Théorie de l'échantillonnage

L'objectif de ce cour est de répondre à la problématique suivante:

Comment à partir des informations sur la population (moyenne, écart-type, La proportion peut-on prévoir celles d'un échantillon?

Pour quoi l'échantillonnage?

La taille de la population peut être très importante et le coût de l'enquête serait trop important (coût et temps);

L'accès à tous les individus de la population est matériellement impossible (complexité, population indéfinie)

Pour quoi l’échantillonnage?

Comment dénombrer ? • Question : combien y a-t-il de personnes atteintes de troubles de la vue parmi les conducteurs automobiles en Algérie ?

• Réponse : 10% ? 40 % ? 85 % ?

• Il est impossible de les compter toutes en examinant toute la population des conducteurs • Il va être nécessaire d’utiliser une procédure particulière (l’échantillonnage) et des méthodes statistiques pour estimer la précision du résultat (incertitude)

Pour quoi l'échantillonnage?

Vocabulaire

- Population : Toutes les personnes à qui les résultats doivent s'appliquer

- Echantillon : Dans la plupart des cas, la taille de la population est trop importante pour que l'on puisse

 étudier tous les individus qui la compose. On étudie un sous-groupe appelé échantillon.

- Unités : il peut s'agir d'unité individuelle (sujet) ou collective (foyer, hôpitaux)

Comment choisir un échantillon?

*Un bon échantillon doit constituer une image réduite de l'ensemble de la population (**représentatif**) dont on va étudier un caractère bien défini. Dans le cas contraire on dit que l'échantillon est **biaisé.***

L'échantillonnage désigne

L'operation destinée à sélectionner une fraction d'une population, afin de conduire des analyses.

Nous distinguerons deux cas

Tirage avec remise (non exhaustif): il y a N^n *échantillons*

Tirage sans remise (exhaustif): il y a C_N^n *échantillons*

Notons que

μ, σ et σ^2 comme les symboles de la moyenne de l'écart-type et de la variance de la population

$\mu_{\bar{X}}, \sigma_{\bar{X}}$ et $\sigma_{\bar{X}}^2$ comme les symboles de la moyenne de l'écart-type et de la variance de la distribution d'échantillonnage des moyennes

la distribution d'échantillonnage des moyennes

Prenons tous les échantillons de taille n tirés d'une population donnée. Pour chaque échantillon , on peut calculer la moyenne qui variera avec l'échantillon. Pour tous les échantillons; on obtient alors une distribution d'échantillonnage des moyennes .

Notons $\bar{X}$ *la variable aléatoire qui à chaque échantillon de taille n, associe sa moyenne*

($\bar{X}$ s'appelle la distribution d'échantillonnage des moyennes)

Notons que

Supposant tous les échantillons de tailles n constitués avec remise à partir d'une population finie de taille N ou encore les échantillons de tailles n constitués avec ou sans remise à partir d'une population infinie de taille N, on a alors:

$$\mu_{\bar{X}} = \mu, \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Notons que

Si les échantillons sont tirés sans remise à partir d'une population finie de taille N, alors:

$$\mu_{\bar{X}} = \mu, \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

$\sqrt{\frac{N-n}{N-1}}$ *s'appelle facteur d'exhaustivité*

Exemple

Soient les notes de Biostatistique d’étudiants 10 12 13 8 16

1. Calculer la moyenne et l’écart-type de la population des notes

2. Former tous les échantillons exhaustifs da la population de taille 2.

3. Calculer l’espérance et l’écart-type de distribution d’échantillonnage de moyennes

4. Recommencer dans le cas d’un échantillon non exhaustif.

Exemple

Soient la population $\{1,2,3\}$ finie de taille N=3
Former les échantillons de tailles n=2

Calculer la moyenne et l'écart-type de la population

2. Former tous les échantillons exhaustifs da la population de taille 2.
3. Calculer l'espérance et l'écart-type de distribution d'échantillonnage de moyennes
4. Recommencer dans le cas d'un échantillon non exhaustif.

La distribution d'échantillonnage des pourcentages

Soit une population donnée. La population de réalisation (succès) d'un évènement est supposée être égale à p. On considère les échantillons de taille n extraits de cette population varie en fonction de l'échantillon choisi. Notons . Notons F *la variable aléatoire, qui à chaque échantillon extrait associe sa proportion de succès (la fréquence de réalisation de l'événement considéré)*

(F s'appelle la distribution d'échantillonnages des Fréquences)

Notons que

Supposant tous les échantillons de tailles n constitués avec remise à partir d'une population finie de taille N ou encore les échantillons de tailles n constitués avec ou sans remise à partir d'une population infinie de taille N, on a alors:

$$\mu_F = p, \sigma_F = \sqrt{\frac{pq}{n}}$$

Notons que

Si les échantillons sont tirés sans remise à partir d'une population finie de taille N, alors:

$$\mu_F = p, \sigma_F = \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}$$

$\sqrt{\frac{N-n}{N-1}}$ *s'appelle facteur d'exhaustivité*

Remarque

Pour des grandes populations le tirage sans remise s’assimile à un tirage avec remise

Exemple

Sur 4 personnes vaccinées, 2 réagissent positivement : On considère tous les échantillons de tailles 3 extraits de cette population.

1. Quelle est la distribution des fréquences d'apparition de l'événement « réaction positive »

Théorème de la limite centrale:

Soient n variables aléatoires indépendantes suivant des lois de probabilités quelconques d'espérances $E(X_i)$*et de variance* σ_i^2
Alors la loi suivie par la variable $X = \sum_{i=1}^{n} X_i$ *peut être approximée pour* n $grand$ $(n \geq 30)$ *une loi Normale* $N(\mu;\sigma)$

$$\mu = \sum_{i=1}^{n} E(X_i)\,, \qquad \sigma = \sqrt{\sum \sigma_i^2}$$

Propriétés:

1. Le théorème de la limite centrale permet d'affirmer que quelque soit la distribution de la population; la distribution d'échantillonnage des moyennes suit la loi $N(\mu; \frac{\sigma}{\sqrt{n}})$ *lorsque* n $grand$ $(n \geq 30)$ *(respectivement* $N(\mu; \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}})$

$$\mu = \sum_{i=1}^{n} E(X_i)\,, \qquad \sigma = \sqrt{\sum \sigma_i^2}$$

Propriétés:

2. Le théorème de la limite centrale permet d'affirmer que quelque soit la distribution de la population; la distribution d'échantillonnage des fréquences suit la loi

$N(\mu; \sqrt{\frac{pq}{n}})$ <u>*lorsque*</u> $n\ grand\ (n \geq 30)$

$(respectivement\ N(\mu; \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}})$

Remarque

Si la loi de survie sur la population est normale(la loi parente), alors la distribution de l'échantillonnage est normale et exactement suit une loi Normale quelque soit la valeur de n