Séance 3 questions - Analyse de données

1. Quel caractère est le plus général : le caractère quantitatif ou le caractère qualitatif ?
Justifier pourquoi.

Le caractère quantitatif est le plus général puisque les paramètres statistiques concernent principalement les variables quantitatives et seulement ponctuellement les variables qualitatives.
En effet, le caractère quantitatif permet de calculer tous les paramètres (moyenne, variance, moments…), il se prête aussi à l'ensemble des opérations statistiques vues (dispersion, forme, position). Les caractères qualitatifs ne permettent quant à eux que des descripteurs limités.

2. Que sont les caractères quantitatifs discrets et caractères quantitatifs continus ? Pourquoi les distinguer ?

Les caractères quantitatifs discrets correspondent aux valeurs dénombrables, séparées (comme le nombre d'enfants), leur moyenne se calcule par une somme

Les caractères quantitatif continus correspondent eux aux valeurs sur un intervalle (longueur, revenu)
Les paramètres deviennent des intégrales

On les distingue d'abord car les les formules ne sont pas les mêmes (somme vs intégrale), aussi parce que les représentations changent (histogrammes continus, classes) et enfin parce que certaines mesures comme les médianes ou les quantiles se calculent différemment pour les deux types.

3. Paramètres de position
   - Pourquoi existe-t-il plusieurs types de moyenne?

Le tableau du cours montre plusieurs moyennes (arithmétique, géométrique, quadratique, harmonique, mobile…) Il en existe plusieurs types afin qu'elles répondent aux situations différentes : la nature de la variable (continue ou discrète), les propriétés voulues et les contextes d'usage (vitesse –> harmonique, produits –> géométrique)

- Pourquoi calculer une médiane ?

On calcule une médiane car contrairement à la moyenne elle n'est pas influencée par les valeurs extrêmes, elle convient aussi à des séries très dissymétriques, engin elle résumer la position centrale même quand la moyenne est trompeuse. On la calcule donc pour obtenir une mesure robuste et insensible aux valeurs aberrantes.

- Quand est-il possible de calculer un mode?

On calcule un mode uniquement lorsque la distribution présente une valeur dominante identifiable. En effet, le mode existe lorsqu'une modalité a l'effectif maximal ou la plus grande densité. Il peut manquer ou être « non unique » (cas des séries pluri-modales) et il dépend du regroupement en classes pour les variables continues

4. Paramètres de concentration

- Quel est l'intérêt de la médiale et de l'indice de C. Gini?

La médiale partage la masse totale en deux parties égales (50 % – 50 %) , elle est toujours plus grande que la médiane et elle permet d'évaluer l'inégalité de distribution d'un caractère.

L'indice de Gini mesure la concentration d'un caractère dans la population, il montre si une petite proportion d'individus concentre une grande part de la masse totale
Il sert donc à mesurer l'inégalité (revenus, tailles, surfaces…).

5. Paramètres de dispersion

- Pourquoi calculer une variance à la place de l'écart à la moyenne? Pourquoi la remplacer par l'écart type?

La variance utilise les carrés ce qui donne des propriétés mathématiques utiles que n'a pas la valeur absolue. L'écart type correspond simplement la racine de la variance : il revient à l'unité de l'origine et est donc plus interprétable. Ainsi la variance comprend plus de rigueur mathématique tandis que l'écart type correspond à une interprétation pratique.

- Pourquoi calculer l'étendue ?

On calcule l'étendue car elle est simple à obtenir (maximum – minimum) et parce qu'elle donne une première idée de la dispersion. Toutefois sa fiabilité reste faible surtout pour les grands effectifs puisqu'elle ne repose que sur les extrême

- À quoi sert-il de créer un quantile ? Quel(s) est (sont) le(s) quantile(s) le(s) plus utilisé(s) ?

Les quantiles divisent une série en parties égales, ils permettent d'étudier la répartition interne des valeurs et de construire des indicateurs robustes.

- Pourquoi construire une boîte de dispersion ? Comment l'interpréter ?

La boîte de dispersion permet de visualiser rapidement à la fois la médiane, les quartiles, l'étendue, les valeurs extrêmes et aussi de comparer facilement plusieurs distributions

On peut l'interpréter ainsi :
Le rectangle : 50 % des valeurs
Ligne interne : médiane ;
Moustaches : valeurs minimum et maximum

—> Elle résume donc à la fois la position, la dispersion et l'asymétrie

6. Paramètres de forme

- Quelle différence faites-vous entre les moments centrés et les moments absolus ? Pourquoi les utiliser ?

Les moments centrés correspondent aux moments calculés par rapport à la moyenne, ils servent à mesurer la variance, l'asymétrie et l'aplatissement
Les moments absolus utilisent la valeur absolue et moins influencé par les valeurs très grandes ou très petites

- Pourquoi les utiliser ?

Pour caractériser la forme de la distribution : symétrie, aplatissement, dissymétrie.

- Pourquoi vérifier la symétrie d'une distribution et comment faire ?

On vérifie la symétrie d'une distribution puisque si une distribution est symétrique elle simplifie l'analyse, la moyenne, la médiane et le mode coïncident dans ce cas et aussi parce que les choix les choix statistiques (tests, modèles) dépendent de la symétrie

On peut utiliser le coefficient d'asymétrie beta 1 (cf formule)
Beta 1 > 0 –> queue à droite
Beta 1 < 0 –> queue à gauche
Beta 1 = 0 –> symétrie .
Ou sinon la comparaison des paramètres :
mode ≈ médiane ≈ moyenne –> distribution symétrique.