Séance 5

Échantillonner consiste à prélever dans une population mère une partie de celle-ci au hasard avec une taille *n* fixée.
On n'utilise pas toute la population soit parce que cette population est trop grande (un nombre trop grand d'individus) donc impossible à étudier soit parce que le coût de l'analyse de cette population est trop élevé.
Il existe deux types de méthodes d'échantillonnage : Les méthodes probabilistes ou aléatoires et les méthodes empiriques (ou non probabilistes).
Il existe 2 méthodes d'échantillonnage : les méthodes aléatoires (c'est-à-dire qui fait appel à un tirage au sort) et les méthodes non aléatoires.
On peut rajouter une potentielle troisième méthode qui n'en est pas une mais qui peut être ajoutée quand même.
La première méthode nécessite de disposer d'une base de sondage fiable (Une liste sans omission ni répétition de tous les individus (avec leur adresse) constituant la population parente et numérotée de 1 à N).
"On effectue alors le tirage au sort de numéros de 1 à N Avec des dés, une roue de loterie, une table de nombres au hasard, la fonction random d'une calculatrice ou d'un logiciel statistique. Les numéros tirés au hasard désignent les individus composant l'échantillon."
Il existe normalement 2 manières de sélectionner les individus au hasard :

- Le tirage avec remise implique que l'on tire un numéro, que l'on note et que l'on ne le raye pas de la liste. On parle aussi d'un échantillonnage non exhaustif.
- Le tirage sans remise suppose le même déroulement des opérations mais un numéro déjà tiré est rayé de la liste et ne peut plus apparaître dans les tirages ultérieurs. On parle aussi d'un échantillonnage exhaustif.

L'estimateur concerne la variable aléatoire. Un estimateur est une fonction des données . Il est construit de telle façon que sa valeur soit proche de la vraie valeur du paramètre. Le but de la théorie de l'estimation est de choisir, parmi toutes les statistiques possibles, celle qui donnera le meilleur estimateur, c'est-à-dire celui qui donnera une estimation ponctuelle la plus proche possible du paramètre, et ceci quelque soit l'échantillon. Le processus s'appelle l'estimation. C'est le processus pour estimer une probabilité qu'un événement arrive.

On utilise un intervalle de fluctuation lorsque la proportion p dans la population est connue ou si l'on fait une hypothèse sur sa valeur (prise de décision à partir d'un échantillon).
On utilise un intervalle de confiance lorsque l'on veut estimer une proportion inconnue p dans une population à partir de la fréquence f observée dans un échantillon (estimation, par exemple dans le cadre d'un sondage).

Un biais correspond à la différence entre l'espérance de l'estimateur et la valeur à estimer dans la population; on l'appelle également erreur d'estimation.
Il est dit sans biais si : mettre formule.
Dans le cas contraire, on dira que l'estimateur est biaisé; on parlera alors d'erreur.

Une statistique portant sur l'intégralité de la population relève de la statistique descriptive exhaustive, également appelée statistique descriptive sur population complète. Le texte met cette approche en contraste avec la statistique inférentielle, qui s'appuie sur un échantillon. Grâce à l'accroissement de la disponibilité des données, notamment dans le

contexte des données massives (big data), il devient de plus en plus courant d'analyser toute la population, diminuant ainsi la nécessité de recourir à l'inférence statistique.

Le choix d'un estimateur joue un rôle essentiel, car :

• il peut être biaisé ou non biaisé,
• sa précision peut varier en fonction de la variance,
• Il influence directement la qualité de l'inférence statistique issue d'un échantillon.

Un mauvais estimateur peut entraîner des interprétations incorrectes concernant la population étudiée.

Estimation d'un paramètre par la méthode des moindres carrés (p.43)

La méthode des moindres carrés est un principe utilisable lorsque les quantités à estimer sont des espérances.

Estimation d'un paramètre par la méthode du maximum de vraisemblance (M.V.)

La méthode du maximum de vraisemblance consiste à choisir comme estimation de la valeur θ0 qui rend f maximale. Si f est supposée deux fois dérivables,alors la valeur θ0 vérifie :

Les tests statistiques sont des outils appartenant à la statistique inférentielle, utilisés pour valider ou rejeter une hypothèse statistique, ainsi que pour déterminer si un écart observé résulte du hasard ou d'un effet réel.

Les principaux éléments constitutifs d'un test statistique comprennent :

• une hypothèse nulle (H_0),
• une hypothèse alternative (H_1),
• une statistique de test,
• et une règle de décision basée sur un seuil de risque.

Ainsi, élaborer un test consiste à formuler ces hypothèses, sélectionner une statistique de test appropriée et définir un seuil de décision.

Il existe ainsi plusieurs critiques concernant la statistique inférentielle : une forte dépendance aux hypothèses liées à l'échantillonnage et aux lois, une fragilité des résultats lorsque les conditions théoriques ne sont pas respectées, ainsi qu'une perte de pertinence dans le contexte des données massives, où il est possible d'observer l'intégralité de la population. Ces observations ne remettent pas en cause l'inférence en soi, mais soulignent qu'elle reste contextuelle et qu'il convient de l'utiliser avec précaution.