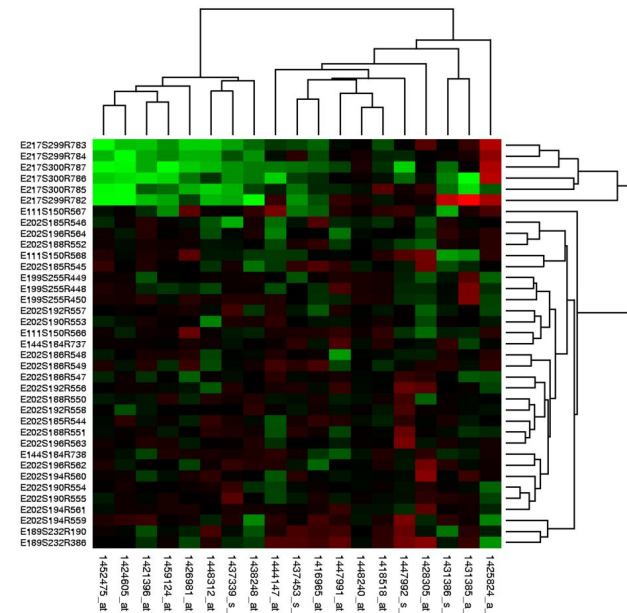
Modèle bivarié (3)

Représentations graphiques

Histogramme



Carte thermique (*heat map*)



Modèle bivarié (4)

Variables discrète et nominale : **table de contingence**

	Alice	Bob	Charlie	David	Classe
A	4	0	1	2	7
B	1	1	1	2	5
C	0	1	2	0	3
D	0	2	1	0	3
E	0	0	0	1	1
F	0	1	0	0	1
	5	5	5	5	20

- Probabilité **jointe** : $p(A \cap \text{Alice}) = 4/20 = 1/5$
- Probabilité **conditionnelle** :
 - $p(A | \text{Alice}) = p(A \cap \text{Alice}) / p(\text{Alice}) = (4/20) / (5/20) = 4/5$
- Probabilités **marginales** :
 - $p(A) = 7/20$ et $p(\text{Alice}) = 5/20 = 1/4$

Modèle bivarié (5)

Propriétés des fonctions de masse et densité jointes

Cas discret

1. $0 \leq p(x_i, y_j) \leq 1$

2. Probabilité totale vaut 1

$$\sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) = 1$$

Cas continu

1. $0 \leq f(x, y)$

2. Probabilité totale vaut 1

$$\int_c^d \int_a^b f(x, y) dx dy = 1$$

Note : $f(x, y)$ peut être plus grand que 1,

car **c'est une densité, pas une probabilité**

Covariance

Mesure du degré avec lequel deux variables aléatoires varient conjointement.

Par exemple : le poids et la taille d'individus

Pour X et Y variables aléatoires de moyennes μ_X et μ_Y

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

Covariance

Propriétés

1. $\text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y)$, pour des constantes a, b, c, d
2. $\text{Cov}(X_1 + X_2, Y) = \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y)$
3. $\text{Cov}(X, X) = \text{Var}(X)$
4. $\text{Cov}(X, Y) = E(XY) - \mu_X \mu_Y$
5. Si X et Y sont indépendants, alors $\text{Cov}(X, Y) = 0$
6. ⚠ La réciproque n'est pas vraie : même si la covariance est nulle, les variables peuvent ne pas être indépendantes

Corrélation (linéaire)

La corrélation entre plusieurs variables aléatoires est une notion de liaison qui contredit leur indépendance.

Le **coefficient de corrélation (de Pearson)** entre X et Y est défini par :

$$\text{Cor}(X, Y) = \rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

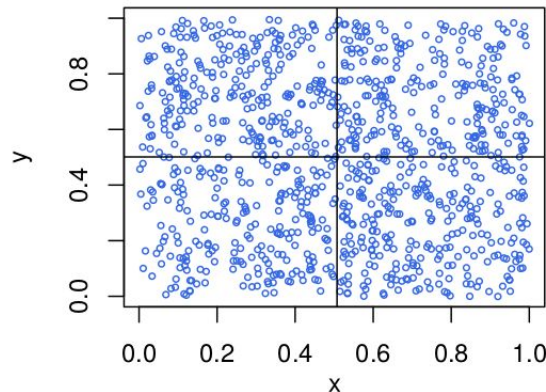
Propriétés :

1. ρ est la covariance des versions centrées réduites de X et Y
2. ρ est adimensionnelle
3. $-1 \leq \rho \leq 1$
 $\rho = 1$ ssi $Y = aX + b$ avec $a > 0$
et $\rho = -1$ ssi $Y = aX + b$ avec $a < 0$

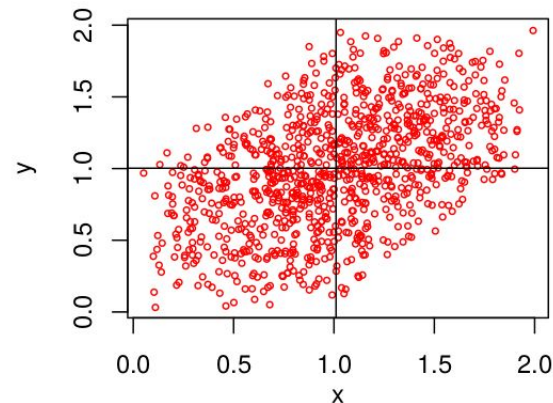
Nuage de points (*scatter plot*)

Représentation : nuage de points (*scatter plot*)

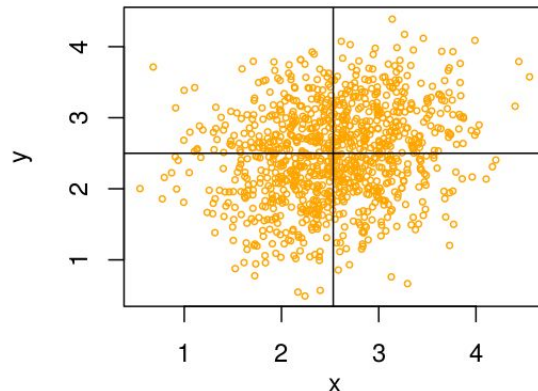
(1, 0) $\text{cor}=0.00$, $\text{sample_cor}=-0.07$



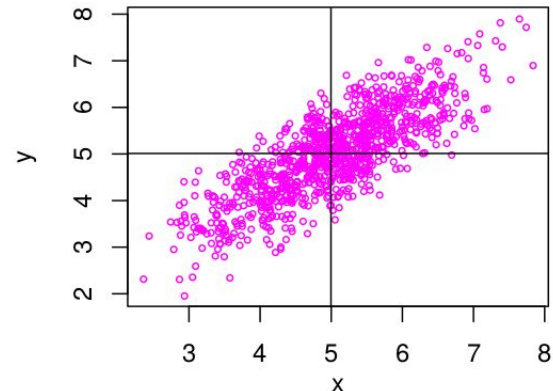
(2, 1) $\text{cor}=0.50$, $\text{sample_cor}=0.48$



(5, 1) $\text{cor}=0.20$, $\text{sample_cor}=0.21$

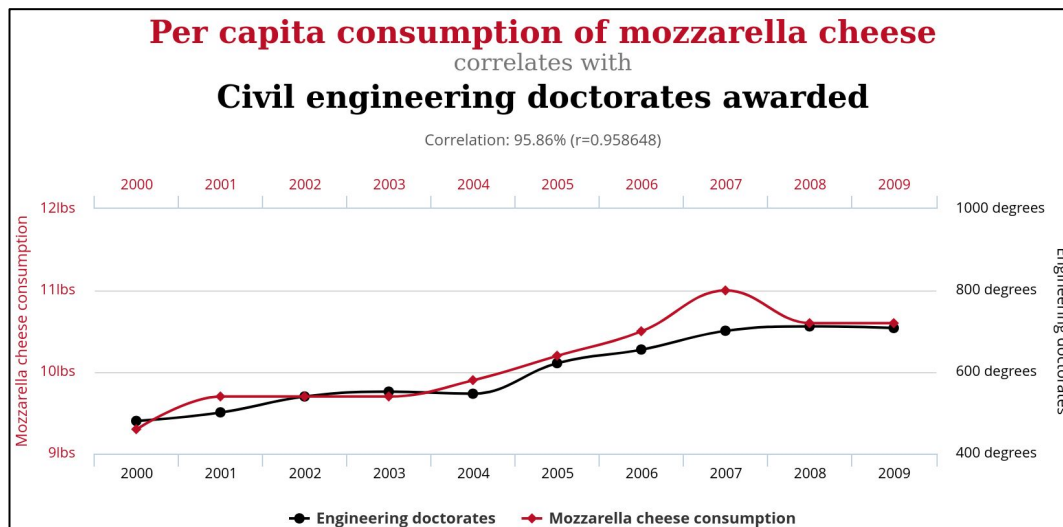
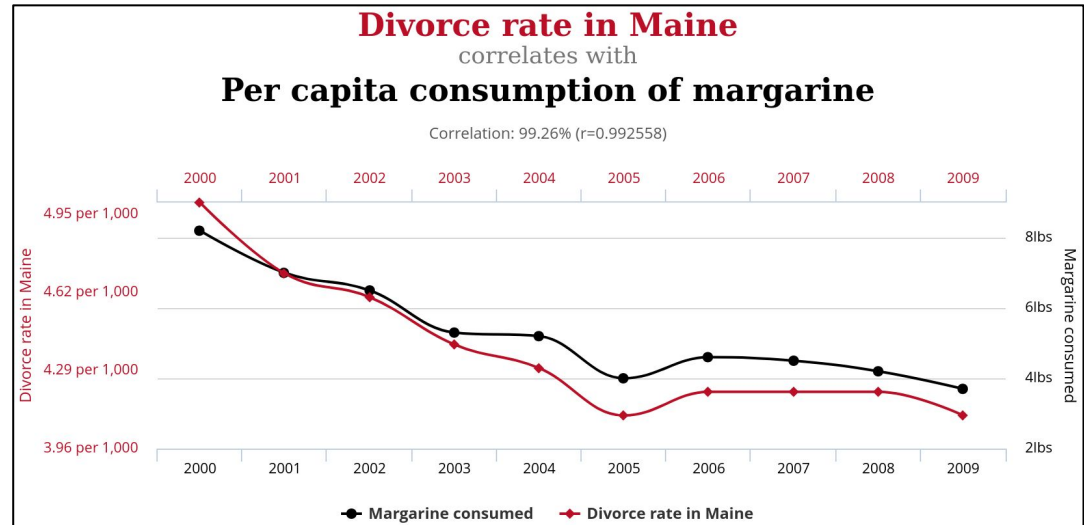


(10, 8) $\text{cor}=0.80$, $\text{sample_cor}=0.81$



Corrélations fallacieuses (*spurious*)

Un coefficient de corrélation élevé n'induit pas nécessairement une **relation de causalité** entre les deux phénomènes mesurés.



Transformation des données

