Chapitre : Statistiques

I. Vocabulaire

Une étude statistique est l'étude d'un ou plusieurs caractères précis sur une population.

Définition (Population) Une population est un ensemble de personnes ou d'objets sur lesquels on peut prélever des renseignements communs.
Un élément de la population est appelé individu.
L'effectif total de la population est le nombre d'individus.

Exemple les élèves d'une classe ; les planches produites par une scierie ; ...

Définition (Caractère) Un caractère est ce qui est étudié sur la population.

Exemple pour chaque des élèves on peut s'intéresser à : sa note à un devoir, son régime (P, DP, E), la couleur de ses yeux, ...

Exemple pour chaque planche de la scierie, on peut s'intéresser à : sa taille, le nombre de défauts.

On observe deux types de caractères :

- les caractères **qualitatifs** sont ceux qui ne donnent pas de valeurs pour lesquelles une moyenne a un sens.

  (exemples : régime, couleur des yeux).

- les caractères **quantitatifs** sont ceux que l'on peut mesurer, compter, ordonner.

  (exemples : la note, le nombre de défauts, la taille).

  Parmi ceux-ci on distingue deux types :

  - ⋆ les caractères **continus** sont ceux dont les valeurs peuvent être pris sur des intervalles. C'est souvent le cas des mesures (longueur, temps, ...).
    (exemples : la taille peut prendre une infinité de valeurs si l'on mesure très précisément)
  - ⋆ les caractères **discrets** sont ceux qui prennent un nombre restreint de valeurs.
    (exemples : les notes, le nombre de défauts)

Pour les caractères qualitatifs il y a en général peu de choses à dire (la couleur d'yeux la plus fréquente). On s'intéresse ici par la suite aux caractères quantitatifs (on peut ordonner et faire la moyenne des notes entre autres).

II. Présentation par tableau

Une série statistique est la donnée brute de la valeur du caractère de chacun des individus composant la population étudiée.

Exemple on obtient une liste de notes : 15 ; 12 ; 08 ; 09 ; 12 ; 14 ; 09 ; 10 ; 17 ; 05.

Pour avoir une vision plus claire on préfère regrouper les valeurs dans un tableau, en les ordonnant. On obtient alors le tableau des effectifs :

Note	05	08	09	10	12	14	15	17
Effectif	1	1	2	1	2	1	1	1

Bien identifier le caractère (note) et l'effectif, selon les tableaux ce n'est pas toujours évident : les deux lignes d'effectif et de caractère pouvant être appelée toutes deux « nombre de ... ») ; c'est le cas par exemple si le caractère étudié est le nombre de pétales dans une population de fleurs.

On lit facilement **l'étendue de la série** : la valeur minimale (5) et la valeur maximale (17).
On peut alors éventuellement s'intéresser aux **effectifs cumulés croissants**.
Il s'agit de compter le nombre d'individus ayant leur caractère inférieur ou égal au caractère indiqué :

Note	05	08	09	10	12	14	15	17
Effectif cumulé croissant	1	2	4	5	7	8	9	10

L'effectif de la dernière colonne est donc l'effectif total. On peut alors voir directement le nombre de personnes ayant eu moins de 10 par exemple.
Au lieu des effectifs, on peut faire le même genre de tableau avec les **fréquences** (éventuellement cumulées aussi).

La fréquence est donnée par : $\dfrac{\text{effectif}}{\text{effectif total}}$.

On peut calculer la fréquence en pourcentage ; elle est donnée par : fréquence$\times$100.

Voici le tableau donnant les fréquence cumulées croissantes en pourcentage :

Note	05	08	09	10	12	14	15	17
fréquences cumulées croissantes en %	10	22	40	50	70	80	90	100

On finit nécessairement par 100.
Dans le cas d'un caractère continu, on utilise des intervalles pour les valeurs (voir page 118).

▶ **Exercices :** p110 (effectifs, fréquences, représentation)

III. Présentation par graphique

Il y a plusieurs représentations possibles.
Dans le cas d'un caractère discret, on peut utiliser des diagrammes en bâton (voir page 114).
Dans le cas d'un caractère continu, on peut utiliser des histogrammes (voir page 110), pour lesquels l'**aire** des rectangles est **proportionnelle** aux effectifs (ou aux fréquences).
Dans tous les cas, on peut aussi faire des diagrammes circulaires ou semi-circulaires (voir page 115 et 117). Ici, c'est l'angle qui est proportionnel aux effectifs.
Pour les effectifs cumulés croissants on peut tracer une courbe (voir page 118), la **courbe représentative des effectifs cumulés croissants**, aussi appelée **polygone** car représentée par une ligne brisée formée de segments.

▶ **Exercices :** 24p124 (diagramme circulaire et en bâtons), 14p119 (fréquences cumulées)

▶ **Exercice :** fiche d'exercices (1)

IV. Moyenne, médiane et quartiles

Définition Supposons que le caractère prenne p valeurs x_1,...,x_p avec des effectifs respectifs n_1, ..., n_p. On note N l'effectif total ($N = n_1 + \cdots + n_p$).
La **moyenne** de la série est alors donnée par :

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

Exemple Dans l'exemple du cours, la note moyenne est

$$\overline{x} = \frac{1 \times 5 + 1 \times 8 + 2 \times 9 + 1 \times 10 + 2 \times 12 + 1 \times 14 + 1 \times 15 + 1 \times 17}{10} = 11{,}1$$

Remarque on peut aussi utiliser les fréquences f_i au lieu des effectifs n_i.

$$\overline{x} = f_1 x_1 + \cdots + f_p x_p$$

▶ **Exercices :** 1,2,3 p115

▶ **Exercices :** 4 p115, 33p126 (centre des classes)

Définition La **médiane** d'une série est la plus petite valeur du caractère notée *Me* telle que au moins 50% des individus ont une valeur du caractère inférieure ou égale à *Me*

La médiane ne se calcule pas, mais se détermine. Pour cela on peut s'aider du tableau des fréquences cumulées. La médiane est la valeur du caractère qui fait atteindre ou dépasser les 50%.

Exemple Dans notre exemple, à la quatrième note la fréquence cumulée atteint exactement 50%. La médiane est donc cette quatrième note, à savoir 10.

Définition Le **premier quartile** est la plus petite valeur du caractère notée Q_1 qui fait atteindre ou dépasser les 25% de fréquence cumulée
Le **troisième quartile** est la plus petite valeur du caractère notée Q_3 qui fait atteindre ou dépasser les 75% de fréquence cumulée

Remarque la médiane pourrait être considérée comme un deuxième quartile.

Exemple Dans notre exemple, $Q_1 = 09$ et $Q_3 = 14$.

Remarque Lorsque l'on ne dispose pas des fréquences cumulées croissantes, il faut au moins disposer des effectifs cumulés croissants. Dans ce cas, en notant N l'effectif total :

- Pour obtenir *Me* on calcule dans un premier temps $\frac{N}{2}$; *Me* est alors la plus petite valeur pour laquelle l'effectif cumulé croissant dépasse $\frac{N}{2}$.
- Pour les quartiles le principe est le même, mais en calculant dans un premier temps $\frac{N}{4}$ et $\frac{3N}{4}$ respectivement.

Définition On appelle écart inter-quartile la différence $Q_3 - Q_1$.

▶ **Exercices :** 11p117 puis 9p117 (médiane et quartiles)

▶ **Exercice :** fiche exercices (2)

▶ **Exercice :** (en DM, expérience statistique)

V. Estimations statistiques

⊛ **Activité :** explication et observation du fichier de tableur avec les fluctuations (lancer de dé).
Une expérience aléatoire est une expérience pour laquelle il est impossible de prévoir le résultat, autrement dit dont le résultat est soumis **au hasard**. Par exemple, le lancer d'un dé ou le tirage d'une boule dans une urne.
Si l'on souhaite répéter une expérience un très grand nombre de fois, comme lancer 5 000 fois un dé, on peut faire soi-même l'expérience avec un dé (à lancer 5 000 fois !) ou utiliser un simulateur comme **un ordinateur ou une calculatrice**.
La série statistique formée des n résultats obtenus lorsqu'on répète n fois une même expérience dans les mêmes conditions est appelée **un échantillon** de taille n.
Lorsque l'on souhaite observer l'apparition d'un caractère en particulier (par exemple obtenir une boule rouge dans une urne contenant 8 boules rouges et 2 boules vertes) on compte le nombre d'apparitions de ce caractère, puis on calcule la fréquence par rapport au nombre de répétitions (on divise donc par $\boldsymbol{n}$).

Propriété D'un échantillon à l'autre, la fréquence obtenue n'est pas toujours la même, elle **fluctue** : c'est ce qu'on appelle la **fluctuation d'échantillonnage**. Lorsque la taille n de l'échantillon augmente l'ampleur des fluctuations **diminue** et les fréquences ont tendances à se **stabiliser**.

1. Intervalle de fluctuation

Définition En statistique, on peut s'intéresser à un caractère qui est soit présent soit non présent (être une femme dans une population humaine, obtenir une somme égale à 7 dans un lancer de deux dés). Si l'on appelle p la probabilité que le caractère soit présent, ainsi la probabilité qu'il ne le soit pas est $\mathbf{1}-\boldsymbol{p}$.
On peut dire que la population est compatible avec un modèle de Bernoulli de probabilité p.

Propriété On admet que si on a un échantillon de taille n d'une population, dont un caractère a la probabilité p d'être présent et tel que $n \geq 25$ et $0{,}2 \leq p \leq 0{,}8$, alors dans ce cas pour 95% des échantillons la fréquence d'apparition du caractère appartient à l'intervalle

$$\left[p-\frac{1}{\sqrt{n}};p+\frac{1}{\sqrt{n}}\right]$$

Définition Cet intervalle est appelé **intervalle de fluctuation au seuil de 95%**.

▶ **Exercice :** 15p140 juste avant la partie 2. de la fiche

2. Prendre une décision à partir d'un échantillon

Pour décider si une fréquence f observée sur un échantillon de taille n est compatible ou non avec un modèle de Bernoulli de probabilité p on teste l'appartenance de f à l'intervalle de fluctuation au seuil de 95% donné ci-dessus. Par suite,

- Si f n'est pas dans l'intervalle de fluctuation, alors on peut **rejeter** l'hypothèse que l'échantillon est compatible avec le modèle.

- Si f est dans l'intervalle de fluctuation, alors on **accepte** l'hypothèse que l'échantillon est compatible avec le modèle.

Remarque Quelle que soit la décision prise il y a toujours le risque que ce ne soit pas la bonne décision.

Exercice 1
En Suède, la proportion de femmes parmi les 349 députés est 44,7%. Elle est de 50,5% dans la population totale du pays. La parité est-elle respectée à la Diète Royale suédoise au seuil de 95% ?

Exercice 2
Dans une usine automobile, on contrôle les défauts de peinture de type « grains ponctuels sur le capot ». Lorsque le processus est sous contrôle, on a 20% de ce type de défauts. Lors du contrôle aléatoire de 50 véhicules, on observe des défauts sur 13 d'entre eux. Faut-il s'inquiéter ?

▶ **Exercices :** 16 à 20 pages 140 et 141