Séance 5 questions - Analyse de données

1. Comment définir l'échantillonnage. Pourquoi ne pas utiliser la population en entier? Quelles sont les méthodes d'échantillonnage? Comment les choisir ?

Définition de l'échantillonnage
L'échantillonnage consiste à prélever un sous-ensemble d'individus (échantillon) dans une population mère, de manière aléatoire ou systématique, afin d'inférer des caractéristiques de la population à partir de cet échantillon.

Pourquoi ne pas utiliser la population en entier ?
L'étude de la population entière est souvent impossible ou trop coûteuse (taille trop grande, contraintes logistiques, financières, ou temporelles). L'échantillonnage permet d'obtenir des résultats fiables en étudiant un sous-ensemble représentatif de la population, appelé échantillon.

Méthodes d'échantillonnage
Il existe deux grandes catégories de méthodes :

Méthodes aléatoires :
- Sondage aléatoire simple (SAS) : Chaque individu a la même probabilité d'être sélectionné (équiprobabilité).
- Tirage avec ou sans remise : Avec remise, un individu peut être sélectionné plusieurs fois ; sans remise, il ne l'est qu'une fois.
- Échantillonnage systématique : Sélection d'individus selon un pas fixe (ex. : tous les 10e individus d'une liste).
- Méthode des quotas : L'échantillon respecte les proportions de sous-groupes connus dans la population (ex. : âge, sexe).

Méthodes non aléatoires :
- Échantillonnage par convenance : Sélection basée sur la facilité d'accès.
- Méthode Monte Carlo : Utilisation de simulations aléatoires pour estimer des paramètres.

Comment choisir une méthode ?

Le choix dépend de :
- L'objectif de l'étude : Précision, représentativité, ou rapidité.
- La disponibilité d'une base de sondage : Liste exhaustive des individus de la population.
- Les contraintes pratiques : Coût, temps, accessibilité.
- La taille de l'échantillon : Un échantillon représentatif et aléatoire est préférable à un grand échantillon biaisé.

2. Comment définir un estimateur et une estimation ?

Estimateur
Un estimateur est une variable aléatoire (fonction des données de l'échantillon) utilisée pour estimer un paramètre inconnu d'une population (ex. : moyenne, variance). Par exemple, la

moyenne de l'échantillon, notée $\bar{X}$, est un estimateur de la moyenne μ de la population.

Estimation
Une estimation est la valeur numérique obtenue en appliquant l'estimateur à un échantillon spécifique. Par exemple, si $\bar{X}=5$ pour un échantillon, alors 5 est une estimation de μ.

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

Intervalle de fluctuation
Définition : Intervalle qui encadre la fréquence observée dans un échantillon, en supposant que la proportion théorique *p* dans la population est connue.
Utilité : Évaluer si la fréquence observée est compatible avec *p*.
Exemple : Pour un sondage, si $p=0,3$ et $n=50$, l'intervalle de fluctuation à 95% est [0,173;0,427][0,173 ; 0,427][0,173;0,427]. Si la fréquence observée est en dehors, on remet en cause l'hypothèse sur *p*.

Intervalle de confiance
Définition : Intervalle qui encadre le paramètre inconnu (ex. : *p*, μ) avec une certaine probabilité, calculé à partir des données de l'échantillon.
Utilité : Estimer la plage de valeurs probables pour le paramètre.
Exemple : Si on observe une fréquence $f=0,4$ dans un échantillon de taille $n=100$, l'intervalle de confiance à 95% pour *p* pourrait être [0,30;0,50][0,30 ; 0,50] [0,30;0,50].
Différence clé

L'intervalle de fluctuation suppose *p* connu et évalue la compatibilité de la fréquence observée.
L'intervalle de confiance suppose *p* inconnu et l'estime à partir de l'échantillon.

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

Définition
Un biais est la différence entre l'espérance de l'estimateur et la vraie valeur du paramètre :

$$\text{Biais}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

Sans biais : $\mathbb{E}(\hat{\theta}) = \theta$

Biaisé : $\mathbb{E}(\hat{\theta}) \neq \theta$

Conséquences
Un biais systématique fausse les estimations. Par exemple, un estimateur qui sous-estime toujours la moyenne est peu fiable.

5. Comment appelle-t-on une statistique travaillant sur la population totale ? Faites le lien avec la notion de données massives ?

Statistique exhaustive (recensement)
Définition : Une enquête exhaustive (ou recensement) consiste à étudier tous les individus d'une population. Elle fournit des résultats exacts, sans inférence.
Exemple : Un recensement national pour compter la population totale.

Lien avec les données massives (Big Data)
Les données massives permettent parfois d'analyser des populations entières (ex. : traces numériques, capteurs IoT), éliminant le besoin d'échantillonnage.
Avantages : Précision, absence de biais d'échantillonnage.
Défis : Coût de collecte, stockage, traitement, et respect de la vie privée.

Comparaison
Échantillonnage : Moins coûteux, mais résultats approximatifs (sous réserve de représentativité).
Données massives : Potentiellement exhaustives, mais complexes à gérer.