

Questions cours 3

1.

Le caractère qualitatif est le concept le plus général en statistique. Un caractère est simplement une propriété observée sur un individu (comme la couleur des yeux, l'âge, ou la taille).

- Caractères Qualitatifs : Ils décrivent des catégories ou des attributs non mesurables numériquement (ex. : "Bleu", "Femme", "Ville").
- Caractères Quantitatifs : Ils sont un cas particulier, décrivant des propriétés qui peuvent être mesurées par un nombre (ex. : 1.75 m, 20 ans).

Par conséquent, tout caractère quantitatif est un caractère, mais l'inverse n'est pas vrai. Le terme "qualitatif" est donc le terme le plus englobant.

2.

Un caractère quantitatif, étant mesurable, se divise en deux types selon l'ensemble de valeurs qu'il peut prendre :

Typ	Définition	Traitement Statistique	Représentation
Discret	Ne peut prendre que des valeurs précises et dénombrables (souvent entières).	Utilisation directe des valeurs observées.	Diagramme en barres ou bâtons.
Continu	Peut prendre toutes les valeurs possibles dans un intervalle donné.	Nécessite un regroupement en classes pour être traité.	Histogramme.

Il est crucial de les distinguer car les méthodes de traitement et de représentation ne sont pas les mêmes.

3.

Ces paramètres visent à identifier le centre d'une distribution de données.

Multiplicité des Moyennes

Il existe plusieurs types de moyennes (arithmétique, quadratique, harmonique, géométrique, mobile) car elles sont adaptées à la nature spécifique de la variable et à l'objectif recherché :

- La moyenne arithmétique est la plus courante, mais elle est très sensible aux valeurs extrêmes (atypiques). D'autres moyennes, comme la moyenne géométrique (souvent utilisée pour les taux de croissance), sont nécessaires pour des contextes mathématiques spécifiques.

Le Rôle de la Médiane

La médiane est un indicateur de position centrale particulièrement robuste.

- Elle est la valeur qui, une fois la série de données classée par ordre croissant, la partage en deux parties égales, chacune contenant 50% des observations.
- On l'appelle parfois la "moyenne du milieu".
- Son principal intérêt est d'analyser les distributions asymétriques où la moyenne arithmétique serait faussée par les valeurs extrêmes. Contrairement à la moyenne, la médiane est une valeur qui, dans la pratique, n'est pas "calculée" au sens strict, mais plutôt identifiée après le tri des données.

Le Calcul du Mode

Le mode est la modalité qui correspond à l'effectif maximal ; c'est la valeur la plus fréquente ou celle qui a la plus forte densité de probabilité. Il s'agit d'une moyenne de fréquence.

- Le mode n'existe pas toujours et, s'il existe, il n'est pas toujours unique (on parle alors de distribution bimodale ou multimodale).
- Il est le seul paramètre de position applicable sans ambiguïté aux données qualitatives nominales.

4.

Ces indicateurs mesurent la manière dont la masse totale d'un caractère est répartie parmi les individus.

Intérêt de la Médiale

La médiale est un concept proche de la médiane, mais elle ne divise pas l'effectif total, elle divise la masse totale de la variable en deux parties égales (chacune représentant 50% des valeurs globales).

- Si l'on étudie les revenus, la médiale est le revenu qui sépare l'ensemble des individus en deux groupes : ceux qui détiennent les 50% des revenus les plus faibles et ceux qui détiennent les 50% des revenus les plus élevés.
- En comparant la médiale et la médiane, on obtient une mesure de la concentration : plus elles sont éloignées, plus la concentration est forte.

L'Indice et la Courbe de Gini

La courbe de Gini a pour but de décrire les effets de la concentration d'une population statistique (souvent pour les revenus ou les richesses).

- Elle se construit sur un repère à partir des fréquences cumulées relatives (en abscisse, représentant les individus) et des fréquences cumulées globales relatives (en ordonnée,

représentant la masse du caractère, ex. : le revenu).

- Shutterstock
- L'Indice de Gini est le résultat de la mesure de l'aire entre cette courbe et la diagonale d'égalité parfaite (où 20% des individus détiennent 20% de la masse, etc.). Plus l'indice est proche de 1, plus la concentration est forte.

5.

Ils mesurent l'étalement des données autour d'un centre (souvent la moyenne).

Variance et Écart-Type

- Pourquoi la Variance ? La variance est considérée comme le meilleur indicateur de dispersion car, en tenant compte de toutes les données et en calculant la moyenne des carrés des écarts à la moyenne, elle annule l'effet des signes négatifs et donne plus de poids aux valeurs éloignées.
- Pourquoi l'Écart-Type ? L'inconvénient de la variance est qu'elle est exprimée dans l'unité du caractère au carré. L'écart-type (σ) est la racine carrée de la variance, et il retrouve l'unité de mesure originale, le rendant plus pertinent et plus parlant pour caractériser la dispersion.
 - Un petit écart-type indique que les données sont fortement regroupées autour de la moyenne (population homogène).
 - L'écart-type est essentiel car il permet d'estimer, pour une distribution normale, le pourcentage de la population appartenant à un intervalle centré sur la moyenne.

L'Étendue

L'étendue est la plus simple des mesures de dispersion : elle est la différence entre la plus grande et la plus petite valeur observée.

- Elle est facile et rapide à calculer, mais elle ne dépend que des valeurs extrêmes et ignore complètement le nombre et la répartition des valeurs intermédiaires. Elle donne seulement une idée générale de l'écart maximal dans la série.

Les Quantiles et l'Écart Interquartile

Les quantiles sont des caractéristiques de position qui permettent de partager la série ordonnée en parties d'égal effectif.

- Les plus utilisés sont les quartiles (qui partagent la série en quatre parties égales) et, bien sûr, le deuxième quartile qui est la médiane.
- L'écart interquartile (différence entre le troisième quartile Q_3 et le premier quartile Q_1) est très utile car il délimite l'intervalle qui contient exactement 50% des observations centrales de la série.

La Boîte de Dispersion (Boîte à Moustaches)

Introduite par J.W. Tukey, la boîte à moustaches (Box Plot) est une représentation graphique schématique d'un caractère quantitatif. *

Shutterstock

Elle fait apparaître visuellement les cinq principales caractéristiques d'une distribution : la valeur minimale, le premier quartile (Q_1), la médiane (Q_2), le troisième quartile (Q_3), et la valeur maximale.

- Interprétation : La boîte elle-même (de Q_1 à Q_3) représente l'écart interquartile.
 - Plus la boîte est étroite, moins les 50% des données centrales sont dispersées.
 - La position du trait de la médiane à l'intérieur de la boîte indique l'asymétrie de la distribution.
 - Elle est particulièrement utile pour comparer visuellement la dispersion et la position centrale de plusieurs séries statistiques.

6.

Les moments permettent de caractériser la forme globale d'une distribution (symétrie, aplatissement).

Moments Centrés vs. Moments Absolus

- Moments Centrés : Ils correspondent à la moyenne des puissances des écarts à la moyenne. L'idée est de décrire la forme de la distribution par rapport à son centre. Ils sont utilisés, entre autres, pour mesurer l'asymétrie (moment centré d'ordre 3).
- Moments Absolus : Ils renvoient à la moyenne des valeurs absolues des écarts à la moyenne. L'utilisation de la valeur absolue permet de décrire la dispersion tout en étant moins influencé par les valeurs extrêmes que les moments centrés.

Vérification de la Symétrie

Vérifier la symétrie d'une distribution est essentiel car cela influence le choix des méthodes statistiques appropriées et l'interprétation des résultats.

- Méthodes :
 - Comparaison des indicateurs : Si la Moyenne, la Médiane et le Mode sont très différents, cela indique une forte asymétrie.

- Outils Graphiques : L'observation de l'histogramme (courbe équilibrée autour de la moyenne) ou de la boîte à moustaches (médiane centrée dans la boîte et moustaches de même longueur) permet de vérifier la symétrie.
- Coefficients d'Asymétrie : Utilisation de coefficients basés sur les moments centrés (ex. : coefficient de Pearson ou de Fisher) pour une mesure quantitative.