Chapitre I : L'échantillonnage

Introduction :

L'étude de propriétés caractéristiques d'un ensemble, quand on ne dispose pas encore de données, nécessite d'examiner, d'observer des éléments de cet ensemble. La manière de recueillir ces données fait l'objet d'une théorie mathématique appelée théorie des sondages ou encore théorie de l'échantillonnage ; Cette théorie concerne l'optimisation de la collecte des données selon divers critères et répond à certaines interrogations sur la façon de procéder à cette collecte en rapport avec l'information disponible et l'effort d'échantillonnage consenti.

1. Echantillonnage :

Définition : L'échantillonnage est le procédé utilisé pour choisir un échantillon ou bien la phase qui consiste à sélectionner les individus que l'on souhaite interroger au sein de la population de base.

Prenons tous les échantillons possibles de taille **n** tirés d'une population donnée. Pour chaque échantillon, on peut calculer une statistique (moyenne, écart-type, variance, etc.…) qui variera avec l'échantillon.

Combien d'échantillons $\boldsymbol{n}$ d'éléments peuvent être choisi d'une population de $\boldsymbol{N}$ éléments ?

On distingue entre deux cas de tirages :

1. Tirage exhaustif (sans remise) : nombre d'échantillons possible est $\boldsymbol{C_N^n}$ (nombre de combinaison de $\boldsymbol{n}$ éléments parmi $\boldsymbol{N}$).
2. Tirage non exhaustif (avec remise) : nombre d'échantillons possible est $\boldsymbol{N^n}$.

2. Méthodes d'échantillonnage :

L'échantillonnage peut se faire avec ou sans remise et une population peut être considérée comme finie ou infinie. Une population finie dans laquelle on procède à un échantillonnage avec remise peut être théoriquement considérée comme infinie ; de même pour des populations finies mais de grandes tailles.

2.1. Méthodes probabilistes (Aléatoires) :

L'échantillonnage probabiliste repose sur un choix d'unités dans la population fait au hasard. Une des caractéristiques de cette méthode est que chaque unité de la population a une probabilité mesurable d'être choisie. On peut effectuer quatre types d'échantillonnage probabiliste :

2.1.1. Echantillonnage aléatoire simple :

Un échantillon aléatoire simple est un échantillon sélectionné de manière à ce que chaque échantillon possible de tailler "**n**" ait la même probabilité d'être sélectionné, On prélève dans la population des individus au hasard, tous les individus ont la même probabilité d'être prélevés, et ils le sont indépendamment les uns des autres.

2.1.2. Echantillonnage aléatoire stratifié :

On suppose que la population soit stratifiée, constituée de sous-populations homogènes, les strates. (Ex : stratification par tranche d'âge). Dans chaque strate, on fait un échantillonnage aléatoire simple, de taille proportionnelle à la taille de strate dans la population (échantillon représentatif). Les individus de la population n'ont pas tous la même probabilité d'être tirés. Le chercheur divise la population en sous-groupes distincts et homogènes (strates) à partir desquels il sélectionnera un échantillon aléatoire simple. La méthode se fait en deux étapes :

1. Choisir une variable de stratification (ex : tranche d'âge).
2. Sélectionner un échantillon aléatoire dans chaque strate.

Exemple :

Supposons que 60% des étudiants de l'école HEC sont des filles et 40% des garçons, pour former un échantillon de 120 étudiants en respectant ces strates, on devrait choisir au hasard 60% x 120 = 72 filles et 40% x 120 = 48 garçons.

2.1.3. Echantillonnage aléatoire par grappe :

On tire au hasard des grappes ou familles d'individus, et on examine tous les individus de la grappe (ex : on tire des immeubles puis on interroge tous les habitants). La méthode est d'autant meilleure que les grappes se ressemblent et que les individus d'une même grappe sont différents, contrairement aux strates.

Le chercheur divise la population en sous-groupes appelés « grappes ». Les grappes ont le même profil, la variance d'une grappe à l'autre étant faible. Il sélectionne par la suite un échantillon aléatoire de grappes et non pas un échantillon aléatoire à l'intérieur de chaque grappe.

Exemple : Les étudiants de première année Master sont répartis en 11 groupes, les groupes sont numérotés de 1 à 11. Supposons que l'on obtienne les nombres 2, 5, 7 et 10, tous les étudiants de ces 4 groupes feront partie Grappe de l'échantillon

2.1.4. Echantillonnage aléatoire systématique :

Dans certaines situations, spécialement lorsque les populations sont importantes, il est coûteux (en temps) de sélectionner un échantillon aléatoire simple en trouvant tout d'abord un nombre aléatoire et ensuite en cherchant dans la liste de la population l'élément correspondant. Une alternative de l'échantillonnage aléatoire simple est ***l'échantillonnage systématique***. Par exemple, si l'on souhaite sélectionner un échantillon de taille **50** parmi une population contenant **5000** éléments, cela revient à sélectionner un élément tous les **(5000/50) = 100** éléments de la population. Constituer un échantillon systématique dans ce cas consiste à sélectionner aléatoirement un élément

parmi les 100 premiers de la liste de la population. Les autres éléments de l'échantillon sont identifiés de la façon suivante : le second élément sélectionné correspond au 100^{e} élément qui suit le premier élément sélectionné dans la liste de la population, le troisième élément sélectionné correspond au 100^{e} élément qui suit le deuxième élément sélectionné dans la liste de la population, et ainsi de suite. En fait, l'échantillon de taille 50 est identifié en se déplaçant systématiquement dans la population et en identifiant les 100^{e}, 200^{e}, 300^{e} …etc. éléments qui suivent le premier élément choisi aléatoirement. L'échantillon de taille 50 est généralement plus facile à identifier de cette manière qu'en utilisant l'échantillonnage aléatoire simple. Puisque le premier élément sélectionné l'est aléatoirement, un échantillon systématique est généralement supposé avoir les propriétés d'un échantillon aléatoire simple, cette hypothèse est particulièrement appropriée lorsque la liste de la population est une énumération aléatoire des éléments de la population[1].

2.2. Méthodes non probabilistes (Raisonnées ou empirique) :

L'échantillonnage non probabiliste repose sur un choix arbitraire des unités, c'est l'enquêteur qui choisit les unités et non le hasard. Ces méthodes sont souvent utilisées dans certaines disciplines. En voici quelques-unes :

2.2.1. Echantillonnage par quota :

Lorsque le chercheur veut reproduire les caractéristiques d'une population (ex. âge, sexe, revenus, etc.) dans son échantillon.

2.2.2. Echantillonnage de convenance (de commodité) :

Cas où les unités d'échantillonnage sont faciles à rejoindre, disponibles et généralement facile à convaincre.

2.2.3. Echantillonnage selon le jugement :

Le chercheur juge que l'échantillon va lui permettre d'atteindre les objectifs de la recherche.

2.2.4. Echantillonnage boule de neige :

Utile dans le cas de la rareté des unités d'échantillonnage ou de l'absence d'un cadre d'échantillonnage valide. On demande à un répondant de nous référer à un autre qui présente les mêmes caractéristiques que les siennes, et ainsi de suite…

3. Distribution d'échantillonnage des moyennes :

Soit une population de taille $\boldsymbol{N}$, on désigne par $\boldsymbol{\mu}$ et $\boldsymbol{\sigma}$ la moyenne et l'écart-type de cette population respectivement. On extrait de la population une série d'échantillons de taille **n**, chacun de ces échantillons a une moyenne $\bar{x}$, les différentes moyennes obtenues ($\bar{X}_1, \bar{X}_2, \bar{X}_3, \ldots\ldots$) constituent une distribution d'échantillonnage de moyenne $\bar{X}$, on désigne par $\mu_{\bar{X}}$ et $\boldsymbol{\sigma}_{\bar{X}}$ la moyenne et l'écart-type de la distribution d'échantillonnage de la moyenne.

Propriétés :

On a :

$$\mu_{\bar{X}} = \mu$$

Si le tirage est exhaustif (sans remise) :

$$\boldsymbol{\sigma}_{\bar{X}} = \frac{\boldsymbol{\sigma}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}.$$

Dans le cas où la population est infinie ou le tirage est non exhaustif (avec remise)

$$\boldsymbol{\sigma}_{\bar{X}} = \frac{\boldsymbol{\sigma}}{\sqrt{n}}.$$

Remarques :

1. Si **n** est petit devant **N**, la distinction entre exhaustivité et non exhaustivité est sans objet car

$$\frac{N-n}{N-1} \simeq 1.$$

2. Si la taille des échantillons est assez grande (en pratique n $\geq$ 30), la distribution d'échantillonnage de la moyenne approche la distribution normale quelle que soit la distribution de la population.
3. Si la population est normalement distribuée, la distribution d'échantillonnage de la moyenne est une loi normale quelle que soit la valeur **n** de la taille des échantillons.

Exemple 1: On a une population finie de 3 éléments $\boldsymbol{P = \{1, 2, 3\}}$**.**

1. Calculer la moyenne et l'écart-type de cette population.

On va effectuer des prélèvements des échantillons de taille (n = 2), on fait le prélèvement dans les deux cas (tirage exhaustif et tirage non exhaustif.

2. Quelle est le nombre des échantillons qui peuvent être prélevés à partir de cette population ?
3. Effectuez le prélèvement de ces échantillons.
4. Etablir une distribution d'échantillonnage des moyennes.
5. Calculez la moyenne de la distribution d'échantillonnage des moyenne $\mu_{\bar{x}}$.
6. Calculez l'écart-type de la distribution d'échantillonnage des moyenne $\sigma_{\bar{x}}$.

Corrigé :

La taille de la population est $N = 3$.

1. La moyenne et la variance de l'échantillon :

La moyenne de la population μ :

$$\mu = \frac{1+2+3}{3} = 2.$$

La variance de la population σ^2 :

$$\sigma^2 = \frac{1}{N}\sum_1^3 (x_i - \mu)^2 = \frac{1}{N}\sum_1^3 x_i{}^2 - \mu^2 = \frac{1}{3}(1+2^2+3^2) - 2^2 = \frac{14}{3} - 4 = \frac{2}{3}.$$

Donc l'écart-type de la population est $\sigma=\sqrt{\frac{2}{3}}$.

Cas 1 : Tirage non exhaustif (avec remise) :

La taille de l'échantillon est $n = 2$

1 Parce que le tirage est avec remise donc le nombre des échantillons possible à être prélevés est une liste de n éléments pris parmi N éléments ça veut dire $N^n=3^2=9$.

2. Les échantillons possibles sont : {(1.1) ; (1.2) ; (1.3) ; (2.1) ; (2.2) ; (2.3) ; (3.1) ; (3.2) ; (3.3)}.

3. La distribution d'échantillonnage de moyenne :

$$\bar{X}=\{\frac{1+1}{2};\ \frac{1+2}{2};\ \frac{1+3}{2};\ \frac{2+1}{2};\ \frac{2+2}{2};\ \frac{2+3}{2};\ \frac{3+1}{2};\frac{3+2}{2};\frac{3+3}{2}\}=\{1\ ;\frac{3}{2}\ ;\ 2\ ;\ \frac{3}{2};\ 2;\frac{5}{2};\ 2;\ \frac{5}{2}\ ;3\}.$$

4. La moyenne de la distribution d'échantillonnage des moyenne μ_X.

$$\mu_X=\frac{1+\frac{3}{2}+2+\frac{3}{2}+2+\frac{5}{2}+2+\frac{5}{2}+3}{9}=\frac{10+\frac{16}{2}}{9}=2=\boldsymbol{\mu}.$$

5. La variance de l'échantillon est $\sigma^2{}_{\bar{x}}$:

$$\sigma^2{}_{\bar{x}}=\frac{1}{n}\sum_1^9(\bar{x}_i-\mu)^2=\frac{1}{n}\sum_1^9\bar{x}^2\text{-}\mu^2=\frac{1}{9}(1+\frac{9}{4}+4+\frac{9}{4}+4+\frac{25}{4}+4+\frac{25}{4}+9)\text{-}4=\frac{1}{3}$$

Donc :

$$\sigma_{\bar{x}}=\sqrt{\frac{1}{3}}.$$

D'après la formule de l'écart-type :

$$\boldsymbol{\sigma_{\bar{X}}}=\frac{\boldsymbol{\sigma}}{\sqrt{n}}=\frac{\sqrt{\frac{2}{3}}}{\sqrt{2}}=\sqrt{\frac{1}{3}}.$$

Cas 2 : Tirage exhaustif (sans remise) :

La taille de l'échantillon est $n = 2$

1. Parce que le tirage est sans remise donc le nombre des échantillons possible à être prélevés est une combinaison de n éléments pris parmi N éléments ça veut dire $C_N^n = C_3^2 = \frac{3!}{3!2!} = 3$.
2. Les échantillons possibles sont : {(1.2) ; (1.3) ; (2.3)}.
3. La distribution d'échantillonnage de moyenne est $\{\frac{3}{2} ; 2 ; \frac{5}{2}\}$
4. La moyenne de la distribution d'échantillonnage des moyenne μ_X.

$$\mu_X = \frac{\frac{3}{2} + 2 + \frac{5}{2}}{3} = 2.$$

5. La variance de l'échantillon $\sigma^2_{\bar{x}}$:

$$\sigma^2_{\bar{x}} = \frac{1}{3}(\frac{9}{4} + 4 + \frac{25}{4}) - 4 = \frac{1}{6}.$$

Donc

$$\sigma_{\bar{x}} = \sqrt{\frac{1}{6}}$$

D'après la formule de l'écart-type :

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{nN-1}} = = \frac{\sqrt{\frac{2}{3}}}{\sqrt{2}}\sqrt{\frac{3-2}{3-1}} = \frac{1}{\sqrt{6}}.$$

<u>Exemple 2</u> :

La moyenne des notes de l'épreuve de statistique de 300 étudiants est égale à 9.8 et l'écart-type est de 3.68.

Trouver la probabilité dans les deux cas (tirage exhaustif et non exhaustif) qu'un échantillon aléatoire de note de 40 étudiants extrait de l'ensemble ait une moyenne

1. Comprise entre 10 et 13.

2. Inférieure à 10

<u>Corrigé</u> :

La population : des étudiants.

La taille de la population : $N = 300$

La moyenne de la population : μ =9.8

L'écart-type de la population : σ =3.68.

La taille de l'échantillon : n=40 (le cas exhaustif est le même que le cas non exhaustif)

La moyenne de l'échantillon $\bar{X}$ ait une moyenne : $\mu_{\bar{x}}$ =9.8= μ et l'écart-type

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3.68}{\sqrt{40}} = 0.092.$$

La distribution de la moyenne de l'échantillon $\bar{X}$ est une loi normale car la taille de l'échantillon n=40 ≥ 30 de moyenne : $\mu_{\bar{x}}$ =9.8= μ et d'écart-type

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3.68}{\sqrt{40}} = 0.58 \ [\bar{X} \rightsquigarrow N(\mu_{\bar{x}}, \frac{\sigma}{\sqrt{n}})].$$

On pose Z=$\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$, alors la variable aléatoire Z, suit aussi une loi Normale centrée et réduite [Z$\rightsquigarrow N(0,1)$]

1. La probabilité qu'un échantillon de 40 étudiants ait une moyenne entre 10 et 13 est :

$$p(10 \leq \bar{X} \leq 13) = p(\bar{X} \leq 13) - P(\bar{X} \leq 10)$$

$$= p(\frac{\bar{X}-9.8}{0.58} \leq \frac{13-9.8}{0.58}) - p(\frac{\bar{X}-9.8}{0.58} \leq \frac{10-9.8}{0,58})$$

$$= p(Z \leq 5.51) - P(Z \leq 0.34)$$

$$= 0.99997 - 0.67 = 0.32997.$$

2. La probabilité qu'un échantillon de 40 étudiants ait une moyenne inférieure à 10.

$$p(\bar{X} \leq 13) = p(\frac{\bar{X}-9.8}{0.58} \leq \frac{10-9.8}{0.58}) = 0.67.$$

3. **Distribution d'échantillonnage des variances** :

Chaque échantillon de taille **n** de la population à une variance :

$$S^2 = \frac{1}{n}\sum_1^n (x_i - \bar{x})^2.$$

Ces variances sont des valeurs observées d'une variable aléatoire. On a :

$$E(S^2) = \frac{n-1}{n}\sigma^2.$$

$$V(S^2) = \frac{n-1}{n^3}((n-1)\mu^4 - (n-3)\sigma^4).$$

4. Distribution d'échantillonnage des fréquences :

La probabilité de la réalisation d'un évènement est supposée être égale à p. on considère les échantillons de taille n extraits, avec remise, d'une population de taille N. à chaque échantillon extrait correspond une fréquence f_n de réalisation de l'évènement considéré.

On a :

Si le tirage est avec remise :

$$E(f_n) = p$$

$$V(f_n) = \frac{p(1-p)}{n}.$$

Si le tirage est sans remise :

$$E(f_n) = p$$

$$V(f_n) = \frac{p(1-p)}{n}.\frac{N-n}{N-1}.$$

<u>Exemple</u>

Un fabricant de clous a déterminé que **3%** des clous produits sont défectueux. On étudie un échantillon aléatoire de **300** clous. Quelle est la probabilité que la proportion de clous défectueux dans l'échantillon soit

comprise entre **2%** et **3,5**% ?

Corrigé

La population : les clous

La probabilité p de trouver un clous défectueux est égale à 0.03.

La taille de l'échantillon : n=300

La proportion ou bien la fréquence f_n suit la loi Normale d'espérance E(f_n) = p=0.03 et de variance V(f_n)=$\frac{p(1-p)}{n}$=$\frac{0.03(1-0.03)}{300}$= 0,00000997

D'où l'on déduit l'écart-type σ_{f_n}=0.031.

Donc la probabilité que la proportion de clous défectueux dans l'échantillon soit comprise entre **2%** et **3,5**% est :

$$P(\mathbf{2\%} \leq f_n \leq \mathbf{3,5\%}) = P(f_n \leq 0.035) - P(f_n \leq 0.02)$$

$$= P\left(\frac{f_n - 0.03}{0.031} \leq \frac{0.035-0.03}{0.031}\right) - P\left(\frac{f_n-0.03}{0.031} \leq \frac{0.02-0.03}{0.031}\right)$$

$$= P(Z \leq 0.16) - P(Z \leq -0.32)$$

$$= 0.5675 - 0.3745$$

$$= 0.193.$$

4. **Distribution d'échantillonnage des différences des moyennes** :

On considère deux populations p_1 et P_2 de moyennes μ_1 et μ_2 et de variance ${\sigma_1}^2$ et ${\sigma_2}^2$. On s'intéresse à la différence $\mu_1 - \mu_2$.

On a :

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2.$$

$$\boldsymbol{\sigma^2}_{\bar{x}_1 - \bar{x}_2} = \boldsymbol{\sigma^2}_{\bar{x}_1} + \boldsymbol{\sigma^2}_{\bar{x}_2}$$

$$\boldsymbol{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\boldsymbol{\sigma^2}_{\bar{x}_1} + \boldsymbol{\sigma^2}_{\bar{x}_2}}.$$

Exemple

La résistance à la rupture du hêtre et du bouleau est respectivement de **4500 kg** et de **4000 kg** avec des écarts-type respectifs de **200 kg** et **300 kg**. Si l'on teste des échantillons de **100** bouleaux et **50** hêtres :

Calculer $\mu_{\bar{x}_1 - \bar{x}_2}$ et $\boldsymbol{\sigma}_{\bar{x}_1 - \bar{x}_2}$?

Corrigé :

La population P_1 est le hêtre.

La population P_1 est le bouleau.

La variable aléatoire : la résistance à la rupture.

La moyenne de la population P_1 est : μ_1 = 4500kg.

La moyenne de la population P_2 est : μ_2= 4000kg.

L'écart-type de la population P_1 est σ_1=200kg.

L'écart-type de la population P_2 est σ_2=300kg.

La taille de l'échantillon extrait de la population P_1 est n_1=100.

La taille de l'échantillon extrait de la population P_2 est n_2=50

$$\mu_{\bar{x}_1-\bar{x}_2}=\mu_1-\mu_2=4500-4000=500.$$

$$\boldsymbol{\sigma_{\bar{x}_1-\bar{x}_2}=\sqrt{\sigma^2_{\bar{x}_1}+\sigma^2_{\bar{x}_2}}=\sqrt{100^2+50^2}}$$

Exercice

On choisit au hasard avec remise six nombres parmi les nombres entiers de **1** à **9**, chacun de ces nombres a la même probabilité d'être choisi.

- Quel est le type d'échantillonnage dans ce cas ?
- Quel est le type de tirage dans ce cas ?
- Quel est le nombre des échantillons possible d'être prélever de la population mère ?
- Calculer la moyenne et l'écart-type de la distribution d'échantillonnage des moyennes.