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

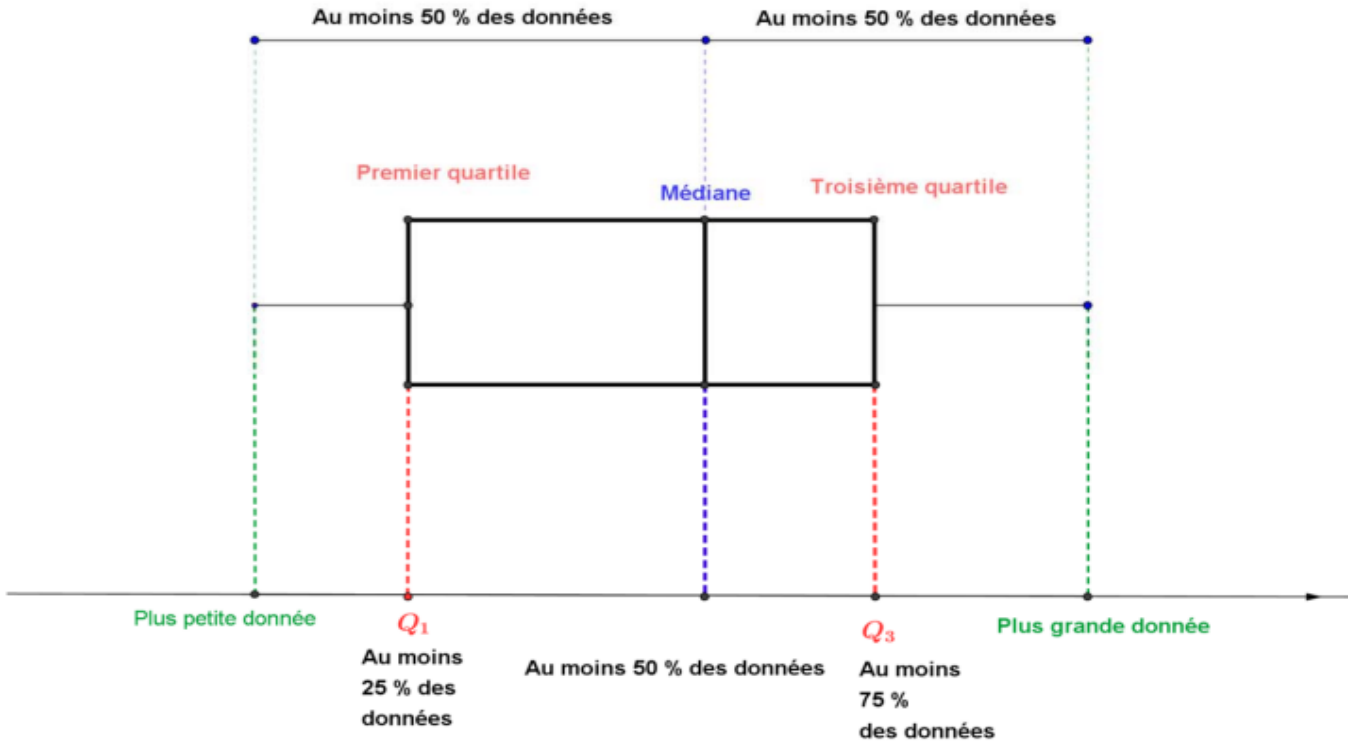
(ou diagramme en boîte, boite de Turkey ou *box-plot*)

- 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



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:

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