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 notera 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) La somme des fréquences est égale à 1.
3) 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.

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 $\frac{58}{500} = 0{,}116 = 11{,}6\%$.

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 parfaitement inutile et inexploitable 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

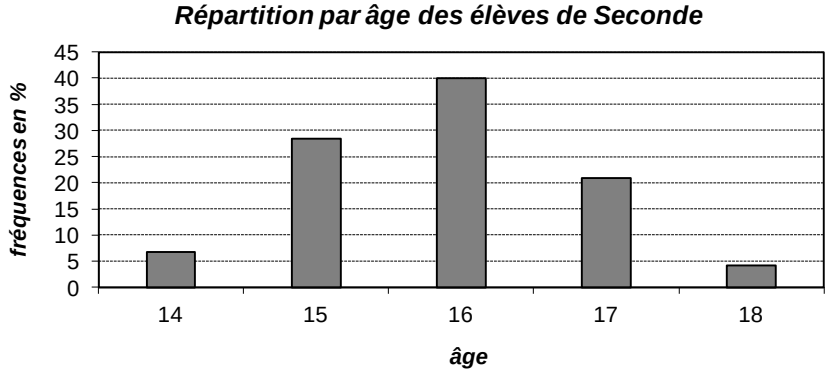
2. Représentations graphiques

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.

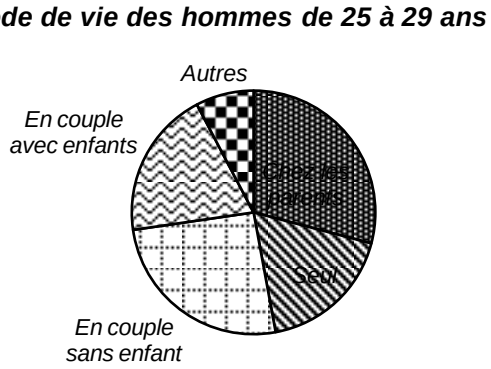
Âge	14	15	16	17	18
Effectif	8	34	48	25	5
Fréquence	0,07	0,28	0,4	0,21	0,04



❖ 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.*

Dans le cas où on connaît les effectifs de chaque valeur, on peut multiplier cette valeur par son effectif plutôt que de l'additionner autant de fois. Ainsi, si on a les valeurs X_1, X_2, … , X_p d'effectifs respectifs n_1,n_2 , …, n_p, on calculera la moyenne (dite pondérée), notée $\bar{X}$, avec la formule :

$$\bar{X} = \frac{n_1X_1 + n_2X_2 + \cdots + n_pX_p}{n_1 + n_2 + \cdots + n_p}$$

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

Théorème (linéarité de la moyenne). Si on multiplie toutes les valeurs X_i par un nombre a, alors la moyenne est multipliée par a. Si on ajoute à toutes les valeurs X_i un nombre b, alors la moyenne est augmentée de b.

Exemple

Dans une entreprise, le salaire moyen des femmes est de 1 206 € par mois. Au bout de six mois, les salaires augmentent de 3 % et six mois plus tard, ils augmentent de 100 €. Le salaire moyen des femmes est alors égal à :

$$1206 + \frac{3}{100} \times 1206 \approx 1342 \text{ €}.$$

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 :

$$\bar{X} = \frac{pM_1 + qM_2}{p + q}.$$

Exemple

Le salaire moyen des 47 hommes de cette entreprise est de 1380 € et que 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 médiane est une valeur qui partage la population en deux moitiés : celle dont les individus ont une valeur inférieure à cette médiane, et l'autre, dont les individus ont une valeur supérieure à cette médiane.*

Si les données ont été réparties en classes, on ne peut déterminer la médiane exacte. En revanche, on appellera ***classe médiane***, la classe qui la contient.

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.
- Si N est pair, on prend les deux valeurs centrales, et on fait leur moyenne, qui sera choisie comme valeur médiane.

Exemple A

Soit la série : $1-1-2-3-4-4-5-7-7-8-9-10-13-15$.
Cette série compte 14 valeurs, donc la médiane est la moyenne des $7^{\text{ème}}$ et $8^{\text{ème}}$ terme, elle est donc égale à $\frac{5+7}{2}=6$.

❖ Quartiles

Définition. *Le premier quartile Q_1 est la plus petite valeur telle que 25 % des données lui soient inférieures ou égales. Le troisième quartile Q_3 est la plus petite valeur telle que 75 % des données lui soient inférieures ou égales.*

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 $\frac{3N}{4}$.

Exemple A

L'entier supérieur à $\frac{14}{4}$ est 4, celui supérieur à $3\times\frac{14}{4}$ est 11 donc
$Q_1=3$ et $Q_3=9$.

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.

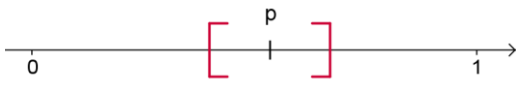
Exemple B

Dans une entreprise qui fabrique des lecteurs MP3 pour effectuer un contrôle dans la fabrication, on peut prélever un échantillon de taille 100. C'est-à-dire que l'on prend au hasard un lecteur pour vérifier si le logo est défectueux, puis on le remet. Et ceci 100 fois de suite.

❖ Intervalle de fluctuation

Soit une population pour laquelle on étudie un caractère pouvant prendre les valeurs « vrai » ou « faux ». 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 se trouve dans I avec une probabilité égale à 0,95.*



Exemple B

Des études ont montré que 25 % des lecteurs MP3 ont le logo défectueux. La proportion effective est donc égale à 0,25.
Lorsque l'on prélève un échantillon de taille $n = 100$ dans cette population, on peut compter le nombre de lecteurs ayant le logo défectueux. Notons n_d ce nombre. La fréquence observée f est $\frac{n_d}{n}$.
Pour un échantillon de taille 100, l'intervalle de fluctuation au seuil de 95 % de la fréquence des lecteurs au logo défectueux est un intervalle de centre 0,25 tel que pour 95% des échantillons de taille 100, les fréquences observées se trouvent dans I.

Théorème. Soit une population pour laquelle on connaît la proportion effective d'un caractère p située entre 0,2 et 0,8. On s'intéresse à la fréquence observée f du caractère dans un échantillon de taille $n \geq 25$ 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 B

On est dans le cas d'application du théorème car $n \geq 25$ et $0,2 \leq p \leq 0,8$. L'intervalle de fluctuation de la fréquence observée des lecteurs MP3 présentant un défaut au niveau du logo pour des échantillons de taille 100 est

$$\left[0,25 - \frac{1}{\sqrt{100}} ; 0,25 + \frac{1}{\sqrt{100}}\right] = [0,15; 0,35].$$

Cela signe qu'avec une probabilité d'au moins 95 %, la fréquence observée est comprise entre 0,15 et 0,35.

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

Exemple B

Un contrôle en sortie de production a lieu sur 100 lecteurs MP3, on en détecte 37 défectueux. La machine nécessite-t-elle d'être réparée ?

Réponse. La fréquence observée du nombre de lecteurs défectueux est $0{,}37$. Or

$$0{,}37 \notin [0{,}15; 0{,}35]$$

par conséquent on peut dire que la machine a besoin d'être réparée, au seuil de confiance de 5%.