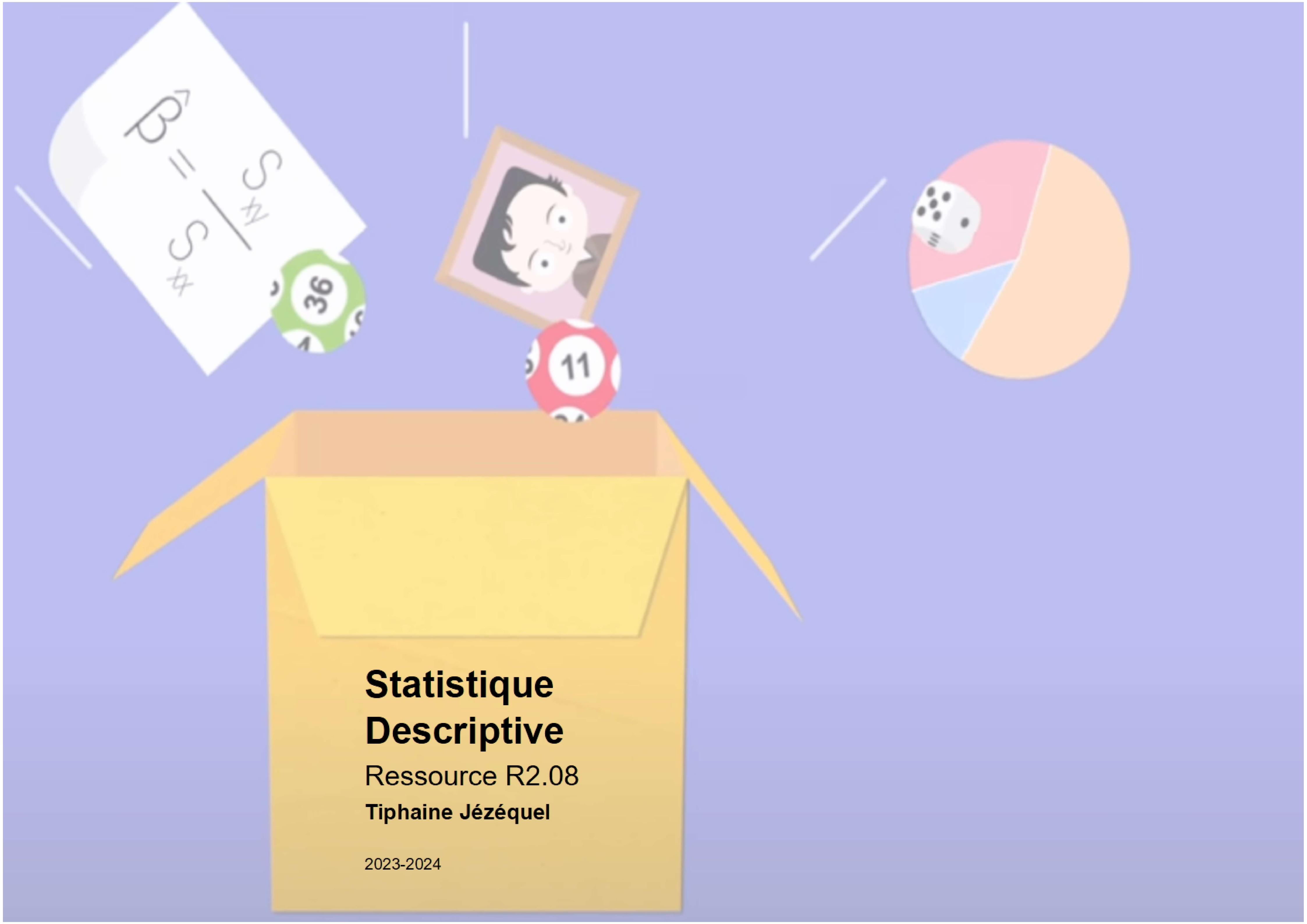


Statistique Descriptive

Ressource R2.08

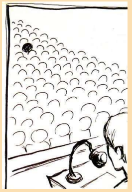
Tiphaine Jézéquel

2023-2024

Organisation de la Ressource

Organisation de la ressource

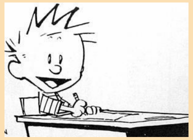
Statistique = vos seules maths jusqu'à la fin de l'année.



- Pendant 4 semaines, chaque semaine :
  - 1h amphi - T. Jézéquel
  - 2h de TD dont 1h sur PC (Python)
    - T. Jézéquel (A,B,C,D), Yannick Favreau (E)

$\hookrightarrow$ A la suite, pendant 4 semaines : encore 2h de TD de statistiques sur Python par semaine pour la **SAé 2.04**.

Évaluation



- QCM Moodle sur les CM à faire chez vous après chaque CM - coeff 1/4
- DS de 1h15 **le 17/05** avec des exercices en Python - coeff 1

Objectifs de la Ressource: outils pour visualiser et analyser des données

Vers la SAé 2.04, lien avec BDD

Dans la SAé 2.04 :

- implémentation d'une BDD
- analyse de ces données avec des outils de statistiques

SAé 2.04 en 2023 : analyse des données de Parcoursup 2022.

- qu'est-ce qui influence le fait qu'il y aie un + grand pourcentage de filles admises dans certaines formations ?

SAé 2.04 en 2022 : analyse des données (notes, villes/bacs d'origine) issues des étudiant.e.s de DUT Info de Lannion de 2017 à 2021

- la longueur du prénom influence-t-elle les notes des étudiant.e.s ?

- **Exemples de jeux de données**

Exemple 1: Nombre d'absences injustifiées par étudiant.e :
$2, 4, 2, 1, 0, 0, 3, 1, 3, 0, 2, 3, 0, 3, 1, 0, 2, 0, 2, 1, 0, 3, 1, 2, 2, 3, 1, 3, 2, 2, 0, 2, 0, 3, 2, 1, 2, 2, 3, 2, 0, 2, 0, 2, 0, 2, 1, 2, 0, 0, 3, 1, 0, 2, 0, 2, 1, 2, 2, 2, 1, 2, 0, 2, 1, 0, 0, 4, 1, 0, 3, 0, 2, 0, 0, 3, 1, 3, 0, 0, 1, 2, 1, 2, 6, 2, 2, 1, 1, 2, 3, 1, 0, 1, 1, 1, 3, 1, 3, 2$

Exemple 2: Taille (exprimée en cm) de chaque étudiant·e :
$188, 169, 181, 180, 159, 180, 164, 184, 177, 159, 187, 174, 170, 173, 175, 150, 171, 166, 173, 182, 167, 173, 174, 175, 166, 173, 180, 188, 165, 173, 178, 182, 175, 186, 162, 193, 173, 184, 184, 184, 182, 187, 193, 161, 193, 169, 163, 179, 179, 184, 182, 182, 172, 175, 193, 170, 176, 166, 177, 163, 191, 171, 189, 183, 178, 197, 166, 187, 180, 172, 181, 167, 177, 177, 186, 174, 168, 182, 183, 182, 170, 186, 193, 167, 184, 159, 169, 192, 172, 167, 178, 176, 177, 175, 172, 175, 182, 176, 179, 188$

$\hookrightarrow$ C'est ce qu'on appelle des *données brutes*.

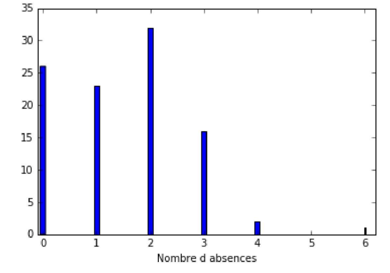
Remarque: Pour des raisons de confidentialité on ne peut pas en général identifier chaque individu de la population.

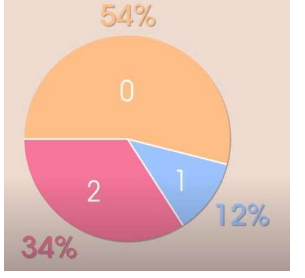
Plan de la Ressource

Cours 1. **Vocabulaire, caractéristiques d'une série de données**

- moyenne, variance, etc.
- ↪ pour résumer une répartition de données ou prévoir des données futures

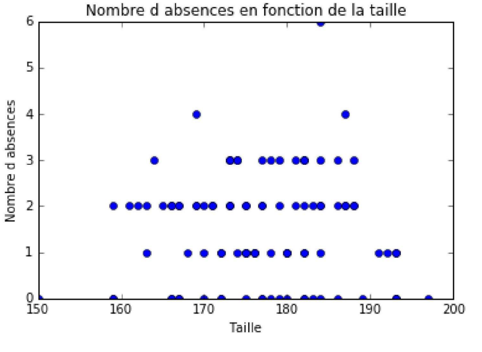
Cours 2. **Représenter la répartition d'une série de données**

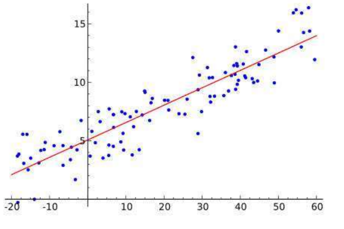




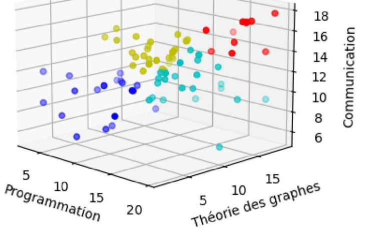
Cours 3. **Corrélations entre 2 séries de données**

- ↪ pour interprétation des données.





Cours 4. **Présentation SAé + approche cas de plus de 2 séries de données**



Cours 1

Vocabulaire, caractéristiques d'une variable statistique

Ressource R2.08 - Statistique Descriptive

2023-2024

Plan

1. La Statistique, c'est quoi ?
2. Variable statistique, Série statistique
3. Moyenne
4. Pourquoi autant de notions ?

5. Variance et écart-type
6. Médiane et quartiles
7. Fonction de répartition

1. La Statistique, c'est quoi ?

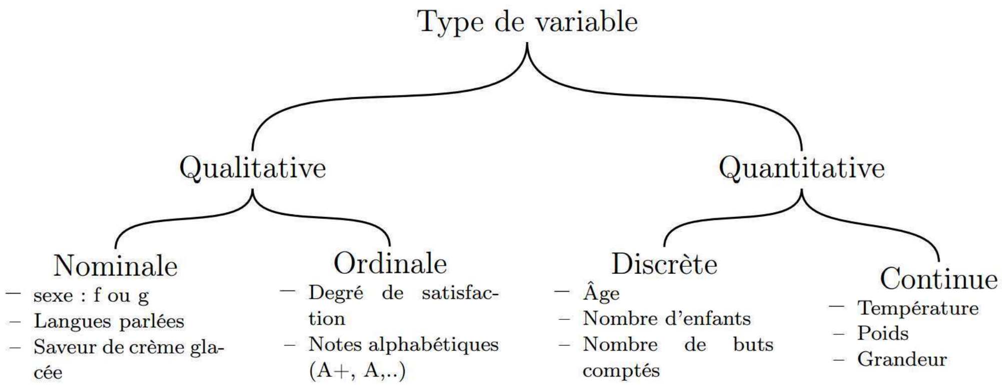
• Définition

La **statistique** est la discipline qui étudie des phénomènes à travers la collecte de données :

- traitement et analyse des données
- interprétation des résultats
- présentation afin de rendre ces données compréhensibles par tous.

La statistique est d'abord un domaine des mathématiques mais aujourd'hui c'est aussi un domaine de la science des données (Data Science).

• Types de données statistiques



Dans cette ressource, on ne travaillera qu'avec des **variables quantitatives**.

• Statistique et Probabilités, quelles différences ?

Ces 2 domaines des mathématiques ont en commun d'être des **sciences de l'aléatoire**. Mais elles n'abordent pas les problèmes de la même manière :

En **probabilités**, on s'intéresse aux **résultats théoriques d'une expérience aléatoire**. On la modélise par une loi qu'on estime correspondre au monde réel.

On peut parler d'une "connaissance a priori" puisque l'étude se fait sans données.

Exemple : expérience aléatoire du lancer de dé, version probabilités

On aura une probabilité 1/6 de tomber sur chacune des valeurs 1,2,3,4,5 ou 6. C'est-à-dire que les résultats suivront une loi uniforme.

On n'a pas besoin de lancer le dé pour le dire !

Et à partir de cette loi théorique, on peut calculer l'espérance (moyenne théorique), la variance etc.
Pour un dé, l'espérance sera par exemple de 3.5.

En **statistique**, on étudie des données issues d'**une expérience aléatoire concrète**. On cherche leurs propriétés, et on peut émettre des hypothèses sur des données futures.

On peut parler d'une "connaissance a posteriori" puisque l'étude se fait après collecte de données.

Exemple 3. expérience aléatoire du lancer de dé, version statistique

On commence par réaliser des lancers de dé :

. . . .

On obtient une série statistique.

Et à partir de cette série statistique, on peut calculer la moyenne, la variance etc.
Pour nos lancers de dé, la moyenne sera par exemple de . . . .

2. Variable statistique, Série statistique

• Variable statistique, population, effectif total

Définitions

Soit P un ensemble, appelé **population**.
Une **variable statistique** est une application :

$$\begin{aligned} X : P &\rightarrow \mathbb{R} \\ i &\rightarrow X(i) \end{aligned}$$

Le nombre $N = Card(P)$ est appelé **l'effectif total**.

Exemple 1: Absences des étudiants

La *population* P est ....
La *variable statistique* $X : P \to \mathbb{R}$, à l'étudiant i associe ....
....
Si les 2 premiers étudiant.es de la liste s'appellent Albert et Maria, alors :

$$\begin{aligned} X(Albert) &= \ldots\ldots, \\ X(\ldots\ldots) &= 4. \end{aligned}$$

L'*effectif total* est de ....

Exercice 2 : Tailles des étudiants

La *population* P est ....
La *variable statistique* $Y : P \to \mathbb{R}$ à l'étudiant i associe ....
$Y(Albert) = \ldots\ldots$
$Y(Maria) = \ldots\ldots$
L'*effectif total* est de ....

Exercice 3 : Lancers de dés

La *population* P est ....
La *variable statistique* $Z : P \to \mathbb{R}$, au lancer numéro i associe ....
$Z(2) = \ldots\ldots$
L'*effectif total* est de ....

• Série statistique

Définition

Soit $X : P \to \mathbb{R}$ une variable statistique.
La liste des $X(i)$ pour $i \in P$ est appelée **série statistique** associée à la variable X.

Exemples: Absences et tailles des étudiants, Lancers de dé

Les *séries statistiques* sont les listes de données brutes du début du cours, et la liste des tirages de dés.

• Tableau statistique

Définition

Un **tableau statistique** est un tableau regroupant les effectifs et/ou fréquences d'une variable statistique pour chaque valeur prise par cette variable.

Exemple 1 :

Valeurs de X	0	1	2	3	4	6
effectifs	26	23	32	16	2	1
fréquences	0.26	0.23	0.32	0.16	0.02	0.01

Exercice : Remplir le tableau pour les tirages de dé :

Valeurs de Z	1	2	3	4	5	6
effectifs	...	...	...	...	...	...
fréquences	...	...	...	...	...	...

3. Moyenne

• Définition

Définition

Soit X une variable statistique, P sa population et N l'effectif total.
La **moyenne** de X, notée $\overline{X}$, est la valeur :

$$\overline{X} = \frac{1}{N} \sum_{k \in P} X(k)$$

Exemple 1 : Absences des étudiants

La population est . . . . , l'effectif total est . . . .

$\overline{X} =$. . . . .

On peut utiliser le **tableau statistique** quand on l'a, pour aller plus vite :

$\overline{X} =$. . . . .

• Moyenne d'une somme

Proposition

Soit X, Y deux séries statistiques sur une même population et $a, b \in$ alors

$$\overline{aX + bY} = a\overline{X} + b\overline{Y}$$

Et en particulier : $\overline{aX + b} = a\overline{X} + b.$

Exemple :

Si les absences données par X sont celles du 1e semestre, et qu'on a une autre variable statistique X_2 avec les absences au 2e semestre. Le total des absences sur l'année d'un étudiant i est $X(i) + X_2(i)$.
Si on sait que la moyenne des absences au 2e semestre est $\overline{X_2} = 1.01$, alors on peut calculer la moyenne des absences des étudiant.es sur l'année :

$$\overline{X + X_2} = \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

Exercice :

Sachant que la taille moyenne des étudiants est de 176.71cm, si on enlève 2cm aux étudiants à chaque absence, quelle sera la moyenne de leurs tailles à la fin du semestre ?

. . . .

4. Pourquoi autant de notions ?

La moyenne est souvent trompeuse : d'où l'intérêt d'utiliser d'autres indicateurs.

Chaine Youtube : **La statistique expliquée à mon chat**
Vidéo : Pourquoi gagnez-vous moins que le salaire moyen ?

www.youtube.com/watch?v=uIx2xvdwIIo

5. Variance et écart-type

Définition

Soit X une variable statistique de moyenne $\overline{X}$, alors on appelle **variance** de X la valeur

$$Var(X) = \frac{1}{N}\sum_{k\in P}(X(k)-\overline{X})^2 = \overline{(X-\overline{X})^2}$$

et l'**écart-type** de X est

$$\sigma_X = \sqrt{Var(X)}$$

Remarque : L'écart-type σ_X donne une estimation de la largeur de l'intervalle $[M-\sigma_X, M+\sigma_X]$ où sont concentrées 2/3 des données.

Exemple 1. On a $\overline{X} = 1.49$, et donc :

$Var(X) = \dots\dots\dots$

Exercice 2. Sachant que $\overline{Y} = 176.71$, faire la même chose avec Y :

$Var(Y) = \dots\dots\dots$

Proposition : Formule de Koenig

Soit X une variable statistique de moyenne $\overline{X}$. On peut calculer la variance avec la formule de Koenig :

$$Var(X) = \overline{X^2} - \overline{X}^2 = \left(\frac{1}{N}\sum_{k\in P}X(k)^2\right) - \overline{X}^2$$

Exemple 1 :

$Var(X) = \dots\dots\dots$

Exercice 2 : Faire la même chose avec Y

$Var(Y) = \dots\dots\dots$

Proposition : Var de aX+b

Soit X une variable statistique et $a, b \in \mathbb{R}$. Alors

$$Var(aX+b) = a^2 Var(X)$$

6. Médiane et quartiles

Définition

La **médiane** m est la valeur qui coupe la population en 2 parts égales

- la part de la population pour laquelle $X(i) \leq m$,
- et la part de la population pour laquelle $X(i) \geq m$.

Si l'effectif total $N = 2p+1$ est impair, alors m est la $(p+1)$-eme valeur.
Si l'effectif total $N = 2p$ est pair, alors m est la moyenne des p-eme et $(p+1)$-eme valeurs.

Exemples :

- dans la série statistique $1, 1, 2, 3, 5, \mathbf{7}, 7, 9, 10, 11, 17$, l'effectif total est $11 = 2\times 5+1$ (impair) donc la médiane est la 6e valeur : 7.
- pour $1, 1, 2, 3, \mathbf{5}, \mathbf{7}, 7, 9, 10, 11$, l'effectif total est $10 = 2\times 5$ (pair) donc la médiane est la moyenne des 5e et 6e valeurs : **6**.

Définition

Les **quartiles** q_1, q_2 et q_3 sont les valeurs qui coupent la population en 4 parts égales

- la part de la population pour laquelle $X(i) \leq q_1$,
- la part de la population pour laquelle $q_1 \leq X(i) \leq q_2$.
- la part de la population pour laquelle $q_2 \leq X(i) \leq q_3$.
- la part de la population pour laquelle $q_3 \leq X(i)$.

De même que pour la médiane, selon les cas où l'effectif N est divisible ou non par 4, les quartiles peuvent être une valeur prise par X, ou la moyenne de 2 valeurs prises par X.

Exemple :

- dans la série statistique $1, 1, 2, 3, \mathbf{5}, \mathbf{7}, 7, 9, 10, 11$, les quartiles sont :

$$q_1 = \dots\dots \qquad q_2 = \dots\dots \qquad q_3 = \dots\dots$$

Remarque : le 2e quartile q_2 est égal à la médiane m.

7. Fonction de répartition

• Définition

Définition

Soit X une variable statistique.
La **fonction de répartition** F_X de X est la fonction

$$F : \quad \longrightarrow [0,1]$$
$$x \longmapsto \sum_{\{i \mid x_i \leq x\}} f_i$$

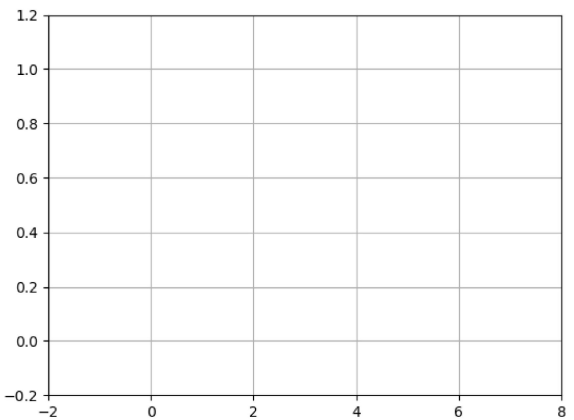
où f_i désigne la fréquence à laquelle apparait la valeur x_i dans la série statistique.

Exemple 1 :

Valeurs de X	0	1	2	3	4	6
Fréquences	0.26	0.23	0.32	0.16	0.02	0.01
Cumul						

$F(2) = \ldots\ldots$, $F(3) = \ldots\ldots$, $F(6) = \ldots\ldots$, $F(4.2) = \ldots\ldots$

I	$]-\infty, 0[$	$[0,1[$	$[1,2[$	$[2,3[$	$[3,4[$	$[4,6[$	$[6,+\infty[$
$F_X(t)$	0	0.26					



• Fonction de répartition pour la médiane et les quartiles

La fonction de répartition peut servir à calculer la médiane et les quartiles :

Fonction de répartition pour la médiane

Soit X une variable statistique, et F_X sa fonction de répartition.
La médiane m de X est le nombre tel que $F_X(m) = 0.5$.

- s'il n'y a pas de x tel que $F_X(x) = 0.5$, m est la première valeur telle que $F_X(x) > 0.5$,
- s'il y a plusieurs valeurs x telles que $F_X(x) = 0.5$, m est la moyenne de ces valeurs.

Fonction de répartition pour les quartiles

Les *quartiles* q_1 et q_3 de X sont les nombres tels que

$$F_X(q_1) = 0.25 \text{ et } F_X(q_3) = 0.75.$$

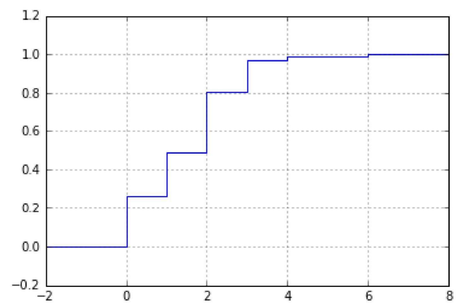
S'il n'y a pas de telles solutions q_i ou s'il y en a plusieurs, on procède comme pour la médiane (avec 0.25 et 0.75 au lieu de 0.5).

Exemple 1 :

$m =$

$q_1 =$

$q_3 =$



Valeurs de X	0	1	2	3	4	6
effectifs	26	23	32	16	2	1
fréquences	0.26	0.23	0.32	0.16	0.02	0.01