Statistiques et échantillonnage

1. Vocabulaire

Définition. *Une étude statistique commence par le recueil de données. Cela se fait sur une population, dont les membres sont appelés individus. On s'intéresse à une particularité, appelée caractère, de ces individus. Ce caractère prend un certain nombre de valeurs qui peuvent être de deux types : quantitatives lorsqu'il s'agit de nombres, qualitatives sinon.*

Exemple

On s'intéresse aux salaires mensuels des employés d'une entreprise. La population étudiée est l'ensemble des salariés et le caractère étudié est le salaire mensuel de chacun d'eux : c'est un caractère quantitatif.

Définition. *On appelle effectif total le nombre d'individus de la population étudiée. On le note N. L'effectif d'une valeur du caractère est le nombre d'individus possédant ce caractère. On le notera n.*
On appelle fréquence d'une valeur du caractère le quotient de son effectif par l'effectif total. Ce nombre sera noté f. On a $f = \frac{n}{N}$.

Remarques.

1. On a toujours $0 \leq f \leq 1$.
2. Dans la vie courante, pour connaître la part occupée par un groupe dans une population, on utilise souvent les pourcentages. Il suffit de multiplier par 100 la fréquence pour obtenir le pourcentage. **On n'écrira pas la multiplication par 100.** (voir ci-dessous).

Exemple

Parmi les 500 élèves de seconde du lycée, 58 ont choisi l'option latin. La fréquence de latinistes parmi les élèves de seconde est donc $\frac{58}{500} = 0{,}116 = 11{,}6\ \%$.

Les valeurs du caractère étudié sont souvent notées x_1, x_2, …, x_p. Les effectifs correspondant sont n_1, n_2, …, n_p.

Valeur du caractère x_i	x_1	x_2	x_3	…	x_p
Effectif n_i	x_1	n_2	x_3	…	n_p

Définition. *Dans certains cas, il peut être intéressant, voire nécessaire, de regrouper les valeurs ; c'est le cas lorsque le caractère peut prendre un très grand nombre de valeurs. On les regroupe alors dans des intervalles, qu'on appelle classes.*

Exemple

On a fait une étude sur la taille des 500 élèves de seconde. Il serait inutile de donner la liste des 500 tailles. On peut par exemple les regrouper par classes d'amplitude 10 cm.

Taille	[140; 150[	[150; 160[	[160; 170[	[170; 180[	[180; 190[	[190; 200[
Effectif	15	51	145	201	91	3

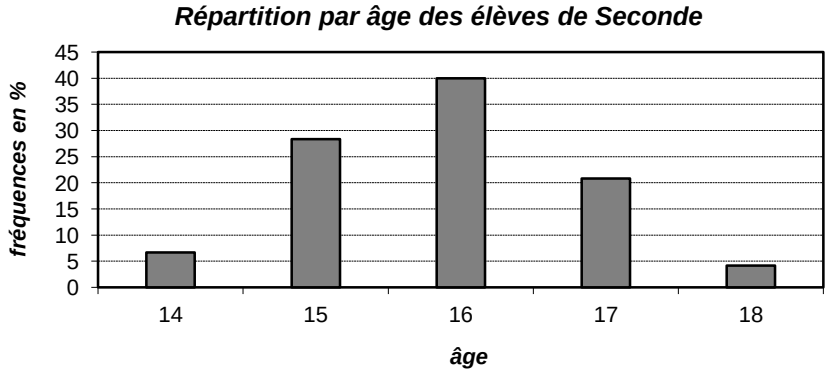
2. Représentations graphiques

Lorsqu'on recueille les données, on les place la plupart du temps dans un tableau. Si l'effectif total est important, ou s'il y a beaucoup de valeurs différentes pour le caractère, ce tableau peut très vite devenir illisible. Pour donner à lire rapidement cette liste de données, on utilise des diagrammes de différents types.

❖ Le diagramme en barre ou en bâtons

On donne la répartition par âge des élèves de Seconde dans un lycée.

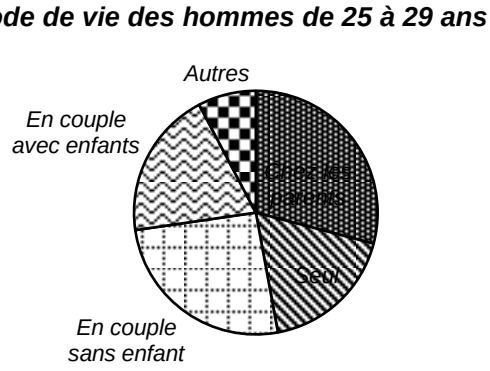
Âge	14	15	16	17	18
Effectif	8	34	48	25	5
Fréquence	0,07	0,28	0,4	0,21	0,04



❖ Le diagramme circulaire

On connaît le mode de vie des hommes de 25 à 29 ans.

	Pourcentage des hommes entre 25 et 29 ans	***Angle au centre du diagramme circulaire***
Chez les parents	29,1 %	
Seul	18 %	
En couple sans enfant	25,6 %	
En couple avec enfants	19,6 %	
Autres	7,7 %	
Total	100 %	360°



3. Mesures en statistiques

❖ La moyenne

Définition. *La moyenne d'une série statistique est la somme de toutes les valeurs divisée par l'effectif total.*

Exemple A

Voici les âges des élèves d'une classe ayant eu mention très bien au brevet :

$$14 - 15 - 15 - 14 - 13 - 13 - 14 - 14 - 14 - 13.$$

La somme des notes est de $14 + 15 + 15 + 14 + 13 + 13 + 14 + 14 + 14 + 13 = 139$.

Comme l'effectif est de 10 élèves, la moyenne des âges est $\frac{139}{10} = 13{,}9$ ans.

Remarque. Si les données sont regroupées par classes, le calcul de la moyenne se fait en prenant le centre des classes.

Si le caractère prend les p valeurs x_1, x_2, …, x_p d'effectifs respectifs x_1, x_2, …, x_p, on calculera la moyenne avec la formule :

$$\frac{n_1x_1 + n_2x_2 + \cdots + n_px_p}{n_1 + n_2 + \cdots + n_p}.$$

Exemple A

Les données précédentes peuvent se regrouper dans ce tableau.

Âge	13	14	15
Effectif	3	5	2

La moyenne peut dont se calculer de la façon suivante.

$$\frac{3 \times 13 + 5 \times 14 + 2 \times 15}{3 + 5 + 2} = \frac{139}{10} = 13{,}9.$$

Théorème (linéarité de la moyenne).

- Si on ajoute à toutes les valeurs d'un caractère quantitatif un nombre a, alors la moyenne est augmentée de a.
- Si on multiplie toutes les valeurs d'un caractère quantitatif par un nombre a, alors la moyenne est multipliée par a.

Exemple

Dans une entreprise, le salaire moyen des femmes est de 1 206 € par mois. Au bout de six mois, les salaires augmentent de 100 €, par conséquent le nouveau salaire moyen des femmes est : 1206 + 100 = 1306 €.

Six mois plus tard, les salaires augmentent de 5 %. Puisque une augmentation de 5 % se traduit par une multiplication par 1,05, le salaire moyen des femmes devient :

$$1306 \times 1{,}05 = 1371{,}3 \text{ €}.$$

Théorème (moyenne par paquet). Si une population est partagée en deux sous-groupes disjoints, d'effectifs p et q, et si les moyennes de ces sous-groupes sont respectivement M_1 et M_2, on peut calculer la moyenne de la série par la formule :

.

Exemple

Le salaire moyen des 47 hommes d'une entreprise est de 1380 € et celui des 35 femmes est de 1206 €. Le salaire moyen dans cette entreprise est donc égal à :

$$\frac{47 \times 1380 + 35 \times 1206}{47 + 35} \approx 1306 \text{ €}.$$

❖ Médiane

Définition. *La <u>médiane</u> est une valeur qui partage la population en deux moitiés : celle dont les individus ont une valeur du caractère inférieure à cette médiane, et l'autre, dont les individus ont une valeur du caractère supérieure à cette médiane.*

En pratique, pour déterminer la médiane, on range les N valeurs par ordre croissant, et on distingue deux cas :

- Si N est impair, on prend la valeur centrale, c'est-à-dire la valeur de rang $\frac{N+1}{2}$.
- Si N est pair, on prend la moyenne des deux valeurs centrales, c'est-à-dire des valeurs de rang $\frac{N}{2}$ et $\frac{N}{2}+1$.

Exemple B

Imaginons un devoir où les notes ont été

3 – 5 – 6 – 7 – 7 – 10 – 11 – 12 – 13 – 13 – 14 – 15 – 18 – 19.

Cette série compte 14 valeurs, donc la médiane est la moyenne des 7^{e} et 8^{e} valeurs, elle est donc égale à $\frac{11+12}{2} = 11{,}5$.

Cela signifie que 50 % des élèves ont eu moins de 11,5 et 50 % des élèves ont eu plus de 11,5.

❖ Quartiles

Définition. *Le <u>premier quartile</u> Q_1 est la plus petite valeur du caractère telle que 25 % de l'effectif ait une valeur du caractère inférieure ou égale à Q_1.*
Le <u>troisième quartile</u> Q_3 est la plus petite valeur telle que 75 % de l'effectif ait une valeur du caractère inférieure ou égale à Q_3.

S'il y a N valeurs, Q_1 est la valeur dont le rang est **le premier entier supérieur ou égal** à $\frac{N}{4}$ et Q_3 est la valeur dont le rang est le **premier entier supérieur ou égal** à $\frac{3}{4}N$.

Exemple B

Puisque $\frac{14}{4} = 3{,}5$, le premier quartile est la 4^{e} valeur de la série, donc $Q_1 = 7$.

Puisque $\frac{3}{4} \times 14 = 10{,}5$, le troisième quartile est la 11^{e} valeur de la série, donc $Q_1 = 14$.

4. Échantillonnage

❖ Échantillon

Définition. *Un échantillon de taille n est obtenu à partir d'une population en réitérant n fois de suite l'opération suivante : on prélève au hasard un de ses éléments, on note la valeur du caractère prélevé et on remet l'élément prélevé dans la population.*

Remarque. Bien souvent pour des raisons pratiques, il n'y a pas de remise lors du prélèvement. On peut néanmoins prouver que les résultats suivants restent vrais lorsqu'il n'y a pas remise après chaque prélèvement, pour peu que l'effectif total soit très grand par rapport à la taille de l'échantillon.

Exemple C

Dans un pays, il y a 27 % de fumeurs, soit une proportion $p = 0{,}27$ de fumeurs.
Réaliser un échantillon de taille $n = 400$ consiste à prendre successivement et avec remise une personne de la population et à noter si elle fume ou pas.
Admettons que l'on ait dénombré 140 fumeurs. La fréquence de fumeur observée sur cet échantillon est donc $f = \frac{140}{400} = 0{,}35 = 35\ \%$.

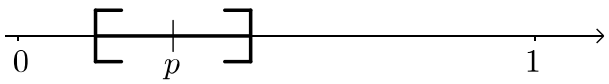
❖ Intervalle de fluctuation

Soit une population pour laquelle on étudie un caractère pouvant prendre uniquement deux valeurs. (« vrai » ou « faux » ; « garçon » ou « fille » ; « rouge » ou « noir » etc.).

On suppose connaître la proportion p de la population pour laquelle le caractère est vrai ; nous la nommerons proportion effective.

Un échantillon de taille n de cette population est prélevé. On peut mesurer la proportion f de l'échantillon pour laquelle le caractère est vrai ; nous la nommerons fréquence ou fréquence observée.

Définition. *Un intervalle de fluctuation au seuil de 95 % d'une fréquence d'un échantillon de taille n est l'intervalle I centré autour de la proportion effective p tel que la fréquence observée f sur l'échantillon se trouve dans I avec une probabilité égale (au moins) à 0,95.*



Théorème. Soit une population pour laquelle on connaît la proportion effective p d'un caractère. On s'intéresse à la fréquence observée f du caractère dans un échantillon de taille n prélevé dans cette population.
L'intervalle de fluctuation au seuil de 95 % de f est l'intervalle $\left[p - \frac{1}{\sqrt{n}}\ ;\ p + \frac{1}{\sqrt{n}}\right]$.
Cela signifie qu'il y a une probabilité supérieure ou égale à 0,95 pour que la fréquence observée du caractère dans un échantillon de taille n se trouve dans l'intervalle

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

Exemple C

L'intervalle de fluctuation de la fréquence observée de fumeurs pour les échantillons de taille 400 est

$$\text{IF} = \left[0{,}27 - \frac{1}{\sqrt{400}} \,;\, 0{,}27 + \frac{1}{\sqrt{400}}\right] = [0{,}22; 0{,}32].$$

Cela signe qu'avec une probabilité d'au moins 95 %, la fréquence observée f sur un échantillon de taille 400 est comprise entre 0,22 et 0,32.

❖ **Prise de décision à partir d'un échantillon**

Exemple C

Une enquête menée sur 400 jeunes de 17 à 25 ans a montré que 140 d'entre eux fument. Le taux de fumeur chez les jeunes est-il dans la moyenne nationale ?

Réponse. La fréquence observée du nombre de fumeur sur cet échantillon est 0,35. Or

$$0{,}35 \notin [0{,}22; 0{,}32]$$

par conséquent on peut dire que le taux de fumeurs n'est pas dans la norme nationale, au seuil de confiance de 5 %.

Remarque. La formule du théorème précédent est une formule d'approximation qui est de plus en plus précise quand n devient grand.
Dans la pratique, on considère qu'elle est exacte quand $n \geq 25$ et $0{,}2 \leq p \leq 0{,}8$.