Questions de cours de la séance 5

1. Comment définir l'échantillonnage. Pourquoi ne pas utiliser la population en entier ?

Quelles sont les méthodes d'échantillonnage ? Comment les choisir ?

2. Comment définir un estimateur et une estimation ?

3. Comment distingueriez-vous l'intervalle de fluctuation et l'intervalle de confiance ?

4. Qu'est-ce qu'un biais dans la théorie de l'estimation ?

5. Comment appelle-t-on une statistique travaillant sur la population totale ? Faites le lien

avec la notion de données massives 1 ?

6. Quels sont les enjeux autour du choix d'un estimateur ?

7. Quelles sont les méthodes d'estimation d'un paramètre ? Comment en sélectionner une ?

8. Quels sont les tests statistiques existants ? À quoi servent-ils ? Comment créer un test ?

9. Que pensez-vous des critiques de la statistique inférentielle ?

1. L'échantillonnage consiste à sélectionner un sous-ensemble d'individus à partir d'une population statistique afin d'en tirer des informations sur l'ensemble. Le cours rappelle que travailler sur la population entière est souvent impossible, pour des raisons pratiques, temporelles, financières ou techniques. Même lorsque cela serait théoriquement envisageable, le coût de traitement des données peut devenir prohibitif.

Les principales méthodes d'échantillonnage présentées sont : l'échantillonnage aléatoire simple, où chaque individu a la même probabilité d'être choisi ; l'échantillonnage stratifié, qui consiste à diviser la population en sous-groupes homogènes ; l'échantillonnage systématique, fondé sur un pas régulier ; l'échantillonnage par grappes, utilisé lorsque les individus sont naturellement regroupés.

Le choix de la méthode dépend de la structure de la population, de la variable étudiée et des contraintes de l'enquête. Le cours insiste sur le fait qu'un bon échantillonnage conditionne la validité de toute inférence statistique.

2. Un estimateur est une variable aléatoire calculée à partir d'un échantillon et destinée à approcher un paramètre inconnu de la population. Une estimation est la valeur numérique prise par cet estimateur pour un échantillon donné. Le cours souligne que l'estimateur relève du raisonnement théorique, tandis que l'estimation est un résultat concret, ancré dans les données observées.

3. L'intervalle de fluctuation correspond à l'intervalle dans lequel une statistique d'échantillon est susceptible de varier lorsque l'on répète les tirages, sous une hypothèse donnée.

L'intervalle de confiance, quant à lui, est un intervalle construit à partir d'un échantillon et destiné à contenir le paramètre inconnu de la population avec une certaine probabilité.

Le cours insiste sur cette distinction fondamentale : l'un porte sur la variabilité des échantillons, l'autre sur l'incertitude concernant le paramètre.

4. Un biais correspond à l'écart systématique entre l'espérance mathématique d'un estimateur et la valeur réelle du paramètre à estimer. Un estimateur est dit sans biais lorsque, en moyenne, il restitue correctement le paramètre de la population. Le cours rappelle que le biais est une propriété théorique essentielle, car un estimateur biaisé peut conduire à des conclusions erronées, même avec de grands échantillons.

5. Une statistique calculée sur l'ensemble de la population est appelée un paramètre. Le cours établit un lien direct avec la notion de données massives : lorsque la totalité des données est accessible, la démarche inférentielle perd en partie sa raison d'être. Toutefois, même dans un contexte de données massives, se posent toujours des questions de qualité, de structure et d'interprétation des données.

6. Le choix d'un estimateur engage plusieurs enjeux : son biais, sa variance, sa précision, sa robustesse face aux valeurs extrêmes. Le cours montre que le meilleur estimateur n'est pas nécessairement celui qui est sans biais, mais celui qui réalise un compromis entre exactitude et stabilité. Ce choix dépend de l'objectif de l'analyse et du contexte des données.

7. Le cours présente notamment : la méthode des moments, qui consiste à égaler moments empiriques et moments théoriques ; la méthode du maximum de vraisemblance, qui cherche à maximiser la probabilité d'observer l'échantillon. Le choix d'une méthode dépend du modèle statistique retenu, des hypothèses formulées sur la distribution et des propriétés souhaitées pour l'estimateur.

8. Les tests statistiques servent à prendre une décision concernant une hypothèse sur la population à partir d'un échantillon. Le cours distingue notamment : les tests paramétriques, les tests non paramétriques. Créer un test implique de formuler une hypothèse nulle et une hypothèse alternative ; de choisir une statistique de test ;de fixer un seuil de signification ; de prendre une décision à partir de la loi de probabilité associée.

9.Le cours présente les critiques de la statistique inférentielle comme légitimes mais nécessaires. Elles portent notamment sur la dépendance aux hypothèses, la simplification du réel et le risque de surinterprétation des résultats. Cependant, ces critiques ne condamnent pas la statistique inférentielle : elles rappellent qu'elle doit être utilisée avec prudence, comme un outil d'aide à la décision et non comme une vérité absolue. Pour le géographe, elle reste un moyen de penser l'incertitude plutôt que de la nier.