Chapitre 1 : Rappel sur la théorie de l'échantillonnage.

I] Définitions :

1- Une population est définie comme l'ensemble de tous les éléments d'intérêt (individus ou unités statistiques), sur lesquels on observe des caractéristiques appelées variables.

2- Un échantillon est définie comme un sous-ensemble de la population = partie de la population.

3- Recensement : observer touts les unités statistiques d'une population finie. Dans un recensement les valeurs des variables sont disponibles sur l'ensemble de la population.

Exemples : - Recensement de la population d'un village.
- Recensement des notes des étudiants obtenues à un examen.

4- Sondage : étudier les unités de l'échantillon (méthodes qui permettent de réaliser un échantillon de bonne qualité)

On veut à partir d'un échantillon déduire des information sur la population. Le problème qui se pose alors est comment choisir une partie de la population qui reproduit le plus fidèlement possible ses caractéristiques, c'est le problème d'échantillonnage (Sondage).

II] Avantages de l'échantillonnage :

- Impossibilité d'étudier toute la population lorsqu'elle est infinie
- le coût : le choix d'un échantillon est de moindre coût qu'un recensement
- le temps : la rapidité nécessaire de certaines prises de décisions empêche le recours à un recensement.

- La théorie de l'échantillonnage a pour objectif l'étude des relations qui existent entre la population mère et les échantillons issus de cette population.
- La théorie permet d'étendre et généraliser les connaissances sur les échantillons à la population globale, ainsi cette théorie permet d'estimer et évaluer les caractéristiques de la population (Tendances centrales : moyenne, quartiles, médiane, ..., ou Dispersion : Variance, ecart-type, étendu, ...) à partir des quantités correspondants estimés sur des échantillons.

La connaissance des paramètres (moyenne, Variance, ...) de la population à partir de la connaissance des quantités correspondants estimés sur des échantillons et vis versa.

C'est le justificatif de l'appélation de la statistique mathématique comme statistique inductive ou statistique inférentielle.

La Théorie de l'echantillonnage est également utile pour étudier les différences entre les échantillons (exemple : lequel est meilleur tester covid-19 à partir de la méthode PCR ou IRM).

Echantillon et nombre aléatoire :

Pour que les conclusions de la théorie de l'échantillonnage et de l'inférence statistique soient valides, les échantillons doivent être choisis représentatifs de la population, une façon de le faire et de procéder à un échantillonnage aléatoire qui garantit l'uniformité c.à.d chaque élément de la population à la même probabilité d'être inclu dans l'échantillon.

Echantillonnage avec remise et sans remise.

Elle sont issus des tirage avec remise et sans remise, un des intérêt est que si une population est finie dans laquelle on procède à un échantillonnage avec remise peut théorique-ment être considérée comme population infinie.

Exemple: on tire 10 billets d'une urne contenant 100 billets

on échantillonne à partir d'une population finie si le tirage est sans remise et elle est infinie si le tirage se fait avec remise

Distribution d'échantillonnage:

Prenons tous les échantillons possible de taille n tirés d'une population donnée de taille N (avec au sans remise) avec $N > n$. Pour chaque echantillon tiré, on peut calculer les caractéristiques (statistique de l'echantillons) (la moyenne, l'ecart-type, ...) qui varient avec l'échantillon on obtient ainsi une distribution dite d'échantillonnage

- Si par exemple on utilise la moyenne comme statistique, la distribution s'appelle distribution d'échantillonnage de la moyenne, De la même manière, on peut obtenir les distributions d'échantillonnage pour l'écart-type, la variance, la médiane, ...

Pour chaque distribution d'échantillonnage, on peut calculer une moyenne, un écart-type, une médiane, ...

Ainsi on peut parler de la moyenne de la distribution d'échantillonnage de la moyenne, de l'écart de la distribution d'échantillonnage de la moyenne, de la moyenne de la distribution d'échantillonnage de la variance, de l'écart de la distribution d'échantillonnage de l'écart, ...

- Si nous dénoterons l'échantillon par $(X_1, \ldots, X_n)$

où : n est la taille de l'échantillon

X_i est une variable aléatoire a la même loi (la même distribution que la v.a X) $\forall i = 1, 2, \ldots n$

X : le caractère que l'on voudrait étudier sur une certaine population.

$(x_1, \ldots, x_n)$ est appelé ensemble des valeurs observées = ensemble des réalisations (x_1 valeur de X_1, x_2 observation de la v.a X_2 ... x_n : réalisation de X_n).

<u>Définition</u> : . une statistique est une quantité que l'on peut déterminer une fois l'échantillon observé

. une statistique est une fonction en $(X_1, \ldots X_n)$ qui forment l'échantillon.

<u>exemple</u> <u>de</u> <u>statistiques</u> :

1- la moyenne empirique de l'échantillon $(X_1, \ldots, X_n)$

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

2- la variance empirique de $(X_1, \dots, X_n)$

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

3- la quasi-variance empirique = la variance empirique corrigée de l'échantillon $(X_1, \dots, X_n)$

$$S'^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

* Soient μ_p et σ_p^2 la moyenne et la variance de la population

a] distribution d'échantillonnage de la moyenne $\overline{X}$

* tirage sans remise = échantillon exhaustif

$E(\overline{X}) = \mu_{\overline{X}} = \mu_p$ → la moyenne de la distribution d'échantillonnage des moyennes $\overline{X}$.

$V(\overline{X}) = \sigma_{\overline{X}}^2 = \frac{\sigma_p^2}{n} \frac{N-n}{N-n}$ → la variance de la distribution d'échantillonnage de $\overline{X}$.

* tirage avec remise = échantillon non exhaustif.

$E(\overline{X}) = \mu_{\overline{X}} = \mu_p$

$V(\overline{X}) = \sigma_{\overline{X}}^2 = \frac{\sigma_p^2}{n}$.

b] distribution d'échantillonnage des variances S^2

* tirage sans remise

$$E(S^2) = \frac{n-1}{n} \sigma_p^2 = \mu_{S^2}$$

tirage avec remise $E(S^2) = \frac{N}{N-1} \frac{n-1}{n} \sigma_p^2$

C) distribution d'échantillonnage des proportions (Echantillon tiré d'une loi de Bernoulli de paramètre p)

prenons une population ou le caractère X (la v.a X de la population) d'une expérience à deux issus :

S : succès avec probabilité (proportion) p $(0 < p < 1)$

et

E : èchec avec probabilité $q = 1-p$ (voir la loi de Bernoulli de paramètre p)

$X \sim B(p)$. $E(X) = p$ = la moyenne de la population $= \mu_p$

$V(X) = pq$ → la variance de la population $= \sigma_p^2$

• Soit f_n = fréquence relative dans l'échantillon ou proportion.

et p = proportion dans la population.

tirage avec remise :

$$E(f_n) = \mu_{f_n} = \mu_p = p = \text{proportion}.$$

$$V(f_n) = \sigma_{f_n}^2 = \frac{\sigma_p^2}{n} = \frac{pq}{n} = \frac{p(1-p)}{n}.$$

• **tirage sans remise :**

$$E(f_n) = \mu_{f_n} = \mu_p = p$$

$$V(f_n) = \sigma_{f_n}^2 = \frac{\sigma_p^2}{n} \; \frac{N-n}{N-1}$$

D) distribution d'échantillonnage des sommes ou des différences :

On tire un échantillon de taille n_1 de la population P_1, on note μ_{S_1} et $\sigma_{S_1}^2$ la moyenne et la variance de la distribution d'échantillonnage.

de même on tire un échantillon de taille n_2 de la population P_2. on note μ_{S_2} et $\sigma^2_{S_2}$ la moyenne et la variance de sa distribution d'échantillonnage.

En procédant à toutes les combinaisons possibles des échantillons des deux population, on peut obtenir une distribution de la somme et une distribution de la différence qui ont comme moyenne et d'écart-type:

- $\mu_{S_1+S_2} = \mu_{S_1} + \mu_{S_2}$ et $\sigma^2_{S_1+S_2} = \sigma^2_{S_1} + \sigma^2_{S_2} \Rightarrow \sigma_{S_1+S_2} = \sqrt{\sigma^2_{S_1} + \sigma^2_{S_2}}$
- $\mu_{S_1-S_2} = \mu_{S_1} - \mu_{S_2}$ et $\sigma^2_{S_1-S_2} = \sigma^2_{S_1} + \sigma^2_{S_2} \Rightarrow \sigma_{S_1-S_2} = \sqrt{\sigma^2_{S_1} + \sigma^2_{S_2}}$

<u>Remarque</u>:

1- $Var(S_1 - S_2) = V(S_1) + V(S_2)$ avec S_1 et S_2 indépendants

2- Si $S_1 = \overline{X}_1$ et $S_2 = \overline{X}_2$ moyenne des échantillons des deux populations.

$$\Rightarrow \mu_{\overline{X}_1+\overline{X}_2} = \mu_{\overline{X}_1} + \mu_{\overline{X}_2} = \mu_{P_1} + \mu_{P_2}$$

$$\sigma^2_{\overline{X}_1+\overline{X}_2} = \sigma^2_{\overline{X}_1} + \sigma^2_{\overline{X}_2} \quad \text{ou} \quad \begin{cases} \sigma^2_{\overline{X}_1} = \dfrac{\sigma^2_{P_1}}{n_1} \text{ et } \sigma^2_{\overline{X}_2} = \dfrac{\sigma^2_{P_2}}{n_2} & \text{si tirage A.R} \\[2ex] \sigma^2_{\overline{X}_1} = \dfrac{\sigma^2_{P_1}}{n_1} \dfrac{N_1 - n_1}{N_1 - 1} \text{ et } \sigma^2_{\overline{X}_2} = \dfrac{\sigma^2_{P_2}}{n_2} \dfrac{N_2 - n_2}{N_2 - 1} & \text{si tirage S.R} \end{cases}$$

- La même formule s'applique pour la somme et la différence de deux proportions, remplacer $\mu_{\overline{X}_1}$ par $\mu_{f_{n_1}}$

et $\sigma^2_{\overline{X}_n}$ par $\sigma^2_{\delta_{n_1}}$

Théorème central limite (T.C.L)

$(X_i)_{i \in \mathbb{N}}$ une suite de v.a.i.i.d (variable aléatoire indépendante de même loi de probabilité) ayant un moment d'ordre 2 $\Leftrightarrow E(X_i^2)$ existe.

tq: $E(X_i) = m$ et $V(X_i) = \sigma^2 \quad \forall i \in \mathbb{N}$

$$\Rightarrow \sum_{i=1}^{n} X_i \xrightarrow{\text{loi}} N(nm, n\sigma^2)$$

(la somme de X_i, $i=\overline{1,n}$ suit une loi Normale de moyenne nm et de Variance $n\sigma^2$) si $n \to \infty$

quelque soit la loi de départ (discrète ou continue) la loi de ∞ st une loi normale)

d'après le T.C.L

$$\frac{\sum X_i - nm}{\sigma\sqrt{n}} \xrightarrow{\text{loi}} N(0,1) \quad \text{(normale centrée et réduite).}$$

$$\Rightarrow \overline{X} \rightsquigarrow N\left(m, \frac{\sigma^2}{n}\right) \text{ alors } \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{\text{loi}} N(0,1)$$

Exemple :

L'approximation de $B(n,p)$ vers une normale.

On a $(X_i)_{i=1,n}$ suite de v.a.i.i.d de loi $B(p)$ $\begin{cases} E(X_i) = p \\ V(X_i) = pq \end{cases}$

$$\Rightarrow \sum_{i=1}^{n} X_i \xrightarrow{\text{loi}} N(np, npq)$$

or : $X_i \rightsquigarrow B(p)$ et X_i sont $\perp\!\!\!\perp \Rightarrow \sum X_i \rightsquigarrow B(n,p)$

donc loi exacte de $\sum X_i \rightsquigarrow B(n,p)$

loi approximative de $\sum X_i \rightsquigarrow N(np, npq)$

Soit $\phi_X(x)$ = la fonction de répartition de la v.a $X \rightsquigarrow N(0,1) \Leftrightarrow \phi_X(x) = P(X < x)$ pour $X \rightsquigarrow N(0,1)$

① - Si $\phi_X(x) = P(X < x) < 0,5$ $\Rightarrow x < 0$

et $\phi_X(x) = 1 - \phi_X(-x)$

② - $\phi_X(x) = 1 - \phi_X(-x)$
$= 1 - P(X < -x)$.