Chapitre :
Statistiques

∽

⊛ **Activité :** p240 (rappels de moyenne, médiane et quartiles)

I. Médiane et écart interquartile

⊛ **Activité :** 2p242

Rappel Dans une série statistique, la **médiane** Me est le plus petit nombre de la série tel qu'au moins 50% des données soient inférieurs ou égales à ce nombre

Définition On considère une série statistique dont on a déterminé :

- Le **premier quartile** Q_1, plus petit nombre de la série tel qu'au moins 25% des données soient inférieures ou égales à ce nombre ;
- Le **troisième quartile** Q_3, plus petit nombre de la série tel qu'au moins 75% des données soient inférieures ou égales à ce nombre ;

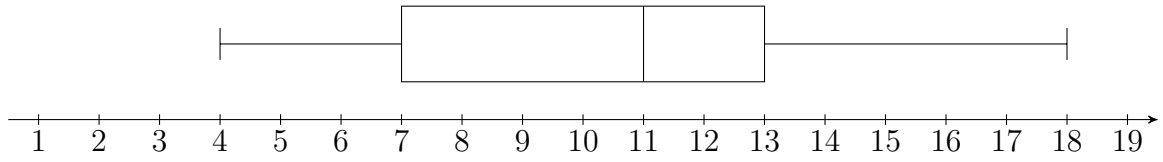
On définit alors l'**intervalle interquartile** comme étant l'intervalle $[Q_1; Q_3]$.
L'**écart interquartile** est la différence $Q_3 - Q_1$.

Remarque L'intervalle interquartile contient environ 50% des valeurs de la série.

Remarque La médiane Me peut être vue comme un deuxième quartile (Q_2).

► **Exercice :** 13,14p249 (médiane et écart interquartile)

Définition Un **diagramme en boîte** (ou boîte à moustache) est un diagramme regroupant les cinq valeurs minimum (min), Q_1, Me, Q_3 et maximum (max) au dessus d'une droite graduée.
La forme d'un diagramme en boîte est la suivante :



Ici, on peut lire que : $min = 4$, $Q_1 = 7$, Me $= 11$, $Q_3 = 13$ et $max = 18$.

► **Exercices :** 25,28p250, 39p252 (construction)

► **Exercices :** 32,34p251,38p252 (comparaisons)

II. Moyenne et écart-type

⊛ **Activité :** 4p243 (utilisation de la calculatrice – voir pages 353 ou 357 – calcul de l'écart-type).
On considère une série statistique, résumé par le tableau suivant :

Valeurs	x_1	x_2	...	x_p	Total
Effectifs	n_1	n_2	...	n_p	N

On rappelle que la **moyenne** de cette série est par définition :

$$\overline{x} = \frac{n_1x_1 + n_2x_2 + \cdots + n_px_p}{N} = \frac{1}{N}\sum_{i=1}^{p} n_ix_i$$

Ou avec les fréquences $f_i = \dfrac{n_i}{f_i}$,

$$\overline{x} = f_1x_1 + f_2x_2 + \cdots + f_px_p = \sum_{i=1}^{p} f_ix_i$$

Définition La **variance** de cette série est le réel V défini par :

$$V = \frac{n_1(x_1 - \overline{x})^2 + n_2(x_2 - \overline{x})^2 + \cdots + n_p(x_p - \overline{x})^2}{N} = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \overline{x})^2$$

L'écart-type σ est alors défini par :

$$\sigma = \sqrt{V}$$

Avec les fréquences $f_i = \dfrac{n_i}{N}$, on a :

$$V = f_1(x_1 - \overline{x})^2 + f_2(x_2 - \overline{x})^2 + \cdots + f_p(x_p - \overline{x})^2$$

⚠ La variance (tout comme l'écart-type), est une valeur **positive**. Il ne faut pas confondre les effectifs n_i (qui sont toujours positifs), avec les valeurs x_i (qui peuvent être négatives et sont dans les carrés dans la formule).

Propriété Il existe une autre formule de la variance permettant une erreur plus faible d'arrondi :

$$V = \frac{n_1x_1^2 + n_2x_2^2 + \cdots + n_px_p^2}{N} - \overline{x}^2 = \left(\frac{1}{N}\sum_{i=1}^{p} n_ix_i^2\right) - \overline{x}^2$$

(C'est la moyenne des carrés moins le carré de la moyenne)

L'écart-type permet de mesurer la dispersion des valeurs autour de la moyenne. Plus le nombre est grand, plus la dispersion est grande. Dans le cas d'une série où toutes les valeurs sont les mêmes, l'écart-type vaut 0.

▶ **Exercices :** 43,44p253 (voir les pages 353 ou 357 selon le modèle de calculatrice)

▶ **Exercices :** 52 (modifications), 53,55p254 (comparaisons)

▶ **Exercice :** 49p254 (logique)

III. Intervalles de fluctuation

⊛ **Activité :** 5p295 (obtention d'un intervalle a seuil de confiance de 95% à l'aide de la loi binomiale)

Supposons que dans une population donnée, une proportion p a un trait particulier (les yeux bleus, un salaire supérieur à 3 000 euros, ...).

On considère un échantillon de taille n de cette population. Dans cet échantillon, une proportion f possède le trait. On souhaite déterminer si l'échantillon correspond à la population globale, ou s'il est particulier. Autrement dit, si f est « assez proche » de p selon un critère à déterminer.

Propriété **(Rappel de seconde)** Si p est la proportion d'un caractère dans une population (avec $0{,}2 \leqslant p \leqslant 0{,}8$), alors pour un échantillon de taille n avec $n \geqslant 25$, la fréquence f du caractère dans l'échantillon appartient à l'intervalle $\left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$ avec une probabilité d'au moins 0,95.
Cet intervalle est appelé intervalle de fluctuation au seuil de confiance de 95% (ou de risque de 5%).

On souhaite une propriété qui ne soit pas limitée sur les valeurs de p et de n. Pour cela nous allons utiliser la loi binomiale.

Soit X une variable aléatoire égale au nombre d'individus d'un échantillon de n personnes qui ont la probabilité p d'avoir le trait étudié. Alors $X \sim \mathscr{B}(n; p)$.
On détermine deux entiers a et b de la manière suivante :

- a est le plus petit entier tel que $P(X \leqslant a) \geqslant 0{,}025$
- b est le plus petit entier tel que $P(X \geqslant b) \leqslant 0{,}025$ (autrement dit que $P(X \leqslant b) \geqslant 0{,}975$)

On a alors $P(a \leqslant X \leqslant b) \geqslant 0{,}95$.
Par suite on observe que $P(a \leqslant X \leqslant b) = P\left(\frac{a}{n} \leqslant \frac{X}{n} \leqslant \frac{b}{n}\right)$, et $\frac{X}{n}$ est la variable aléatoire égale à la fréquence des individus ayant le trait étudié.
Ainsi, sur un grand nombre d'expériences sur des échantillons de taille n, 95% des échantillons environ auront une fréquence observée comprise entre $\frac{a}{n}$ et $\frac{b}{n}$.
Ainsi :

Propriété L'intervalle de fluctuation au seuil de confiance de 95% de la fréquence est

$$\left[\frac{a}{n}; \frac{b}{n}\right]$$

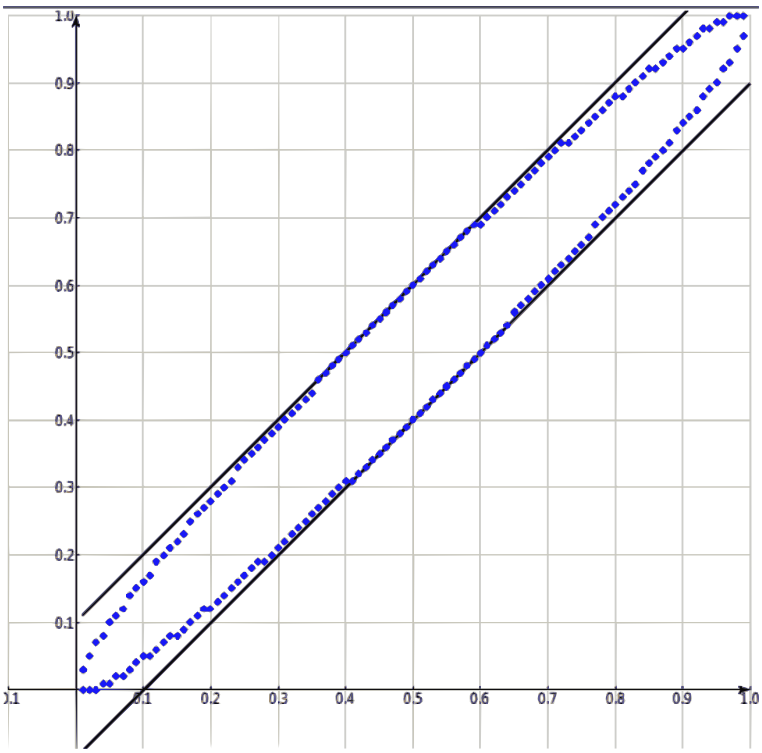
Pour revenir à la fréquence f observée, on adopte donc la règle de décision suivante :

Méthode On fait l'hypothèse que l'échantillon suit la même loi que la population totale, autrement dit que la proportion du trait étudié est p. Soit I l'intervalle de fluctuation de la fréquence à 95% dans des échantillons de taille n pour une proportion p.

- Si $f \notin I$, alors on rejette l'hypothèse au seuil de risque de 5%.
- Sinon, ($f \in I$), on ne rejette pas l'hypothèse au seuil 5% (on dit parfois que l'on valide l'hypothèse).

On peut décider d'autres valeurs de seuils. Par exemple pour 99%, on cherche le plus petit a tel que $P(X \leqslant a) \geqslant 0{,}005$ et le plus petit b tel que $P(X \leqslant b) \geqslant 0{,}995$, pour que $P((a \leqslant X \leqslant b) \geqslant 0{,}99$.

Voici finalement une illustration de la taille des intervalles de fluctuation en fonction de la valeur de p (en abscisse) pour des échantillons de taille 100 selon la méthode de seconde (entre les deux droites puisque la taille de l'intervalle sera toujours constant égal à $\dfrac{2}{\sqrt{100}} = 0{,}2$) et la méthode utilisant la loi binomiale (entre deux points).



On observe qu'effectivement, plus p est petit ou au contraire grand, moins l'intervalle donné en seconde est précis par rapport à celui de la méthode de première. Ce dernier réduit ; en effet, si par exemple extrême $p = 0$, tous les échantillons devraient n'avoir qu'une fréquence nulle, il ne devrait pas y avoir de fluctuation.

► **Exercices :** 68p311, 72,73p312

► **Exercices :** 81,83,84p313 (prise de décision)