Statistiques

La **population** est l'ensemble sur lequel porte l'observation : on étudie un caractère bien précisé sur les **individus** de cette population. On collecte et on dépouille des données. Un **échantillon** est une partie de la population.
La liste des valeurs (ou **modalités**) prises par le caractère constitue la **série statistique**. Lorsque le caractère étudié prend des valeurs numériques, le caractère est **quantitatif**, sinon le caractère est **qualitatif**.
Un caractère quantitatif est **discret** lorsqu'il ne prend que quelques valeurs isolées.
Un caractère quantitatif est **continu** lorsqu'il peut prendre toutes les valeurs d'un intervalle.

L'**effectif** d'une modalité du caractère est le nombre d'individus de la population ayant cette valeur.
La **fréquence** de la modalité d'un caractère est le quotient de l'effectif de cette modalité par l'effectif total: $f = \frac{effectif}{effectif\ total}$.
L'ensemble des fréquences de toutes les modalités du caractère s'appelle la **distribution des fréquences** de la série statistique.
Dans le cas d'une série statistique à caractère quantitatif continu, on regroupe les modalités en intervalles. Ces intervalles sont appelés **classe**. On parle alors d'effectifs et de fréquence d'une classe.
Les **effectifs cumulés croissants** d'une modalité A d'une série statistique à caractère quantitatif est la somme des effectifs de toutes les modalités inférieures ou égales à A.
Les **fréquences cumulées croissantes** d'une modalité A d'une série statistique à caractère quantitatif est la somme des fréquences de toutes les modalités inférieures ou égales à A. Les **effectifs cumulés décroissants** d'une modalité A d'une série statistique à caractère quantitatif est la somme des effectifs de toutes les modalités supérieures ou égales à A. Les **fréquences cumulées décroissantes** d'une modalité A d'une série statistique à caractère quantitatif est la somme des fréquences de toutes les modalités supérieures ou égales à A.

Il faut savoir tracer:

- des histogrammes
- des diagrammes circulaires ou semi-circulaires
- des nuages de points
- la courbe des effectifs cumulés croissants ou décroissants
- la courbe des fréquences cumulées croissantes ou décroissantes

Pour la suite, considérons la série statistique suivante:

valeurs	x_1	x_2	**...**	x_p
Effectifs	n_1	n_2	...	n_p
Fréquences	f_1	f_2	...	f_p

Indicateurs de position

La **médiane** Me de la série statistique est un nombre qui partage la population en deux parties de telle sorte que :

- Au moins 50% des individus prennent une valeur inférieure ou égale à la médiane.
- Au moins 50% des individus prennent une valeur supérieure ou égale à la médiane.
- L'usage veut que si plusieurs valeurs sont possibles, on prenne la moyenne de ces valeurs.

Le **premier quartile** Q_1 de la série statistique est un nombre qui partage la population en deux parties de telle sorte que :

- Au moins 25% des individus prennent une valeur inférieure ou égale à Q_1.
- Au moins 75% des individus prennent une valeur supérieure ou égale à Q_1.
- L'usage veut que si plusieurs valeurs sont possibles, on prenne la plus petite de ces valeurs.

Le **troisième quartile** Q_3 de la série statistique est un nombre qui partage la population en deux parties de telle sorte que :

- Au moins 75% des individus prennent une valeur inférieure ou égale à Q_3.
- Au moins 25% des individus prennent une valeur supérieure ou égale à Q_3.
- L'usage veut que si plusieurs valeurs sont possibles, on prenne la plus petite de ces valeurs.

Pour trouver les quartiles et la médiane, il est préférable d'ordonner la série statistique.

La moyenne arithmétique de cette série statistique est le nombre $\overline{x} = \frac{\sum_{i=1}^{p} n_i \times x_i}{\sum_{i=1}^{p} n_i} = \frac{n_1 \times x_1 + n_2 \times x_2 + ... n_p \times x_p}{n_1 + n_2 + ... + n_p}$.

On a aussi $\overline{x} = \sum_{i=1}^{p} f_i \times x_i = f_1 \times x_1 + f_2 \times x_2 + \ldots f_p \times x_p$.

La moyenne est linéaire, c'est-à-dire que $\forall m \in \mathbb{R}, \forall p \in \mathbb{R}, \overline{m \times x + p} = m \times \overline{x} + p$.

Indicateurs de dispersion

L'**étendue** d'une série statistique est la différence entre la modalité la plus grande et la modalité la plus petite.

L'**écart interquartile** est le nombre $Q_3 - Q_1$.

La **Variance** d'une série statistique, notée V mesure la moyenne des carrés des écarts à la moyenne.

$V = \sum_{i=1}^{p} \left(x_i - \bar{x}\right)^2 = \left(x_1 - \bar{x}\right)^2 + \left(x_2 - \bar{x}\right)^2 + \ldots + \left(x_p - \bar{x}\right)^2$.

On a aussi $V = \left(\sum_{i=1}^{p} x_i^2\right) - \bar{x}^2$

L'**écart type** d'une série statistique, noté σ mesure la dispersion autour de la moyenne. Il est égal à la racine carrée de la varaince. $\sigma = \sqrt{V}$.

L'écart type a la même unité que les modalités de la série statistiques.

Probabilités

Une **expérience** est dite **aléatoire** lorsqu'elle a plusieurs **issues** (ou résultats) possibles et que l'on ne peut ni prévoir ni calculer laquelle de ces issues sera réalisée.

Dans une expérience aléatoire, l'**univers**, noté Ω, est l'ensemble de toutes les issues.

Un **événement** est un ensemble constitué d'issues de l'univers.

Un **événement élémentaire** est un événement constitué d'une seule issue.

L'**événement certain**, noté Ω, est l'événement constitué de toutes les issues de l'univers.

L'**événement impossible**, $\emptyset$, ne contient aucune issue. Les issues qui sont dans l'événement A ou l'événement B constituent l'événement $A \cup B$, **réunion** des événement A et B.

Les issues qui sont dans l'événement A et l'événement B constituent l'événement $A \cap B$, **intersection** des événement A et B.

L'**événement contraire** d'un événement A est constitué de toutes les issues de Ω qui ne sont pas dans l'événement A. Il est noté $\overline{A}$.

On a donc $A \cup \overline{A} = \Omega$ et $A \cap \overline{A} = \emptyset$.

Loi de probabilité

Soient Ω l'univers d'une expérience aléatoire et e_1, e_2, …, e_n les n issues de cette expérience.

A chaque issue e_i est associée un nombre réel p_i appelé probabilité.

p_i est un nombre réel positif ou nul tel que pour tout i, on a $0 \leq p_i \leq 1$ et $p_1 + p_2 + \cdots + p_n = 1$.

On dit alors qu'on a défini une loi de probabilité sur Ω.

Equiprobabilité

On dit qu'il y a équiprobabilité sur l'univers Ω lorsque toutes les issues ont la même probabilité.

On a alors $p_i = \frac{1}{n}$. On dit alors que la loi de probabilité est équirépartie.

Dans ce cas, la probabilité d'un événement A qui contient k issues est $p(A) = \frac{k}{n}$.

Propriété

- $\mathbb{P}(\emptyset) = 0$
- $\mathbb{P}(\Omega) = 1$

- Soient A et B deux événements. On a $\mathbb{P}(A \cup B) + \mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B)$.
  Ou encore $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
- Soit A un événement. On a $\mathbb{P}(A) + \mathbb{P}(\overline{A}) = 1$ ou encore $\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A)$.

Loi des grands nombres

Pour une expérience aléatoire, dans le modèle défini par une loi de probabilité, les distributions de fréquences obtenues dans des séries de tailles n se rapprochent de la loi de probabilité quand n devient grand.

Echantillonnage

Les distributions des fréquences varient d'un échantillon à l'autre pour une observation sur le même caractère : c'est ce qu'on appelle la **fluctuation d'échantillonnage**.
Même pour des échantillons de même taille, les fréquences peuvent fluctuer.

Lorsque la taille n de l'échantillon augmente, l'ampleur des fluctuations des distributions de fréquences calculées sur ces échantillons de taille n diminue et les fréquences tendent à se stabiliser vers des nombres appelés probabilité des modalités du caractère.

On appelle **intervalle de fluctuation au seuil de 95%** l'intervalle $\left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$ avec p la fréquence théorique (la probabilité) d'obtention du caractère et n la taille de l'échantillon.

Soit f la fréquence d'obtention du caractère dans un échantillon.

- Si f appartient à l'intervalle de fluctuation au seuil de 95% alors l'échantillon étudié est représentatif de la population.
- Si f n'appartient pas à l'intervalle de fluctuation au seuil de 95% alors l'échantillon étudié n'est pas représentatif de la population.