Chapitre 1: Statistiques descriptives univariées

Arnaud Rousselle
arnaud.rousselle@iut-dijon.u-bourgogne.fr
http://arousselle.perso.math.cnrs.fr/

Année universitaire 2022-2023
Semestre 1

Contexte et objectifs

- Développer un panel de méthodes permettant de représenter et synthétiser des données collectées
- Les données sont obtenues soit sur une population entière soit sur un échantillon choisi au hasard dans cette population
- On souhaite fournir
  - une visualisation (représentation graphique) d'un phénomène aléatoire
  - une description simple du phénomène à l'aide d'un nombre limité de valeurs (paramètres statistiques)
- statistiques descriptives $\neq$ statistiques inférentielles

Vocabulaire statistique

Définition

On appelle :

1. *population tout ensemble étudié par la statistique ;*
2. *individu tout élément de la population ;*
3. *effectif total le nombre d'individus dans la population.*

Notations

On note $\mathcal{P}$ la population et N l'effectif total de cette population.

Exemple

1. *Si la population $\mathcal{P}$ est l'ensemble des étudiants de la promotion, un individu est un étudiant de la promotion.*
2. *Si la population $\mathcal{P}$ est l'ensemble des jours du mois de septembre 2016, chaque jour de ce mois est un individu et l'effectif total est $N = 30$.*

Vocabulaire statistique

Définition

Soit $\mathcal{P}$ une population.
On appelle variable statistique (ou caractère) une quantité ou qualité définie sur $\mathcal{P}$ susceptible de varier d'un individu à l'autre. On appelle modalités les différentes valeurs ou aspects pris par cette variable.
On distingue :

1. *les variables qualitatives pour lesquelles les modalités ne sont pas objectivement comparables ;*
2. *les variables quantitatives (ou ordinales) dont les modalités sont mesurables et comparables deux à deux.*

Parmi les variables quantitatives, on distingue :

1. *les variables quantitatives discrètes dont les valeurs possibles sont isolées ;*
2. *les variables quantitatives continues pouvant prendre toutes les valeurs contenues dans un intervalle.*

Vocabulaire statistique

Notation

Les variables statistiques sont désignées par des lettres majuscules, généralement X ou Y.

Exemple

1. *Si la population $\mathcal{P}$ est l'ensemble des étudiants de la promotion, une variable statistique peut être la couleur des yeux (variable qualitative) ou son âge (variable quantitative discrète).*
2. *Si la population $\mathcal{P}$ est l'ensemble des jours de septembre 2016, une variable statistique peut être la hauteur totale des précipitations journalières relevées à Dijon (variable quantitative continue).*

Cas quantitatif discret sans regroupement en classe

Données

- X : variable statistique quantitative discrète sur une population $\mathcal{P}$
- N : effectif total
- $x_1 < x_2 < \cdots < x_r$: les modalités
- $n_1, n_2, \ldots, n_r$: les effectifs associés à ces modalités

Pour obtenir le tableau statistique complet, on ajoute

- $f_i = \dfrac{n_i}{N}$: la fréquence de la modalité x_i
- $N_i = \sum_{k=1}^{i} n_k = n_1 + n_2 + \cdots + n_i$: l' Effectif Cumulé Croissant (ECC) jusqu'à la modalité x_i
- $F_i = \sum_{k=1}^{i} f_k = \dfrac{N_i}{N}$: la Fréquence Cumulée Croissante (FCC)

Cas quantitatif discret sans regroupement en classe

Exemple

Une enquête réalisée auprès d'une population de 100 femmes de 40 ans a recensé le nombre d'enfant(s) de chacune.

Modalités x_i	*0*	*1*	*2*	*3*	*4*	*5*	*6*
Effectifs n_i	*9*	*28*	*32*	*24*	*4*	*2*	*1*

Cas quantitatif discret sans regroupement en classe

Exemple

Une enquête réalisée auprès d'une population de 100 femmes de 40 ans a recensé le nombre d'enfant(s) de chacune.

Modalités x_i	*0*	*1*	*2*	*3*	*4*	*5*	*6*
Effectifs n_i	*9*	*28*	*32*	*24*	*4*	*2*	*1*

- *Population : le contingent de femmes interrogées*
- *Effectif total :* $N = 100$
- *Variable statistique étudiée* X *: le nombre d'enfant(s) par femme*
- *Nature : quantitative discrète*

Cas quantitatif discret sans regroupement en classe

Exemple

Une enquête réalisée auprès d'une population de 100 femmes de 40 ans a recensé le nombre d'enfant(s) de chacune.

Modalités x_i	0	1	2	3	4	5	6
Effectifs n_i	9	28	32	24	4	2	1
Fréquences $f_i = \frac{n_i}{N}$	0,09	0,28	0,32	0,24	0,04	0,02	0,01
ECC $N_i = \sum_{k=1}^{i} n_k$	9	37	69	93	97	99	100
FCC $F_i = \frac{N_i}{N}$	0,09	0,37	0,69	0,93	0,97	0,99	1

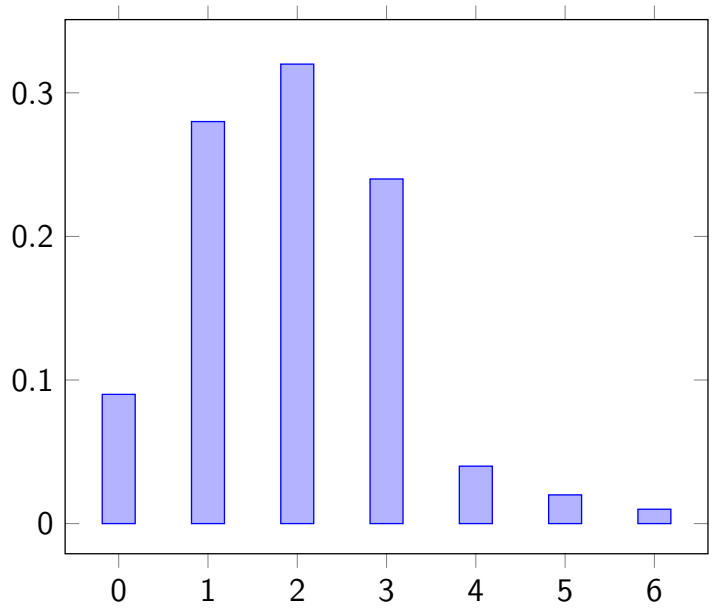
- *Population : le contingent de femmes interrogées*
- *Effectif total :* $N = 100$
- *Variable statistique étudiée X : le nombre d'enfant(s) par femme*
- *Nature : quantitative discrète*

Cas quantitatif discret sans regroupement en classe

Représentation graphique 1 : Histogramme des fréquences

Diagramme en bâtons : la hauteur du bâton associé à x_i = fréquence f_i

Exemple (suite)



On observe dans ce cas une traine (ou queue de distribution) étalée vers la droite.

Cas quantitatif discret sans regroupement en classe

Notation

$\mathbf{P}[X < t]$: *la fréquence totale des modalités x_i telles que $x_i < t$*
autrement dit, la proportion de données $< t$
$\mathbf{P}[X \leq t]$: *la proportion de données $\leq t$*
$\mathbf{P}[X > t]$: *la proportion de données $> t$*
$\mathbf{P}[X \geq t]$: *la proportion de données $\geq t$*
$\mathbf{P}[t_1 < X < t_2]$: *la proportion de données strictement comprises entre t_1 et t_2*
...

Fonction de répartition - Représetation graphique 2

$$F : \mathbf{R} \longrightarrow [0,1]$$
$$t \longmapsto F(t) = \mathbf{P}[X \leq t]$$

Cas quantitatif discret sans regroupement en classe

Proposition

Soit X une variable statistique discrète dont les modalités sont $x_1, \ldots, x_r$.

1. *La fonction de répartition F de X est croissante, en escalier telle que*

$$F(t) = \begin{cases} 0 & si\ t < x_1 \\ F_i & si\ x_i \leq t < x_{i+1},\ i = 1, \ldots, r-1 \\ 1 & si\ t \geq x_r \end{cases} .$$

2. *On a :*

$$\mathbf{P}[X > t] = 1 - \mathbf{P}[X \leq t] = 1 - F(t),$$

$$\mathbf{P}[t_1 < X \leq t_2] = \mathbf{P}[X \leq t_2] - \mathbf{P}[X \leq t_1] = F(t_2) - F(t_1),$$

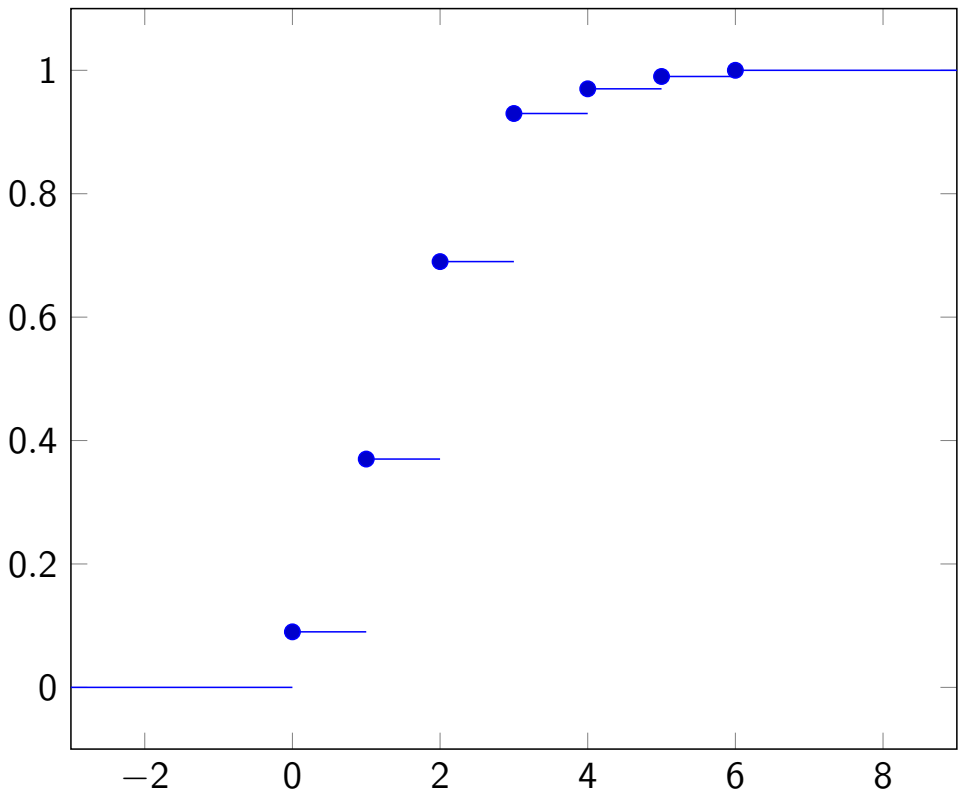
$$\mathbf{P}[X = x_i] = f_i \quad et \quad \mathbf{P}[X = t] = 0\ si\ t \notin \{x_1, \ldots, x_r\},$$

$$\mathbf{P}[X < t] = \mathbf{P}[X \leq t] - \mathbf{P}[X = t]$$

$$= \begin{cases} F(x_i) - f(x_i) = F(x_{i-1}) & si\ t = x_i \\ F(t) & si\ t \notin \{x_1, \ldots, x_r\} \end{cases} .$$

Cas quantitatif discret sans regroupement en classe

Exemple (suite)



Cas quantitatif discret sans regroupement en classe

Paramètres de position

Moyenne arithmétique :

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{r} n_i x_i = \frac{n_1 x_1 + \cdots + n_r x_r}{N} = \sum_{i=1}^{r} f_i x_i.$$

Mode : toute modalité x_i dont l'effectif est maximal parmi tous les effectifs.
Médiane : toute valeur me telle que

$$\mathbf{P}\left[X \leq me\right] \geq 0,5 \quad \textbf{et} \quad \mathbf{P}\left[X \geq me\right] \geq 0,5.$$

Quantile d'ordre $p \in]0,1[$: toute valeur q_p telle que :

$$\mathbf{P}\left[X \leq q_p\right] \geq p \quad \textbf{et} \quad \mathbf{P}\left[X \geq q_p\right] \geq 1-p.$$

Quantiles d'ordre 0,25 et 0,75 : 1^{er} et 3^{e} quartiles
Quantiles d'ordre 0,1 et 0,9 : 1^{er} et 9^{e} déciles

Cas quantitatif discret sans regroupement en classe

Exemple (suite)

Moyenne :

$$\overline{X} = \frac{9 \times 0 + 28 \times 1 + 32 \times 2 + 24 \times 3 + 4 \times 4 + 2 \times 5 + 1 \times 6}{100} = 1,96.$$

Mode : 2 (l'effectif correspondant est 32 est maximal)

Médiane : $me = 2$
En effet, $\mathbf{P}[X \leq 2] = F(2) = 0,69 \geq 0,5$ *et*
$\mathbf{P}[X \geq 2] = 1 - F(1) = 1 - 0,37 = 0,63 \geq 0,5$

Quartiles : $q_{0,25} = 1$ *et* $q_{0,75} = 3$

Cas quantitatif discret sans regroupement en classe

Paramètres de dispersion

Étendue : $x_r - x_1$ où x_1 est la plus petite modalité et x_r la plus grande modalité
Étendue inter-quartiles : $q_{0,75} - q_{0,25}$ où $q_{0,25}$ et $q_{0,75}$ sont les premier et troisième quartiles.
Variance :

$$\mathrm{V}[X] = \frac{1}{N}\sum_{i=1}^{r} n_i (x_i - \overline{X})^2 = \frac{1}{N}\sum_{i=1}^{r} n_i x_i^2 - \overline{X}^2$$

Écart-type : $\sigma = \sqrt{\mathrm{V}[X]}$
Écart-type relatif ou coefficient de variation : $\mathrm{c_v} = \dfrac{\sigma}{\overline{X}}$ si $\overline{X} \neq 0$
Écart absolu moyen :

$$\mathrm{EAM} = \frac{1}{N}\sum_{i=1}^{r} n_i |x_i - \overline{X}|$$

Cas quantitatif discret sans regroupement en classe

Exemple (suite)

Étendue : $x_7 - x_1 = 6 - 0 = 6$
Étendue inter-quartiles : $q_{0,75} - q_{0,25} = 3 - 1 = 2$
Variance :

$$\mathbf{V}[X] = \frac{9 \times (0 - 1,96)^2 + \cdots + 1 \times (6 - 1,96)^2}{100} \simeq 1,3784.$$

Écart-type : $\sigma = \sqrt{1,3784} \simeq 1,1741$

Coefficient de variation : $c_v = \dfrac{\sqrt{1,3784}}{1,96} \simeq 0,5990$

Écart absolu moyen :

$$\text{EAM} = \frac{9 \times |0 - 1,96| + \cdots + 1 \times |6 - 1,96|}{100} = 0,8904$$

Cas quantitatif continu (ou discret avec regroupements en classes)

Changements

- On n'a pas des modalités isolées x_i mais des classes $]b_{i-1}, b_i]$
- Les classes peuvent être d'amplitudes différentes
  - plus de travail pour le prendre en compte
- La fonction de répartition est continue (interpolation linéaire) et $\mathbf{P}[X = t] = 0$ pour tout t
- En supposant les données réparties uniformément dans chaque classe, on choisira comme représentant d'une classe son centre pour le calcul de certains paramètres

Cas quantitatif continu (ou discret avec regroupements en classes)

À partir des données, pour obtenir le tableau statistique complet, on ajoute

- $f_i = \dfrac{n_i}{N}$: la fréquence de la modalité x_i
- $N_i = \sum_{k=1}^{i} n_k$ l'ECC (jusqu'à la fin de la classe C_i, *i.e.* en b_i)
- $F_i = \sum_{k=1}^{i} f_k = \dfrac{N_i}{N}$: la FCC
- $a_i = b_i - b_{i-1}$: l'amplitude de la classe $C_i =]b_{i-1}, b_i]$
- $a_i^r = \dfrac{a_i}{A}$: l'amplitude relative de la classe C_i, A à choisir
- $n_i^r = \dfrac{n_i}{a_i^r}$: l'effectif relatif de la classe C_i
- $f_i^r = \dfrac{f_i}{a_i^r} = \dfrac{n_i^r}{N}$: la fréquence relative de la classe C_i
- $c_i = \dfrac{b_i + b_{i-1}}{2}$: le centre de la classe C_i

Cas quantitatif continu (ou discret avec regroupements en classes)

Exemple

Durant le mois de janvier 2016, on a relevé les précipitations journalières (exprimées en mm) sur Dijon.

Hauteur des précipitations (en mm)	$[0;1[$	$[1;3[$	$[3;6[$	$[6;9[$	$[9;14[$
Nombre de jours	*17*	*5*	*5*	*2*	*2*

Cas quantitatif continu (ou discret avec regroupements en classes)

Exemple

Durant le mois de janvier 2016, on a relevé les précipitations journalières (exprimées en mm) sur Dijon.

Hauteur des précipitations (en mm)	$[0;1[$	$[1;3[$	$[3;6[$	$[6;9[$	$[9;14[$
Nombre de jours	*17*	*5*	*5*	*2*	*2*

Population : *l'ensemble des jours sur le mois de janvier 2016*
Variable statistique étudiée X : *hauteur des précipitations relevées à Dijon /jour*
Nature : *quantitative continue*
Choix de l'amplitude unitaire : $A = 1$

Cas quantitatif continu (ou discret avec regroupements en classes)

Exemple

Durant le mois de janvier 2016, on a relevé les précipitations journalières (exprimées en mm) sur Dijon.

i	C_i	c_i	$a_i = a_i^r$	n_i	f_i	n_i^r	f_i^r	N_i	F_i
1	$[0;1[$	0,5	1	17	0,55	17	0,55	17	0,55
2	$[1;3[$	2	2	5	0,16	2,5	0,08	22	0,71
3	$[3;6[$	4,4	3	5	0,16	1,67	0,05	27	0,87
4	$[6;9[$	7,5	3	2	0,06	0,67	0,02	29	0,94
5	$[9;14[$	11,5	5	2	0,06	0,4	0,01	31	1

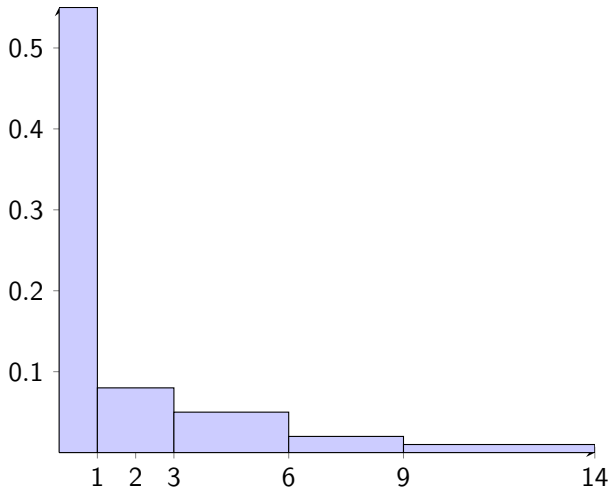
Population : *l'ensemble des jours sur le mois de janvier 2016*
Variable statistique étudiée X : *hauteur des précipitations relevées à Dijon /jour*
Nature : *quantitative continue*
Choix de l'amplitude unitaire : $A = 1$

Cas quantitatif continu (ou discret avec regroupements en classes)

Représentation graphique 1 - Histogramme des fréquences (relatives)

Diagramme en rectangles tel que l'**aire** du rectangle associé à une classe représente sa fréquence.
Le rectangle associé à une classe a pour largeur son amplitude et pour **hauteur** sa fréquence **relative**.

Exemple (suite)



Sur l'histogramme des fréquences (relatives), on observe une traine (ou queue de distribution) étalée vers la droite

Cas quantitatif continu (ou discret avec regroupements en classes)

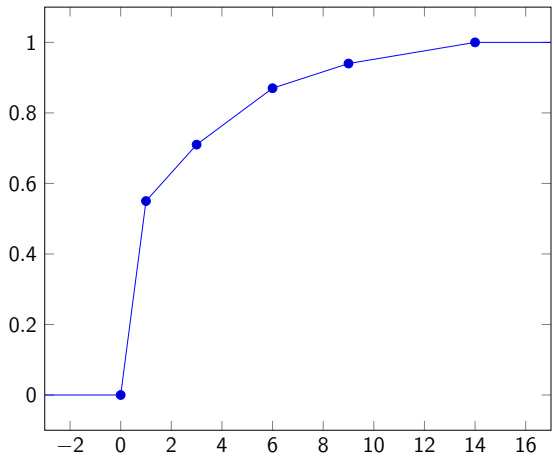
Représentation graphique 2 - Fonction de répartition

Polygone des fréquences cumulées : ligne polygonale dont l'ordonnée est nulle pour $x \leq b_0$, vaut 1 pour $x \geq b_r$, et interpolant linéairement les points

$$(b_0, 0), (b_1, F_1), \ldots, (b_{r-1}, F_{r-1}), (b_r, F_r = 1)$$

$\rightsquigarrow$ Approximation de la fonction de répartition en supposant une répartition uniforme des données au sein de chaque classe

Exemple (suite)



Cas quantitatif continu (ou discret avec regroupements en classes)

Proposition

Soit X une variable aléatoire continue.

1. *La fonction de répartition de X est croissante et* **continue**. *En particulier, on a pour tout $x \in \mathbf{R}$:*

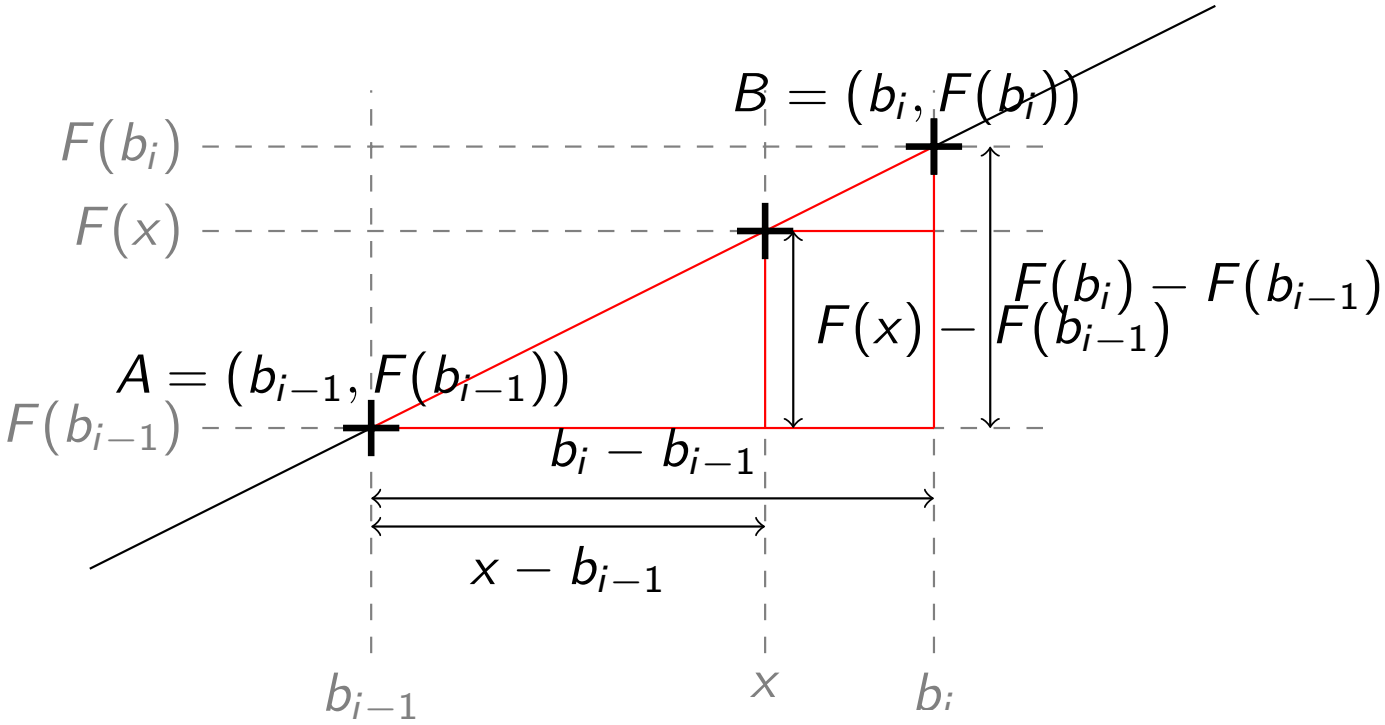
$$\mathbf{P}[X < x] = \mathbf{P}[X \leq x] = F(x) \quad \text{et} \quad \mathbf{P}[X = x] = 0.$$

2. *Pour tout x dans la classe $C_i = [b_{i-1}, b_i[$, on a :*

$$\begin{aligned} F(x) &= \frac{x - b_{i-1}}{b_i - b_{i-1}} \left(F(b_i) - F(b_{i-1})\right) + F(b_{i-1}) \\ &= \frac{f_i^r}{A}(x - b_{i-1}) + F(b_{i-1}), \end{aligned}$$

Cas quantitatif continu (ou discret avec regroupements en classes)

La méthode d'interpolation linéaire



Cas quantitatif continu (ou discret avec regroupements en classes)

Exemple (suite)

Déterminons $F(2) = \mathbf{P}[X \leq 2]$ *par interpolation linéaire dans l'Exemple précédent.*
Meilleur encadrement de 2 par des bornes de classes : $1 < 2 < 3$
La méthode d'interpolation linéaire donne :

$$\frac{F(2) - F(1)}{F(3) - F(1)} = \frac{2-1}{3-1} = \frac{1}{2}$$

donc

$$F(2) - F(1) = \frac{1}{2}(F(3) - F(1))$$

donc

$$F(2) = \frac{1}{2}(F(3) - F(1)) + F(1) = \frac{1}{2}(0,71 - 0,55) + 0,55 = 0,63.$$

Ainsi, $\mathbf{P}[X \leq 2] = F(2) = 0,63$ *et on a relevé au plus 2mm de précipitations durant 63% des jours.*

Cas quantitatif continu (ou discret avec regroupements en classes)

Paramètres de position - ce qui change peu

Moyenne arithmétique :

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{r} n_i c_i = \frac{n_1 c_1 + \cdots + n_r c_r}{N} = \sum_{i=1}^{r} f_i c_i .$$

Paramètres de position - ce qui change plus

Classe modale : toute classe C_i dont l'effectif relatif est maximal parmi tous les effectifs relatifs.

Médiane : l'unique valeur me telle que $\mathbf{P}[X \leq me] = 0,5$
$\rightsquigarrow$ interpolation linéaire

Quantile d'ordre $p \in]0,1[$: l'unique valeur q_p telle que :

$$\mathbf{P}[X \leq q_p] = p$$

Cas quantitatif continu (ou discret avec regroupements en classes)

Exemple (suite)

Classe modale : $C_1 = [0;1[$ *(effectif relatif le plus élevé)*
Moyenne (arithmétique) :

$$\overline{X} = \frac{n_1 c_1 + \cdots + n_5 c_5}{N} \simeq 2,5483871;$$

Médianne : $me = \dfrac{10}{11}$
En effet, il faut trouver la valeur me telle que $F(me) = \mathbf{P}[X \leq me] = 0,5$.
On a $F(0) = 0$ *et* $F(1) = 0,55$ *et cette valeur est dans la classe* $C_1 = [0;1[$.
On détermine me par interpolation linéaire :

$$\frac{me - 0}{1 - 0} = \frac{F(me) - F(0)}{F(1) - F(0)}$$

soit

$$me = \frac{0,5}{0,55} = \frac{10}{11} \simeq 0,9090$$

Cas quantitatif continu (ou discret avec regroupements en classes)

Paramètres de dispersion - peu de changements

Étendue : $b_r - b_0$ où b_0 est la borne inférieure de la 1re classe et b_r la borne suprérieure de la dernière classe

Étendue inter-quartiles : $q_{0,75} - q_{0,25}$ où $q_{0,25}$ et $q_{0,75}$ sont les premier et troisième quartiles.

Variance :

$$\mathrm{V}[X] = \frac{1}{N}\sum_{i=1}^{r} n_i(c_i - \overline{X})^2 = \frac{1}{N}\sum_{i=1}^{r} n_i {c_i}^2 - \overline{X}^2$$

Écart-type : $\sigma = \sqrt{\mathrm{V}[X]}$

Écart-type relatif ou coefficient de variation : $\mathrm{c_v} = \dfrac{\sigma}{\overline{X}}$ si $\overline{X} \neq 0$

Écart absolu moyen :

$$\mathrm{EAM} = \frac{1}{N}\sum_{i=1}^{r} n_i |c_i - \overline{X}|$$

Cas quantitatif continu (ou discret avec regroupements en classes)

Exemple (suite)

Étendue : $b_5 - b_0 = 14 - 0 = 14$

Variance :

$$\mathrm{V}[X] = \frac{n_1(c_1 - \overline{X})^2 + \cdots + n_5(c_5 - \overline{X})^2}{N} = \frac{n_1 c_1^2 + \cdots + n_5 c_5^2}{N} - \overline{X}^2 \simeq 9,71540;$$

Écart-type :

$$\sigma = \sqrt{\mathrm{V}[X]} \simeq 3,11695;$$

Coefficient de variation :

$$\mathrm{c_v} = \frac{\sigma}{\overline{X}} \simeq 1,22310;$$

Écart-absolu moyen :

$$\mathrm{EAM} = \frac{n_1|c_1 - \overline{X}| + \cdots + n_5|c_5 - \overline{X}|}{N} \simeq 2,42352.$$