Cours : séance 3 (lundi 20 janvier)

Cinquième partie 5. **Statistiques et probabilités.**

Voici le contexte pour faire une statistique.

Il y a une *population*, c'est à dire un ensemble homogène (de personnes, d'êtres vivants, d'objets...) La population considérée est toujours finie, soyons réaliste. Un membre de cette population est un *individu*. Le nombre d'individus de l'ensemble de la population, ou d'un sous-groupe de la population, est appelé son *effectif*.

Sur cette population on effectue une *mesure* (la taille, l'âge, le poids...), on obtient donc une *série* (de mesures). (En fait on peut aussi considérer des mesures non quantitative : des caractères qualitatifs.) Attention : la série (de valeurs) est la *liste* des valeurs pour chacun des individus de la population, en particulier deux individus distincts peuvent avoir la même mesure (le même caractère), cela donne une répétition dans la série. Ainsi dans la série $0; 0; 1; 0; 1; 1; 1; 0; 1$ la valeur 0 apparaît 4 fois, et la valeur 1 apparaît 5 fois : ce n'est pas la même série que $0; 1$ - bien que les valeurs apparaissant dans les deux séries soient les mêmes ! La série de valeurs constitue la *donnée* brute à partir de laquelle on travaille. La première chose à faire est souvent de remettre dans l'ordre les termes de la série. Pour revenir au précédent exemple la série réorganisée devient $0; 0; 0; 0; 1; 1; 1; 1; 1$: l'ordre a changé mais les propriétés statistiques sont les mêmes.

En statistique on s'intéresse notamment à l'ensemble (on dit plutôt la *classe*) des individus de la population pour lesquels la mesure considérée prend une certaine valeur, ou bien à la classe des individus de la population pour lesquels la mesure considérée est dans un certain intervalle de valeurs. Ce qui importe, c'est l'effectif de ces classes.

Bien prendre garde à toujours distinguer la population étudiée de la mesure effectuée : la mesure est (souvent) un nombre (mais pas toujours), les individus de la population ne sont pas des nombres a priori ("Je ne suis pas un numéro !").

On se pose plusieurs questions sur les valeurs mesurées. Par exemple on se demande :

(1) combien d'individus de la population ont une mesure donnée ? Ceci se représente par un diagramme en bâton : il y a deux axes, sur l'axe des abscisses on porte les mesures pertinentes, puis on dessine des rectangles (les bâtons) centrés sur la valeur choisie de la mesure, deux à deux disjoints, de hauteur égale à l'effectif.

(2) quelle proportion (en pourcentage) des individus de la population ont une mesure donnée ? On peut représenter ces données par un diagramme circulaire. Pour chaque groupe d'individus ayant une mesure donnée on calcule le pourcentage p du total, puis on place un secteur angulaire d'ouverture $\frac{p \times 360}{100}$ degrés (ce secteur est coloré de manière à identifier à quelle mesure il correspond). Les secteurs sont contigus et ne se chevauchent pas, comme au total il y a 100%, la figure obtenue est un disque complet.

(3) combien d'individus de la population ont une mesure comprise entre deux bornes données ? (Le groupe des individus de mesure comprise entre deux bornes données est appelée une *classe*.) Ceci se représente par un histogramme : il y a deux axes, sur l'axe des abscisses on porte les intervalles de mesures pertinents, puis on dessine des rectangles (les bâtons) posés sur

l'intervalle de mesure choisi, de hauteur égale à l'effectif de la classe. Ces bâtons sont contigus si les intervalles sont contigus !

On réalise que pour analyser un jeu de données obtenu sur une "grande" population, la bonne approche est de découper la série des mesures possibles en une réunion d'intervalles (ou sous-ensembles plus généraux), et de considérer alors les classes correspondant à chaque intervalle de mesure ou élément de la partition. On compte alors l'effectif de chaque classe ainsi obtenue.

Supposons donc choisi un découpage de la série de mesures possibles, ce qui nous donne un découpage de la population en classes comme ci-dessus.

La *fréquence* d'apparition dans une série donnée d'une mesure (caractère) est définie comme le quotient

$$\boxed{f = \frac{\text{effectif de la classe des individus ayant ce caractère}}{\text{effectif total de la population}}},$$

que l'on exprime souvent en pourcentage. Plus généralement la *fréquence* d'apparition dans une série donnée d'un intervalle de mesure est définie comme le quotient

$$\boxed{f = \frac{\text{effectif de la classe des individus ayant une mesure dans l'intervalle}}{\text{effectif total de la population}}}.$$

La *moyenne* d'une série de mesures numériques est $\boxed{m = \frac{\text{somme des mesures pour chaque individu}}{\text{effectif total de la population}}}$ (à ne pas confondre avec la moyenne des mesures possibles...). Par définition :

$$\text{moyenne} \times \text{effectif total} = \text{ la somme des mesures pour chaque individu de la population}$$

Autrement dit : les mesures varient suivant les individus et forment ensemble une somme totale de mesures, et pour obtenir la même somme totale avec une valeur commune de la mesure pour tous les individus, il faudrait que la mesure de chaque individu soit précisément la moyenne. Dit autrement : si on modifie les valeurs d'une série en conservant la somme de toutes les valeurs, alors la façon d'obtenir une valeur égale pour tous les individus est de supposer qu'on leur a donné pour mesure la moyenne (principe des vases communicants). En version géométrique : si on représente le diagramme en bâton (donc avec des hauteurs de bâton variable suivant les individus) et si on trace la droite horizontale H correspondant à la moyenne, on a des bâtons qui finissent au dessus de la droite H (excédent) et des bâtons qui finissent au dessous de la droite H (déficit) alors la somme des parties excédentaires compense exactement la somme des parties déficitaires.

L'expression du numérateur peut être un peu simplifiée : Si on regroupe dans la somme des mesures pour chaque individu les termes correspondant aux individus qui ont la même mesure on obtient : la somme des mesures possibles $\times$ (le nombre d'individus qui ont cette mesure), ce qui est plus "ramassé".

Si la série est présentée par classe, on peut aussi définir la *moyenne par classe* :

$$\boxed{m = \frac{\text{somme des valeurs centrales pour chaque classe} \times \text{l'effectif de la classe}}{\text{effectif total de la population}}}.$$

La valeur centrale pour une classe définie par un intervalle de mesure $[a; b[$ est $\dfrac{a+b}{2}$.

Exercice 5.1 – Vocabulaire : population, effectif, caractère. C'est donc un exercice de lecture d'énoncé.

1. La population étudiée est l'ensemble des élèves du collège.

2. Les caractères étudiés sont : le sexe (qualitatif), l'âge (quantitatif), la couleur des yeux (qualitatif), la taille (quantitatif).

3. Une valeur possible pour le sexe est "Féminin". Une valeur possible pour l'âge est "13 ans".

4. En admettant qu'il n'y ait que des filles et des garçons, la population totale est donc de $223+217 = 440$ élèves. Le caractère retenu ici pour faire ces deux groupes (F/G) est le sexe.

Exercice 5.3 – Utilisation d'un tableau pour présenter des données.

1. Fréquence veut juste dire ici fraction du total des visiteurs (au cours de la première heure, un jour donné). La somme des fréquences doit valoir 1. Donc la fréquence des enfants est $f = 1 - 0,45 - 0,1 - 0,2 = 0,25$ (un quart des visiteurs sont des enfants).

Origine	Adultes	Enfants	Etudiants	Groupes
Fréquence	$0,45$	$0,25$	$0,1$	$0,2$

2. Il s'agit simplement d'une conversion en pourcentage : on transforme la fraction $\frac{a}{b}$ en $\frac{p}{100} = p\%$, donc $p = 100 \times \frac{a}{b}$.

Origine	Adultes	Enfants	Etudiants	Groupes
Fréquence	$0,45$	$0,25$	$0,1$	$0,2$
Pourcentage	45%	25%	10%	20%

3. On applique le pourcentage à l'effectif total : $p\%$ de $1700 = \frac{p \times 1700}{100}$, ou alors on utilise la fréquence : $f \times 1700$. On peut tout faire de tête en remplissant dans l'ordre les cases 10%, 20%, 25%, 45%.

Origine	Adultes	Enfants	Etudiants	Groupes
Fréquence	$0,45$	$0,25$	$0,1$	$0,2$
Pourcentage	45%	25%	10%	20%
Effectif	765	425	170	340

Exercice 5.4 – Deux types de diagrammes pour présenter des données.

Diagramme en bâtons : Il faut choisir dans quel ordre on dispose les différents bâtons sur l'axe des x. On peut choisir de mettre les bâtons selon l'effectif croissant. Ou de mettre les bâtons dans l'ordre alphabétique du nom des candidats correspondant, en isolant l'abstention soit tout à gauche, soit tout à droite. Notons que l'axe des x n'a pas besoin d'être gradué ! L'effectif, lui, se lit sur l'axe des y.

Camembert : Il faut choisir dans quel ordre (circulaire) on dispose les différents secteurs (parts de camembert) sur le disque. Puis il faut calculer le pourcentage de 360 pour chaque secteur. L'angle α du secteur circulaire représentant est proportionnel à la fréquence du caractère représenté : $\frac{\alpha}{360} =$

$\frac{\text{effectif du caractère}}{\text{effectif total}}$, ainsi $\alpha = \frac{360 \times \text{effectif du caractère}}{\text{effectif total}}$, ou encore $\alpha = \text{effectif du caractère} \times \frac{360}{\text{effectif total}}$.

L'effectif total est 432, et $\frac{360}{432} = \frac{5}{6}$. L'angle pour M. sera donc de $96 \times \frac{5}{6} = 80$ degrés, pour S. de $72 \times \frac{5}{6} = 60$ degrés, pour B. de $60 \times \frac{5}{6} = 50$ degrés, pour D. de $156 \times \frac{5}{6} = 130$ degrés, pour Abstention de $48 \times \frac{5}{6} = 40$ degrés.

Exercice 5.5 – Moyenne et vases communicants.

1. Il y a 6 notes indiquées, la moyenne que l'on trouverait est $m = \frac{11+8+12+13+9+10}{6} = \frac{63}{6} = 10,5$, ce qui ne cadre pas avec la note indiquée sur le bulletin.

2. Il y avait en fait 7 notes, dont une (x) qu'il faut retrouver pour obtenir la moyenne de 10. On a donc

$$10 = \frac{11+8+12+13+9+10+x}{7} = \frac{63+x}{7}, \text{ et donc } 63+x = 7\times 10 = 70, \text{ d'où } x = 70-63 = 7.$$

Ainsi le jeune Adrien a oublié de signaler un 7/20 à ses parents.

Exercice 5.7 – La moyenne tient compte de l'effectif.

1. Ce qui est calculé ici c'est simplement la moyenne des salaires dans cette entreprise, sans tenir compte de l'effectif de chaque catégorie de salariés. Ce n'est donc que rarement la moyenne des salaires versés par l'entreprise / reçus par l'ensemble des salariés. Cela donne le bon résultat s'il y a autant d'ouvriers que d'employés et de cadres - ce qui est bien rare. Mais s'il y a 80% d'ouvriers, 15% d'employés et 5% de cadres, alors l'effectif (largement majoritaire) des ouvriers tire la moyenne vers le bas.

2. Si l'on note x le nombre d'ouvrier, y le nombre d'employés et z le nombre de cadres, la moyenne est en fait donnée par la formule $m = \frac{1480\times x + 1620\times y + 3450\times z}{x+y+z}$. Puisqu'ici $x = 38, y = 24, z = 6$ on trouve $m = \frac{115820}{68} \simeq 1703$ euros. Alors que si on prend $x = y = z$ on obtient $m = \frac{1480+1620+3450}{3} \simeq 2183$.

Exercice 5.8 – Moyenne par classes.

Dans cette situation, la distinction entre ouvrier et cadre n'est pas pertinente. En effet la question est simplement celle des salaires. L'information intéressante est l'effectif de chaque *classe* de salaire. On regroupe les salariés en trois classes : la classe C_1 où le salaire est dans $[1000; 1500[$, la classe C_2 où le salaire est dans $[1500; 3000[$ et la classe C_3 où le salaire est dans $[3000; 8000[$.

1. On fait comme si tous les individus d'une classe donnée avaient pour salaire le centre de l'intervalle de salaire correspondant à la classe. Pour C_1 : $s_1 = 1,25$; pour C_2 : $s_2 = 2,25$; pour C_3 : $s_1 = 5,5$. Si on note N_1, N_2, N_3 les effectifs des classes C_1, C_2, C_3 la formule pour la moyenne "par classe" donne

$$\text{mpc} = \frac{\text{effectif de } C_1 \times \text{salaire "central"} s_1 + \text{effectif de } C_2 \times \text{salaire "central"} s_2 + \text{effectif de } C_3 \times \text{salaire "central"} s_3}{\text{effectif total}}$$
$$= \frac{N_1\times s_1 + N_2 \times s_2 + N_3 \times s_3}{N_1+N_2+N_3}$$

$\longrightarrow$ Pour la prochaine séance : *Finir 5.8, cours de probabilités.*