Chapitre 4 : Introduction à la statistique descriptive

La statistique, qu'est-ce que c'est ?

La statistique (que l'on écrit au singulier en français, bien que cela se traduise par *statistics* en anglais) est une discipline qui consiste à analyser et interpréter des données, dans le but :

La statistique, qu'est-ce que c'est ?

La statistique (que l'on écrit au singulier en français, bien que cela se traduise par *statistics* en anglais) est une discipline qui consiste à analyser et interpréter des données, dans le but :

- De prédire des réalisations futures d'expériences à partir de données passées, par exemple en vue de prendre des décisions (*inférence statistique*),

La statistique, qu'est-ce que c'est ?

La statistique (que l'on écrit au singulier en français, bien que cela se traduise par *statistics* en anglais) est une discipline qui consiste à analyser et interpréter des données, dans le but :

- De prédire des réalisations futures d'expériences à partir de données passées, par exemple en vue de prendre des décisions (*inférence statistique*),
- De décrire les caractéristiques d'une population à partir d'une petite proportion de celle-ci (*échantillonnage statistique*),

La statistique, qu'est-ce que c'est ?

La statistique (que l'on écrit au singulier en français, bien que cela se traduise par *statistics* en anglais) est une discipline qui consiste à analyser et interpréter des données, dans le but :

- De prédire des réalisations futures d'expériences à partir de données passées, par exemple en vue de prendre des décisions (*inférence statistique*),
- De décrire les caractéristiques d'une population à partir d'une petite proportion de celle-ci (*échantillonnage statistique*),
- De décrire des caractéristiques des données, en les classant par exemple selon des critères (*statistique descriptive*).

La statistique, qu'est-ce que c'est ?

La statistique (que l'on écrit au singulier en français, bien que cela se traduise par *statistics* en anglais) est une discipline qui consiste à analyser et interpréter des données, dans le but :

- De prédire des réalisations futures d'expériences à partir de données passées, par exemple en vue de prendre des décisions (*inférence statistique*),
- De décrire les caractéristiques d'une population à partir d'une petite proportion de celle-ci (*échantillonnage statistique*),
- De décrire des caractéristiques des données, en les classant par exemple selon des critères (*statistique descriptive*).

Nous allons nous concentrer sur la statistique descriptive, et le reste sera abordé dans le cours ≪ probabilités et statistique 2 ≫.

Un peu de terminologie

On suppose qu'on dispose d'une famille de données $X = (x_1, \ldots, x_n)$, toutes à valeurs dans un même espace E. La famille de données peut être vue comme une variable aléatoire $X : \Omega \to E$, pour $\Omega = \{1, \ldots, n\}$. En statistique, l'ensemble Ω est appelé population, ses éléments sont appelés individus, et la variable aléatoire X représente une caractéristique de la population.

Un peu de terminologie

On suppose qu'on dispose d'une famille de données $X = (x_1, \ldots, x_n)$, toutes à valeurs dans un même espace E. La famille de données peut être vue comme une variable aléatoire $X : \Omega \to E$, pour $\Omega = \{1, \ldots, n\}$. En statistique, l'ensemble Ω est appelé population, ses éléments sont appelés individus, et la variable aléatoire X représente une caractéristique de la population.

Exemple : les notes d'élèves à un examen.

Sylvain DUBOIS (1)	8	Mélodie COURBET (7)	9
Anna CONDA (2)	0.75	Cyril PLESSIS (8)	7
Maud ZARELLA (3)	11	Harry COVERT (9)	8.5
Rémy DUTERTRE (4)	11.25	Kim ONO (10)	6.25
Charles ATTAN (5)	8.5	Phil DEFERT (11)	7.5
Barbara GARDET (6)	13.5	Luc ARNE (12)	11

Un peu de terminologie

On suppose qu'on dispose d'une famille de données $X = (x_1, \ldots, x_n)$, toutes à valeurs dans un même espace E. La famille de données peut être vue comme une variable aléatoire $X : \Omega \to E$, pour $\Omega = \{1, \ldots, n\}$. En statistique, l'ensemble Ω est appelé population, ses éléments sont appelés individus, et la variable aléatoire X représente une caractéristique de la population.

Exemple : les notes d'élèves à un examen.

Sylvain DUBOIS (1)	8	Mélodie COURBET (7)	9
Anna CONDA (2)	0.75	Cyril PLESSIS (8)	7
Maud ZARELLA (3)	11	Harry COVERT (9)	8.5
Rémy DUTERTRE (4)	11.25	Kim ONO (10)	6.25
Charles ATTAN (5)	8.5	Phil DEFERT (11)	7.5
Barbara GARDET (6)	13.5	Luc ARNE (12)	11

- La population est l'ensemble des élèves, que l'on identifie à $\{1, \ldots, 12\}$. L'ensemble des données est $(8, 0.75, 11, 11.25, 8.5, 13.5, 9, 7, 8.5, 6.25, 7.5, 11)$.

Un peu de terminologie

On suppose qu'on dispose d'une famille de données $X = (x_1, \ldots, x_n)$, toutes à valeurs dans un même espace E. La famille de données peut être vue comme une variable aléatoire $X : \Omega \to E$, pour $\Omega = \{1, \ldots, n\}$. En statistique, l'ensemble Ω est appelé population, ses éléments sont appelés individus, et la variable aléatoire X représente une caractéristique de la population.

Exemple : les notes d'élèves à un examen.

Sylvain DUBOIS (1)	8	Mélodie COURBET (7)	9
Anna CONDA (2)	0.75	Cyril PLESSIS (8)	7
Maud ZARELLA (3)	11	Harry COVERT (9)	8.5
Rémy DUTERTRE (4)	11.25	Kim ONO (10)	6.25
Charles ATTAN (5)	8.5	Phil DEFERT (11)	7.5
Barbara GARDET (6)	13.5	Luc ARNE (12)	11

- La population est l'ensemble des élèves, que l'on identifie à $\{1, \ldots, 12\}$. L'ensemble des données est $(8, 0.75, 11, 11.25, 8.5, 13.5, 9, 7, 8.5, 6.25, 7.5, 11)$.
- La caractéristique est la note : par exemple, la note de Cyril PLESSIS est $X(8) = x_8 = 7$.

Ici on travaille sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ explicite : il s'agit de $\Omega = \{1, \dots, n\}$, muni de sa tribu triviale et de sa probabilité uniforme. La loi de la variable aléatoire $X : \Omega \to E$ est donc connue à l'avance : pour tout $x \in E$, l'événement $\{X = x\}$ se réécrit

$$\{X = x\} = \{i \in \{1, \dots, n\} \mid x_i = x\},$$

par conséquent, $\mathbb{P}(X = x)$ correspond à la proportion de la population dont la caractéristique vaut exactement x :

$$\mathbb{P}(X = x) = \frac{|\{i \in \{1, \dots, n\} \mid x_i = x\}|}{n}.$$

Ici on travaille sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ explicite : il s'agit de $\Omega = \{1, \dots, n\}$, muni de sa tribu triviale et de sa probabilité uniforme. La loi de la variable aléatoire $X : \Omega \to E$ est donc connue à l'avance : pour tout $x \in E$, l'événement $\{X = x\}$ se réécrit

$$\{X = x\} = \{i \in \{1, \dots, n\} \mid x_i = x\},$$

par conséquent, $\mathbb{P}(X = x)$ correspond à la proportion de la population dont la caractéristique vaut exactement x :

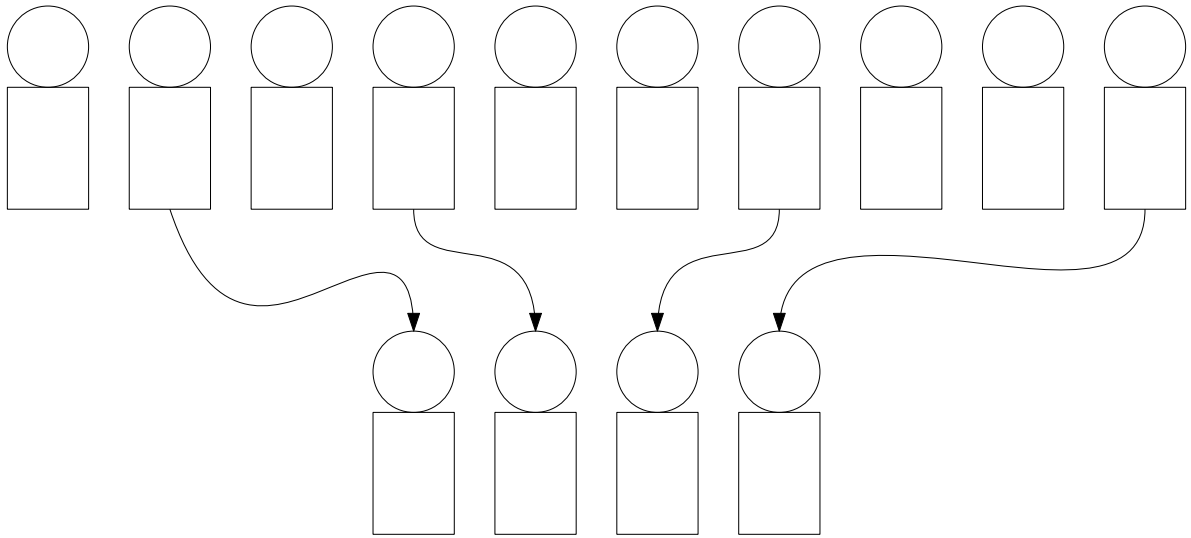
$$\mathbb{P}(X = x) = \frac{|\{i \in \{1, \dots, n\} \mid x_i = x\}|}{n}. \tag{1}$$

Exemple : Un archer tire 10 flèches en direction d'une cible. Pour chaque flèche, on note 1 si elle atteint la cible et 0 sinon. Le résultat des tirs donne $(1, 0, 0, 1, 0, 1, 1, 1, 0, 1)$. Cela peut se représenter par une variable aléatoire $X : \{1, \dots, 10\} \to \{0, 1\}$, et ici sa loi est donnée par

$$\mathbb{P}(X = 1) = \frac{6}{10} = \frac{3}{5}, \quad \mathbb{P}(X = 0) = \frac{4}{10} = \frac{2}{5}.$$

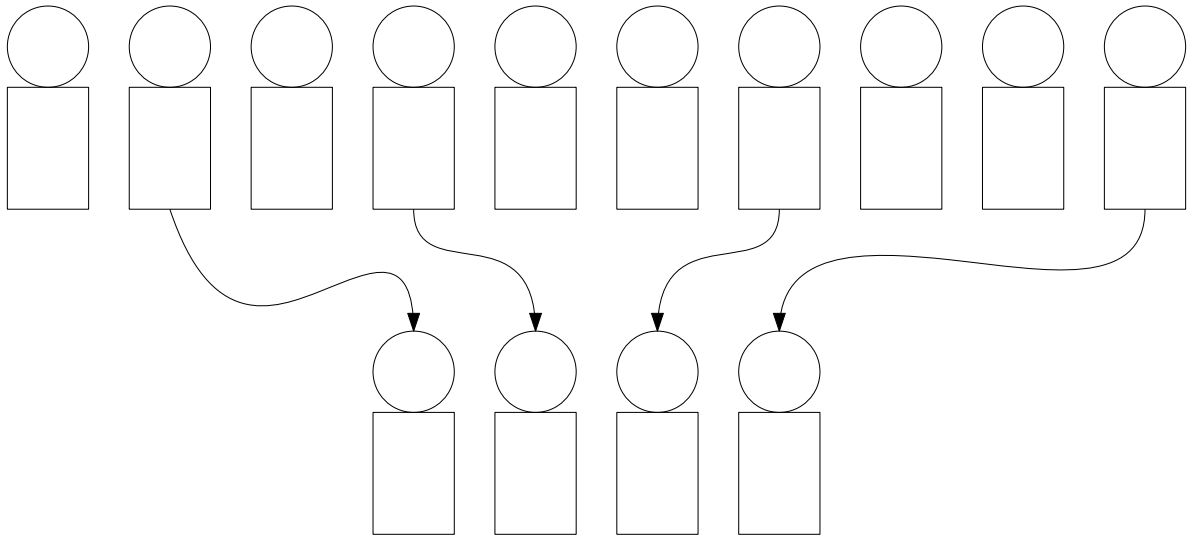
Echantillon ou population ?

Parfois la population est trop grande (ex : population mondiale, européenne, française) pour effectuer une mesure auprès de tout le monde. On choisit alors un sous-ensemble $\Omega' \subset \Omega$ de taille nettement inférieure, appelé échantillon.



Echantillon ou population ?

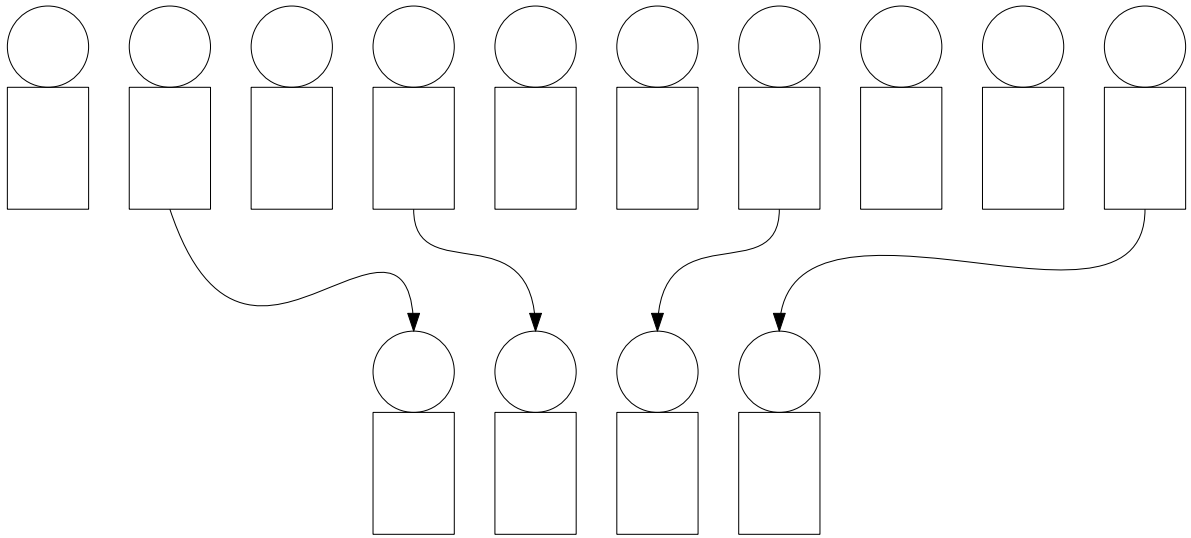
Parfois la population est trop grande (ex : population mondiale, européenne, française) pour effectuer une mesure auprès de tout le monde. On choisit alors un sous-ensemble $\Omega' \subset \Omega$ de taille nettement inférieure, appelé échantillon.



- Cela définit un nouvel espace de probabilité $(\Omega', \mathcal{F}', \mathbb{P}')$ en munissant Ω' de sa mesure de probabilité uniforme, et la caractéristique devient une variable aléatoire $X' : \Omega' \to E$.

Echantillon ou population ?

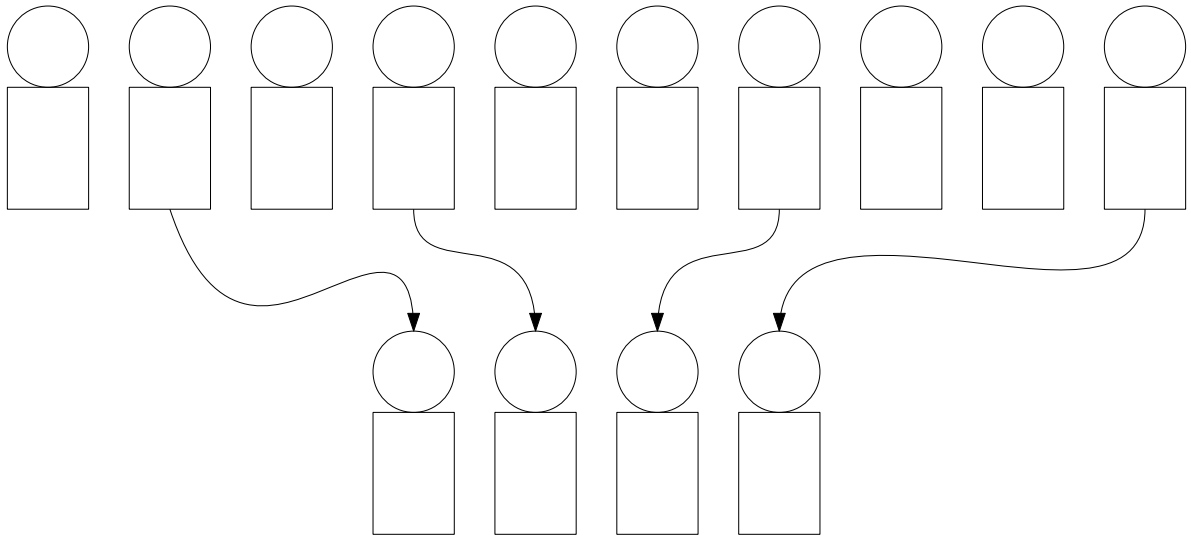
Parfois la population est trop grande (ex : population mondiale, européenne, française) pour effectuer une mesure auprès de tout le monde. On choisit alors un sous-ensemble $\Omega' \subset \Omega$ de taille nettement inférieure, appelé échantillon.



- Cela définit un nouvel espace de probabilité $(\Omega', \mathcal{F}', \mathbb{P}')$ en munissant Ω' de sa mesure de probabilité uniforme, et la caractéristique devient une variable aléatoire $X' : \Omega' \to E$.
- Le principe de l'échantillonnage est de trouver un moyen de choisir Ω' de sorte que la loi de X' soit $\ll$ proche $\gg$ de celle de X.

Echantillon ou population ?

Parfois la population est trop grande (ex : population mondiale, européenne, française) pour effectuer une mesure auprès de tout le monde. On choisit alors un sous-ensemble $\Omega' \subset \Omega$ de taille nettement inférieure, appelé échantillon.



- Cela définit un nouvel espace de probabilité $(\Omega', \mathcal{F}', \mathbb{P}')$ en munissant Ω' de sa mesure de probabilité uniforme, et la caractéristique devient une variable aléatoire $X' : \Omega' \to E$.
- Le principe de l'échantillonnage est de trouver un moyen de choisir Ω' de sorte que la loi de X' soit « proche » de celle de X.
- Dans ce cours on va considérer en général que l'échantillon est la population entière : $\Omega' = \Omega$.

Variables qualitatives et quantitatives

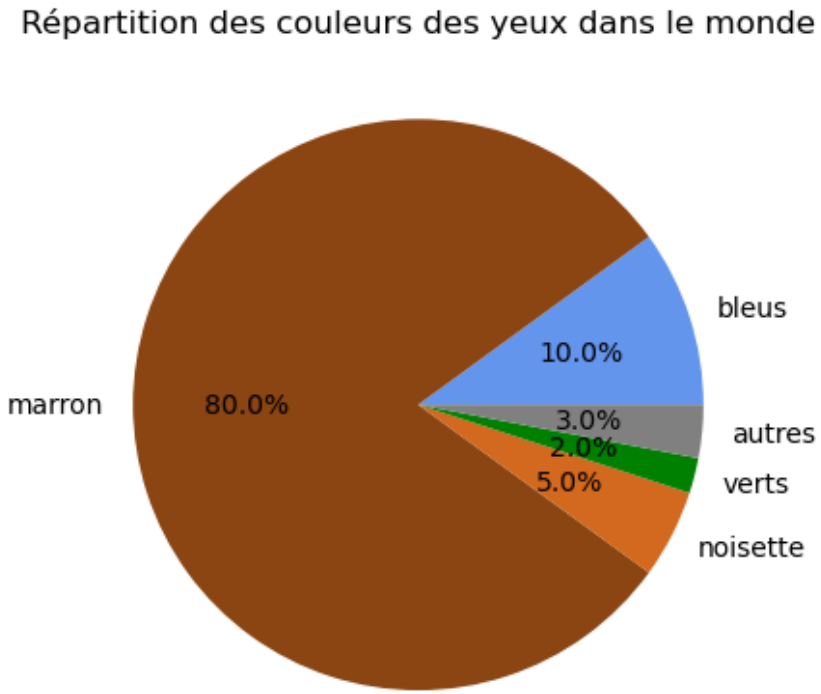
Si $(x_1, \ldots, x_n)$ est un échantillon, on appelle variables (ou données) les éléments de l'échantillon. Elles peuvent être de deux natures :

1. Quantitatives : à valeurs dans un espace mathématique usuel ($\mathbb{N}$, $\mathbb{R}$, $\mathbb{R}^n$)
2. Qualitatives/catégorielles : à valeurs dans un ensemble fini de catégories prédéfinies (couleurs, catégories sociales...)

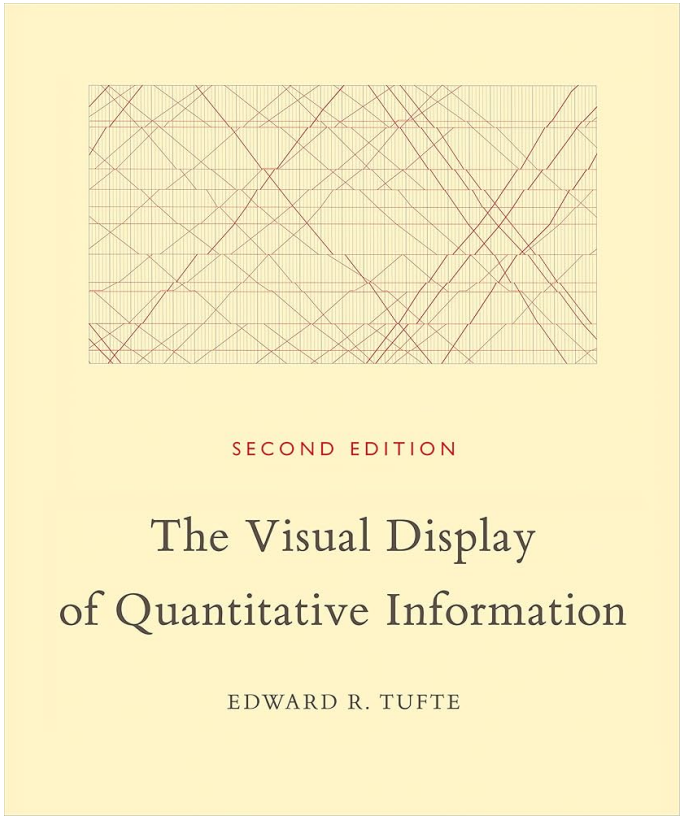
Variables qualitatives et quantitatives

Si $(x_1, \ldots, x_n)$ est un échantillon, on appelle variables (ou données) les éléments de l'échantillon. Elles peuvent être de deux natures :

1. Quantitatives : à valeurs dans un espace mathématique usuel ($\mathbb{N}$, $\mathbb{R}$, $\mathbb{R}^n$)
2. Qualitatives/catégorielles : à valeurs dans un ensemble fini de catégories prédéfinies (couleurs, catégories sociales...)



Comment représenter des variables ?



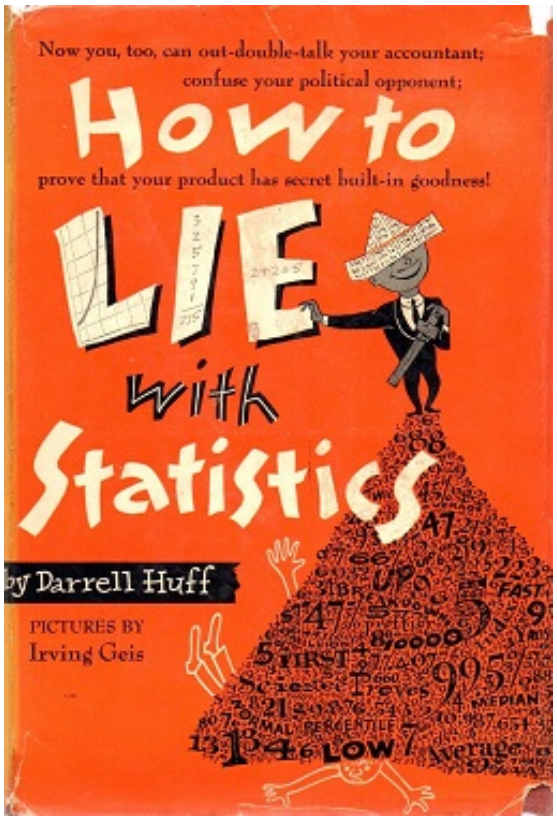


Diagramme sectoriel (communément appelé “camembert”)

Ce que vous pensez en cours de maths :

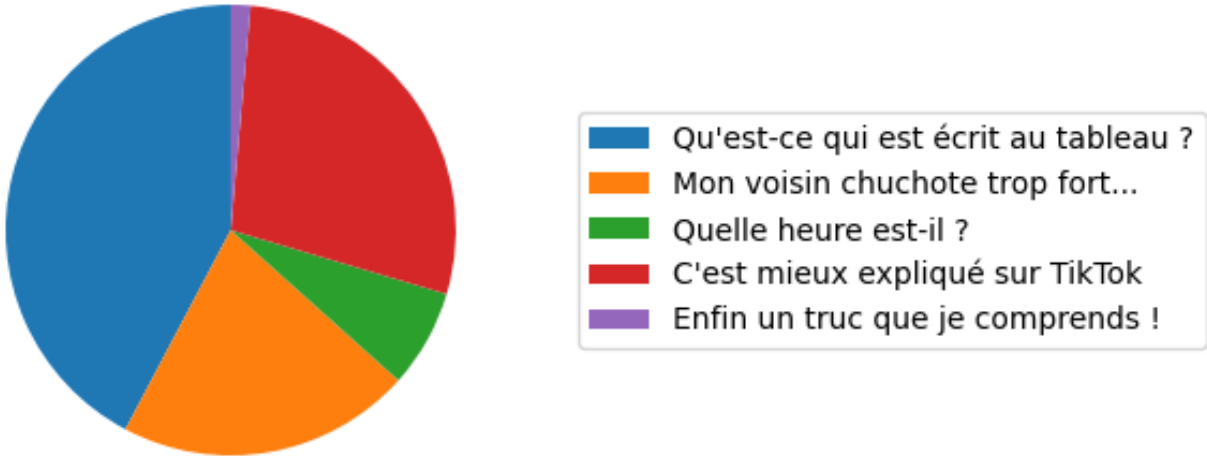
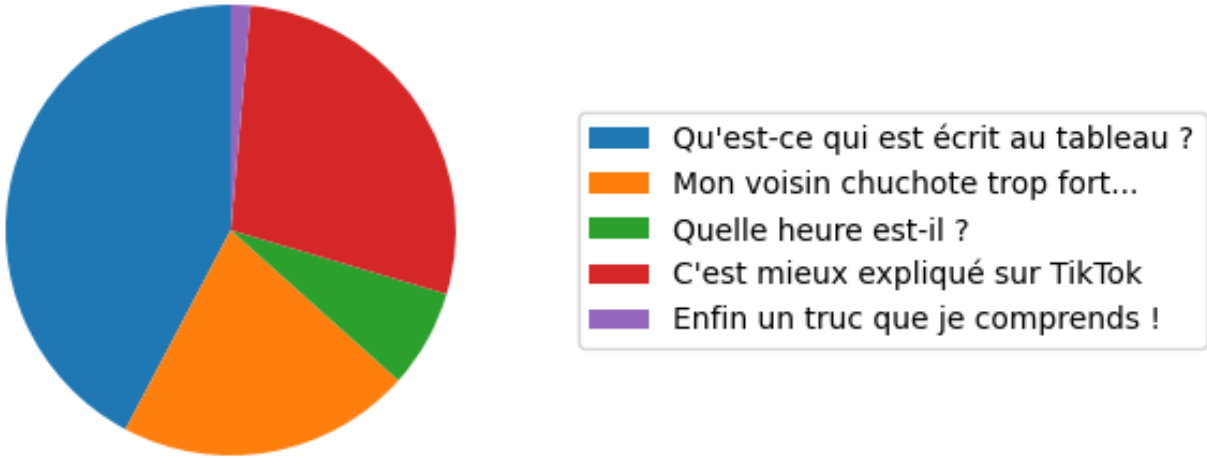


Diagramme sectoriel (communément appelé “camembert”)

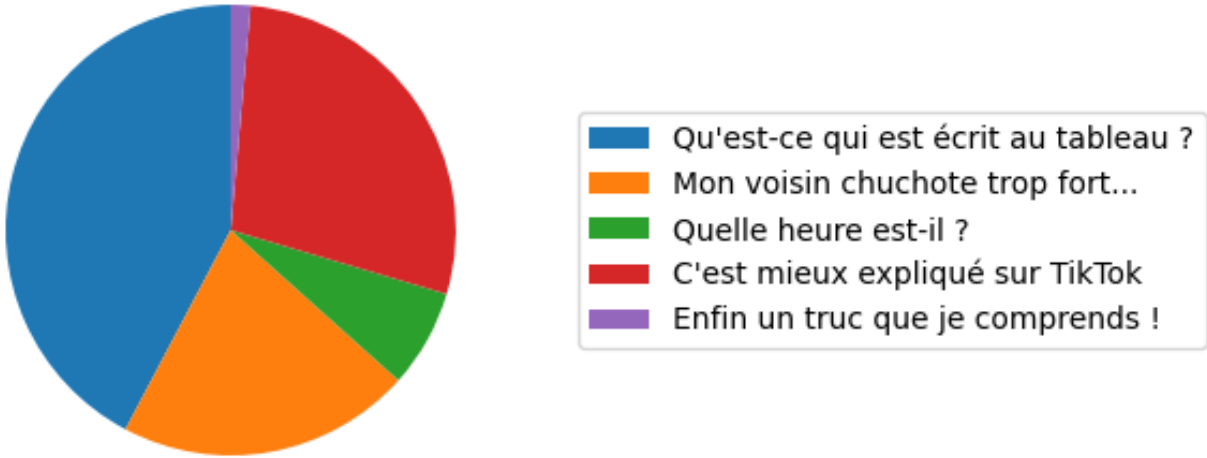
Ce que vous pensez en cours de maths :



- Surtout utile pour des variables qualitatives.

Diagramme sectoriel (communément appelé “camembert”)

Ce que vous pensez en cours de maths :



- Surtout utile pour des variables qualitatives.
- Permet de bien visualiser les proportions absolues.

Diagramme en bâtons

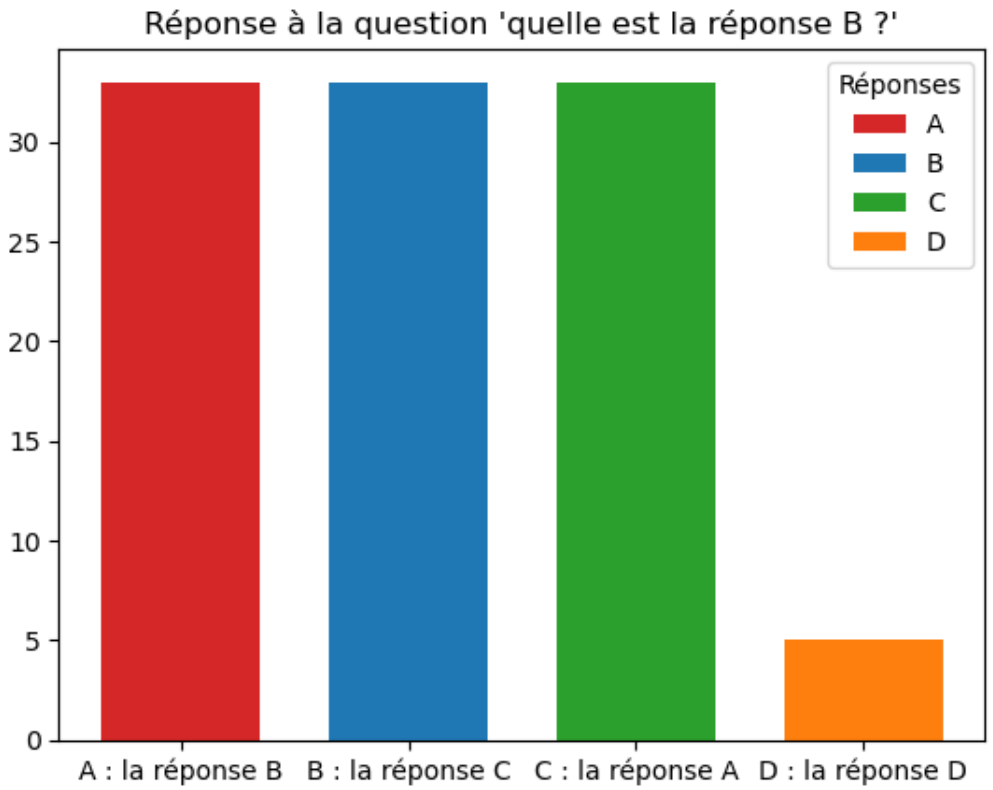
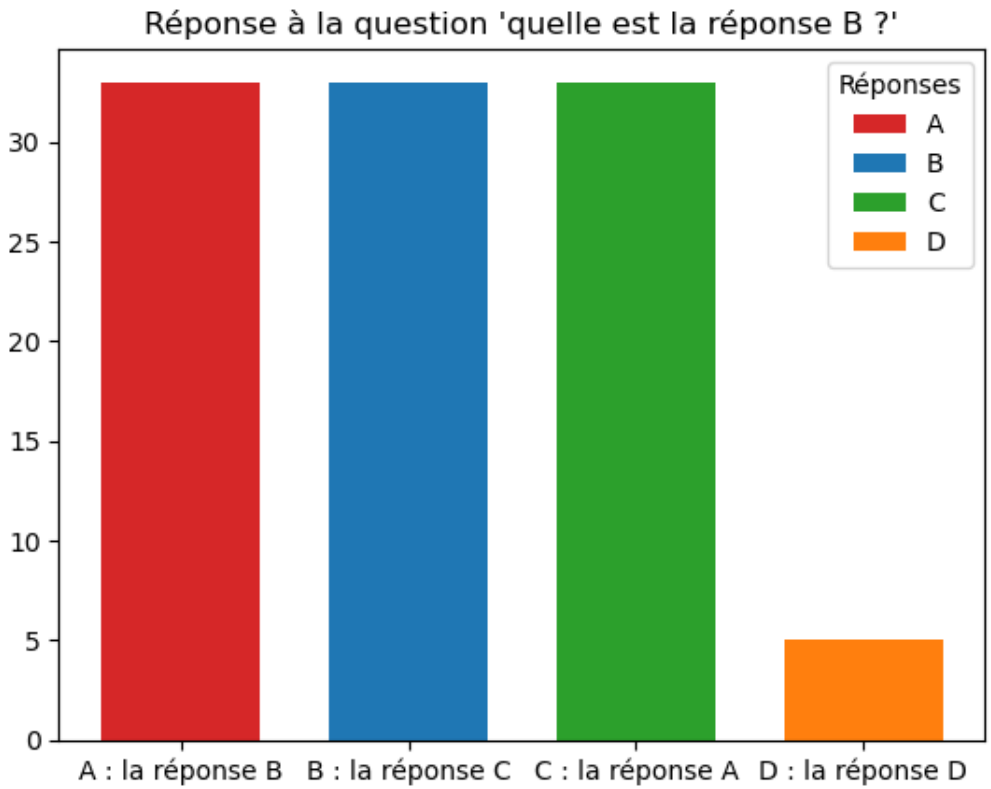
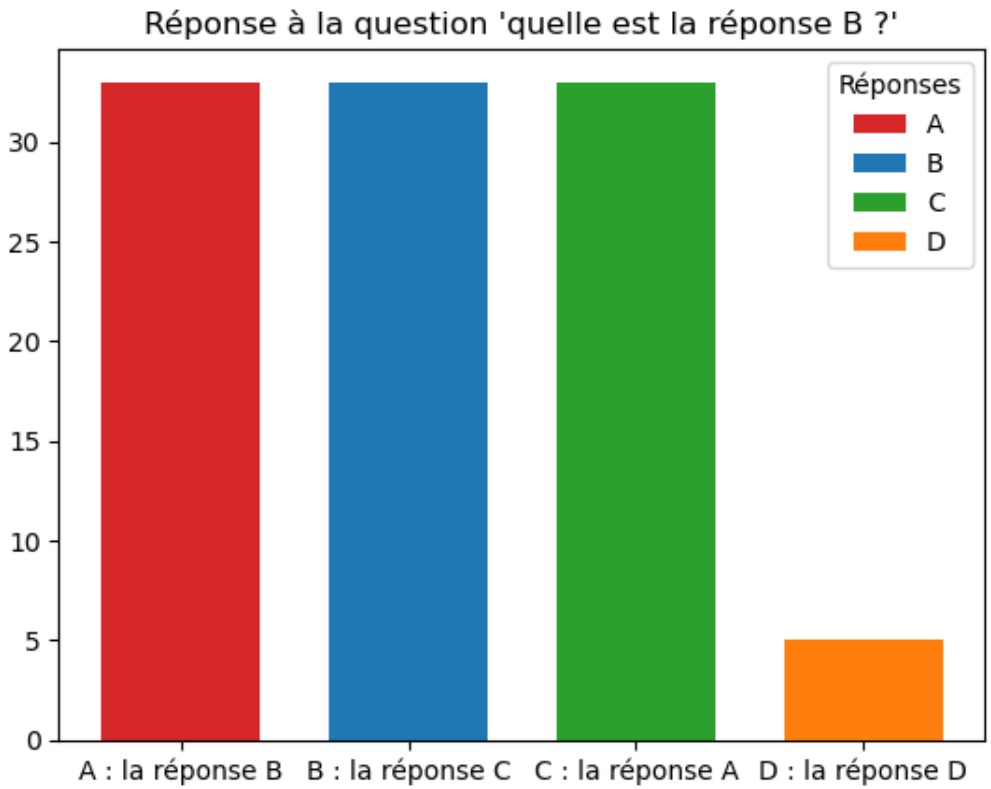


Diagramme en bâtons



- Fonctionne aussi pour des variables qualitatives.

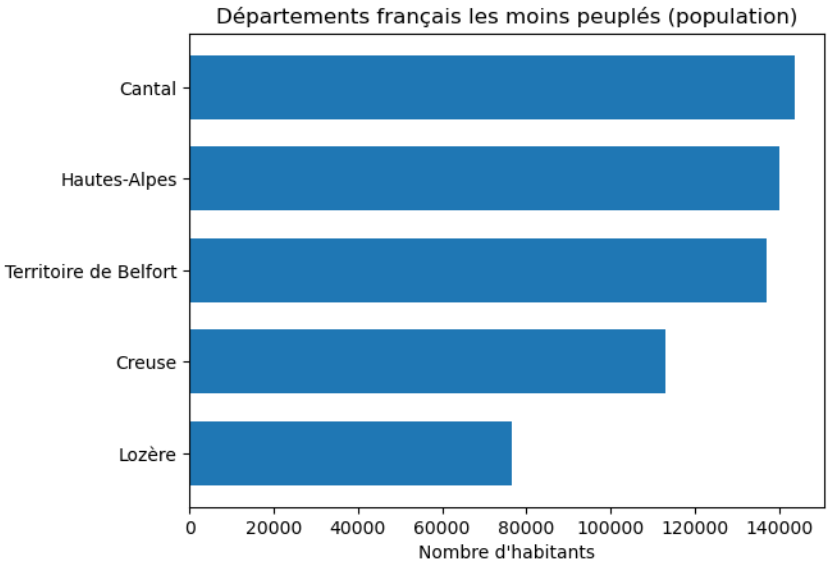
Diagramme en bâtons

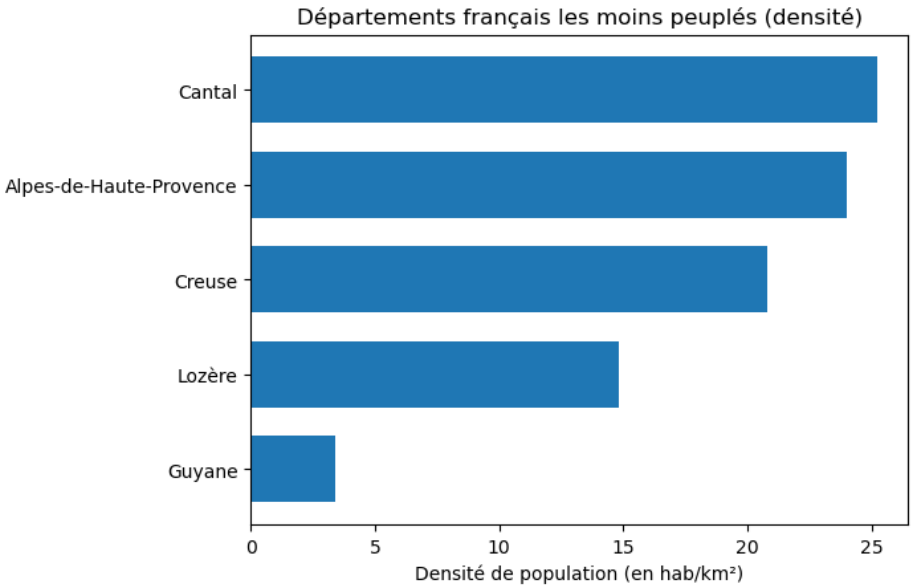


- Fonctionne aussi pour des variables qualitatives.
- Permet de bien visualiser les proportions relatives.

Diagramme en bâtons (2)

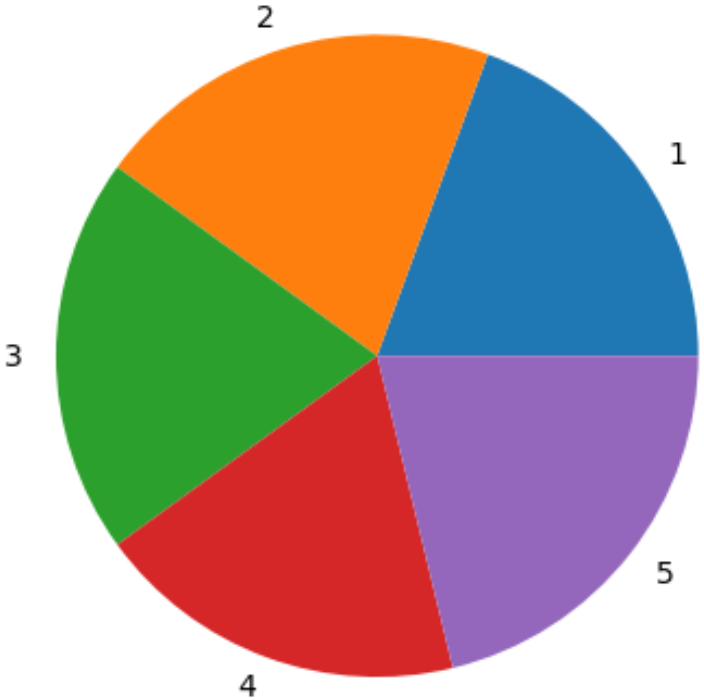
On peut aussi s'en servir pour des variables quantitatives :

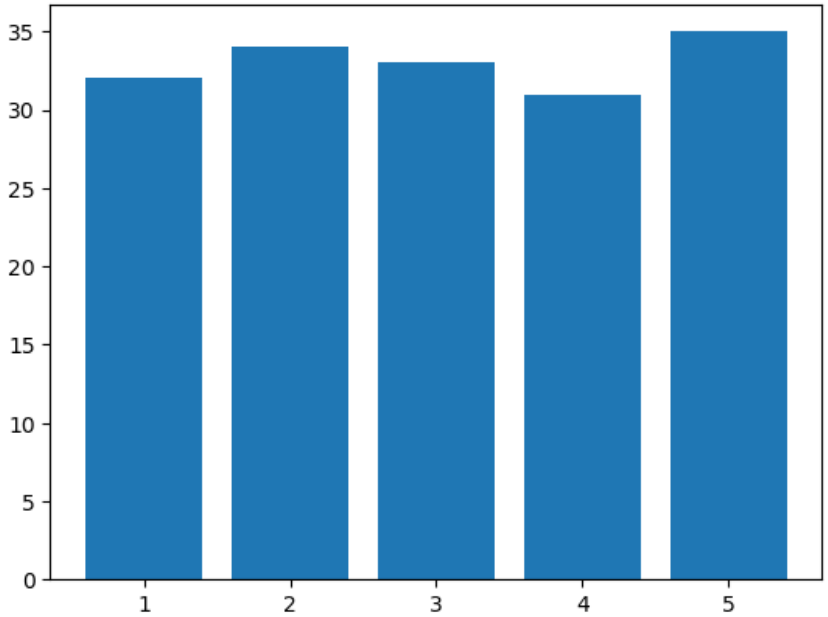




Comparaison des diagrammes

Dans certains cas les diagrammes en bâtons sont plus utiles :





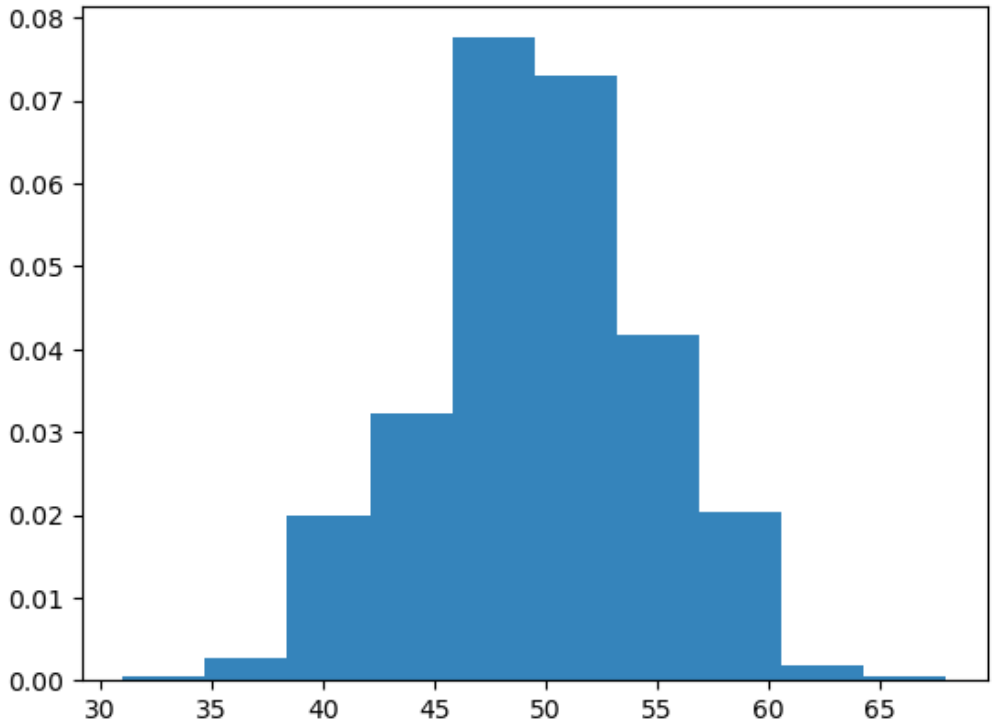
Histogrammes

L'histogramme généralise le diagramme à barres pour des variables quantitatives discrètes ou continues : on divise l'espace en intervalles de tailles égales, et on affiche le nombre ou la proportion d'éléments qui appartiennent à chaque intervalle.

Histogrammes

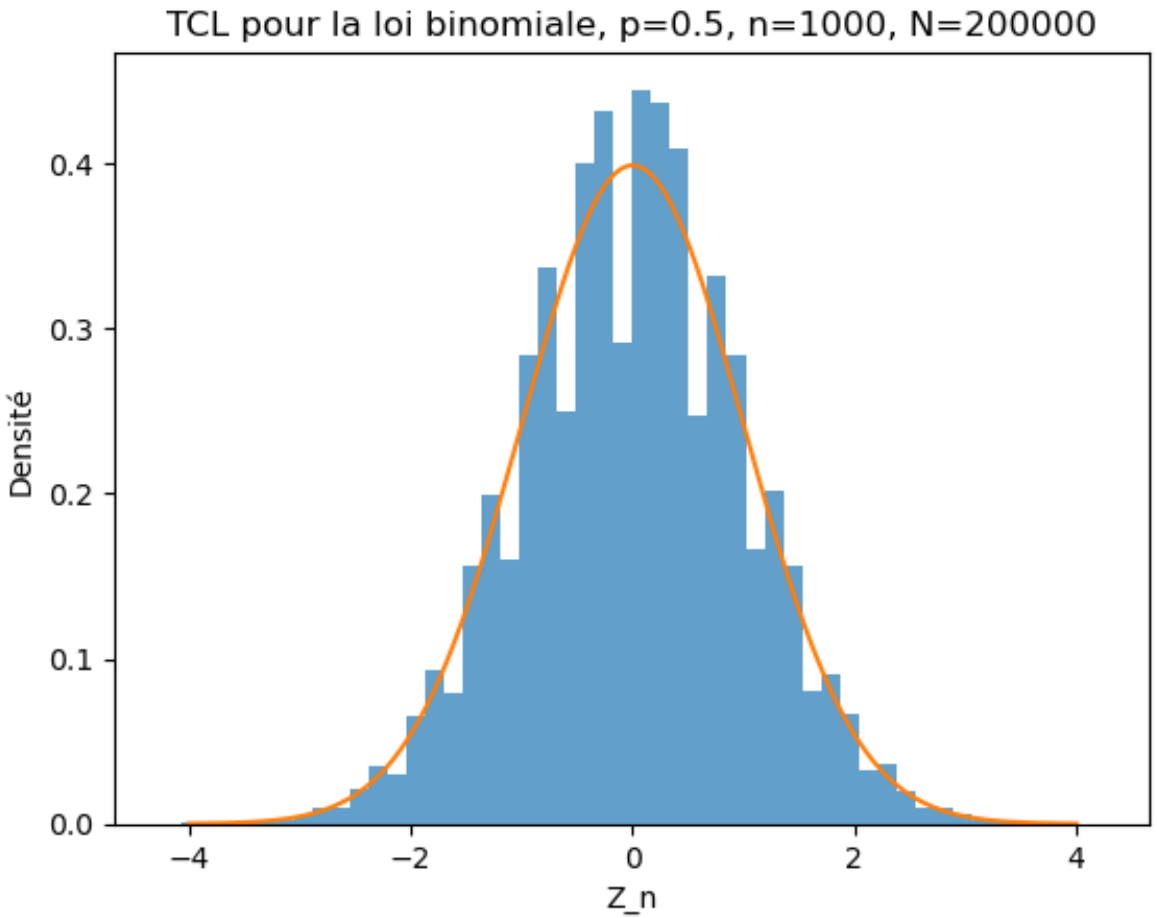
L'histogramme généralise le diagramme à barres pour des variables quantitatives discrètes ou continues : on divise l'espace en intervalles de tailles égales, et on affiche le nombre ou la proportion d'éléments qui appartiennent à chaque intervalle.

Exemple : On lance 100 fois une pièce équilibrée de manière indépendante, et on compte le nombre de fois où la pièce tombe sur face. On répète 1000 fois cette expérience, ce qui produit une famille $(k_1, k_2, \ldots, k_{1000})$ d'entiers naturels.

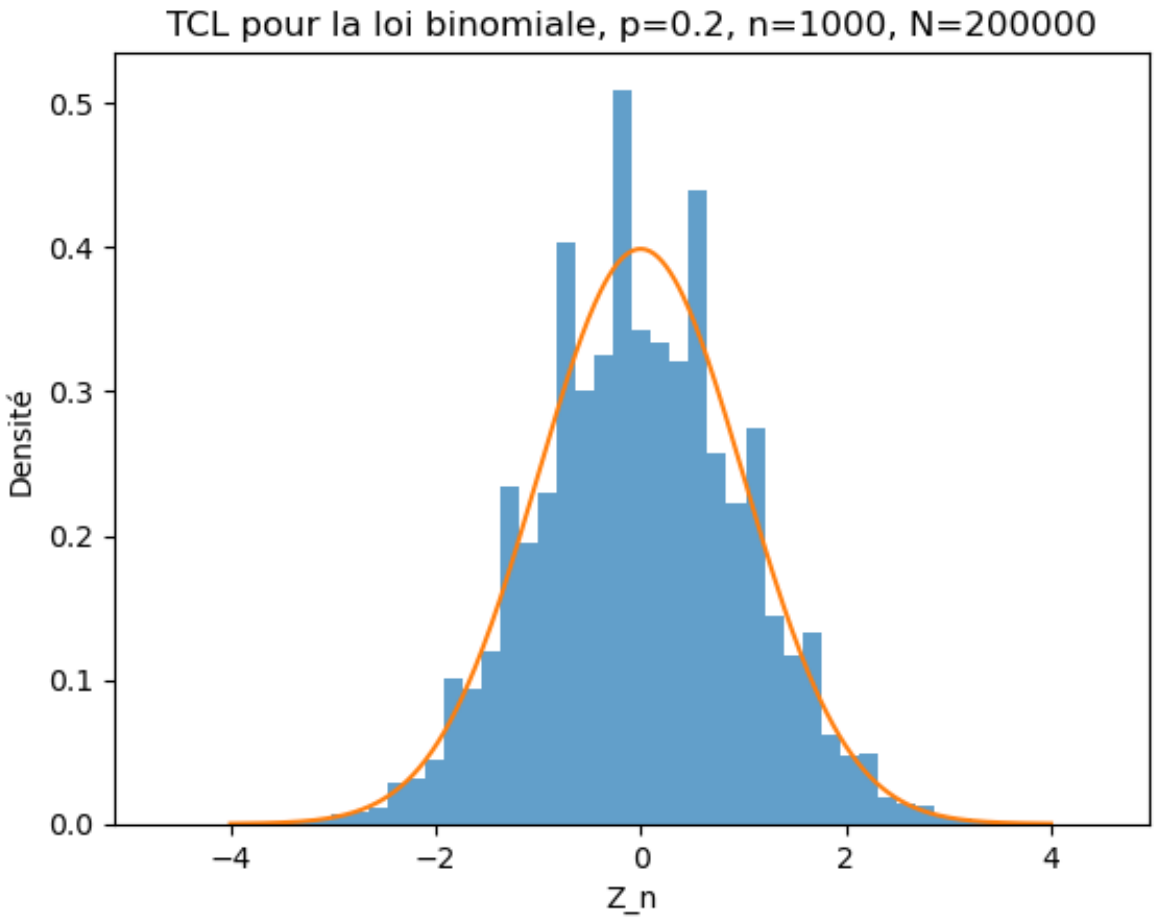


Retour vers le TCL

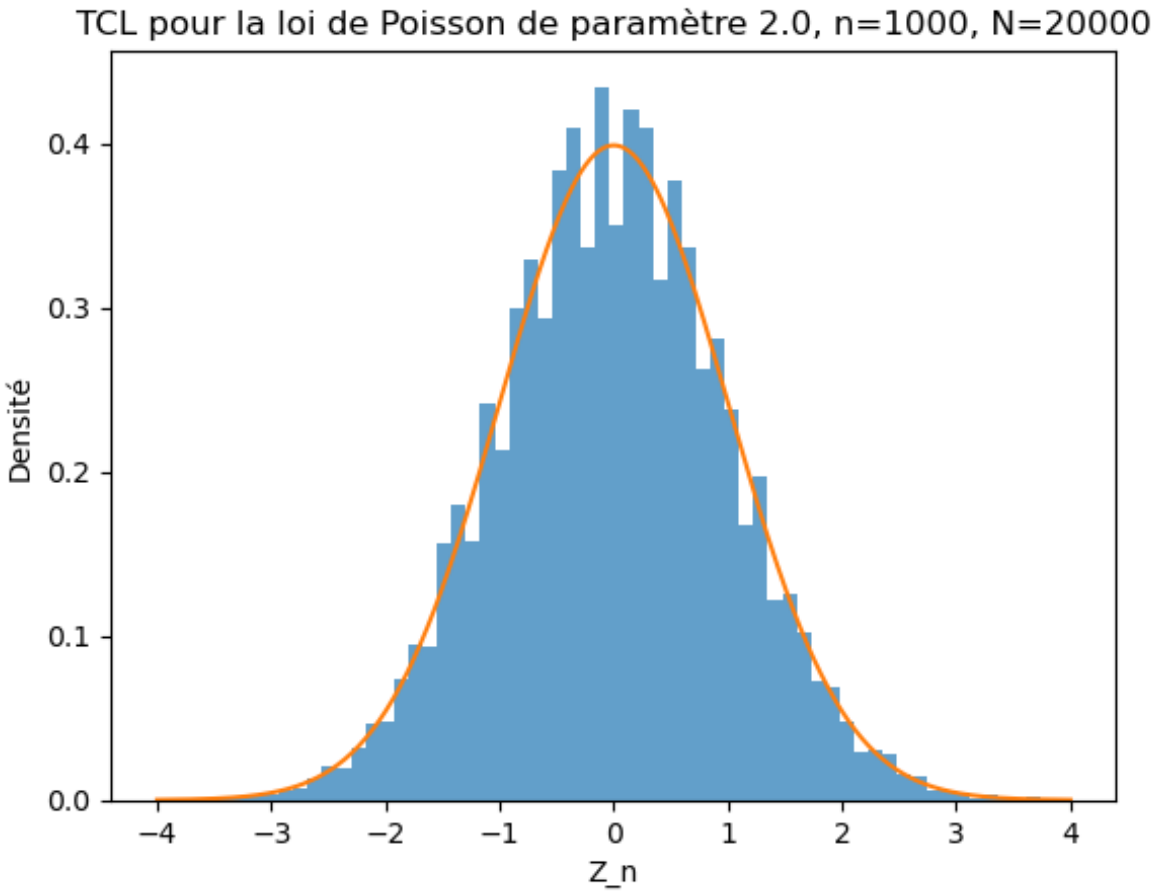
On reprend l'exemple précédent, pour un nombre n de pièces et un nombre N d'expériences. $(k_1, \ldots, k_N)$ est la réalisation d'une famille de variables aléatoires $(X_1, \ldots, X_N)$ indépendantes de loi $\mathrm{Binom}(n, 1/2)$, donc d'après le TCL si on les renormalise on doit retrouver asymptotiquement une loi normale standard.



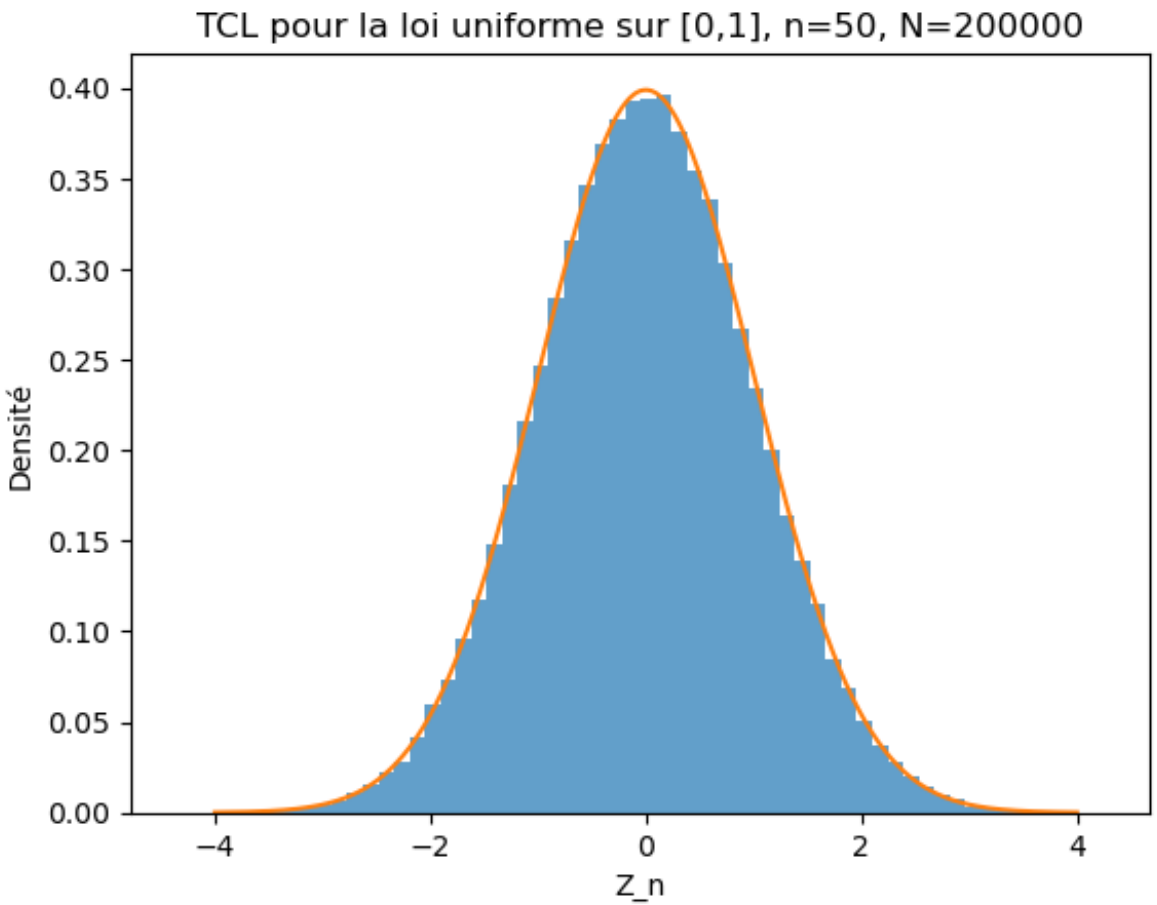
Le TCL est partout



Le TCL est partout



Le TCL est partout



Soit $x = (x_1, \ldots, x_n)$ un échantillon statistique dont les éléments sont à valeurs réelles.

1. Son étendue est définie par la différence entre ses valeurs extrêmes :

$$e(x) = \max(x_1, \ldots, x_n) - \min(x_1, \ldots, x_n).$$

Soit $x = (x_1, \ldots, x_n)$ un échantillon statistique dont les éléments sont à valeurs réelles.

1. Son étendue est définie par la différence entre ses valeurs extrêmes :

$$e(x) = \max(x_1, \ldots, x_n) - \min(x_1, \ldots, x_n).$$

2. Sa moyenne empirique est définie par

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Soit $x = (x_1, \ldots, x_n)$ un échantillon statistique dont les éléments sont à valeurs réelles.

1. Son étendue est définie par la différence entre ses valeurs extrêmes :

$$e(x) = \max(x_1, \ldots, x_n) - \min(x_1, \ldots, x_n).$$

2. Sa moyenne empirique est définie par

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

3. Sa variance empirique est définie par

$$\widehat{\sigma}_x^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 \right) - \overline{x}^2.$$

La racine carrée de la variance empirique est l'écart-type $\widehat{\sigma}_x$.

Soit $x = (x_1, \ldots, x_n)$ un échantillon statistique dont les éléments sont à valeurs réelles.

1. Son étendue est définie par la différence entre ses valeurs extrêmes :

$$e(x) = \max(x_1, \ldots, x_n) - \min(x_1, \ldots, x_n).$$

2. Sa moyenne empirique est définie par

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

3. Sa variance empirique est définie par

$$\widehat{\sigma}_x^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 \right) - \overline{x}^2.$$

   La racine carrée de la variance empirique est l'écart-type $\widehat{\sigma}_x$.

Remarque : Si x est représenté par une variable aléatoire $X : \Omega \to \mathbb{R}$ on constate que $\overline{x} = \mathbb{E}[X]$ et $\widehat{\sigma}_x^2 = \mathrm{Var}(X)$.

Exemple : les manchots

On se donne une population de 20 manchots Adélie et on souhaite analyser leur masse. On obtient l'échantillon suivant (les masses sont données en grammes)

$$\begin{aligned} x =&(3750, 3850, 3800, 3300, 3250, 3400, 3450, 3950, 3650, 4000,\\ &3600, 3500, 3900, 3725, 4100, 3625, 3550, 3475, 3700, 3825). \end{aligned}$$

On a

$$\overline{x} = \frac{1}{20}(3750 + 3850 + \ldots + 3700 + 3825) = 3670.$$

$$\widehat{\sigma}_x^2 = \frac{1}{20}\left((3750 - 3670)^2 + (3850 - 3670)^2 + \ldots + (3825 - 3670)^2\right) = 51600.$$

On obtient un écart-type de $\widehat{\sigma}_x = 227$.

Exemple : les manchots

Imaginons qu'on ajoute un manchot supplémentaire, dont la masse est de 6kg.
On a un nouvel échantillon, ou une nouvelle population :

$$\begin{aligned} x = &(3750, 3850, 3800, 3300, 3250, 3400, 3450, 3950, 3650, 4000, \\ &3600, 3500, 3900, 3725, 4100, 3625, 3550, 3475, 3700, 3825, 6000). \end{aligned}$$

On obtient de nouveaux estimateurs : $\overline{x} = 3780$, $\widehat{\sigma}_x^2 = 295531$ et $\widehat{\sigma}_x = 543$.

Exemple : les manchots

Imaginons qu'on ajoute un manchot supplémentaire, dont la masse est de 6kg. On a un nouvel échantillon, ou une nouvelle population :

$$\begin{aligned} x = &(3750, 3850, 3800, 3300, 3250, 3400, 3450, 3950, 3650, 4000, \\ &3600, 3500, 3900, 3725, 4100, 3625, 3550, 3475, 3700, 3825, 6000). \end{aligned}$$

On obtient de nouveaux estimateurs : $\overline{x} = 3780$, $\widehat{\sigma}_x^2 = 295531$ et $\widehat{\sigma}_x = 543$.

Si la masse du dernier manchot était de 4kg au lieu de 6, on aurait $\overline{x} = 3685$, $\widehat{\sigma}_x^2 = 54082$ et $\widehat{\sigma}_x = 233$.

Exemple : les manchots

Imaginons qu'on ajoute un manchot supplémentaire, dont la masse est de 6kg. On a un nouvel échantillon, ou une nouvelle population :

$$x = (3750, 3850, 3800, 3300, 3250, 3400, 3450, 3950, 3650, 4000,$$
$$3600, 3500, 3900, 3725, 4100, 3625, 3550, 3475, 3700, 3825, 6000).$$

On obtient de nouveaux estimateurs : $\overline{x} = 3780$, $\widehat{\sigma}_x^2 = 295531$ et $\widehat{\sigma}_x = 543$.

Si la masse du dernier manchot était de 4kg au lieu de 6, on aurait $\overline{x} = 3685$, $\widehat{\sigma}_x^2 = 54082$ et $\widehat{\sigma}_x = 233$.

- Plus la nouvelle valeur est éloignée de la moyenne empirique, plus la moyenne et la variance vont être altérées.

Quantiles

Soit $x = (x_1, \ldots, x_n)$ un échantillon. On le réordonne en un nouvel échantillon $x^* = (x_1^*, \ldots, x_n^*)$ avec $x_1^* \leqslant \ldots \leqslant x_n^*$. Le quantile empirique pour un paramètre $\alpha \in [0, 1]$ est défini par

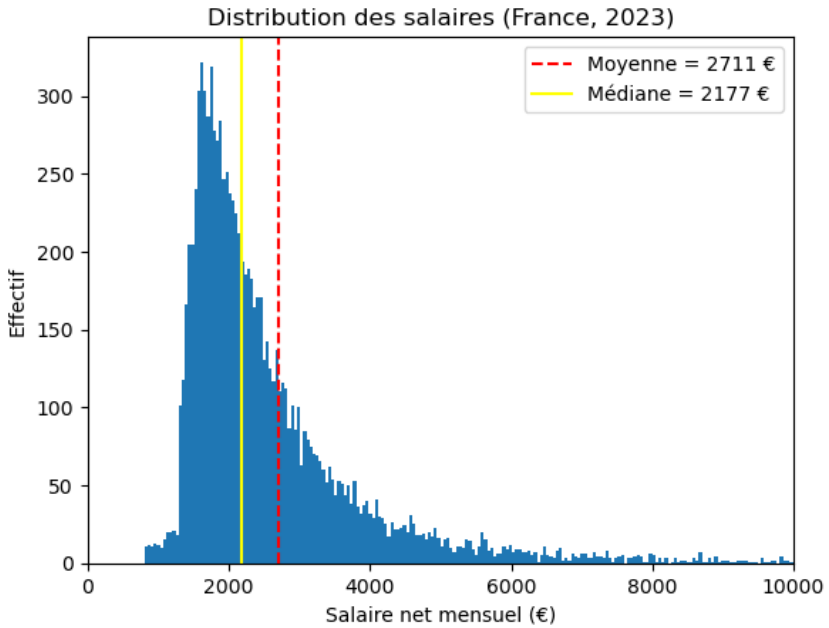
$$q_x(\alpha) = \begin{cases} \frac{x^*_{n\alpha} + x^*_{n\alpha+1}}{2} & \text{si } n\alpha \in \mathbb{N}, \\ x^*_{\lfloor n\alpha \rfloor + 1} & \text{sinon.} \end{cases}$$

Quantiles

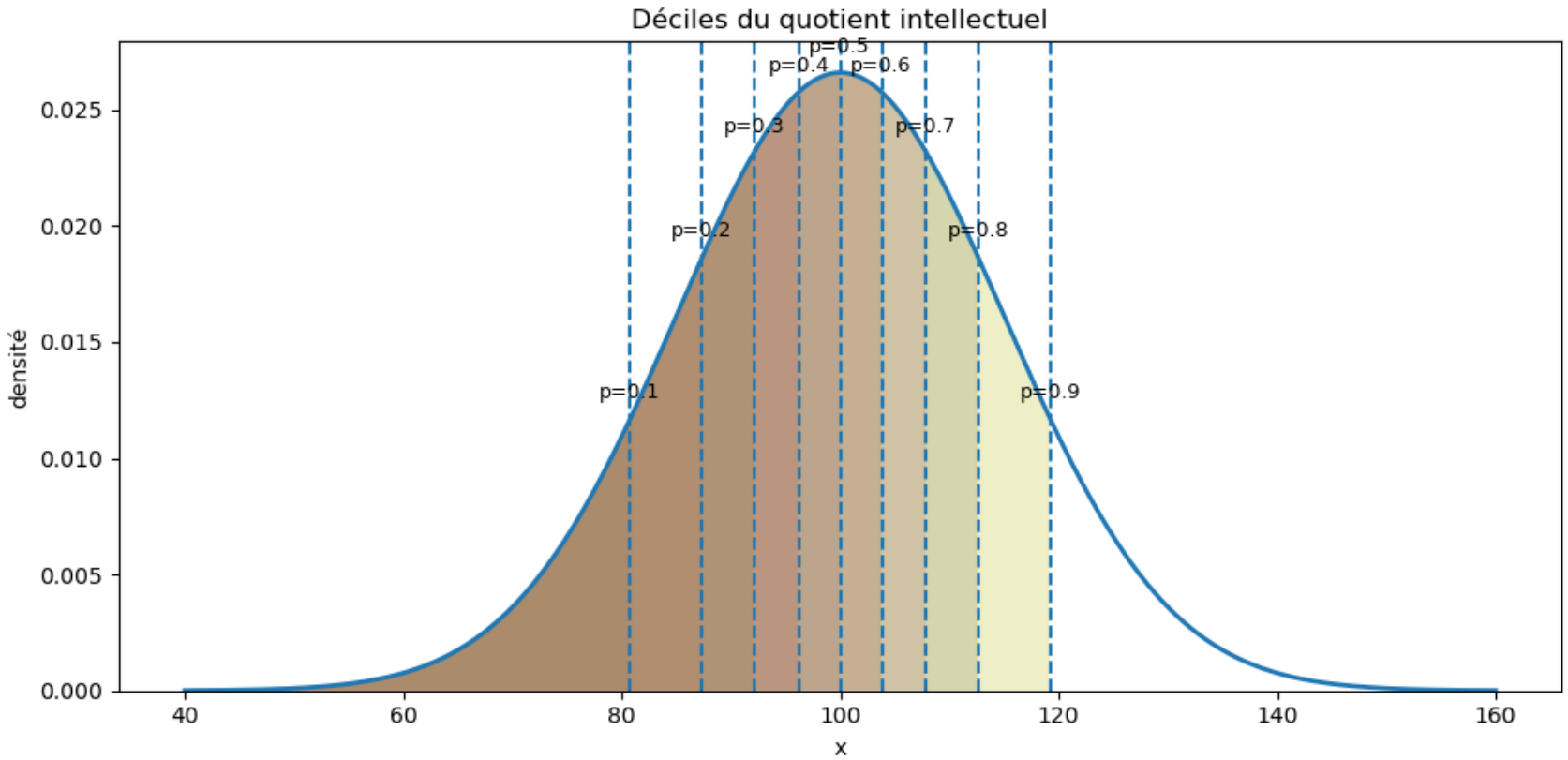
Soit $x = (x_1, \ldots, x_n)$ un échantillon. On le réordonne en un nouvel échantillon $x^* = (x_1^*, \ldots, x_n^*)$ avec $x_1^* \leqslant \ldots \leqslant x_n^*$. Le quantile empirique pour un paramètre $\alpha \in [0, 1]$ est défini par

$$q_x(\alpha) = \begin{cases} \frac{x^*_{n\alpha} + x^*_{n\alpha+1}}{2} & \text{si } n\alpha \in \mathbb{N}, \\ x^*_{\lfloor n\alpha \rfloor + 1} & \text{sinon.} \end{cases}$$

Exemple : si $\alpha = 1/2$, on obtient la médiane. Si $n/2 \in \mathbb{N}$, c'est-à-dire si n est pair, la médiane est la moyenne entre $x^*_{n/2}$ et $x^*_{n/2+1}$, et si n est impair, la médiane est la valeur du milieu, c'est-à-dire $x_{\lfloor n/2 \rfloor + 1}$.



On peut de la même manière définir les quartiles $q_{k/4}$ pour $k \in \{1, 2, 3\}$, et les déciles $q_{k/10}$ pour $k \in \{1, \ldots, 9\}$.



Les quantiles sont moins sensibles que la moyenne ou la variance aux valeurs extrêmes : reprenons l'échantillon de manchots

$$x = (3750, 3850, 3800, 3300, 3250, 3400, 3450, 3950, 3650, 4000, 3600, 3500, 3900, 3725, 4100, 3625, 3550, 3475, 3700, 3825).$$

L'échantillon ordonné est

$$x^* = (3250, 3300, 3400, 3450, 3475, 3500, 3550, 3600, 3625, 3650, 3700, 3725, 3750, 3800, 3825, 3850, 3900, 3950, 4000, 4100).$$

La médiane est de 3675.

Les quantiles sont moins sensibles que la moyenne ou la variance aux valeurs extrêmes : reprenons l'échantillon de manchots

$$x = (3750, 3850, 3800, 3300, 3250, 3400, 3450, 3950, 3650, 4000, 3600, 3500, 3900, 3725, 4100, 3625, 3550, 3475, 3700, 3825).$$

L'échantillon ordonné est

$$x^* = (3250, 3300, 3400, 3450, 3475, 3500, 3550, 3600, 3625, 3650, 3700, 3725, 3750, 3800, 3825, 3850, 3900, 3950, 4000, 4100).$$

La médiane est de 3675.

- Si on ajoute une valeur supérieure à 3700, la médiane devient 3700, peu importe si on ajoute une valeur de 4000 ou de 6000.

Les quantiles sont moins sensibles que la moyenne ou la variance aux valeurs extrêmes : reprenons l'échantillon de manchots

$$x = (3750, 3850, 3800, 3300, 3250, 3400, 3450, 3950, 3650, 4000, 3600, 3500, 3900, 3725, 4100, 3625, 3550, 3475, 3700, 3825).$$

L'échantillon ordonné est

$$x^* = (3250, 3300, 3400, 3450, 3475, 3500, 3550, 3600, 3625, 3650, 3700, 3725, 3750, 3800, 3825, 3850, 3900, 3950, 4000, 4100).$$

La médiane est de 3675.

- Si on ajoute une valeur supérieure à 3700, la médiane devient 3700, peu importe si on ajoute une valeur de 4000 ou de 6000.
- Si on ajoute une valeur supérieure à 3675 et une inférieure à 3675, la médiane reste la même, quelles que soient les valeurs en question.

Récapitulatif du cours

Un espace de probabilité est un triplet $(\Omega, \mathcal{F}, \mathbb{P})$, où :

- Ω est un ensemble quelconque.
- $\mathcal{F}$ est une tribu sur Ω, en particulier un sous-ensemble de $\mathcal{P}(\Omega)$.
- $\mathbb{P}$ est une fonction $\mathbb{P} : \mathcal{F} \to [0, 1]$ qui vérifie certaines propriétés relativement aux opérations sur les sous-ensembles de Ω.

À connaître :

- Les propriétés de $\mathbb{P}$ vis-à-vis des opérations ensemblistes $({}^c, \cup, \cap)$
- Les principales notions du dénombrements : arrangements, combinaisons, permutations.
- La notion d'indépendance de 2 événements
- La notion d'indépendance de 3 événements ou plus
- La définition de la probabilité conditionnelle et le théorème de Bayes.

Une variable aléatoire est une fonction $X : \Omega \to E$, où $(\Omega, \mathcal{F}, \mathbb{P})$ est un espace de probabilité abstrait. Elle est discrète si E est fini ou dénombrable. Dans ce cas, sa loi est donnée par $\mathbb{P}(X = x)$ pour $x \in E$.

À connaître :

- Formule de transfert
- Définition de l'espérance et des moments
- Définition et calcul de la variance
- Connaître la définition, l'espérance et la variance des lois usuelles (Dirac, Bernoulli, binomiale, Poisson, uniforme discrète)
- Définition et utilisation de la fonction génératrice
- Loi, indépendance et conditionnement de plusieurs variables aléatoires

Il existe plusieurs notions différentes de convergence pour des variables aléatoires.

À connaître :

- Définition de la convergence en probabilité
- Définition de la convergence dans L^p
- Définition de la convergence en loi
- Liens entre les convergences
- Caractérisation fonctionnelle des convergences
- Loi des grands nombres
- Théorème central limite

Une variable aléatoire $X : \{1, \ldots, n\} \to E$ représente une caractéristique d'une population $\Omega = \{1, \ldots, n\}$ constituée d'individus. Elle définit un échantillon $x = (x_1, \ldots, x_n)$ avec $x_i = X(i)$.

À connaître :

- Différence entre variables qualitatives/quantitatives
- Les différents types de représentations de variables
- Les estimateurs : étendue, moyenne/variance empirique, quantiles

Bonus : Autour de l'espérance

Peut-on définir l'espérance de variables aléatoires à valeurs dans d'autres espaces que $\mathbb{R}$?

Peut-on définir l'espérance de variables aléatoires à valeurs dans d'autres espaces que $\mathbb{R}$?

- Si $(X_1, X_2) \in \mathbb{R}^2$, on peut définir

$$\mathbb{E}[(X_1, X_2)] = \sum_{\omega \in \Omega} (X_1(\omega), X_2(\omega))\mathbb{P}(\omega) = (\mathbb{E}[X_1], \mathbb{E}[X_2]),$$

et cela fonctionne de la même manière. On peut généraliser cela à n'importe quel $\mathbb{R}$-espace vectoriel.

Peut-on définir l'espérance de variables aléatoires à valeurs dans d'autres espaces que $\mathbb{R}$?

- Si $(X_1, X_2) \in \mathbb{R}^2$, on peut définir

$$\mathbb{E}[(X_1, X_2)] = \sum_{\omega \in \Omega} (X_1(\omega), X_2(\omega))\mathbb{P}(\omega) = (\mathbb{E}[X_1], \mathbb{E}[X_2]),$$

et cela fonctionne de la même manière. On peut généraliser cela à n'importe quel $\mathbb{R}$-espace vectoriel.Mais cela n'aura pas vraiment d'utilité dans ce cours !

Peut-on définir l'espérance de variables aléatoires à valeurs dans d'autres espaces que $\mathbb{R}$?

- Si $(X_1, X_2) \in \mathbb{R}^2$, on peut définir

$$\mathbb{E}[(X_1, X_2)] = \sum_{\omega \in \Omega} (X_1(\omega), X_2(\omega))\mathbb{P}(\omega) = (\mathbb{E}[X_1], \mathbb{E}[X_2]),$$

  et cela fonctionne de la même manière. On peut généraliser cela à n'importe quel $\mathbb{R}$-espace vectoriel.Mais cela n'aura pas vraiment d'utilité dans ce cours !

- Si X appartient à un ensemble qui n'est pas un sous-ensemble d'un espace vectoriel, il va y avoir un problème. Prenons par exemple des variables qualitatives : une pièce a une chance sur deux de tomber sur pile et une chance sur deux de tomber sur face. Si on se donne $X : \Omega \to \{\text{pile}, \text{face}\}$, alors on devrait avoir

$$\mathbb{E}[X] = \frac{1}{2}\text{pile} + \frac{1}{2}\text{face},$$

  ce qui n'a de sens que si on construit artificiellement un espace vectoriel engendré par les objets « pile » et « face ».