Éléments de compréhension des statistiques

Jeffery P.

Doctorant au Laboratoire des Sciences du Numérique de Nantes (LS2N)

2019

Boites de dispersion

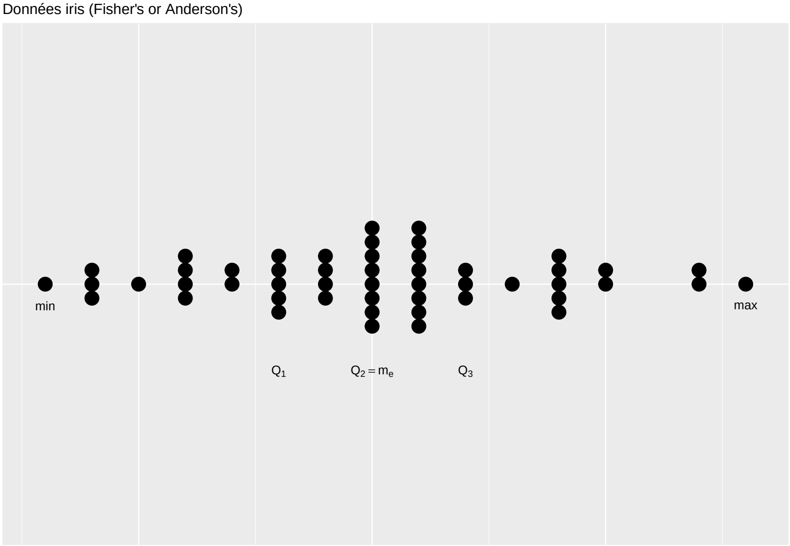
En statistique, il est souvent utile de visualiser les distributions.

On peut par exemple tracer des **boites de dispersion** (en anglais **boxplot**). Pour cela, on représente :

- Un rectangle délimité par les 1er et 3ème quartiles (Q_1 et Q_3)
- La médiane (Q_2) à l'intérieur de ce rectangle
- Un segment inférieur jusqu'au 1er décile (D_1), ou jusqu'au minimum
- Un segment supérieur jusqu'au 9ème décile (D_9), ou jusqu'au maximum

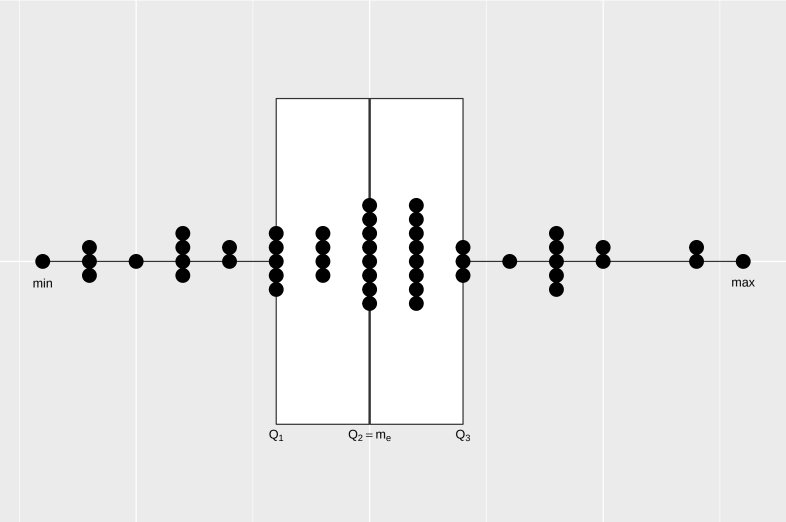
Boites de dispersion

Boites de dispersion



Boites de dispersion

Données iris (Fisher's or Anderson's)



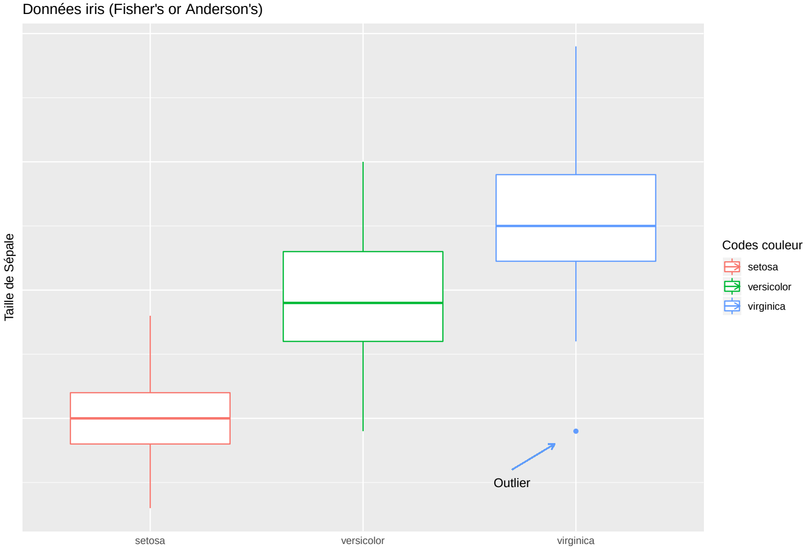
Boites de dispersion : quelques intérêts

Grâce aux boites de dispersion, on peut rapidement avoir une idée de la répartition des données

- Appréhender l'échelle
- Savoir si les données sont rassemblées autour de la médiane ou dispersées...

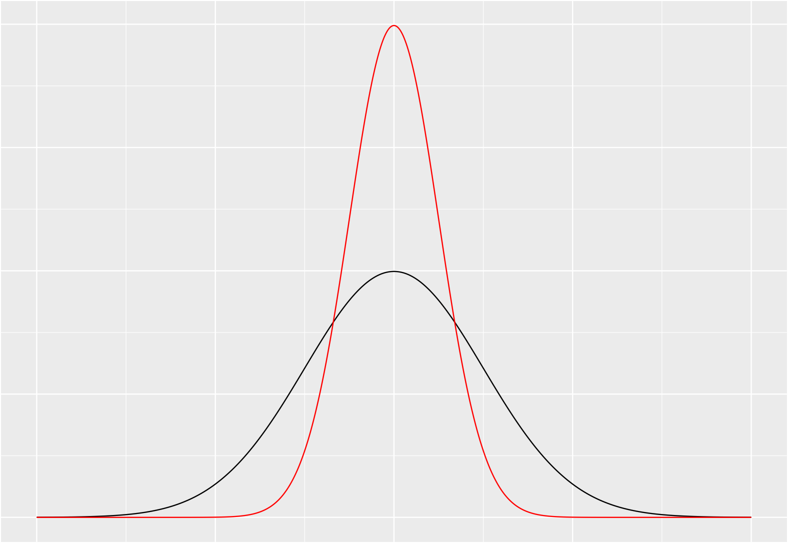
De plus, les boites de dispersion permettent de comparer plusieurs distributions entre elles !

Boites de dispersion



Mesure de dispersion

Que pensez vous des deux courbes suivantes ?

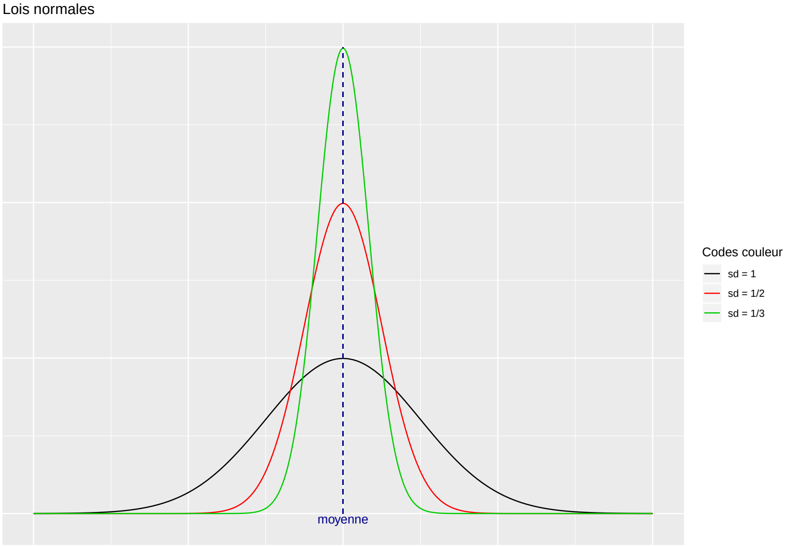


Mesure de dispersion

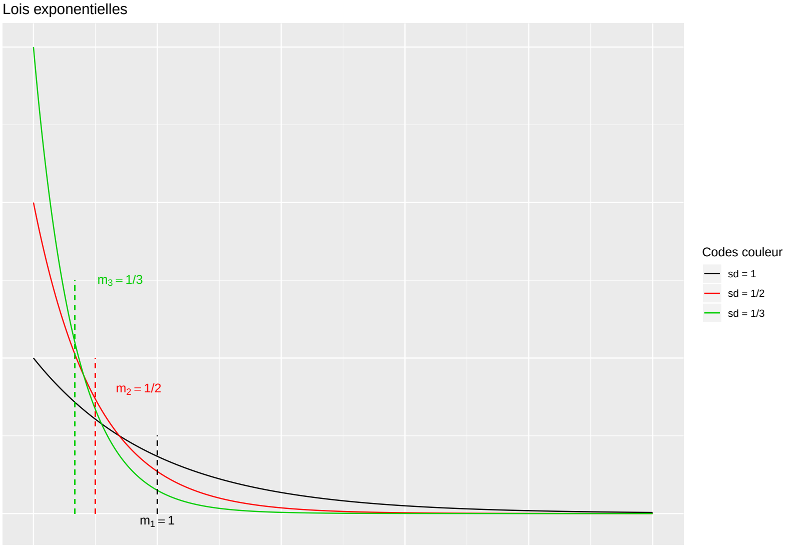
En statistique la dispersion se nomme **variance.** Il s'agit d'une valeur strictement positive qui ne se calcule que pour des variables **quantitatives**

- La dispersion traduit le fait que les données sont plus ou moins resserrées autour de la valeur moyenne $\bar{x}$
- Plus la variance est élevée, plus les données seront éloignées de leur moyenne (i.e., plus l'histogramme sera aplati)
- La variance est nulle si toutes les observations ont la même valeur

Mesure de dispersion : illustration



Mesure de dispersion : illustration



Le calcul de la variance d'observation

La variance d'**observation** se note souvent σ^2 ou var. Elle se calcule différemment en fonction des cas :

1. Sur données brutes :

$$\sigma^2 = \frac{1}{\mathrm{N}} \sum (x_i - \bar{x})^2$$

2. Sur données traitées avec effectif :

$$\sigma^2 = \frac{1}{\mathrm{N}} \sum n_i (x_i - \bar{x})^2$$

3. Sur données traitées regroupées en classes :

$$\sigma^2 = \frac{1}{\mathrm{N}} \sum n_i (c_i - \bar{x})^2$$

Éléments de compréhension de la variance

Exemple sur données brutes :

$$\sigma^2 = \frac{1}{\mathrm{N}} \sum (x_i - \bar{x})^2$$

Pour obtenir la variance :

1. On calcule les écarts entre chaque observation et la valeur moyenne
2. On élève tous ces écarts au carré (puissance 2)
3. On calcule la moyenne des valeurs observées

Éléments de compréhension de la variance

On peut montrer que :

$$\sigma^2 = \frac{1}{\mathrm{N}} \sum (x_i - \bar{x})^2 = \frac{1}{\mathrm{N}} \sum x_i^2 - \bar{x}^2$$

Ou encore :

$$\sigma^2 = \bar{x^2} - \bar{x}^2, \quad \text{où } \bar{x^2} = \frac{1}{\mathrm{N}} \sum x_i^2$$

Donc :

$$\sigma^2 = \text{moyenne des carrés} - \text{carré de la moyenne}$$

Pourquoi élever au carré ?

Si on n'élevait pas les écarts au carré, les valeurs pourraient se compenser. Par exemple avec les observations suivantes $2, 3, 3, 4$

- On a $\bar{x} := \frac{\sum x_i}{\mathrm{N}} = \frac{2+3+3+4}{4} = 12/4 = 3$

Sans élever les écarts au carré, on trouverait :

$$\sigma^2 = \frac{1}{\mathrm{N}} \sum (x_i - \bar{x}) = \frac{1}{4}\big((2-3) + (3-3) + (3-3) + (4-3)\big) = 0$$

. . .et pourtant les observations n'ont pas toutes la même valeur

Le calcul de la variance d'échantillon

En statistique inférentielle on corrige la variance par le facteur $\frac{\mathrm{N}}{\mathrm{N}-1}$. On obtient ainsi une variance dite d'**échantillon**

En pratique, on applique les mêmes formules mais on divise par $\mathrm{N}-1$ plutôt que par N

1. Sur données brutes :

$$\mathrm{s}^2 = \frac{1}{\mathrm{N}-1}\sum(x_i - \bar{x})^2$$

2. Sur données traitées avec effectif :

$$\mathrm{s}^2 = \frac{1}{\mathrm{N}-1}\sum n_i(x_i - \bar{x})^2$$

3. Sur données traitées regroupées en classes :

$$\mathrm{s}^2 = \frac{1}{\mathrm{N}-1}\sum n_i(c_i - \bar{x})^2$$

L'écart-type

- L'écart type est la racine carré de la variance que l'on note naturellement σ ou s (= $\sqrt{\sigma^2}$ ou $\sqrt{\mathrm{s}^2}$)
- Puisque qu'il existe une variance d'observation et d'échantillon, on calcule également un écart-type d'observation et un écart-type d'échantillon

Calcul de l'écart-type

1. Sur données brutes :

$$\sigma = \sqrt{\frac{1}{\mathrm{N}} \sum (x_i - \bar{x})^2} \quad \text{ou} \quad \mathrm{s} = \sqrt{\frac{1}{\mathrm{N}-1} \sum (x_i - \bar{x})^2}$$

2. Sur données traitées avec effectif :

$$\sigma = \sqrt{\frac{1}{\mathrm{N}} \sum n_i (x_i - \bar{x})^2} \quad \text{ou} \quad \mathrm{s} = \sqrt{\frac{1}{\mathrm{N}-1} \sum n_i (x_i - \bar{x})^2}$$

3. Sur données traitées regroupées en classes :

$$\sigma = \sqrt{\frac{1}{\mathrm{N}} \sum n_i (c_i - \bar{x})^2} \quad \text{ou} \quad \mathrm{s} = \sqrt{\frac{1}{\mathrm{N}-1} \sum n_i (c_i - \bar{x})^2}$$

Calcul de l'écart-type : remarques

- Comme la variance, l'écart-type est strictement positif (attention aux calculs !)
- Puisqu'il se déduit de la variance (par fonction croissante), l'écart-type est aussi une mesure de dispersion qui s'interprète de la même façon
- L'écart-type est une mesure homogène aux données : si les observations sont des poids en kg, l'écart-type est aussi en kg
- Étant données plusieurs distributions, on peut comparer les dispersions avec la valeur des écart-types… mais il faut prendre quelques précautions

Le coefficient de variation

L'écart type est sensible à l'échelle des données. On utilise alors le coefficient de variation (noté cv) pour le mettre à l'échelle. Il est définit par le pourcentage suivant :

$$\mathrm{cv} = \frac{\mathrm{s}}{\bar{x}} \times 100$$

- Le cv est réservé aux distributions strictement positives
- Le cv est une mesure d'**homogénéité** des données, plus il est faible, plus les données sont homogènes
- En pratique si $\mathrm{cv} < 15\%$, on dit que les données sont homogènes
- Le cv est très utile pour comparer des distributions avec des unités différentes et/ou avec des moyennes très éloignées