Indices de dispersion

J.-M., Galharret

LMJL, Faculté de Psychologie, Nantes Université

Introduction

Définition : Une caractéristique de dispersion est un nombre qui permet d'avoir une estimation de l'écartement des valeurs les unes par rapport aux autres, ou bien par rapport à une valeur centrale.

On peut distinguer plusieurs types d'indices de dispersion :

- Différence entre deux valeurs d'un jeu de données.
- Différences entre toutes les valeurs d'un jeu de données et une valeur centrale.

Différence entre deux valeurs

Avantages/Désavantages :

Avantages :

- Ils sont très faciles à calculer !
- Ils sont très faciles à interpréter !

Désavantages :

- Ils sont très sensibles à la fluctuation d'échantillonage.
- Ils n'ont pas de bonnes propriétés de calculs.

Etendue

C'est l'indice le plus simple à calculer : il est égal à la valeur maximale moins la valeur minimale du jeu de données $(x_i)_{i=1,\ldots,N}$.

$$E = \max(x) - \min(x)$$

Remarque :

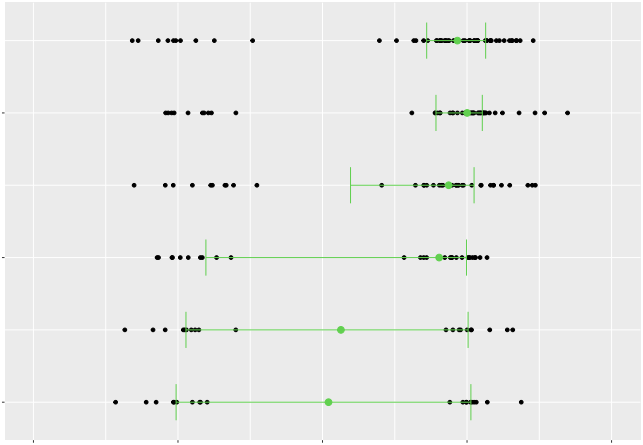
Cet indicateur n'a d'intéret que lorsque la taille de l'échantillon est suffisament grande.

Intervalles interquartiles/ déciles

Il existe plusieurs indices de dispersion basés sur les quartiles et les déciles :

- Interquartile $IQR = Q_3 - Q_1$
- Semi-interquartile $\frac{1}{2}(Q_3 - Q_1)$
- Interquartile relatif $\frac{Q_3 - Q_1}{Q_2}$
- Interdécile : $D_9 - D_1$
- Intercentile : $C_{99} - C_1$ (comme pour l'étendue cet indicateur ne sera utilisé qu'avec des échantillons de grande taille)

Exemple



Différence entre toutes les valeurs et un indice de tendance centrale

L'écart absolu

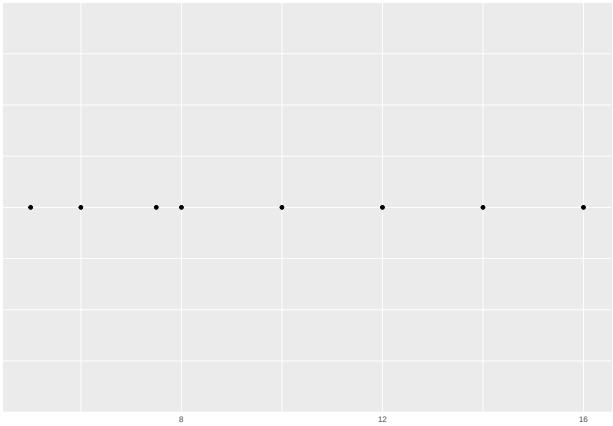
Définition L'écart absolu entre des valeurs $(x_i)_{i=1,\ldots,N}$ et une valeur A est défini par

$$\frac{1}{N}\sum_{i=1}^{N}|x_i - A|$$

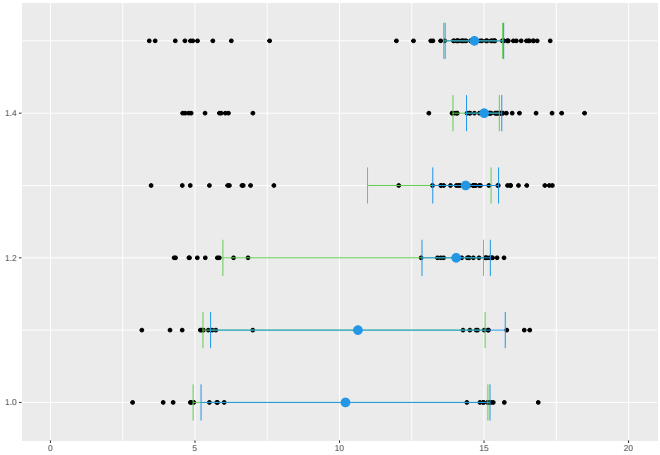
- En général on calcule cet écart absolu relativement à la moyenne de la série ou bien à sa médiane (MAD : Median Absolute Deviation).
- L'écart absolu relativement à la médiane est plus faible que celui à la moyenne.

Exemple :

X	écart
5.0	4.0
6.0	3.0
7.5	1.5
8.0	1.0
10.0	1.0
12.0	3.0
14.0	5.0
16.0	7.0



On représente toujours sur les mêmes données l'intervalle $[\tilde{x} - MAD, \tilde{x} + MAD]$



Désavantage :

- On ne peut pas estimer l'écart absolu des valeurs sur la population à partir de celui d'un échantillon.
- L'écart absolu n'a pas de bonnes propriétés de calculs.

La variance et l'écart type

Définition La variance des valeurs $(x_i)_{i=1,\ldots,N}$ est définie par

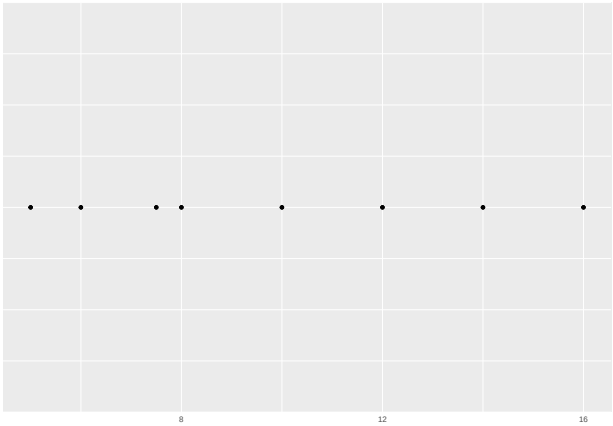
$$V(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

L'inconvenient de la variance est qu'il n'est pas sur la même échelle que les valeurs. Par exemple si on a des données en € alors la variance sera exprimée en $€^2$ ce qui n'est pas très pratique.

On définit donc l'écart type d'observation comme la racine carrée de la variance $\sigma(x) = \sqrt{V(x)}$.

Exemple :

X	écart
5.0	23.04
6.0	14.44
7.5	5.29
8.0	3.24
10.0	0.04
12.0	4.84
14.0	17.64
16.0	38.44



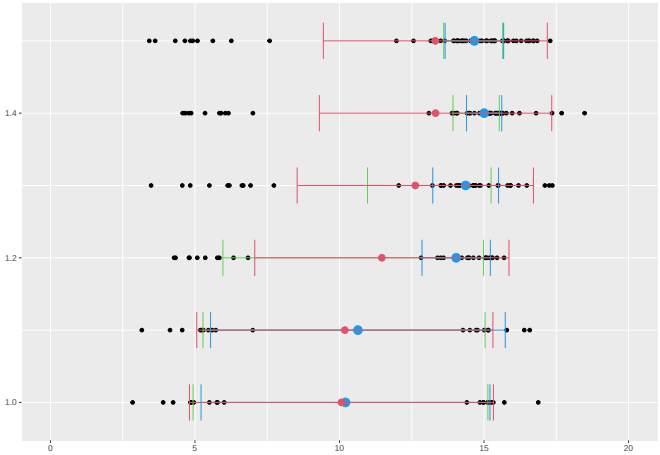
Avantage de l'écart type :

- On peut estimer l'écart type des valeurs sur la population à partir des observations :

$$s(x) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

- L'écart type a de bonnes propriétés de calculs.
- On a un résultat théorique sur la proportion de valeurs dans l'intervalle $[\bar{x} - k \times s(x), \bar{x} + k \times s(x)]$: celle-ci est supérieure à $1 - \frac{1}{k^2}$.

Dans cet exemple on a représenté des échantillons de valeurs pour lesquelles on a calculé la moyenne et l'intervalle $[\bar{x} - s, \bar{x} + s]$



Désavantage

Comme la moyenne l'écart type est sensible aux valeurs extrèmes.

Retour sur l'exemple des salaires :

Les salaires des 100 employés ont pour moyenne 3018 et pour écart type 302. Si on ajoute le salaire du patron alors on obtient une moyenne de 12889 et pour écart type 99204.

Identifier des valeurs extrèmes :

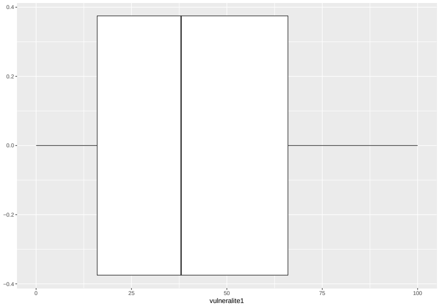
Boxplot

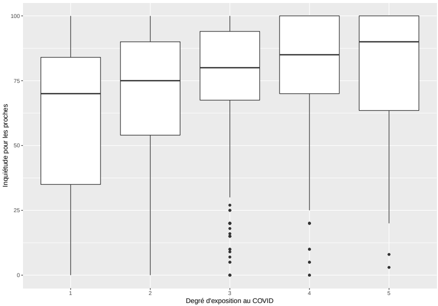
- Ces représentations graphiques ont été introduites par Tukey.
- On y représente $Q1, Q2, Q3$.
- Toute valeur x de l'échantillon telle que

$$x > Q3 + 1.5IQR \text{ ou } x < Q1 - 1.5IQR$$

est considérée comme une valeur aberrante (outlier).

Exemples


0.4
0.2
0.0
−0.2
−0.4
0
25
50
75
100
vulneralite1


Inquiétude pour les proches
100
75
50
25
0
1
2
3
4
5
Degré d'exposition au COVID

Autres critères d'identification

- ***Moyenne/ Ecart type*** Une valeur x est considérée comme aberrante lorsque
$$|x - \bar{x}| > 3s$$
- ***Médiane/ MAD*** Rappel : on a $MAD = \frac{1}{N}\sum_{i=1}^{N}|x_i - \tilde{x}|$. Une valeur x est considérée comme une valeur aberrante si
$$|x - \tilde{x}| > \frac{2MAD}{0.6745}$$