Statistiques descriptives

Cours 4

Paramètres de dispersion

Paramètres de dispersion

Informations sur la répartition des valeurs autour de la valeur centrale de référence

- Etendue ou le rang
- Quartiles/déciles
- Variance et écart type

Etendue

- L'étendue d'une série statistique quantitative est la différence entre la plus grande valeur de la variable et la plus petite valeur.

Rang	Genre	QI
1	F	151
2	F	111
3	M	111
4	F	109
5	F	105
6	M	102
7	M	98
8	F	96

L'étendue
151- 96 = 55

Etendue

- L'étendue d'une série statistique quantitative est la différence entre la plus grande valeur de la variable et la plus petite valeur.
- Si l'étendue est très petite, alors il y a peu d'écart entre toutes les valeurs de la série. Celle-ci est homogène.
- Si au contraire l'étendue est grande, alors l'écart est important entre la plus petite et la plus grande valeur.

Quantiles

Les quantiles ont différents noms selon le nombre de parts dans la population

- Si la population est séparée en 4, ce sont des quartiles
- Si la population est séparée en 5, ce sont des quintiles
- Si la population est séparée en 10, ce sont des déciles
- Si la population est séparée en 100, ce sont des centiles

Quantiles: les quartiles

- Les quartiles permettent de séparer une série statistique en quatre groupes de plus ou moins même effectif
- Quartiles Q1, Q2, Q3 tels que
  - ✓ Au moins 25% des valeurs prises par la série sont inférieures ou égales à Q1
  - ✓ Au moins 25% des valeurs prises par la série sont supérieures ou égales à Q3
    - ou moins 75 % des données soient inférieures ou égales à Q_3.
  - ✓ Q2 est la médiane
  - ✓ L'écart interquartile (Q3-Q1) est un paramètre de dispersion absolue qui correspond à l'étendue de la distribution une fois que l'on a retiré les 25% des valeurs les plus faibles et les 25% des valeurs les plus fortes.
  - ✓ [Q3 ; Q1] est l'intervalle interquartile, il contient au moins (environ) 50% des valeurs de la série

Quantiles: les quartiles

Exemple les notes à un DS:

3-5-5-6-7-8-8-9-9-10-10-10-10-11-11-12-13-13-13-14-15-16-19

On écrit la série sous forme d'un tableau

Notes	3	5	6	7	8	9	10	11	12	13	14	15	16	19
Eff.	1	2	1	1	2	2	4	2	1	3	1	1	1	1
Eff. Cum.	1	3	4	5	7	9	13	15	16	19	20	21	22	23

Quantiles: les quartiles

Exemple les notes à un DS:

3-5-5-6-7-8-8-9-9-10-10-10-10-11-11-12-13-13-13-14-15-16-19

- N= 23 valeurs
- Position médiane = (23+1)/2=12, valeur=10
- Position Q1= 23/4~5,75, on prend l'entier juste supérieur 6, Q1=8
  - ✓ 7 valeurs , 30% des notes ≤ 8
- Position Q3=3x23/4~17,25 on prend l'entier juste supérieur 18, Q3=13
  - ✓ 7 valeurs , 30% des notes ≥ 13
- [Q3;Q1] 14 valeurs , 60% des notes entre 8 et 13

Quantiles: boite à moustaches

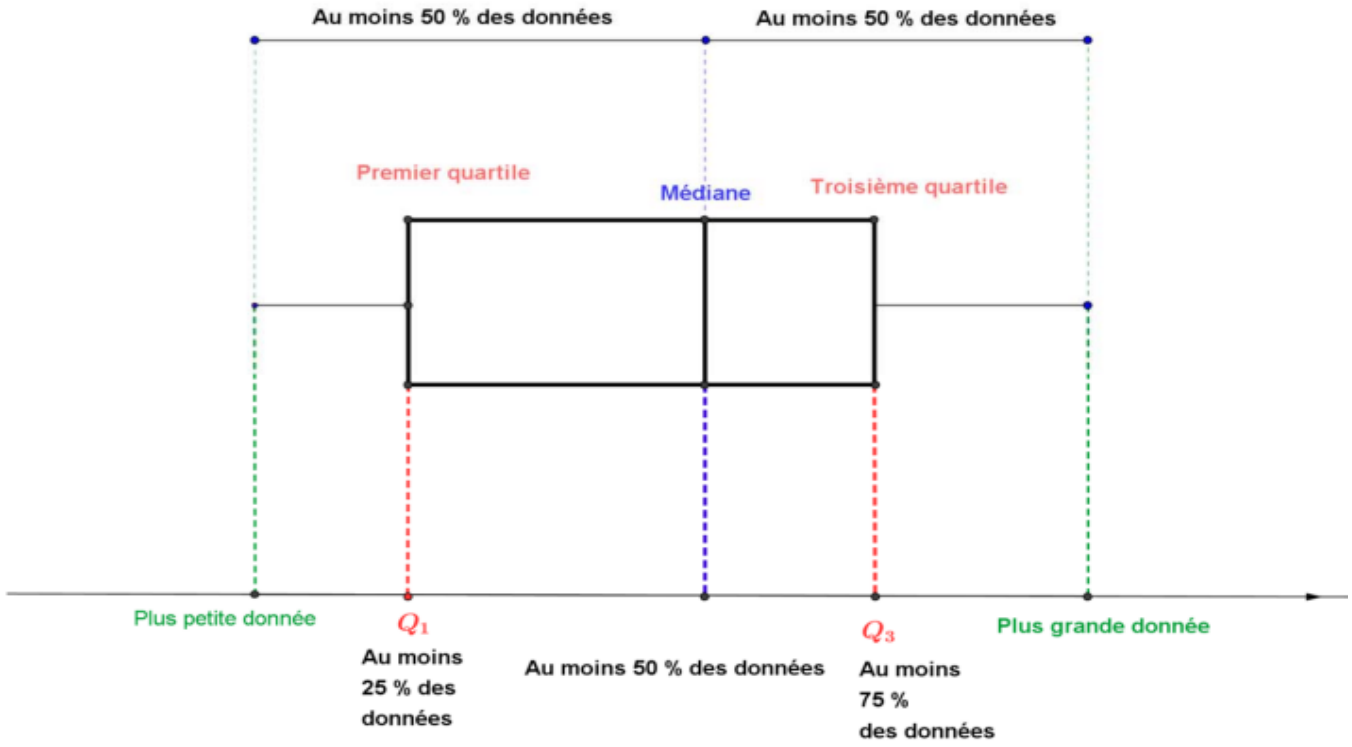
(ou diagramme en boîte, boite de Turkey ou *box-plot*)

- Moyen rapide de figurer le profile essentiel d'une série statistique quantitative.
- Un autre intérêt est de pouvoir faire facilement des comparaisons entre des groupes de données.
- Plusieurs échantillons peuvent être représentés simultanément et comparés par des *box-plots* les uns à côté des autres.

Quantiles: boite à moustaches

(ou diagramme en boîte, boite de Turkey ou *box-plot*)

- Définition (Wikipedia):
  Boite à moustaches pour quartiles: Il s'agit de tracer un rectangle allant du 1er quartile au troisième quartile coupé par la médiane



Quartiles: remarque sur les déciles

- Les déciles permettent de séparer une série statistique en dix groupes de plus ou moins même effectif
  - ✓ Le premier décile d'une série la plus petite valeur D1 des termes de la série pour laquelle au moins un dixième (10%) des données sont inférieures ou égales à D1
  - ✓ Le neuvième décile (D9) d'une série est la plus petite valeur des termes de la série pour laquelle au moins neuf dixièmes (90%) des données sont inférieures ou égales à D9
  - ✓ L'intervalle interdécile est [D1;D9]

Variance

Moyenne: somme des valeurs numériques divisée par le nombre de ces valeurs numériques

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

- La variance est un indicateur de la dispersion d'une série par rapport à sa moyenne

$$\mathrm{V(x)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$$

ou

$$\mathrm{V(x)} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - m^2$$

La variance se définit comme la somme pondérée des carrés des écarts des valeurs de la série à la moyenne.

Ecart-type

- L'écart-type sert à mesurer la dispersion, ou l'étalement, d'un ensemble de valeur autour de leur moyenne

$$\sigma = \sqrt{V(x)}$$

  - ✓ L'écart-type n'est jamais négatif.
  - ✓ L'écart-type est sensible aux valeurs aberrantes.
    Une seule valeur aberrante peut accroître l'écart-type et, par le fait même, déformer le portrait de la dispersion.

- Si l'écart-type est faible, cela signifie que les valeurs sont assez concentrées autour de la moyenne

- Si l'écart-type est élevé, cela veut dire au contraire que les valeurs sont plus dispersées autour de la moyenne.

Exemple

Répartition des notes d'une classe, plus l'écart type est faible, plus la classe est homogène.

Remarque

- Quand il s'agit décrire la dispersion d'une distribution on divise la sommes des carrés des écarts à la moyenne par n
- Lorsque la variance (ou l'écart-type) est calculée sur un échantillon et doit servir à faire une inférence sur la variance de la population dont l'échantillon est extrait on sait que l'estimation est biaisée
- Il faudra corriger par n-1 au lieu de n ainsi:

$$\sigma_c = \sqrt{\frac{n}{n-1}}\sigma$$

Notion de corrélation

Coefficient de corrélation de Bravais-Pearson, ou encore coefficient de corrélation linéaire

- ✓ Il permet d'analyser les relations linéaires. Il existe d'autre coefficient de corrélation comme celle de Spearman (des relations non-linéaires)

- Le coefficient de corrélation permet de mesurer la liaison ou le lien entre 2 ensembles de données.
  - ✓ Est-ce que le niveau d'études atteint dépend du milieu social ?
  - ✓ Est-ce que la mémorisation des mots d'un texte dépend de la longueur des mots ?
  - ✓ Est-ce que le loisir préféré des étudiants dépend de leur sexe ?

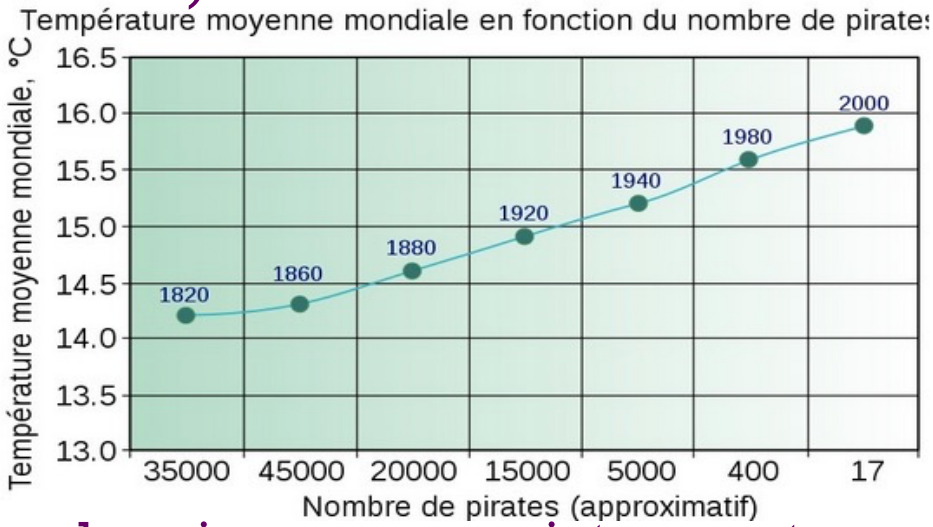
- Toutes ces questions mettent en jeu deux variables. Ces deux variables sont observées sur la même population.

- Attention ce n'est pas parce qu'une corrélation existe entre 2 séries statistiques qu'il y a un lien de cause à effet entre les deux !

Notion de corrélation

- Exemple:

✓ Le réchauffement planétaire, les tremblements de terre, les cyclones et les autres catastrophes naturelles sont une conséquence directe du nombre décroissant de Pirates depuis les années 1800 (cf.. Pastafarisme)



✓ S'il y a plus de naissances au printemps et en automne, on risque de trouver une corrélation entre le nombre de passage de cigognes et le nombre de naissance.

Notion de corrélation

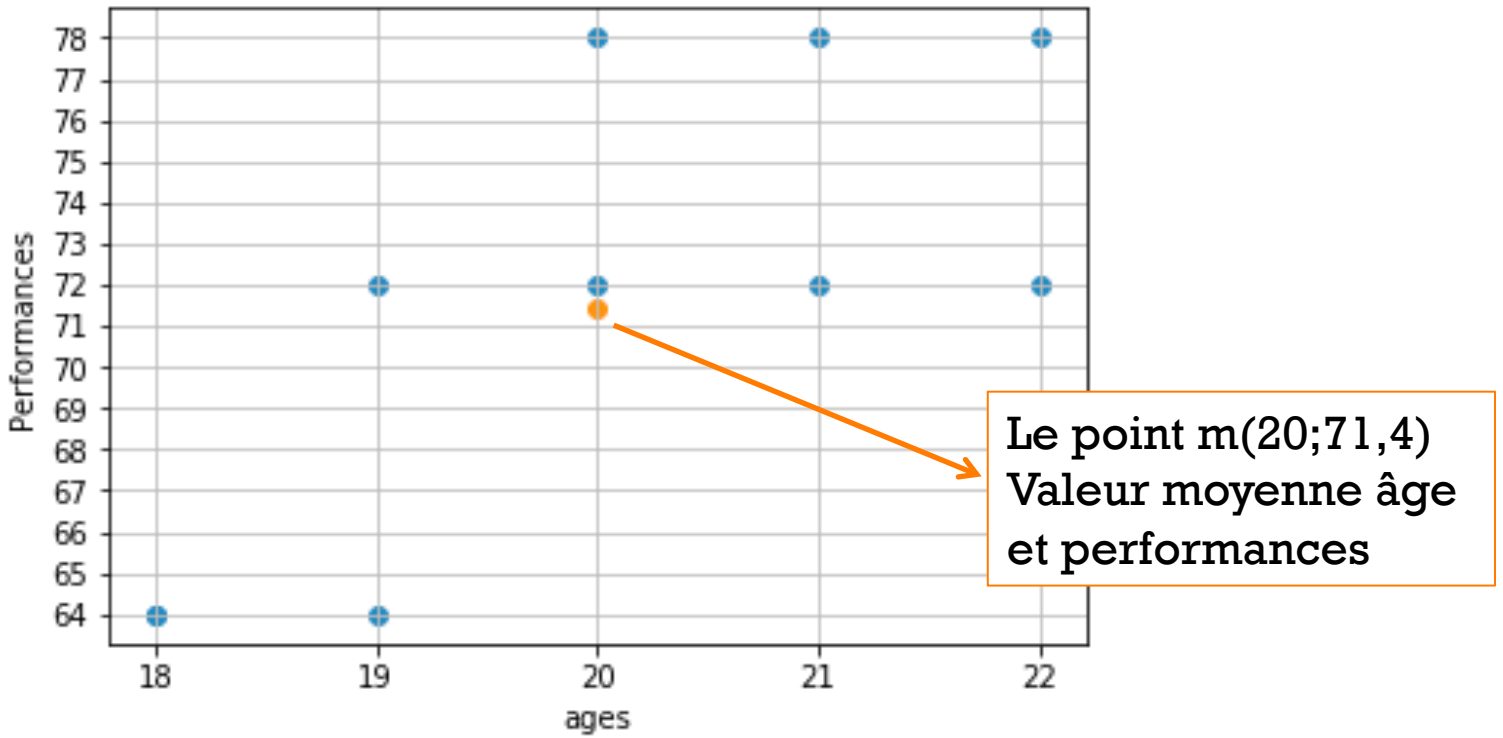
- Exemple:
  On considère une population de 10 personnes, auquel on attribue une note indiquant leurs performances mémorielles.
  On note leur âge x et leur performances mémorielles y.
  Les données mesurées sont les suivantes :

x	18	21	21	19	22	20	19	18	22	20
y	64	72	78	64	72	78	72	64	78	72

- Vous savez comment étudier d'une part l'âge des individus, et d'autre part leur performances mémorielles, de manière complètement dissociée.
  - ✓ Intérêt réside dans la possibilité de mettre en évidence un lien entre l'âge et la mémoire, ce qui nécessite l'étude simultanée de ces deux propriétés.

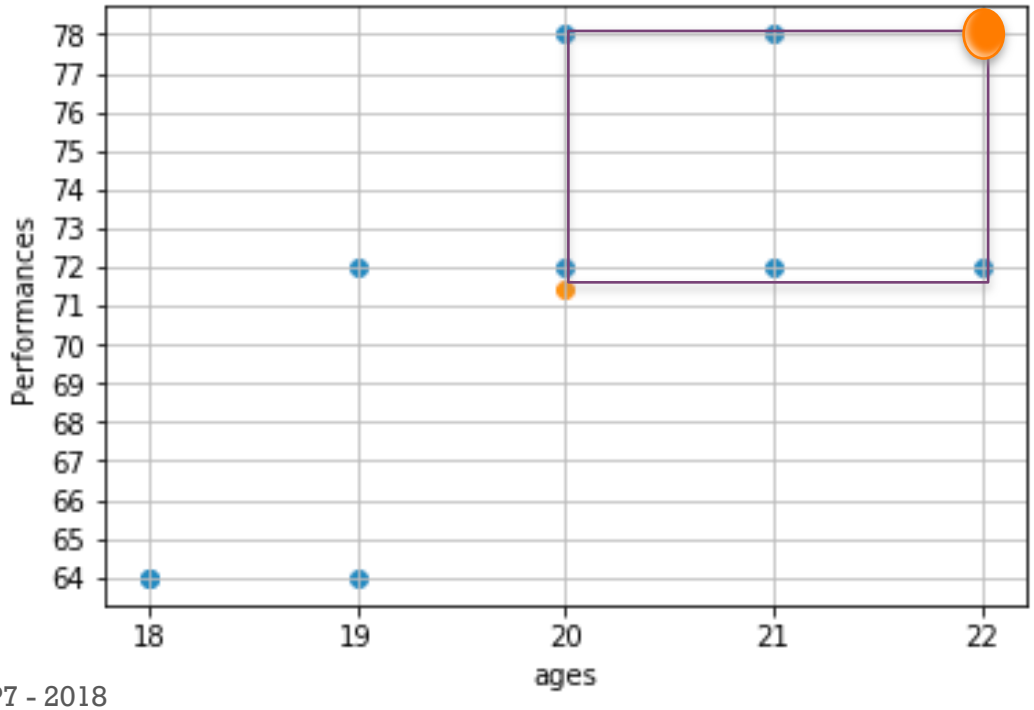
Notion de corrélation

- Une façon de synthétiser efficacement les données et de se faire une idée du lien entre les deux variables consiste à réaliser un nuage de points



Notion de corrélation

- A chaque point $(x_i ; y_i)$ on associe son écart par rapport au point moyen $(m_x ; m_y)=(20 ; 71,4)$.
  On obtient un couple d'écarts : $(x_i - m_x ; y_i - m_y)$.
  Exemple au point (22 ; 78) on associe le couple d'écarts:
  (22-20 ; 78-71,4) = (2 ; 6,6)



Notion de corrélation

- Le coefficient de Pearson permet de détecter la présence ou l'absence d'une relation linéaire entre deux caractères quantitatifs continus.

- Pour calculer ce coefficient il faut tout d'abord calculer la covariance.
  C'est la moyenne du produit des écarts à la moyenne:
  Cov(x ; y) = m ((x − m (y)) × (y − m (y))
  (le « m » indique la moyenne)

$$cov(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}) \times (y_i - \bar{y})$$

Notion de corrélation

- La covariance peut aussi s'écrire:

  Cov(x ; y) = m (xy) − m (x) m (y) .

  Cette seconde expression est plus efficace pour les calculs

$$cov(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i \times y_i) - (\bar{x} \times \bar{y})$$

On définit le coefficient de corrélation linéaire de deux caractères x et y comme covariance de x et y divisée par le produit des écarts-types de x et y

$$r(x,y) = \frac{cov(x,y)}{\sigma_x \times \sigma_y}$$

Notion de corrélation

Propriétés et interprétation de r(x,y)

- r(x,y) varie entre -1 et +1
  - ✓ si *r* est proche de 0, il n'y a pas de relation linéaire entre x et y;
  - ✓ si *r* est proche de -1, il existe une forte relation linéaire négative entre x et y;
  - ✓ si *r* est proche de 1, il existe une forte relation linéaire positive entre x et y.
- Le signe de *r* indique donc le sens de la relation tandis que la valeur absolue de *r* indique l'intensité de la relation c.à.d. la capacité à prédire les valeurs de y en fonctions de celles de x
- Reprenons notre exemple des âges x et des performances mémorielles y

Notion de corrélation

$(m_x , m_y)=(20;71,4)$

											Moy.
x	18	21	21	19	22	20	19	18	22	20	20
$x-m_x$	-2	1	1	-1	2	0	-1	-2	2	0	0
y	64	72	78	64	72	78	72	64	78	72	71,4
$y-m_y$	-7,4	0,6	6,6	-7,4	0,6	6,6	0,6	-7,4	6,6	0,6	0
$(x-m_x) \times (y-m_{y)}$	14,8	0,6	6,6	7,4	1,2	0	-0,6	14,8	13,2	0	5,8

On donne $\sigma_x = 1,41$ $\sigma_y = 5,44$ d'où $\underline{r= 0,75}$

Notion de corrélation

- Nous sommes en présence d'une corrélation positive forte, qui semble indiquer qu'il existe une relation linéaire (de type y=ax+b) reliant les performances mémorielles et l'âges des individus dans la population étudiée

Limites du coefficient de Pearson

- En principe, le coefficient de Pearson n'est applicable que pour mesurer la relation entre deux variables x et y ayant une distribution de type gaussien et ne comportant pas de valeur exceptionnelles.
  - ✓ Si ces conditions ne sont pas vérifiées (cas fréquent ...) l'emploi de ce coefficient peut aboutir à des conclusions erronées sur la présence ou l'absence d'une relation.
- On notera également que l'absence d'une relation linéaire ne signifie pas l'absence de toute relation entre les deux caractères étudiés.