Indices de tendance centrale

J-M., GALHARRET

LMJL, Faculté de Psychologie, Nantes Université

Introduction

Introduction

- Les indices de tendance centrale : ils résument dans une certaine mesure les valeurs les plus typiques de l'échantillon.
- Les indices de dispersion : ils mesurent la répartition des valeurs autour des tendances centrales. Ils sont centraux dans la description des données.
- Pour résumer ou décrire un jeu de données on doit toujours associer un indice de tendance centrale et un indice de dispersion ! Mais nous verrons que ce n'est pas suffisant....

Conditions de Yule

Yule a défini des conditions pour qu'un indice soit de bonne qualité, il doit :

- être objectif (Y1),
- tenir compte de toutes les observations (Y2),
- avoir une signification concrète (Y3),
- être simple à calculer (Y4),
- être peu sensible aux fluctuations d'échantillonnage (Y5),
- se prêter au calcul algébrique (Y6).

La moyenne

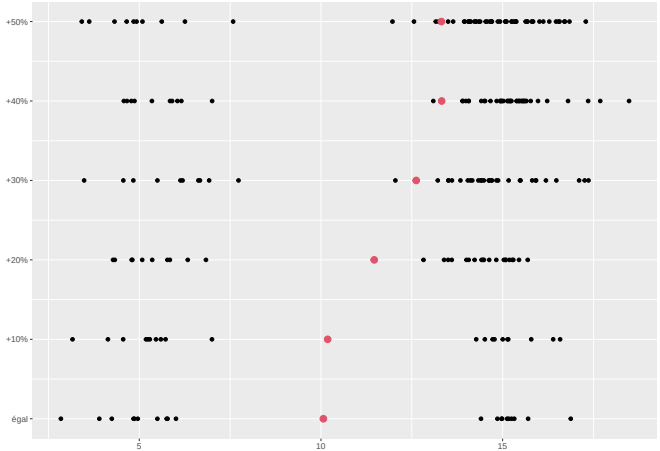
La moyenne arithmétique

- La moyenne est l'indice de tendance centrale le plus utilisé.
- Il ne concerne que les variables mesurées (sur une échelle discrète ou bien une échelle continue).

Définition Si on considère un échantillon de valeurs $(x_i)_{i=1,\ldots,N}$ alors la moyenne de ces valeurs est

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

Interprétation de la moyenne



Avantages/désavantages

- Elle répond bien à tous les critères de Yule.
- La somme des écarts des données à leur moyenne est nulle i.e. $\sum_{i=1}^{N}(x_i - \bar{x}) = 0$.
- La somme des carrés des écarts à une valeur A est minimale lorsque $A = \bar{x}$, i.e. $\sum_{i=1}^{N}(x_i - A)^2$ est minimale lorsque $A = \bar{x}$
- Désavantage : Elle est sensible aux valeurs extrèmes !

Exemple :

Si on considère 100 salariés d'une entreprise qui gagnent entre 2000 € et 4000 € bruts on aura une moyenne comprise entre 2000 et 4000 euros brut.

Admettons qu'on ajoute le salaire du patron (1000000 d'€ brut!) alors la moyenne sera comprise entre 11881.2 et 13861.4 ! La moyenne a augmenté d'environ 10000 €.

Calcul pour des valeurs regroupées en tableaux :

- Variable discrète : les valeurs sont regroupées dans un tableau d'effectifs. Le calcul de la la moyenne est alors

$$\frac{\sum_{j=1}^{J} n_j x_j}{\sum_{j=1}^{J} n_j}.$$

où J est le nombre de valeurs différentes parmi les $(x_i)_{i=1,\dots,N}$.

- Variable regroupée en classe : ici on ne peut donner qu'une approximation de la moyenne. On reprendra le calcul précédent en considérant que les valeurs sont les milieux de classe.

Exemple :

modalité	effectif	fréquence (%)
0	1	2
1	1	2
2	6	12
3	12	24
4	13	26
5	6	12
6	5	10
7	5	10
8	1	2

La moyenne vaut 4.08.
Remarque : On peut aussi utiliser les fréquences pour calculer les moyennes on a alors : $\bar{x} = \sum_{j=1}^{J} f_j x_j$.

On reprend l'exemple précédent et on regroupe les valeurs en classe de longeur 3 :

classe	effectif	fréquence (%)
[0, 3)	8	16
[3, 6)	31	62
[6, 8]	11	22

Une approximation de la moyenne est alors 4.68

Remarque : Cette approximation est éloignée de la valeur précédente car on a regroupé les valeurs dans peu de classes.

La médiane

La médiane

Définition : La médiane $\tilde{x}$ d'un jeu de données est la valeur qui partage une série de données en deux séries de même effectif

Concrètement on classe les observations dans l'ordre croissant, on note alors $\tilde{x}_1 \leq \tilde{x}_2 \leq ... \leq \tilde{x}_N$.

- Si $0.5 \times (N+1)$ est un entier la médiane est $\tilde{x}_{0.5\times(N+1)}$
- Sinon, la médiane est $\frac{\tilde{x}_{\lceil 0.5\times(N+1)\rceil}+\tilde{x}_{\lceil 0.5\times(N+1)\rceil+1}}{2}$

Exemples

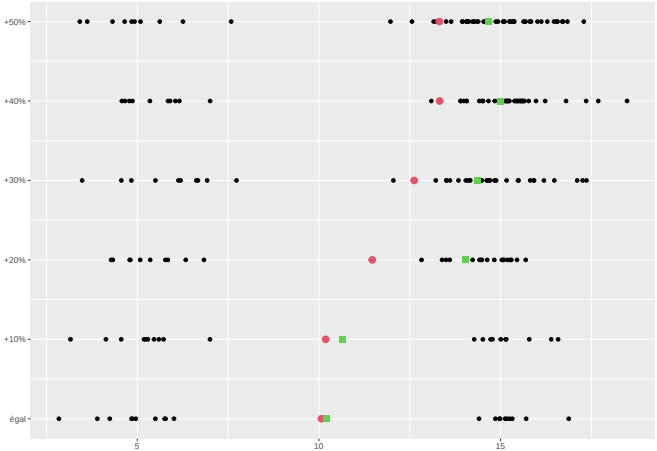
On considère un jeu de 5 données

9 9 13 9 3

On considère un jeu de 6 données

11 11 15 11 4 9

Retour sur l'exemple



Avantages

Elle possède les propriétés suivantes :

- La somme des valeurs absolues des écarts à une valeur A est minimale lorsque $A = \tilde{x}$, i.e. $\sum_{i=1}^{N} |x_i - A|$ est minimale lorsque $A = \tilde{x}$
- Avantage par rapport à la moyenne : Elle est peu sensible aux valeurs extrèmes !

Désavantages

Elle ne répond pas bien à tous les critères de Yule.

- Elle ne tient pas compte des valeurs mais de leur rang.
- Elle est plus sensible aux fluctuations d'échantillon que la moyenne.
- Elle n'est pas pratique pour faire des calculs : par exemple si on a deux jeux de données on peut facilement calculer la moyenne de l'ensemble des données en utilisant la moyenne des deux jeux de données, ce n'est pas le cas avec la médiane.

Exemple

Exemple (retour salaire) Dans le jeu de données initial on avait 100 valeurs donc la médiane était la moyenne entre la 50ième et la 51ième valeur. En ajoutant le salaire de 1000000€ la médiane du jeu de données est alors la 51ième valeur.

Donc l'impact de cette valeur extrême est très faible sur la médiane.

Calcul pour les valeurs regroupées en tableau :

- Variable discrète : les valeurs sont regroupées dans un tableau d'effectifs. Le calcul de la la médiane est identique à celui des données brutes.
- Variable regroupée en classe : ici on ne peut donner qu'une approximation de la médiane. On pourra déterminer la classe dans laquelle se trouve la médiane et prendre le milieu de la classe.

Remarque : L'interpolation linéaire est une méthode plus sophistiquée pour estimer la médiane pour une variable regroupée en classe, elle repose sur la même hypothèse (jamais vérifiée en pratique) selon laquelle les valeurs sont réparties linéairement dans les classes !

Autres indices

Le mode

Définition Le mode est une valeur du jeu de données ayant une fréquence maximale.

Avantages/désavantages

- Il est très facile à déterminer (graphiqement ou dans un tableau d'effectifs ou de fréquences.)
- Il apporte des informations supplémentaires par rapport à la moyenne et à la médiane.

Remarque :

Lorsque les valeurs sont regroupées en classe on peut trouver la classe modale, mais on ne peut pas être sûr que le mode appartienne à cette classe (voir exemple).

Les percentiles

Définition : Le $p^{\text{ième}}$ percentile d'un jeu de données est la plus petite valeur u de la série telle qu'on a au moins $p\%$ des observations sont inférieures ou égales à u.

Concrètement on fait comme pour la médiane ($p = 0.5!$) on classe les observations dans l'ordre croissant, on note alors $\tilde{x}_1 \leq \tilde{x}_2 \leq ... \leq \tilde{x}_N$.

- Si $p \times (N+1)$ est un entier le $p^{\text{ième}}$ percentile est $\tilde{x}_{p\times(N+1)}$
- Sinon, le $p^{\text{ième}}$ percentile est $\frac{\tilde{x}_{\lceil p\times(N+1)\rceil}+\tilde{x}_{\lceil p\times(N+1)\rceil+1}}{2}$

Parmi les percentiles on utilise fréquemment:

- les quartiles Q_1, Q_2, Q_3 (respectivement 25%, 50%, 75%)
- les déciles $D_1, ..., D_9$ (respectivement 10%, 20%, ..., 90%).

Conclusion

- Comparer des séries statistiques uniquement par leurs moyennes et médianes n'est pas suffisant.
- Ajouter le mode permet d'améliorer cette comparaison.
- Calculer d'autres indices qui vont permettre de juger de distribution des valeurs autour de ces valeurs centrales.

Exemple

On a deux séries de 200 valeurs, les indices de tendance centrales sont donnés ci-dessous :

	Moyenne	Médiane	Mode
Série 1	10.065	10	15
Série 2	9.730	10	11

Représentation des deux séries a moyenne est répresentée en rouge, la médiane en vert et le mode en bleu.