Part I

Statistiques univariées

démarche scientifique qui a pour but d'observer une $\underbrace{\text{grandeur / une qualité}}_{\text{Caractère}}$ sur une population

Relevés: $x_1 \dots x_N$ // variable statistique.

Salaire des Français

caractère ↗ ↖ population.

$\underbrace{\text{le nombre de voitures}}_{\text{caractère}}$ sur des $\underbrace{\text{tranches d'une heure}}_{\text{population statistique}}$ (à un péage)

"On observe le caractère sur la population statistique"

1 Statistiques descriptives

La statistique est une méthode scientifique qui consiste à organiser, analyser et interpréter des observations faites sur un ou plusieurs caractères (*poids*, *consommation*, $\cdots$) des individus (*chèvres*, *machines*, *pièces manufacturées*, $\cdots$) d'une population de taille N.

qualitatives
- couleur
- absentéisme
- type de cheveux.

quantitatives
- mesure en cm
- temps en sec.
- salaire en €

discrètes
valeurs indexables par $\mathbb{N}$

continues
toutes les valeurs d'un intervalle

1, 2, 3, 3, 4, 4, 4, 5, 5, 6

1.3 Variable discrète quantitative

Prenons l'exemple suivant d'une série de notes d'un examen noté sur dix points.

discrète

trier

x_i	1	2	3	4	5	6
n_i	1	1	2	3	2	1
f_i	0.1	0.1	0.2	0.3	0.2	0.1

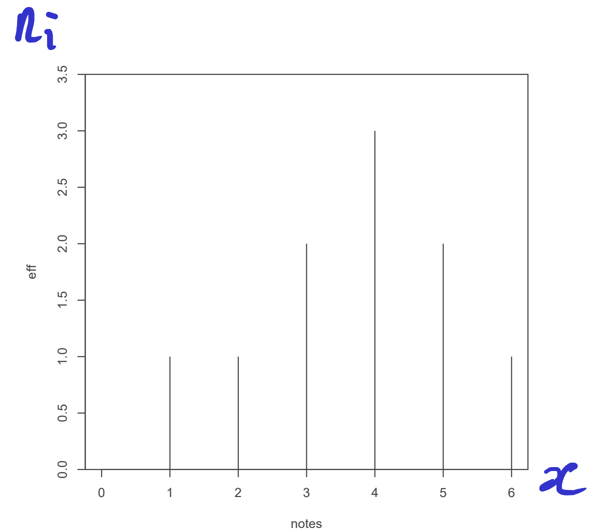
= nombre fini des valeurs du caractère

= modalités

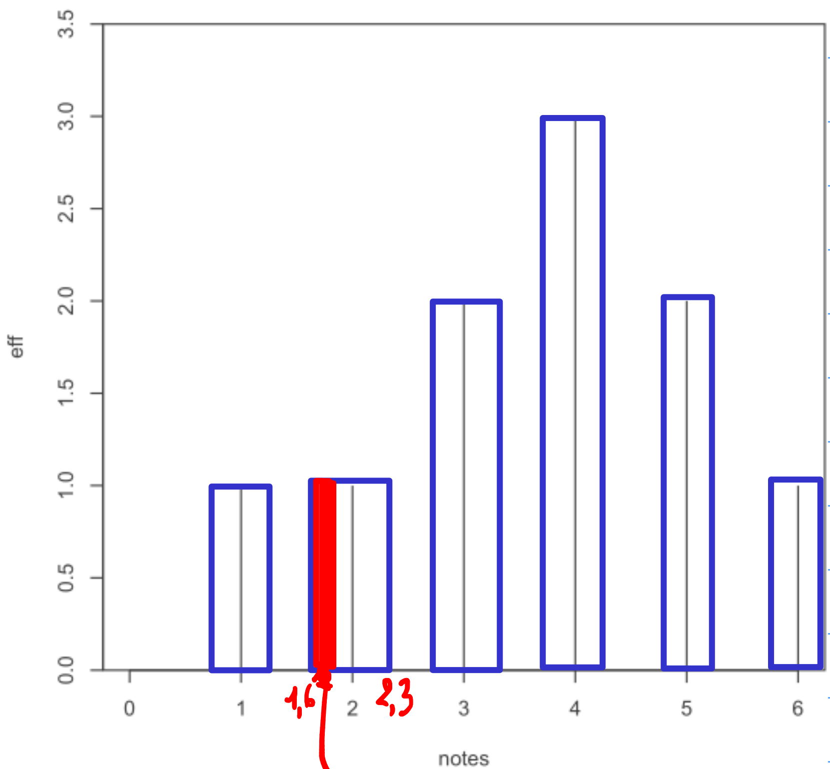
$\frac{n_i}{N}$ $N = 10$

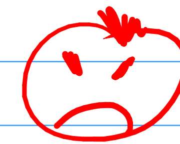
Il est à noter qu'il est totalement équivalent de parler en terme d'effectif n_i ou en terme de fréquences f_i.
On représentera graphiquement la série précédente par un diagramme en bâtons :

ni en fonction de xi



Note: Excel...





Aire = effectif ! ⚠

Il y a des notes entre 1,6 et 1,9 ?

On pourra également tracer la fonction de répartition. Pour cela on ajoute la ligne des effectifs cumulés (*eff. c.*) et celle des fréquences cumulées (*f.c.*) au tableau précédent.

x_i	1	2	3	4	5	6
n_i	1	1	2	3	2	1
f_i	0.1	0.1	0.2	0.3	0.2	0.1

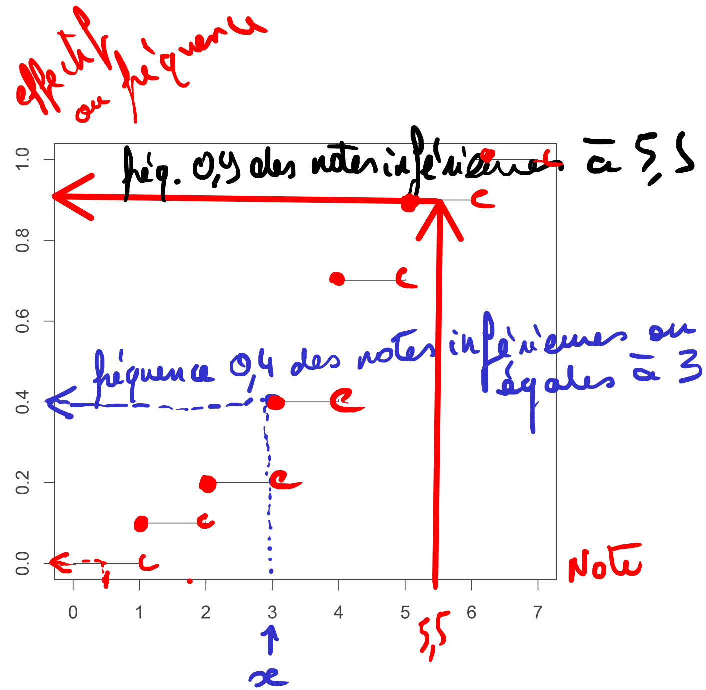
On cumule

nombre de notes inférieures ou égales à 3

x_i	1	2	3	4	5	6
eff.c.	1	2	4	7	9	10
f.c.	0.1	0.2	0.4	0.7	0.9	1

toutes les notes.

Ces lignes permettent de répondre facilement à la question *combien d'élèves ont eu une note inférieure ou égale à 4.* Le graphe des fréquences cumulées en fonction des valeurs de X est appelé fonction de répartition.



classes modales = modalités

1.4 Variable continue quantitative

On a mesuré le poids de $N = 20$ personnes. Ici le caractère x, le poids, est une variable continue. On regroupe les mesures en classe et on obtient le tableau suivant :

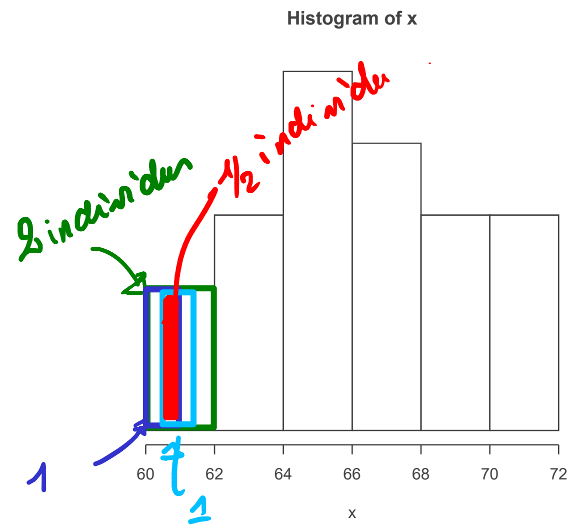
x_i	[60,62[	[62,64[	[64,66[	[66,68[	[68,70[	[70,72[
n_i	2	3	5	4	3	3

4 relevés entre 66 et 68 kg.

Note : Après regroupement en classe, on perd les mesures exactes. On suppose l'équi-répartition dans les classes :

- 2 individus dans [60,62[→ 1 individu dans [61;62[
  → 1 individu dans [60;61[→ 1 individu dans [60,5;61,5[

Il est également possible de s'exprimer en terme de fréquences. On représente ces données par un histogramme :



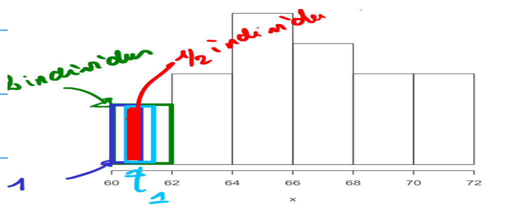
Remarque 1.1 *Sur l'histogramme, l'effectif se lit en terme d'aire des rectangles. Dans cet exemple (et c'est souvent le cas), les largeurs des classes sont les mêmes et donc les hauteurs des rectangles sont proportionnelles aux effectifs.*

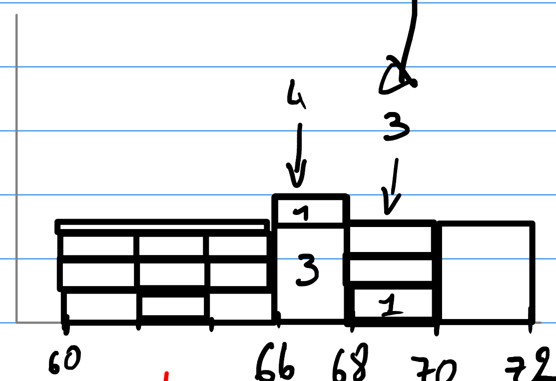
l'effectif ou la fréquence est donnée par l'aire du rectangle.

⚠ Si les classes ne sont pas de même taille !

x_i	[60,62[	[62,64[	[64,66[	[66,68[	[68,70[	[70,72[
n_i	2	3	5	4	3	3

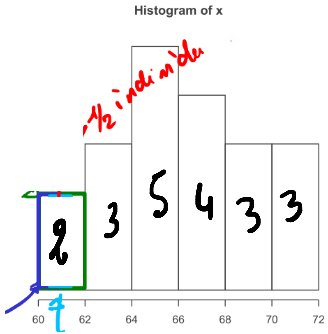
x_i	[60;66[	[66,68[	[68,70[	[70,72[
n_i	10	4	3	3



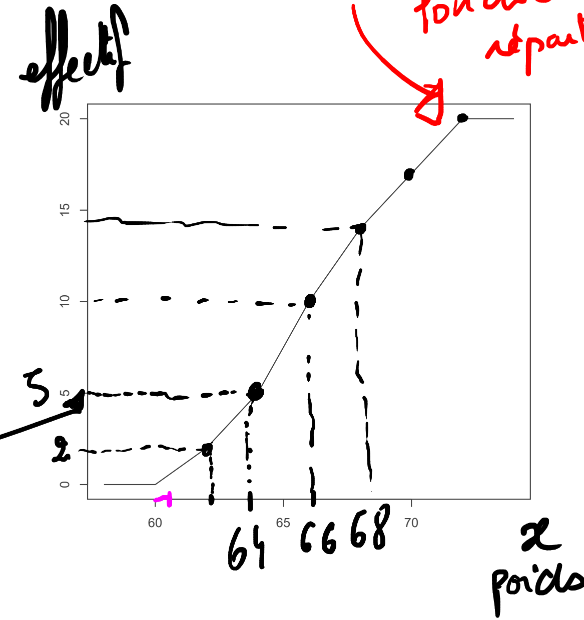


Le changement de classes modifie peu l'aspect.

équirépartition = effectif proportionnel à la longueur d'intervalle
⇒ segments dans la fonction de répartition



Dans cette organisation en classe, on suppose qu'au sein d'une classe l'effectif est equiréparti, i.e. il y aura autant de personnes ayant un poids compris entre 60 et 61 kg qu'entre 61 et 62 kg. On a donc la fonction de répartition suivante :



effectif ayant poids inférieur ou égal à 64.

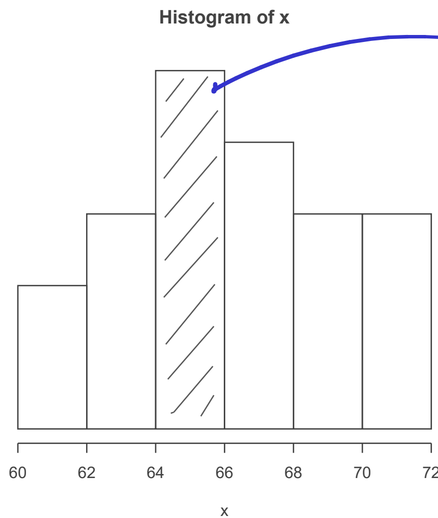
1.6 Caractéristiques de position

1.6.1 Le mode

Le mode, ou classe modale, d'une série statistique correspond intuitivement à la valeur du caractère (modalité.s) qui regroupe le plus grand effectif.

- Pour une variable discrète, c'est la modalité avec le plus grand effectif. Dans l'exemple 1.3, le mode est $x_4 = 4$ avec un effectif de 3.

- Pour une variable continue, c'est la classe $[a_i, a_{i+1}[$ dont la barre sur l'histogramme est la plus haute.



Aire = effectif

$$\text{Aire} = l([a_i, a_{i+1}[) \times h = (a_{i+1} - a_i) \times h = n_i$$

$$\Leftrightarrow h = \frac{n_i}{a_{i+1} - a_i}.$$

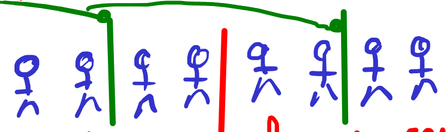
Autrement dit, la classe modale est la classe $[a_i; a_{i+1}[$ qui maximise la quantité

$$\frac{n_i}{l([a_i; a_{i+1}[)},$$

où l est la longueur de l'intervalle. Ici, la classe modale est $[64; 66[$.

(population ordonnée)

1er et 3è quartile



médiane = valeur du caractère qui partage la pop. en 2.

Variables quantitatives

1.6.2 Médiane, quartiles, quantiles

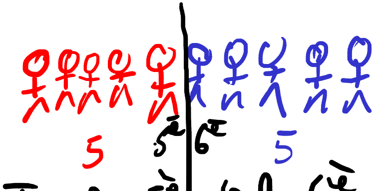
La médiane est une valeur du caractère qui sépare le classement en deux groupes de même taille. Le premier quartile (resp. le troisième) partage le classement dans les proportions 0.25 et 0.75 (resp. 0.75 et 0.25). On peut voir les quartiles comme les médianes des demi-groupes séparés par la médiane.

Plus généralement, le quantile d'ordre α, noté q_α est une valeur du caractère qui sépare le classement dans les proportions α et $(1-\alpha)$.

5ème 6ème

- les notes

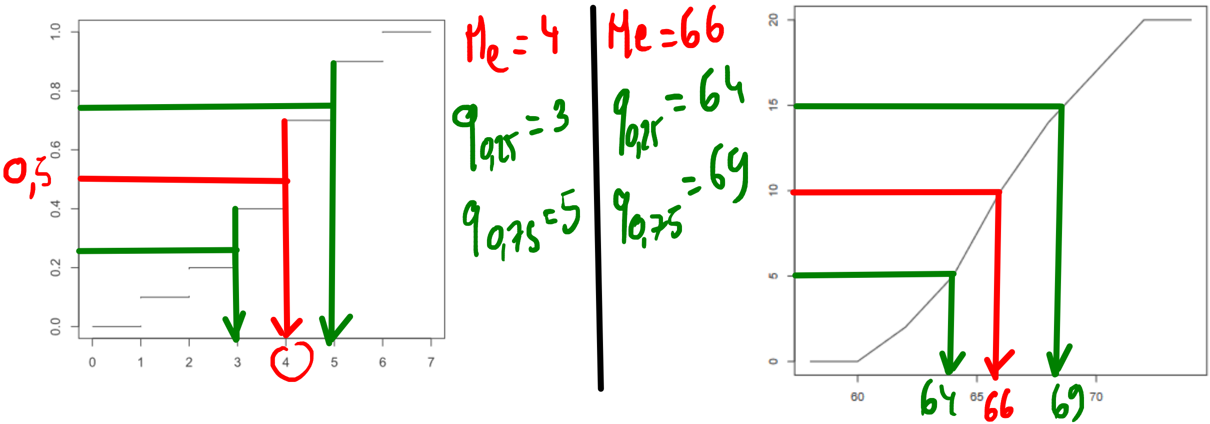
x_i	1	2	3	4	5	6
n_i	1	1	2	3	2	1



Pour la médiane : on prend une valeur entre le 5è et le 6è :

$$\frac{x_5+x_6}{2} = \frac{4+4}{2} = 4$$

Proposition 1.2 *La médiane est un antécédant de 0.5 par la fonction de répartition. Le quantile d'ordre α est un antécédant par la fonction de répartition de α.*



x_i	1	2	3	4	5	6
n_i	1	1	2	3	2	1

x_i	[60,62[	[62,64[	[64,66[	[66,68[	[68,70[	[70,72[
n_i	2	3	5	4	3	3

$$\bar{x} = \frac{1\times1+2\times1+3\times2+4\times3+5\times2+6\times1}{10}$$

$$\bar{x} = \frac{61\times2+63\times3+65\times5+67\times4+69\times3+71\times3}{20}$$

1.6.3 Moyenne

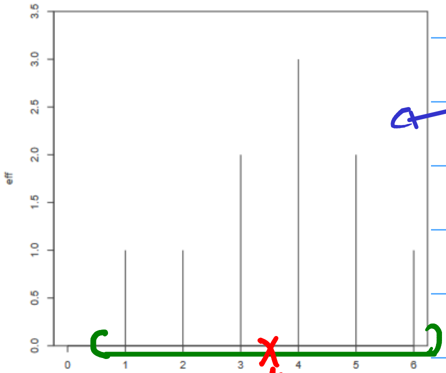
La moyenne est la valeur du caractère qu'auraient tous les individus de la population s'ils étaient identiques. Il s'agit du point d'équilibre (centre de gravité) du diagramme en batons ou de l'histogramme. En pratique, il s'agit de la moyenne des modalités pondérée par l'effectif :

$$m(x) = \bar{x} = \sum_i x_i \frac{n_i}{N}.$$

Dans l'exemple 1.3, la moyenne vaut 3.7.

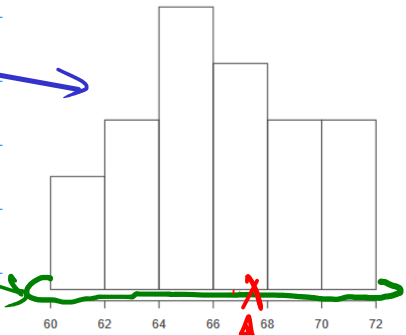
Remarque 1.3 *Dans le cas d'une série continue regroupée en classes, il convient de remplacer les valeurs x_i du caractère par c_i, les valeurs du centre de la classe.*

La moyenne est le centre de gravité des représentations graphiques :



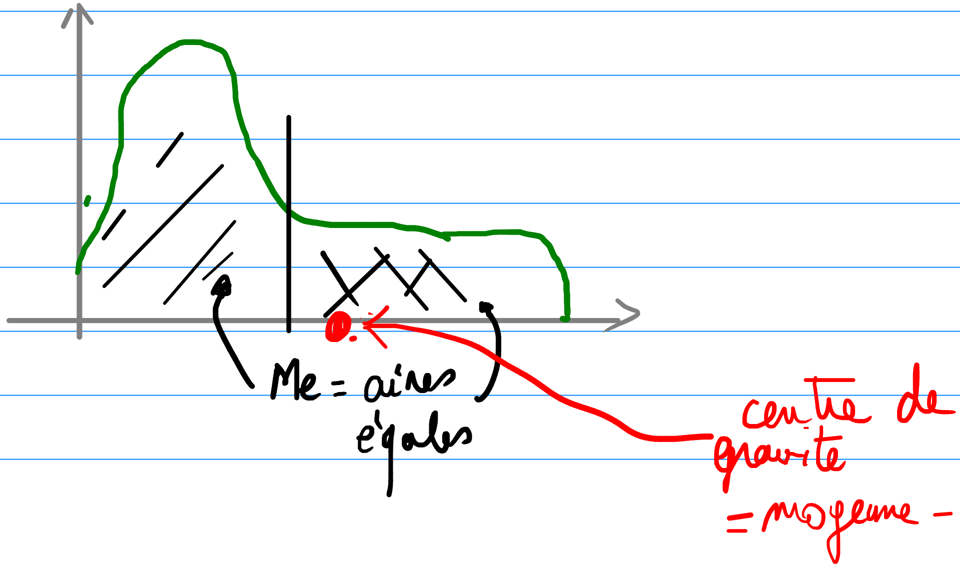
verres

plateau



centre de gravité (graphique) = bonne approximation de la moyenne

Les séries statistique ne sont pas toujours symétriques



Proposition 1.4 (Linéarité de la moyenne) *Soit a, b, c des réels, x, y des séries statistiques. On a*

$$m(ax + by + c) = am(x) + bm(y) + c.$$

1.7 Caractéristiques de dispersion

Il arrive que deux séries aient les mêmes caractéristiques de position mais soient totalement différentes. En effet, ces derniers indicateurs (hors quantiles) ne s'intéressent pas aux différences au sein de la population et donnent le stéréotype d'une population où tous les individus sont identiques. Pour aller plus loin dans l'étude statistique il nous faut donc introduire les notions suivantes concernant la dispersion des données.

1.7.1 Étendue

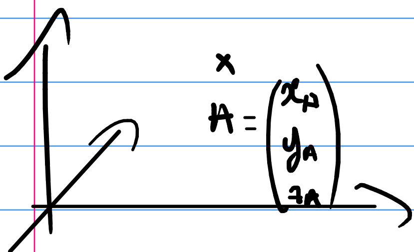
Il s'agit de l'écart entre la plus grande et la plus petite valeur prise par la série : $max\{x_i\} - min\{x_i\}$.

maximum – minimum.

Note : la distance euclidienne

$$d(A,B) = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2 + (z_B - z_A)^2}$$

$$B = \begin{pmatrix} x_B \\ y_B \\ z_B \end{pmatrix}$$



$$A = \begin{pmatrix} x_A \\ y_A \\ z_A \end{pmatrix}$$

1.7.2 Variance et écart-type

La variance et l'écart–type sont des mesures des écarts à la moyenne : $(x_i - \bar{x})$. La variance pondère ces écarts au carré :

$$Var(x) = \sum_i (x_i - \bar{x})^2 \frac{n_i}{N}.$$

La variance est donc très influencée par les valeurs éloignées de la moyenne.

$$\sigma_x = \sqrt{Var(x)} = \sqrt{\sum_i (x_i - \bar{x})^2 \frac{n_i}{n}}$$

coefficienté par l'effectif.

Somme des (écarts entre observation et moyenne) au carré

L'écart-type est la distance moyenne entre les observations x_i et $\bar{x}$.

Proposition 1.5 *En pratique, on pourra calculer*

$$Var(x) = m(x^2) - (\bar{x})^2,$$

où

$$m(x^2) = \sum_i x_i^2 \frac{n_i}{N}.$$

Le résultat se vérifie en développant et en factorisant dans la formule initiale de la variance.

Preuve :

$$Var(x) = \sum_i (x_i - \bar{x})^2 \frac{n_i}{N} = (x_1 - \bar{x})^2 \frac{n_1}{N} + (x_2 - \bar{x})^2 \frac{n_2}{N} + \dots + (x_N - \bar{x})^2 \frac{n_N}{N}$$

$$= \frac{1}{N} \sum_i (x_i - \bar{x})^2 n_i = \frac{1}{N} \sum_i \left(x_i^2 - 2\bar{x}x_i + \bar{x}^2\right) n_i$$

$$= \frac{1}{N} \sum_i \left(x_i^2 n_i - 2\bar{x} x_i n_i + \bar{x}^2 n_i \right)$$

$$= \frac{1}{N} \left(\sum_i x_i^2 n_i + \sum_i -2\bar{x} x_i n_i + \sum_i \bar{x}^2 n_i \right)$$

constant donc factorisable

$$= \underbrace{\frac{1}{N} \sum_i x_i^2 n_i}_{\text{moyennes des valeurs de } x_i \text{ au carré}} + \frac{1}{N} \sum_i -2\bar{x}\, x_i n_i + \frac{1}{N} \sum_i \bar{x}^2 n_i$$

$$= m(x^2) + \frac{1}{N} \times (-2\bar{x}) \sum_i x_i n_i + \frac{1}{N} \bar{x}^2 \underbrace{\sum_i n_i}_{N}$$

$\bar{x}$

somme des effectifs = effectif total

$$= m(x^2) - 2\bar{x}\,\bar{x} + \frac{N}{N}\bar{x}^2$$

$$= m(x^2) - 2\bar{x}^2 + \bar{x}^2 = m(x^2) - \bar{x}^2$$

$\blacksquare$

On définit l'écart-type σ par la racine carrée de la variance : $\sigma = \sqrt{Var(X)}$. Ainsi, l'écart-type renvoie à la distance pondérée entre les valeurs de la série et la moyenne. Il en découle la propriété suivante.

Proposition 1.6 *Soit a, b des réels. On a*

$$\sigma_{ax+b} = |a|\sigma_x, \qquad Var(ax+b) = a^2 Var(x).$$

Rappel

Exemple notes :

le 5ᵉ a eu 4

le 6ᵉ a eu 4

x_i	1	2	3	4	5	6
n_i	1	1	2	3	2	1

$N = 10$.

Me : la note qui sépare en 2 le groupe :
↳ 2 groupes de 5

5ᵉ étu. 6ᵉ étu

$$Me = \frac{4+4}{2} = 4$$

$q_{0,25}$ est la note qui sépare le groupe en 1/4 - 3/4.
⇒ entre le 2ᵉ et le 3ᵉ

2ᵉ 3ᵉ

$$q_{0,25} = \frac{2+3}{2} = 2,5$$

Le 3è quartile est la note qui sépare la population en 3/4 - 1/4. Sur 10, c'est "l'individu" 7,5.

$$q_{0,75} = \frac{\overset{7^{è}}{5} + \overset{8^{è}}{5}}{2} = 5.$$

min : 1 $\quad q_{0,25} = 2,5 \quad Me = 4 \quad q_{0,75} = 5 \quad$ max = 6.

<u>Pb</u> : comparer à une formation de 40 personnes ?

boîte de distribution / box plot

(i) Tracer un rectangle qui s'étend du premier quartile au troisième.

(ii) Séparer ce rectangle en deux à la hauteur de la médiane. On obtient alors une boîte.

(iii) On complète ce rectangles par deux segments. Pour cela, on calcule

$$a = q_{0.25} - 1.5IQ \quad \text{et} \quad b = q_{0.75} + 1.5IQ,$$

avec la distance inter-quartile

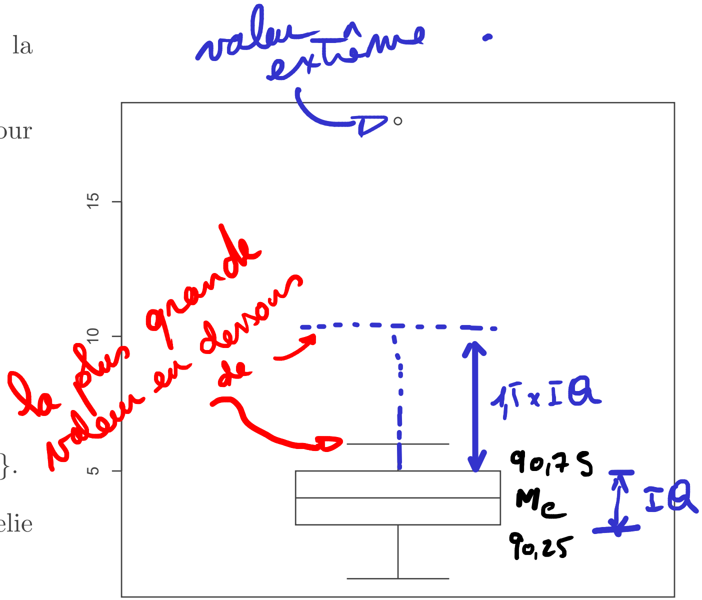
$$IQ = q_{0.75} - q_{0.25}.$$

On repère les valeurs :

$$x_a = \min\{x_i : x_i \geq a\} \quad \text{et} \quad x_b = \max\{x_i : x_i \leq b\}.$$

Ces valeurs sont appelées *valeurs adjacentes*. On relie ces valeurs aux cotés de la boîte.

(iv) Les valeurs qui ne sont pas comprises entre les valeurs adjacentes sont représentées par des points et sont appelées *valeurs extrêmes*.



min : 1 $\quad q_{0,25} = 2,5 \quad Me = 4 \quad q_{0,75} = 5 \quad max = 6.$

$$IQ = q_{0,75} - q_{0,25} = 5 - 2,5 = 2,5$$

$$1,5 \times IQ = 1,5 \times 2,5 = 3,75$$

2 2 2 3 . . .

$q_{0,25} = 3$

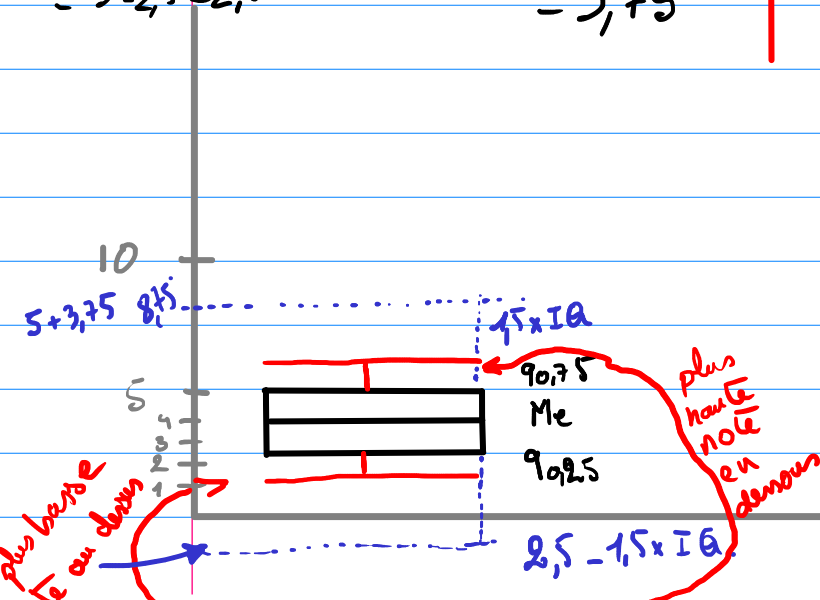
$Me = 5,5$

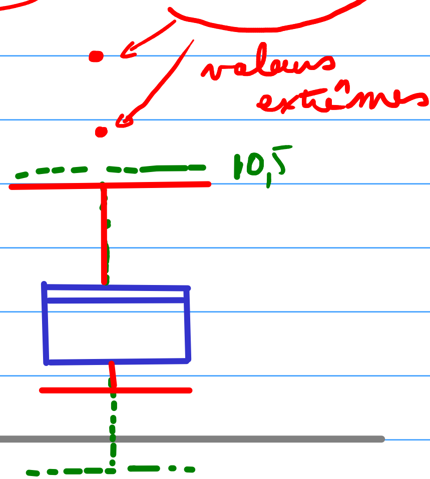
$q_{0,75} = 6$

$IQ : 6 - 3 = 3$

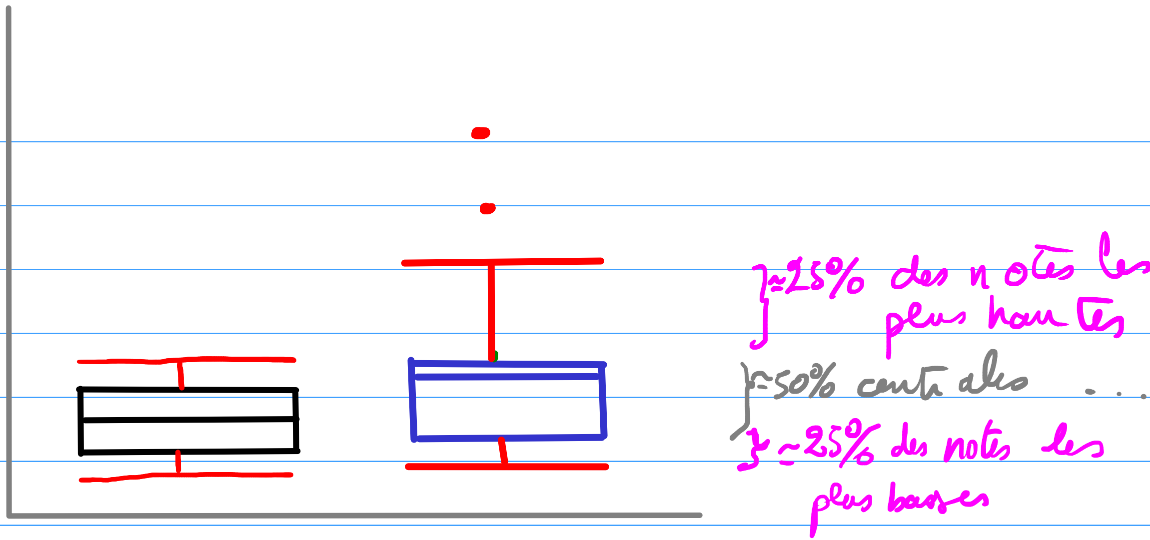
$1,5\, IQ = 4,5$

10 - 12 - 15





x_i	1	2	3	4	5	6
n_i	1	1	2	3	2	1



Tous les indicateurs du groupe à droite sont plus hauts que ceux du groupe à gauche.

=> le 2e groupe a globalement mieux réussi.