Chapitre : Statistiques

∽

⊛ **Activité :** Rappels et tests page 126

I. Diagrammes en boîte

Rappel On considère une série statistique.

- Le **premier quartile**, noté Q_1, est la plus petite valeur de la série telle qu'au moins 25% des valeurs lui soient inférieures ou égales.
- Le **troisième quartile**, noté Q_3, est la plus petite valeur de la série telle qu'au moins 75% des valeurs lui soient inférieures ou égales.
- La **médiane**, notée Me, est la plus petite valeur de la série telle qu'au moins 50% des valeurs lui soient inférieures ou égales.
- On appelle intervalle interquartile l'intervalle $[Q_1 ; Q_3]$.
- On appelle écart interquartile le nombre $Q_3 - Q_1$.

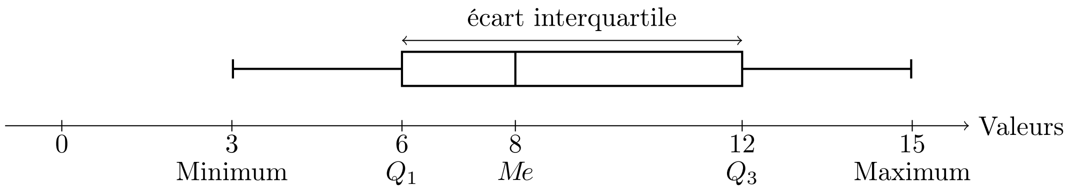
Exemple Soit le tableau statistique suivant :

Valeurs	3	5	6	8	11	12	15
f.c.c. (en %)	7	23	36	52	72	84	100

Où f.c.c. est la fréquence cumulée croissante ; On peut aussi utiliser les effectifs cumulés croissants (e.c.c.).
Alors :

- $Me = 8$ (première valeur pour laquelle la fréquence cumulée dépasse 50%)
- $Q_1 = 6$ (première valeur pour laquelle la fréquence cumulée dépasse 25%)
- $Q_3 = 12$ (première valeur pour laquelle la fréquence cumulée dépasse 75%)

On peut représenter une série statistique à une variable par un **diagramme en boîte** de la manière suivante, en reprenant les valeurs de l'exemple plus haut :



La droite est graduée. En particulier, il faut respecter une échelle choisie au départ. Ne pas oublier non plus de donner un titre à l'axe.

Remarque On peut alors lire dans ce diagramme plusieurs zones de valeurs correspondant à 25% de la population, ou à 50% de la population (à mettre en évidence sur le diagramme).

Méthode Voir page 136 pour l'utilisation de la calculatrice

▶ **Exercices :** 30 à 37p139

▶ **Exercice :** 39p139 (lecture de diagrammes et détermination de valeurs)

▶ **Exercice :** (en DM ?) 62p147, 65p148

II. Variance et écart-type

⊛ **Activité :** page 127

Définition Soit la série statistique suivante :

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

On note N l'effectif total ($N = n_1 + n_2 + \cdots + n_p$).
On rappelle que la moyenne, notée ici $\overline{x}$, est donnée par :

$$\overline{x} = \frac{n_1 \times x_1 + n_2 \times x_2 + \cdots + n_p \times x_p}{N}$$

La **variance** V de la série est la moyenne des carrés des écarts entre les valeurs et la moyenne :

$$V = \frac{n_1 \times (x_1 - \overline{x})^2 + n_2 \times (x_2 - \overline{x})^2 + \cdots + n_p \times (x_p - \overline{x})^2}{N}$$

Remarque La variance est toujours un nombre positif.
Cet indicateur statistique permet d'évaluer les écarts des valeurs autour de la moyenne.
Plus les écarts à la moyenne sont grands, plus la variance est grande, et réciproquement.

Définition La variance V étant positive, on appelle **écart-type** le nombre $\sigma = \sqrt{V}$.

Exemple Soit la série statistique suivante :

Valeurs	1	4	6
Effectifs	2	1	3

L'effectif total est $N = 2 + 1 + 3 = 6$.
La moyenne est $\overline{x} = \dfrac{2 \times 1 + 1 \times 4 + 3 \times 6}{6} = \dfrac{24}{6} = 4$.
La variance est alors $V = \dfrac{2 \times (1-4)^2 + 1 \times (4-4)^2 + 3 \times (6-4)^2}{6} = \dfrac{2 \times 9 + 3 \times 4}{6} = \dfrac{30}{6} = 5$.
Enfin l'écart-type est $\sigma = \sqrt{V} = \sqrt{5}$.

▶ **Exercices :** 19 à 22p138

▶ **Exercices :** 26 à 28p138 (Formule de König)

★ **Approfondissement :** 40,41,42p140, 48p142, 49p143

III. Échantillonnage

1. Intervalle de fluctuation

Voir : cours page 184 (rappel de seconde et représentation de la loi binomiale)

Définition On s'intéresse à un caractère de proportion p dans une population.
On prélève un échantillon aléatoire de taille n et on considère la variable aléatoire X, nombre d'individus de cet échantillon ayant ce caractère. La variable aléatoire X suit la loi binomiale $\mathscr{B}(n; p)$.
On appelle intervalle de fluctuation au seuil de confiance 95% de la fréquence l'intervalle $\left[\frac{a}{n}; \frac{b}{n}\right]$ où :

- a est le plus petit entier tel que $\mathbb{P}(X \leqslant a) > 0{,}025$;
- b est le plus petit entier tel que $\mathbb{P}(x \leqslant b) \geqslant 0{,}975$.

Il faut retenir que la probabilité que la fréquence observée pour un échantillon (autrement dit la valeur observée de X divisée par n) a une probabilité au moins égale à 0,95 d'appartenir à cet intervalle.
On obtient les valeurs de a et de b à l'aide de la calculatrice :

Casio : InvBinomCD(k,n,p) où l'on remplace k par 0,025 puis par 0,975 ;

TI : On peut afficher le tableau des valeurs des $\mathbb{P}(X \leqslant k)$ en définissant la fonction :
Y1=binomFRep(n,p,X)
Puis l'on cherche les valeurs de X qui font dépasser 0,025 puis 0,975.

Exemple Pour $n = 100$ et $p = 0{,}5$, on trouve $a = 40$ et $b = 60$.
L'intervalle de fluctuation au seuil de confiance 95% est donc $\left[\frac{40}{100}; \frac{60}{100}\right]$ soit $[0{,}4; 0{,}6]$.

Remarque Il s'agit d'un intervalle de fréquences, les valeurs sont donc nécessairement comprises entre 0 et 1.

▶ **Exercices :** exercice 37p192 (voir exercice corrigé Ep191), 120,121,124pp208-209

2. Prise de décision

On fait l'hypothèse que la proportion d'un caractère étudié dans une population est p.
Pour vérifier cette hypothèse on prélève un échantillon de taille n.
Une fois l'échantillon prélevé on observe une fréquence f d'apparition du caractère étudié dans l'échantillon. D'après la section précédente, la probabilité que f appartienne à l'intervalle de fluctuation au seuil 95% est de 0,95. Par conséquent on utilise la règle de décision suivante :

- Si f appartient à l'intervalle de fluctuation, l'hypothèse selon laquelle la proportion est p dans la population est validée.
- Si f n'appartient pas à l'intervalle de fluctuation, l'hypothèse est rejetée. Il y a alors un risque d'erreur de 5%.

Exemple Comme dans l'exemple précédent on suppose que $p = 0{,}5$ et $n = 100$. Supposons maintenant que dans un échantillon de 100 individus, 36 individus aient le caractère étudié.
Alors $f = \frac{36}{100} = 0{,}36$.
On rappelle que l'intervalle de fluctuation au seuil de confiance 95% est $I = [0{,}4; 0{,}6]$.
On observe que $f \notin I$, donc on rejette l'hypothèse que la proportion p soit égale à 0,5 dans la population.

Remarque En seconde on a déjà vu un intervalle de fluctuation. La formule était la suivante : $\left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$.
Cet intervalle est proche de celui de première, mais moins précis. De plus il son utilisation est soumise à des conditions portant sur n et p : il faut que $n \geqslant 25$ et $0{,}2 \leqslant p \leqslant 0{,}8$.
L'intervalle de première n'est soumis à aucune condition, par contre sa détermination est techniquement plus difficile (bien que ça devienne moins vrai avec les calculatrices actuelles).

▶ **Exercices :** 132 à 137pp211-212