Notes de cours Techniques d'échantillonnage

Pr. Mokhtar KOUKI

Université de Carthage (ESSAI)
mokhtar.kouki@essai.ucar.tn

Novembre - Décembre 2023

Contenu I

Introduction

Echantillonnage (sampling en anglais)

L'échantillonnage est le processus de sélection d'un sous ensemble d'unités statistiques appratenant à une population étudiée (cible) afin de pouvoir estimer des caractéristiques (indicateurs) sur toute la population.

Population cible

La population cible est la population objet de l'étude. Si à titre d'exemple, on veut étudier le taux d'activité, la population ciblé correspond à la population des personnes en âge d'activité. En Tunisie, la population en âge d'activité est la population agée de 15 ans et plus.

Echantillon

Un échantillon est un sous-ensemble **représentatif** de la population cible. Il doit constitue une image **fidèle**. Une propriété qui permet **d'extrapoler** ou **de généraliser** les résultats, obtenus à partir de l'échantillon, sur la population étudiée.

Indicateurs

On considère l'observation d'un caractère Y sur une population cible de taille N et sur un échantillon s de taille n. On peut considérer la mesure des indicateurs suivants sur la population et sur l'échantillon s.

Indicateur	Population (vrai)	Echantillon (estimateur)
Moyenne	$m = \frac{1}{N}\sum_{i=1}^{N} Y_i$	$\hat{m} = \frac{1}{n}\sum_{i \in s}^{n} Y_i$
Total	$T = \sum_{i=1}^{N} Y_i = N \cdot m$	$\hat{T} = N\frac{1}{n}\sum_{i \in s}^{n} Y_i = N \cdot \hat{m}$
Variance	$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - m)^2$	$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i \in s}^{n}(Y_i - \hat{m})^2$

On distingue deux types d'échantillonnage : échantillonnage probabilistes et échantillonnage non probabiliste.

Echantillonnage probabiliste / Aléatoire

- Echantillonnage aléatoire simple
- Echantillonnage systématique
- Echantillonnage aléaoire stratifié : Stratification
- Echantillonnage à probabilité inégale, Sondgae par grappe
- Echantillonnage aléatoire à plusieurs degrés

On distingue deux types d'échantillonnage : échantillonnage probabilistes et échantillonnage non probabiliste.

Echantillonnage probabiliste / Aléatoire

- Echantillonnage aléatoire simple
- Echantillonnage systématique
- Echantillonnage aléaoire stratifié : Stratification
- Echantillonnage à probabilité inégale, Sondgae par grappe
- Echantillonnage aléatoire à plusieurs degrés

Echantillonnage non probabiliste / Non Aléatoire

- Echantillonnage par Quota
- Echantillonnage par Convenance
- Echantillonnage raisonné ou discrétionnaire

Etapes à suivre

- Définition l'objectif de l'étude
- Choix de la population cible
- Adoption d'une méthode d'échantillonnage
- Détermination d'une taille "optimale" de l'échantillon
- Préparation et test du questionnaire /
- Collecte des données
- Traitement et Analyse des données

Echantillonnage alétoire simple

Pour étudier le caractère Y (i.e. Dépenses par ménage), on considère une population de taille N à partir d'un échantillon de taille n.
On consdière la variable aléatoire D_i qui vaut 1 si l'individu appartient à l'échantillon et 0 sinon :

$$D_i = \begin{cases} 1 & \text{si } i \in S \\ 0 & \text{sinon} \end{cases}$$

La variable D_i suit une loi de Bernoulli de paramètre $P = \dfrac{n}{N}$.

$$\begin{aligned} V(\overline{Y}) &= V(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n^2} V\left(\sum_{i=1}^{N} Y_i D_i\right) \\ &= \frac{1}{n^2}\left\{\sum_{i=1}^{N} Y_i^2 V(D_i) + 2\sum_{i<j}\sum_{j=1}^{N} Y_i Y_j Cov(D_i, D_j)\right\} \end{aligned}$$

$$
\begin{aligned}
V(D_i) &= \frac{n}{N}\frac{N-n}{N} = \frac{n(N-n)}{N^2} \\
Cov(D_i, D_j) &= E(D_iD_j) - E(D_i)E(D_j) = \frac{n}{N}\frac{n-1}{N-1} - \left(\frac{n}{N}\right)^2 \\
&= \frac{n}{N}\left\{\frac{n-1}{N-1} - \frac{n}{N}\right\} = -\frac{n(N-n)}{N^2(N-1)}
\end{aligned}
$$

Ce qui permet d'écrire :

$$
\begin{aligned}
V(\overline{Y}) &= \frac{1}{n^2}\left\{\sum_{i=1}^{N} Y_i^2 \frac{n}{N}\frac{N-n}{N} - 2\sum_{i<j}\sum_{j=1}^{N} Y_iY_j \frac{n(N-n)}{N^2(N-1)}\right\} \\
&= \frac{N-n}{nN^2}\left\{\sum_{i=1}^{N} Y_i^2 - 2\frac{1}{N-1}\sum_{i<j}\sum_{j=1}^{N} Y_iY_j\right\}
\end{aligned}
$$

Or, on sait que :

$$\left(\sum_{i=1}^{N} Y_i\right)^2 = \sum_{i=1}^{N} Y_i^2 + 2\sum_{i<j}\sum_{j=1}^{N} Y_i Y_j$$

Ce qui donne :

$$2\sum_{i<j}\sum_{j=1}^{N} Y_i Y_j = \left(\sum_{i=1}^{N} Y_i\right)^2 - \sum_{i=1}^{N} Y_i^2$$

Et la variance de $\overline{Y}$ peut être écrite sous la forme :

$$\begin{aligned} V(\overline{Y}) &= \frac{N-n}{nN^2}\left\{\sum_{i=1}^{N} Y_i^2 - \frac{1}{N-1}\left(\left(\sum_{i=1}^{N} Y_i\right)^2 - \sum_{i=1}^{N} Y_i^2\right)\right\} \\ &= \frac{N-n}{nN^2}\left\{\sum_{i=1}^{N} Y_i^2 \frac{N}{N-1} - \frac{N^2}{N-1} m^2\right\} \end{aligned}$$

En conclusion :

$$V(\overline{Y}) = \frac{N-n}{nN^2}\left\{\sum_{i=1}^{N} Y_i^2 - \frac{1}{N-1}\left(\left(\sum_{i=1}^{N} Y_i\right)^2 - \sum_{i=1}^{N} Y_i^2\right)\right\}$$

$$= \frac{N-n}{nN}\dot{1}N-1\sum_{i=1}^{N}(Y_i-m)^2 = (1-f)\frac{\sigma^2}{n}$$

avec

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i-m)^2, \; m = \frac{1}{N}\sum_{i=1}^{N} Y_i, \text{ et } f = \frac{n}{N}$$

f : est appelé taux de sondage

Pour la taille de l'échantillon, on peut adopter deux scénarios :

- Le scénario fondé sur le budget C On considère qu'on dispose d'un budget C pour réaliser une enquête (un sondage) et que le coût unitaire par unité statistique est égal à c. La taille de l'échantillon est définie par :

$$n = \frac{C}{c}$$

Pour la taille de l'échantillon, on peut adopter deux scénarios :

- Le scénario fondé sur le budget C On considère qu'on dispose d'un budget C pour réaliser une enquête (un sondage) et que le coût unitaire par unité statistique est égal à c. La taille de l'échantillon est définie par :
$$n = \frac{C}{c}$$
- Le scénario fondé sur un niveau de confiance $1-\alpha$ et une marge d'erreur égale à ϵ. La taille de l'échantillon est donnée par la relation :
$$P(\left|\overline{Y} - m\right| < \epsilon) = 1 - \alpha$$
ou bien
$$P(\sqrt{n}\frac{\left|\overline{Y} - m\right|}{\sigma\sqrt{1-f}} < \sqrt{n}\frac{\epsilon}{\sigma\sqrt{1-f}}) = 1 - \alpha$$

La taille "minimale" de l'échantillon est égale à :

$$n^* = \frac{\sigma^2 \cdot z_{1-\alpha/2}^2}{\epsilon^2 + \frac{1}{N}\sigma^2 \cdot z_{1-\alpha/2}^2} \tag{1}$$

où $z_{1-\alpha/2}$ est le quantile la loi normale centrée-réduite de niveau $1 - \alpha/2$.
Pour N assez grand, la taille de l'échantillon se réduit à :

$$n^* = \frac{\sigma^2 \cdot z_{1-\alpha/2}^2}{\epsilon^2} \tag{2}$$

Application : Pour $\sigma = 2$, $\alpha = 5\%$, $\epsilon = 15\%$, et $N = 2000$, $n \equiv 510$.

La taille "minimale" de l'échantillon est égale à :

$$n^* = \frac{\sigma^2 \cdot z^2_{1-\alpha/2}}{\epsilon^2 + \frac{1}{N}\sigma^2 \cdot z^2_{1-\alpha/2}} \tag{1}$$

où $z_{1-\alpha/2}$ est le quantile la loi normale centrée-réduite de niveau $1-\alpha/2$.
Pour N assez grand, la taille de l'échantillon se réduit à :

$$n^* = \frac{\sigma^2 \cdot z^2_{1-\alpha/2}}{\epsilon^2} \tag{2}$$

Application : Pour $\sigma = 2$, $\alpha = 5\%$, $\epsilon = 15\%$, et $N = 2000$, $n \equiv 510$.

Observations :

- Dans le cas d'une proportion, la variance est :

$$V(\hat{P}) = \frac{P(1-P)}{n}(1-f)$$

- Dans le cas d'un total, la variance est :

$$V(\hat{T}) = N^2 \frac{\sigma^2}{n}(1-f)$$

Exercice 1

Une étude prélimoinaire donnait une prévalence du diabète en Tunisie de l'ordre de 20%. Combien de personnes faut-il examiner pour estimer le nomnbre de diabétiques sachant un coefficient de variation de 8%.

Exercice 2

On considère 2000 plantations d'oliviers. Pour un échantillon de 100 plantations, on mesure la quantité récoltée (Q, en tonnes) d'olives. Construire un intervalle de confiance de niveau 95% pour la récolte totale, sachant que :

$$\sum_{i=1}^{100} Q_i = 4000, \sum_{i=1}^{100} Q_i^2 = 200000$$

Exercice 3

Combien de personnes faut-il interroger pour estimer le pourcentage de fumeurs dans le Grand-Tunis à 2 points de pourcentage et avec un niveau de confiance égal à 95%.

Echantillonnage par stratification

Concepts et définition

On considère que la population est partitionnée en H sous-ensembles appelée **strates** de tailles respectives $N_h, h = 1, \cdots, H$. Les indicateurs au niveau de la population peuvent être écrits sous la forme :

$$
\begin{aligned}
m &= \frac{1}{N}\sum_{i=1}^{N} Y_i = \frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{N_h} Y_{hj} = \sum_{h=1}^{H}\frac{N_h}{N} m_h \text{où } m_h = \frac{o}{N_h}\sum_{j=1}^{N_h} Y_{hj} \\
\sigma^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - m)^2 = \frac{1}{N-1}\sum_{h=1}^{H}\sum_{j=1}^{N_h}(Y_h j - m_h - (m - m_h))^2 \\
&= \frac{1}{N-1}\left[\sum_{h=1}^{H}\sum_{j=1}^{N_h}(Y_h j - m_h)^2 + \sum_{h=1}^{H} N_h (m_h - m)^2\right] \\
&= \frac{N}{N-1}\left[\sum_{h=1}^{H}\frac{N_h}{N}\cdot\frac{1}{N_h}\sum_{j=1}^{N_h}(Y_{hj} - m_h)^2 + \sum_{h=1}^{H}\frac{N_h}{N}(m_h - m)^2\right] \\
&\equiv \textit{Variance Intra} + \textit{Variance inter}
\end{aligned}
$$

L'estimateur, sans biais, de la moyenne générale est défini par :

$$\hat{m} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{m}_h \text{ avec } \hat{m}_h = \frac{1}{n_h} \sum_{j \in s_h} Y_{hj}$$

n_h est la taille de l'échantillon prélevé dans la strate h.

$$\begin{aligned}
V(\hat{m}) &= V\left(\frac{1}{N} \sum_{h=1}^{H} N_h \hat{m}_h\right) = \frac{1}{N^2} \sum_{h=1}^{H} N_h^2 \frac{N_h - n_h}{N_h} \frac{\sigma_h^2}{n_h} \\
&= \frac{1}{N^2} \sum_{h=1}^{H} N_h \frac{N_h - n_h}{n_h} \sigma_h^2 = \frac{1}{N^2} \sum_{h=1}^{H} N_h \left(\frac{N_h}{n_h} - 1 \right) \sigma_h^2 \\
&= \frac{1}{N^2} \sum_{h=1}^{H} \frac{N_h^2}{n_h} \sigma_h^2 - \frac{1}{N^2} \sum_{h=1}^{H} N_h \sigma_h^2 \\
&= \sum_{h=1}^{H} \frac{w_h^2}{n_h} \sigma_h^2 - \frac{1}{N} \sum_{h=1}^{H} w_h \sigma_h^2
\end{aligned}$$

avec $w_h = \frac{N_h}{N}$

Allocation proportionnelle

On considère $N_h, h = 1, \cdots, H$ la taille des strates et n la taille de l'échantillon. Un allocation proportionnelle de l'échantillon entre les strates est définie par :

$$\frac{n_h}{n} = \frac{N_h}{N} \Rightarrow n_h = \frac{N_h}{N} n$$

et la variance de l'estimateur de la moyenne empirique est égale à :

$$V(\hat{m}) = \frac{N-n}{nN^2} \sum_{h=1}^{H} N_h \sigma_h^2 = \frac{N-n}{N} \frac{\sigma_{intra}^2}{n}$$

avec

$$\sigma_{intra}^2 = \frac{1}{N} \sum_{h=1}^{H} N_h \sigma_h^2$$

On suppose que N_h est assez grand tel que $\frac{N-1}{N} \equiv 1$

Allocation optimale au sens de Neymann

Pour l'allocation optimale au sens de Neymann, il s'agit de résoudre le programme suivant :

$$Min_{n_h} \sum_{h=1}^{H} N_h \left(\frac{N_h}{n_h} - 1 \right) \sigma_h^2$$

$$\text{s.c.} \sum_{h=1}^{H} n_h = n$$

Le lagrangien de ce programme est :

$$L = \sum_{h=1}^{H} N_h \left(\frac{N_h}{n_h} - 1 \right) \sigma_h^2 - \lambda(n - \sum_{h=1}^{H} n_h)$$

Les conditions nécessaires de premiers ordres sont définies par :

$$\frac{N_h^2 \sigma_h^2}{n_h^2} = \lambda, h = 1, \cdots, H$$

$$\sum_{h=1}^{H} n_h = n$$

Ce qui permet d'écrite :

$$n_h = \frac{N_h \sigma_h}{\sqrt{\lambda}}$$

$$\sum_{l=1}^{H} N_l \sigma_l = n\sqrt{\lambda}$$

Et la taille optimale de chaque strate est donnée par la relation suivante :

$$n_h^* = \frac{N_h \sigma_h}{\sum_{l=1}^{H} N_l \sigma_l} n$$

Allocation optimale prenant en compte le budget

Pour l'allocation optimale prenant en compte le budget, il s'agit de résoudre le programme suivant :

$$Min_{n_h} \sum_{h=1}^{H} N_h \left(\frac{N_h}{n_h} - 1 \right) \sigma_h^2$$

$$\text{s.c.} \sum_{h=1}^{H} c_h n_h = C$$

avec c_h le coût unitaire dans la strate h et C correspond au budget total. Le lagrangien de ce programme est :

$$L = \sum_{h=1}^{H} N_h \left(\frac{N_h}{n_h} - 1 \right) \sigma_h^2 - \lambda (C - \sum_{h=1}^{H} c_h n_h)$$

Les conditions nécessaires de premiers ordres sont définies par :

$$\frac{N_h^2 \sigma_h^2}{n_h^2} = \lambda c_h, h = 1, \cdots, H$$

$$\sum^{H} c_h n_h = C$$

Ce qui permet d'écrite :

$$n_h = \frac{N_h \sigma_h}{\sqrt{\lambda c_h}}$$

$$\sum_{l=1}^{H} c_l \frac{N_l \sigma_l}{\sqrt{c_l}} = C\sqrt{\lambda}$$

Et la taille optimale de chaque strate est donnée par la relation suivante :

$$n_h^* = \frac{N_h \sigma_h \sqrt{c_h}}{\sum_{l=1}^{H} N_l \sigma_l \sqrt{c_l}} \frac{C}{c_h}$$

Cas particuliers

- Si $c_h = c \forall h$

$$n_h^* = \frac{N_h \sigma_h}{\sum_{l=1}^{H} N_l \sigma_l} n, n = \frac{C}{c}$$

- Si $c_h = c \forall h$ et $\sigma_h = \sigma \forall h$

$$n_h^* = \frac{N_h}{\sum_{l=1}^{H} N_l} n = \frac{N_h}{N} n, n = \frac{C}{c}$$

Allocation optimale pour une taille minimale de l'echantillon

Considérant une valeur V_0 pour la variance de $\hat{m}$ et $a_h = \frac{n_h}{n}$. La taille de l'échantillon peut être exprimée en fonction des a_h et de V_0.

$$\begin{aligned} V_0 &= \frac{1}{N^2}\sum_{h=1}^{H} N_h\left(\frac{N_h}{n_h}-1\right)\sigma_h^2 \\ &= \frac{1}{N^2}\left[\sum_{h=1}^{H}\frac{N_h^2}{a_h n}\sigma_h^2 - \sum_{h=1}^{H} N_h\sigma_h^2\right] \\ &= \frac{1}{n}\sum_{h=1}^{H}\frac{w_h^2}{a_h}\sigma_h^2 - \frac{1}{N^2}\sum_{h=1}^{H} w_h\sigma_h^2 \end{aligned}$$

Ce qui permet d'écrire n sous la forme :

$$n = \frac{\sum_{h=1}^{H}\frac{w_h^2}{a_h}\sigma_h^2}{V_0 + \frac{1}{N^2}\sum_{h=1}^{H} w_h\sigma_h^2}$$

Pour l'allocation optimale minimisant la taille de l'échantillon, il s'agit de résoudre le programme suivant :

$$Min_{a_h} \sum_{h=1}^{H} \frac{w_h^2}{a_h} \sigma_h^2$$

$$\text{s.c.} \sum_{h=1}^{H} a_h = 1$$

Le lagrangien de ce programme est :

$$L = \sum_{h=1}^{H} \frac{w_h^2}{a_h} \sigma_h^2 - \lambda(1 - \sum_{h=1}^{H} a_h)$$

La résolution aboutit aux valeurs de a_h :

$$a_h^* = \frac{w_h \sigma_h}{\sum_{l=1}^{H} w_l \sigma_l} = \frac{N_h \sigma_h}{\sum_{l=1}^{H} N_l \sigma_l}$$

et

$$n^* = \frac{\left(\sum_{h=1}^{H} w_h \sigma_h\right)^2}{V_0 + \frac{1}{N} \sum_{h=1}^{H} w_h \sigma_h^2} \tag{3}$$

Exercice 1

On considère une population de 5 unités statistiques pour lesquelles on observe la variable Y. Les valeurs enregistrées sont 10, 12, 15, 9, 14. On prélève un échantillon (sans remise) de 2 unités.

- Calculer la moyenne et la variance de la population
- Pour tous les échantillons possibles, calculer la moyenne arithmétique ($\overline{m} = \overline{Y}$)
- Calculer l'espérance mathématique de $\hat{m}$ ($E(\hat{m})$). Conclure

Exercice 2

Pour une population divisée en 4 strates de tailles respectives $N_1 = 180$, $N_2 = 90$, $N_3 = 130$ et $N_4 = 200$. Les écarts types, au sein de chaque strate, sont respectivement égaux à 100, 130, 160 et 220.

- Déterminer l'allocation optimale au sens de Neymann pour un échantillon de 60 unités.
- Déterminer l'allocation optimale pour un budget total de 300 DT, sachant que les coûts unitaires sont respectivement 2 DT, 1 DT, 3 DT et 4 DT.
- Déterminer l'allocation correspondant à une taille minimale de l'échantillon et pour une variance de la moyenne égale à 200.
- Comparer les stratégies de sondage en terme de précision.

Exercice 3 (Ardilly & Tillé(1994))

On veut estimer un chiffre d'affaires moyen relatif à une population d'entreprises. Les entreprises sont répertoriées en 3 classes. Les informations sont résumées dans le tableau suivant :

Chiffre d'affaires en millions d'euros	Nombre d'entreprises
inférieur à 1	1000
De 1 à 10	100
De 10 à 100	10

On suppose que le chiffre d'affaires suit une loi uniforme (dans chaque classe) et on prélève un échantillon de 111 entreprises. Calculer la variance de l'estimateur de la moyenne du chiffre d'affaires pour une allocation proportionnelle et pour une allocation optimale au sens de Neymann.

Exercice 4 (Inspiré de Ardilly & Tillé(1994))

Dans une population de grande taille, on cherche à estimer l'âge moyen (m). On fait une enquête auprès d'un échatillon de taille 100. On considère que les taux de sondage sont négligeables. Le tableau suivant fournit les informations disponibles :

Strate	$\frac{N_h}{N}$	$\hat{m}_h$	σ_h^2	n_h	c_h
Moins de 40 ans	50%	25	16	40	1
De 40 à 50 ans	30%	45	9	20	2
plus de 50 ans	20%	58	25	40	4

- Calculer l'estimateur stratifié de m, noté $\hat{m}$
- Calculer la variance de $\hat{m}$
- Déterminer l'allocation proportionnelle et calculer la variance de $\hat{m}$ qui en découle
- Déterminer l'allocation au sens de Neymann et l'allocation optimale pour un budget de 300 unités monéaires.
- Déterminer l'allocation optimale correspondant à une taille minimale de l'échantillon et pour une variance $V_0 = 0.14$.

Echantillonnage à probabilité inégales

Définition

On considère le tirage d'un échatillon de taille n à partir d'une population de tailles N. Le tirage est dit à probabilité inégale, si la probabilité d'inclusion d'un individu varie selon les unités statistiques.

$$P(i \in s) = \pi_i, i = 1, \cdots, N$$

On suppose que les probabilités d'inclusion (π_i) sont proportionnelles à une variable auxilière X et les conditions suivantes sont vérifiées :

$$\pi_i = \frac{X_i}{\sum_{i=1}^{N} X_i} n, \ \sum_{i=1}^{N} \pi_i = n \text{ (n fixe)}, \ \pi_{ij} > 0, \tag{4}$$

$$\sum_{j \neq i} \pi_{ij} = E(\sum_{j \neq i} D_i D_j) = E(D_i(n - D_i)) = (n-1)\pi_i \tag{5}$$

$$\sum_{j \neq i} \pi_i \pi_j = \pi_i(n - \pi_i) \tag{6}$$

$$\sum_{i=1}^{N} \sum_{j \neq i} \pi_{ij} = (n-1)n \tag{7}$$

En pratique, $\pi_i = \min\{1, \frac{X_i}{\sum_{i=1}^{N} X_i} n\}$

Estimateur de Horvitz-Thompson du Total

Pour un échantillon s de taille n, l'estimateur de Horvitz-Thompson du total (T) est défini par :

$$\hat{T}_{HT} = \sum_{i \in s} \frac{Y_i}{\pi_i} = \sum_{i=1}^{N} \frac{Y_i}{\pi_i} D_i \qquad (8)$$

avec D_i une variable indicatrice qui vaut 1 si l'unité i appartient à l'échantillon est 0 sinon et $P(D_i = 1) = \pi_i$. On :

$$\begin{aligned} E(\hat{T}_{HT}) &= E\left(\sum_{i=1}^{N} \frac{Y_i}{\pi_i} D_i\right) \\ &= \sum_{i=1}^{N} \frac{Y_i}{\pi_i} E(D_i) = \sum_{i=1}^{N} \frac{Y_i}{\pi_i} \pi_i = T \end{aligned}$$

$$
\begin{aligned}
V(\hat{T}_{HT}) &= V\left(\sum_{i=1}^{N} \frac{Y_i}{\pi_i} D_i\right) \\
&= \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i^2} V(D_i) + \sum_{i=1}^{N} \sum_{j \neq i} \frac{Y_i}{\pi_i} \frac{Y_j}{\pi_j} cov(D_i, D_j) \\
&= \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i} (1 - \pi_i) + 2 \sum_{i=1}^{N} \sum_{j < i} \frac{Y_i}{\pi_i} \frac{Y_j}{\pi_j} (\pi_{ij} - \pi_i \pi_j)
\end{aligned}
$$

Or,

$$
\begin{aligned}
\sum_{j \neq i} (\pi_{ij} - \pi_i \pi_j) &= E(\sum_{j \neq i} D_i D_j - D_i D_j)) = E(D_i(n - D_i) - D_i \\
&= (n-1)\pi_i - \pi_i(n - \pi_i) = -\pi_i(1 - \pi_i)
\end{aligned}
$$

Ce qui permet d'ecrire

$$
1 - \pi_i = \sum_{j \neq i} \frac{(\pi_{ij} - \pi_i \pi_j)}{-\pi_i}
$$

Ce qui permet d'écrire :

$$\begin{aligned}
\sum_{i=1}^{N} \frac{Y_i^2}{\pi_i}(1-\pi_i) &= \sum_{i=1}^{N}\sum_{j\neq i}\left(\frac{Y_i}{\pi_i}\right)^2(\pi_i\pi_j-\pi_{ij}) \\
&= \sum_{i=1}^{N}\sum_{j<i}\left\{\left(\frac{Y_i}{\pi_i}\right)^2+\left(\frac{Y_j}{\pi_j}\right)^2\right\}(\pi_i\pi_j-\pi_{ij}) \qquad (9)
\end{aligned}$$

Et la variance de l'estimateur de Horvitz-Thompson peut s'ecrire :

$$\begin{aligned}
V(\hat{T}_{HT}) &= \sum_{i=1}^{N}\sum_{j<i}\left\{\left(\frac{Y_i}{\pi_i}\right)^2+\left(\frac{Y_j}{\pi_j}\right)^2\right\}(\pi_i\pi_j-\pi_{ij}) \\
&\quad +2\sum_{i=1}^{N}\sum_{j<i}\frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j}(\pi_{ij}-\pi_i\pi_j) \\
&= \sum_{i=1}^{N}\sum_{j<i}\left\{\frac{Y_i}{\pi_i}-\frac{Y_j}{\pi_j}\right\}^2(\pi_i\pi_j-\pi_{ij}) \qquad (10)
\end{aligned}$$

avec comme condition $\pi_{ij} < \pi_i\pi_j$.

L'estimateur sans biais de la variance de l'estimateur de T est défini par :

$$\hat{V}_1(\hat{T}_{HT}) = \sum_{i=1}^{n} \frac{Y_i^2}{\pi_i^2}(1-\pi_i) + 2\sum_{i=1}^{n}\sum_{j<i} \frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j \pi_{ij}}(\pi_{ij}-\pi_i\pi_j) \tag{11}$$

$$\hat{V}_2(\hat{T}_{HT}) = \sum_{i=1}^{n}\sum_{j<i} \left\{\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right\}^2 \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}} \tag{12}$$

$$\hat{V}_3(\hat{T}_{HT}) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{Y_i}{\pi_i} - \hat{T}\right)^2 \tag{13}$$

Propriété

La taille de l'échantillon n est fixe, si et seulement si

$$\sum_{i=1}^{N}(\pi_{ij}-\pi_i\pi_j)=0$$

En effet

$$\sum_{i=1}^{N}(\pi_{ij}-\pi_i\pi_j)=\sum_{i=1}^{N}cov(D_i,Dj)=cov(\sum_{i=1}^{N}D_i,D_j)=cov(n,D_i)=0$$

Estimateur de la moyenne

L'estimateur sans biais de la moyenne de la population est défini par :

$$\hat{m}_{HT}=\frac{1}{N}\hat{T}=\frac{1}{N}\sum_{i=1}^{n}\frac{Y_i}{\pi_i}, \tag{14}$$

Sa variance est égale à :

$$V(\hat{m})=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j<i}\left\{\frac{Y_i}{\pi_i}-\frac{Y_j}{\pi_j}\right\}^2(\pi_i\pi_j-\pi_{ij}) \tag{15}$$

Estimateur généralisé de la moyenne : Estimateur de Hajek

L'estimateur généralisé de la moyenne de la population est défini comme suit :

$$\hat{m}_H = \frac{\displaystyle\sum_{i=1}^{n} \frac{Y_i}{\pi_i}}{\displaystyle\sum_{i=1}^{n} \frac{1}{\pi_i}} \tag{16}$$

$\hat{m}_H$ est un estimateur convergent de m. **Observation :** Le dénominateur

$$\hat{N} = \sum_{i=1}^{n} \frac{1}{\pi_i}$$

est un estimateur sans bias de N.

Dans le cas d'un tirage avec remise, on considère les unités distinctes de l'échantillon , $r \leq n$. A tire d'exemple :

$$\hat{m}_{HT}^{ar} = \frac{\sum_{i=1}^{r} \frac{Y_i}{\pi_i}}{\sum_{i=1}^{r} \frac{1}{\pi_i}}$$

r le nombre d'unités distinctes.

Inégalité de Jensen

Soient X une variable aléatoire réelle d'espérance mathématiqe m et de variance σ^2 et $f()$ une fonction deux-fois continuement différenciable. L'espérance de $f(X)$ est inférieure (res. supérieure) à $f(m)$) si $f()$ concave (res. convexe).

Démonstration : On considère le développement à l'ordre 2 de $f()$ au voisinage de m :

$$f(X) = f(m) + f'(m)(X - m) + \frac{1}{2}f''((m))(X - m)^2 + 0(X)$$

Et :

$$E(f(X)) \equiv f(m) + f'(m)E(X - m) + \frac{1}{2}f''(m)E(X - m)^2 = f(m) + \frac{1}{2}f''(m)\sigma^2$$

CQFD : Si $f()$ est concave (res. convexe) $f''(m) < 0$ (res. $f''(m) > 0$)

Inégalité de Jensen

Soient X une variable aléatoire réelle d'espérance mathématiqe m et de variance σ^2 et $f()$ une fonction deux-fois continuement différenciable. L'espérance de $f(X)$ est inférieure (res. supérieure) à $f(m)$) si $f()$ concave (res. convexe).

Démonstration : On considère le développement à l'ordre 2 de $f()$ au voisinage de m :

$$f(X) = f(m) + f'(m)(X-m) + \frac{1}{2}f''((m))(X-m)^2 + 0(X)$$

Et :

$$E(f(X)) \equiv f(m) + f'(m)E(X-m) + \frac{1}{2}f''(m)E(X-m)^2 = f(m) + \frac{1}{2}f''(m)\sigma^2$$

CQFD : Si $f()$ est concave (res. convexe) $f''(m) < 0$ (res. $f''(m) > 0$)

Variance approchée de $f(X)$: Si $f()$ est continuement différentiable la variance de $f(X)$ peut être apprché par :

$$V(f(X)) \equiv f'(m)^2 V(X) = f'(m)^2\sigma^2$$

Appplication : construction d'un estimateur d'un rapport

Exercice 1

Dans une population de taille 10, on tire, sans remise, un échantillon de taille 3. Les valeurs observées d'une variable X sont $X_1 = 4$,$X_2 = 8$ et $X_3 = 14.5$. Les probabilités de tirage sont respectivement $P(1,2) = 0.16$, $P(1,3) = 0.24$ et $P(2,3) = 0.34$.

a Calculer la probabilité d'inclusion de chaque unité de l'échantillon

b Calculer l'estimateur de Horvitz-Thompson du total.

c Calculer l'estimateur sans bias de cet estimateur.

d Sous l'hypothèse de la normalité des X_i, construire un intervalle de confiance de niveau 0.95% de l'espérance mathématique de X.

Exercice 2

On considère une population de taille $N = 3$ (i=1, 2 et 3). Et considère un tirage **avec remise** de 2 unités. Les données sont résumés dans le tableau suivant :

Unité	Y	P_i=Probabilité de tirage
1	25	0.3
2	12	0.2
3	32	0.5

- déterminer les échantillons possibles et les probabilités de tirage de chaque échantillon
- En déduire les probabilités d'inclusion de chaque unité statistique
- Pour chaque échantillon, claculer les estimateurs du Total, de Horvitz-Thompson et de Hajek.
- Calcuer l'espérance de l'estimateur de chaque estimateur. Conclure.

Solution

Echantillon (i,j)	$P(i,j)$	(Y_1, Y_2)	$\hat{T}_{HT}$	$\hat{T}_H$
1,1	0.09	25,25	49.020	75
2,2	0.04	12,12	33.333	36
3,3	0.25	32,32	42.667	96
1,2	0.06	25,12	82.353	52.138
2,1	0.06	12,25	82.353	52.138
1,3	0.15	25,32	91.686	83.5
3,1	0.15	32,25	91.686	83.5
2,3	0.1	12,32	76	55.459
3,2	0.1	32,12	76	55.459
		$E(\hat{T})$	69	74.588

Les probabilités d'inclusion de chaque unité statistiques sont : $\pi_1 = 0.51$, $\pi_2 = 0.36$ et $\pi_3 = 0.75$

Echantillonnage à deux degrés

Définition

On considère que la population étudiée est répartie en M unités primaires (UP), appelée **grappes**, de tailles respectives $N_j, j = 1, \cdots, M$ (i.e. Gouvernorats). Chaque unité primaire est composée d'unités secondaires (US). La valeur de la variable d'intérêt pour l'individu i de l'unité primaire j est notée Y_{ij}.

L'échantillonnage, aléatoire simple, à deux degrès consiste en un échantillon de m unités primaires et de chaque unité primaire échantillonnée on tire un échantillon de taille n_j.

Le total est défini par :

$$T = \sum_{j=1}^{M} \sum_{i=1}^{N_j} Y_{ij} = \sum_{j=1}^{M} T_j$$

avec

$$T_j = \sum_{i=1}^{N_j} Y_{ij}$$

L'estimateur du total dans chaque unité primaire est défini par :

$$\hat{T}_j = \sum_{i \in s_j} \frac{Y_{ij}}{\pi_{ij}}$$

L'estimateur du total dans chaque unité primaire est défini par :

$$\hat{T}_j = \sum_{i \in s_j} \frac{Y_{ij}}{\pi_{ij}} = \sum_{i \in s_j} \frac{Y_{ij}}{\frac{n_j}{N_j}} =$$

L'estimateur du total dans chaque unité primaire est défini par :

$$\hat{T}_j = \sum_{i \in s_j} \frac{Y_{ij}}{\pi_{ij}} = \sum_{i \in s_j} \frac{Y_{ij}}{\frac{n_j}{N_j}} = N_j \cdot \frac{1}{n_j} \sum_{i \in s_j} Y_{ij} = N_j \hat{m}_j \tag{17}$$

avec π_{ij} la probabilité d'inclusion de l'unité i apppartenant à l'UP j, s_j l'échantillon d'unités secondaires appartenant à l'unité primaire j et $\hat{m}_j$ est l'estimateur de la moyenne de l'unité primaire j.

L'estimateur du total dans chaque unité primaire est défini par :

$$\hat{T}_j = \sum_{i \in s_j} \frac{Y_{ij}}{\pi_{ij}} = \sum_{i \in s_j} \frac{Y_{ij}}{\frac{n_j}{N_j}} = N_j \cdot \frac{1}{n_j} \sum_{i \in s_j} Y_{ij} = N_j \hat{m}_j \qquad (17)$$

avec π_{ij} la probabilité d'inclusion de l'unité i apppartenant à l'UP j, s_j l'échantillon d'unités secondaires appartenant à l'unité primaire j et $\hat{m}_j$ est l'estimateur de la moyenne de l'unité primaire j.

L'estimateur sans biais du total (T) est défini par :

$$\hat{T} = \sum_{j \in s} \frac{\hat{T}_j}{\frac{m}{M}} = \frac{M}{m} \sum_{j \in s} \left(\frac{N_j}{n_j} \sum_{i \in s_j} Y_{ij} \right) = \frac{M}{m} \sum_{j \in s} N_j \hat{m}_j \qquad (18)$$

$\frac{m}{M}$ étant la probabilité d'inclusion de chaque unité primaire.

En effet,

L'estimateur du total dans chaque unité primaire est défini par :

$$\hat{T}_j = \sum_{i \in s_j} \frac{Y_{ij}}{\pi_{ij}} = \sum_{i \in s_j} \frac{Y_{ij}}{\frac{n_j}{N_j}} = N_j \cdot \frac{1}{n_j} \sum_{i \in s_j} Y_{ij} = N_j \hat{m}_j \tag{17}$$

avec π_{ij} la probabilité d'inclusion de l'unité i apppartenant à l'UP j, s_j l'échantillon d'unités secondaires appartenant à l'unité primaire j et $\hat{m}_j$ est l'estimateur de la moyenne de l'unité primaire j.

L'estimateur sans biais du total (T) est défini par :

$$\hat{T} = \sum_{j \in s} \frac{\hat{T}_j}{\frac{m}{M}} = \frac{M}{m} \sum_{j \in s} \left(\frac{N_j}{n_j} \sum_{i \in s_j} Y_{ij} \right) = \frac{M}{m} \sum_{j \in s} N_j \hat{m}_j \tag{18}$$

$\frac{m}{M}$ étant la probabilité d'inclusion de chaque unité primaire.

En effet,

$$E(\hat{T}) = E_{UP} E_{US/UP} \left(M \sum_{j \in s} \frac{\hat{T}_j}{m} \right)$$

L'estimateur du total dans chaque unité primaire est défini par :

$$\hat{T}_j = \sum_{i \in s_j} \frac{Y_{ij}}{\pi_{ij}} = \sum_{i \in s_j} \frac{Y_{ij}}{\frac{n_j}{N_j}} = N_j \cdot \frac{1}{n_j} \sum_{i \in s_j} Y_{ij} = N_j \hat{m}_j \quad (17)$$

avec π_{ij} la probabilité d'inclusion de l'unité i apppartenant à l'UP j, s_j l'échantillon d'unités secondaires appartenant à l'unité primaire j et $\hat{m}_j$ est l'estimateur de la moyenne de l'unité primaire j.

L'estimateur sans biais du total (T) est défini par :

$$\hat{T} = \sum_{j \in s} \frac{\hat{T}_j}{\frac{m}{M}} = \frac{M}{m} \sum_{j \in s} \left(\frac{N_j}{n_j} \sum_{i \in s_j} Y_{ij} \right) = \frac{M}{m} \sum_{j \in s} N_j \hat{m}_j \quad (18)$$

$\frac{m}{M}$ étant la probabilité d'inclusion de chaque unité primaire.

En effet,

$$\begin{aligned} E(\hat{T}) &= E_{UP} E_{US/UP} \left(M \sum_{j \in s} \frac{\hat{T}_j}{m} \right) \\ &= E_{UP} \left(\frac{M}{m} \sum_{j \in s} T_j \right) \end{aligned}$$

L'estimateur du total dans chaque unité primaire est défini par :

$$\hat{T}_j = \sum_{i \in s_j} \frac{Y_{ij}}{\pi_{ij}} = \sum_{i \in s_j} \frac{Y_{ij}}{\frac{n_j}{N_j}} = N_j \cdot \frac{1}{n_j} \sum_{i \in s_j} Y_{ij} = N_j \hat{m}_j \qquad (17)$$

avec π_{ij} la probabilité d'inclusion de l'unité i apppartenant à l'UP j, s_j l'échantillon d'unités secondaires appartenant à l'unité primaire j et $\hat{m}_j$ est l'estimateur de la moyenne de l'unité primaire j.
L'estimateur sans biais du total (T) est défini par :

$$\hat{T} = \sum_{j \in s} \frac{\hat{T}_j}{\frac{m}{M}} = \frac{M}{m} \sum_{j \in s} \left(\frac{N_j}{n_j} \sum_{i \in s_j} Y_{ij} \right) = \frac{M}{m} \sum_{j \in s} N_j \hat{m}_j \qquad (18)$$

$\frac{m}{M}$ étant la probabilité d'inclusion de chaque unité primaire.
En effet,

$$\begin{aligned} E(\hat{T}) &= E_{UP} E_{US/UP} \left(M \sum_{j \in s} \frac{\hat{T}_j}{m} \right) \\ &= E_{UP} \left(\frac{M}{m} \sum_{j \in s} T_j \right) = \frac{M}{m} \sum_{j=1}^{M} T_j \frac{m}{M} = \sum_{j=1}^{M} T_j = T \end{aligned}$$

D'après l'équation de la décomposition de la variance on a :

$$V(\hat{T}) = E_{UP}(V_{US/UP}(\hat{T})) + V_{UP}(E_{US/UP}(\hat{T})) \tag{19}$$

On a :

$$V_{US/UP}(\hat{T}) \quad = \quad \left(\frac{M}{m}\right)^2 \sum_{j \in s} V_{US/UP}(\hat{T}_j)$$

D'après l'équation de la décomposition de la variance on a :

$$V(\hat{T}) = E_{UP}(V_{US/UP}(\hat{T})) + V_{UP}(E_{US/UP}(\hat{T})) \qquad (19)$$

On a :

$$\begin{aligned} V_{US/UP}(\hat{T}) &= \left(\frac{M}{m}\right)^2 \sum_{j\in s} V_{US/UP}(\hat{T}_j) \\ &= \left(\frac{M}{m}\right)^2 \sum_{j\in s} N_j^2 (1 - \frac{n_j}{N_j}) \frac{\sigma_j^2}{n_j} \qquad (20) \end{aligned}$$

σ_j^2 la variance de l'unité principale j.

Et

$$E_{UP}(V_{US/UP}(\hat{T})) = \left(\frac{M}{m}\right)^2 \frac{m}{M} \sum_{j=1}^{M} N_j^2 (1 - \frac{n_j}{N_j}) \frac{\sigma_j^2}{n_j}$$

D'après l'équation de la décomposition de la variance on a :

$$V(\hat{T}) = E_{UP}(V_{US/UP}(\hat{T})) + V_{UP}(E_{US/UP}(\hat{T})) \quad (19)$$

On a :

$$\begin{aligned} V_{US/UP}(\hat{T}) &= \left(\frac{M}{m}\right)^2 \sum_{j\in s} V_{US/UP}(\hat{T}_j) \\ &= \left(\frac{M}{m}\right)^2 \sum_{j\in s} N_j^2 (1 - \frac{n_j}{N_j}) \frac{\sigma_j^2}{n_j} \end{aligned} \quad (20)$$

σ_j^2 la variance de l'unité principale j.

Et

$$\begin{aligned} E_{UP}(V_{US/UP}(\hat{T})) &= \left(\frac{M}{m}\right)^2 \frac{m}{M} \sum_{j=1}^{M} N_j^2 (1 - \frac{n_j}{N_j}) \frac{\sigma_j^2}{n_j} \\ &= \frac{M}{m} \sum_{j=1}^{M} N_j^2 (1 - \frac{n_j}{N_j}) \frac{\sigma_j^2}{n_j} \end{aligned} \quad (21)$$

$$E_{US/UP}(\hat{T}) = \frac{M}{m} \sum_{j \in s} E_{US/UP}(\hat{T}_j)$$

$$E_{US/UP}(\hat{T}) = \frac{M}{m} \sum_{j \in s} E_{US/UP}(\hat{T}_j) = \frac{M}{m} \sum_{j \in s} T_j$$

Et

$$V_{US}(E_{US/UP}(\hat{T})) = M^2 V_{US}(\frac{1}{m} \sum_{j \in s} T_j)$$

$$E_{US/UP}(\hat{T}) = \frac{M}{m}\sum_{j\in s} E_{US/UP}(\hat{T}_j) = \frac{M}{m}\sum_{j\in s} T_j$$

Et

$$V_{US}(E_{US/UP}(\hat{T})) = M^2 V_{US}(\frac{1}{m}\sum_{j\in s} T_j) = M^2(1-\frac{m}{M})\frac{\sigma_I^2}{m}$$

avec

$$\sigma_I^2 = \frac{1}{M-1}\sum_{j=1}^{M}\left(T_j - \overline{T}\right)^2, \text{ et } \overline{T} = \frac{1}{M}\sum_{j=1}^{M} T_j$$

$$E_{US/UP}(\hat{T}) = \frac{M}{m}\sum_{j\in s} E_{US/UP}(\hat{T}_j) = \frac{M}{m}\sum_{j\in s} T_j$$

Et

$$V_{US}(E_{US/UP}(\hat{T})) = M^2 V_{US}(\frac{1}{m}\sum_{j\in s} T_j) = M^2(1-\frac{m}{M})\frac{\sigma_I^2}{m}$$

avec

$$\sigma_I^2 = \frac{1}{M-1}\sum_{j=1}^{M}\left(T_j - \overline{T}\right)^2, \text{ et } \overline{T} = \frac{1}{M}\sum_{j=1}^{M} T_j$$

Conclusion

$$V(\hat{T}) = M^2\left(1-\frac{m}{M}\right)\frac{\sigma_I^2}{m} + \frac{M}{m}\sum_{j=1}^{M} N_j^2\left(1-\frac{n_j}{N_j}\right)\frac{\sigma_j^2}{n_j} \tag{22}$$

Observations

$$E\left(\sum_{j\in s} n_j\right) = \sum_{j=1}^{M} n_j \frac{m}{M} = m \cdot \overline{n}$$

$$\hat{N} = \frac{M}{m}\sum_{j\in s}\frac{N_j}{n_j}\left(\sum_{i\in s_j} 1\right) = \frac{M}{m}\sum_{j\in s}\frac{N_j}{n_j}\left(\sum_{i\in s_j} 1\right) = \frac{M}{m}\sum_{j\in s}\frac{N_j}{n_j}n_j = \frac{M}{m}\sum_{j\in s} N_j$$

Un estimateur sans biais de $V(\hat{T})$ est défini par :

$$V(\hat{T}) = M^2\left(1-\frac{m}{M}\right)\frac{\hat{\sigma}_I^2}{m} + \frac{M}{m}\sum_{j=1}^{M} N_j^2\left(1-\frac{n_j}{N_j}\right)\frac{\hat{\sigma}_j^2}{n_j} \qquad (23)$$

avec

$$\hat{\sigma}_I^2 = \frac{1}{m-1}\sum_{j\in s}\left(\hat{T}_i - \frac{\hat{T}}{M}\right)^2$$

$$\hat{\sigma}_j^2 = \frac{1}{n_j-1}\sum_{i\in s_j}(Y_{ij}-\hat{m}_j)^2 \text{ et } \hat{m}_j = \frac{1}{n_j}\sum_{i\in s_j} Y_{ij}$$

Echantillonnage en grappe

L'échantillonnage en grappe est un cas particulier de l'échantillonnage à deux degrés. Il s'agit de prendre toutes unités unités de l'échantillon des Unités Primaires.

Echantillonnage en grappe

L'échantillonnage en grappe est un cas particulier de l'échantillonnage à deux degrès. Il s'agit de prendre toutes unités unités de l'échantillon des Unités Primaires. Ainsi :

$$\hat{T} = \sum_{j \in s} \frac{T_j}{\frac{m}{M}} \tag{24}$$

$$V(\hat{T}) = M^2 \left(1 - \frac{m}{M}\right) \frac{\sigma_I^2}{m} \tag{25}$$

$$n = \sum_{j \in s} N_j \tag{26}$$

$$\hat{N} = \sum_{j \in s} \frac{N_j}{\pi_j} = \frac{M}{m} \sum_{j \in s} N_j \tag{27}$$

$$\tag{28}$$

Observation : La variance intra-grappe est nulle

Exercices

Exercice 1

On considère le nombre de clients qui fréquentent des magasins dans une journée ; classées selon la taille : Petite, Moyenne et Grande. On sait qu'il y a 1200 de petite taille, 800 de taille moyenne et 400 de grande taille. On prélève un échantillon, par sondage aléatoire simple, de 100 magasins de chaque classe et on enregistre le nombre de clients (noté Y) par magasin par jour. Les données sont résumées dans le tableau suivant :

Classe de magasin	Petite	Moyenne	Grande
$\sum_{i \in s_h} Y_{ih}$	6000	10000	6000
$\sum_{i \in s_h} (Y_{ih} - \hat{\overline{Y}}_h)^2$	3600	5200	4800

1. Calculer un estimateur du nombre moyen de clients par magasin
2. Construire un intervalle de confiance à 95% du nombre moyen de clients par magasin
3. Calculer la variance correspondant à une allocation proportionnelle et par rapport à une allocation au sens de Neyman de l'échantillon total.
4. Quels sont les gains de précision par rapport à l'allocation intiale.

Exercice 2

Quelle taille d'échantillon est nécessaire pour estimer la proportion de personnes ayant du sang du groupe O dans une population de 2000 personnes pour être à 0,02 près de la vraie proportion et avec 95% de niveau de confiance ?

Exercice 3

Pour estimer la masse salariale totale d'une population donnée, on considère la répartition de la population en 4 strates (selon le groupe d'âge). Le tableau suivant fournit des données sur les 4 strates.

Strate	N_j	$\sum_{i=1}^{N_j} Y_{ij}$	$\sum_{i=1}^{N_j} Y_{ij}^2$	$\sum_{i \in s_j} Y_{ij}$	$\sum_{i \in s_j} Y_{ij}^2$
18 - 30	50	800	14000	30	1800
30 - 45	200	3500	80000	350	9800
45 - 55	60	1400	45000	250	8000
56 et plus	20	620	25000	110	5400

1. Calculer les variances dans chaque strate
2. Déterminer une allocation optimale selon la méthode de Neyman d'un échantillon de 50.
3. Calculer un estimateur de la masse salariale en se basant sur la répartition déterminée en 2.
4. Donner une estimation de la masse salariale en adoptant un échantillonnage aléatoire simple (sans remise) de même taille. Construire un intervalle de confiance pour le total de niveau 95%.
5. Quelle est la méthode la plus efficace. Justifier.

MERCI POUR VOTRE ATTENTION

Pr. Mokhtar KOUKI

mokhtar.kouki@essai.ucar.tn