Chapitre :
Statistiques

∽

⊛ **Activité :** QCM page 60 (rappels sur la moyenne et sur la médiane et les quartiles)

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

1. **Données brutes :**
   on donne la liste des valeurs, par exemple les quantités de pluie tombées à Rome chaque mois pendant une année :
   $$80; 73; 77; 47; 35; 20; 7; 35; 76; 88; 127; 109$$
   Il faut alors ordonner la liste par ordre croissant :
   $$7; 20; 35; 35; 47; 73; 76; 77; 80; 88; 109; 127$$
   Par suite, l'effectif total est $N = 12$. Donc :
   - Me : $\dfrac{N}{2} = 6$, donc Me est la sixième valeur, soit $Me = 73$.
     (une autre règle de seconde indique, lorsque N est pair, de faire la moyenne des valeurs centrales, ici $Me = \dfrac{73 + 76}{2} = 74{,}5$)
   - Q_1 : $\dfrac{N}{4} = 3$, donc Q_1 est la troisième valeur, soit $Me = 35$.
   - Q_3 : $\dfrac{N}{4} \times 3 = 9$, donc Q_3 est la neuvième valeur, soit $Me = 80$.
2. **Données par valeurs et effectifs :**
   Imaginons que l'on ait le tableau statistique suivant :

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

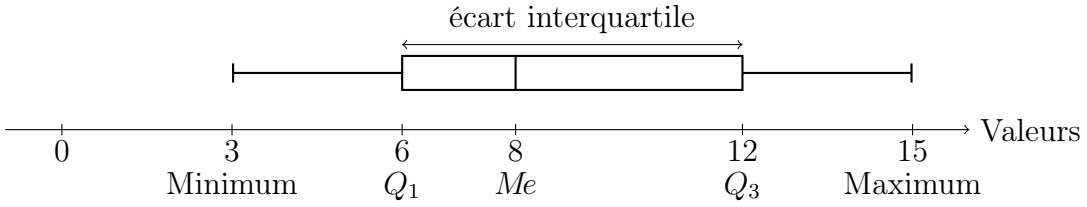
   Où f.c.c. est la fréquence cumulée croissante.
   Alors :
   - $Me = 8$ (première valeur pour laquelle la fréquence cumulée dépasse 50%)
   - $Q_1 = 6$ (première valeur pour laquelle la fréquence cumulée dépasse 25%)

- $Q_3 = 12$ (première valeur pour laquelle la fréquence cumulée dépasse 75%)

On peut aussi utiliser les effectifs cumulés croissants (e.c.c.)

▶ **Exercices :** 18,20,25p70 (détermination à la main ou à la calculatrice)

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).

▶ **Exercices :** 2p65 (lecture), 1p65 (construction à la calculatrice)

▶ **Exercices :** 27,28p71 (tableau à construire)

▶ **Exercices :** 34 et 35 p72 (comparaison)

II. Variance et écart-type

Définition Soit $(x_1,x_2,\ldots,x_N)$ une série statistique simple. On note $\overline{x}$ sa moyenne :

$$\overline{x} = \frac{x_1 + \cdots + x_N}{N}$$

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

$$V = \frac{(x_1 - \overline{x})^2 + \cdots + (x_N - \overline{x})^2}{N}$$

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 sont importants, plus la variance est grande.

Exemple On consisdère les notes de deux élèves :

- 10 ; 11 ; 12. Moyenne : 11. Variance : $\dfrac{(10-11)^2+(11-11)^2+(12-11)^2}{3} = \dfrac{2}{3} \simeq 0{,}67$
- 8 ; 11 ; 14. Moyenne : 11. Variance : $\dfrac{(8-11)^2+(11-11)^2+(14-11)^2}{3} = \dfrac{18}{3} = 6$

On remarque que les notes de l'élève 2 a des notes plus dispersées que l'élève 1, ce qui se confirme en voyant que la variance pour l'élève 2 est plus grande que celle de l'élève 1.

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

Exemple L'écart type des notes de l'élève 1 est $\sqrt{\frac{2}{3}}$.

▶ **Exercice :** 3p67

▶ **Exercice :** 40p73 (utilisation de la calculatrice pour plus de rapidité)

▶ **Exercice :** 48p74

Définition Dans le cas d'une série statistique avec effectifs, chacune des k valeurs x_i étant associée à un effectif n_i, alors la variance est définie par :

$$V = \frac{n_1(x_1 - \overline{x})^2 + \cdots + n_k(x_k - \overline{x})^2}{n_1 + \cdots + n_k}$$

où l'on rappelle que

$$\overline{x} = \frac{n_1 x_1 + \cdots + n_k x_k}{n_1 + \cdots + n_k}$$

▶ **Exercices :** 44,45p74

▶ **Exercices :** 52p75 (formule de Koenig) et 53p76 (utilisation de la formule)

▶ **Exercice :** (en DM ?) 54p76

★ **Approfondissement :** 58p78

III. Échantillonnage

1. Intervalle de fluctuation

⊛ **Activité :** 2p180 (représentation de la loi binomiale)

⊛ **Activité :** 4p181 sauf question 3. (détermination d'un intervalle de fluctuation)

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.

▶ **Exercice :** 1p183 (avec une table de lois binomiales)

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 acceptée.
- 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 $f = \dfrac{36}{100} = 0{,}36$.
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 : $\left[p - \dfrac{1}{\sqrt{n}}; p + \dfrac{1}{\sqrt{n}}\right]$. 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.

▶ **Exercice :** 3p185

▶ **Exercices :** 7 à 11p187

▶ **Exercices :** 15 à 17p189