Éléments de théorie des sondages

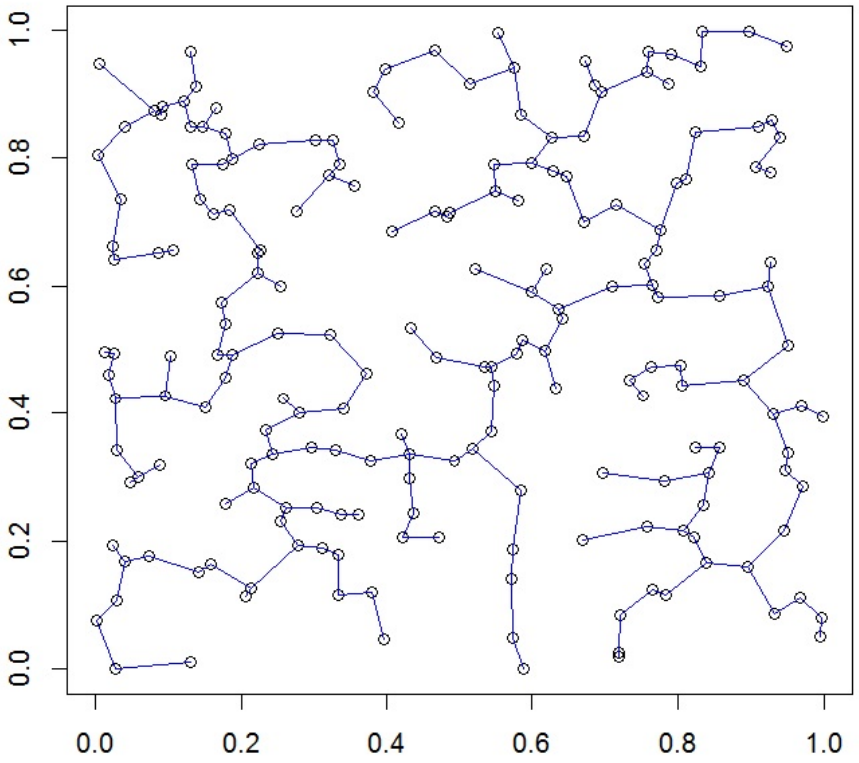
Christophe Chesneau

Université de Caen-Normandie

Éléments de théorie des sondages

Christophe Chesneau

https://chesneau.users.lmno.cnrs.fr/



Caen, le 17 Mai 2024

Table des matières

~ Note ~

Ce document résume les notions abordées dans le cours *Théorie des sondages* du Master 2 orienté statistique de l'université de Caen.

Un des objectifs est de donner des pistes de réflexion à la mise en place de sondage.

N'hésitez pas à me contacter pour tout commentaire :

christophe.chesneau@gmail.com

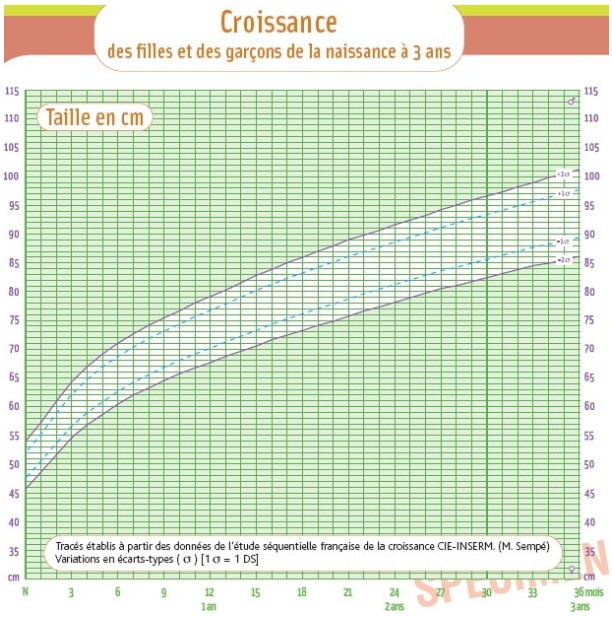
Bonne lecture !

1 Introduction

1.1 Exemples

Quelques exemples de résultats liés à des sondages sont donnés ci-dessous :

1. Le salaire moyen pour une première embauche d'un jeunes diplômé (Bac+5) titulaire d'un diplôme en sciences technologiques est de 31700€ brut.
2. 84% des français ne croient pas que leurs impôts vont baisser en 2019.
3. Parmi des amateurs de bières, la question suivante a été posée : Quel est votre type de bière préféré ? Réponses : Blondes : 33.61%, Ambrées : 25.58%, Brunes : 15.92%, Blanches : 9.64%, Un peu toutes : 15.24%
4. La prise de poids moyenne pour un individu fumeur est de
   - 2.26 kilogrammes après deux mois sans tabac,
   - 4.67 kilogrammes après un an sans tabac.
5. Les courbes de croissance des filles et des garçons de 0 à 3 ans :



1.2 Concepts de base et notations

Population et individus : On appelle population un ensemble fini d'objets sur lesquels une étude se porte. Ces objets sont appelés individus/unités statistiques. Une population est notée

$$U = \{u_1, \ldots, u_N\},$$

où N est le nombre d'individus dans la population et, pour tout $i \in \{1, \ldots, N\}$, u_i est le i-ème individu.

Base de sondage : On appelle base de sondage une liste qui répertorie tous les individus d'une population.

Caractère : Un caractère est une qualité que l'on étudie chez les individus d'une population.

Un caractère est noté Y. Pour tout $i \in \{1, \ldots, N\}$, on note y_i la valeur de Y pour l'individu u_i.

Moyenne-population :

On appelle moyenne-population le réel :

$$\overline{y}_U = \frac{1}{N} \sum_{i=1}^{N} y_i.$$

Le paramètre $\overline{y}_U$ est une valeur centrale de Y.

Écart-type corrigé-population :

On appelle écart-type corrigé-population le réel :

$$s_U = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (y_i - \overline{y}_U)^2}.$$

Le paramètre s_U mesure la dispersion de Y autour de $\overline{y}_U$.

Calcul/évaluation des paramètres-population : Pour calculer/évaluer les paramètres-population, deux méthodes sont possibles :

- *le recensement :* on a accès à tous les individus et on peut mesurer les valeurs de Y pour chacun d'entre eux. Toutefois, cela n'est pas toujours possible pour des raisons de coût, de temps ou à cause de certaines contraintes comme la destruction des individus étudiés.
- *le sondage :* on étudie les valeurs de Y sur un ensemble d'individus issus de la population.

Échantillon : On appelle échantillon un ensemble d'individus issus d'une population.

Un échantillon est noté ω. Le nombre d'individus dans un échantillon est noté n.

Deux questions centrales :

Pour constituer un échantillon représentatif de la population,

- comment faut-il procéder ?
- combien d'individus faut-il choisir ?

Plan de sondage :

On appelle plan de sondage une procédure permettant de sélectionner un échantillon dans une population. Un plan de sondage est dit :

- aléatoire si chaque individu de la population a une probabilité connue de se retrouver dans l'échantillon,
- simple si chaque individu a la même probabilité qu'un autre d'être sélectionné ; les probabilités sont égales (PE),
- sans remise (SR) si un même individu ne peut apparaître qu'une seule fois dans l'échantillon,
- avec remise (AR) si un même individu peut apparaître plusieurs fois dans l'échantillon et si l'ordre dans lequel apparaissent les individus compte.

Remarques :

- Mathématiquement, sans autre précision, un échantillon s'obtient par tirage avec remise (AR) des individus. Ainsi, un échantillon de n individus est la liste des n individus obtenus par n prélèvements indépendants. Un individu peut donc être prélevé plusieurs fois.
- Les formules habituelles d'estimation sont associées à un plan de sondage aléatoire de type PEAR (Probabilités Égales + Avec Remise). Pour simplifier la situation, elles sont généralement utilisées dans le cas SR (Sans Remise) lorsque n est beaucoup plus petit que N. Une convention existante est $N \geq 10n$.

2 Plan de sondage aléatoire simple sans remise (PESR)

2.1 Contexte

Loi de probabilité :

On prélève un échantillon de n individus suivant un plan de sondage aléatoire simple sans remise (PESR pour Probabilités Egales Sans Remise) dans une population U. Soit W la *var* égale à l'échantillon obtenu. Alors la loi de W est donnée par

$$\mathbb{P}(W = \omega) = \frac{1}{\binom{N}{n}}, \qquad \omega \in W(\Omega),$$

où $\mathbb{P}$ désigne la probabilité uniforme et $W(\Omega)$ désigne l'ensemble de tous les échantillons de n individus possibles avec un tel plan de sondage.

Explication : Pour fixer les idées, on considère la situation simplifiée suivante : on prélève au hasard et simultanément n individus de la population pour former un échantillon. L'univers associé à cette expérience aléatoire est $\Omega = \{\text{combinaisons de } n \text{ individus parmi } N\}$. Comme Ω est fini et qu'il y a équiprobabilité, l'utilisation de la probabilité uniforme $\mathbb{P}$ est justifiée. Il vient

$$\mathbb{P}(W = \omega) = \frac{\text{Card}(\{W = \omega\})}{\text{Card}(\Omega)}, \qquad \omega \in W(\Omega).$$

Or on a $\text{Card}(\Omega) = \binom{N}{n}$ et $\text{Card}(\{W = \omega\}) = 1$, d'où le résultat.

Situations de référence : Les différents types de prélèvements décrits ci-dessous rentrent dans le cadre d'un PESR :

I on prélève au hasard et simultanément n individus de la population pour former un échantillon,

II on prélève au hasard et un à un n individus de la population pour former un échantillon, l'ordre n'étant pas pris en compte.

Quelques commandes R : Pour illustrer un plan de sondage aléatoire de type PESR avec le logiciel R, on propose l'animation :

```
library(animation)
sample.simple(nrow = 10, ncol = 10, size = 15, p.col = c("blue", "red"), p.cex =
c(1, 3))
```

Par exemple, pour faire un tirage sans remise de $n = 20$ individus dans une population de $N = 200$ individus, on peut utiliser

◦ la commande `sample` :

```
sample(1:200, 20, replace = F)
```

◦ la commande `srswor` de la librairie `sampling` :

```
library(sampling)
t = srswor(20, 200)
x = 1:200
x[t != 0]
```

L'abréviation `srswor` signifie `Simple Random Sampling WithOut Replacement`.

Précisons que `t = srswor(20, 200)` renvoie un vecteur de taille 200 constitué de 20 chiffres 1 et de 180 chiffres 0. Les 1 sont positionnés aux indices des individus prélevés et les 0 aux autres.

Un autre exemple : on considère la population U constituée de $N = 9$ garçons et on prélève un échantillon de $n = 3$ individus suivant un plan de sondage aléatoire de type PESR :

```
U = c("Bob", "Nico", "Ali", "Fabien", "Malik", "John", "Jean", "Chris", "Karl")
library(sampling)
t = srswor(3, 9)
w = U[t != 0]
w
```

Dans la suite :

◦ pour les résultats, on considère un plan de sondage aléatoire de type PESR et la *var* W égale à l'échantillon obtenu,

◦ pour les preuves, pour raison de simplicité, on se place dans la situation de référence I,

◦ pour les commandes R, on utilisera dorénavant la librairie `sampling`.

Taux de sondage :

On appelle taux de sondage le réel :

$$f = \frac{n}{N}.$$

Probabilités d'appartenance :

- pour tout $i \in \{1, \dots, N\}$, la probabilité que l'individu u_i appartienne à W est

$$\mathbb{P}(u_i \in W) = \frac{n}{N} \qquad (= f).$$

- pour tout $(i, j) \in \{1, \dots, N\}^2$ avec $i \neq j$, la probabilité que les individus u_i et u_j appartiennent à W est

$$\mathbb{P}((u_i, u_j) \in W) = \frac{n(n-1)}{N(N-1)}.$$

Preuve :

- Par la définition de la probabilité uniforme, on a

$$\mathbb{P}(u_i \in W) = \frac{\text{Card}(\{u_i \in W\})}{\text{Card}(\Omega)}.$$

On a $\text{Card}(\Omega) = \binom{N}{n}$. Il reste à calculer $\text{Card}(\{u_i \in W\})$. Le nombre de possibilités pour que u_i soit dans l'échantillon est égal au nombre de possibilités de prélever $n-1$ individus parmi les $N-1$ autres que u_i. D'où $\text{Card}(\{u_i \in W\}) = \binom{N-1}{n-1}$. On en déduit que

$$\mathbb{P}(u_i \in W) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{\frac{(N-1)!}{(n-1)!((N-1)-(n-1))!}}{\frac{N!}{n!(N-n)!}} = \frac{n!}{(n-1)!}\frac{(N-1)!}{N!} = \frac{n}{N}.$$

- Avec un raisonnement similaire, on a

$$\mathbb{P}((u_i, u_j) \in W) = \frac{\text{Card}(\{(u_i, u_j) \in W\})}{\text{Card}(\Omega)}.$$

On a $\text{Card}(\Omega) = \binom{N}{n}$. Il reste à calculer $\text{Card}(\{(u_i, u_j) \in W\})$.

Le nombre de possibilités pour que u_i et u_j soient dans l'échantillon est égal au nombre de possibilités pour prélever simultanément $n-2$ individus parmi les $N-2$ autres que u_i et u_j.

D'où $\text{Card}(\{(u_i,u_j)\in W\}) = \binom{N-2}{n-2}$. On en déduit que

$$\mathbb{P}((u_i,u_j)\in W) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{\frac{(N-2)!}{(n-2)!((N-2)-(n-2))!}}{\frac{N!}{n!(N-n)!}} = \frac{n!}{(n-2)!}\frac{(N-2)!}{N!} = \frac{n(n-1)}{N(N-1)}.$$

□

2.2 Estimateurs

Estimation aléatoire de $\overline{y}_U$:

Un estimateur aléatoire de $\overline{y}_U$ est

$$\overline{y}_W = \frac{1}{n}\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i\in W\}},$$

où $\mathbb{1}$ désigne la fonction indicatrice définie par : $\mathbb{1}_A = \begin{cases} 1 & \text{si l'événement } A \text{ est réalisé,} \\ 0 & \text{sinon.} \end{cases}$

Remarques : On peut également écrire cet estimateur

- sous la forme :

$$\overline{y}_W = \frac{1}{n}\sum_{i\in S} y_i,$$

où $S = \{(i_1,\ldots,i_n)\in\{1,\ldots,N\}^n,\ i_1\neq\ldots\neq i_n;\ \ u_{i_1}\in W,\ldots,u_{i_n}\in W\}$,

- sous la forme :

$$\overline{y}_W = \frac{1}{n}\sum_{i=1}^{N} y_i \sum_{m=1}^{n} \mathbb{1}_{\{W_m=u_i\}},$$

où W_m est la *var* égale au m-ème individu de l'échantillon.

En effet, comme $W=(W_1,\ldots,W_n)$ et tous les individus sont différents, on a
$\sum_{m=1}^{n} \mathbb{1}_{\{W_m=u_i\}} = \mathbb{1}_{\{u_i\in W\}}$.
On peut montrer que, pour tout $i\in\{1,\ldots,N\}$ et $m\in\{1,\ldots,n\}$, on a $\mathbb{P}(u_i\in W_m) = 1/N$.

Espérance de $\overline{y}_W$:

L'estimateur $\overline{y}_W$ est sans biais pour $\overline{y}_U$:

$$\mathbb{E}(\overline{y}_W) = \overline{y}_U.$$

Preuve : On propose deux preuves différentes :

Preuve I : En utilisant la linéarité de l'espérance, $\mathbb{E}(\mathbb{1}_A) = \mathbb{P}(A)$ et $\mathbb{P}(u_i \in W) = n/N$, il vient

$$\begin{aligned}\mathbb{E}(\overline{y}_W) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in W\}}\right) = \frac{1}{n}\sum_{i=1}^{N} y_i \mathbb{E}\left(\mathbb{1}_{\{u_i \in W\}}\right) \\ &= \frac{1}{n}\sum_{i=1}^{N} y_i \mathbb{P}(u_i \in W) = \frac{1}{n}\sum_{i=1}^{N} y_i \frac{n}{N} = \frac{1}{N}\sum_{i=1}^{N} y_i = \overline{y}_U.\end{aligned}$$

Preuve II : On pose $M = \binom{N}{n}$ et $W(\Omega) = \{\omega_1, \ldots, \omega_M\}$, où, pour tout $m \in \{1, \ldots, M\}$, ω_m désigne un échantillon de n individus de U. La formule du transfert donne :

$$\begin{aligned}\mathbb{E}(\overline{y}_W) &= \sum_{m=1}^{M} \overline{y}_{\omega_m} \mathbb{P}(W = \omega_m) = \frac{1}{\binom{N}{n}} \sum_{m=1}^{M} \overline{y}_{\omega_m} = \frac{1}{\binom{N}{n}} \sum_{m=1}^{M} \frac{1}{n} \sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in \omega_m\}} \\ &= \frac{1}{n\binom{N}{n}} \sum_{i=1}^{N} y_i \sum_{m=1}^{M} \mathbb{1}_{\{u_i \in \omega_m\}}.\end{aligned}$$

Comme il y a autant d'échantillons contenant u_i que de possibilités pour prélever simultanément $n-1$ individus parmi les $N-1$ autres que u_i, on a $\sum_{m=1}^{M} \mathbb{1}_{\{u_i \in \omega_m\}} = \binom{N-1}{n-1}$. Donc

$$\mathbb{E}(\overline{y}_W) = \frac{\binom{N-1}{n-1}}{n\binom{N}{n}} \sum_{i=1}^{N} y_i = \frac{\frac{(N-1)!}{(n-1)!((N-1)-(n-1))!}}{n\frac{N!}{n!(N-n)!}} \sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N} y_i = \overline{y}_U.$$

□

Variance de $\overline{y}_W$:

La variance de $\overline{y}_W$ est

$$\mathbb{V}(\overline{y}_W) = (1-f)\frac{s_U^2}{n}.$$

Preuve : Par la formule de la variance d'une somme de *var*, on obtient

$$\begin{aligned}\mathbb{V}(\overline{y}_W) &= \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in W\}}\right) = \frac{1}{n^2}\mathbb{V}\left(\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in W\}}\right) \\ &= \frac{1}{n^2}\left(\sum_{i=1}^{N} \mathbb{V}\left(y_i \mathbb{1}_{\{u_i \in W\}}\right) + 2\sum_{i=2}^{N}\sum_{j=1}^{i-1} \mathbb{C}\left(y_i \mathbb{1}_{\{u_i \in W\}}, y_j \mathbb{1}_{\{u_j \in W\}}\right)\right) \\ &= \frac{1}{n^2}\left(\sum_{i=1}^{N} y_i^2 \mathbb{V}\left(\mathbb{1}_{\{u_i \in W\}}\right) + 2\sum_{i=2}^{N}\sum_{j=1}^{i-1} y_i y_j \mathbb{C}\left(\mathbb{1}_{\{u_i \in W\}}, \mathbb{1}_{\{u_j \in W\}}\right)\right).\end{aligned}$$

Or, en utilisant $\mathbb{P}(u_i \in W) = n/N$, on a

$$\begin{aligned}\mathbb{V}\left(\mathbb{1}_{\{u_i \in W\}}\right) &= \mathbb{E}\left(\mathbb{1}^2_{\{u_i \in W\}}\right) - \left(\mathbb{E}\left(\mathbb{1}_{\{u_i \in W\}}\right)\right)^2 = \mathbb{P}(u_i \in W) - \left(\mathbb{P}(u_i \in W)\right)^2 \\ &= \frac{n}{N} - \left(\frac{n}{N}\right)^2 = \frac{n}{N}\left(1 - \frac{n}{N}\right).\end{aligned}$$

De plus, comme $\mathbb{P}(\{u_i \in W\} \cap \{u_j \in W\}) = \mathbb{P}((u_i, u_j) \in W) = n(n-1)/(N(N-1))$, il vient

$$\begin{aligned}\mathbb{C}\left(\mathbb{1}_{\{u_i \in W\}}, \mathbb{1}_{\{u_j \in W\}}\right) &= \mathbb{E}\left(\mathbb{1}_{\{u_i \in W\}}\mathbb{1}_{\{u_j \in W\}}\right) - \mathbb{E}\left(\mathbb{1}_{\{u_i \in W\}}\right)\mathbb{E}\left(\mathbb{1}_{\{u_j \in W\}}\right) \\ &= \mathbb{P}(\{u_i \in W\} \cap \{u_j \in W\}) - \mathbb{P}(u_i \in W)\mathbb{P}(u_j \in W) \\ &= \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2 = \frac{n}{N}\left(\frac{n-1}{N-1} - \frac{n}{N}\right).\end{aligned}$$

En combinant ces égalités, on obtient

$$\begin{aligned}\mathbb{V}(\overline{y}_W) &= \frac{1}{n^2}\left(\frac{n}{N}\left(1 - \frac{n}{N}\right)\sum_{i=1}^N y_i^2 + 2\frac{n}{N}\left(\frac{n-1}{N-1} - \frac{n}{N}\right)\sum_{i=2}^N\sum_{j=1}^{i-1} y_i y_j\right) \\ &= \frac{1}{nN}\left(\left(1 - \frac{n}{N}\right)\sum_{i=1}^N y_i^2 + \left(\frac{n-1}{N-1} - \frac{n}{N}\right)\left(2\sum_{i=2}^N\sum_{j=1}^{i-1} y_i y_j\right)\right).\end{aligned}$$

En utilisant la décomposition :

$$2\sum_{i=2}^N\sum_{j=1}^{i-1} y_i y_j = \left(\sum_{i=1}^N y_i\right)^2 - \sum_{i=1}^N y_i^2,$$

on obtient

$$\begin{aligned}\mathbb{V}(\overline{y}_W) &= \frac{1}{nN}\left(\left(1 - \frac{n}{N}\right)\sum_{i=1}^N y_i^2 + \left(\frac{n-1}{N-1} - \frac{n}{N}\right)\left(\left(\sum_{i=1}^N y_i\right)^2 - \sum_{i=1}^N y_i^2\right)\right) \\ &= \frac{1}{nN}\left(\left(1 - \frac{n}{N} - \frac{n-1}{N-1} + \frac{n}{N}\right)\sum_{i=1}^N y_i^2 + \left(\frac{n-1}{N-1} - \frac{n}{N}\right)\left(\sum_{i=1}^N y_i\right)^2\right) \\ &= \frac{1}{nN}\left(\frac{N-n}{N-1}\sum_{i=1}^N y_i^2 - \frac{N-n}{N(N-1)}\left(\sum_{i=1}^N y_i\right)^2\right) \\ &= \frac{N-n}{nN}\left(\frac{1}{N-1}\left(\sum_{i=1}^N y_i^2 - N\left(\frac{1}{N}\sum_{i=1}^N y_i\right)^2\right)\right).\end{aligned}$$

D'autre part, on a

$$\begin{aligned} s_U^2 &= \frac{1}{N-1}\sum_{i=1}^N (y_i - \overline{y}_U)^2 = \frac{1}{N-1}\left(\sum_{i=1}^N y_i^2 - 2\overline{y}_U \sum_{i=1}^N y_i + N\overline{y}_U^2\right) \\ &= \frac{1}{N-1}\left(\sum_{i=1}^N y_i^2 - 2N\overline{y}_U^2 + N\overline{y}_U^2\right) = \frac{1}{N-1}\left(\sum_{i=1}^N y_i^2 - N\overline{y}_U^2\right) \\ &= \frac{1}{N-1}\left(\sum_{i=1}^N y_i^2 - N\left(\frac{1}{N}\sum_{i=1}^N y_i\right)^2\right). \end{aligned}$$

Il s'ensuit

$$\mathbb{V}(\overline{y}_W) = \frac{N-n}{nN} s_U^2 = \left(1 - \frac{n}{N}\right)\frac{s_U^2}{n} = (1-f)\frac{s_U^2}{n}.$$

□

Erreur quadratique moyenne de $\overline{y}_W$:

L'erreur quadratique moyenne de $\overline{y}_W$ est le réel :

$$EQM(\overline{y}_W)[PESR] = \mathbb{E}\left((\overline{y}_W - \overline{y}_U)^2\right) = (1-f)\frac{s_U^2}{n}.$$

La quantité $EQM(\overline{y}_W)[PESR]$ est une mesure de l'erreur que commet $\overline{y}_W$ dans l'estimation de $\overline{y}_U$.

On constate que :

- plus n est grand/l'échantillon est grand, plus $\overline{y}_W$ estime bien $\overline{y}_U$,
- plus U est homogène/plus s_U^2 est petit, plus $\overline{y}_W$ estime bien $\overline{y}_U$.

Estimation aléatoire de s_U :

Un estimateur aléatoire de s_U est

$$s_W = \sqrt{\frac{1}{n-1}\sum_{i=1}^N (y_i - \overline{y}_W)^2 \mathbb{1}_{\{u_i \in W\}}}.$$

Propriété de s_W^2 :

L'estimateur s_W^2 est sans biais pour s_U^2 :

$$\mathbb{E}\left(s_W^2\right) = s_U^2.$$

Preuve : En remarquant que $\sum_{i=1}^{N} \mathbb{1}_{\{u_i \in W\}} = n$, il vient

$$\begin{aligned} s_W^2 &= \frac{1}{n-1}\sum_{i=1}^{N}(y_i - \overline{y}_W)^2 \mathbb{1}_{\{u_i \in W\}} \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{N} y_i^2 \mathbb{1}_{\{u_i \in W\}} - 2\overline{y}_W \sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in W\}} + \overline{y}_W^2 \sum_{i=1}^{N} \mathbb{1}_{\{u_i \in W\}}\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{N} y_i^2 \mathbb{1}_{\{u_i \in W\}} - 2n\overline{y}_W^2 + n\overline{y}_W^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{N} y_i^2 \mathbb{1}_{\{u_i \in W\}} - n\overline{y}_W^2\right). \end{aligned}$$

On a $\mathbb{P}(u_i \in W) = n/N$ et

$$\mathbb{E}\left(\overline{y}_W^2\right) = \mathbb{V}\left(\overline{y}_W\right) + \left(\mathbb{E}\left(\overline{y}_W\right)\right)^2 = (1-f)\frac{s_U^2}{n} + \overline{y}_U^2.$$

D'où

$$\begin{aligned} \mathbb{E}\left(s_W^2\right) &= \mathbb{E}\left(\frac{1}{n-1}\left(\sum_{i=1}^{N} y_i^2 \mathbb{1}_{\{u_i \in W\}} - n\overline{y}_W^2\right)\right) = \frac{1}{n-1}\left(\sum_{i=1}^{N} y_i^2 \mathbb{E}\left(\mathbb{1}_{\{u_i \in W\}}\right) - n\mathbb{E}\left(\overline{y}_W^2\right)\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{N} y_i^2 \mathbb{P}\left(u_i \in W\right) - n\mathbb{E}\left(\overline{y}_W^2\right)\right) \\ &= \frac{1}{n-1}\left(\frac{n}{N}\sum_{i=1}^{N} y_i^2 - n\left((1-f)\frac{s_U^2}{n} + \overline{y}_U^2\right)\right) \\ &= \frac{1}{n-1}\left(\frac{n}{N}\left(\sum_{i=1}^{N} y_i^2 - N\overline{y}_U^2\right) - \left(1 - \frac{n}{N}\right)s_U^2\right) \\ &= \frac{n(N-1)}{(n-1)N}\left(\frac{1}{N-1}\left(\sum_{i=1}^{N} y_i^2 - N\overline{y}_U^2\right)\right) - \frac{1}{n-1}\left(1 - \frac{n}{N}\right)s_U^2. \end{aligned}$$

Or

$$\begin{aligned} s_U^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y}_U)^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} y_i^2 - 2\overline{y}_U \sum_{i=1}^{N} y_i + N\overline{y}_U^2\right) \\ &= \frac{1}{N-1}\left(\sum_{i=1}^{N} y_i^2 - 2N\overline{y}_U^2 + N\overline{y}_U^2\right) = \frac{1}{N-1}\left(\sum_{i=1}^{N} y_i^2 - N\overline{y}_U^2\right). \end{aligned}$$

Par conséquent,

$$\begin{aligned}\mathbb{E}\left(s_W^2\right) &= \frac{n(N-1)}{(n-1)N}s_U^2 - \frac{1}{n-1}\left(1-\frac{n}{N}\right)s_U^2 \\ &= \frac{n(N-1)-N+n}{(n-1)N}s_U^2 = \frac{nN-n-N+n}{(n-1)N}s_U^2 = \frac{(n-1)N}{(n-1)N}s_U^2 = s_U^2.\end{aligned}$$

□

2.3 Estimations ponctuelles

Estimation ponctuelle de $\overline{y}_U$:

Soit ω un échantillon de n individus de U. Une estimation ponctuelle de $\overline{y}_U$ est la moyenne-échantillon :

$$\overline{y}_\omega = \frac{1}{n}\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in \omega\}}.$$

Quelques commandes R : Un exemple de calcul de $\overline{y}_\omega$ avec R est décrit ci-dessous :

```
U = c("Bob", "Nico", "Ali", "Fabien", "Malik", "John", "Jean", "Chris", "Karl")
y = c(72, 89, 68, 74, 81, 87, 76, 61, 84)
n = 3
library(sampling)
t = srswor(n, 9)
bar_y_w = (1 / n) * sum(y * t)
bar_y_w
```

Erreur d'estimation :

Soit ω un échantillon de n individus de U. L'erreur d'estimation que commet $\overline{y}_\omega$ en estimant $\overline{y}_U$ est le réel :

$$e_\omega = |\overline{y}_\omega - \overline{y}_U|.$$

Probabilité d'erreur :

La probabilité de se tromper de plus de $(100 \times \beta)\%$, $\beta \in]0,1[$, en estimant $\overline{y}_U$ par $\overline{y}_W$ est le réel :

$$p_\beta = \frac{1}{\binom{N}{n}} \sum_{\omega \in W(\Omega)} \mathbb{1}_{\{e_\omega \geq \beta \overline{y}_U\}}.$$

Estimation ponctuelle de s_U :

Soit ω un échantillon de n individus de U. Une estimation ponctuelle de s_U est l'écart-type corrigé-échantillon :

$$s_\omega = \sqrt{\frac{1}{n-1}\sum_{i=1}^{N}(y_i - \overline{y}_\omega)^2 \mathbb{1}_{\{u_i \in \omega\}}}.$$

Tout comme la moyenne-population, on peut aussi s'intéresser à l'erreur d'estimation et la probabilité d'erreur, lesquelles se définissent de manière similaire.

Quelques commandes R : Un exemple de calcul de s_ω avec R est décrit ci-dessous :

```
U = c("Bob", "Nico", "Ali", "Fabien", "Malik", "John", "Jean", "Chris", "Karl")
y = c(72, 89, 68, 74, 81, 87, 76, 61, 84)
n = 3
library(sampling)
t = srswor(n, 9)
bar_y_w = (1 / n) * sum(y * t)
s_w = sqrt(sum((y - bar_y_w)^2 * t) / (n - 1))
s_w
```

Estimation ponctuelle de l'écart-type de $\overline{y}_W$:

Soit ω un échantillon de n individus de U. Une estimation ponctuelle de l'écart-type de $\overline{y}_W$ est le réel :

$$s(\overline{y}_\omega) = \sqrt{(1-f)\frac{s_\omega^2}{n}}.$$

2.4 Intervalles de confiance

Résultat limite (Théorème de Hajek) : Si n, N et $N-n$ sont suffisamment grands, alors on a

$$Z = \frac{\overline{y}_W - \overline{y}_U}{\sqrt{(1-f)\frac{s_W^2}{n}}} \approx \mathcal{N}(0,1).$$

Intervalle de confiance pour $\overline{y}_U$:

Soit ω un échantillon de n individus de U. Un intervalle de confiance pour $\overline{y}_U$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, est

$$\begin{aligned} i_{\overline{y}_U} &= [\overline{y}_\omega - z_\alpha s(\overline{y}_\omega),\ \overline{y}_\omega + z_\alpha s(\overline{y}_\omega)] \\ &= \left[\overline{y}_\omega - z_\alpha \sqrt{(1-f)\frac{s_\omega^2}{n}},\ \overline{y}_\omega + z_\alpha \sqrt{(1-f)\frac{s_\omega^2}{n}}\right], \end{aligned}$$

où z_α est le réel vérifiant $\mathbb{P}(|Z| \geq z_\alpha) = \alpha$, $Z \sim \mathcal{N}(0,1)$.

Il y a $100(1-\alpha)$ chances sur 100 que $\overline{y}_U$ appartienne à l'intervalle $i_{\overline{y}_U}$.

Quelques commandes R : Un exemple de fonction R pour calculer l'intervalle de confiance pour $\overline{y}_U$ au niveau $100(1-\alpha)\%$ est décrit ci-dessous :

```
icPESR = function(y, N, niveau) {
n = length(y)
bar_y_w = mean(y)
z = qnorm(1 - (1 - niveau) / 2)
s2_w = sd(y)^2
var_bar_y_w = (1 - n / N) * (s2_w / n)
a = bar_y_w - z * sqrt(var_bar_y_w)
b = bar_y_w + z * sqrt(var_bar_y_w)
print(c(a, b)) }
icPESR(y = c(2.1, 2.3, 4.1, 2.6, 7.1, 8.6), N = 100, niveau = 0.95)
```

Cela renvoie : 2.329876, 6.603457.

2.5 Taille d'échantillon

Incertitude absolue :

Soit ω un échantillon de n individus de U. On appelle incertitude absolue sur $\overline{y}_U$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, la demi-longueur de $i_{\overline{y}_U}$:

$$d_\omega = z_\alpha s(\overline{y}_\omega) = z_\alpha \sqrt{(1-f)\frac{s_\omega^2}{n}}.$$

Plus d_ω est petit, plus l'estimation de $\overline{y}_U$ par $\overline{y}_\omega$ est précise.

Incertitude relative :

Soit ω un échantillon de n individus de U et d_ω l'incertitude absolue sur $\overline{y}_U$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$. On appelle incertitude relative sur $\overline{y}_U$ au niveau $100(1-\alpha)\%$ le pourcentage $(100 \times d^*_\omega)\%$ où d^*_ω est le réel :

$$d^*_\omega = \frac{d_\omega}{\overline{y}_\omega}.$$

Taille d'échantillon :

Soit ω un échantillon prélevé lors d'une étude préliminaire. La taille d'échantillon n à choisir pour avoir :

- une incertitude absolue sur $\overline{y}_U$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, inférieure ou égale à d_0 est le plus petit n tel que
$$d_\omega \leq d_0 \qquad \Leftrightarrow \qquad n \geq \frac{N z^2_\alpha s^2_\omega}{N d^2_0 + z^2_\alpha s^2_\omega},$$
- une incertitude relative sur $\overline{y}_U$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, inférieure ou égale à $(100 \times d_1)\%$ est le plus petit n tel que
$$d^*_\omega \leq d_1 \qquad \Leftrightarrow \qquad n \geq \frac{N z^2_\alpha s^2_\omega}{N(\overline{y}_\omega d_1)^2 + z^2_\alpha s^2_\omega}.$$

Quelques commandes R : Un exemple de fonction R pour calculer la taille n d'un échantillon à partir de l'incertitude absolue sur $\overline{y}_U$ au niveau $100(1-\alpha)\%$ est décrit ci-dessous :

```
n_ech = function(N, s2, d0, niveau) {
z = qnorm(1 - (1 - niveau) / 2)
n = N * s2 * z^2 / (N * d0^2 + s2 * z^2)
print (ceiling(n)) }
n_ech(N = 1000, s2 = 625, d0 = 3, niveau = 0.95)
```

Cela renvoie 211.

2.6 Sélection des individus

Méthode du tri aléatoire : La méthode du tri aléatoire est un un plan de sondage aléatoire de type PESR. Pour la mettre en œuvre,

- on génère N nombres $x_1, \ldots, x_N$ (indépendamment des uns des autres) suivant la loi uniforme $\mathcal{U}([0,1])$,

- pour tout $i \in \{1, \ldots, N\}$, on affecte à l'individu u_i le nombre x_i,
- on sélectionne les n individus correspondant au n plus grandes valeurs de $x_1, \ldots, x_N$.

Quelques commandes R : Un exemple de commandes R sur la méthode du tri aléatoire est décrit ci-dessous :

```
N = 100
n = 10
x = runif(N)
z = NULL
u = x
for (i in 1:10){
z[i] = which.max(u)
u[which.max(u)] = 0 }
z
```

2.7 Exercices corrigés

Exercice 1 : *L'objectif de cet exercice est d'illustrer certains résultats théoriques du cours sur les plans de sondage aléatoire de type PESR avec un exemple.* On étudie un caractère Y dans une population de 5 individus : $U = \{u_1, \ldots, u_5\}$. Pour tout $i \in \{1, \ldots, 5\}$, soit y_i la valeur de Y pour l'individu u_i. Les résultats sont :

y_1	y_2	y_3	y_4	y_5
3	4	6	8	13

1. Calculer la moyenne-population $\overline{y}_U$ et l'écart-type corrigé-population s_U.
2. On prélève au hasard et simultanément 2 individus dans cette population formant ainsi un échantillon. Chaque individu a la même probabilité qu'un autre d'être sélectionné. On est donc dans le cadre PESR.
   (a) Quel est est le taux de sondage ? Combien d'échantillons peut-on former ? Expliciter les.
   (b) Pour chaque échantillon ω, calculer la moyenne-échantillon $\overline{y}_\omega$ et l'écart-type corrigé-échantillon s_ω.
   (c) Soit $\overline{y}_W$ la *var* égale à la moyenne-échantillon, l'aléatoire étant dans l'échantillon considéré. Déterminer sa loi, puis calculer son espérance et sa variance.

(d) Soit s_W la *var* égale à l'écart-type corrigé-échantillon, l'aléatoire étant dans l'échantillon considéré. Calculer l'espérance de s_W^2.

(e) Retrouver les résultats des deux questions précédentes avec les formules du cours.

(f) Calculer les erreurs dans l'estimation de $\overline{y}_U$.

(g) Quelle est la probabilité de se tromper de plus de 20% dans l'estimation de $\overline{y}_U$?

Solution :

1. En prenant la moyenne et l'écart-type corrigé des données, on obtient

$$\overline{y}_U = 6.8, \qquad s_U = 3.9623.$$

2. (a) Le taux de sondage est

$$f = \frac{n}{N} = \frac{2}{5} = 0.4.$$

Vu le mode de prélèvement, le nombre d'échantillons possibles est

$$\binom{5}{2} = \frac{5!}{2!(5-2)!} = 10.$$

Ils sont :

$\{u_1, u_2\}$	$\{u_1, u_3\}$	$\{u_1, u_4\}$	$\{u_1, u_5\}$	$\{u_2, u_3\}$
$\{u_2, u_4\}$	$\{u_2, u_5\}$	$\{u_3, u_4\}$	$\{u_3, u_5\}$	$\{u_4, u_5\}$

(b) On a, en prenant 4 chiffres après la virgule :

ω	Y	$\overline{y}_\omega$	s_ω
$\{u_1, u_2\}$	$\{3, 4\}$	3.5	0.7071
$\{u_1, u_3\}$	$\{3, 6\}$	4.5	2.1213
$\{u_1, u_4\}$	$\{3, 8\}$	5.5	3.5355
$\{u_1, u_5\}$	$\{3, 13\}$	8	7.0710
$\{u_2, u_3\}$	$\{4, 6\}$	5	1.4142
$\{u_2, u_4\}$	$\{4, 8\}$	6	2.8284
$\{u_2, u_5\}$	$\{4, 13\}$	8.5	6.3639
$\{u_3, u_4\}$	$\{6, 8\}$	7	1.4142
$\{u_3, u_5\}$	$\{6, 13\}$	9.5	4.9497
$\{u_4, u_5\}$	$\{8, 13\}$	10.5	3.5355

(c) Soit $\overline{y}_W$ la *var* égale à la moyenne-échantillon. L'ensemble des valeurs possibles pour $\overline{y}_W$ est

$$\overline{y}_W(\Omega) = \{3.5, 4.5, 5.5, 8, 5, 6, 8.5, 7, 9.5, 10.5\}.$$

Comme il y a 10 échantillons différents et qu'ils sont équiprobables, la loi de $\overline{y}_W$ est donnée par

k	3.5	4.5	5.5	8	5	6	8.5	7	9.5	10.5
$\mathbb{P}(\overline{y}_W = k)$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

En utilisant la loi de $\overline{y}_W$, l'espérance de $\overline{y}_W$ est

$$\begin{aligned}\mathbb{E}(\overline{y}_W) &= \sum_{k \in \overline{y}_W(\Omega)} k\mathbb{P}(\overline{y}_W = k) \\ &= \frac{1}{10}(3.5 + 4.5 + 5.5 + 8 + 5 + 6 + 8.5 + 7 + 9.5 + 10.5) \\ &= 6.8.\end{aligned}$$

En utilisant la formule de König-Huyghens, la variance de $\overline{y}_W$ est

$$\mathbb{V}(\overline{y}_W) = \mathbb{E}(\overline{y}_W^2) - \left(\mathbb{E}(\overline{y}_W)\right)^2.$$

Or on a $\mathbb{E}(\overline{y}_W) = 6.8$ et

$$\begin{aligned}\mathbb{E}(\overline{y}_W^2) &= \sum_{k \in \overline{y}_W(\Omega)} k^2\mathbb{P}(\overline{y}_W = k) \\ &= \frac{1}{10}(3.5^2 + 4.5^2 + 5.5^2 + 8^2 + 5^2 + 6^2 + 8.5^2 + 7^2 + 9.5^2 + 10.5^2) \\ &= 50.95.\end{aligned}$$

D'où

$$\mathbb{V}(\overline{y}_W) = 50.95 - 6.8^2 = 4.71.$$

(d) Soit s_W la *var* égale à l'écart-type corrigé-échantillon. L'ensemble des valeurs possibles pour s_W est

$$s_W(\Omega) = \{0.7071, 1.4142, 2.1213, 2.8284, 3.5355, 4.9497, 6.3639, 7.0710\}.$$

Comme il y a 10 échantillons différents et qu'ils sont équiprobables, la loi de s_W est donnée par

k	0.7071	1.4142	2.1213	2.8284	3.5355	4.9497	6.3639	7.0710
$\mathbb{P}(s_W = k)$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

En utilisant la loi de s_W^2, l'espérance de s_W^2 est

$$\begin{aligned}\mathbb{E}(s_W^2) &= \sum_{k \in s_W(\Omega)} k^2 \mathbb{P}(s_W = k) \\ &= \frac{1}{10}(0.7071^2 + 2 \times 1.4142^2 + 2.1213^2 + 2.8284^2 + 2 \times 3.5355^2 + 4.9497^2 \\ &+ 6.3639^2 + 7.0710^2) \\ &= 15.6997.\end{aligned}$$

(e) En utilisant les formules du cours, on retrouve les résultats précédents (en prenant en compte les approximations) :

$$\mathbb{E}(\overline{y}_W) = \overline{y}_U = 6.8, \qquad \mathbb{V}(\overline{y}_W) = (1-f)\frac{s_U^2}{n} = (1-0.4)\frac{3.9623^2}{2} = 4.71$$

et

$$\mathbb{E}(s_W^2) = s_U^2 = 15.6998.$$

(f) On utilise la formule d'erreur d'estimation :

$$e_\omega = |\overline{y}_\omega - \overline{y}_U| = |\overline{y}_\omega - 6.8|.$$

On a, en prenant 4 chiffres après la virgule :

ω	$\overline{y}_\omega$	e_ω
$\{u_1, u_2\}$	3.5	3.3
$\{u_1, u_3\}$	4.5	2.3
$\{u_1, u_4\}$	5.5	1.3
$\{u_1, u_5\}$	8	1.2
$\{u_2, u_3\}$	5	1.8
$\{u_2, u_4\}$	6	0.8
$\{u_2, u_5\}$	8.5	1.7
$\{u_3, u_4\}$	7	0.2
$\{u_3, u_5\}$	9.5	2.7
$\{u_4, u_5\}$	10.5	3.7

(g) On a $20\% = (100 \times \beta)\%$ avec $\beta = 0.2$. Le nombre de e_ω dépassant $\beta \times \overline{y}_U = 0.2 \times 6.8 = 1.36$ est de 6. Donc la probabilité de se tromper de plus de $(100 \times \beta)\%$ dans l'estimation de $\overline{y}_U$ par $\overline{y}_W$ est

$$p = \frac{1}{\binom{N}{n}} \sum_{\omega \in W(\Omega)} \mathbb{1}_{\{e_\omega \geq \beta \times \overline{y}_U\}} = \frac{6}{10} = 0.6.$$

Il y a 60% chances de se tromper de plus de 20% en estimant $\overline{y}_U$ par $\overline{y}_W$.

Exercice 2 : On prélève 25 sacs de farine de maïs dans une usine en contenant 200 suivant un plan de sondage aléatoire de type PESR. On pèse ces 25 sacs. Les valeurs obtenues donnent une moyenne de 13.5 kilogrammes et un écart-type corrigé de 1.3 kilogrammes.

Déterminer un intervalle de confiance pour la moyenne des poids des 200 sacs de farine de maïs au niveau 95%.

Solution : On a $95\% = 100(1-\alpha)\%$ avec $\alpha = 0.05$. On a $\mathbb{P}(|Z| \geq z_\alpha) = \alpha = 0.05$, $Z \sim \mathcal{N}(0,1)$, avec $z_\alpha = 1.96$.

Un intervalle de confiance pour la moyenne des poids des 200 sacs de farine $\overline{y}_U$ au niveau 95% est

$$\begin{aligned} i_{\overline{y}_U} &= \left[\overline{y}_\omega - z_\alpha\sqrt{(1-f)\frac{s_\omega^2}{n}},\ \overline{y}_\omega + z_\alpha\sqrt{(1-f)\frac{s_\omega^2}{n}}\right] \\ &= \left[13.5 - 1.96\sqrt{\left(1-\frac{25}{200}\right)\frac{1.3^2}{25}},\ 13.5 + 1.96\sqrt{\left(1-\frac{25}{200}\right)\frac{1.3^2}{25}}\right] \\ &= [13.0233,\ 13.9766]. \end{aligned}$$

Ainsi, il y a 95 chances sur 100 que $[13.0233,\ 13.9766]$ contienne $\overline{y}_U$, l'unité étant le kilogramme.

Exercice 3 : On dispose d'une liste de 500 foyers avec, pour chacun d'entre eux, le nombre d'individus y vivant. Sur un échantillon de 8 foyers constitué par un plan de sondage aléatoire de type PESR, les résultats sont :

3	6	1	2	4	4	1	8

1. Calculer le taux de sondage.
2. Donner une estimation ponctuelle de la moyenne des effectifs des 500 foyers.
3. Donner une estimation ponctuelle de l'écart-type corrigé de l'estimateur de la moyenne des effectifs des 500 foyers.
4. Déterminer un intervalle de confiance au niveau 95% pour la moyenne-population.
5. Déterminer la taille d'échantillon à choisir pour avoir une incertitude absolue sur la moyenne-population inférieure ou égale à 1 au niveau 95%.

Solution :

1. On a $n = 8$ et $N = 500$. Le taux de sondage est

$$f = \frac{n}{N} = \frac{8}{500} = 0.016.$$

2. Une estimation ponctuelle de la moyenne des effectifs des 500 foyers est la moyenne échantillon :

$$\overline{y}_\omega = 3.625.$$

3. Une estimation ponctuelle de l'écart-type corrigé de l'estimateur de la moyenne des effectifs des 500 foyers est

$$s(\overline{y}_\omega) = \sqrt{(1-f)\frac{s_\omega^2}{n}} = \sqrt{(1-0.016)\frac{2.4458^2}{8}} = 0.8577.$$

4. On a $95\% = 100(1-\alpha)\%$ avec $\alpha = 0.05$. On a $\mathbb{P}(|Z| \geq z_\alpha) = \alpha = 0.05$, $Z \sim \mathcal{N}(0,1)$, avec $z_\alpha = 1.96$. Un intervalle de confiance pour $\overline{y}_U$ au niveau 95% est

$$\begin{aligned} i_{\overline{y}_U} &= [\overline{y}_\omega - z_\alpha s(\overline{y}_\omega),\ \overline{y}_\omega + z_\alpha s(\overline{y}_\omega)] \\ &= [3.625 - 1.96 \times 0.8577,\ 3.625 + 1.96 \times 0.8577] \\ &= [1.9439,\ 5.3060]. \end{aligned}$$

   Ainsi, il y a 95 chances sur 100 que $[1.9439,\ 5.3060]$ contienne $\overline{y}_U$.

5. On a $95\% = 100(1-\alpha)\%$ avec $\alpha = 0.05$. On souhaite déterminer le plus petit n tel que :

$$d_\omega = z_\alpha \sqrt{(1-f)\frac{s_\omega^2}{n}} \leq d_0 \qquad \Leftrightarrow \qquad n \geq \frac{N z_\alpha^2 s_\omega^2}{N d_0^2 + z_\alpha^2 s_\omega^2},$$

   avec $d_0 = 1$, $z_\alpha = 1.96$, ω est l'échantillon considéré précédemment, $s_\omega = 2.4458$ et $N = 500$. On a

$$\frac{500 \times 1.96^2 \times 2.4458^2}{500 \times 1^2 + 1.96^2 \times 2.4458^2} = 21.97044.$$

   Donc $n = 22$ convient.

2.8 Synthèse

Paramètres-population et les paramètres-échantillon correspondants :

	Population U	Échantillon ω
Taille	N	n
Taux de sondage	$\square$	$f = \dfrac{n}{N}$
Moyenne	$\overline{y}_U = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} y_i$	$\overline{y}_\omega = \dfrac{1}{n}\displaystyle\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in \omega\}}$
Écart-type corrigé	$s_U = \sqrt{\dfrac{1}{N-1}\displaystyle\sum_{i=1}^{N}(y_i - \overline{y}_U)^2}$	$s_\omega = \sqrt{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{N}(y_i - \overline{y}_\omega)^2 \mathbb{1}_{\{u_i \in \omega\}}}$
Écart-type de $\overline{y}_W$	$\sigma(\overline{y}_W) = \sqrt{(1-f)\dfrac{s_U^2}{n}}$	$s(\overline{y}_\omega) = \sqrt{(1-f)\dfrac{s_\omega^2}{n}}$

Autre notions utilisées autour de $\overline{y}_U$ (niveau : $100(1-\alpha)\%$, $\alpha \in]0,1[$) :

Intervalle de confiance	$i_{\overline{y}_U} = \left[\overline{y}_\omega - z_\alpha\sqrt{(1-f)\dfrac{s_\omega^2}{n}},\ \overline{y}_\omega + z_\alpha\sqrt{(1-f)\dfrac{s_\omega^2}{n}}\right]$
Incertitude absolue	$d_\omega = z_\alpha\sqrt{(1-f)\dfrac{s_\omega^2}{n}}$
Incertitude relative	$d_\omega^* = \dfrac{d_\omega}{\overline{y}_\omega}$
Taille n telle que $d_\omega \leq d_0$	$n \geq \dfrac{N z_\alpha^2 s_\omega^2}{N d_0^2 + z_\alpha^2 s_\omega^2}$
Taille n telle que $d_\omega^* \leq d_1$	$n \geq \dfrac{N z_\alpha^2 s_\omega^2}{N(\overline{y}_\omega d_1)^2 + z_\alpha^2 s_\omega^2}$

Rappel : $\mathbb{P}(|Z| \geq z_\alpha) = \alpha$, $Z \sim \mathcal{N}(0,1)$.

3 Total, proportion et effectif dans le cadre PESR

On reprend le cadre mathématique d'un plan de sondage aléatoire de type PESR.

3.1 Estimation du total

Total :

On appelle total-population le réel :

$$\tau_U = \sum_{i=1}^{N} y_i = N\overline{y}_U.$$

Estimation aléatoire de τ_U :

Un estimateur aléatoire de τ_U est

$$\tau_W = N\overline{y}_W = N\frac{1}{n}\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in W\}}.$$

Espérance de τ_W :

L'estimateur τ_W est sans biais pour τ_U :

$$\mathbb{E}(\tau_W) = \tau_U.$$

Preuve : Comme $\mathbb{E}(\overline{y}_W) = \overline{y}_U$, on a

$$\mathbb{E}(\tau_W) = \mathbb{E}(N\overline{y}_W) = N\mathbb{E}(\overline{y}_W) = N\overline{y}_U = \tau_U.$$

□

Variance de τ_W :

La variance de τ_W est

$$\mathbb{V}(\tau_W) = N^2(1-f)\frac{s_U^2}{n}.$$

Preuve : Comme $\mathbb{V}(\overline{y}_W) = (1-f)s_U^2/n$, on a

$$\mathbb{V}(\tau_W) = \mathbb{V}(N\overline{y}_W) = N^2\mathbb{V}(\overline{y}_W) = N^2(1-f)\frac{s_U^2}{n}.$$

□

Erreur quadratique moyenne de τ_W :

L'erreur quadratique moyenne de τ_W est le réel :

$$EQM(\tau_W)[PESR] = N^2(1-f)\frac{s_U^2}{n}.$$

Estimation ponctuelle de τ_U :

Soit ω un échantillon de n individus de U. Une estimation ponctuelle de τ_U est le total-échantillon :

$$\tau_\omega = N\overline{y}_\omega = N\frac{1}{n}\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in \omega\}}.$$

Estimation ponctuelle de l'écart-type de τ_W :

Soit ω un échantillon de n individus de U. Une estimation ponctuelle de l'écart-type de τ_W est le réel :

$$s(\tau_\omega) = \sqrt{N^2(1-f)\frac{s_\omega^2}{n}}.$$

Intervalle de confiance pour τ_U :

Soit ω un échantillon de n individus de U. Un intervalle de confiance pour τ_U au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, est

$$\begin{aligned} i_{\tau_U} &= [\tau_\omega - z_\alpha s(\tau_\omega),\ \tau_\omega + z_\alpha s(\tau_\omega)] \\ &= \left[\tau_\omega - z_\alpha\sqrt{N^2(1-f)\frac{s_\omega^2}{n}},\ \tau_\omega + z_\alpha\sqrt{N^2(1-f)\frac{s_\omega^2}{n}}\right] = N \times i_{\overline{y}_U}, \end{aligned}$$

où z_α est le réel vérifiant $\mathbb{P}(|Z| \geq z_\alpha) = \alpha$, $Z \sim \mathcal{N}(0,1)$.

On peut également définir l'incertitude absolue ou relative sur τ_U, ainsi que la taille d'échantillon souhaitée pour une incertitude donnée.

3.2 Estimation d'une proportion

Contexte : On suppose que le caractère Y est binaire : $Y(\Omega) = \{0, 1\}$. Cela correspond à un codage.

Par exemple, $Y = 1$ peut caractériser :

- le succès à une épreuve,
- la présence d'un élément caractéristique.

Ainsi, les données brutes $y_1, \ldots, y_N$ sont constituées uniquement de 0 et de 1.

Proportion :

On appelle proportion-population la proportion des individus dans U vérifiant $Y = 1$:

$$p_U = \frac{1}{N}\sum_{i=1}^{N} y_i \qquad (= \overline{y}_U).$$

Estimation d'une proportion :

Un estimateur aléatoire de p_U est

$$p_W = \overline{y}_W = \frac{1}{n}\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in W\}}.$$

Espérance de p_W :

L'estimateur p_W est sans biais pour p_U :

$$\mathbb{E}(p_W) = p_U.$$

Variance de p_W :

La variance de p_W est

$$\mathbb{V}(p_W) = (1-f)\frac{s_U^2}{n} = (1-f)\frac{N}{n(N-1)} p_U(1-p_U).$$

Preuve : Comme $y_i \in \{0, 1\}$ pour tout $i \in \{1, \ldots, N\}$, on a $y_i^2 = y_i$ et

$$\begin{aligned} s_U^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y}_U)^2 = \frac{N}{N-1}\left(\frac{1}{N}\sum_{i=1}^{N} y_i^2 - 2\overline{y}_U \frac{1}{N}\sum_{i=1}^{N} y_i + \overline{y}_U^2\right) \\ &= \frac{N}{N-1}\left(\frac{1}{N}\sum_{i=1}^{N} y_i - \left(\frac{1}{N}\sum_{i=1}^{N} y_i\right)^2\right) = \frac{N}{N-1}(p_U - p_U^2) = \frac{N}{N-1} p_U(1-p_U). \end{aligned}$$

□

Erreur quadratique moyenne de p_W :

L'erreur quadratique moyenne de p_W est le réel :

$$EQM(p_W)[PESR] = (1-f)\frac{N}{n(N-1)}p_U(1-p_U).$$

Estimation ponctuelle de p_U :

Soit ω un échantillon de n individus de U. Une estimation ponctuelle de p_U est la proportion-échantillon :

$$p_\omega = \overline{y}_\omega = \frac{1}{n}\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in \omega\}}.$$

Estimation ponctuelle de l'écart-type de p_W :

Soit ω un échantillon de n individus de U. Une estimation ponctuelle de l'écart-type de p_W est le réel :

$$s(p_\omega) = \sqrt{(1-f)\frac{p_\omega(1-p_\omega)}{n-1}}.$$

Intervalle de confiance pour p_U :

Soit ω un échantillon de n individus de U. Un intervalle de confiance pour p_U au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, est

$$\begin{aligned} i_{p_U} &= [p_\omega - z_\alpha s(p_\omega),\ p_\omega + z_\alpha s(p_\omega)] \\ &= \left[p_\omega - z_\alpha\sqrt{(1-f)\frac{p_\omega(1-p_\omega)}{n-1}},\ p_\omega + z_\alpha\sqrt{(1-f)\frac{p_\omega(1-p_\omega)}{n-1}}\right], \end{aligned}$$

où z_α est le réel vérifiant $\mathbb{P}(|Z| \geq z_\alpha) = \alpha$, $Z \sim \mathcal{N}(0,1)$.

Quelques commandes R : Un exemple de fonction R pour calculer l'intervalle de confiance pour p_U au niveau $100(1-\alpha)\%$ est décrit ci-dessous :

```
icPESR = function(y, N, niveau) {
n = length(y)
p_w = mean(y)
z = qnorm(1 - (1 - niveau) / 2)
var_p_w = (1 - n / N) * (p_w * (1 - p_w) / (n - 1))
a = p_w - z * sqrt(var_p_w)
b = p_w + z * sqrt(var_p_w)
print(c(a, b)) }
icPESR(y = c(0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0), N = 100, niveau = 0.90)
```

Cela renvoie : 0.3176725, 0.7592506.

Incertitude absolue :

Soit ω un échantillon de n individus de U. On appelle incertitude absolue sur p_U au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, la demi-longueur de i_{p_U} :

$$d_\omega = z_\alpha s(p_\omega) = z_\alpha \sqrt{(1-f)\frac{p_\omega(1-p_\omega)}{n-1}}.$$

Plus d_ω est petit, plus l'estimation de p_U par p_ω est précise.

Incertitude relative :

Soit ω un échantillon de n individus de U et d_ω l'incertitude absolue sur p_U au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$. On appelle incertitude relative sur p_U au niveau $100(1-\alpha)\%$ le pourcentage $(100 \times d^*_\omega)\%$ où d^*_ω est le réel :

$$d^*_\omega = \frac{d_\omega}{p_\omega}.$$

Taille d'échantillon :

Soit ω un échantillon prélevé lors d'une étude préliminaire. La taille d'échantillon n à choisir pour avoir :

- une incertitude absolue sur p_U au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, inférieure ou égale à d_0 est le plus petit n tel que

$$d_\omega \leq d_0 \qquad \Rightarrow \qquad n \geq \frac{N z_\alpha^2 p_\omega (1-p_\omega)}{N d_0^2 + z_\alpha^2 p_\omega (1-p_\omega)},$$

- une incertitude relative sur p_U au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, inférieure ou égale à $(100 \times d_1)\%$ est le plus petit n tel que

$$d_\omega^* \leq d_1 \qquad \Rightarrow \qquad n \geq \frac{N z_\alpha^2 p_\omega (1-p_\omega)}{N (p_\omega d_1)^2 + z_\alpha^2 p_\omega (1-p_\omega)}.$$

On peut aussi remplacer $p_\omega(1-p_\omega)$ par $1/4$, ce qui évite une étude avec un échantillon préliminaire pour l'incertitude absolue.

Quelques commandes R : Un exemple de fonction R pour calculer la taille n d'un échantillon à partir de l'incertitude absolue sur p_U au niveau $100(1-\alpha)\%$ est décrit ci-dessous :

```
n_ech = function(N, p_w, d0, niveau) {
z = qnorm(1 - (1 - niveau) / 2)
n = N * p_w * (1 - p_w) * z^2 / (N * d0^2 + p_w * (1 - p_w) * z^2)
print(ceiling(n)) }
n_ech(N = 1000, p_w = 0.45, d0 = 0.2, niveau = 0.95)
```

Cela renvoie 24.

3.3 Estimation d'un effectif

Contexte : On suppose que le caractère Y est binaire : $Y(\Omega) = \{0, 1\}$. Cela correspond à un codage.

Effectif :

On appelle effectif-population le nombre des individus dans U vérifiant $Y = 1$:

$$\eta_U = Np_U.$$

Estimation aléatoire de η_U :

Un estimateur aléatoire de η_U est

$$\eta_W = Np_W = N\frac{1}{n}\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in W\}}.$$

Espérance de η_W :

L'estimateur η_W est sans biais pour η_U :

$$\mathbb{E}(\eta_W) = \eta_U.$$

Preuve : Comme $\mathbb{E}(p_W) = p_U$, on a

$$\mathbb{E}(\eta_W) = \mathbb{E}(Np_W) = N\mathbb{E}(p_W) = Np_U = \eta_U.$$

□

Variance de η_W :

La variance de η_W est

$$\mathbb{V}(\eta_W) = N^2(1-f)\frac{N}{n(N-1)}p_U(1-p_U).$$

Preuve : Comme $\mathbb{V}(p_W) = (1-f)(N/n(N-1))p_U(1-p_U)$, on a

$$\mathbb{V}(\eta_W) = \mathbb{V}(Np_W) = N^2\mathbb{V}(p_W) = N^2(1-f)\frac{N}{n(N-1)}p_U(1-p_U).$$

□

Erreur quadratique moyenne de η_W :

L'erreur quadratique moyenne de τ_W est le réel :

$$EQM(\eta_W)[PESR] = N^2(1-f)\frac{N}{n(N-1)}p_U(1-p_U).$$

Estimation ponctuelle de η_U :

Soit ω un échantillon de n individus de U. Une estimation ponctuelle de η_U est le total-échantillon :

$$\eta_\omega = Np_\omega = N\frac{1}{n}\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i\in\omega\}}.$$

Estimation ponctuelle de l'écart-type de η_W :

Soit ω un échantillon de n individus de U. Une estimation ponctuelle de l'écart-type de η_W est le réel :

$$s(\eta_\omega) = \sqrt{N^2(1-f)\frac{p_\omega(1-p_\omega)}{n-1}}.$$

Intervalle de confiance pour η_U :

Soit ω un échantillon de n individus de U. Un intervalle de confiance pour η_U au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, est

$$\begin{aligned} i_{\eta_U} &= [\eta_\omega - z_\alpha s(\eta_\omega),\ \eta_\omega + z_\alpha s(\eta_\omega)] \\ &= \left[\eta_\omega - z_\alpha\sqrt{N^2(1-f)\frac{p_\omega(1-p_\omega)}{n-1}},\ \eta_\omega + z_\alpha\sqrt{N^2(1-f)\frac{p_\omega(1-p_\omega)}{n-1}}\right] = N\times i_{p_U}, \end{aligned}$$

où z_α est le réel vérifiant $\mathbb{P}(|Z|\geq z_\alpha)=\alpha$, $Z\sim\mathcal{N}(0,1)$.

On peut également définir l'incertitude absolue ou relative sur η_U, ainsi que la taille d'échantillon souhaitée pour une incertitude donnée.

3.4 Exercices corrigés

Exercice 1 : Sur un campus universitaire, un jour donné, on s'intéresse au total des montants dépensés par les 1765 étudiants du campus pour le repas du midi. On note ce total τ_U. Sur un échantillon ω de 279 étudiants prélevé suivant un plan de sondage aléatoire de type PESR, on obtient : $\overline{y}_\omega = 4.25$ € et $s_\omega = 2.15$ €.

1. Préciser le caractère étudié.
2. Calculer le taux de sondage.
3. Donner une estimation ponctuelle de τ_U.
4. Déterminer un intervalle de confiance pour τ_U au niveau 95%.

Solution :

1. On étudie le caractère $Y =$ "dépense d'un étudiant du campus pour le repas du midi" en €.
2. Le taux de sondage est
$$f = \frac{n}{N} = \frac{279}{1765} = 0.1580.$$
3. Une estimation ponctuelle de τ_U est
$$\tau_\omega = N\overline{y}_\omega = 1765 \times 4.25 = 7501.25.$$
4. On a $95\% = 100(1-\alpha)\%$ avec $\alpha = 0.05$. On a $\mathbb{P}(|Z| \geq z_\alpha) = \alpha = 0.05$, $Z \sim \mathcal{N}(0,1)$, avec $z_\alpha = 1.96$. Un intervalle de confiance pour τ_U au niveau 95% est
$$\begin{aligned} i_{\tau_U} &= \left[\tau_\omega - z_\alpha\sqrt{N^2(1-f)\frac{s_\omega^2}{n}},\ \tau_\omega + z_\alpha\sqrt{N^2(1-f)\frac{s_\omega^2}{n}}\right] \\ &= \left[7501.25 - 1.96\sqrt{1765^2\left(1-\frac{279}{1765}\right)\frac{2.15^2}{279}},\right. \\ &\quad \left. 7501.25 + 1.96\sqrt{1765^2\left(1-\frac{279}{1765}\right)\frac{2.15^2}{279}}\right] \\ &= [7092.673,\ 7909.827]. \end{aligned}$$

Ainsi, il y a 95 chances sur 100 que $[7092.673,\ 7909.827]$ contienne τ_U, l'unité étant le €.

Exercice 2 : Sur un campus universitaire de 1765 étudiants, un échantillon de 250 étudiants est prélevé suivant un plan de sondage aléatoire de type PESR. Parmi ces 250 étudiants, 189 admettent regarder la télévision plus de 1 heure par jour. On note p_U la proportion des 1765 étudiants qui admettent cela.

1. Calculer le taux de sondage.
2. Donner une estimation ponctuelle de p_U.
3. Déterminer un intervalle de confiance pour p_U au niveau 95%.
4. Déterminer la taille d'échantillon à choisir pour avoir une incertitude relative sur p_U inférieure ou égale à 5% au niveau 95%.

Solution :

1. Le taux de sondage est
$$f = \frac{n}{N} = \frac{250}{1765} = 0.1416.$$

2. Une estimation ponctuelle de p_U est
$$p_\omega = \frac{189}{250} = 0.756.$$

3. On a $95\% = 100(1-\alpha)\%$ avec $\alpha = 0.05$. On a $\mathbb{P}(|Z| \geq z_\alpha) = \alpha = 0.05$, $Z \sim \mathcal{N}(0,1)$, avec $z_\alpha = 1.96$. Un intervalle de confiance pour p_U au niveau 95% est
$$\begin{aligned} i_{p_U} &= \left[p_\omega - z_\alpha \sqrt{(1-f)\frac{p_\omega(1-p_\omega)}{n-1}},\ p_\omega + z_\alpha \sqrt{(1-f)\frac{p_\omega(1-p_\omega)}{n-1}} \right] \\ &= \left[0.756 - 1.96\sqrt{\left(1 - \frac{250}{1765}\right)\frac{0.756(1-0.756)}{250-1}}, \right. \\ &\quad \left. 0.756 + 1.96\sqrt{\left(1 - \frac{250}{1765}\right)\frac{0.756(1-0.756)}{250-1}} \right] \\ &= [0.7065,\ 0.8054]. \end{aligned}$$

   Ainsi, il y a 95 chances sur 100 que $[0.7065,\ 0.8054]$ contienne p_U.

4. On a $95\% = 100(1-\alpha)\%$ avec $\alpha = 0.05$. On souhaite déterminer le plus petit n tel que :
$$d_\omega^* \leq d_1 \qquad \Rightarrow \qquad n \geq \frac{N z_\alpha^2 p_\omega(1-p_\omega)}{N(p_\omega d_1)^2 + z_\alpha^2 p_\omega(1-p_\omega)},$$
   avec $d_1 = 0.05$, $z_\alpha = 1.96$, ω est l'échantillon considéré précédemment, $p_\omega = 0.756$ et $N = 1765$.
   On a
$$n \geq \frac{1765 \times 1.96^2 \times 0.756(1-0.756)}{1765 \times (0.756 \times 0.05)^2 + 1.96^2 \times 0.756(1-0.756)} = 387.1626.$$

Donc la taille d'échantillon à choisir pour avoir une incertitude relative sur p_U inférieure ou égale à 5% au niveau 95% est de $n = 388$.

3.5 Synthèse : proportion

Paramètres-population et les paramètres-échantillon correspondants :

	Population U	Échantillon ω
Taille	N	n
Taux de sondage	$\square$	$f = \dfrac{n}{N}$
Proportion	$p_U = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} y_i$	$p_\omega = \dfrac{1}{n}\displaystyle\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in \omega\}}$
Écart-type de p_W	$\sigma(p_W) = \sqrt{(1-f)\dfrac{N}{n(N-1)} p_U(1-p_U)}$	$s(p_\omega) = \sqrt{(1-f)\dfrac{p_\omega(1-p_\omega)}{n-1}}$

Autre notions utilisées autour de p_U (niveau : $100(1-\alpha)\%$, $\alpha \in]0,1[$) :

Intervalle de confiance	$i_{p_U} = \left[p_\omega - z_\alpha \sqrt{(1-f)\dfrac{p_\omega(1-p_\omega)}{n-1}},\ p_\omega + z_\alpha \sqrt{(1-f)\dfrac{p_\omega(1-p_\omega)}{n-1}}\right]$
Incertitude absolue	$d_\omega = z_\alpha \sqrt{(1-f)\dfrac{p_\omega(1-p_\omega)}{n-1}}$
Incertitude relative	$d_\omega^* = \dfrac{d_\omega}{p_\omega}$
Taille n telle que $d_\omega \leq d_0$	$n \geq \dfrac{N z_\alpha^2 p_\omega(1-p_\omega)}{N d_0^2 + z_\alpha^2 p_\omega(1-p_\omega)}$
Taille n telle que $d_\omega^* \leq d_1$	$n \geq \dfrac{N z_\alpha^2 p_\omega(1-p_\omega)}{N(p_\omega d_1)^2 + z_\alpha^2 p_\omega(1-p_\omega)}$

Rappel : $\mathbb{P}(|Z| \geq z_\alpha) = \alpha$, $Z \sim \mathcal{N}(0,1)$.

4 Plan de sondage aléatoire simple avec remise (PEAR)

4.1 Contexte

Loi de probabilité :

On prélève un échantillon de n individus suivant un plan de sondage aléatoire simple avec remise (PEAR pour Probabilités Egales Avec Remise) dans une population U. Soit W la *var* égale à l'échantillon obtenu :

$$W = (W_1, \ldots, W_n),$$

où, pour tout $m \in \{1, \ldots, n\}$, W_m est la *var* égale au m-ème individu de l'échantillon. Alors, pour tout $m \in \{1, \ldots, n\}$, la loi de W_i est donnée par

$$\mathbb{P}(W_m = u_i) = \frac{1}{N}, \qquad i \in \{1, \ldots, N\},$$

où $\mathbb{P}$ désigne la probabilité uniforme.

Preuve : L'univers associé à cette expérience aléatoire est $\Omega = \{u_1, \ldots, u_N\}^n$. Comme Ω est fini et que chaque individu a la même probabilité d'être prélevé, on considère la probabilité uniforme $\mathbb{P}$:

$$\mathbb{P}(W_m = u_i) = \frac{\text{Card}(\{W_m = u_i\})}{\text{Card}(\Omega)}.$$

On a $\text{Card}(\Omega) = N^n$. Les prélèvements étant avec remise, il y a N possibilités pour chacun des $n-1$ individus autres que u_i. Donc $\text{Card}(\{W_m = u_i\}) = N^{n-1}$. Il vient

$$\mathbb{P}(W_m = u_i) = \frac{N^{n-1}}{N^n} = \frac{1}{N}.$$

□

Situation de référence : On prélève au hasard et avec remise n individus pour former un échantillon. Chaque individu a la même probabilité qu'un autre d'être sélectionné.

Cette démarche est intéressante quand n est petit ou pour servir d'élément de comparaison avec une situation de type PESR.

Conditions habituelles d'estimation : Les formules habituelles d'estimation sont associées à un plan de sondage aléatoire de type PEAR. Elles sont aussi utilisées dans le cas SR (Sans Remise) lorsque n est beaucoup plus petit que N. Une convention existante est $N \geq 10n$.

Quelques commandes R : Par exemple, pour faire un tirage avec remise de $n = 20$ individus dans une population de $N = 200$ individus, on peut utiliser

- la commande `sample` :

```
sample(1:200, 20, replace = T)
```

- la commande `srswr` de la librairie `sampling` :

```
library(sampling)
t = srswr(20, 200)
x = 1:200
x[t != 0]
```

L'abréviation `srswr` signifie `Simple Random Sampling With Replacement`.

Précisons que `t = srswr(20, 200)` renvoie un vecteur de taille 200 constitué de chiffres entre 0 et 20. Les chiffres non nuls $m \in \{1, \ldots, 20\}$ sont positionnés aux indices des individus prélevés m fois et les 0 aux autres.

Un autre exemple : on considère la population U constituée de $N = 9$ garçons et on prélève un échantillon de $n = 3$ individus suivant un plan de sondage aléatoire de type PEAR :

```
U = c("Bob", "Nico", "Ali", "Fabien", "Malik", "John", "Jean", "Chris", "Karl")
library(sampling)
t = srswr(3, 9)
w = U[t != 0]
w
```

Dans la suite :

- pour les résultats, on considère un plan de sondage aléatoire de type PEAR et la *var* $W = (W_1, \ldots, W_m)$ égale à l'échantillon obtenu,
- pour les preuves, pour raison de simplicité, on se place dans la situation de référence.

Probabilités d'appartenance :

- pour tout $i \in \{1, \dots, N\}$,
$$\mathbb{P}(u_i \in W) = 1 - \left(1 - \frac{1}{N}\right)^n.$$
- pour tout $(i, j) \in \{1, \dots, N\}^2$ avec $i \neq j$,
$$\mathbb{P}((u_i, u_j) \in W) = 1 - 2\left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n.$$

Preuve :

- On a
$$\mathbb{P}(u_i \in W) = 1 - \mathbb{P}(u_i \notin W).$$
Par la définition de la probabilité uniforme, on a
$$\mathbb{P}(u_i \notin W) = \frac{\text{Card}(\{u_i \notin W\})}{\text{Card}(\Omega)}.$$
On a $\text{Card}(\Omega) = N^n$. Il reste à calculer $\text{Card}(\{u_i \notin W\})$. Le nombre de possibilités pour que u_i ne soit pas dans l'échantillon est égal au nombre de possibilités de choisir, pour chacun des n prélèvements, un individu parmi les $N - 1$ autres que u_i. D'où $\text{Card}(\{u_i \notin W\}) = (N-1)^n$. On en déduit que
$$\mathbb{P}(u_i \notin W) = \frac{(N-1)^n}{N^n} = \left(1 - \frac{1}{N}\right)^n.$$
Au final, on a
$$\mathbb{P}(u_i \in W) = 1 - \left(1 - \frac{1}{N}\right)^n.$$
- La formule d'inclusion-exclusion donne
$$\begin{aligned}\mathbb{P}((u_i, u_j) \in W) &= \mathbb{P}(\{u_i \in W\} \cap \{u_j \in W\}) \\ &= \mathbb{P}(u_i \in W) + \mathbb{P}(u_j \in W) - \mathbb{P}(\{u_i \in W\} \cup \{u_j \in W\}).\end{aligned}$$
Calculons chacune de ces probabilités. On a
$$\mathbb{P}(u_i \in W) = \mathbb{P}(u_j \in W) = 1 - \left(1 - \frac{1}{N}\right)^n.$$

D'autre part,

$$\mathbb{P}(\{u_i \in W\} \cup \{u_j \in W\}) = 1 - \mathbb{P}(\overline{\{u_i \in W\} \cup \{u_j \in W\}}) = 1 - \mathbb{P}(\{u_i \notin W\} \cap \{u_j \notin W\}).$$

Or

$$\mathbb{P}(\{u_i \notin W\} \cap \{u_j \notin W\}) = \frac{\text{Card}(\{u_i \notin W\} \cap \{u_j \notin W\})}{\text{Card}(\Omega)}.$$

On a $\text{Card}(\Omega) = N^n$. Il reste à calculer $\text{Card}(\{u_i \notin W\} \cap \{u_j \notin W\})$. Le nombre de possibilités pour que u_i et u_j ne soient pas dans l'échantillon est égal au nombre de possibilités de choisir, pour chacun des n prélèvements, un individu parmi les $N-2$ autres que u_i et u_j. D'où $\text{Card}(\{u_i \notin W\} \cap \{u_j \notin W\}) = (N-2)^n$. On en déduit que

$$\mathbb{P}(\{u_i \notin W\} \cap \{u_j \notin W\}) = \frac{(N-2)^n}{N^n} = \left(1 - \frac{2}{N}\right)^n.$$

Au final, on a

$$\begin{aligned}\mathbb{P}((u_i, u_j) \in W) &= 1 - \left(1 - \frac{1}{N}\right)^n + 1 - \left(1 - \frac{1}{N}\right)^n - \left(1 - \left(1 - \frac{2}{N}\right)^n\right) \\ &= 1 - 2\left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n.\end{aligned}$$

□

4.2 Estimateurs

Estimation aléatoire de $\overline{y}_U$:

Un estimateur aléatoire de $\overline{y}_U$ est

$$\overline{y}_W = \frac{1}{n} \sum_{i=1}^{N} y_i \sum_{m=1}^{n} \mathbb{1}_{\{W_m = u_i\}}.$$

Remarques : On peut également écrire cet estimateur

- sous la forme :

$$\overline{y}_W = \frac{1}{n} \sum_{i \in S} y_i,$$

où $S = \{(i_1, \ldots, i_n) \in \{1, \ldots, N\}^n; \quad u_{i_1} \in W, \ldots, u_{i_n} \in W\}$,

- sous la forme :

$$\overline{y}_W = \frac{1}{n}\sum_{m=1}^{n} Z_m, \qquad Z_m = \sum_{i=1}^{N} y_i \mathbb{1}_{\{W_m = u_i\}}.$$

On peut montrer que $Z_1, \ldots, Z_n$ sont des *var iid* avec

$$\mathbb{E}(Z_1) = \overline{y}_U, \qquad \mathbb{V}(Z_1) = \frac{N-1}{N} s_U^2.$$

On est donc dans les conditions habituelles d'estimation en posant $\mathbb{E}(Z_1) = \overline{y}_U = \mu$ et $\mathbb{V}(Z_1) = ((N-1)/N)s_U^2 = \sigma^2$.

Sous l'hypothèse que Y suit une loi normale et $n \geq 1$, il est raisonnable de penser que Z_m suit une loi normale.

Espérance de $\overline{y}_W$:

L'estimateur $\overline{y}_W$ est sans biais pour $\overline{y}_U$:

$$\mathbb{E}(\overline{y}_W) = \overline{y}_U.$$

Preuve : En utilisant la linéarité de l'espérance, $\mathbb{E}(\mathbb{1}_A) = \mathbb{P}(A)$ et $\mathbb{P}(W_m = u_i) = 1/N$, il vient

$$\begin{aligned}\mathbb{E}(\overline{y}_W) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{N} y_i \sum_{m=1}^{n} \mathbb{1}_{\{W_m=u_i\}}\right) = \frac{1}{n}\sum_{i=1}^{N} y_i \sum_{m=1}^{n} \mathbb{E}\left(\mathbb{1}_{\{W_m=u_i\}}\right) \\ &= \frac{1}{n}\sum_{i=1}^{N} y_i \sum_{m=1}^{n} \mathbb{P}(W_m = u_i) = \frac{1}{n}\sum_{i=1}^{N} y_i \frac{n}{N} = \frac{1}{N}\sum_{i=1}^{N} y_i = \overline{y}_U.\end{aligned}$$

□

Variance de $\overline{y}_W$:

La variance de $\overline{y}_W$ est

$$\mathbb{V}(\overline{y}_W) = \frac{N-1}{N}\frac{s_U^2}{n}.$$

Preuve : Les prélèvements étant avec remise et $\mathbb{P}(W_m = u_i) = 1/N$, les *var* $\mathbb{1}_{\{W_1=u_i\}}, \ldots, \mathbb{1}_{\{W_n=u_i\}}$ sont *iid*. Par conséquent, on a

$$\begin{aligned}\mathbb{V}(\overline{y}_W) &= \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{N} y_i \sum_{m=1}^{n} \mathbb{1}_{\{W_m=u_i\}}\right) = \frac{1}{n^2}\mathbb{V}\left(\sum_{m=1}^{n}\sum_{i=1}^{N} y_i \mathbb{1}_{\{W_m=u_i\}}\right) \\ &= \frac{1}{n^2}\sum_{m=1}^{n} \mathbb{V}\left(\sum_{i=1}^{N} y_i \mathbb{1}_{\{W_m=u_i\}}\right) = \frac{1}{n}\mathbb{V}\left(\sum_{i=1}^{N} y_i \mathbb{1}_{\{W_1=u_i\}}\right).\end{aligned}$$

En utilisant la formule de König-Huyghens et le fait que, pour tout $(i,j) \in \{1,\dots,N\}^2$ avec $i \neq j$, $\mathbb{1}_{\{W_1=u_i\}}\mathbb{1}_{\{W_1=u_j\}} = 0$, on obtient

$$\begin{aligned}
\mathbb{V}\left(\sum_{i=1}^N y_i \mathbb{1}_{\{W_1=u_i\}}\right) &= \mathbb{E}\left(\left(\sum_{i=1}^N y_i \mathbb{1}_{\{W_1=u_i\}}\right)^2\right) - \left(\mathbb{E}\left(\sum_{i=1}^N y_i \mathbb{1}_{\{W_1=u_i\}}\right)\right)^2 \\
&= \mathbb{E}\left(\sum_{i=1}^N \sum_{j=1}^N y_i y_j \mathbb{1}_{\{W_1=u_i\}}\mathbb{1}_{\{W_1=u_j\}}\right) - \left(\sum_{i=1}^N y_i \mathbb{E}\left(\mathbb{1}_{\{W_1=u_i\}}\right)\right)^2 \\
&= \sum_{i=1}^N y_i^2 \mathbb{E}\left(\mathbb{1}_{\{W_1=u_i\}}\right) - \left(\sum_{i=1}^N y_i \mathbb{E}\left(\mathbb{1}_{\{W_1=u_i\}}\right)\right)^2 \\
&= \sum_{i=1}^N y_i^2 \mathbb{P}(W_1=u_i) - \left(\sum_{i=1}^N y_i \mathbb{P}(W_1=u_i)\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^N y_i^2 - \left(\frac{1}{N}\sum_{i=1}^N y_i\right)^2.
\end{aligned}$$

D'autre part, on a

$$\begin{aligned}
s_U^2 &= \frac{1}{N-1}\sum_{i=1}^N (y_i - \overline{y}_U)^2 = \frac{1}{N-1}\left(\sum_{i=1}^N y_i^2 - 2\overline{y}_U \sum_{i=1}^N y_i + N\overline{y}_U^2\right) \\
&= \frac{1}{N-1}\left(\sum_{i=1}^N y_i^2 - 2N\overline{y}_U^2 + N\overline{y}_U^2\right) = \frac{1}{N-1}\left(\sum_{i=1}^N y_i^2 - N\overline{y}_U^2\right) \\
&= \frac{1}{N-1}\left(\sum_{i=1}^N y_i^2 - N\left(\frac{1}{N}\sum_{i=1}^N y_i\right)^2\right).
\end{aligned}$$

Il s'ensuit

$$\mathbb{V}\left(\sum_{i=1}^N y_i \mathbb{1}_{\{W_1=u_i\}}\right) = \frac{N-1}{N} s_U^2.$$

Au final, il vient

$$\mathbb{V}(\overline{y}_W) = \frac{N-1}{N}\frac{s_U^2}{n}.$$

□

Erreur quadratique moyenne de $\overline{y}_W$:

L'erreur quadratique moyenne de $\overline{y}_W$ est le réel :

$$EQM(\overline{y}_W)[PEAR] = \frac{N-1}{N}\frac{s_U^2}{n}.$$

On constate que :

- plus n est grand/l'échantillon est grand, plus $\overline{y}_W$ estime bien $\overline{y}_U$,
- plus U est homogène/plus s_U^2 est petit, plus $\overline{y}_W$ estime bien $\overline{y}_U$.

Remarque : L'estimation de $\overline{y}_U$ par $\overline{y}_W$ est plus précise avec un plan de sondage aléatoire de type PESR que d'un plan de sondage aléatoire de type PEAR. En effet, en évaluant les erreurs quadratiques moyennes, on a :

$$EQM(\overline{y}_W)[PESR] = (1-f)\frac{s_U^2}{n} = \left(1-\frac{n}{N}\right)\frac{s_U^2}{n}$$

et

$$EQM(\overline{y}_W)[PEAR] = \frac{N-1}{N}\frac{s_U^2}{n} = \left(1-\frac{1}{N}\right)\frac{s_U^2}{n}.$$

Donc

$$EQM(\overline{y}_W)[PESR] \leq EQM(\overline{y}_W)[PEAR].$$

L'estimation de $\overline{y}_U$ par $\overline{y}_W$ commet donc moins d'erreur dans le cadre PESR que dans le cadre PEAR.

Estimation aléatoire de s_U :

Un estimateur aléatoire de s_U est

$$s_W = \sqrt{\frac{1}{n-1}\sum_{i=1}^{N}(y_i-\overline{y}_W)^2\sum_{m=1}^{n}\mathbb{1}_{\{W_m=u_i\}}}.$$

Propriété de s_W^2 :

L'estimateur s_W^2 est sans biais pour $((N-1)/N)s_U^2$:

$$\mathbb{E}\left(s_W^2\right) = \frac{N-1}{N}s_U^2.$$

Preuve : En remarquant que $\sum_{i=1}^{N}\sum_{m=1}^{n}\mathbb{1}_{\{W_m=u_i\}} = n$, il vient

$$\begin{aligned}
s_W^2 &= \frac{1}{n-1}\sum_{i=1}^{N}(y_i-\overline{y}_W)^2\sum_{m=1}^{n}\mathbb{1}_{\{W_m=u_i\}} \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{N}y_i^2\sum_{m=1}^{n}\mathbb{1}_{\{W_m=u_i\}} - 2\overline{y}_W\sum_{i=1}^{N}y_i\sum_{m=1}^{n}\mathbb{1}_{\{W_m=u_i\}} + \overline{y}_W^2\sum_{i=1}^{N}\sum_{m=1}^{n}\mathbb{1}_{\{W_m=u_i\}}\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{N}y_i^2\sum_{m=1}^{n}\mathbb{1}_{\{W_m=u_i\}} - 2n\overline{y}_W^2 + n\overline{y}_W^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{N}y_i^2\sum_{m=1}^{n}\mathbb{1}_{\{W_m=u_i\}} - n\overline{y}_W^2\right).
\end{aligned}$$

On a $\mathbb{P}(W_m = u_i) = 1/N$ et

$$\mathbb{E}\left(\overline{y}_W^2\right) = \mathbb{V}\left(\overline{y}_W^2\right) + \left(\mathbb{E}\left(\overline{y}_W\right)\right)^2 = \frac{N-1}{N}\frac{s_U^2}{n} + \overline{y}_U^2.$$

D'où

$$\begin{aligned}
\mathbb{E}\left(s_W^2\right) &= \mathbb{E}\left(\frac{1}{n-1}\left(\sum_{i=1}^N y_i^2 \sum_{m=1}^n \mathbb{1}_{\{W_m=u_i\}} - n\overline{y}_W^2\right)\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^N y_i^2 \sum_{m=1}^n \mathbb{E}\left(\mathbb{1}_{\{W_m=u_i\}}\right) - n\mathbb{E}\left(\overline{y}_W^2\right)\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^N y_i^2 \sum_{m=1}^n \mathbb{P}\left(W_m=u_i\right) - n\mathbb{E}\left(\overline{y}_W^2\right)\right) \\
&= \frac{1}{n-1}\left(\frac{n}{N}\sum_{i=1}^N y_i^2 - n\left(\frac{N-1}{N}\frac{s_U^2}{n} + \overline{y}_U^2\right)\right) \\
&= \frac{1}{n-1}\left(\frac{n}{N}\left(\sum_{i=1}^N y_i^2 - N\overline{y}_U^2\right) - \left(1-\frac{1}{N}\right)s_U^2\right) \\
&= \frac{n(N-1)}{(n-1)N}\left(\frac{1}{N-1}\left(\sum_{i=1}^N y_i^2 - N\overline{y}_U^2\right)\right) - \frac{1}{n-1}\left(1-\frac{1}{N}\right)s_U^2.
\end{aligned}$$

En remarquant que

$$\begin{aligned}
s_U^2 &= \frac{1}{N-1}\sum_{i=1}^N (y_i - \overline{y}_U)^2 = \frac{1}{N-1}\left(\sum_{i=1}^N y_i^2 - 2\overline{y}_U\sum_{i=1}^N y_i + N\overline{y}_U^2\right) \\
&= \frac{1}{N-1}\left(\sum_{i=1}^N y_i^2 - 2N\overline{y}_U^2 + N\overline{y}_U^2\right) = \frac{1}{N-1}\left(\sum_{i=1}^N y_i^2 - N\overline{y}_U^2\right).
\end{aligned}$$

D'où

$$\mathbb{E}\left(s_W^2\right) = \frac{n(N-1)}{(n-1)N}s_U^2 - \frac{1}{n-1}\left(1-\frac{1}{N}\right)s_U^2 = \frac{n(N-1)-(N-1)}{(n-1)N}s_U^2 = \frac{N-1}{N}s_U^2.$$

□

4.3 Estimations ponctuelles

Estimation ponctuelle de $\overline{y}_U$:

Soit $\omega = (\omega_1, \ldots, \omega_n)$ un échantillon de n individus de U. Une estimation ponctuelle de $\overline{y}_U$ est la moyenne-échantillon :

$$\overline{y}_\omega = \frac{1}{n} \sum_{i=1}^{N} y_i \sum_{m=1}^{n} \mathbb{1}_{\{\omega_m = u_i\}}.$$

Quelques commandes R : Un exemple de calcul de $\overline{y}_\omega$ avec R est décrit ci-dessous :

```
U = c("Bob", "Nico", "Ali", "Fabien", "Malik", "John", "Jean", "Chris", "Karl")
y = c(72, 89, 68, 74, 81, 87, 76, 61, 84)
n = 3
library(sampling)
t = srswr(n, 9)
bar_y_w = (1 / n) * sum(y * t)
bar_y_w
```

Erreur d'estimation :

Soit ω un échantillon de n individus de U. L'erreur d'estimation que commet $\overline{y}_\omega$ en estimant $\overline{y}_U$ est le réel :

$$e_\omega = |\overline{y}_\omega - \overline{y}_U|.$$

Probabilité d'erreur :

La probabilité de se tromper de plus de $(100 \times \beta)\%$, $\beta \in]0, 1[$, en estimant $\overline{y}_U$ par $\overline{y}_W$ est le réel :

$$p_\beta = \frac{1}{N^n} \sum_{\omega \in W(\Omega)} \mathbb{1}_{\{e_\omega \geq \beta \overline{y}_U\}}.$$

Estimation ponctuelle de s_U :

Soit $\omega = (\omega_1, \ldots, \omega_n)$ un échantillon de n individus de U. Une estimation ponctuelle de s_U est l'écart-type corrigé-échantillon :

$$s_\omega = \sqrt{\frac{1}{n-1} \sum_{i=1}^{N} (y_i - \overline{y}_\omega)^2 \sum_{m=1}^{n} \mathbb{1}_{\{\omega_m = u_i\}}}.$$

Quelques commandes R : Un exemple de calcul de s_ω avec R est décrit ci-dessous :

```
U = c("Bob", "Nico", "Ali", "Fabien", "Malik", "John", "Jean", "Chris", "Karl")
y = c(72, 89, 68, 74, 81, 87, 76, 61, 84)
n = 3
library(sampling)
t = srswr(n, 9)
bar_y_w = (1 / n) * sum(y * t)
s_w = sqrt(sum((y - bar_y_w)^2 * t) / (n - 1))
s_w
```

Estimation ponctuelle de l'écart-type de $\overline{y}_W$:

Soit ω un échantillon de n individus de U. Une estimation ponctuelle de l'écart-type de $\overline{y}_W$ est le réel :

$$s(\overline{y}_\omega) = \sqrt{\frac{s_\omega^2}{n}}.$$

4.4 Intervalles de confiance

Résultat en loi : Si on peut admettre que Y suit une loi normale, alors

$$T = \frac{\overline{y}_W - \overline{y}_U}{\sqrt{\frac{s_W^2}{n}}} \sim \mathcal{T}(\nu),$$

où $\mathcal{T}(\nu)$ désigne la loi de Student à $\nu = n - 1$ degrés de liberté.

Si n est grand, on peut utiliser l'approximation $\mathcal{T}(\nu) \approx \mathcal{N}(0, 1)$.

T-intervalle de confiance pour $\overline{y}_U$:

Soit ω un échantillon de n individus de U. On suppose que Y suit une loi normale. Un intervalle de confiance pour $\overline{y}_U$ au niveau $100(1-\alpha)\%$, $\alpha \in]0, 1[$, est

$$\begin{aligned} i_{\overline{y}_U} &= [\overline{y}_\omega - t_\alpha(\nu) s(\overline{y}_\omega),\ \overline{y}_\omega + t_\alpha(\nu) s(\overline{y}_\omega)] \\ &= \left[\overline{y}_\omega - t_\alpha(\nu)\sqrt{\frac{s_\omega^2}{n}},\ \overline{y}_\omega + t_\alpha(\nu)\sqrt{\frac{s_\omega^2}{n}}\right], \end{aligned}$$

où $t_\alpha(\nu)$ est le réel vérifiant $\mathbb{P}(|T| \geq t_\alpha(\nu)) = \alpha$, $T \sim \mathcal{T}(\nu)$, $\nu = n - 1$.

Quelques commandes R : Un exemple de fonction R pour calculer le T-intervalle de confiance pour $\overline{y}_U$ au niveau $100(1-\alpha)\%$ est décrit ci-dessous :

```
icPEAR = function(y, N, niveau) {
n = length(y)
nu = n - 1
bar_y_w = mean(y)
t = qt(1 - (1 - niveau) / 2, nu)
s2_w = sd(y)^2
var_bar_y_w = s2_w / n
a = bar_y_w - t * sqrt(var_bar_y_w)
b = bar_y_w + t * sqrt(var_bar_y_w)
print(c(a, b)) }
icPEAR(y = c(2.1, 2.3, 4.1, 2.6, 7.1, 8.6), N = 100, niveau = 0.95)
```

Cela renvoie : 1.576111, 7.357222.

Une autre possibilité utilisant des fonctions existantes est

```
y = c(2.1, 2.3, 4.1, 2.6, 7.1, 8.6)
t.test(y, conf.level = 0.95)$conf.int
```

Cela renvoie la même chose que précédemment : 1.576111, 7.357222.

Résultat limite : Si n est suffisamment grand, sans hypothèse de loi normale sur Y, on a l'approximation :

$$Z = \frac{\overline{y}_W - \overline{y}_U}{\sqrt{\frac{s_W^2}{n}}} \approx \mathcal{N}(0,1).$$

Intervalle de confiance (limite) pour $\overline{y}_U$:

Soit ω un échantillon de n individus de U. Un intervalle de confiance pour $\overline{y}_U$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, est

$$\begin{aligned} i_{\overline{y}_U} &= [\overline{y}_\omega - z_\alpha s(\overline{y}_\omega),\ \overline{y}_\omega + z_\alpha s(\overline{y}_\omega)] \\ &= \left[\overline{y}_\omega - z_\alpha \sqrt{\frac{s_\omega^2}{n}},\ \overline{y}_\omega + z_\alpha \sqrt{\frac{s_\omega^2}{n}}\right], \end{aligned}$$

où z_α est le réel vérifiant $\mathbb{P}(|Z| \geq z_\alpha) = \alpha$, $Z \sim \mathcal{N}(0,1)$.

Quelques commandes R : Un exemple de fonction R pour calculer l'intervalle de confiance limite pour $\overline{y}_U$ au niveau $100(1-\alpha)\%$ est décrit ci-dessous :

```
icPEAR2 = function(y, N, niveau) {
n = length(y)
bar_y_w = mean(y)
t = qnorm(1 - (1 - niveau) / 2)
s2_w = sd(y)^2
var_bar_y_w = s2_w / n
a = bar_y_w - t * sqrt(var_bar_y_w)
b = bar_y_w + t * sqrt(var_bar_y_w)
print(c(a, b)) }
icPEAR2(y = c(2.1, 2.3, 4.1, 2.6, 7.1, 8.6, 2.1, 2.3, 4.1, 2.6, 7.1, 8.6),
N = 100, niveau = 0.95)
```

Cela renvoie : 2.980777, 5.952557.

Résultat en loi : Si on peut admettre que Y suit une loi normale, alors

$$K = (n-1)\frac{s_W^2}{s_U^2} \sim \chi^2(\nu),$$

où $\chi^2(\nu)$ désigne la loi du Chi-deux à $\nu = n-1$ degrés de liberté.

Intervalle de confiance pour s_U^2 :

Soit ω un échantillon de n individus de U. On suppose que Y suit une loi normale. Un intervalle de confiance pour s_U^2 au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, est

$$i_{s_U^2} = \left[\frac{n-1}{b_\alpha(\nu)}s_\omega^2,\ \frac{n-1}{a_\alpha(\nu)}s_\omega^2\right],$$

où $a_\alpha(\nu)$ et $b_\alpha(\nu)$ sont les réels vérifiant :

$$\mathbb{P}(K \geq a_\alpha(\nu)) = 1 - \frac{\alpha}{2}, \qquad \mathbb{P}(K \geq b_\alpha(\nu)) = \frac{\alpha}{2},$$

$K \sim \chi^2(\nu)$, $\nu = n-1$.

Remarque : Les tests statistiques habituels s'appliquent (T-Test, Z-Test, ...).

4.5 Taille d'échantillon

Incertitude absolue :

Soit ω un échantillon de n individus de U. On appelle incertitude absolue sur $\overline{y}_U$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, la demi-longueur de $i_{\overline{y}_U}$ (limite) :

$$d_\omega = z_\alpha s(\overline{y}_\omega) = z_\alpha \sqrt{\frac{s^2_\omega}{n}}.$$

Plus d_ω est petit, plus l'estimation de $\overline{y}_U$ par $\overline{y}_\omega$ est précise.

Incertitude relative :

Soit ω un échantillon de n individus de U et d_ω l'incertitude absolue sur $\overline{y}_U$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$. On appelle incertitude relative sur $\overline{y}_U$ au niveau $100(1-\alpha)\%$ le pourcentage $(100 \times d^*_\omega)\%$ où d^*_ω est le réel :

$$d^*_\omega = \frac{d_\omega}{\overline{y}_\omega}.$$

Taille d'échantillon :

Soit ω un échantillon prélevé lors d'une étude préliminaire. La taille d'échantillon n à choisir pour avoir :

- une incertitude absolue sur $\overline{y}_U$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, inférieure ou égale à d_0 est le plus petit n tel que
$$d_\omega \leq d_0 \qquad \Leftrightarrow \qquad n \geq \left(\frac{z_\alpha s_\omega}{d_0}\right)^2,$$
- une incertitude relative sur $\overline{y}_U$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, inférieure ou égale à $(100 \times d_1)\%$ est le plus petit n tel que
$$d^*_\omega \leq d_1 \qquad \Leftrightarrow \qquad n \geq \left(\frac{z_\alpha s_\omega}{\overline{y}_\omega d_1}\right)^2.$$

Quelques commandes R : Un exemple de fonction R pour calculer la taille n d'un échantillon à partir de l'incertude absolue de $\overline{y}_U$ au niveau $100(1-\alpha)\%$ est décrit ci-dessous :

```
n_ech = function(N, s2, d0, niveau) {
z = qnorm(1 - (1 - niveau) / 2)
n = s2 * z^2 / d0^2
print(ceiling(n)) }
n_ech(N = 100, s2 = 63, d0 = 3, niveau = 0.95)
```

Cela renvoie 27.

4.6 Exercices corrigés

Exercice 1 : On considère le caractère Y = "âge" en années dans la population de 4 individus : U = {Marcel, Christian, Jean, Seb} = $\{u_1, \ldots, u_4\}$. Pour tout $i \in \{1, \ldots, 4\}$, soit y_i la valeur de Y pour l'individu u_i. Les résultats, en années, sont :

y_1	y_2	y_3	y_4
33	34	29	37

1. Calculer la moyenne-population $\overline{y}_U$ et l'écart-type corrigé-population s_U.
2. On prélève au hasard et avec remise 2 individus dans U formant ainsi un échantillon. Chaque individu a la même probabilité qu'un autre d'être sélectionné. On est donc dans le cadre d'un plan de sondage aléatoire de type PEAR.
   (a) Combien d'échantillons peut-on former ? Expliciter les.
   (b) Calculer la probabilité que Marcel appartienne à un tel échantillon.
   (c) Pour chaque échantillon ω, calculer la moyenne-échantillon $\overline{y}_\omega$ et l'écart-type corrigé-échantillon s_ω.
   (d) Soit $\overline{y}_W$ la *var* égale à la moyenne-échantillon, l'aléatoire étant dans l'échantillon considéré. Déterminer sa loi, puis calculer son espérance et sa variance.
   (e) Soit s_W la *var* égale à l'écart-type corrigé-échantillon, l'aléatoire étant dans l'échantillon considéré. Calculer l'espérance de s_W^2.
   (f) Retrouver les résultats des deux questions précédentes avec les formules du cours.

Solution :

1. On a
$$\overline{y}_U = 33.25, \qquad s_U = 3.3040.$$
2. (a) Vu le mode de prélèvement, le nombre d'échantillons possible est
$$4^2 = 16.$$

Ils sont :

(u_1, u_1)	(u_1, u_2)	(u_1, u_3)	(u_1, u_4)
(u_2, u_1)	(u_2, u_2)	(u_2, u_3)	(u_2, u_4)
(u_3, u_1)	(u_3, u_2)	(u_3, u_3)	(u_3, u_4)
(u_4, u_1)	(u_4, u_2)	(u_4, u_3)	(u_4, u_4)

(b) Il y a 7 échantillons contenant u_1 = Marcel. Comme il y a un total de 16 échantillons possibles, la probabilité que Marcel appartienne à un tel échantillon est $7/16 = 0.4375$.

On peut retrouver ce résultat avec la formule :

$$\mathbb{P}(u_1 \in W) = 1 - \left(1 - \frac{1}{N}\right)^n = 1 - \left(1 - \frac{1}{4}\right)^2 = \frac{7}{16} = 0.4375.$$

(c) On a, en prenant 4 chiffres après la virgule :

ω	Y	$\overline{y}_\omega$	s_ω
(u_1, u_1)	$(33, 33)$	33	0
(u_1, u_2)	$(33, 34)$	33.5	0.7071
(u_1, u_3)	$(33, 29)$	31	2.8284
(u_1, u_4)	$(33, 37)$	35	2.8284
(u_2, u_1)	$(34, 33)$	33.5	0.7071
(u_2, u_2)	$(34, 34)$	34	0
(u_2, u_3)	$(34, 29)$	31.5	3.5355
(u_2, u_4)	$(34, 37)$	35.5	2.1213
(u_3, u_1)	$(29, 33)$	31	2.8284
(u_3, u_2)	$(29, 34)$	31.5	3.5355
(u_3, u_3)	$(29, 29)$	29	0
(u_3, u_4)	$(29, 37)$	33	5.6568
(u_4, u_1)	$(37, 33)$	35	2.8284
(u_4, u_2)	$(37, 34)$	35.5	2.1213
(u_4, u_3)	$(37, 29)$	33	5.6568
(u_4, u_4)	$(37, 37)$	37	0

(d) Soit $\overline{y}_W$ la *var* égale à la moyenne-échantillon. L'ensemble des valeurs possibles pour $\overline{y}_W$ est

$$\overline{y}_W(\Omega) = \{29, 31, 31.5, 33, 33.5, 34, 35, 35.5, 37\}.$$

Comme il y a 16 échantillons différents et qu'ils sont équiprobables, la loi de $\overline{y}_W$ est donnée par

k	29	31	31.5	33	33.5	34	35	35.5	37
$\mathbb{P}(\overline{y}_W = k)$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

En utilisant la loi de $\overline{y}_W$, l'espérance de $\overline{y}_W$ est

$$\begin{aligned}\mathbb{E}(\overline{y}_W) &= \sum_{k\in\overline{y}_W(\Omega)} k\mathbb{P}(\overline{y}_W = k)\\ &= \frac{1}{16}(29 + 31\times 2 + 31.5\times 2 + 33\times 3 + 33.5\times 2 + 34 + 35\times 2\\ &+ 35.5\times 2 + 37)\\ &= 33.25.\end{aligned}$$

En utilisant la formule de König-Huyghens, la variance de $\overline{y}_W$ est

$$\mathbb{V}(\overline{y}_W) = \mathbb{E}(\overline{y}_W^2) - \left(\mathbb{E}(\overline{y}_W)\right)^2.$$

Or on a $\mathbb{E}(\overline{y}_W) = 33.25$ et

$$\begin{aligned}\mathbb{E}(\overline{y}_W^2) &= \sum_{k\in\overline{y}_W(\Omega)} k^2\mathbb{P}(\overline{y}_W = k)\\ &= \frac{1}{16}(29^2 + 31^2\times 2 + 31.5^2\times 2 + 33^2\times 3 + 33.5^2\times 2 + 34^2\\ &+ 35^2\times 2 + 35.5^2\times 2 + 37^2)\\ &= 1109.656.\end{aligned}$$

D'où

$$\mathbb{V}(\overline{y}_W) = 1109.656 - 33.25^2 = 4.0935.$$

(e) Soit s_W la *var* égale à l'écart-type corrigé-échantillon. L'ensemble des valeurs possibles pour s_W est

$$s_W(\Omega) = \{0, 0.7071, 2.1213, 2.8284, 3.5355, 5.6568\}.$$

Comme il y a 16 échantillons différents et qu'ils sont équiprobables, la loi de s_W est donnée par

k	0	0.7071	2.1213	2.8284	3.5355	5.6568
$\mathbb{P}(s_W = k)$	$\frac{4}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{4}{16}$	$\frac{2}{16}$	$\frac{2}{16}$

L'espérance de s_W^2 est

$$\begin{aligned}\mathbb{E}(s_W^2) &= \sum_{k\in s_W(\Omega)} k^2\mathbb{P}(s_W = k)\\ &= \frac{1}{16}(0^2\times 4 + 0.7071^2\times 2 + 2.1213^2\times 2 + 2.8284^2\times 4 + 3.5355^2\times 2\\ &+ 5.6568^2\times 2)\\ &= 8.1873.\end{aligned}$$

(f) En utilisant les formules du cours, on retrouve les résultats précédents (en prenant en compte les approximations) :

$$\mathbb{E}(\overline{y}_W) = \overline{y}_U = 33.25, \qquad \mathbb{V}(\overline{y}_W) = \frac{N-1}{N}\frac{s_U^2}{n} = \frac{3}{4}\times\frac{3.3040^2}{2} = 4.0936$$

et

$$\mathbb{E}(s_W^2) = \frac{N-1}{N}s_U^2 = \frac{3}{4}\times 3.3040^2 = 8.1873.$$

Exercice 2 : Sur les 80 sacs de pommes de terre d'une petite production, on prélève un échantillon de 17 sacs suivant un plan de sondage aléatoire de type PEAR. On pèse ces 17 sacs. Les valeurs obtenues donnent une moyenne de 22.53 kilogrammes et un écart-type corrigé de 1.25 kilogrammes. On suppose que le poids en kilogrammes d'un sac de pommes de terre issu de cette production peut être modélisé par une *var* Y suivant une loi normale.

1. Déterminer un intervalle de confiance pour la moyenne des poids des 80 sacs de la production au niveau 90%.
2. Déterminer la taille d'échantillon à choisir pour avoir une incertitude absolue sur la moyenne des poids des 80 sacs inférieure ou égale à 0.5 au niveau 90%.

Solution :

1. On a $90\% = 100(1-\alpha)\%$ avec $\alpha = 0.1$. On a $\mathbb{P}(|T| \geq t_\alpha(\nu)) = \alpha = 0.1$, $T \sim \mathcal{T}(\nu)$, $\nu = n-1 = 17-1 = 16$ avec $t_\alpha(\nu) = 1.746$.

Un intervalle de confiance pour $\overline{y}_U$ au niveau 90% est

$$\begin{aligned} i_{\overline{y}_U} &= \left[\overline{y}_\omega - t_\alpha(\nu)\sqrt{\frac{s_\omega^2}{n}},\ \overline{y}_\omega + t_\alpha(\nu)\sqrt{\frac{s_\omega^2}{n}}\right] \\ &= \left[22.53 - 1.746\sqrt{\frac{1.25^2}{17}},\ 22.53 + 1.746\sqrt{\frac{1.25^2}{17}}\right] \\ &= [22.0006,\ 23.0593]. \end{aligned}$$

Ainsi, il y a 90 chances sur 100 que $[22.0006,\ 23.0593]$ contienne $\overline{y}_U$.

2. On a $90\% = 100(1-\alpha)\%$ avec $\alpha = 0.1$. On souhaite déterminer le plus petit n tel que :

$$d_\omega = z_\alpha\sqrt{\frac{s_\omega^2}{n}} \leq d_0 \qquad \Leftrightarrow \qquad n \geq \left(\frac{z_\alpha s_\omega}{d_0}\right)^2,$$

avec $d_0 = 0.5$, $z_\alpha = 1.645$, ω est l'échantillon considéré précédemment et $s_\omega = 1.25$. On a

$$\left(\frac{1.645 \times 1.25}{0.5}\right)^2 = 16.9127.$$

Donc $n = 17$ convient.

Exercice 3 : On demande à 60 élèves de maternelle de reproduire 16 dessins. On s'intéresse au temps en secondes mis par un élève. On considère un échantillon de 7 élèves suivant un plan de sondage aléatoire de type PEAR. Les résultats, en secondes, sont :

376	389	407	401	397	360	410

On suppose que le temps en secondes que met un élève de maternelle pour reproduire ces 16 dessins peut être modélisé par une *var* Y suivant une loi normale.

1. Déterminer un intervalle de confiance pour la moyenne des temps des 60 élèves au niveau 99%.
2. Proposer des commandes R donnant le résultat de la question précédente.

Solution :

1. On a $99\% = 100(1-\alpha)\%$ avec $\alpha = 0.01$. On a

$$\overline{y}_\omega = 391.4286, \qquad s_\omega = 17.9894.$$

On a $\mathbb{P}(|T| \geq t_\alpha(\nu)) = \alpha = 0.01$, $T \sim \mathcal{T}(\nu)$, $\nu = n - 1 = 7 - 1 = 6$ avec $t_\alpha(\nu) = 3.707$.

Un intervalle de confiance pour $\overline{y}_U$ au niveau 99% est

$$\begin{aligned} i_{\overline{y}_U} &= \left[\overline{y}_\omega - t_\alpha(\nu)\sqrt{\frac{s^2_\omega}{n}},\ \overline{y}_\omega + t_\alpha(\nu)\sqrt{\frac{s^2_\omega}{n}}\right] \\ &= \left[391.4286 - 3.707\sqrt{\frac{17.9894^2}{7}},\ 391.4286 + 3.707\sqrt{\frac{17.9894^2}{7}}\right] \\ &= [366.2234,\ 416.6338]. \end{aligned}$$

Ainsi, il y a 99 chances sur 100 que $[366.2234,\ 416.6338]$ contienne $\overline{y}_U$.

2. On propose :

```
y = c(376, 389, 407, 401, 397, 360, 410)
t.test(y, conf.level = 0.99)$conf.int
```

4.7 Synthèse

Paramètres-population et les paramètres-échantillon correspondants :

	Population U	Échantillon $\omega = (\omega_1, \ldots, \omega_n)$
Taille	N	n
Moyenne	$\overline{y}_U = \frac{1}{N}\sum_{i=1}^{N} y_i$	$\overline{y}_\omega = \frac{1}{n}\sum_{i=1}^{N} y_i \sum_{m=1}^{n} \mathbb{1}_{\{\omega_m = u_i\}}$
Écart-type corrigé	$s_U = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y}_U)^2}$	$s_\omega = \sqrt{\frac{1}{n-1}\sum_{i=1}^{N}(y_i - \overline{y}_\omega)^2 \sum_{m=1}^{n} \mathbb{1}_{\{\omega_m = u_i\}}}$
Écart-type de $\overline{y}_W$	$\sigma(\overline{y}_W) = \sqrt{\frac{N-1}{N}\frac{s_U^2}{n}}$	$s(\overline{y}_\omega) = \sqrt{\frac{s_\omega^2}{n}}$

Autre notions utilisées autour de $\overline{y}_U$ (niveau : $100(1-\alpha)\%$, $\alpha \in]0,1[$) :

Intervalle de confiance	$i_{\overline{y}_U} = \left[\overline{y}_\omega - z_\alpha\sqrt{\frac{s_\omega^2}{n}},\ \overline{y}_\omega + z_\alpha\sqrt{\frac{s_\omega^2}{n}}\right]$
Incertitude absolue	$d_\omega = z_\alpha\sqrt{\frac{s_\omega^2}{n}}$
Incertitude relative	$d_\omega^* = \frac{d_\omega}{\overline{y}_\omega}$
Taille n telle que $d_\omega \leq d_0$	$n \geq \left(\frac{z_\alpha s_\omega}{d_0}\right)^2$
Taille n telle que $d_\omega^* \leq d_1$	$n \geq \left(\frac{z_\alpha s_\omega}{\overline{y}_\omega d_1}\right)^2$

Rappel : $\mathbb{P}(|Z| \geq z_\alpha) = \alpha$, $Z \sim \mathcal{N}(0,1)$.

5 Total, proportion et effectif dans le cadre PEAR

On reprend le cadre mathématique d'un plan de sondage aléatoire de type PEAR.

5.1 Estimation du total

Total :

On appelle total-population le réel :

$$\tau_U = \sum_{i=1}^{N} y_i = N\overline{y}_U.$$

Estimation aléatoire de τ_U :

Un estimateur aléatoire de τ_U est

$$\tau_W = N\overline{y}_W = N\frac{1}{n}\sum_{i=1}^{N} y_i \sum_{m=1}^{n} \mathbb{1}_{\{W_m = u_i\}}.$$

Espérance de τ_W :

L'estimateur τ_W est sans biais pour τ_U :

$$\mathbb{E}(\tau_W) = \tau_U.$$

Variance de τ_W :

La variance de τ_W est

$$\mathbb{V}(\tau_W) = N^2 \frac{N-1}{N}\frac{s_U^2}{n}.$$

Erreur quadratique moyenne de τ_W :

L'erreur quadratique moyenne de τ_W est le réel :

$$EQM(\tau_W)[PEAR] = N^2 \frac{N-1}{N}\frac{s_U^2}{n}.$$

Estimation ponctuelle de τ_U :

Soit $\omega = (\omega_1, \ldots, \omega_n)$ un échantillon de n individus de U. Une estimation ponctuelle de τ_U est le total-échantillon :

$$\tau_\omega = N\overline{y}_\omega = N\frac{1}{n}\sum_{i=1}^{N} y_i \sum_{m=1}^{n} \mathbb{1}_{\{\omega_m = u_i\}}.$$

Estimation ponctuelle de l'écart-type de τ_W :

Soit $\omega = (\omega_1, \ldots, \omega_n)$ un échantillon de n individus de U. Une estimation ponctuelle de l'écart-type de τ_W est le réel :

$$s(\tau_\omega) = \sqrt{N^2 \frac{s_\omega^2}{n}}.$$

Intervalle de confiance pour τ_U :

Soit $\omega = (\omega_1, \ldots, \omega_n)$ un échantillon de n individus de U. On suppose que Y suit une loi normale. Un intervalle de confiance pour τ_U au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, est

$$\begin{aligned} i_{\tau_U} &= [\tau_\omega - t_\alpha(\nu)s(\tau_\omega),\ \tau_\omega + t_\alpha(\nu)s(\tau_\omega)] \\ &= \left[\tau_\omega - t_\alpha(\nu)\sqrt{N^2\frac{s_\omega^2}{n}},\ \tau_\omega + t_\alpha(\nu)\sqrt{N^2\frac{s_\omega^2}{n}}\right] = N \times i_{\overline{y}_U}, \end{aligned}$$

où $t_\alpha(\nu)$ est le réel vérifiant $\mathbb{P}(|T| \geq t_\alpha(\nu)) = \alpha$, $T \sim \mathcal{T}(\nu)$, $\nu = n-1$.

Intervalle de confiance (limite) pour τ_U :

Soit $\omega = (\omega_1, \ldots, \omega_n)$ un échantillon de n individus de U. Un intervalle de confiance pour τ_U au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, est

$$\begin{aligned} i_{\tau_U} &= [\tau_\omega - z_\alpha s(\tau_\omega),\ \tau_\omega + z_\alpha s(\tau_\omega)] \\ &= \left[\tau_\omega - z_\alpha\sqrt{N^2\frac{s_\omega^2}{n}},\ \tau_\omega + z_\alpha\sqrt{N^2\frac{s_\omega^2}{n}}\right] = N \times i_{\overline{y}_U}, \end{aligned}$$

où z_α est le réel vérifiant $\mathbb{P}(|Z| \geq z_\alpha) = \alpha$, $Z \sim \mathcal{N}(0,1)$.

On peut également définir l'incertitude absolue ou relative sur τ_U, ainsi que la taille d'échantillon souhaitée pour une incertitude donnée.

5.2 Estimation d'une proportion

Contexte : On suppose que le caractère Y est binaire : $Y(\Omega) = \{0, 1\}$. Cela correspond à un codage.

Proportion :

On appelle proportion-population la proportion des individus dans U vérifiant $Y = 1$:

$$p_U = \frac{1}{N}\sum_{i=1}^{N} y_i \qquad (= \overline{y}_U).$$

Estimation d'une proportion :

Un estimateur aléatoire de p_U est

$$p_W = \overline{y}_W = \frac{1}{n}\sum_{i=1}^{N} y_i \sum_{m=1}^{n} \mathbb{1}_{\{W_m = u_i\}}.$$

Espérance de p_W :

L'estimateur p_W est sans biais pour p_U :

$$\mathbb{E}(p_W) = p_U.$$

Variance de p_W :

La variance de p_W est

$$\mathbb{V}(p_W) = \frac{N-1}{N}\frac{s_U^2}{n} = \frac{p_U(1-p_U)}{n}.$$

Erreur quadratique moyenne de p_W :

L'erreur quadratique moyenne de p_W est le réel :

$$EQM(p_W)[PEAR] = \frac{p_U(1-p_U)}{n}.$$

Estimation ponctuelle de p_U :

Soit $\omega = (\omega_1, \ldots, \omega_n)$ un échantillon de n individus de U. Une estimation ponctuelle de p_U est la proportion-échantillon :

$$p_\omega = \overline{y}_\omega = \frac{1}{n}\sum_{i=1}^{N} y_i \sum_{m=1}^{n} \mathbb{1}_{\{\omega_m = u_i\}}.$$

Estimation ponctuelle de l'écart-type de p_W :

Soit $\omega = (\omega_1, \ldots, \omega_n)$ un échantillon de n individus de U. Une estimation ponctuelle de l'écart-type de p_W est le réel :

$$s(p_\omega) = \sqrt{\frac{p_\omega(1-p_\omega)}{n-1}}.$$

Intervalle de confiance pour p_U :

Soit $\omega = (\omega_1, \ldots, \omega_n)$ un échantillon de n individus de U. Un intervalle de confiance pour p_U au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, est

$$\begin{aligned} i_{p_U} &= [p_\omega - z_\alpha s(p_\omega),\ p_\omega + z_\alpha s(p_\omega)] \\ &= \left[p_\omega - z_\alpha \sqrt{\frac{p_\omega(1-p_\omega)}{n-1}},\ p_\omega + z_\alpha \sqrt{\frac{p_\omega(1-p_\omega)}{n-1}}\right], \end{aligned}$$

où z_α est le réel vérifiant $\mathbb{P}(|Z| \geq z_\alpha) = \alpha$, $Z \sim \mathcal{N}(0,1)$.

Quelques commandes R : Un exemple de fonction R pour calculer l'intervalle de confiance pour p_U au niveau $100(1-\alpha)\%$ est décrit ci-dessous :

```
icPEAR = function(y, niveau) {
n = length(y)
p_w = mean(y)
z = qnorm(1 - (1 - niveau) / 2)
var_p_w = p_w * (1 - p_w) / (n - 1)
a = p_w - z * sqrt(var_p_w)
b = p_w + z * sqrt(var_p_w)
print(c(a, b)) }
icPEAR(y = c(0, 1,0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0), niveau = 0.90)
```

Cela renvoie : 0.3017508, 0.7751723.

Incertitude absolue :

Soit $\omega = (\omega_1, \ldots, \omega_n)$ un échantillon de n individus de U. On appelle incertitude absolue sur p_U au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, la demi-longueur de i_{p_U} :

$$d_\omega = z_\alpha s(p_\omega) = z_\alpha \sqrt{\frac{p_\omega(1-p_\omega)}{n-1}}.$$

Plus d_ω est petit, plus l'estimation de p_U par p_ω est précise.

Incertitude relative :

Soit $\omega = (\omega_1, \ldots, \omega_n)$ un échantillon de n individus de U et d_ω l'incertitude absolue sur p_U au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$. On appelle incertitude relative sur p_U au niveau $100(1-\alpha)\%$ le pourcentage $(100 \times d^*_\omega)\%$ où d^*_ω est le réel :

$$d^*_\omega = \frac{d_\omega}{p_\omega}.$$

Taille d'échantillon :

Soit $\omega = (\omega_1, \ldots, \omega_n)$ un échantillon prélevé lors d'une étude préliminaire. La taille d'échantillon n à choisir pour avoir :

- une incertitude absolue sur p_U au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, inférieure ou égale à d_0 est le plus petit n tel que
$$d_\omega \leq d_0 \qquad \Rightarrow \qquad n \geq \frac{z^2_\alpha p_\omega (1-p_\omega)}{d_0^2},$$
- une incertitude relative sur p_U au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, inférieure ou égale à $(100 \times d_1)\%$ est le plus petit n tel que
$$d^*_\omega \leq d_1 \qquad \Rightarrow \qquad n \geq \frac{z^2_\alpha p_\omega (1-p_\omega)}{(p_\omega d_1)^2}.$$

On peut aussi remplacer $p_\omega(1-p_\omega)$ par $1/4$, ce qui évite une étude avec un échantillon préliminaire pour l'incertitude absolue.

Quelques commandes R : Un exemple de fonction R pour calculer la taille n d'un échantillon à partir de l'incertitude relative sur p_U au niveau $100(1-\alpha)\%$ est décrit ci-dessous :

```
n_ech = function(p_w, d1, niveau) {
z = qnorm(1 - (1 - niveau) / 2)
n = p_w * (1 - p_w) * z^2 / (d1 * p_w)^2
print(ceiling(n)) }
n_ech(p_w = 0.61, d1 = 0.5, niveau = 0.95)
```

Cela renvoie 10.

5.3 Estimation d'un effectif

Contexte : On suppose que le caractère Y est binaire : $Y(\Omega) = \{0, 1\}$. Cela correspond à un codage.

Effectif :

On appelle effectif-population le nombre des individus dans U vérifiant $Y = 1$:

$$\eta_U = Np_U.$$

Estimation aléatoire de η_U :

Un estimateur aléatoire de η_U est

$$\eta_W = Np_W = N\frac{1}{n}\sum_{i=1}^{N} y_i \sum_{m=1}^{n} \mathbb{1}_{\{W_m = u_i\}}.$$

Espérance de η_W :

L'estimateur η_W est sans biais pour η_U :

$$\mathbb{E}(\eta_W) = \eta_U.$$

Variance de η_W :

La variance de η_W est

$$\mathbb{V}(\eta_W) = N^2 \frac{p_U(1-p_U)}{n}.$$

Erreur quadratique moyenne de η_W :

L'erreur quadratique moyenne de τ_W est le réel :

$$EQM(\eta_W)[PEAR] = N^2 \frac{p_U(1-p_U)}{n}.$$

Estimation ponctuelle de η_U :

Soit $\omega = (\omega_1, \ldots, \omega_n)$ un échantillon de n individus de U. Une estimation ponctuelle de η_U est le total-échantillon :

$$\eta_\omega = Np_\omega = N\frac{1}{n}\sum_{i=1}^{N} y_i \sum_{m=1}^{n} \mathbb{1}_{\{\omega_m = u_i\}}.$$

Estimation ponctuelle de l'écart-type de η_W :

Soit $\omega = (\omega_1, \ldots, \omega_n)$ un échantillon de n individus de U. Une estimation ponctuelle de l'écart-type de η_W est le réel :

$$s(\eta_\omega) = \sqrt{N^2 \frac{p_\omega(1-p_\omega)}{n-1}}.$$

Intervalle de confiance pour η_U :

Soit $\omega = (\omega_1, \ldots, \omega_n)$ un échantillon de n individus de U. Un intervalle de confiance pour η_U au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, est

$$\begin{aligned} i_{\eta_U} &= [\eta_\omega - z_\alpha s(\eta_\omega),\ \eta_\omega + z_\alpha s(\eta_\omega)] \\ &= \left[\eta_\omega - z_\alpha \sqrt{N^2 \frac{p_\omega(1-p_\omega)}{n-1}},\ \eta_\omega + z_\alpha \sqrt{N^2 \frac{p_\omega(1-p_\omega)}{n-1}}\right] = N \times i_{p_U}, \end{aligned}$$

où z_α est le réel vérifiant $\mathbb{P}(|Z| \geq z_\alpha) = \alpha$, $Z \sim \mathcal{N}(0,1)$.

On peut également définir l'incertitude absolue ou relative sur η_U, ainsi que la taille d'échantillon souhaitée pour une incertitude donnée.

5.4 Exercices corrigés

Exercice 1 : Sur un campus universitaire, un jour donné, on s'intéresse au total des montants dépensés par les 1765 étudiants du campus pour le café. On note ce total τ_U. Sur un échantillon ω de 279 étudiants prélevé suivant un plan de sondage aléatoire de type PEAR, on obtient : $\overline{y}_\omega = 1.25$ € et $s_\omega = 0.25$ €.

1. Préciser le caractère étudié.
2. Donner une estimation ponctuelle de τ_U.
3. Déterminer un intervalle de confiance pour τ_U au niveau 95%.

Solution :

1. On étudie le caractère $Y =$ "dépense d'un étudiant du campus pour le café" en €.
2. Une estimation ponctuelle de τ_U est

$$\tau_\omega = N\overline{y}_\omega = 1765 \times 1.25 = 2206.25.$$

3. On a $95\% = 100(1-\alpha)\%$ avec $\alpha = 0.05$. A priori, on n'a pas l'hypothèse de normalité sur Y. On a $\mathbb{P}(|Z| \geq z_\alpha) = \alpha = 0.05$, $Z \sim \mathcal{N}(0,1)$, avec $z_\alpha = 1.96$.

Un intervalle de confiance pour τ_U au niveau 95% est

$$\begin{aligned} i_{\tau_U} &= \left[\tau_\omega - z_\alpha\sqrt{N^2\frac{s_\omega^2}{n}},\ \tau_\omega + z_\alpha\sqrt{N^2\frac{s_\omega^2}{n}}\right] \\ &= \left[2206.25 - 1.96\sqrt{1765^2\frac{0.25^2}{279}},\ 2206.25 + 1.96\sqrt{1765^2\frac{0.25^2}{279}}\right] \\ &= [2154.473,\ 2258.027]. \end{aligned}$$

Ainsi, il y a 95 chances sur 100 que $[2154.473,\ 2258.027]$ contienne τ_U, l'unité étant le €.

Exercice 2 : Sur un campus universitaire de 1765 étudiants, un échantillon de 250 étudiants est prélevé suivant un plan de sondage aléatoire de type PEAR. Parmi ces 250 étudiants, 144 admettent jouer aux jeux vidéos plus de 30 minutes par jour. On note p_U la proportion des 1765 étudiants qui admettent cela.

1. Donner une estimation ponctuelle de p_U.
2. Déterminer un intervalle de confiance pour p_U au niveau 95%.
3. Déterminer la taille d'échantillon à choisir pour avoir une incertitude relative sur p_U inférieure ou égale à 5% au niveau 95%.

Solution :

1. Une estimation ponctuelle de p_U est
$$p_\omega = \frac{144}{250} = 0.576.$$

2. On a $95\% = 100(1-\alpha)\%$ avec $\alpha = 0.05$.

   On a $\mathbb{P}(|Z| \geq z_\alpha) = \alpha = 0.05$, $Z \sim \mathcal{N}(0,1)$, avec $z_\alpha = 1.96$.

   Un intervalle de confiance pour p_U au niveau 95% est

$$\begin{aligned} i_{p_U} &= \left[p_\omega - z_\alpha\sqrt{\frac{p_\omega(1-p_\omega)}{n-1}},\ p_\omega + z_\alpha\sqrt{\frac{p_\omega(1-p_\omega)}{n-1}}\right] \\ &= \left[0.576 - 1.96\sqrt{\frac{0.576(1-0.576)}{250-1}},\ 0.576 + 1.96\sqrt{\frac{0.576(1-0.576)}{250-1}}\right] \\ &= [0.5146,\ 0.6373]. \end{aligned}$$

   Ainsi, il y a 95 chances sur 100 que $[0.5146,\ 0.6373]$ contienne p_U.

3. On a $95\% = 100(1-\alpha)\%$ avec $\alpha = 0.05$. On souhaite déterminer le plus petit n tel que :
$$d_\omega^* \leq d_1 \qquad \Rightarrow \qquad n \geq \frac{z_\alpha^2 p_\omega(1-p_\omega)}{(p_\omega d_1)^2},$$

avec $d_1 = 0.05$, $z_\alpha = 1.96$, ω est l'échantillon considéré précédemment et $p_\omega = 0.576$.

On a

$$n \geq \frac{1.96^2 \times 0.576(1 - 0.576)}{(0.576 \times 0.05)^2} = 1131.138.$$

Donc la taille d'échantillon à choisir pour avoir une incertitude relative sur p_U inférieure ou égale à 5% au niveau 95% est de $n = 1132$.

5.5 Synthèse : proportion

Paramètres-population et les paramètres-échantillon correspondants :

	Population U	Échantillon $\omega = (\omega_1, \ldots, \omega_n)$
Taille	N	n
Proportion	$p_U = \frac{1}{N}\sum_{i=1}^{N} y_i$	$p_\omega = \frac{1}{n}\sum_{i=1}^{N} y_i \sum_{m=1}^{n} \mathbb{1}_{\{\omega_m = u_i\}}$
Écart-type de p_W	$\sigma(p_W) = \sqrt{\frac{p_U(1-p_U)}{n}}$	$s(p_\omega) = \sqrt{\frac{p_\omega(1-p_\omega)}{n-1}}$

Autre notions utilisées autour de p_U (niveau : $100(1-\alpha)\%$, $\alpha \in]0,1[$) :

Intervalle de confiance	$i_{p_U} = \left[p_\omega - z_\alpha\sqrt{\frac{p_\omega(1-p_\omega)}{n-1}},\ p_\omega + z_\alpha\sqrt{\frac{p_\omega(1-p_\omega)}{n-1}}\right]$
Incertitude absolue	$d_\omega = z_\alpha\sqrt{\frac{p_\omega(1-p_\omega)}{n-1}}$
Incertitude relative	$d^*_\omega = \frac{d_\omega}{p_\omega}$
Taille n telle que $d_\omega \leq d_0$	$n \geq \frac{z^2_\alpha p_\omega(1-p_\omega)}{d_0^2}$
Taille n telle que $d^*_\omega \leq d_1$	$n \geq \frac{z^2_\alpha p_\omega(1-p_\omega)}{(p_\omega d_1)^2}$

Rappel : $\mathbb{P}(|Z| \geq z_\alpha) = \alpha$, $Z \sim \mathcal{N}(0,1)$.

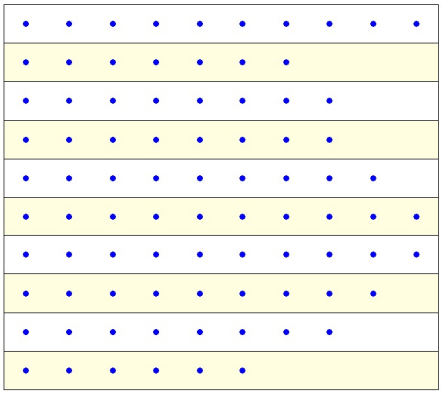
6 Plan de sondage aléatoire stratifié (ST)

6.1 Contexte

Idée : Les plans de sondages aléatoire de types PESR ou PEAR sont adaptés lorsque la population est homogène. Si la population n'est pas homogène mais qu'un découpage en plusieurs sous-populations homogènes est possible, un plan de sondage aléatoire pour chacune de ces sous-populations peut améliorer la précisions dans l'estimation des paramètres.

Strate :

On considère une partition de H éléments de U notée $(U_1, \ldots, U_H)$. Ainsi, on a $U = \bigcup_{h=1}^{H} U_h$ et, pour tout $(h, k) \in \{1, \ldots, H\}^2$ avec $h \neq k$, on a $U_h \cap U_k = \emptyset$.
On appelle strate un élément U_h de $(U_1, \ldots, U_H)$.



Plan de sondage aléatoire stratifié (ST) :

Un échantillon ω de n individus de $U = (U_1, \ldots, U_H)$ est prélevé suivant un plan de sondage aléatoire de type stratifié (ST) si on peut l'écrire sous la forme :

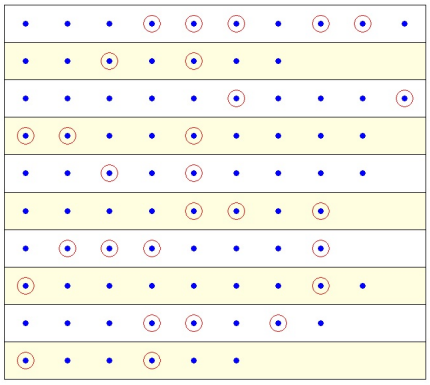
$$\omega = (\omega_1, \ldots, \omega_H),$$

où, pour tout $h \in \{1, \ldots, H\}$, ω_h est un échantillon de n_h individus de U_h prélevé suivant un plan de sondage aléatoire de type PESR.

Dans ce contexte, il y a

$$\binom{N_1}{n_1} \times \ldots \times \binom{N_H}{n_H}$$

échantillons possibles.



Quelques commandes R : Pour illustrer un plan de sondage aléatoire de type ST avec le logiciel R, on propose l'animation :

```
library(animation)
sample.strat(col = c("lightyellow", "white"))
```

Un autre exemple : On considère une population U partagée en 3 strates : U_1, U_2 et U_3. On fait un plan de sondage ST avec $n_1 = 3$, $n_2 = 2$ et $n_3 = 3$:

```
U_1 = c("Bob", "Nico", "Ali", "Fabien", "Malik", "John", "Jean", "Chris", "Karl")
U_2 = c("Jean", "Bill", "Omar", "Raul", "Mia")
U_3 = c("Paul", "Chael", "Nathan", "Sam", "Tom", "Tim", "Leo", "Kevin")
U = c(U_1, U_2, U_3)
n_h = c(3, 2, 3)
library(sampling)
t_1 = srswor(n_h[1], length(U_1))
w_1 = U_1[t_1 != 0]
t_2 = srswor(n_h[2], length(U_2))
w_2 = U_2[t_2 != 0]
t_3 = srswor(n_h[3], length(U_3))
w_3 = U_3[t_3 != 0]
c(w_1, w_2, w_3)
```

Le même exemple avec la commande `strata` de la librairie `sampling` :

```
U_1 = c("Bob", "Nico", "Ali", "Fabien", "Malik", "John", "Jean", "Chris", "Karl")
U_2 = c("Jean", "Bill", "Omar", "Raul", "Mia")
U_3 = c("Paul", "Chael", "Nathan", "Sam", "Tom", "Tim", "Leo", "Kevin")
dat = cbind.data.frame(c(U_1, U_2, U_3), c(rep(1, length(U_1)), rep(2,
length(U_2)), rep(3, length(U_3))))
names(dat) = c("noms", "souspop")
library(sampling)
s = strata(dat, "souspop", size = c(3, 2, 3), method = "srswor")
s
U = c(U_1, U_2, U_3)
U[s[ ,2]]
```

Remarque : Ce sont les plans de sondage aléatoire de type ST qui sont classiquement utilisés pour les enquêtes de l'INSEE auprès des entreprises.

Paramètres-population :

On adopte les notations suivantes :

◦ concernant la population U, rien ne change :

	Taille	Moyenne	Écart-type corrigé	Échantillon
U	N	$\overline{y}_U = \frac{1}{N}\sum_{i=1}^{N} y_i$	$s_U = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y}_U)^2}$	ω

◦ concernant la strate U_h :

	Taille	Moyenne	Écart-type corrigé	Échantillon
U_h	N_h	$\overline{y}_{U_h} = \frac{1}{N_h}\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in U_h\}}$	$s_{U_h} = \sqrt{\frac{1}{N_h-1}\sum_{i=1}^{N}(y_i - \overline{y}_{U_h})^2 \mathbb{1}_{\{u_i \in U_h\}}}$	ω_h

Paramètres-population avec les strates :

En utilisant la stratification $U = (U_1, \ldots, U_H)$, on a :

	Taille	Moyenne	Écart-type corrigé
U	$N = \sum_{h=1}^{H} N_h$	$\overline{y}_U = \frac{1}{N} \sum_{h=1}^{H} N_h \overline{y}_{U_h}$	$s_U = \sqrt{\frac{1}{N-1} \left(\sum_{h=1}^{H} (N_h - 1) s_{U_h}^2 + \sum_{h=1}^{H} N_h (\overline{y}_{U_h} - \overline{y}_U)^2 \right)}$

Preuve : On a

- comme $(U_1, \ldots, U_H)$ est une partition de U, on a $\sum_{h=1}^{H} N_h = N$,
- comme $(U_1, \ldots, U_H)$ est une partition de U, on a

$$\sum_{h=1}^{H} \mathbb{1}_{\{u_i \in U_h\}} = \mathbb{1}_{\{u_i \in \bigcup_{h=1}^{H} U_h\}} = \mathbb{1}_{\{u_i \in U\}} = 1.$$

Donc

$$\frac{1}{N} \sum_{h=1}^{H} N_h \overline{y}_{U_h} = \frac{1}{N} \sum_{h=1}^{H} N_h \frac{1}{N_h} \sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in U_h\}} = \frac{1}{N} \sum_{i=1}^{N} y_i \sum_{h=1}^{H} \mathbb{1}_{\{u_i \in U_h\}} = \frac{1}{N} \sum_{i=1}^{N} y_i = \overline{y}_U.$$

- En utilisant de nouveau $\sum_{h=1}^{H} \mathbb{1}_{\{u_i \in U_h\}} = 1$, on a

$$\begin{aligned}
\sum_{i=1}^{N} (y_i - \overline{y}_U)^2 &= \sum_{i=1}^{N} (y_i - \overline{y}_U)^2 \sum_{h=1}^{H} \mathbb{1}_{\{u_i \in U_h\}} = \sum_{h=1}^{H} \sum_{i=1}^{N} (y_i - \overline{y}_U)^2 \mathbb{1}_{\{u_i \in U_h\}} \\
&= \sum_{h=1}^{H} \sum_{i=1}^{N} \left((y_i - \overline{y}_{U_h}) + (\overline{y}_{U_h} - \overline{y}_U) \right)^2 \mathbb{1}_{\{u_i \in U_h\}} \\
&= \sum_{h=1}^{H} \sum_{i=1}^{N} (y_i - \overline{y}_{U_h})^2 \mathbb{1}_{\{u_i \in U_h\}} + 2 \sum_{h=1}^{H} \sum_{i=1}^{N} (y_i - \overline{y}_{U_h})(\overline{y}_{U_h} - \overline{y}_U) \mathbb{1}_{\{u_i \in U_h\}} \\
&+ \sum_{h=1}^{H} \sum_{i=1}^{N} (\overline{y}_{U_h} - \overline{y}_U)^2 \mathbb{1}_{\{u_i \in U_h\}}.
\end{aligned}$$

Étudions chacun des termes de cette somme. Pour le premier terme, on a

$$\sum_{h=1}^{H} \sum_{i=1}^{N} (y_i - \overline{y}_{U_h})^2 \mathbb{1}_{\{u_i \in U_h\}} = \sum_{h=1}^{H} (N_h - 1) s_{U_h}^2.$$

Pour le deuxième terme, en utilisant $\sum_{i=1}^{N} \mathbb{1}_{\{u_i \in U_h\}} = N_h$ et $\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in U_h\}} = N_h \overline{y}_{U_h}$, il vient

$$\begin{aligned}\sum_{h=1}^{H}\sum_{i=1}^{N}(y_i - \overline{y}_{U_h})(\overline{y}_{U_h} - \overline{y}_U)\mathbb{1}_{\{u_i \in U_h\}} &= \sum_{h=1}^{H}(\overline{y}_{U_h} - \overline{y}_U)\sum_{i=1}^{N}(y_i - \overline{y}_{U_h})\mathbb{1}_{\{u_i \in U_h\}} \\ &= \sum_{h=1}^{H}(\overline{y}_{U_h} - \overline{y}_U)\left(\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in U_h\}} - \overline{y}_{U_h}\sum_{i=1}^{N}\mathbb{1}_{\{u_i \in U_h\}}\right) \\ &= \sum_{h=1}^{H}(\overline{y}_{U_h} - \overline{y}_U)\left(\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in U_h\}} - N_h \overline{y}_{U_h}\right) = 0.\end{aligned}$$

Pour le troisième terme, en utilisant encore $\sum_{i=1}^{N} \mathbb{1}_{\{u_i \in U_h\}} = N_h$, on a

$$\begin{aligned}\sum_{h=1}^{H}\sum_{i=1}^{N}(\overline{y}_{U_h} - \overline{y}_U)^2 \mathbb{1}_{\{u_i \in U_h\}} &= \sum_{h=1}^{H}(\overline{y}_{U_h} - \overline{y}_U)^2 \sum_{i=1}^{N}\mathbb{1}_{\{u_i \in U_h\}} \\ &= \sum_{h=1}^{H} N_h (\overline{y}_{U_h} - \overline{y}_U)^2.\end{aligned}$$

Au final, on a

$$\sum_{i=1}^{N}(y_i - \overline{y}_U)^2 = \sum_{h=1}^{H}(N_h - 1)s_{U_h}^2 + \sum_{h=1}^{H} N_h (\overline{y}_{U_h} - \overline{y}_U)^2.$$

D'où

$$s_U = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y}_U)^2} = \sqrt{\frac{1}{N-1}\left(\sum_{h=1}^{H}(N_h - 1)s_{U_h}^2 + \sum_{h=1}^{H} N_h (\overline{y}_{U_h} - \overline{y}_U)^2\right)}.$$

□

Dispersion des valeurs de Y :

On pose

$$I = (N-1)s_U^2, \qquad I_{intra} = \sum_{h=1}^{H}(N_h - 1)s_{U_h}^2, \qquad I_{inter} = \sum_{h=1}^{H} N_h(\overline{y}_{U_h} - \overline{y}_U)^2, \qquad \eta^2 = \frac{I_{inter}}{I}.$$

Alors

- $I = I_{intra} + I_{inter}$,
- I_{intra} est un indicateur sur la dispersion des valeurs de Y au sein des strates,
- I_{inter} est un indicateur sur la dispersion des valeurs de Y entre les strates,
- la dispersion de Y entre les strates constitue $(100 \times \eta^2)\%$ de la dispersion des valeurs de Y dans U. Plus η^2 est proche de 1, plus la mise en œuvre d'un plan de sondage aléatoire de type ST est justifié.

Loi de probabilité :

Soit W_h la *var* égale à l'échantillon de taille n_h obtenu dans la strate U_h par un plan de sondage aléatoire de type PESR. Alors on a :

$$\mathbb{P}(W_h = \omega) = \frac{1}{\binom{N_h}{n_h}}, \qquad \omega \in W_h(\Omega).$$

Probabilités d'appartenance :

- pour tout $i \in \{1, \ldots, N\}$, on a

$$\mathbb{P}(u_i \in W_h) = \frac{n_h}{N_h}\mathbb{1}_{\{u_i \in U_h\}}.$$

- pour tout $(i, j) \in \{1, \ldots, N\}^2$ avec $i \neq j$ tels que u_i et u_j appartiennent à U_h, on a

$$\mathbb{P}((u_i, u_j) \in W_h) = \frac{n_h(n_h - 1)}{N_h(N_h - 1)}\mathbb{1}_{\{(u_i, u_j) \in U_h\}}.$$

Dans la suite :

- pour les résultats, on considère un plan de sondage aléatoire de type ST et la *var* $W = (W_1, \ldots, W_H)$ égale à l'échantillon obtenu,
- pour les commandes R, on utilisera dorénavant la librairie `sampling`.

6.2 Estimateurs

Estimation aléatoire de $\overline{y}_U$:

Un estimateur aléatoire de $\overline{y}_U$ est

$$\overline{y}_W = \frac{1}{N}\sum_{h=1}^{H} N_h \overline{y}_{W_h}, \qquad \overline{y}_{W_h} = \frac{1}{n_h}\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in W_h\}}.$$

Espérance de $\overline{y}_{W_h}$:

Pour tout $h \in \{1, \ldots, H\}$, on a

$$\mathbb{E}(\overline{y}_{W_h}) = \overline{y}_{U_h}.$$

Preuve : On a

$$\begin{aligned}\mathbb{E}(\overline{y}_{W_h}) &= \mathbb{E}\left(\frac{1}{n_h}\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in W_h\}}\right) = \frac{1}{n_h}\sum_{i=1}^{N} y_i \mathbb{E}\left(\mathbb{1}_{\{u_i \in W_h\}}\right) = \frac{1}{n_h}\sum_{i=1}^{N} y_i \mathbb{P}(u_i \in W_h) \\ &= \frac{1}{n_h}\sum_{i=1}^{N} y_i \frac{n_h}{N_h}\mathbb{1}_{\{u_i \in U_h\}} = \frac{1}{N_h}\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in U_h\}} = \overline{y}_{U_h}.\end{aligned}$$

□

Variance de $\overline{y}_{W_h}$:

Pour tout $h \in \{1, \ldots, H\}$, on a

$$\mathbb{V}(\overline{y}_{W_h}) = (1 - f_h)\frac{s^2_{U_h}}{n_h},$$

avec $f_h = n_h/N_h$.

Preuve : Par la formule de la variance d'une somme de *var*, on obtient

$$\begin{aligned}\mathbb{V}(\overline{y}_{W_h}) &= \mathbb{V}\left(\frac{1}{n_h}\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in W_h\}}\right) = \frac{1}{n_h^2}\mathbb{V}\left(\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in W_h\}}\right) \\ &= \frac{1}{n_h^2}\left(\sum_{i=1}^{N}\mathbb{V}\left(y_i \mathbb{1}_{\{u_i \in W_h\}}\right) + 2\sum_{i=2}^{N}\sum_{j=1}^{i-1}\mathbb{C}\left(y_i \mathbb{1}_{\{u_i \in W_h\}}, y_j \mathbb{1}_{\{u_j \in W_h\}}\right)\right) \\ &= \frac{1}{n_h^2}\left(\sum_{i=1}^{N} y_i^2 \mathbb{V}\left(\mathbb{1}_{\{u_i \in W_h\}}\right) + 2\sum_{i=2}^{N}\sum_{j=1}^{i-1} y_i y_j \mathbb{C}\left(\mathbb{1}_{\{u_i \in W_h\}}, \mathbb{1}_{\{u_j \in W_h\}}\right)\right).\end{aligned}$$

Or, en utilisant $\mathbb{P}(u_i \in W_h) = (n_h/N_h)\mathbb{1}_{\{u_i \in U_h\}}$, on a

$$\begin{aligned}\mathbb{V}\left(\mathbb{1}_{\{u_i \in W_h\}}\right) &= \mathbb{E}\left(\mathbb{1}^2_{\{u_i \in W_h\}}\right) - \left(\mathbb{E}\left(\mathbb{1}_{\{u_i \in W_h\}}\right)\right)^2 = \mathbb{P}(u_i \in W_h) - \left(\mathbb{P}(u_i \in W_h)\right)^2 \\ &= \frac{n_h}{N_h}\mathbb{1}_{\{u_i \in U_h\}} - \left(\frac{n_h}{N_h}\right)^2 \mathbb{1}_{\{u_i \in U_h\}} = \frac{n_h}{N_h}\left(1 - \frac{n_h}{N_h}\right)\mathbb{1}_{\{u_i \in U_h\}}.\end{aligned}$$

De plus, comme

$\mathbb{P}(\{u_i \in W_h\} \cap \{u_j \in W_h\}) = \mathbb{P}((u_i, u_j) \in W_h) = n_h(n_h - 1)/(N_h(N_h - 1))\mathbb{1}_{\{(u_i,u_j) \in U_h\}}$, il vient

$$\begin{aligned}\mathbb{C}\left(\mathbb{1}_{\{u_i \in W_h\}}, \mathbb{1}_{\{u_j \in W_h\}}\right) &= \mathbb{E}\left(\mathbb{1}_{\{u_i \in W_h\}}\mathbb{1}_{\{u_j \in W_h\}}\right) - \mathbb{E}\left(\mathbb{1}_{\{u_i \in W_h\}}\right)\mathbb{E}\left(\mathbb{1}_{\{u_j \in W_h\}}\right) \\ &= \mathbb{P}(\{u_i \in W_h\} \cap \{u_j \in W_h\}) - \mathbb{P}(u_i \in W_h)\mathbb{P}(u_j \in W_h) \\ &= \frac{n_h(n_h - 1)}{N_h(N_h - 1)}\mathbb{1}_{\{(u_i,u_j) \in U_h\}} - \frac{n_h}{N_h}\mathbb{1}_{\{u_i \in U_h\}}\frac{n_h}{N_h}\mathbb{1}_{\{u_j \in U_h\}} \\ &= \frac{n_h}{N_h}\left(\frac{n_h - 1}{N_h - 1} - \frac{n_h}{N_h}\right)\mathbb{1}_{\{u_i \in U_h\}}\mathbb{1}_{\{u_j \in U_h\}}.\end{aligned}$$

En combinant ces égalités, on obtient

$$\begin{aligned}&\mathbb{V}(\overline{y}_{W_h}) \\ &= \frac{1}{n_h^2}\left(\frac{n_h}{N_h}\left(1 - \frac{n_h}{N_h}\right)\sum_{i=1}^{N} y_i^2 \mathbb{1}_{\{u_i \in U_h\}} + 2\frac{n_h}{N_h}\left(\frac{n_h - 1}{N_h - 1} - \frac{n_h}{N_h}\right)\sum_{i=2}^{N}\sum_{j=1}^{i-1} y_i \mathbb{1}_{\{u_i \in U_h\}} y_j \mathbb{1}_{\{u_j \in U_h\}}\right) \\ &= \frac{1}{n_h N_h}\left(\left(1 - \frac{n_h}{N_h}\right)\sum_{i=1}^{N} y_i^2 \mathbb{1}_{\{u_i \in U_h\}} + \left(\frac{n_h - 1}{N_h - 1} - \frac{n_h}{N_h}\right)\left(2\sum_{i=2}^{N}\sum_{j=1}^{i-1} y_i \mathbb{1}_{\{u_i \in U_h\}} y_j \mathbb{1}_{\{u_j \in U_h\}}\right)\right).\end{aligned}$$

On a $2\sum_{i=2}^{N}\sum_{j=1}^{i-1} y_i \mathbb{1}_{\{u_i\in U_h\}} y_j \mathbb{1}_{\{u_j\in U_h\}} = \left(\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i\in U_h\}}\right)^2 - \sum_{i=1}^{N} y_i^2 \mathbb{1}_{\{u_i\in U_h\}}$. D'où

$$\begin{aligned}
\mathbb{V}(\overline{y}_{W_h}) &= \frac{1}{n_h N_h}\Bigg(\left(1-\frac{n_h}{N_h}\right)\sum_{i=1}^{N} y_i^2 \mathbb{1}_{\{u_i\in U_h\}} \\
&\quad + \left(\frac{n_h-1}{N_h-1}-\frac{n_h}{N_h}\right)\left(\left(\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i\in U_h\}}\right)^2 - \sum_{i=1}^{N} y_i^2 \mathbb{1}_{\{u_i\in U_h\}}\right)\Bigg) \\
&= \frac{1}{n_h N_h}\Bigg(\left(1-\frac{n_h}{N_h}-\frac{n_h-1}{N_h-1}+\frac{n_h}{N_h}\right)\sum_{i=1}^{N} y_i^2 \mathbb{1}_{\{u_i\in U_h\}} \\
&\quad + \left(\frac{n_h-1}{N_h-1}-\frac{n_h}{N_h}\right)\left(\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i\in U_h\}}\right)^2\Bigg) \\
&= \frac{1}{n_h N_h}\left(\frac{N_h-n_h}{N_h-1}\sum_{i=1}^{N} y_i^2 \mathbb{1}_{\{u_i\in U_h\}} - \frac{N_h-n_h}{N_h(N_h-1)}\left(\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i\in U_h\}}\right)^2\right) \\
&= \frac{N_h-n_h}{n_h N_h}\left(\frac{1}{N_h-1}\left(\sum_{i=1}^{N} y_i^2 \mathbb{1}_{\{u_i\in U_h\}} - N_h\left(\frac{1}{N_h}\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i\in U_h\}}\right)^2\right)\right).
\end{aligned}$$

D'autre part, on a

$$\begin{aligned}
s_{U_h}^2 &= \frac{1}{N_h-1}\sum_{i=1}^{N}(y_i-\overline{y}_{U_h})^2 \mathbb{1}_{\{u_i\in U_h\}} \\
&= \frac{1}{N_h-1}\left(\sum_{i=1}^{N} y_i^2 \mathbb{1}_{\{u_i\in U_h\}} - 2\overline{y}_{U_h}\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i\in U_h\}} + N_h\overline{y}_{U_h}^2\right) \\
&= \frac{1}{N_h-1}\left(\sum_{i=1}^{N} y_i^2 - 2N\overline{y}_{U_h}^2 + N_h\overline{y}_{U_h}^2\right) = \frac{1}{N_h-1}\left(\sum_{i=1}^{N} y_i^2 \mathbb{1}_{\{u_i\in U_h\}} - N\overline{y}_{U_h}^2\right) \\
&= \frac{1}{N_h-1}\left(\sum_{i=1}^{N} y_i^2 \mathbb{1}_{\{u_i\in U_h\}} - N_h\left(\frac{1}{N_h}\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i\in U_h\}}\right)^2\right).
\end{aligned}$$

Il s'ensuit

$$\mathbb{V}(\overline{y}_{W_h}) = \frac{N_h-n_h}{n_h N_h}s_{U_h}^2 = \left(1-\frac{n_h}{N_h}\right)\frac{s_{U_h}^2}{n_h} = (1-f_h)\frac{s_{U_h}^2}{n_h}.$$

□

Espérance de $\overline{y}_W$:

L'estimateur $\overline{y}_W$ est sans biais pour $\overline{y}_U$:

$$\mathbb{E}(\overline{y}_W) = \overline{y}_U.$$

Preuve : En utilisant $\mathbb{E}\left(\overline{y}_{W_h}\right) = \overline{y}_{U_h}$, il vient

$$\mathbb{E}(\overline{y}_W) = \mathbb{E}\left(\frac{1}{N}\sum_{h=1}^{H} N_h \overline{y}_{W_h}\right) = \frac{1}{N}\sum_{h=1}^{H} N_h \mathbb{E}\left(\overline{y}_{W_h}\right) = \frac{1}{N}\sum_{h=1}^{H} N_h \overline{y}_{U_h} = \overline{y}_U.$$

□

Variance de $\overline{y}_W$:

On a

$$\mathbb{V}(\overline{y}_W) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 (1-f_h)\frac{s_{U_h}^2}{n_h}.$$

Preuve : Comme $(U_1, \ldots, U_H)$ forme une partition de U, les *var* $\overline{y}_{W_1}, \ldots, \overline{y}_{W_H}$ sont indépendantes. Cela combiné à $\mathbb{V}\left(\overline{y}_{W_h}\right) = (1-f_h)s_{U_h}^2/n_h$ donne

$$\mathbb{V}(\overline{y}_W) = \mathbb{V}\left(\frac{1}{N}\sum_{h=1}^{H} N_h \overline{y}_{W_h}\right) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 \mathbb{V}\left(\overline{y}_{W_h}\right) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 (1-f_h)\frac{s_{U_h}^2}{n_h}.$$

□

Erreur quadratique moyenne de $\overline{y}_W$:

L'erreur quadratique moyenne de $\overline{y}_W$ est le réel :

$$EQM(\overline{y}_W)[ST] = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 (1-f_h)\frac{s_{U_h}^2}{n_h}.$$

Estimation aléatoire de s_{U_h} :

Un estimateur aléatoire de s_{U_h} est

$$s_{W_h} = \sqrt{\frac{1}{n_h - 1}\sum_{i=1}^{N} (y_i - \overline{y}_{W_h})^2 \mathbb{1}_{\{u_i \in W_h\}}}.$$

Propriété de $s^2_{W_h}$:

L'estimateur $s^2_{W_h}$ est sans biais pour $s^2_{U_h}$:

$$\mathbb{E}\left(s^2_{W_h}\right) = s^2_{U_h}.$$

Preuve : En remarquant que $\sum_{i=1}^{N} \mathbb{1}_{\{u_i \in W_h\}} = n_h$, il vient

$$\begin{aligned}
s^2_{W_h} &= \frac{1}{n_h - 1} \sum_{i=1}^{N} (y_i - \overline{y}_{W_h})^2 \mathbb{1}_{\{u_i \in W_h\}} \\
&= \frac{1}{n_h - 1} \left(\sum_{i=1}^{N} y_i^2 \mathbb{1}_{\{u_i \in W_h\}} - 2\overline{y}_{W_h} \sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in W_h\}} + \overline{y}^2_{W_h} \sum_{i=1}^{N} \mathbb{1}_{\{u_i \in W_h\}} \right) \\
&= \frac{1}{n_h - 1} \left(\sum_{i=1}^{N} y_i^2 \mathbb{1}_{\{u_i \in W_h\}} - 2n_h \overline{y}^2_{W_h} + n_h \overline{y}^2_{W_h} \right) = \frac{1}{n_h - 1} \left(\sum_{i=1}^{N} y_i^2 \mathbb{1}_{\{u_i \in W_h\}} - n_h \overline{y}^2_{W} \right).
\end{aligned}$$

En utilisant $\mathbb{P}(u_i \in W_h) = (n_h/N_h)\mathbb{1}_{\{u_i \in U_h\}}$ et

$$\mathbb{E}\left(\overline{y}^2_{W_h}\right) = \mathbb{V}\left(\overline{y}_{W_h}\right) + \left(\mathbb{E}\left(\overline{y}_{W_h}\right)\right)^2 = (1 - f_h)\frac{s^2_{U_h}}{n_h} + \overline{y}^2_{U_h},$$

on a

$$\begin{aligned}
\mathbb{E}\left(s^2_{W_h}\right) &= \mathbb{E}\left(\frac{1}{n_h - 1} \left(\sum_{i=1}^{N} y_i^2 \mathbb{1}_{\{u_i \in W_h\}} - n_h \overline{y}^2_{W_h} \right) \right) \\
&= \frac{1}{n_h - 1} \left(\sum_{i=1}^{N} y_i^2 \mathbb{E}\left(\mathbb{1}_{\{u_i \in W_h\}}\right) - n_h \mathbb{E}\left(\overline{y}^2_{W_h}\right) \right) \\
&= \frac{1}{n_h - 1} \left(\sum_{i=1}^{N} y_i^2 \mathbb{P}\left(u_i \in W_h\right) - n_h \mathbb{E}\left(\overline{y}^2_{W_h}\right) \right) \\
&= \frac{1}{n_h - 1} \left(\frac{n_h}{N_h} \sum_{i=1}^{N} y_i^2 \mathbb{1}_{\{u_i \in U_h\}} - n_h \left((1 - f_h)\frac{s^2_{U_h}}{n_h} + \overline{y}^2_{U_h} \right) \right) \\
&= \frac{1}{n_h - 1} \left(\frac{n_h}{N_h} \left(\sum_{i=1}^{N} y_i^2 \mathbb{1}_{\{u_i \in U_h\}} - N_h \overline{y}^2_{U_h} \right) - \left(1 - \frac{n_h}{N_h}\right) s^2_{U_h} \right) \\
&= \frac{n_h(N_h - 1)}{(n_h - 1)N_h} \left(\frac{1}{N_h - 1} \left(\sum_{i=1}^{N} y_i^2 - N_h \overline{y}^2_{U_h} \right) \right) - \frac{1}{n_h - 1} \left(1 - \frac{n_h}{N_h}\right) s^2_{U_h}.
\end{aligned}$$

En remarquant que

$$\begin{aligned} s^2_{U_h} &= \frac{1}{N_h-1}\sum_{i=1}^{N}(y_i-\overline{y}_{U_h})^2\mathbb{1}_{\{u_i\in U_h\}} \\ &= \frac{1}{N_h-1}\left(\sum_{i=1}^{N}y_i^2\mathbb{1}_{\{u_i\in U_h\}} - 2\overline{y}_{U_h}\sum_{i=1}^{N}y_i\mathbb{1}_{\{u_i\in U_h\}} + N_h\overline{y}^2_{U_h}\right) \\ &= \frac{1}{N_h-1}\left(\sum_{i=1}^{N}y_i^2\mathbb{1}_{\{u_i\in U_h\}} - 2N_h\overline{y}^2_{U_h} + N_h\overline{y}^2_{U_h}\right) = \frac{1}{N_h-1}\left(\sum_{i=1}^{N}y_i^2\mathbb{1}_{\{u_i\in U_h\}} - N_h\overline{y}^2_{U_h}\right). \end{aligned}$$

D'où

$$\begin{aligned} \mathbb{E}\left(s^2_{W_h}\right) &= \frac{n_h(N_h-1)}{(n_h-1)N_h}s^2_{U_h} - \frac{1}{n_h-1}\left(1-\frac{n_h}{N_h}\right)s^2_{U_h} \\ &= \frac{n_h(N_h-1)-N_h+n_h}{(n_h-1)N_h}s^2_{U_h} = \frac{n_hN_h-n_h-N_h+n_h}{(n_h-1)N_h}s^2_{U_h} = \frac{(n_h-1)N_h}{(n_h-1)N_h}s^2_{U_h} = s^2_{U_h}. \end{aligned}$$

□

6.3 Estimations ponctuelles

Estimation ponctuelle de $\overline{y}_{U_h}$:

Soit ω_h un échantillon de n_h individus de U_h. Une estimation ponctuelle de $\overline{y}_{U_h}$ est la moyenne-échantillon :

$$\overline{y}_{\omega_h} = \frac{1}{n_h}\sum_{i=1}^{N}y_i\mathbb{1}_{\{u_i\in\omega_h\}}.$$

Estimation ponctuelle de s_{U_h} :

Soit ω_h un échantillon de n_h individus de U_h. Une estimation ponctuelle de s_{U_h} est l'écart-type corrigé-échantillon :

$$s_{\omega_h} = \sqrt{\frac{1}{n_h-1}\sum_{i=1}^{N}(y_i-\overline{y}_{\omega_h})^2\mathbb{1}_{\{u_i\in\omega_h\}}}.$$

Estimation ponctuelle de $\overline{y}_U$:

Soit $\omega = (\omega_1, \ldots, \omega_H)$ un échantillon de $n = \sum_{h=1}^{H} n_h$ individus de U. Une estimation ponctuelle de $\overline{y}_U$ est la moyenne-échantillon (stratifiée) :

$$\overline{y}_\omega = \frac{1}{N} \sum_{h=1}^{H} N_h \overline{y}_{\omega_h}.$$

Quelques commandes R : Un exemple de calcul de $\overline{y}_\omega$ avec R est décrit ci-dessous :

```
Y_1 = c(35, 43, 36, 39, 28, 28, 29, 25, 38, 27, 26, 32, 29, 40, 35, 41, 38, 31,
45, 34, 15, 4, 41, 49, 25, 10)
Y_2 = c(27, 15, 4, 41, 49, 25, 10, 30, 32, 29, 40, 35, 41, 36, 31, 45)
Y_3 = c(8, 14, 12, 12, 15, 30, 32, 21, 20, 34, 7, 11, 24, 32, 29, 42, 35, 41, 37,
31, 42)
n_h = c(3, 2, 4)
library(sampling)
t_1 = srswor(n_h[1], length(Y_1))
t_2 = srswor(n_h[2], length(Y_2))
t_3 = srswor(n_h[3], length(Y_3))
bar_y_w_1 = (1 / n_h[1]) * sum(Y_1 * t_1)
bar_y_w_2 = (1 / n_h[2]) * sum(Y_2 * t_2)
bar_y_w_3 = (1 / n_h[3]) * sum(Y_3 * t_3)
bar_y_w_h = c(bar_y_w_1, bar_y_w_2, bar_y_w_3)
N_h = c(length(Y_1), length(Y_2), length(Y_3))
N = sum(N_h)
bar_y_w = sum(N_h * bar_y_w_h) / N
bar_y_w
```

Estimation ponctuelle de s_U :

Soit $\omega = (\omega_1, \ldots, \omega_H)$ un échantillon de $n = \sum_{h=1}^{H} n_h$ individus de U. Une estimation ponctuelle de s_U est l'écart-type corrigé-échantillon :

$$s_\omega = \sqrt{\frac{1}{N-1} \left(\sum_{h=1}^{H} (N_h - 1) s^2_{\omega_h} + \sum_{h=1}^{H} N_h (\overline{y}_{\omega_h} - \overline{y}_\omega)^2 \right)}.$$

Estimation ponctuelle de l'écart-type de $\overline{y}_W$:

Soit $\omega = (\omega_1, \ldots, \omega_H)$ un échantillon de $n = \sum_{h=1}^{H} n_h$ individus de U. Une estimation ponctuelle de l'écart-type de $\overline{y}_W$ est le réel :

$$s(\overline{y}_\omega) = \sqrt{\frac{1}{N^2} \sum_{h=1}^{H} N_h^2 (1 - f_h) \frac{s_{\omega_h}^2}{n_h}}.$$

Quelques commandes R : Un exemple de calcul de s_ω avec R est décrit ci-dessous :

```
Y_1 = c(35, 43, 36, 39, 28, 28, 29, 25, 38, 27, 26, 32, 29, 40, 35, 41, 38, 31,
45, 34, 15, 4, 41, 49, 25, 10)
Y_2 = c(27, 15, 4, 41, 49, 25, 10, 30, 32, 29, 40, 35, 41, 36, 31, 45)
Y_3 = c(8, 14, 12, 12, 15, 30, 32, 21, 20, 34, 7, 11, 24, 32, 29, 42, 35, 41, 37,
31, 42)
n_h = c(3, 2, 4)
t_1 = srswor(n_h[1], length(Y_1))
t_2 = srswor(n_h[2], length(Y_2))
t_3 = srswor(n_h[3], length(Y_3))
bar_y_w_1 = (1 / n_h[1]) * sum(Y_1 * t_1)
bar_y_w_2 = (1 / n_h[2]) * sum(Y_2 * t_2)
bar_y_w_3 = (1 / n_h[3]) * sum(Y_3 * t_3)
bar_y_w_h = c(bar_y_w_1, bar_y_w_2, bar_y_w_3)
s_w_1 = sqrt(sum((Y_1 - bar_y_w_1)^2 * t_1) / (n_h[1] - 1))
s_w_2 = sqrt(sum((Y_2 - bar_y_w_2)^2 * t_2) / (n_h[2] - 1))
s_w_3 = sqrt(sum((Y_3 - bar_y_w_3)^2 * t_3) / (n_h[3] - 1))
s_w_h = c(s_w_1, s_w_2, s_w_3)
N_h = c(length(Y_1), length(Y_2), length(Y_3))
N = sum(N_h)
bar_y_w = sum(N_h * bar_y_w_h) / N
s_bar_y_w = sqrt((1 / N^2) * sum(N_h^2 * (1-n_h / N_h) * (s_w_h^2 / n_h)))
s_bar_y_w
```

Question : Comment doit-on choisir les nombres d'individus $n_1, \ldots, n_H$ dans chaque strate pour que l'estimation de $\overline{y}_U$ soit la plus précise possible ? Deux réponses possibles sont apportées par :

- le plan de sondage STP,
- le plan de sondage STO.

6.4 Plan de sondage aléatoire stratifié proportionnel (STP)

Plan de sondage STP :

On appelle plan de sondage aléatoire stratifié proportionnel (STP) tout plan de sondage aléatoire stratifié (ST) tel que les entiers $n_1, \ldots, n_H$ vérifient, pour tout $h \in \{1, \ldots, H\}$, $f_h = f$, soit

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

Choix pratique :

En pratique, pour tout $h \in \{1, \ldots, H\}$, on prend le plus petit entier n_h tel que

$$n_h \geq \frac{n}{N} N_h.$$

Si on a $\sum_{h=1}^{H} n_h \neq n$, on ajuste en ajoutant ou enlevant une unités pour les échantillons les plus nombreux.

Réécriture de $\overline{y}_W$:

On a

$$\begin{aligned}\overline{y}_W &= \frac{1}{N} \sum_{h=1}^{H} N_h \overline{y}_{W_h} = \frac{1}{n} \sum_{h=1}^{H} n_h \overline{y}_{W_h} = \frac{1}{n} \sum_{i=1}^{N} y_i \sum_{h=1}^{H} \mathbb{1}_{\{u_i \in W_h\}} \\ &= \frac{1}{n} \sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in W\}}.\end{aligned}$$

On retrouve le même estimateur de la moyenne que celui présenté dans le cadre PESR.

Erreur quadratique moyenne de $\overline{y}_W$:

L'erreur quadratique moyenne de $\overline{y}_W$ est le réel :

$$EQM(\overline{y}_W)[STP] = (1-f) \frac{1}{n} \sum_{h=1}^{H} \frac{N_h}{N} s^2_{U_h}.$$

Comparaison de plans de sondage aléatoires de type PESR et STP :

Si N et N_h sont suffisamment grands, on peut montrer que

$$EQM(\overline{y}_W)[STP] \leq EQM(\overline{y}_W)[PESR].$$

Réécriture de $\overline{y}_\omega$:

Soit $\omega = (\omega_1, \ldots, \omega_H)$ un échantillon de $n = \sum_{h=1}^{H} n_h$ individus de U. Une estimation ponctuelle de $\overline{y}_U$ est la moyenne-échantillon :

$$\overline{y}_\omega = \frac{1}{N}\sum_{h=1}^{H} N_h \overline{y}_{\omega_h} = \frac{1}{n}\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in \omega\}}.$$

On retrouve la même estimation ponctuelle de $\overline{y}_U$ que celle présentée dans le cadre PESR.

Estimation ponctuelle de l'écart-type de $\overline{y}_W$:

Soit $\omega = (\omega_1, \ldots, \omega_H)$ un échantillon de $n = \sum_{h=1}^{H} n_h$ individus de U. Une estimation ponctuelle de l'écart-type de $\overline{y}_W$ est le réel :

$$s(\overline{y}_\omega) = \sqrt{\frac{1}{N^2}\sum_{h=1}^{H} N_h^2 (1-f_h)\frac{s^2_{\omega_h}}{n_h}} = \sqrt{(1-f)\frac{1}{n}\sum_{h=1}^{H}\frac{N_h}{N} s^2_{\omega_h}}.$$

6.5 Plan de sondage aléatoire stratifié optimal (STO)

Plan de sondage STO :

On appelle plan de sondage aléatoire stratifié optimal (STO) tout plan de sondage aléatoire stratifié (ST) tel que les entiers $n_1, \ldots, n_H$ minimisent

$$f(n_1, \ldots, n_H) = \frac{1}{N^2} \sum_{h=1}^{H} N_h^2 (1 - f_h) \frac{s_{U_h}^2}{n_h},$$

sous la contrainte $\sum_{h=1}^{H} n_h = n$.
Notons que $f(n_1, \ldots, n_H) = EQM(\overline{y}_W)[ST]$.

En utilisant une fonction lagrangienne, on obtient :

$$n_h = n \frac{N_h s_{U_h}}{\sum_{\ell=1}^{H} N_\ell s_{U_\ell}}.$$

Choix pratique : Soit $\omega = (\omega_1, \ldots, \omega_H)$ un échantillon prélevé lors d'une étude préliminaire. En pratique, pour tout $h \in \{1, \ldots, H\}$, on prend le plus petit entier n_h tel que

$$n_h \geq n \frac{N_h s_{\omega_h}}{\sum_{\ell=1}^{H} N_\ell s_{\omega_\ell}}.$$

Il dépend ainsi de la taille de strate U_h et de la dispersion des valeurs de Y dans la strate U_h.

- Si $n_h \geq N_h$, alors on prend $n_h = N_h$ et on recalcule les autres tailles sans prendre en compte l'échantillon ω_h :

$$n_k \geq (n - n_h) \frac{N_k s_{\omega_k}}{\sum_{\substack{\ell=1 \\ \ell \neq h}}^{H} N_\ell s_{\omega_\ell}}.$$

  On procède de même si $n_k \geq N_k$.
- Si $\sum_{h=1}^{H} n_h \neq n$, on ajuste en enlevant une unité pour les échantillons les plus nombreux.

Erreur quadratique moyenne de $\overline{y}_W$:

L'erreur quadratique moyenne de $\overline{y}_W$ est le réel :

$$EQM(\overline{y}_W)[STO] = \frac{1}{n}\left(\sum_{h=1}^{H}\frac{N_h}{N}s_{U_h}\right)^2 - \frac{1}{N}\sum_{h=1}^{H}\frac{N_h}{N}s^2_{U_h}.$$

Comparaison de plans de sondage aléatoires de types STP et STO :

On peut montrer que

$$EQM(\overline{y}_W)[STO] \leq EQM(\overline{y}_W)[STP].$$

Remarque : Si l'on dispose d'une information permettant la stratification de la population, on a tout intérêt à l'utiliser pour améliorer l'estimation de $\overline{y}_U$. Le plan de sondage aléatoire de type STO donne les meilleurs résultats.

6.6 Intervalles de confiance

Résultat limite : Si n, N et $N-n$ sont suffisamment grands, alors on a

$$Z = \frac{\overline{y}_W - \overline{y}_U}{\sqrt{\frac{1}{N^2}\sum_{h=1}^{H}N_h^2(1-f_h)\frac{s^2_{W_h}}{n_h}}} \approx \mathcal{N}(0,1).$$

Intervalle de confiance pour $\overline{y}_U$:

Soit ω un échantillon de n individus de U. Un intervalle de confiance pour $\overline{y}_U$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, est

$$\begin{aligned} i_{\overline{y}_U} &= [\overline{y}_\omega - z_\alpha s(\overline{y}_\omega),\ \overline{y}_\omega + z_\alpha s(\overline{y}_\omega)] \\ &= \left[\overline{y}_\omega - z_\alpha\sqrt{\frac{1}{N^2}\sum_{h=1}^{H}N_h^2(1-f_h)\frac{s^2_{\omega_h}}{n_h}},\ \overline{y}_\omega + z_\alpha\sqrt{\frac{1}{N^2}\sum_{h=1}^{H}N_h^2(1-f_h)\frac{s^2_{\omega_h}}{n_h}}\right], \end{aligned}$$

où z_α est le réel vérifiant $\mathbb{P}(|Z| \geq z_\alpha) = \alpha$, $Z \sim \mathcal{N}(0,1)$.

Il y a $100(1-\alpha)$ chances sur 100 que $\overline{y}_U$ appartienne à l'intervalle $i_{\overline{y}_U}$.

Quelques commandes R : Un exemple de calcul de $i_{\overline{y}_U}$ avec R est décrit ci-dessous :

```
icST= function(N_h, y, niveau) {
N = sum(N_h)
n_h = unlist(lapply(y, length))
bar_y_w_h = unlist(lapply(y, mean))
s_w_h = unlist(lapply(y, sd))
bar_y_w = sum(N_h * bar_y_w_h) / N
var_bar_y_w = (1 / N^2) * sum(N_h^2 * (1-n_h / N_h) * (s_w_h^2 / n_h))
z = qnorm(1 - (1 - niveau) / 2)
a = bar_y_w - z * sqrt(var_bar_y_w)
b = bar_y_w + z * sqrt(var_bar_y_w)
print(c(a, b)) }
N_h = c(155, 62, 93)
y_1 = c(35, 43, 36, 15, 30, 32, 21, 28, 29, 25, 38, 27, 26, 41, 49, 25, 10, 30,
31, 45, 34)
y_2 = c(27, 12, 12, 15, 49, 25, 10, 30)
y_3 = c(8, 14, 12, 12, 15, 30, 32, 21, 20, 34, 7, 11, 24)
y = list(y_1, y_2, y_3)
icST(N_h, y, 0.95)
```

6.7 Taille d'échantillon

Incertitude absolue :

Soit ω un échantillon de n individus de U. On appelle incertitude absolue sur $\overline{y}_U$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, la demi-longueur de $i_{\overline{y}_U}$:

$$d_\omega = z_\alpha s(\overline{y}_\omega) = z_\alpha \sqrt{\frac{1}{N^2}\sum_{h=1}^{H} N_h^2 (1-f_h)\frac{s_{\omega_h}^2}{n_h}}.$$

Plus d_ω est petit, plus l'estimation de $\overline{y}_U$ par $\overline{y}_\omega$ est précise.

Incertitude relative :

Soit ω un échantillon de n individus de U et d_ω l'incertitude absolue sur $\overline{y}_U$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$. On appelle incertitude relative sur $\overline{y}_U$ au niveau $100(1-\alpha)\%$ le pourcentage $(100 \times d^*_\omega)\%$ où d^*_ω est le réel :

$$d^*_\omega = \frac{d_\omega}{\overline{y}_\omega}.$$

Taille d'échantillon à partir de l'incertitude absolue :

Soit ω un échantillon prélevé lors d'une étude préliminaire. La taille d'échantillon n à choisir pour avoir une incertitude absolue sur $\overline{y}_U$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, inférieure ou égale à d_0 est le plus petit n tel que $d_\omega \leq d_0$. En particulier,

- pour un plan de sondage aléatoire de type STP :

$$n \geq \frac{N z_\alpha^2 \sum_{h=1}^{H} N_h s^2_{\omega_h}}{N^2 d_0^2 + z_\alpha^2 \sum_{h=1}^{H} N_h s^2_{\omega_h}},$$

- pour un plan de sondage aléatoire de type STO :

$$n \geq \frac{z_\alpha^2 \left(\sum_{h=1}^{H} N_h s_{\omega_h} \right)^2}{N^2 d_0^2 + z_\alpha^2 \sum_{h=1}^{H} N_h s^2_{\omega_h}}.$$

Quelques commandes R : Un exemple de fonction R pour calculer la taille n d'un échantillon à partir de l'incertitude absolue sur $\overline{y}_U$ pour un plan de sondage aléatoire de type STP au niveau $100(1-\alpha)\%$ est décrit ci-dessous :

```
n_ech = function(N_h, s_w_h, d0, niveau) {
N = sum(N_h)
z = qnorm(1 - (1 - niveau) / 2)
n = (N * z^2 * sum(N_h * s_w_h^2 )) / (N^2 * d0^2 + z^2 * sum(N_h * s_w_h^2))
print(ceiling(n)) }
N_h = c(15, 12, 134)
s_w_h = c(0.225, 1.271, 0.124)
n_ech(N_h, s_w_h, d0 = 0.1, niveau = 0.95)
```

Cela renvoie 40.

Taille d'échantillon à partir de l'incertitude relative :

Soit ω un échantillon prélevé lors d'une étude préléminaire. La taille d'échantillon n à choisir pour avoir une incertitude relative sur $\overline{y}_U$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, inférieure ou égale à $(100\times d_1)\%$ est le plus petit n tel que $d^*_\omega \leq d_1$. En particulier,

- pour un plan de sondage aléatoire de type STP :

$$n \geq \frac{N z_\alpha^2 \sum_{h=1}^{H} N_h s^2_{\omega_h}}{N^2 (d_1 \overline{y}_\omega)^2 + z_\alpha^2 \sum_{h=1}^{H} N_h s^2_{\omega_h}},$$

- pour un plan de sondage aléatoire de type STO :

$$n \geq \frac{z_\alpha^2 \left(\sum_{h=1}^{H} N_h s_{\omega_h} \right)^2}{N^2 (d_1 \overline{y}_\omega)^2 + z_\alpha^2 \sum_{h=1}^{H} N_h s^2_{\omega_h}}.$$

6.8 Exercices corrigés

Exercice 1 : On considère le caractère Y = "âge" en années dans la population de 5 individus : $U = \{$Paul, John, Charles, Alexandre, Dimitri$\} = \{u_1, \ldots, u_5\}$. Pour tout $i \in \{1, \ldots, 5\}$, soit y_i la valeur de Y pour l'individu u_i. Les résultats, en années, sont :

y_1	y_2	y_3	y_4	y_5
17	14.5	26	22.5	23

1. Calculer la moyenne-population $\overline{y}_U$ et l'écart-type corrigé-population s_U.
2. Dans un premier temps, on prélève un échantillon de 2 individus suivant un plan de sondage aléatoire de type PESR.
   (a) Quel est le taux de sondage ? Combien d'échantillons peut-on former ? Expliciter les.
   (b) Pour chaque échantillon ω, calculer la moyenne-échantillon $\overline{y}_\omega$.
   (c) Soit $\overline{y}_W$ la *var* égale à la moyenne-échantillon, l'aléatoire étant dans l'échantillon considéré. Déterminer sa loi, puis calculer son espérance, sa variance et son erreur quadratique moyenne : EQM$(\overline{y}_W)$.

3. Dans un deuxième temps, on prélève un échantillon de 2 individus suivant un plan de sondage aléatoire de type ST avec :
   - les 2 strates : $U_1 = \{\text{Paul, John}\}$ et $U_2 = \{\text{Charles, Alexandre, Dimitri}\}$,
   - un individu par strate.

   (a) Combien d'échantillons peut-on former ? Expliciter les.

   (b) Pour chaque échantillon ω, calculer la moyenne-échantillon $\overline{y}_\omega$.

   (c) Soit $\overline{y}_W$ la *var* égale à la moyenne-échantillon stratifié, l'aléatoire étant dans l'échantillon considéré. Déterminer sa loi, puis calculer son espérance, sa variance et son erreur quadratique moyenne : $\text{EQM}(\overline{y}_W)$.

4. Quel plan de sondage donne une meilleure précision dans l'estimation de $\overline{y}_U$?

Solution :

1. On a
$$\overline{y}_U = 20.6, \qquad s_U = 4.7090.$$

2. (a) Le taux de sondage est
$$f = \frac{n}{N} = \frac{2}{5} = 0.4.$$
Vu le mode de prélèvement, le nombre d'échantillons possibles est
$$\binom{5}{2} = \frac{5!}{2!(5-2)!} = 10.$$
Ils sont :

$\{u_1, u_2\}$	$\{u_1, u_3\}$	$\{u_1, u_4\}$	$\{u_1, u_5\}$	$\{u_2, u_3\}$
$\{u_2, u_4\}$	$\{u_2, u_5\}$	$\{u_3, u_4\}$	$\{u_3, u_5\}$	$\{u_4, u_5\}$

   (b) On a :

ω	Y	$\overline{y}_\omega$
$\{u_1, u_2\}$	$\{17, 14.5\}$	15.75
$\{u_1, u_3\}$	$\{17, 26\}$	21.5
$\{u_1, u_4\}$	$\{17, 22.5\}$	19.75
$\{u_1, u_5\}$	$\{17, 23\}$	20
$\{u_2, u_3\}$	$\{14.5, 26\}$	20.25
$\{u_2, u_4\}$	$\{14.5, 22.5\}$	18.5
$\{u_2, u_5\}$	$\{14.5, 23\}$	18.75
$\{u_3, u_4\}$	$\{26, 22.5\}$	24.25
$\{u_3, u_5\}$	$\{26, 23\}$	24.5
$\{u_4, u_5\}$	$\{22.5, 23\}$	22.75

(c) Soit $\overline{y}_W$ la *var* égale à la moyenne-échantillon. L'ensemble des valeurs possibles pour $\overline{y}_W$ est

$$\overline{y}_W(\Omega) = \{15.75, 18.5, 18.75, 19.75, 20, 20.25, 21.5, 22.75, 24.25, 24.5\}.$$

Comme il y a 10 échantillons différents et qu'ils sont équiprobables, la loi de $\overline{y}_W$ est donnée par

k	15.75	18.5	18.75	19.75	20	20.25	21.5	22.75	24.25	24.5
$\mathbb{P}(\overline{y}_W = k)$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

En utilisant la loi de $\overline{y}_W$, l'espérance de $\overline{y}_W$ est

$$\begin{aligned}
\mathbb{E}(\overline{y}_W) &= \sum_{k \in \overline{y}_W(\Omega)} k\mathbb{P}(\overline{y}_W = k) \\
&= \frac{1}{10}(15.75 + 18.5 + 18.75 + 19.75 + 20 + 20.25 + 21.5 + 22.75 + 24.25 + 24.5) \\
&= 20.6 \qquad (= \overline{y}_U)
\end{aligned}$$

En utilisant la formule de König-Huyghens, la variance de $\overline{y}_W$ est

$$\mathbb{V}(\overline{y}_W) = \mathbb{E}(\overline{y}_W^2) - \left(\mathbb{E}(\overline{y}_W)\right)^2.$$

Or on a $\mathbb{E}(\overline{y}_W) = 20.6$ et

$$\begin{aligned}\mathbb{E}(\overline{y}_W^2) &= \sum_{k\in\overline{y}_W(\Omega)} k^2\mathbb{P}(\overline{y}_W = k)\\ &= \frac{1}{10}(15.75^2 + 18.5^2 + 18.75^2 + 19.75^2 + 20^2 + 20.25^2\\ &+ 21.5^2 + 22.75^2 + 24.25^2 + 24.5^2)\\ &= 431.0125.\end{aligned}$$

D'où

$$\mathbb{V}(\overline{y}_W) = 431.0125 - 20.6^2 = 6.652 \qquad \left(= (1-f)\frac{s_U^2}{n}\right)$$

et

$$EQM(\overline{y}_W) = \mathbb{V}(\overline{y}_W) = 6.652.$$

3. (a) Vu le mode de prélèvement, le nombre d'échantillons possibles est

$$\binom{2}{1}\binom{3}{1} = 2 \times 3 = 6.$$

Ils sont :

$\{u_1, u_3\}$	$\{u_1, u_4\}$	$\{u_1, u_5\}$	$\{u_2, u_3\}$	$\{u_2, u_4\}$	$\{u_2, u_5\}$

(b) Dans le cadre d'un plan de sondage aléatoire de type ST, on rappelle que

$$\overline{y}_\omega = \frac{1}{N}\sum_{h=1}^{H} N_h \overline{y}_{\omega_h}.$$

Ici, $N = 5$, $H = 2$, $N_1 = 2$, $N_2 = 3$, $\overline{y}_{\omega_1}$ est la valeur de Y pour l'individu prélevé dans la Strate U_1 et $\overline{y}_{\omega_2}$ est la valeur de Y pour l'individu prélevé dans la Strate U_2.

Par exemple, avec $\omega = \{u_1, u_3\}$, on a

$$\overline{y}_\omega = \frac{2}{5}17 + \frac{3}{5}26 = 22.4.$$

On a

ω	Y	$\overline{y}_\omega$
$\{u_1, u_3\}$	$\{17, 26\}$	22.4
$\{u_1, u_4\}$	$\{17, 22.5\}$	20.3
$\{u_1, u_5\}$	$\{17, 23\}$	20.6
$\{u_2, u_3\}$	$\{14.5, 26\}$	21.4
$\{u_2, u_4\}$	$\{14.5, 22.5\}$	19.3
$\{u_2, u_5\}$	$\{14.5, 23\}$	19.6

(c) Soit $\overline{y}_W$ la *var* égale à la moyenne-échantillon dans le cadre ST. L'ensemble des valeurs possibles pour $\overline{y}_W$ est

$$\overline{y}_W(\Omega) = \{19.3, 19.6, 20.3, 20.6, 21.4, 22.4\}.$$

Comme il y a 6 échantillons différents et qu'ils sont équiprobables, la loi de $\overline{y}_W$ est donnée par

k	19.3	19.6	20.3	20.6	21.4	22.4
$\mathbb{P}(\overline{y}_W = k)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

En utilisant la loi de $\overline{y}_W$, l'espérance de $\overline{y}_W$ est

$$\begin{aligned}
\mathbb{E}(\overline{y}_W) &= \sum_{k \in \overline{y}_W(\Omega)} k\mathbb{P}(\overline{y}_W = k) \\
&= \frac{1}{6}(22.4 + 20.3 + 20.6 + 21.4 + 19.3 + 19.6) \\
&= 20.6 \qquad (= \overline{y}_U)
\end{aligned}$$

En utilisant la formule de König-Huyghens, la variance de $\overline{y}_W$ est

$$\mathbb{V}(\overline{y}_W) = \mathbb{E}(\overline{y}_W^2) - \left(\mathbb{E}(\overline{y}_W)\right)^2.$$

Or on a $\mathbb{E}(\overline{y}_W) = 20.6$ et

$$\begin{aligned}
\mathbb{E}(\overline{y}_W^2) &= \sum_{k \in \overline{y}_W(\Omega)} k^2\mathbb{P}(\overline{y}_W = k) \\
&= \frac{1}{6}(22.4^2 + 20.3^2 + 20.6^2 + 21.4^2 + 19.3^2 + 19.6^2) \\
&= 425.47.
\end{aligned}$$

D'où

$$\mathbb{V}(\overline{y}_W) = 425.47 - 20.6^2 = 1.11 \qquad \left(= \frac{1}{N^2} \sum_{h=1}^{H} N_h^2 (1 - f_h) \frac{s_{U_h}^2}{n_h} \right)$$

et

$$EQM(\overline{y}_W)[ST] = \mathbb{V}(\overline{y}_W) = 1.11.$$

Remarque : On a bien

$$\frac{1}{N^2} \sum_{h=1}^{H} N_h^2 (1 - f_h) \frac{s_{U_h}^2}{n_h} = \frac{1}{5^2} \left(2^2 \left(1 - \frac{1}{2} \right) \frac{1.7677^2}{1} + 3^2 \left(1 - \frac{1}{3} \right) \frac{1.8929^2}{1} \right) = 1.10991.$$

4. Par les résultats des questions 2 (c) et 3 (c), on a

$$EQM(\overline{y}_W)[ST] = 1.11 \leq 6.652 = EQM(\overline{y}_W)[PESR].$$

Donc le plan de sondage aléatoire de type ST donne une meilleure précision dans l'estimation de $\overline{y}_U$ que le plan de sondage aléatoire de type PESR.

Exercice 2 : Une population U est partagée en 3 strates U_1, U_2 et U_3 de tailles respectives : $N_1 = 12$, $N_2 = 28$ et $N_3 = 50$. On prélève un échantillon de $n = 20$ individus suivant un plan de sondage aléatoire de type ST avec :

- $n_1 = 2$ individus pour U_1,
- $n_2 = 6$ individus pour U_2,
- $n_3 = 12$ individus pour U_3.

On mesure un caractère quantitatif Y sur chacun d'entre eux. Les résultats obtenus sont :

Pour U_1	1450	1598				
Pour U_2	718	626	922	823	901	823
Pour U_3	201	268	225	231	453	387
	401	368	325	331	253	197

1. Donner une estimation ponctuelle de la moyenne-population $\overline{y}_U$.
2. Donner une estimation ponctuelle de l'écart-type de l'estimateur de $\overline{y}_U$.
3. Déterminer un intervalle de confiance pour $\overline{y}_U$ au niveau 95%.

Solution :

1. Dans le cadre d'un plan de sondage aléatoire de type ST, une estimation ponctuelle de la moyenne-population $\overline{y}_U$ est
$$\overline{y}_\omega = \frac{1}{N}\sum_{h=1}^{H} N_h \overline{y}_{\omega_h}.$$
Ici, $H = 3$, $N_1 = 12$, $N_2 = 28$, $N_3 = 50$, $N = \sum_{h=1}^{H} N_h = 90$,
$$\overline{y}_{\omega_1} = 1524, \qquad \overline{y}_{\omega_2} = 802.1667, \qquad \overline{y}_{\omega_3} = 303.3333.$$
Ainsi, une estimation ponctuelle de la moyenne-population $\overline{y}_U$ est
$$\overline{y}_\omega = \frac{1}{90}(12 \times 1524 + 28 \times 802.1667 + 50 \times 303.3333) = 621.2815.$$

2. Dans le cadre d'un plan de sondage aléatoire de type ST, une estimation ponctuelle de l'écart-type de $\overline{y}_W$ est
$$s(\overline{y}_\omega) = \sqrt{\frac{1}{N^2}\sum_{h=1}^{H} N_h^2 (1-f_h)\frac{s_{\omega_h}^2}{n_h}}.$$
Ici on a
$$s_{\omega_1} = 104.6518, \qquad s_{\omega_2} = 112.352, \qquad s_{\omega_3} = 85.9622$$
et
$$f_1 = \frac{n_1}{N_1} = \frac{2}{12}, \qquad f_2 = \frac{n_2}{N_2} = \frac{6}{28}, \qquad f_3 = \frac{n_3}{N_3} = \frac{12}{50}.$$
Donc
$$\begin{aligned} s^2(\overline{y}_\omega) = {} & \frac{1}{90^2}\left(12^2\left(1-\frac{2}{12}\right)\frac{104.6518^2}{2}\right) + \frac{1}{90^2}\left(28^2\left(1-\frac{6}{28}\right)\frac{112.352^2}{6}\right) \\ & + \frac{1}{90^2}\left(50^2\left(1-\frac{12}{50}\right)\frac{85.96229^2}{12}\right) = 385.566. \end{aligned}$$
Il vient
$$s(\overline{y}_\omega) = \sqrt{385.566} = 19.63583.$$

3. On a $95\% = 100(1-\alpha)\%$ avec $\alpha = 0.05$. On a $\mathbb{P}(|Z| \geq z_\alpha) = \alpha = 0.05$, $Z \sim \mathcal{N}(0,1)$, avec $z_\alpha = 1.96$.

En utilisant les résultats des questions 1 et 2, un intervalle de confiance pour $\overline{y}_U$ au niveau 95% est

$$i_{\overline{y}_U} = \left[\overline{y}_\omega - z_\alpha\sqrt{\frac{1}{N^2}\sum_{h=1}^{H} N_h^2(1-f_h)\frac{s^2_{\omega_h}}{n_h}},\ \overline{y}_\omega + z_\alpha\sqrt{\frac{1}{N^2}\sum_{h=1}^{H} N_h^2(1-f_h)\frac{s^2_{\omega_h}}{n_h}}\right]$$
$$= [621.2815 - 1.96 \times 19.63583,\ 621.2815 + 1.96 \times 19.63583] = [582.7953,\ 659.7677].$$

Ainsi, il y a 95 chances sur 100 que $[582.7953,\ 659.7677]$ contienne $\overline{y}_U$.

Exercice 3 : Une population U est partagée en 4 strates U_1, U_2, U_3 et U_4. On prélève un échantillon de 77 individus suivant un plan de sondage aléatoire de type ST et on mesure un caractère quantitatif Y sur chacun d'entre eux. On dispose des informations suivantes :

Strate U_h	U_1	U_2	U_3	U_4
Taille N_h	310	220	130	110
Écart-type corrigé s_{U_h}	9.5	6.1	3.5	2.1

1. Quelle est l'effectif total de la population ?
2. On considère un plan de sondage aléatoire de type STP.
   (a) Déterminer les tailles des échantillons pour chacune des strates.
   (b) Calculer l'erreur quadratique moyenne de l'estimateur de la moyenne-population.
3. On considère maintenant un plan de sondage aléatoire de type STO.
   (a) Déterminer les tailles des échantillons pour chacune des strates.
   (b) Calculer l'erreur quadratique moyenne de l'estimateur de la moyenne-population.
4. Comparer les résultats des 2 plans de sondage considérés.

Solution :

1. On a $H = 4$. L'effectif total de la population est

$$N = \sum_{h=1}^{H} N_h = 770.$$

2. On considère un plan de sondage aléatoire de type STP.
   (a) Par la définition du type STP, on prend les plus petits entiers n_1, n_2, n_3 et n_4 tels que :

$$n_1 \geq \frac{n}{N}N_1 = 0.1 \times 310 = 31, \qquad n_2 \geq \frac{n}{N}N_2 = 0.1 \times 220 = 22,$$

$$n_3 \geq \frac{n}{N} N_1 = 0.1 \times 130 = 13, \qquad n_4 \geq \frac{n}{N} N_2 = 0.1 \times 110 = 11.$$

D'où :

$$n_1 = 31. \qquad n_2 = 22, \qquad n_3 = 13, \qquad n_4 = 11.$$

On a $\sum_{h=1}^{H} n_h = 77 = n$, il n'y a pas d'ajustement à faire.

(b) L'erreur quadratique moyenne de l'estimateur de la moyenne-population $\overline{y}_W$ est

$$\begin{aligned} EQM(\overline{y}_W)[STP] &= (1-f)\frac{1}{n}\sum_{h=1}^{H}\frac{N_h}{N}s_{U_h}^2 \\ &= \left(1 - \frac{77}{770}\right)\frac{1}{77}\left(\frac{310}{770}9.5^2 + \frac{220}{770}6.1^2 + \frac{130}{770}3.5^2 + \frac{110}{770}2.1^2\right) \\ &= 0.5804. \end{aligned}$$

3. On considère maintenant un plan de sondage aléatoire de type STO.

(a) Par la définition du type STO, on prend les plus petits entiers n_1, n_2, n_3 et n_4 tels que :

$$n_1 \geq n\frac{N_1 s_{U_1}}{\sum_{\ell=1}^{H} N_\ell s_{U_\ell}} = 77 \times \frac{310 \times 9.5}{310 \times 9.5 + 220 \times 6.1 + 130 \times 3.5 + 110 \times 2.1} = 45.5992,$$

$$n_2 \geq n\frac{N_2 s_{U_2}}{\sum_{\ell=1}^{H} N_\ell s_{U_\ell}} = 77 \times \frac{220 \times 6.1}{310 \times 9.5 + 220 \times 6.1 + 130 \times 3.5 + 110 \times 2.1} = 20.7790,$$

$$n_3 \geq n\frac{N_3 s_{U_3}}{\sum_{\ell=1}^{H} N_\ell s_{U_\ell}} = 77 \times \frac{130 \times 3.5}{310 \times 9.5 + 220 \times 6.1 + 130 \times 3.5 + 110 \times 2.1} = 7.0450,$$

et

$$n_4 \geq n\frac{N_4 s_{U_4}}{\sum_{\ell=1}^{H} N_\ell s_{U_\ell}} = 77 \times \frac{110 \times 2.1}{310 \times 9.5 + 220 \times 6.1 + 130 \times 3.5 + 110 \times 2.1} = 3.5767.$$

D'où :

$$n_1 = 46, \qquad n_2 = 21, \qquad n_3 = 8, \qquad n_4 = 4.$$

Comme $\sum_{h=1}^{H} n_h = 79 \neq 77$, on propose l'ajustement :

$$n_1 = 45, \qquad n_2 = 20, \qquad n_3 = 8, \qquad n_4 = 4.$$

(b) L'erreur quadratique moyenne de l'estimateur de la moyenne-population $\overline{y}_W$ est

$$\begin{aligned} EQM(\overline{y}_W)[STO] &= \frac{1}{n}\left(\sum_{h=1}^{H}\frac{N_h}{N}s_{U_h}\right)^2 - \frac{1}{N}\sum_{h=1}^{H}\frac{N_h}{N}s_{U_h}^2 \\ &= \frac{1}{77}\left(\frac{310}{770}9.5 + \frac{220}{770}6.1 + \frac{130}{770}3.5 + \frac{110}{770}2.1\right)^2 \\ &\quad - \frac{1}{770}\left(\frac{310}{770}9.5^2 + \frac{220}{770}6.1^2 + \frac{130}{770}3.5^2 + \frac{110}{770}2.1^2\right) \\ &= 0.4772. \end{aligned}$$

4. On remarque que les plans de sondage amènent à des tailles différentes pour le choix des échantillons. De plus, par rapport au type STP, le sondage aléatoire de type STO conduit à une meilleure performance de $\overline{y}_W$ dans l'estimation de $\overline{y}_U$.

6.9 Synthèse

Paramètres-strates et les paramètres-échantillon correspondants, $\omega = (\omega_1, \ldots, \omega_H)$:

	Strate U_h	Échantillon ω_h
Taille	N_h	n_h
Taux de sondage	$\square$	$f_h = \dfrac{n_h}{N_h}$
Moyenne	$\overline{y}_{U_h} = \dfrac{1}{N_h}\displaystyle\sum_{i=1}^{N_h} y_i$	$\overline{y}_{\omega_h} = \dfrac{1}{n_h}\displaystyle\sum_{i=1}^{N_h} y_i \mathbb{1}_{\{u_i \in \omega_h\}}$
Écart-type corrigé	$s_{U_h} = \sqrt{\dfrac{1}{N_h - 1}\displaystyle\sum_{i=1}^{N_h}(y_i - \overline{y}_{U_h})^2}$	$s_{\omega_h} = \sqrt{\dfrac{1}{n_h - 1}\displaystyle\sum_{i=1}^{N_h}(y_i - \overline{y}_{\omega_h})^2 \mathbb{1}_{\{u_i \in \omega_h\}}}$
Écart-type de $\overline{y}_{W_h}$	$\sigma(\overline{y}_{W_h}) = \sqrt{(1 - f_h)\dfrac{s^2_{U_h}}{n_h}}$	$s(\overline{y}_{\omega_h}) = \sqrt{(1 - f_h)\dfrac{s^2_{\omega_h}}{n_h}}$

Paramètres-population et les paramètres-échantillon correspondants :

	Population U	Échantillon ω
Taille	N	n
Moyenne	$\overline{y}_U = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} y_i$	$\overline{y}_\omega = \dfrac{1}{N}\displaystyle\sum_{h=1}^{H} N_h \overline{y}_{\omega_h}$
Écart-type de $\overline{y}_W$	$\sigma(\overline{y}_W) = \sqrt{\dfrac{1}{N^2}\displaystyle\sum_{h=1}^{H} N_h^2 (1 - f_h)\dfrac{s^2_{U_h}}{n_h}}$	$s(\overline{y}_\omega) = \sqrt{\dfrac{1}{N^2}\displaystyle\sum_{h=1}^{H} N_h^2 (1 - f_h)\dfrac{s^2_{\omega_h}}{n_h}}$

Plans de sondage aléatoires de types STP et STO :

	STP	STO	STO (applicable)
n_h	$\frac{n}{N}N_h$	$n\frac{N_h s_{U_h}}{\sum_{\ell=1}^{H} N_\ell s_{U_\ell}}$	$n\frac{N_h s_{\omega_h}}{\sum_{\ell=1}^{H} N_\ell s_{\omega_\ell}}$

Autre notions utilisées autour de $\overline{y}_U$ (niveau : $100(1-\alpha)\%$, $\alpha \in]0,1[$) :

Intervalle de confiance	$i_{\overline{y}_U} = \left[\overline{y}_\omega - z_\alpha\sqrt{\frac{1}{N^2}\sum_{h=1}^{H} N_h^2(1-f_h)\frac{s_{\omega_h}^2}{n_h}},\ \overline{y}_\omega + z_\alpha\sqrt{\frac{1}{N^2}\sum_{h=1}^{H} N_h^2(1-f_h)\frac{s_{\omega_h}^2}{n_h}}\right]$
Incertitude absolue	$d_\omega = z_\alpha\sqrt{\frac{1}{N^2}\sum_{h=1}^{H} N_h^2(1-f_h)\frac{s_{\omega_h}^2}{n_h}}$
Incertitude relative	$d_\omega^* = \frac{d_\omega}{\overline{y}_\omega}$
Taille n telle que $d_\omega \leq d_0$	∘ pour un plan de sondage aléatoire de type STP : $n \geq \frac{N z_\alpha^2 \sum_{h=1}^{H} N_h s_{\omega_h}^2}{N^2 d_0^2 + z_\alpha^2 \sum_{h=1}^{H} N_h s_{\omega_h}^2}$, ∘ pour un plan de sondage aléatoire de type STO : $n \geq \frac{z_\alpha^2\left(\sum_{h=1}^{H} N_h s_{\omega_h}\right)^2}{N^2 d_0^2 + z_\alpha^2 \sum_{h=1}^{H} N_h s_{\omega_h}^2}$.
Taille n telle que $d_\omega^* \leq d_1$	∘ pour un plan de sondage aléatoire de type STP : $n \geq \frac{N z_\alpha^2 \sum_{h=1}^{H} N_h s_{\omega_h}^2}{N^2 (d_1\overline{y}_\omega)^2 + z_\alpha^2 \sum_{h=1}^{H} N_h s_{\omega_h}^2}$, ∘ pour un plan de sondage aléatoire de type STO : $n \geq \frac{z_\alpha^2\left(\sum_{h=1}^{H} N_h s_{\omega_h}\right)^2}{N^2 (d_1\overline{y}_\omega)^2 + z_\alpha^2 \sum_{h=1}^{H} N_h s_{\omega_h}^2}$.

Rappel : $\mathbb{P}(|Z| \geq z_\alpha) = \alpha$, $Z \sim \mathcal{N}(0,1)$.

7 Total, proportion et effectif dans le cadre ST

On reprend le cadre mathématique d'un plan de sondage aléatoire de type ST.

7.1 Estimation du total

Total :

On appelle total-population le réel :

$$\tau_U = \sum_{i=1}^{N} y_i = N\overline{y}_U = \sum_{h=1}^{H} N_h \overline{y}_{W_h}.$$

Estimation aléatoire de τ_U :

Un estimateur aléatoire de τ_U est

$$\tau_W = N\overline{y}_W.$$

Espérance de τ_W :

L'estimateur τ_W est sans biais pour τ_U :

$$\mathbb{E}(\tau_W) = \tau_U.$$

Variance de τ_W :

La variance de τ_W est

$$\mathbb{V}(\tau_W) = \sum_{h=1}^{H} N_h^2 (1 - f_h) \frac{s_{U_h}^2}{n_h}.$$

Erreur quadratique moyenne de τ_W :

L'erreur quadratique moyenne de τ_W est le réel :

$$EQM(\tau_W)[PESR] = \sum_{h=1}^{H} N_h^2 (1 - f_h) \frac{s_{U_h}^2}{n_h}.$$

Estimation ponctuelle de τ_U :

Soit ω un échantillon de n individus de U. Une estimation ponctuelle de τ_U est le total-échantillon :

$$\tau_\omega = N\overline{y}_\omega.$$

Estimation ponctuelle de l'écart-type de τ_W :

Soit ω un échantillon de n individus de U. Une estimation ponctuelle de l'écart-type de τ_W est le réel :

$$s(\tau_\omega) = \sqrt{\sum_{h=1}^{H} N_h^2 (1 - f_h) \frac{s_{\omega_h}^2}{n_h}}.$$

Intervalle de confiance pour τ_U :

Soit ω un échantillon de n individus de U. Un intervalle de confiance pour τ_U au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, est

$$\begin{aligned} i_{\tau_U} &= [\tau_\omega - z_\alpha s(\tau_\omega),\ \tau_\omega + z_\alpha s(\tau_\omega)] \\ &= \left[\tau_\omega - z_\alpha \sqrt{\sum_{h=1}^{H} N_h^2 (1 - f_h) \frac{s_{\omega_h}^2}{n_h}},\ \tau_\omega + z_\alpha \sqrt{\sum_{h=1}^{H} N_h^2 (1 - f_h) \frac{s_{\omega_h}^2}{n_h}}\right] = N \times i_{\overline{y}_U}, \end{aligned}$$

où z_α est le réel vérifiant $\mathbb{P}(|Z| \geq z_\alpha) = \alpha$, $Z \sim \mathcal{N}(0,1)$.

On peut également définir l'incertitude absolue ou relative sur τ_U, ainsi que la taille d'échantillon souhaitée pour une incertitude donnée.

7.2 Estimation d'une proportion

Contexte : On suppose que le caractère Y est binaire : $Y(\Omega) = \{0,1\}$. Cela correspond à un codage.

Proportion :

On appelle proportion-population la proportion des individus dans U vérifiant $Y = 1$:

$$p_U = \frac{1}{N} \sum_{i=1}^{N} y_i \qquad (= \overline{y}_U).$$

Estimation d'une proportion :

Un estimateur aléatoire de p_U est

$$p_W = \frac{1}{N} \sum_{h=1}^{H} N_h p_{U_h},$$

avec $p_{U_h} = \overline{y}_{W_h}$.

Espérance de p_W :

L'estimateur p_W est sans biais pour p_U :

$$\mathbb{E}(p_W) = p_U.$$

Variance de p_W :

La variance de p_W est

$$\mathbb{V}(p_W) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2(1-f_h)\frac{s_{U_h}^2}{n_h} = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2(1-f_h)\frac{N_h}{n_h(N_h-1)}p_{U_h}(1-p_{U_h}).$$

Erreur quadratique moyenne de p_W :

L'erreur quadratique moyenne de p_W est le réel :

$$EQM(p_W)[ST] = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2(1-f_h)\frac{N_h}{n_h(N_h-1)}p_{U_h}(1-p_{U_h}).$$

Estimation ponctuelle de p_U :

Soit ω un échantillon de n individus de U. Une estimation ponctuelle de p_U est la proportion-échantillon :

$$p_\omega = \overline{y}_\omega = \frac{1}{N}\sum_{h=1}^{H} N_h p_{\omega_h}, \qquad p_{\omega_h} = \overline{y}_{\omega_h}.$$

Estimation ponctuelle de l'écart-type de p_W :

Soit ω un échantillon de n individus de U. Une estimation ponctuelle de l'écart-type de p_W est le réel :

$$s(p_\omega) = \sqrt{\frac{1}{N^2}\sum_{h=1}^{H} N_h^2(1-f_h)\frac{p_{\omega_h}(1-p_{\omega_h})}{n_h-1}}.$$

Intervalle de confiance pour p_U :

Soit ω un échantillon de n individus de U. Un intervalle de confiance pour p_U au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, est

$$\begin{aligned} i_{p_U} &= [p_\omega - z_\alpha s(p_\omega),\ p_\omega + z_\alpha s(p_\omega)] \\ &= \left[p_\omega - z_\alpha\sqrt{\frac{1}{N^2}\sum_{h=1}^{H} N_h^2(1-f_h)\frac{p_{\omega_h}(1-p_{\omega_h})}{n_h-1}},\ p_\omega + z_\alpha\sqrt{\frac{1}{N^2}\sum_{h=1}^{H} N_h^2(1-f_h)\frac{p_{\omega_h}(1-p_{\omega_h})}{n_h-1}}\right], \end{aligned}$$

où z_α est le réel vérifiant $\mathbb{P}(|Z| \geq z_\alpha) = \alpha$, $Z \sim \mathcal{N}(0,1)$.

Quelques commandes R : Un exemple de calcul de i_{p_U} avec R est décrit ci-dessous :

```
icpST= function(N_h, y, niveau) {
N = sum(N_h)
n_h = unlist(lapply(y, length))
bar_y_h = unlist(lapply(y,mean))
p_w = sum(N_h * bar_y_h) / N
var_p_w = (1 / N^2) * sum(N_h^2 * (1-n_h / N_h) *
(p_w * (1 - p_w) / (n_h - 1)))
z = qnorm(1 - (1 - niveau) / 2)
a = p_w - z * sqrt(var_p_w)
b = p_w + z * sqrt(var_p_w)
print(c(a, b)) }
N_h = c(181, 54, 73)
y_1 = c(0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 0, 1, 0, 0)
y_2 = c(1, 1, 0, 1, 1, 0, 0, 1, 1, 0)
y_3 = c(0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0)
y = list(y_1, y_2, y_3)
icpST(N_h, y, 0.95)
```

Cela renvoie : 0.25874, 0.60120.

Pour demander un niveau de 99%, on fait :

```
icpST(N_h, y, 0.99)
```

Cela renvoie : 0.20494, 0.65501.

Plan de sondage STP :

En pratique, pour tout $h \in \{1, \dots, H\}$, on considère le plus petit entier n_h tel que

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

Plan de sondage STO :

Soit $\omega = (\omega_1, \ldots, \omega_H)$ un échantillon prélevé lors d'une étude préliminaire. En pratique, pour tout $h \in \{1, \ldots, H\}$, on prend le plus petit entier n_h tel que

$$n_h \geq n \frac{N_h \sqrt{p_{\omega_h}(1 - p_{\omega_h})}}{\sum_{\ell=1}^{H} N_\ell \sqrt{p_{\omega_\ell}(1 - p_{\omega_\ell})}}.$$

Incertitude absolue :

Soit ω un échantillon de n individus de U. On appelle incertitude absolue sur p_U au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, la demi-longueur de i_{p_U} :

$$d_\omega = z_\alpha s(\overline{y}_\omega) = z_\alpha \sqrt{\frac{1}{N^2} \sum_{h=1}^{H} N_h^2 (1 - f_h) \frac{p_{\omega_h}(1 - p_{\omega_h})}{n_h - 1}}.$$

Plus d_ω est petit, plus l'estimation de p_U par p_ω est précise.

Incertitude relative :

Soit ω un échantillon de n individus de U et d_ω l'incertitude absolue sur p_U au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$. On appelle incertitude relative sur i_{p_U} au niveau $100(1-\alpha)\%$ le pourcentage $(100 \times d_\omega^*)\%$ où d_ω^* est le réel :

$$d_\omega^* = \frac{d_\omega}{p_\omega}.$$

Taille d'échantillon à partir de l'incertitude absolue :

Soit ω un échantillon prélevé lors d'une étude préliminaire. La taille d'échantillon n à choisir pour avoir une incertitude absolue sur p_U au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, inférieure ou égale à d_0 est le plus petit n tel que $d_\omega \leq d_0$. En particulier, cela entraîne,

- pour un plan de sondage aléatoire de type STP :

$$n \geq \frac{N z_\alpha^2 \sum_{h=1}^{H} N_h p_{\omega_h}(1-p_{\omega_h})}{N^2 d_0^2 + z_\alpha^2 \sum_{h=1}^{H} N_h p_{\omega_h}(1-p_{\omega_h})},$$

- pour un plan de sondage aléatoire de type STO :

$$n \geq \frac{z_\alpha^2 \left(\sum_{h=1}^{H} N_h \sqrt{p_{\omega_h}(1-p_{\omega_h})} \right)^2}{N^2 d_0^2 + z_\alpha^2 \sum_{h=1}^{H} N_h p_{\omega_h}(1-p_{\omega_h})}.$$

Quelques commandes R : Un exemple de fonction R pour calculer la taille n d'un échantillon à partir de l'incertitude absolue sur p_U pour un plan de sondage aléatoire de type STP au niveau $100(1-\alpha)\%$ est décrit ci-dessous :

```
n_ech = function(N_h, p_w_h, d0, niveau) {
N = sum(N_h)
z = qnorm(1 - (1 - niveau) / 2)
n = (N * z^2 * sum(N_h * p_w_h * (1 - p_w_h) )) /
(N^2 * d0^2 + z^2 * sum(N_h * p_w_h * (1 - p_w_h)))
print(ceiling(n)) }
N_h = c(15, 12, 134)
p_w_h = c(0.75, 0.21, 0.55)
n_ech(N_h, p_w_h, d0 = 0.3, niveau = 0.95)
```

Cela renvoie 10.

Taille d'échantillon à partir de l'incertitude relative :

Soit ω un échantillon prélevé lors d'une étude préliminaire. La taille d'échantillon n à choisir pour avoir une incertitude relative sur p_U au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, inférieure ou égale à $(100\times d_1)\%$ est le plus petit n tel que $d^*_\omega \leq d_1$. En particulier, cela entraîne,

- pour un plan de sondage aléatoire de type STP :

$$n \geq \frac{N z_\alpha^2 \sum_{h=1}^{H} N_h p_{\omega_h}(1-p_{\omega_h})}{N^2(d_1 p_\omega)^2 + z_\alpha^2 \sum_{h=1}^{H} N_h p_{\omega_h}(1-p_{\omega_h})},$$

- pour un plan de sondage aléatoire de type STO :

$$n \geq \frac{z_\alpha^2 \left(\sum_{h=1}^{H} N_h \sqrt{p_{\omega_h}(1-p_{\omega_h})}\right)^2}{N^2(d_1 p_\omega)^2 + z_\alpha^2 \sum_{h=1}^{H} N_h p_{\omega_h}(1-p_{\omega_h})}.$$

7.3 Estimation d'un effectif

Contexte : On suppose que le caractère Y est binaire : $Y(\Omega) = \{0,1\}$. Cela correspond à un codage.

Effectif :

On appelle effectif-population le nombre des individus dans U vérifiant $Y = 1$:

$$\eta_U = N p_U.$$

Estimation aléatoire de η_U :

Un estimateur aléatoire de η_U est

$$\eta_W = N p_W.$$

Espérance de η_W :

L'estimateur η_W est sans biais pour η_U :

$$\mathbb{E}(\eta_W) = \eta_U.$$

Variance de η_W :

La variance de η_W est

$$\mathbb{V}(\eta_W) = \sum_{h=1}^{H} N_h^2(1-f_h)\frac{s_{\omega_h}^2}{n_h} = \sum_{h=1}^{H} N_h^2(1-f_h)\frac{N_h}{n_h(N_h-1)}p_{U_h}(1-p_{U_h}).$$

Erreur quadratique moyenne de η_W :

L'erreur quadratique moyenne de η_W est le réel :

$$EQM(\eta_W)[ST] = \sum_{h=1}^{H} N_h^2(1-f_h)\frac{N_h}{n_h(N_h-1)}p_{U_h}(1-p_{U_h}).$$

Estimation ponctuelle de η_U :

Soit ω un échantillon de n individus de U. Une estimation ponctuelle de η_U est la proportion-échantillon :

$$\eta_\omega = Np_\omega = \sum_{h=1}^{H} N_h p_{\omega_h}.$$

Estimation ponctuelle de l'écart-type de η_W :

Soit ω un échantillon de n individus de U. Une estimation ponctuelle de l'écart-type de η_W est le réel :

$$s(\eta_\omega) = \sqrt{\sum_{h=1}^{H} N_h^2(1-f_h)\frac{p_{\omega_h}(1-p_{\omega_h})}{n_h-1}}.$$

Intervalle de confiance pour η_U :

Soit ω un échantillon de n individus de U. Un intervalle de confiance pour η_U au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, est

$$\begin{aligned} i_{\eta_U} &= [\eta_\omega - z_\alpha s(\eta_\omega),\ \eta_\omega + z_\alpha s(\eta_\omega)] \\ &= \left[\eta_\omega - z_\alpha\sqrt{\sum_{h=1}^{H} N_h^2(1-f_h)\frac{p_{\omega_h}(1-p_{\omega_h})}{n_h-1}},\ \eta_\omega + z_\alpha\sqrt{\sum_{h=1}^{H} N_h^2(1-f_h)\frac{p_{\omega_h}(1-p_{\omega_h})}{n_h-1}}\right] \\ &= N \times i_{p_U}, \end{aligned}$$

où z_α est le réel vérifiant $\mathbb{P}(|Z| \geq z_\alpha) = \alpha$, $Z \sim \mathcal{N}(0,1)$.

On peut également définir l'incertitude absolue ou relative sur η_U, ainsi que la taille d'échantillon souhaitée pour une incertitude donnée.

7.4 Exercices corrigés

Exercice 1 : Sur les 6000 employés d'une entreprise, on souhaite connaître la proportion p_U d'entre eux qui sont propriétaires de leur logement. On décide de former 3 strates en fonction du revenu des employés. On considère alors :

- la strate U_1 : ensemble des employés à revenu faible,
- la strate U_2 : ensemble des employés à revenu modeste,
- la strate U_3 : ensemble des employés à revenu fort.

On dispose des informations suivantes :

U_h	U_1	U_2	U_3
N_h	2800	2200	1000
n_h	210	200	110
p_{ω_h}	0.11	0.55	0.85

1. Donner une estimation ponctuelle de p_U.
2. Donner une estimation ponctuelle de l'écart-type de l'estimateur de p_U.
3. Déterminer un intervalle de confiance pour p_U au niveau 95%.

Solution :

1. On a $H = 3$. Une estimation ponctuelle de p_U est

$$p_\omega = \frac{1}{N}\sum_{h=1}^{H} N_h p_{\omega_h} = \frac{1}{6000}(2800 \times 0.11 + 2200 \times 0.55 + 1000 \times 0.85) = 0.39466.$$

2. On a

$$\begin{aligned} s^2(p_\omega) &= \frac{1}{N^2}\sum_{h=1}^{H} N_h^2(1-f_h)\frac{p_{\omega_h}(1-p_{\omega_h})}{n_h-1} \\ &= \frac{1}{6000^2}\left(2800^2\left(1-\frac{210}{2800}\right)\frac{0.11(1-0.11)}{210-1} + 2200^2\left(1-\frac{200}{2200}\right)\frac{0.55(1-0.55)}{200-1}\right. \\ &\left. + 1000^2\left(1-\frac{110}{1000}\right)\frac{0.85(1-0.85)}{110-1}\right) \\ &= 0.0002752. \end{aligned}$$

Donc

$$s(p_\omega) = 0.016589.$$

3. On a $95\% = 100(1-\alpha)\%$ avec $\alpha = 0.05$. On a $\mathbb{P}(|Z| \geq z_\alpha) = \alpha = 0.05$, $Z \sim \mathcal{N}(0,1)$, avec $z_\alpha = 1.96$. Un intervalle de confiance pour τ_U au niveau 95% est

$$\begin{aligned} i_{p_U} &= [p_\omega - z_\alpha s(p_\omega),\ p_\omega + z_\alpha s(p_\omega)] \\ &= [0.39466 - 1.96 \times 0.016589,\ 0.39466 - 1.96 \times 0.016589] \\ &= [0.36214,\ 0.42717]. \end{aligned}$$

Ainsi, il y a 95 chances sur 100 que $[0.36214,\ 0.42717]$ contienne p_U.

Exercice 2 : On veut estimer le taux de réussite à la session d'examens de juin dans une université qui comprend 950 inscrits en première année, 700 en deuxième, 430 en troisième et 400 en quatrième. On veut estimer le taux de réussite à partir des résultats de 500 étudiants.

1. On prélève un échantillon de 500 étudiants suivant un plan de sondage aléatoire de type PESR. On trouve un taux de réussite de 72%. Donner un intervalle de confiance du taux de réussite global au niveau 95%.
2. Est-ce que l'estimation aurait été meilleure avec un plan de sondage aléatoire de type ST avec pour strates les années d'étude ?
3. Combien d'étudiants aurait-il fallu prendre par année pour faire un plan de sondage aléatoire de type STP ?
4. Avec un échantillon de 500 étudiants prélevé suivant un plan de sondage aléatoire de type STP, on obtient :

$$p_{\omega_1} = 0.62, \qquad p_{\omega_2} = 0.72, \qquad p_{\omega_3} = 0.78, \qquad p_{\omega_4} = 0.83.$$

Donner une estimation ponctuelle de taux de réussite global.

Solution :

1. Soit p_U le taux de réussite global. Par l'énoncé, on a $p_\omega = 0.72$. On a $95\% = 100(1-\alpha)\%$ avec $\alpha = 0.05$. On a $\mathbb{P}(|Z| \geq z_\alpha) = \alpha = 0.05$, $Z \sim \mathcal{N}(0,1)$, avec $z_\alpha = 1.96$.

Un intervalle de confiance pour p_U au niveau 95% est

$$\begin{aligned} i_{p_U} &= \left[p_\omega - z_\alpha \sqrt{(1-f)\frac{p_\omega(1-p_\omega)}{n-1}},\ p_\omega + z_\alpha \sqrt{(1-f)\frac{p_\omega(1-p_\omega)}{n-1}} \right] \\ &= \left[0.72 - 1.96 \sqrt{\left(1 - \frac{500}{2480}\right)\frac{0.72(1-0.72)}{500-1}}, \right. \\ &\quad \left. 0.72 + 1.96 \sqrt{\left(1 - \frac{500}{2480}\right)\frac{0.72(1-0.72)}{500-1}} \right] \\ &= [0.6847,\ 0.7552]. \end{aligned}$$

Ainsi, il y a 95 chances sur 100 que $[0.6847,\ 0.7552]$ contienne p_U.

2. Oui, il est fort probable qu'un plan de sondage aléatoire de type ST avec pour strates les années d'étude aurait amené une meilleure estimation.

3. Pour faire un plan de sondage aléatoire de type STP, il faut choisir les plus petites tailles d'échantillons : $n_1, \ldots, n_4$ telles que, pour tout $h \in \{1, \ldots, 4\}$,

$$n_h \geq \frac{n}{N} N_h.$$

Il vient

$$n_1 \geq \frac{500}{2480} 950 = 191.5323, \qquad n_2 \geq \frac{500}{2480} 700 = 141.129, \qquad n_3 \geq \frac{500}{2480} 430 = 86.69355,$$

et

$$n_1 \geq \frac{500}{2480} 400 = 80.64516.$$

Donc $n_1 = 192$, $n_2 = 142$, $n_3 = 87$ et $n_4 = 81$. On a $\sum_{h=1}^{4} n_h = 502 \neq 500$, on ajuste : $n_1 = 191$, $n_2 = 141$, $n_3 = 87$ et $n_4 = 81$.

4. Une estimation ponctuelle de p_U est

$$p_\omega = \frac{1}{N} \sum_{h=1}^{4} N_h p_{\omega_h} = \frac{1}{2480}(950 \times 0.62 + 700 \times 0.72 + 430 \times 0.78 + 400 \times 0.83) = 0.7098.$$

7.5 Synthèse : proportion

Paramètres-strates et les paramètres-échantillon correspondants, $\omega = (\omega_1, \ldots, \omega_H)$:

	Strate U_h	Échantillon ω_h
Taille	N_h	n_h
Taux de sondage	$\square$	$f_h = \dfrac{n_h}{N_h}$
Proportion	$p_{U_h} = \dfrac{1}{N_h}\displaystyle\sum_{i=1}^{N_h} y_i$	$p_{\omega_h} = \dfrac{1}{n_h}\displaystyle\sum_{i=1}^{N_h} y_i \mathbb{1}_{\{u_i \in \omega_h\}}$
Écart-type de p_{W_h}	$\sigma(p_{W_h}) = \sqrt{(1-f_h)\dfrac{N_h}{n_h(N_h-1)} p_{U_h}(1-p_{U_h})}$	$s(p_{\omega_h}) = \sqrt{(1-f_h)\dfrac{p_{U_h}(1-p_{U_h})}{n_h - 1}}$

Paramètres-population et les paramètres-échantillon correspondants :

	Population U	Échantillon ω
Taille	N	n
Moyenne	$p_U = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} y_i$	$p_\omega = \dfrac{1}{N}\displaystyle\sum_{h=1}^{H} N_h p_{\omega_h}$
Écart-type de p_W	$\sigma(p_W) = \sqrt{\dfrac{1}{N^2}\displaystyle\sum_{h=1}^{H} N_h^2 \sigma^2(p_{W_h})}$	$s(p_\omega) = \sqrt{\dfrac{1}{N^2}\displaystyle\sum_{h=1}^{H} N_h^2 (1-f_h)\dfrac{p_{\omega_h}(1-p_{\omega_h})}{n_h - 1}}$

Plans de sondage aléatoires de types STP et STO :

	STP	STO	STO (applicable)
n_h	$\frac{n}{N}N_h$	$n\frac{N_h\sqrt{p_{U_h}(1-p_{U_h})}}{\sum_{\ell=1}^{H}N_\ell\sqrt{p_{U_\ell}(1-p_{U_\ell})}}$	$n\frac{N_h\sqrt{p_{\omega_h}(1-p_{\omega_h})}}{\sum_{\ell=1}^{H}N_\ell\sqrt{p_{\omega_\ell}(1-p_{\omega_\ell})}}$

Autre notions utilisées autour de p_U (niveau : $100(1-\alpha)\%$, $\alpha\in]0,1[$) :

Intervalle de confiance	$i_{p_U}=\left[p_\omega - z_\alpha\sqrt{\frac{1}{N^2}\sum_{h=1}^{H}N_h^2(1-f_h)\frac{p_{\omega_h}(1-p_{\omega_h})}{n_h-1}},\ p_\omega + z_\alpha\sqrt{\frac{1}{N^2}\sum_{h=1}^{H}N_h^2(1-f_h)\frac{p_{\omega_h}(1-p_{\omega_h})}{n_h-1}}\right]$
Incertitude absolue	$d_\omega = z_\alpha\sqrt{\frac{1}{N^2}\sum_{h=1}^{H}N_h^2(1-f_h)\frac{p_{\omega_h}(1-p_{\omega_h})}{n_h-1}}$
Incertitude relative	$d_\omega^* = \frac{d_\omega}{\overline{y}_\omega}$
Taille n telle que $d_\omega \le d_0$	∘ pour un plan de sondage aléatoire de type STP : $n \ge \frac{Nz_\alpha^2\sum_{h=1}^{H}N_hp_{\omega_h}(1-p_{\omega_h})}{N^2d_0^2+z_\alpha^2\sum_{h=1}^{H}N_hp_{\omega_h}(1-p_{\omega_h})}$, ∘ pour un plan de sondage aléatoire de type STO : $n \ge \frac{z_\alpha^2\left(\sum_{h=1}^{H}N_h\sqrt{p_{\omega_h}(1-p_{\omega_h})}\right)^2}{N^2d_0^2+z_\alpha^2\sum_{h=1}^{H}N_hp_{\omega_h}(1-p_{\omega_h})}$.
Taille n telle que $d_\omega^* \le d_1$	∘ pour un plan de sondage aléatoire de type STP : $n \ge \frac{Nz_\alpha^2\sum_{h=1}^{H}N_hs_{\omega_h}^2}{N^2(d_1p_\omega)^2+z_\alpha^2\sum_{h=1}^{H}N_hs_{\omega_h}^2}$, ∘ pour un plan de sondage aléatoire de type STO : $n \ge \frac{z_\alpha^2\left(\sum_{h=1}^{H}N_h\sqrt{p_{\omega_h}(1-p_{\omega_h})}\right)^2}{N^2(d_1p_\omega)^2+z_\alpha^2\sum_{h=1}^{H}N_hp_{\omega_h}(1-p_{\omega_h})}$.

Rappel : $\mathbb{P}(|Z|\ge z_\alpha)=\alpha$, $Z\sim\mathcal{N}(0,1)$.

8 Plan de sondage aléatoire à probabilités inégales sans remise (PISR)

8.1 Contexte

Probabilités inégales (PI) :

Un plan de sondage aléatoire est dit à probabilités inégales (PI) si au moins 2 individus n'ont pas la même probabilité d'être sélectionné. De plus, il est dit PISR si il est à probabilités inégales et si un même indiivdu ne peut apparaître qu'une seule fois dans l'échantillon.
Ainsi, en notant W la *var* égale à l'echantillon obtenu, il existe deux individus u_i et u_j tels que

$$\mathbb{P}(u_i \in W) \neq \mathbb{P}(u_j \in W).$$

Quelques commandes R : Un exemple de sondage aléatoire de type PISR est décrit-ci-dessous :

```
U = c("Bob", "Nico", "Ali", "Fabien", "Malik", "John", "Jean", "Chris", "Karl")
p = c(0.1, 0.1, 0.1, 0.1, 0.1, 0.9, 0.9, 0.9, 0.9)
t = sample(U, 3, replace = F, prob = p)
t
```

Notations ; probabilités d'appartenance : On adopte les notations suivantes :

- la probabilité que l'individu ω_i appartienne à W :

$$\pi_i = \mathbb{P}(u_i \in W).$$

- la probabilité que les individus ω_i et ω_j appartiennent à W :

$$\pi_{i,j} = \mathbb{P}((u_i, u_j) \in W).$$

Dans la suite : On se place dans le cadre d'un plan de sondage aléatoire de type PISR.

Propriétés des probabilités d'appartenance :

On a

- $\sum_{i=1}^{N} \pi_i = n$,
- $\sum_{\substack{j=1 \\ j\neq i}}^{N} \pi_{i,j} = (n-1)\pi_i$,
- $\sum_{\substack{j=1 \\ j\neq i}}^{N} (\pi_{i,j} - \pi_i\pi_j) = -\pi_i(1-\pi_i)$.

Preuve :

- Soit W_m est la *var* égale au m-ème individu de l'échantillon : $W = (W_1, \ldots, W_n)$. Comme tous les individus sont différents, on a

$$\pi_i = \mathbb{P}(u_i \in W) = \mathbb{P}\left(\bigcup_{m=1}^{n} \{W_m = u_i\}\right) = \sum_{m=1}^{n} \mathbb{P}(W_m = u_i).$$

  Avec des arguments identiques, comme $\mathbb{P}(W_m \in U) = 1$,

$$\sum_{i=1}^{N} \pi_i = \sum_{i=1}^{N}\sum_{m=1}^{n} \mathbb{P}(W_m = u_i) = \sum_{m=1}^{n}\left(\sum_{i=1}^{N} \mathbb{P}(W_m = u_i)\right) = \sum_{m=1}^{n} \mathbb{P}(W_m \in U) = n.$$

- Pour $i \neq j$, on a

$$\begin{aligned}\pi_{i,j} = \mathbb{P}((u_i, u_j) \in W) &= \mathbb{P}\left(\bigcup_{m=1}^{n}\bigcup_{\substack{\ell=1 \\ \ell\neq m}}^{n} \{W_m = u_i\} \cap \{W_\ell = u_j\}\right) \\ &= \sum_{m=1}^{n}\sum_{\substack{\ell=1 \\ \ell\neq m}}^{n} \mathbb{P}(\{W_m = u_i\} \cap \{W_\ell = u_j\}).\end{aligned}$$

Comme, pour $\ell \neq m$, on a $\{W_m = u_i\} \subseteq \{W_\ell \in U - \{u_i\}\}$, il vient

$$\begin{aligned}
\sum_{\substack{j=1\\ j\neq i}}^{N} \pi_{i,j} &= \sum_{\substack{j=1\\ j\neq i}}^{N} \sum_{m=1}^{n} \sum_{\substack{\ell=1\\ \ell\neq m}}^{n} \mathbb{P}(\{W_m = u_i\} \cap \{W_\ell = u_j\}) \\
&= \sum_{m=1}^{n} \sum_{\substack{\ell=1\\ \ell\neq m}}^{n} \mathbb{P}\left(\{W_m = u_i\} \cap \bigcup_{\substack{j=1\\ j\neq i}}^{N} \{W_\ell = u_j\}\right) \\
&= \sum_{m=1}^{n} \sum_{\substack{\ell=1\\ \ell\neq m}}^{n} \mathbb{P}\left(\{W_m = u_i\} \cap \{W_\ell \in U - \{u_i\}\}\right) \\
&= \sum_{m=1}^{n} \sum_{\substack{\ell=1\\ \ell\neq m}}^{n} \mathbb{P}\left(W_m = u_i\right) = (n-1) \sum_{m=1}^{n} \mathbb{P}(W_m = u_i) = (n-1)\pi_i.
\end{aligned}$$

- Par les éqgalités : $\sum_{i=1}^{N} \pi_i = n$ et $\sum_{\substack{j=1\\ j\neq i}}^{N} \pi_{i,j} = (n-1)\pi_i$, on obtient

$$\begin{aligned}
\sum_{\substack{j=1\\ j\neq i}}^{N} (\pi_{i,j} - \pi_i\pi_j) &= \sum_{\substack{j=1\\ j\neq i}}^{N} \pi_{i,j} - \pi_i \sum_{\substack{j=1\\ j\neq i}}^{N} \pi_j = \sum_{\substack{j=1\\ j\neq i}}^{N} \pi_{i,j} - \pi_i \left(\sum_{j=1}^{N} \pi_j - \pi_i\right) \\
&= (n-1)\pi_i - \pi_i(n - \pi_i) = -\pi_i(1-\pi_i).
\end{aligned}$$

□

8.2 Estimateurs

Estimation aléatoire de $\overline{y}_U$ (estimateur de Horvitz-Thompson) :

Un estimateur aléatoire de $\overline{y}_U$ est l'estimateur de Horvitz-Thompson :

$$\overline{y}_W = \frac{1}{N} \sum_{i=1}^{N} \frac{y_i}{\pi_i} \mathbb{1}_{\{u_i \in W\}}.$$

Espérance de $\overline{y}_W$:

L'estimateur $\overline{y}_W$ est sans biais pour $\overline{y}_U$:

$$\mathbb{E}(\overline{y}_W) = \overline{y}_U.$$

Preuve : En utilisant la linéarité de l'espérance, $\mathbb{E}\left(\mathbb{1}_A\right) = \mathbb{P}(A)$ et $\mathbb{P}(u_i \in W) = \pi_i$, il vient

$$\begin{aligned}\mathbb{E}(\overline{y}_W) &= \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\pi_i}\mathbb{1}_{\{u_i\in W\}}\right) = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\pi_i}\mathbb{E}\left(\mathbb{1}_{\{u_i\in W\}}\right)\\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\pi_i}\mathbb{P}(u_i\in W) = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\pi_i}\pi_i = \frac{1}{N}\sum_{i=1}^{N}y_i = \overline{y}_U.\end{aligned}$$

□

Variance de $\overline{y}_W$:

La variance de $\overline{y}_W$ est

$$\mathbb{V}(\overline{y}_W) = \frac{1}{N^2}\left(\sum_{i=1}^{N}\frac{y_i^2}{\pi_i^2}\pi_i(1-\pi_i) + \sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}(\pi_{i,j}-\pi_i\pi_j)\right).$$

Preuve : Par la formule de la variance d'une somme de *var*, on obtient

$$\begin{aligned}\mathbb{V}(\overline{y}_W) &= \mathbb{V}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\pi_i}\mathbb{1}_{\{u_i\in W\}}\right) = \frac{1}{N^2}\mathbb{V}\left(\sum_{i=1}^{N}\frac{y_i}{\pi_i}\mathbb{1}_{\{u_i\in W\}}\right)\\ &= \frac{1}{N^2}\left(\sum_{i=1}^{N}\mathbb{V}\left(\frac{y_i}{\pi_i}\mathbb{1}_{\{u_i\in W\}}\right) + \sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\mathbb{C}\left(\frac{y_i}{\pi_i}\mathbb{1}_{\{u_i\in W\}},\frac{y_j}{\pi_j}\mathbb{1}_{\{u_j\in W\}}\right)\right)\\ &= \frac{1}{N^2}\left(\sum_{i=1}^{N}\frac{y_i^2}{\pi_i^2}\mathbb{V}\left(\mathbb{1}_{\{u_i\in W\}}\right) + \sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\mathbb{C}\left(\mathbb{1}_{\{u_i\in W\}},\mathbb{1}_{\{u_j\in W\}}\right)\right).\end{aligned}$$

Or

$$\begin{aligned}\mathbb{V}\left(\mathbb{1}_{\{u_i\in W\}}\right) &= \mathbb{E}\left(\mathbb{1}_{\{u_i\in W\}}^2\right) - \left(\mathbb{E}\left(\mathbb{1}_{\{u_i\in W\}}\right)\right)^2 = \mathbb{P}(u_i\in W) - \left(\mathbb{P}(u_i\in W)\right)^2\\ &= \pi_i - \pi_i^2 = \pi_i(1-\pi_i).\end{aligned}$$

De plus

$$\begin{aligned}\mathbb{C}\left(\mathbb{1}_{\{u_i\in W\}},\mathbb{1}_{\{u_j\in W\}}\right) &= \mathbb{E}\left(\mathbb{1}_{\{u_i\in W\}}\mathbb{1}_{\{u_j\in W\}}\right) - \mathbb{E}\left(\mathbb{1}_{\{u_i\in W\}}\right)\mathbb{E}\left(\mathbb{1}_{\{u_j\in W\}}\right)\\ &= \mathbb{P}(\{u_i\in W\}\cap\{u_j\in W\}) - \mathbb{P}(u_i\in W)\mathbb{P}(u_j\in W) = \pi_{i,j} - \pi_i\pi_j.\end{aligned}$$

En combinant ces égalités, on obtient

$$\mathbb{V}(\overline{y}_W) = \frac{1}{N^2}\left(\sum_{i=1}^{N}\frac{y_i^2}{\pi_i^2}\pi_i(1-\pi_i) + \sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}(\pi_{i,j}-\pi_i\pi_j)\right).$$

□

Autre expression de la variance de $\overline{y}_W$:

La variance de $\overline{y}_W$ est

$$\mathbb{V}(\overline{y}_W) = \frac{1}{N^2}\sum_{i=2}^{N}\sum_{j=1}^{i-1}(\pi_i\pi_j-\pi_{i,j})\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2.$$

Preuve : En utilisant l'égalité : $\pi_i(1-\pi_i) = -\sum_{\substack{j=1\\j\neq i}}^{N}(\pi_{i,j}-\pi_i\pi_j)$, on obtient

$$\begin{aligned}
\mathbb{V}(\overline{y}_W) &= \frac{1}{N^2}\left(\sum_{i=1}^{N}\frac{y_i^2}{\pi_i^2}\pi_i(1-\pi_i) + \sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}(\pi_{i,j}-\pi_i\pi_j)\right)\\
&= -\frac{1}{N^2}\left(\sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\frac{y_i^2}{\pi_i^2}(\pi_{i,j}-\pi_i\pi_j) - \sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}(\pi_{i,j}-\pi_i\pi_j)\right)\\
&= -\frac{1}{N^2}\left(\sum_{i=2}^{N}\sum_{j=1}^{i-1}\left(\frac{y_i^2}{\pi_i^2}+\frac{y_j^2}{\pi_j^2}\right)(\pi_{i,j}-\pi_i\pi_j) - 2\sum_{i=2}^{N}\sum_{j=1}^{i-1}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}(\pi_{i,j}-\pi_i\pi_j)\right)\\
&= -\frac{1}{N^2}\sum_{i=2}^{N}\sum_{j=1}^{i-1}(\pi_{i,j}-\pi_i\pi_j)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2 = \frac{1}{N^2}\sum_{i=2}^{N}\sum_{j=1}^{i-1}(\pi_i\pi_j-\pi_{i,j})\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2.
\end{aligned}$$

□

Erreur quadratique moyenne de $\overline{y}_W$:

L'erreur quadratique moyenne de $\overline{y}_W$ est le réel :

$$EQM(\overline{y}_W)[PISR] = \mathbb{E}\left((\overline{y}_W-\overline{y}_U)^2\right) = \frac{1}{N^2}\sum_{i=2}^{N}\sum_{j=1}^{i-1}(\pi_i\pi_j-\pi_{i,j})\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2.$$

8.3 Estimations ponctuelles

Estimation ponctuelle de $\overline{y}_U$:

Soit ω un échantillon de n individus de U. Une estimation ponctuelle de $\overline{y}_U$ est la moyenne pondérée-échantillon :

$$\overline{y}_\omega = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\pi_i}\mathbb{1}_{\{u_i\in\omega\}}.$$

Quelques commandes R : Un exemple de calcul de $\overline{y}_\omega$ avec R est décrit ci-dessous :

```
U = c("Bob", "Nico", "Ali", "Fabien", "Malik", "John", "Jean", "Chris", "Karl")
y = c(72, 89, 68, 74, 81, 87, 76, 61, 84)
pi_i = c(0.2, 0.4, 0.6, 0.3, 0.4, 0.7, 0.2, 0.1, 0.6)
N = 9
n = 3
library(sampling)
t = srswor(n, 9)
bar_y_w = (1 / N) * sum(y * t/ pi_i)
bar_y_w
```

Cela renvoie 68.14815.

Estimation ponctuelle de l'écart-type de $\overline{y}_W$:

Soit ω un échantillon de n individus de U. Deux estimations ponctuelles différentes de l'écart-type de $\overline{y}_W$ sont données par :

- le réel :

$$s_1(\overline{y}_\omega) = \sqrt{\frac{1}{N^2}\left(\sum_{i=1}^{N}\frac{y_i^2}{\pi_i^2}\frac{\pi_i(1-\pi_i)}{\pi_i}\mathbb{1}_{\{u_i\in\omega\}} + \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\frac{(\pi_{i,j}-\pi_i\pi_j)}{\pi_{i,j}}\mathbb{1}_{\{(u_i,u_j)\in\omega\}}\right)}.$$

- le réel :

$$s_2(\overline{y}_\omega) = \sqrt{\frac{1}{N^2}\sum_{i=2}^{N}\sum_{j=1}^{i-1}\frac{(\pi_i\pi_j-\pi_{i,j})}{\pi_{i,j}}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2\mathbb{1}_{\{(u_i,u_j)\in\omega\}}}.$$

Celles-ci reposent sur les deux expressions de $\mathbb{V}(\overline{y}_W)$.

Intervalle de confiance pour $\overline{y}_U$:

Soit ω un échantillon de n individus de U. Un intervalle de confiance pour $\overline{y}_U$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, est

$$i_{\overline{y}_U} = \left[\overline{y}_\omega - z_\alpha s_1(\overline{y}_\omega),\ \overline{y}_\omega + z_\alpha s_1(\overline{y}_\omega)\right],$$

où z_α est le réel vérifiant $\mathbb{P}(|Z| \geq z_\alpha) = \alpha$, $Z \sim \mathcal{N}(0,1)$.

Un autre est $i_{\overline{y}_U} = [\overline{y}_\omega - z_\alpha s_2(\overline{y}_\omega),\ \overline{y}_\omega + z_\alpha s_2(\overline{y}_\omega)]$.

8.4 Cas particuliers

Plan de sondage aléatoire de type PESR :

Pour tout $i \in \{1,\dots,n\}$, on a

$$\pi_i = \mathbb{P}(u_i \in W) = \frac{n}{N}.$$

L'estimateur de Horvitz-Thompson devient :

$$\overline{y}_W = \frac{1}{N}\sum_{i=1}^{N} \frac{y_i}{\pi_i}\mathbb{1}_{\{u_i \in W\}} = \frac{1}{n}\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in W\}}.$$

On retrouve l'estimateur classique.

Plan de sondage aléatoire stratifié :

Pour tout $i \in \{1,\dots,n\}$, on a

$$\pi_i = \mathbb{P}(u_i \in W_h) = \frac{n_h}{N_h}\mathbb{1}_{\{u_i \in U_h\}}.$$

L'estimateur de Horvitz-Thompson devient :

$$\begin{aligned}\overline{y}_W &= \frac{1}{N}\sum_{i=1}^{N} \frac{y_i}{\pi_i}\mathbb{1}_{\{u_i \in W\}} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N} \frac{y_i}{\pi_i}\mathbb{1}_{\{u_i \in W_h\}} \\ &= \frac{1}{N}\sum_{h=1}^{H} N_h \frac{1}{n_h}\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in W_h\}} = \frac{1}{N}\sum_{h=1}^{H} N_h \overline{y}_{W_h}.\end{aligned}$$

On retrouve l'estimateur classique.

Plan de sondage aléatoire proportionnel à la taille :

Pour tout $i \in \{1, \dots, n\}$, on suppose l'existance d'un caractère secondaire X tel que sa valeur pour l'individu ω_i, notée x_i, est à peu près proportionnelle à y_i.

Pour tout $i \in \{1, \dots, n\}$, on suppose l'existance d'un réel α tel que

$$\pi_i = \mathbb{P}(u_i \in W) = \alpha x_i.$$

Comme, par définition, on a $\sum_{i=1}^{N} \pi_i = n$, il vient

$$\alpha = \frac{n}{\sum_{i=1}^{N} x_i}.$$

L'estimateur de Horvitz-Thompson devient :

$$\overline{y}_W = \left(\frac{1}{n} \sum_{j=1}^{N} x_i \right) \frac{1}{N} \sum_{i=1}^{N} \frac{y_i}{x_i} \mathbb{1}_{\{u_i \in W\}}.$$

En pratique : Comme on peut avoir $\pi_i = n x_i / \sum_{j=1}^{N} x_i > 1$ avec la méthode procédente, un ajustement doit être fait. On considère alors l'ensemble

$$A = \left\{ i \in \{1, \dots, N\};\ x_i > \frac{1}{n} \sum_{j=1}^{N} x_j \right\}$$

et $m = \text{Card}(A)$ et, à la place de π_i, on prend :

$$\pi_i^* = \begin{cases} 1 & \text{si } i \in A, \\ (n-m) \dfrac{x_i}{\sum_{j \in \{1,\dots,N\} - A} x_j} & \text{si } i \in U - A. \end{cases}$$

Quelques commandes R : Ces probabilités sont calculées avec les commandes R :

```
library(sampling)
a = 1:20
p = inclusionprobabilities(a, 12)
p
```

On peut comprendre la sortie de `p` en faisant :

```
a * 12 / sum(a)
p2 = NULL
p2[1:17] = (12 - 3) * a[1:17] / sum(a[1:17])
p2[18:20] = 1
p2
```

8.5 Sélection des individus

Plan de sondage aléatoire de Poisson :

Pour le mettre en œuvre, le plan de sondage aléatoire de Poisson,

- on considère n probabilités $\pi_1, \ldots, \pi_n$,
- on génère N nombres $x_1, \ldots, x_N$ (indépendamment des uns des autres) suivant la loi uniforme $\mathcal{U}([0,1])$,
- pour tout $i \in \{1, \ldots, N\}$, on sélectionne l'individu u_i s'il vérifie $x_i < \pi_i$,
- les individus sélectionnés constituent l'échantillon.

Remarques : On peut montrer que, pour tout $i \in \{1, \ldots, n\}$, $\pi_i = \mathbb{P}(u_i \in W) = \pi_i$.

Un inconvénient de cette méthode est que l'on ne sait pas a l'avance la taille n de l'échantillon sélectionné.

En revanche, la méthode est simple et rapide.

Sur le plan de la modélisation, on suppose que les *var* $\mathbb{1}_{\{u_1 \in W\}}, \ldots, \mathbb{1}_{\{u_n \in W\}}$ sont indépendantes. Ainsi, on a $\pi_{i,j} = \pi_i \pi_j$

$$\mathbb{P}(\omega \in W) = \prod_{k=1}^{N} \pi_k^{\mathbb{1}_{\{u_k \in \omega\}}} \prod_{k=1}^{N} (1 - \pi_k)^{\mathbb{1}_{\{u_k \notin \omega\}}}.$$

Quelques commandes R : Un exemple de commandes R sur le plan de sondage aléatoire de Poisson est décrit ci-dessous :

```
library(sampling)
pi_i = c(0.2, 0.7, 0.8, 0.5, 0.4, 0.4)
N = length(pi_i)
y = c(23.4, 5.64, 31.45, 25.4, 15.94, 21.45)
t = UPpoisson(pi_i)
(1:N)[t == 1]
bar_y_w = (1 / N) * sum((1 / pi_i[t == 1]) * y[t == 1])
bar_y_w
```

Cela renvoie 15.01875.

Plan de sondage aléatoire systématique à probabilités inégales :

Pour le mettre en œuvre, le plan de sondage aléatoire systématique à probabilités inégales,

- on considère N probabilités $\pi_1, \ldots, \pi_N$ et, pour tout $k \in \{1, \ldots, N\}$, on pose

$$C_k = \sum_{i=1}^{k} \pi_i, \qquad C_0 = 0,$$

- on génère un nombre x_1 suivant la loi uniforme $\mathcal{U}([0,1])$,
- pour tout $i \in \{1, \ldots, N\}$, on sélectionne l'individu u_i s'il vérifie : il existe un entier $j \in \{0, \ldots, n-1\}$ tel que

$$C_{i-1} \leq x_1 + j < C_i.$$

- les n individus sélectionnés constituent l'échantillon.

Remarques : On peut montrer que, pour tout $i \in \{1, \ldots, n\}$, $\pi_i = \mathbb{P}(u_i \in W) = \pi_i$.

Contrairement au plan de sondage aléatoire de Poisson, le plan de sondage aléatoire systématique à probabilités inégales est de taille fixe : n, pour l'échantillon.

Quelques commandes R : Un exemple de commandes R sur le plan de sondage aléatoire systématique à probabilités inégales avec un échantillon de $n = 3$ individus est décrit ci-dessous :

```
library(sampling)
pi_i = c(0.2, 0.7, 0.8, 0.5, 0.4, 0.4)
```

Remarquons que `sum(pi_i) = 3` $= n$.

```
N = length(pi_i)
y = c(23.4, 5.64, 31.45, 25.4, 15.94, 21.45)
t = UPsystematic(pi_i)
(1:N)[t == 1]
bar_y_w = (1 / N) * sum((1 / pi_i[t == 1]) * y[t == 1])
bar_y_w
```

Cela renvoie 34.51875.

8.6 Exercices corrigés

Exercice 1 : Dans une population de 3 individus $U = \{u_1, u_2, u_3\}$, on prélève au hasard et sans remise 2 individus pour former un échantillon. La *var* W égale à l'échantillon obtenu vérifie :

$$\mathbb{P}(W = \{u_1, u_2\}) = \frac{1}{4}, \qquad \mathbb{P}(W = \{u_1, u_3\}) = \frac{1}{4}, \qquad \mathbb{P}(W = \{u_2, u_3\}) = \frac{1}{2}.$$

On étudie un caractère Y dans U. Pour tout $i \in \{1, 2, 3\}$, soit y_i la valeur de Y pour l'individu u_i. Les résultats sont :

y_1	y_2	y_3
2	5	11

1. Calculer, pour tout $i \in \{1, 2, 3\}$, $\pi_i = \mathbb{P}(u_i \in W)$. Est-ce que l'on a affaire à un plan de sondage aléatoire de type PISR ?
2. On considère la *var* :
$$\overline{y}_W = \frac{1}{2} \sum_{i=1}^{3} y_i \mathbb{1}_{\{u_i \in W\}}.$$
   (a) Déterminer l'ensemble des valeurs possibles de $\overline{y}_W$, ainsi que sa loi.
   (b) Calculer la moyenne-population $\overline{y}_U$. Est-ce que $\mathbb{E}(\overline{y}_W) = \overline{y}_U$?

3. On considère la *var* :
$$\overline{y}_W^* = \frac{1}{3}\sum_{i=1}^{3}\frac{y_i}{\pi_i}\mathbb{1}_{\{u_i\in W\}}.$$
   (a) Déterminer l'ensemble des valeurs possibles de $\overline{y}_W^*$, ainsi que sa loi.
   (b) Est-ce que $\mathbb{E}(\overline{y}_W^*) = \overline{y}_U$?
   (c) Calculer $\mathbb{V}(\overline{y}_W^*)$.
4. (a) Calculer la matrice de variance-covariance du vecteur de *vars* $(\mathbb{1}_{\{u_1\in W\}}, \mathbb{1}_{\{u_2\in W\}}, \mathbb{1}_{\{u_3\in W\}})$.
   (b) Retrouver la valeur de $\mathbb{V}(\overline{y}_W^*)$ à l'aide de la matrice précédente et de la formule du cours.

Solution :

1. On a
$$\pi_1 = \mathbb{P}(u_1 \in W) = \sum_{j=1}^{3}\mathbb{P}(W = \{u_1, u_j\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2},$$
$$\pi_2 = \mathbb{P}(u_2 \in W) = \sum_{j=1}^{3}\mathbb{P}(W = \{u_2, u_j\}) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}$$
et
$$\pi_3 = \mathbb{P}(u_3 \in W) = \sum_{j=1}^{3}\mathbb{P}(W = \{u_3, u_j\}) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}.$$
Les probabilités d'inclusion du première ordre étant inégales et la sélection étant sans remise, on a bien affaire à un plan de sondage aléatoire de type PISR.

2. (a) Par l'énnoncé, les échantillons possibles sont :

$\{u_1, u_2\}$	$\{u_1, u_3\}$	$\{u_2, u_3\}$

Pour le premier échantillon (donc si $\omega = \{u_1, u_2\}$), il vient
$$\overline{y}_\omega = \frac{1}{2}\sum_{i=1}^{3} y_i\mathbb{1}_{\{u_i\in\omega\}} = \frac{1}{2}(2+5) = 3.5.$$
Celui-ci est btenu avec une probabilité $\mathbb{P}(W = \{u_1, u_2\}) = \frac{1}{4}$. En procédant de même, on complète le tableau suivant :

ω	$\mathbb{P}(W=\omega)$	Y	$\overline{y}_\omega$
$\{u_1, u_2\}$	$\frac{1}{4}$	$\{2, 5\}$	3.5
$\{u_1, u_3\}$	$\frac{1}{4}$	$\{2, 11\}$	6.5
$\{u_2, u_3\}$	$\frac{1}{2}$	$\{5, 11\}$	8

Ainsi, on a

$$\overline{y}_W(\Omega) = \{3.5, 6.5, 8\}.$$

La loi de $\overline{y}_W$ est donnée par

k	3.5	6.5	8
$\mathbb{P}(\overline{y}_W = k)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$

(b) La moyenne-population est

$$\overline{y}_U = \frac{1}{3}\sum_{i=1}^{3} y_i = 6.$$

Or on a

$$\mathbb{E}(\overline{y}_W) = \sum_{k\in\overline{y}_W(\Omega)} k\mathbb{P}(\overline{y}_W = k) = \frac{1}{4}\times 3.5 + \frac{1}{4}\times 6.5 + \frac{1}{2}\times 8 = 6.5.$$

On a donc $\mathbb{E}(\overline{y}_W) = 6.5 \neq 6 = \overline{y}_U$. Ainsi, l'estimateur $\overline{y}_W$ n'est pas sans biais pour $\overline{y}_U$.

3.

(a) Pour le premier échantillon (donc si $\omega = \{u_1, u_2\}$), il vient

$$\overline{y}^*_\omega = \frac{1}{3}\sum_{i=1}^{3} \frac{y_i}{\pi_i}\mathbb{1}_{\{u_i\in\omega\}} = \frac{1}{3}\left(\frac{2}{1/2} + \frac{5}{3/4}\right) = 3.555556.$$

Celui-ci est obtenu avec une probabilité de $\mathbb{P}(W = \{u_1, u_2\}) = \frac{1}{4}$. En procédant de même, on complète le tableau suivant :

ω	$\mathbb{P}(W = \omega)$	Y	π	$\overline{y}^*_\omega$
$\{u_1, u_2\}$	$\frac{1}{4}$	$\{2, 5\}$	$\{\frac{1}{2}, \frac{3}{4}\}$	3.555556
$\{u_1, u_3\}$	$\frac{1}{4}$	$\{2, 11\}$	$\{\frac{1}{2}, \frac{3}{4}\}$	6.222222
$\{u_2, u_3\}$	$\frac{1}{2}$	$\{5, 11\}$	$\{\frac{3}{4}, \frac{3}{4}\}$	7.111111

Ainsi, on a

$$\overline{y}^*_W(\Omega) = \{3.555556, 6.222222, 7.111111\}.$$

La loi de $\overline{y}^*_W$ est donnée par

k	3.555556	6.222222	7.111111
$\mathbb{P}(\overline{y}^*_W = k)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$

(b) En utilisant la loi de $\overline{y}_W^*$, l'espérance de $\overline{y}_W^*$ est

$$\mathbb{E}(\overline{y}_W^*) = \sum_{k \in \overline{y}_W^*(\Omega)} k\mathbb{P}(\overline{y}_W^* = k) = \frac{1}{4} \times 3.555556 + \frac{1}{4} \times 6.222222 + \frac{1}{2} \times 7.111111 = 6.$$

On a donc $\mathbb{E}(\overline{y}_W^*) = \overline{y}_U$; l'esimateur $\overline{y}_W^*$ est sans biais pour $\overline{y}_U$.

(c) En utilisant la formule de König-Huyghens, la variance de $\overline{y}_W^*$ est

$$\mathbb{V}(\overline{y}_W^*) = \mathbb{E}\left((\overline{y}_W^*)^2\right) - \left(\mathbb{E}(\overline{y}_W^*)\right)^2.$$

Or on a $\mathbb{E}(\overline{y}_W^*) = 6$ et

$$\begin{aligned}\mathbb{E}\left((\overline{y}_W^*)^2\right) &= \sum_{k \in \overline{y}_W^*(\Omega)} k^2\mathbb{P}(\overline{y}_W^* = k) = \frac{1}{4} \times 3.555556^2 + \frac{1}{4} \times 6.222222^2 + \frac{1}{2} \times 7.111111^2 \\ &= 38.12346.\end{aligned}$$

D'où

$$\mathbb{V}(\overline{y}_W) = 38.12346 - 6^2 = 2.123456.$$

4. (a) La matrice de variance-covariance du vecteur de *vars* $(\mathbb{1}_{\{u_1 \in W\}}, \mathbb{1}_{\{u_2 \in W\}}, \mathbb{1}_{\{u_3 \in W\}})$ est

$$\mathbb{C}_{ov} = \begin{pmatrix} \mathbb{C}\left(\mathbb{1}_{\{u_1 \in W\}}, \mathbb{1}_{\{u_1 \in W\}}\right) & \mathbb{C}\left(\mathbb{1}_{\{u_1 \in W\}}, \mathbb{1}_{\{u_2 \in W\}}\right) & \mathbb{C}\left(\mathbb{1}_{\{u_1 \in W\}}, \mathbb{1}_{\{u_3 \in W\}}\right) \\ \mathbb{C}\left(\mathbb{1}_{\{u_2 \in W\}}, \mathbb{1}_{\{u_1 \in W\}}\right) & \mathbb{C}\left(\mathbb{1}_{\{u_2 \in W\}}, \mathbb{1}_{\{u_2 \in W\}}\right) & \mathbb{C}\left(\mathbb{1}_{\{u_2 \in W\}}, \mathbb{1}_{\{u_3 \in W\}}\right) \\ \mathbb{C}\left(\mathbb{1}_{\{u_3 \in W\}}, \mathbb{1}_{\{u_1 \in W\}}\right) & \mathbb{C}\left(\mathbb{1}_{\{u_3 \in W\}}, \mathbb{1}_{\{u_2 \in W\}}\right) & \mathbb{C}\left(\mathbb{1}_{\{u_3 \in W\}}, \mathbb{1}_{\{u_3 \in W\}}\right) \end{pmatrix},$$

avec, pour exemples de calcul :

$$\begin{aligned}\mathbb{C}\left(\mathbb{1}_{\{u_1 \in W\}}, \mathbb{1}_{\{u_1 \in W\}}\right) &= \mathbb{V}\left(\mathbb{1}_{\{u_1 \in W\}}\right) = \mathbb{E}\left(\mathbb{1}_{\{u_1 \in W\}}^2\right) - \left(\mathbb{E}\left(\mathbb{1}_{\{u_1 \in W\}}\right)\right)^2 \\ &= \mathbb{E}\left(\mathbb{1}_{\{u_1 \in W\}}\right) - \left(\mathbb{E}\left(\mathbb{1}_{\{u_1 \in W\}}\right)\right)^2 = \pi_1 - \pi_1^2 \\ &= \frac{1}{2} - \frac{1}{4} = \frac{1}{4}\end{aligned}$$

et

$$\begin{aligned}\mathbb{C}\left(\mathbb{1}_{\{u_1 \in W\}}, \mathbb{1}_{\{u_2 \in W\}}\right) &= \mathbb{E}\left(\mathbb{1}_{\{u_1 \in W\}}\mathbb{1}_{\{u_2 \in W\}}\right) - \mathbb{E}\left(\mathbb{1}_{\{u_1 \in W\}}\right)\mathbb{E}\left(\mathbb{1}_{\{u_2 \in W\}}\right) \\ &= \mathbb{P}(W = \{u_1, u_2\}) - \pi_1\pi_2 = \frac{1}{4} - \frac{1}{2} \times \frac{3}{4} = -\frac{1}{8}.\end{aligned}$$

En procédant ainsi (et en utilisant la symmétrie pour compléter la matrice plus rapidement), on obtient la matrice de variance-covariance :

$$\begin{pmatrix} \frac{1}{4} & -\frac{1}{8} & -\frac{1}{8} \\ -\frac{1}{8} & \frac{3}{16} & -\frac{1}{16} \\ -\frac{1}{8} & -\frac{1}{16} & \frac{3}{16} \end{pmatrix}.$$

(b) Par une formule du cours, on peut écrire :

$$\mathbb{V}(\overline{y}_W^*) = \frac{1}{3^2}\left(\sum_{i=1}^{3}\frac{y_i^2}{\pi_i^2}\pi_i(1-\pi_i) + \sum_{i=1}^{3}\sum_{\substack{j=1\\ j\neq i}}^{3}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}(\pi_{i,j}-\pi_i\pi_j)\right),$$

avec $\pi_{i,j} = \mathbb{P}(W = \{u_i, u_j\})$. Les éléments $\pi_i(1-\pi_i)$ et $(\pi_{i,j} - \pi_i\pi_j)$ sont déjà calculés ; ce sont les composantes de la matrice de variance-covariance ; les éléments de la forme $\pi_i(1-\pi_i)$ correspondants aux variances, et ceux de la forme $(\pi_{i,j} - \pi_i\pi_j)$ correspondants aux covariances. Dès lors, on a

$$\begin{aligned}\mathbb{V}(\overline{y}_W^*) = \frac{1}{3^2}\Bigg(&\frac{2^2}{(1/2)^2}\times\frac{1}{4} + \frac{5^2}{(3/4)^2}\times\frac{3}{16} + \frac{11^2}{(3/4)^2}\times\frac{3}{16} + 2\times\frac{2}{1/2}\times\frac{5}{3/4}\times\left(-\frac{1}{8}\right)\\ &+ 2\times\frac{2}{1/2}\times\frac{11}{3/4}\times\left(-\frac{1}{8}\right) + 2\times\frac{5}{3/4}\times\frac{11}{3/4}\times\left(-\frac{1}{16}\right)\Bigg)\\ &= 2.123456.\end{aligned}$$

On retrouve bien le même résultat.

Exercice 2 : Dans une population de 6 individus $U = \{u_1, u_2, u_3, u_4, u_5, u_6\}$, on prélève au hasard et sans remise 4 individus pour former un échantillon. Est-ce que la *var* W égale à l'échantillon obtenu peut vérifier les conditions suivantes :

$$\mathbb{P}(\{u_1,u_2\}\in W) = 1, \qquad \mathbb{P}(\{u_3,u_4\}\in W) = \frac{2}{3}, \qquad \mathbb{P}(\{u_3,u_5\}\in W) = \frac{1}{6}, \qquad \mathbb{P}(\{u_4,u_6\}\in W) = \frac{1}{6}$$

et toutes les autres paires d'individus $\{u_i, u_j\}$ non-indiquées précédemment vérifient $\mathbb{P}(\{u_i,u_j\}\in W) = 0$?

Solution : La réponse est Non. D'une part, le fait que $\mathbb{P}(\{u_1,u_2\}\in W) = 1$ implique que u_1 et u_2 sont nécessairement dans l'échantillon. D'autre part, le fait que toutes les paires d'individus $\{u_i, u_j\}$ non-

indiquées vérifient $\mathbb{P}(\{u_i, u_j\} \in W) = 0$ implique, entre autres, que

$$\mathbb{P}(\{u_1, u_3\} \in W) = 0, \quad \mathbb{P}(\{u_1, u_4\} \in W) = 0 \quad \mathbb{P}(\{u_1, u_5\} \in W) = 0 \quad \mathbb{P}(\{u_1, u_6\} \in W) = 0.$$

Ainsi, la présence imposée de u_1 dans l'échantillon entraîne l'impossibilité pour u_3, u_4, u_5 et u_6 de faire partie de cet échantillon. On ne peut donc pas constituer un échantillon de 4 individus avec une telle *var* W.

Exercice 3 : Dans une population de 6 individus $U = \{u_1, u_2, u_3, u_4, u_5, u_6\}$, on prélève au hasard et sans remise 4 individus pour former un échantillon. La *var* W égale à l'échantillon obtenu vérifie :

$$\mathbb{P}(W = \{u_1, u_2, u_3, u_4\}) = \frac{2}{3}, \qquad \mathbb{P}(W = \{u_1, u_2, u_3, u_5\}) = \frac{1}{6}, \qquad \mathbb{P}(W = \{u_1, u_2, u_4, u_6\}) = \frac{1}{6}.$$

Tous les autres échantillons d'individus $\{u_i, u_j, u_k, u_\ell\}$ non-indiquées précédemment vérifient $\mathbb{P}(W = \{u_i, u_j, u_k, u_\ell\}) = 0$.

1. Calculer, pour tout $i \in \{1, 2, 3, 4, 5, 6\}$,

$$\pi_i = \mathbb{P}(u_i \in W).$$

   Est-ce que l'on a affaire à un plan de sondage aléatoire de type PESR ?

2. On étudie un caractère Y dans U. Pour tout $i \in \{1, 2, 3, 4, 5, 6\}$, soit y_i la valeur de Y pour l'individu u_i. Les résultats sont :

y_1	y_2	y_3	y_4	y_5	y_6
75	51	34	22	12	8

   On considère la *var* :

$$\overline{y}_W = \frac{1}{6} \sum_{i=1}^{6} \frac{y_i}{\pi_i} \mathbb{1}_{\{u_i \in W\}}.$$

   (a) Déterminer la loi de $\overline{y}_W$.

   (b) Calculer la moyenne-population $\overline{y}_U$. Est-ce que $\mathbb{E}(\overline{y}_W) = \overline{y}_U$?

   (c) Calculer $\mathbb{V}(\overline{y}_W)$.

Solution :

1. On a

$$\begin{aligned}\pi_1 = \mathbb{P}(u_1 \in W) &= \mathbb{P}(W = \{u_1, u_2, u_3, u_4\}) + \mathbb{P}(W = \{u_1, u_2, u_3, u_5\}) + \mathbb{P}(W = \{u_1, u_2, u_4, u_6\}) \\ &= \frac{2}{3} + \frac{1}{6} + \frac{1}{6} = 1,\end{aligned}$$

$$\begin{aligned}\pi_2 = \mathbb{P}(u_2 \in W) &= \mathbb{P}(W = \{u_1, u_2, u_3, u_4\}) + \mathbb{P}(W = \{u_1, u_2, u_3, u_5\}) + \mathbb{P}(W = \{u_1, u_2, u_4, u_6\}) \\ &= \frac{2}{3} + \frac{1}{6} + \frac{1}{6} = 1,\end{aligned}$$

$$\pi_3 = \mathbb{P}(u_3 \in W) = \mathbb{P}(W = \{u_1, u_2, u_3, u_4\}) + \mathbb{P}(W = \{u_1, u_2, u_3, u_5\}) = \frac{2}{3} + \frac{1}{6} = \frac{5}{6},$$

$$\pi_4 = \mathbb{P}(u_4 \in W) = \mathbb{P}(W = \{u_1, u_2, u_3, u_4\}) + \mathbb{P}(W = \{u_1, u_2, u_4, u_6\}) = \frac{2}{3} + \frac{1}{6} = \frac{5}{6},$$

$$\pi_5 = \mathbb{P}(u_5 \in W) = \mathbb{P}(W = \{u_1, u_2, u_3, u_5\}) = \frac{1}{6}$$

et

$$\pi_6 = \mathbb{P}(u_6 \in W) = \mathbb{P}(W = \{u_1, u_2, u_4, u_6\}) = \frac{1}{6}.$$

Ainsi, on a

π_1	π_2	π_3	π_4	π_5	π_6
1	1	$\frac{5}{6}$	$\frac{5}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Comme ces probabilités diffèrent, le plan de sondage n'est pas à probabilités égales (PE), donc il n'est pas PESR. C'est un plan de sondage PISR.

2.

(a) Par l'énoncé, les échantillons possibles sont :

$\{u_1, u_2, u_3, u_4\}$	$\{u_1, u_2, u_3, u_5\}$	$\{u_1, u_2, u_4, u_6\}$

Pour le premier échantillon (donc si $\omega = \{u_1, u_2, u_3, u_4\}$), il vient

$$\overline{y}_\omega = \frac{1}{6}\sum_{i=1}^{6} \frac{y_i}{\pi_i} \mathbb{1}_{\{u_i \in \omega\}} = \frac{1}{6}\left(\frac{75}{1} + \frac{51}{1} + \frac{34}{5/6} + \frac{22}{5/6}\right) = 32.2.$$

Celui-ci est obtenu avec une probabilité de $\mathbb{P}(W = \{u_1, u_2, u_3, u_4\}) = \frac{2}{3}$. En procédant de même, on complète le tableau suivant :

ω	$\mathbb{P}(W=\omega)$	Y	π	$\overline{y}_\omega$
$\{u_1,u_2,u_3,u_4\}$	$\frac{2}{3}$	$\{75,51,34,22\}$	$\{1,1,\frac{5}{6},\frac{5}{6}\}$	32.2
$\{u_1,u_2,u_3,u_5\}$	$\frac{1}{6}$	$\{75,51,34,12\}$	$\{1,1,\frac{5}{6},\frac{1}{6}\}$	39.8
$\{u_1,u_2,u_4,u_6\}$	$\frac{1}{6}$	$\{75,51,22,8\}$	$\{1,1,\frac{5}{6},\frac{1}{6}\}$	33.4

Ainsi, on a

$$\overline{y}_W(\Omega)=\{32.2,33.4,39.8\}.$$

La loi de $\overline{y}_W$ est donnée par

k	32.2	33.4	39.8
$\mathbb{P}(\overline{y}_W=k)$	$\frac{2}{3}$	$\frac{1}{6}$	$\frac{1}{6}$

(b) La moyenne-population est

$$\overline{y}_U=33.66667.$$

En utilisant la loi de $\overline{y}_W$, l'espérance de $\overline{y}_W$ est

$$\mathbb{E}(\overline{y}_W)=\sum_{k\in\overline{y}_W(\Omega)}k\mathbb{P}(\overline{y}_W=k)=\frac{2}{3}\times 32.2+\frac{1}{6}\times 33.4+\frac{1}{6}\times 39.8=33.66667.$$

On a donc $\mathbb{E}(\overline{y}_W)=\overline{y}_U$; l'esimateur $\overline{y}_W$ est sans biais pour $\overline{y}_U$.

(c) En utilisant la formule de König-Huyghens, la variance de $\overline{y}_W$ est

$$\mathbb{V}(\overline{y}_W)=\mathbb{E}(\overline{y}_W^2)-\left(\mathbb{E}(\overline{y}_W)\right)^2.$$

Or on a $\mathbb{E}(\overline{y}_W)=33.66667$ et

$$\mathbb{E}(\overline{y}_W^2)=\sum_{k\in\overline{y}_W(\Omega)}k^2\mathbb{P}(\overline{y}_W=k)=\frac{2}{3}\times 32.2^2+\frac{1}{6}\times 33.4^2+\frac{1}{6}\times 39.8^2=1141.16.$$

D'où

$$\mathbb{V}(\overline{y}_W)=1141.16-33.66667^2=7.715331.$$

9 Plan de sondage aléatoire par grappe (G)

9.1 Contexte

Idée : On suppose que l'on a affaire à une population homogène, laquelle est répartie en de nombreux groupes homogènes a priori. Ces groupes peuvent être naturellement formés, pour des raisons géographiques, par exemple. Pour gagner du temps et de l'argent, l'idée du plan de sondage par grappe est de faire un PESR qui portent sur ces groupes (et non sur les individus directement) et de considérer la totalité des individus de ceux-ci pour former l'échantillon.

Groupe :

On considère une partition de M groupes (éléments) de U notée $(G_1, \ldots, G_M)$. Ainsi, on a $U = \bigcup_{j=1}^{M} G_j$ et, pour tout $(j,k) \in \{1,\ldots,M\}^2$ avec $j \neq k$, on a $G_j \cap G_k = \emptyset$.
On appelle groupe un élément G_j de $(G_1, \ldots, G_M)$.

Plan de sondage aléatoire par grappe (G) :

On sélectionne au hazard et sans remise m groupes parmi les M groupes, puis on prend tous les individus des groupes sélectionnés pour formé un échantillon d'individus. Ainsi, l'échantillon obtenu peut s'écrire sous la forme :

$$\omega = (\omega_1, \ldots, \omega_m),$$

où, pour tout $j \in \{1,\ldots,m\}$, ω_j est un échantillon d'individus contenant tous les individus du j-ième groupe sélectionné.

La taille de ω peut être encore noté n, mais il faut remarquer que ce n est obtenu après sélection des groupes. On peut donc avoir une idée de son ordre de grandeur dès le début, mais on ne peut pas le fixer précisément dès le début du processus de sélection.

Quelques commandes R : Pour faire un plan de sondage aléatoire par grappe, on peut utiliser la fonction `cluster` de la librarie `sampling`. Pour un exemple de commandes R, on peut utiliser le jeu de données `swissmunicipalities` de la librairie `sampling`. Dans ce jeu de données, il y a un caractère qualitatif `REG` qui divisent la population en $M = 7$ groupes. On souhaite faire un plan de sondage aléatoire G avec $m = 7$.

Les commandes sont décrite ci-dessous :

```
library(sampling)
data(swissmunicipalities)
cl = cluster(swissmunicipalities, clustername = c("REG"), size = 3, method =
"srswor")
getdata(swissmunicipalities, cl)$Surfacescult
```

Probabilités d'appartenance de groupe :

- pour tout $j \in \{1, \dots, M\}$, la probabilité que G_j appartienne à W est

$$\mathbb{P}(G_j \in W) = \frac{m}{M}.$$

- pour tout $(j, k) \in \{1, \dots, M\}^2$ avec $j \neq k$, la probabilité que G_j et G_k appartiennent à W est

$$\mathbb{P}((G_j, G_k) \in W) = \frac{m(m-1)}{M(M-1)}.$$

La preuve est la même que pour les probabilités d'appartenance des individus dans le cadre PESR ; il suffit de remplacer u_i par G_j, n par m et N par M.

9.2 Estimateurs

Estimation aléatoire de $\overline{y}_U$:

Un estimateur aléatoire de $\overline{y}_U$ est

$$\overline{y}_W = \frac{M}{mN} \sum_{j=1}^{M} t_j \mathbb{1}_{\{G_j \in W\}},$$

où

$$t_j = \sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in G_j\}}.$$

Espérance de $\overline{y}_{W_h}$:

On a

$$\mathbb{E}(\overline{y}_W) = \overline{y}_U.$$

Preuve : Comme $\mathbb{P}(G_j \in W) = m/M$ et $\sum_{j=1}^{M} \mathbb{1}_{\{u_i \in G_j\}} = 1$, il vient

$$\mathbb{E}(\overline{y}_W) = \mathbb{E}\left(\frac{M}{mN}\sum_{j=1}^{M} t_j \mathbb{1}_{\{G_j \in W\}}\right) = \frac{M}{mN}\sum_{j=1}^{M} t_j \mathbb{E}\left(\mathbb{1}_{\{G_j \in W\}}\right) = \frac{M}{mN}\sum_{j=1}^{M} t_j \mathbb{P}(G_j \in W)$$
$$= \frac{M}{mN}\sum_{j=1}^{M} t_j \frac{m}{N} = \frac{1}{N}\sum_{i=1}^{N} y_i \sum_{j=1}^{M} \mathbb{1}_{\{u_i \in G_j\}} = \frac{1}{N}\sum_{i=1}^{N} y_i = \overline{y}_U.$$

□

Variance de $\overline{y}_W$:

On a

$$\mathbb{V}(\overline{y}_W) = \frac{M^2}{N^2}\left(1 - \frac{m}{M}\right)\frac{1}{m}\Xi_U^2,$$

où

$$\Xi_U^2 = \frac{1}{M-1}\sum_{j=1}^{M}\left(t_j - \frac{1}{M}\sum_{k=1}^{M} t_k\right)^2.$$

On peut remarquer que Ξ_U^2 est la variance-corrigée population associée aux valeurs $t_1, \ldots, t_M$.

Preuve : On a

$$\mathbb{V}(\overline{y}_W) = \mathbb{V}\left(\frac{M}{mN}\sum_{j=1}^{M} t_j \mathbb{1}_{\{G_j \in W\}}\right) = \frac{M^2}{N^2}\mathbb{V}\left(\frac{1}{m}\sum_{j=1}^{M} t_j \mathbb{1}_{\{G_j \in W\}}\right).$$

En procédant comme pour la variance de l'estimateur de la moyenne population pour un plan de sondage PESR (en remplaçant $y_1, \ldots, y_N$ par $t_1, \ldots, t_M$, u_i par G_j, n par m et N par M), il vient

$$\begin{aligned}
\mathbb{V}\left(\frac{1}{m}\sum_{j=1}^{M} t_j \mathbb{1}_{\{G_j \in W\}}\right) &= \left(1 - \frac{m}{M}\right)\frac{1}{m}\frac{1}{M-1}\left(\sum_{j=1}^{M} t_j^2 - \frac{1}{M}\left(\sum_{j=1}^{M} t_j\right)^2\right) \\
&= \left(1 - \frac{m}{M}\right)\frac{1}{m}\frac{1}{M-1}\sum_{j=1}^{M}\left(t_j - \frac{1}{M}\sum_{k=1}^{M} t_k\right)^2 \\
&= \left(1 - \frac{m}{M}\right)\frac{1}{m}\frac{1}{M-1}\sum_{j=1}^{M}\Xi_U^2.
\end{aligned}$$

D'où

$$\mathbb{V}(\overline{y}_W) = \frac{M^2}{N^2}\left(1 - \frac{m}{M}\right)\frac{1}{m}\Xi_U^2.$$

□

Remarque : On peut aussi écrire

$$\Xi_U^2 = \frac{1}{M-1}\sum_{j=1}^{M}\left(t_j - \frac{N}{M}\overline{y}_U\right)^2.$$

Erreur quadratique moyenne de $\overline{y}_W$:

L'erreur quadratique moyenne de $\overline{y}_W$ est le réel :

$$EQM(\overline{y}_W)[G] = \mathbb{E}\left((\overline{y}_W - \overline{y}_U)^2\right) = \frac{M^2}{N^2}\left(1 - \frac{m}{M}\right)\frac{1}{m}\Xi_U^2.$$

Estimation aléatoire de Ξ_U :

Un estimateur aléatoire de Ξ_U est

$$\Xi_W = \sqrt{\frac{1}{m-1}\sum_{j=1}^{M}\left(t_j - \frac{1}{m}\sum_{k=1}^{M} t_k \mathbb{1}_{\{G_k \in W\}}\right)^2 \mathbb{1}_{\{G_j \in W\}}}.$$

Propriété de Ξ_W^2 :

L'estimateur Ξ_W^2 est sans biais pour Ξ_U^2 :

$$\mathbb{E}\left(\Xi_W^2\right) = \Xi_U^2.$$

Preuve : La preuve est identique à celle de $\mathbb{E}(s_W^2) = s_U^2$ dans le cadre PESR (en remplaçant $y_1, \ldots, y_N$ par $t_1, \ldots, t_M$, u_i par G_j, n par m et N par M).

□

9.3 Estimations ponctuelles

Estimation ponctuelle de $\overline{y}_U$:

Une estimation ponctuelle de $\overline{y}_U$ est

$$\overline{y}_\omega = \frac{M}{mN}\sum_{j=1}^{M} t_j \mathbb{1}_{\{G_j \in \omega\}},$$

où

$$t_j = \sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in G_j\}}.$$

Estimation ponctuelle de Ξ_U :

Soit ω un échantillon de m groupes de U. Une estimation ponctuelle de Ξ_U est l'écart-type corrigé échantillon associée aux valeurs $t_1, \ldots, t_M$:

$$\Xi_\omega = \sqrt{\frac{1}{m-1}\sum_{j=1}^{M}\left(t_j - \frac{1}{m}\sum_{k=1}^{M} t_k \mathbb{1}_{\{G_k \in \omega\}}\right)^2 \mathbb{1}_{\{G_j \in \omega\}}}.$$

Remarques :

- On peut écrire Ξ_ω comme

$$\Xi_\omega = \sqrt{\frac{1}{m-1}\sum_{j=1}^{M}\left(t_j - \frac{N}{M}\overline{y}_\omega\right)^2 \mathbb{1}_{\{G_j \in \omega\}}}.$$

- En posant $T_1 = \sum_{j=1}^{M} t_j^2 \mathbb{1}_{\{G_j \in \omega\}}$ et $T_2 = \sum_{j=1}^{M} t_j \mathbb{1}_{\{G_j \in \omega\}}$, on a aussi

$$\Xi_\omega = \sqrt{\frac{1}{m-1}\left(T_1 - \frac{1}{m}T_2^2\right)}.$$

Estimation ponctuelle de l'écart-type de $\overline{y}_W$:

Soit ω un échantillon de n individus de U. Une estimation ponctuelle de l'écart-type de $\overline{y}_W$ est le réel :

$$s(\overline{y}_\omega) = \sqrt{\frac{M^2}{N^2}\left(1 - \frac{m}{M}\right)\frac{1}{m}\Xi_\omega^2}.$$

Quelques commandes R : Un exemple de fonction R pour calculer $\overline{y}_w$, Ξ_w et $s(\overline{y}_w)$ est décrit ci-dessous :

```
m = 2
M = 3
N = 50
y_1 = c(12.2, 5.4, 7.9, 9.1, 10.2, 11.7, 12.3)
y_2 = c(9.8, 10.2, 8.9, 10.1, 11.1, 12.1, 12.1)
y_3 = c(10.9, 7.1, 8.8, 12.1, 13.1, 9.8, 2.6)
tot = c(sum(y_1), sum(y_2), sum(y_3))
library(sampling)
t = srswor(m, M)
bar_y_w = (M / (m * N)) * sum(tot * t)
Xi_w = sqrt(sum ((tot - (N / M) * bar_y_w)^2 * t) / (m - 1))
s_bar_y_w = sqrt((M^2 / N^2) * (1 - m / M) * (1 / m) * Xi_w^2)
bar_y_w; Xi_w; s_bar_y_w
```

Cela renvoie 4.161, 7.000357 and 0.171473. Donc on a $\overline{y}_w = 4.161$, $\Xi_w = 7.000357$ et $s(\overline{y}_w) = 0.171473$. Cela peut changer à chaque expérience, puisque la sélection des groupes est aléatoire.

9.4 Intervalles de confiance

Résultat limite : Si m, M et $M - m$ sont suffisamment grands, alors on a

$$Z = \frac{\overline{y}_W - \overline{y}_U}{\sqrt{\frac{M^2}{N^2}\left(1 - \frac{m}{M}\right)\frac{1}{m}\Xi_W^2}} \approx \mathcal{N}(0,1).$$

Intervalle de confiance pour $\overline{y}_U$:

Soit ω un échantillon de m groupes de U. Un intervalle de confiance pour $\overline{y}_U$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, est

$$\begin{aligned} i_{\overline{y}_U} &= [\overline{y}_\omega - z_\alpha s(\overline{y}_\omega),\ \overline{y}_\omega + z_\alpha s(\overline{y}_\omega)] \\ &= \left[\overline{y}_\omega - z_\alpha\sqrt{\frac{M^2}{N^2}\left(1 - \frac{m}{M}\right)\frac{1}{m}\Xi_\omega^2},\ \overline{y}_\omega + z_\alpha\sqrt{\frac{M^2}{N^2}\left(1 - \frac{m}{M}\right)\frac{1}{m}\Xi_\omega^2}\right], \end{aligned}$$

où z_α est le réel vérifiant $\mathbb{P}(|Z| \geq z_\alpha) = \alpha$, $Z \sim \mathcal{N}(0,1)$.

Il y a $100(1-\alpha)$ chances sur 100 que $\overline{y}_U$ appartienne à l'intervalle $i_{\overline{y}_U}$.

Quelques commandes R : Un exemple de fonction R pour calculer l'intervalle de confiance pour $\overline{y}_U$ au niveau $100(1-\alpha)\%$ est décrit ci-dessous :

```
icG = function(tot, m, M, N, niveau) {
bar_y_w = (M / (m * N)) * sum(tot)
z = qnorm(1 - (1 - niveau) / 2)
Xi2_w = var(tot)
var_bar_y_w = (M^2 / N^2) * (1 - m / M) * (1 / m) * Xi2_w
a = bar_y_w - z * sqrt(var_bar_y_w)
b = bar_y_w + z * sqrt(var_bar_y_w)
print(c(a, b)) }
icG(tot = c(4.4, 4.3, 5.3), m = 3, M = 5, N = 12, niveau = 0.95)
```

Cela renvoie : 1.780209, 2.108680.

9.5 Taille de groupe

Incertitude absolue :

Soit ω un échantillon de m groupes de U. On appelle incertitude absolue sur $\overline{y}_U$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, la demi-longueur de $i_{\overline{y}_U}$:

$$d_\omega = z_\alpha s(\overline{y}_\omega) = z_\alpha \sqrt{\frac{M^2}{N^2}\left(1-\frac{m}{M}\right)\frac{1}{m}\Xi_\omega^2}.$$

Plus d_ω est petit, plus l'estimation de $\overline{y}_U$ par $\overline{y}_\omega$ est précise.

Incertitude relative :

Soit ω un échantillon de m groupes de U et d_ω l'incertitude absolue sur $\overline{y}_U$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$. On appelle incertitude relative sur $\overline{y}_U$ au niveau $100(1-\alpha)\%$ le pourcentage $(100 \times d_\omega^*)\%$ où d_ω^* est le réel :

$$d_\omega^* = \frac{d_\omega}{\overline{y}_\omega}.$$

Taille de groupe :

Soit ω un échantillon prélevé lors d'une étude préliminaire. La taille de groupe m à choisir pour avoir :

- une incertitude absolue sur $\overline{y}_U$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, inférieure ou égale à d_0 est le plus petit m tel que
$$d_\omega \leq d_0 \qquad \Leftrightarrow \qquad m \geq \frac{z_\alpha^2 M^2 \Xi_\omega^2}{N^2 d_0^2 + z_\alpha^2 M \Xi_\omega^2},$$
- une incertitude relative sur $\overline{y}_U$ au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, inférieure ou égale à $(100\times d_1)\%$ est le plus petit m tel que
$$d_\omega^* \leq d_1 \qquad \Leftrightarrow \qquad m \geq \frac{z_\alpha^2 M^2 \Xi_\omega^2}{N^2 (\overline{y}_\omega d_1)^2 + z_\alpha^2 M \Xi_\omega^2}.$$

9.6 Exercices corrigés

Exercice 1 : L'objectif est d'estimer le revenu moyen des ménages dans un arrondissement d'une grande ville composée de 60 îlots de maisons (un îlot est un "pâté de maisons", de taille variable). Pour cela, on sélectionne 3 îlots par un plan de sondage PESR et on interroge tous les ménages qui y résident. On sait, en outre, que 5000 menages resident dans cet arrondissement. Le résultats sont les suivants :

— Numéro de l'îlot : 1. Revenu total des ménages : 2100 euros.

— Numéro de l'îlot : 2. Revenu total des ménages : 2000 euros.

— Numéro de l'îlot : 3. Revenu total des ménages : 1500 euros.

1. Quel est le plan de sondage considéré ?
2. Donner une estimation ponctuelle du revenu moyens (population) des ménages de l'arrondissement.
3. Donner un intervalle de confiance pour le revenu moyens (population) des ménages de l'arrondissement au niveau 95%.

Solution :

1. Il s'agit d'un plan de sondage aléatoire par grappe ; une fois les îlots sélectionnés suivant un plan de sondage PESR, on considère tous les ménages de ces îlots.
2. Dans le cadre d'un plan de sondage aléatoire par grappe, une estimation ponctuelle de la moyenne-

population est

$$\overline{y}_\omega = \frac{M}{mN}\sum_{j=1}^{M} t_j \mathbb{1}_{\{G_j \in \omega\}},$$

où

$$t_j = \sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in G_j\}}.$$

Ici, on a $M = 60$, $m = 3$ et $N = 5000$, $t_1 = 2100$, $t_2 = 2000$ et $t_3 = 1500$. Par conséquent, une estimation ponctuelle du revenu moyens (population) des ménages de l'arrondissement est donnée par

$$\overline{y}_\omega = \frac{60}{3 \times 5000}(2100 + 2000 + 1500) = 22.4.$$

Ainsi, cette estimation est de 22.4 euros.

3. Dans le cadre d'un plan de sondage aléatoire par grappe, une estimation ponctuelle de l'écart-type de l'estimateur de la moyenne-population est le réel :

$$s(\overline{y}_\omega) = \sqrt{\frac{M^2}{N^2}\left(1 - \frac{m}{M}\right)\frac{1}{m}\Xi_\omega^2},$$

où Ξ_ω est l'écart-type corrigé échantillon associée aux valeurs $t_1, \ldots, t_M$. L'écart-type corrigé des valeurs 2100, 2000 et 1500 est $\Xi_\omega = 321.455$. Dès lors, une estimation ponctuelle de l'écart-type de l'estimateur du revenu moyens (population) des ménages de l'arrondissement est

$$s(\overline{y}_\omega) = \sqrt{\frac{60^2}{5000^2}\left(1 - \frac{3}{60}\right)\frac{1}{3}321.455^2} = 2.170715.$$

D'autre part, on a $95\% = 100(1-\alpha)\%$ avec $\alpha = 0.05$. On a $\mathbb{P}(|Z| \geq z_\alpha) = \alpha = 0.05$, $Z \sim \mathcal{N}(0,1)$, avec $z_\alpha = 1.96$. Donc l'intervalle de confiance recherché est

$$\begin{aligned} i_{\overline{y}_U} &= [\overline{y}_\omega - z_\alpha s(\overline{y}_\omega),\ \overline{y}_\omega + z_\alpha s(\overline{y}_\omega)] \\ &= [22.4 - 1.96 \times 2.170715,\ 22.4 + 1.96 \times 2.170715] \\ &= [18.1454,\ 26.6546]. \end{aligned}$$

Ainsi, il y a 95 chances sur 100 que [18.1454, 26.6546] contienne le revenu moyens (population) des ménages de l'arrondissement, l'unité étant l'euro.

Exercice 2 : Dans une ville, une mairie fait une enquête sur le bien-être de ses habitants. Sur $N = 20000$ ménages répartis en $M = 400$ quartiers, elle sélectionne $m = 80$ quartiers par un plan de sondage PESR. Pour chaque ménage des quartiers sélectionnés, on demande de noter entre 0 et 10 le niveau de bien-être dans la ville. On a observé, sur les m quartiers sélectionnées,

$$\sum_{j=1}^{m} t_j = 29800, \qquad \sum_{j=1}^{m} t_j^2 = 58804000,$$

où t_j désigne la somme des notes des ménages du j-ème quartier.

1. Expliquer en une phrase l'information : $\sum_{j=1}^{m} t_j = 29800$.
2. Donner une estimation ponctuelle de la note moyenne d'un ménage.
3. Déterminer un intervalle de confiance au niveau 95% pour la note moyenne d'un ménage.

Solution :

1. L'information : $\sum_{j=1}^{m} t_j = 29800$ indique que la somme des notes des ménages des $m = 80$ quartiers sélectionnés est égale à 29800.
2. On a affaire à un sondage par grappe. Par conséquent, avec les notations de l'exercice, une estimation ponctuelle de la note moyenne d'un ménage est

$$\overline{y}_\omega = \frac{M}{mN} \sum_{j=1}^{m} t_j = \frac{400}{80 \times 20000} \times 29800 = 7.45.$$

3. On a $95\% = 100(1-\alpha)\%$ avec $\alpha = 0.05$. On a $\mathbb{P}(|Z| \geq z_\alpha) = \alpha = 0.05$, $Z \sim \mathcal{N}(0,1)$, avec $z_\alpha = 1.96$. Un intervalle de confiance pour la note moyenne d'un ménage $\overline{y}_U$ au niveau 95% est

$$\begin{aligned} i_{\overline{y}_U} &= \left[\overline{y}_\omega - z_\alpha \sqrt{\frac{M^2}{N^2}\left(1 - \frac{m}{M}\right)\frac{1}{m}\Xi_\omega^2},\ \overline{y}_\omega + z_\alpha \sqrt{\frac{M^2}{N^2}\left(1 - \frac{m}{M}\right)\frac{1}{m}\Xi_\omega^2} \right] \\ &= \left[7.45 - 1.96\sqrt{\frac{400^2}{20000^2}\left(1 - \frac{80}{400}\right)\frac{1}{80}\Xi_\omega^2},\ 7.45 + 1.96\sqrt{\frac{400^2}{20000^2}\left(1 - \frac{80}{400}\right)\frac{1}{80}\Xi_\omega^2} \right], \end{aligned}$$

avec Ξ_ω^2 que l'on peut écrire comme

$$\Xi_\omega^2 = \frac{1}{m-1}\left(\sum_{j=1}^{m} t_j^2 - \frac{1}{m}\left(\sum_{j=1}^{m} t_j \right)^2 \right) = \frac{1}{80-1}\left(58804000 - \frac{1}{80} \times 29800^2 \right) = 603841.8.$$

Après calcul, on trouve

$$i_{\overline{y}_U} = [4.403875,\ 10.49612]\,,$$

que l'on peut tronquer comme [4.403875, 10] (la note maximale étant 10). Ainsi, il y a 95 chances sur 100 que [4.403875, 10] contienne $\overline{y}_U$. Les ménags sont donc plutôt satisfait de leur ville.

10 Formulaire

10.1 Formules dans le cadre PESR

Paramètres-population et les paramètres-échantillon correspondants :

	Population U	Échantillon ω
Taille	N	n
Taux de sondage	$\square$	$f = \frac{n}{N}$
Moyenne	$\overline{y}_U = \frac{1}{N}\sum_{i=1}^{N} y_i$	$\overline{y}_\omega = \frac{1}{n}\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in \omega\}}$
Écart-type corrigé	$s_U = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y}_U)^2}$	$s_\omega = \sqrt{\frac{1}{n-1}\sum_{i=1}^{N}(y_i - \overline{y}_\omega)^2 \mathbb{1}_{\{u_i \in \omega\}}}$
Écart-type de $\overline{y}_W$	$\sigma(\overline{y}_W) = \sqrt{(1-f)\frac{s_U^2}{n}}$	$s(\overline{y}_\omega) = \sqrt{(1-f)\frac{s_\omega^2}{n}}$

Autre notions utilisées autour de $\overline{y}_U$ (niveau : $100(1-\alpha)\%$, $\alpha \in]0,1[$) :

Intervalle de confiance	$i_{\overline{y}_U} = \left[\overline{y}_\omega - z_\alpha\sqrt{(1-f)\frac{s_\omega^2}{n}},\ \overline{y}_\omega + z_\alpha\sqrt{(1-f)\frac{s_\omega^2}{n}}\right]$
Incertitude absolue	$d_\omega = z_\alpha\sqrt{(1-f)\frac{s_\omega^2}{n}}$
Incertitude relative	$d_\omega^* = \frac{d_\omega}{\overline{y}_\omega}$
Taille n telle que $d_\omega \leq d_0$	$n \geq \frac{N z_\alpha^2 s_\omega^2}{N d_0^2 + z_\alpha^2 s_\omega^2}$
Taille n telle que $d_\omega^* \leq d_1$	$n \geq \frac{N z_\alpha^2 s_\omega^2}{N(\overline{y}_\omega d_1)^2 + z_\alpha^2 s_\omega^2}$

10.2 Formules dans le cadre PESR : proportion

Paramètres-population et les paramètres-échantillon correspondants :

	Population U	Échantillon ω
Taille	N	n
Taux de sondage	$\square$	$f = \dfrac{n}{N}$
Proportion	$p_U = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} y_i$	$p_\omega = \dfrac{1}{n}\displaystyle\sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in \omega\}}$
Écart-type de p_W	$\sigma(p_W) = \sqrt{(1-f)\dfrac{N}{n(N-1)} p_U(1-p_U)}$	$s(p_\omega) = \sqrt{(1-f)\dfrac{p_\omega(1-p_\omega)}{n-1}}$

Autre notions utilisées autour de p_U (niveau : $100(1-\alpha)\%$, $\alpha \in]0,1[$) :

Intervalle de confiance	$i_{p_U} = \left[p_\omega - z_\alpha \sqrt{(1-f)\dfrac{p_\omega(1-p_\omega)}{n-1}},\ p_\omega + z_\alpha \sqrt{(1-f)\dfrac{p_\omega(1-p_\omega)}{n-1}} \right]$
Incertitude absolue	$d_\omega = z_\alpha \sqrt{(1-f)\dfrac{p_\omega(1-p_\omega)}{n-1}}$
Incertitude relative	$d^*_\omega = \dfrac{d_\omega}{p_\omega}$
Taille n telle que $d_\omega \leq d_0$	$n \geq \dfrac{N z_\alpha^2 p_\omega(1-p_\omega)}{N d_0^2 + z_\alpha^2 p_\omega(1-p_\omega)}$
Taille n telle que $d^*_\omega \leq d_1$	$n \geq \dfrac{N z_\alpha^2 p_\omega(1-p_\omega)}{N(p_\omega d_1)^2 + z_\alpha^2 p_\omega(1-p_\omega)}$

10.3 Formules dans le cadre PEAR

Paramètres-population et les paramètres-échantillon correspondants :

	Population U	Échantillon $\omega = (\omega_1, \ldots, \omega_n)$
Taille	N	n
Moyenne	$\overline{y}_U = \frac{1}{N}\sum_{i=1}^{N} y_i$	$\overline{y}_\omega = \frac{1}{n}\sum_{i=1}^{N} y_i \sum_{m=1}^{n} \mathbb{1}_{\{\omega_m = u_i\}}$
Écart-type corrigé	$s_U = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y}_U)^2}$	$s_\omega = \sqrt{\frac{1}{n-1}\sum_{i=1}^{N}(y_i - \overline{y}_\omega)^2 \sum_{m=1}^{n} \mathbb{1}_{\{\omega_m = u_i\}}}$
Écart-type de $\overline{y}_W$	$\sigma(\overline{y}_W) = \sqrt{\frac{N-1}{N}\frac{s_U^2}{n}}$	$s(\overline{y}_\omega) = \sqrt{\frac{s_\omega^2}{n}}$

Autre notions utilisées autour de $\overline{y}_U$ (niveau : $100(1-\alpha)\%$, $\alpha \in]0,1[$) :

Intervalle de confiance	$i_{\overline{y}_U} = \left[\overline{y}_\omega - z_\alpha\sqrt{\frac{s_\omega^2}{n}},\ \overline{y}_\omega + z_\alpha\sqrt{\frac{s_\omega^2}{n}}\right]$
Incertitude absolue	$d_\omega = z_\alpha\sqrt{\frac{s_\omega^2}{n}}$
Incertitude relative	$d_\omega^* = \frac{d_\omega}{\overline{y}_\omega}$
Taille n telle que $d_\omega \leq d_0$	$n \geq \left(\frac{z_\alpha s_\omega}{d_0}\right)^2$
Taille n telle que $d_\omega^* \leq d_1$	$n \geq \left(\frac{z_\alpha s_\omega}{\overline{y}_\omega d_1}\right)^2$

10.4 Formules dans le cadre PEAR : proportion

Paramètres-population et les paramètres-échantillon correspondants :

	Population U	Échantillon $\omega = (\omega_1, \ldots, \omega_n)$
Taille	N	n
Proportion	$p_U = \frac{1}{N}\sum_{i=1}^{N} y_i$	$p_\omega = \frac{1}{n}\sum_{i=1}^{N} y_i \sum_{m=1}^{n} \mathbb{1}_{\{\omega_m = u_i\}}$
Écart-type de p_W	$\sigma(p_W) = \sqrt{\frac{p_U(1-p_U)}{n}}$	$s(p_\omega) = \sqrt{\frac{p_\omega(1-p_\omega)}{n-1}}$

Autre notions utilisées autour de p_U (niveau : $100(1-\alpha)\%$, $\alpha \in]0,1[$) :

Intervalle de confiance	$i_{p_U} = \left[p_\omega - z_\alpha\sqrt{\frac{p_\omega(1-p_\omega)}{n-1}},\ p_\omega + z_\alpha\sqrt{\frac{p_\omega(1-p_\omega)}{n-1}}\right]$
Incertitude absolue	$d_\omega = z_\alpha\sqrt{\frac{p_\omega(1-p_\omega)}{n-1}}$
Incertitude relative	$d^*_\omega = \frac{d_\omega}{p_\omega}$
Taille n telle que $d_\omega \leq d_0$	$n \geq \frac{z^2_\alpha p_\omega(1-p_\omega)}{d_0^2}$
Taille n telle que $d^*_\omega \leq d_1$	$n \geq \frac{z^2_\alpha p_\omega(1-p_\omega)}{(p_\omega d_1)^2}$

10.5 Formules dans le cadre ST

Paramètres-strates et les paramètres-échantillon correspondants, $\omega = (\omega_1, \ldots, \omega_H)$:

	Strate U_h	Échantillon ω_h
Taille	N_h	n_h
Taux de sondage	$\square$	$f_h = \dfrac{n_h}{N_h}$
Moyenne	$\overline{y}_{U_h} = \dfrac{1}{N_h}\displaystyle\sum_{i=1}^{N_h} y_i$	$\overline{y}_{\omega_h} = \dfrac{1}{n_h}\displaystyle\sum_{i=1}^{N_h} y_i \mathbb{1}_{\{u_i \in \omega_h\}}$
Écart-type corrigé	$s_{U_h} = \sqrt{\dfrac{1}{N_h - 1}\displaystyle\sum_{i=1}^{N_h}(y_i - \overline{y}_{U_h})^2}$	$s_{\omega_h} = \sqrt{\dfrac{1}{n_h - 1}\displaystyle\sum_{i=1}^{N_h}(y_i - \overline{y}_{\omega_h})^2 \mathbb{1}_{\{u_i \in \omega_h\}}}$
Écart-type de $\overline{y}_{W_h}$	$\sigma(\overline{y}_{W_h}) = \sqrt{(1 - f_h)\dfrac{s^2_{U_h}}{n_h}}$	$s(\overline{y}_{\omega_h}) = \sqrt{(1 - f_h)\dfrac{s^2_{\omega_h}}{n_h}}$

Paramètres-population et les paramètres-échantillon correspondants :

	Population U	Échantillon ω
Taille	N	n
Moyenne	$\overline{y}_U = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} y_i$	$\overline{y}_\omega = \dfrac{1}{N}\displaystyle\sum_{h=1}^{H} N_h \overline{y}_{\omega_h}$
Écart-type de $\overline{y}_W$	$\sigma(\overline{y}_W) = \sqrt{\dfrac{1}{N^2}\displaystyle\sum_{h=1}^{H} N_h^2 (1 - f_h)\dfrac{s^2_{U_h}}{n_h}}$	$s(\overline{y}_\omega) = \sqrt{\dfrac{1}{N^2}\displaystyle\sum_{h=1}^{H} N_h^2 (1 - f_h)\dfrac{s^2_{\omega_h}}{n_h}}$

Plans de sondage aléatoires de types STP et STO :

	STP	STO	STO (applicable)
n_h	$\dfrac{n}{N}N_h$	$n\dfrac{N_h s_{U_h}}{\sum_{\ell=1}^{H} N_\ell s_{U_\ell}}$	$n\dfrac{N_h s_{\omega_h}}{\sum_{\ell=1}^{H} N_\ell s_{\omega_\ell}}$

Autre notions utilisées autour de $\overline{y}_U$ (niveau : $100(1-\alpha)\%$, $\alpha \in]0,1[$) :

Intervalle de confiance	$i_{\overline{y}_U} = \left[\overline{y}_\omega - z_\alpha\sqrt{\dfrac{1}{N^2}\sum_{h=1}^{H} N_h^2(1-f_h)\dfrac{s_{\omega_h}^2}{n_h}},\ \overline{y}_\omega + z_\alpha\sqrt{\dfrac{1}{N^2}\sum_{h=1}^{H} N_h^2(1-f_h)\dfrac{s_{\omega_h}^2}{n_h}}\right]$
Incertitude absolue	$d_\omega = z_\alpha\sqrt{\dfrac{1}{N^2}\sum_{h=1}^{H} N_h^2(1-f_h)\dfrac{s_{\omega_h}^2}{n_h}}$
Incertitude relative	$d_\omega^* = \dfrac{d_\omega}{\overline{y}_\omega}$
Taille n telle que $d_\omega \leq d_0$	∘ pour un plan de sondage aléatoire de type STP : $n \geq \dfrac{N z_\alpha^2 \sum_{h=1}^{H} N_h s_{\omega_h}^2}{N^2 d_0^2 + z_\alpha^2 \sum_{h=1}^{H} N_h s_{\omega_h}^2}$, ∘ pour un plan de sondage aléatoire de type STO : $n \geq \dfrac{z_\alpha^2\left(\sum_{h=1}^{H} N_h s_{\omega_h}\right)^2}{N^2 d_0^2 + z_\alpha^2 \sum_{h=1}^{H} N_h s_{\omega_h}^2}$.
Taille n telle que $d_\omega^* \leq d_1$	∘ pour un plan de sondage aléatoire de type STP : $n \geq \dfrac{N z_\alpha^2 \sum_{h=1}^{H} N_h s_{\omega_h}^2}{N^2 (d_1\overline{y}_\omega)^2 + z_\alpha^2 \sum_{h=1}^{H} N_h s_{\omega_h}^2}$, ∘ pour un plan de sondage aléatoire de type STO : $n \geq \dfrac{z_\alpha^2\left(\sum_{h=1}^{H} N_h s_{\omega_h}\right)^2}{N^2 (d_1\overline{y}_\omega)^2 + z_\alpha^2 \sum_{h=1}^{H} N_h s_{\omega_h}^2}$.

10.6 Formules dans le cadre ST : proportion

Paramètres-strates et les paramètres-échantillon correspondants, $\omega = (\omega_1, \ldots, \omega_H)$:

	Strate U_h	Échantillon ω_h
Taille	N_h	n_h
Taux de sondage	$\square$	$f_h = \dfrac{n_h}{N_h}$
Proportion	$p_{U_h} = \dfrac{1}{N_h}\displaystyle\sum_{i=1}^{N_h} y_i$	$p_{\omega_h} = \dfrac{1}{n_h}\displaystyle\sum_{i=1}^{N_h} y_i \mathbb{1}_{\{u_i \in \omega_h\}}$
Écart-type de p_{W_h}	$\sigma(p_{W_h}) = \sqrt{(1-f_h)\dfrac{N_h}{n_h(N_h-1)} p_{U_h}(1-p_{U_h})}$	$s(p_{\omega_h}) = \sqrt{(1-f_h)\dfrac{p_{U_h}(1-p_{U_h})}{n_h - 1}}$

Paramètres-population et les paramètres-échantillon correspondants :

	Population U	Échantillon ω
Taille	N	n
Proportion	$p_U = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} y_i$	$p_\omega = \dfrac{1}{N}\displaystyle\sum_{h=1}^{H} N_h p_{\omega_h}$
Écart-type de p_W	$\sigma(p_W) = \sqrt{\dfrac{1}{N^2}\displaystyle\sum_{h=1}^{H} N_h^2 \sigma^2(p_{W_h})}$	$s(p_\omega) = \sqrt{\dfrac{1}{N^2}\displaystyle\sum_{h=1}^{H} N_h^2 (1-f_h)\dfrac{p_{\omega_h}(1-p_{\omega_h})}{n_h - 1}}$

Plans de sondage aléatoires de types STP et STO :

	STP	STO	STO (applicable)
n_h	$\dfrac{n}{N}N_h$	$n\dfrac{N_h\sqrt{p_{U_h}(1-p_{U_h})}}{\sum_{\ell=1}^{H}N_\ell\sqrt{p_{U_\ell}(1-p_{U_\ell})}}$	$n\dfrac{N_h\sqrt{p_{\omega_h}(1-p_{\omega_h})}}{\sum_{\ell=1}^{H}N_\ell\sqrt{p_{\omega_\ell}(1-p_{\omega_\ell})}}$

Autre notions utilisées autour de p_U (niveau : $100(1-\alpha)\%$, $\alpha\in]0,1[$) :

Intervalle de confiance	$i_{p_U}=\left[p_\omega - z_\alpha\sqrt{\dfrac{1}{N^2}\sum_{h=1}^{H}N_h^2(1-f_h)\dfrac{p_{\omega_h}(1-p_{\omega_h})}{n_h-1}},\ p_\omega + z_\alpha\sqrt{\dfrac{1}{N^2}\sum_{h=1}^{H}N_h^2(1-f_h)\dfrac{p_{\omega_h}(1-p_{\omega_h})}{n_h-1}}\right]$
Incertitude absolue	$d_\omega = z_\alpha\sqrt{\dfrac{1}{N^2}\sum_{h=1}^{H}N_h^2(1-f_h)\dfrac{p_{\omega_h}(1-p_{\omega_h})}{n_h-1}}$
Incertitude relative	$d_\omega^* = \dfrac{d_\omega}{\overline{y}_\omega}$
Taille n telle que $d_\omega \leq d_0$	∘ pour un plan de sondage aléatoire de type STP : $n \geq \dfrac{Nz_\alpha^2\sum_{h=1}^{H}N_hp_{\omega_h}(1-p_{\omega_h})}{N^2d_0^2+z_\alpha^2\sum_{h=1}^{H}N_hp_{\omega_h}(1-p_{\omega_h})}$, ∘ pour un plan de sondage aléatoire de type STO : $n \geq \dfrac{z_\alpha^2\left(\sum_{h=1}^{H}N_h\sqrt{p_{\omega_h}(1-p_{\omega_h})}\right)^2}{N^2d_0^2+z_\alpha^2\sum_{h=1}^{H}N_hp_{\omega_h}(1-p_{\omega_h})}$.
Taille n telle que $d_\omega^* \leq d_1$	∘ pour un plan de sondage aléatoire de type STP : $n \geq \dfrac{Nz_\alpha^2\sum_{h=1}^{H}N_hs_{\omega_h}^2}{N^2(d_1p_\omega)^2+z_\alpha^2\sum_{h=1}^{H}N_hs_{\omega_h}^2}$, ∘ pour un plan de sondage aléatoire de type STO : $n \geq \dfrac{z_\alpha^2\left(\sum_{h=1}^{H}N_h\sqrt{p_{\omega_h}(1-p_{\omega_h})}\right)^2}{N^2(d_1p_\omega)^2+z_\alpha^2\sum_{h=1}^{H}N_hp_{\omega_h}(1-p_{\omega_h})}$.

10.7 Formules dans le cadre G

Paramètres-groupe et les paramètres-échantillon correspondants, $\omega = (\omega_1, \ldots, \omega_m)$:

	Groupe G_j	Échantillon ω_j
Taille	M	m
Total par groupe	$t_j = \sum_{i=1}^{N} y_i \mathbb{1}_{\{u_i \in G_j\}}$	

Paramètres-population et les paramètres-échantillon correspondants :

	Population U	Échantillon ω
Taille	N	n
Moyenne	$\overline{y}_U = \frac{1}{N} \sum_{i=1}^{N} y_i$	$\overline{y}_\omega = \frac{M}{mN} \sum_{j=1}^{M} t_j \mathbb{1}_{\{G_j \in \omega\}}$
Écart-type corrigé	$\Xi_U = \sqrt{\frac{1}{M-1} \sum_{j=1}^{M} \left(t_j - \frac{N}{M} \overline{y}_U \right)^2}$	$\Xi_\omega = \sqrt{\frac{1}{m-1} \sum_{j=1}^{M} \left(t_j - \frac{1}{m} \sum_{k=1}^{M} t_k \mathbb{1}_{\{G_k \in \omega\}} \right)^2 \mathbb{1}_{\{G_j \in \omega\}}}$
Écart-type de $\overline{y}_W$		$s(\overline{y}_\omega) = \sqrt{\frac{M^2}{N^2} \left(1 - \frac{m}{M} \right) \frac{1}{m} \Xi_\omega^2}$

Autre notions utilisées autour de $\overline{y}_U$ (niveau : $100(1-\alpha)\%$, $\alpha \in]0,1[$) :

Intervalle de confiance	$i_{\overline{y}_U} = \left[\overline{y}_\omega - z_\alpha\sqrt{\frac{M^2}{N^2}\left(1-\frac{m}{M}\right)\frac{1}{m}\Xi_\omega^2},\ \overline{y}_\omega + z_\alpha\sqrt{\frac{M^2}{N^2}\left(1-\frac{m}{M}\right)\frac{1}{m}\Xi_\omega^2}\right]$
Incertitude absolue	$d_\omega = z_\alpha s(\overline{y}_\omega) = z_\alpha\sqrt{\frac{M^2}{N^2}\left(1-\frac{m}{M}\right)\frac{1}{m}\Xi_\omega^2}$
Incertitude relative	$d_\omega^* = \frac{d_\omega}{\overline{y}_\omega}$
Taille m telle que $d_\omega \leq d_0$	$m \geq \frac{z_\alpha^2 M^2 \Xi_\omega^2}{N^2 d_0^2 + z_\alpha^2 M \Xi_\omega^2}$
Taille m telle que $d_\omega^* \leq d_1$	$m \geq \frac{z_\alpha^2 M^2 \Xi_\omega^2}{N^2(\overline{y}_\omega d_1)^2 + z_\alpha^2 M \Xi_\omega^2}$

10.8 Table : Loi normale

Soit $Z \sim \mathcal{N}(0,1)$. La table ci-dessous donne, pour un α choisi, la valeur z_α telle que $\mathbb{P}(|Z| \geq z_\alpha) = \alpha$.

α	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	∞	2.576	2.326	2.170	2.054	1.960	1.881	1.812	1.751	1.695
0.10	1.645	1.598	1.555	1.514	1.476	1.440	1.405	1.372	1.341	1.311
0.20	1.282	1.254	1.227	1.200	1.175	1.150	1.126	1.103	1.080	1.058
0.30	1.036	1.015	0.994	0.974	0.954	0.935	0.915	0.896	0.878	0.860
0.40	0.842	0.824	0.806	0.789	0.772	0.755	0.739	0.722	0.706	0.690
0.50	0.674	0.659	0.643	0.628	0.613	0.598	0.583	0.568	0.553	0.539
0.60	0.524	0.510	0.496	0.482	0.468	0.454	0.440	0.426	0.412	0.399
0.70	0.385	0.372	0.358	0.345	0.332	0.319	0.305	0.292	0.279	0.266
0.80	0.253	0.240	0.228	0.215	0.202	0.189	0.176	0.164	0.151	0.138
0.90	0.126	0.113	0.100	0.088	0.075	0.063	0.050	0.038	0.025	0.013

10.9 Table : Loi de Student à ν degrés de liberté

Soit $T \sim \mathcal{T}(\nu)$. La table ci-dessous donne, pour un α et un ν choisis, la valeur $t_\alpha(\nu)$ telle que $\mathbb{P}(|T| \geq t_\alpha(\nu)) = \alpha$.

ν \ α	0.90	0.50	0.30	0.20	0.10	0.05	0.02	0.01	0.001
1	0.158	1.000	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.142	0.816	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	0.137	0.765	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.134	0.741	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.132	0.727	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.131	0.718	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.130	0.711	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.130	0.706	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.129	0.703	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.129	0.700	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.129	0.697	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.128	0.695	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.128	0.694	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.128	0.692	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.128	0.691	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.128	0.690	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.128	0.689	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.127	0.688	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.127	0.688	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.127	0.687	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.127	0.686	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.127	0.686	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.127	0.685	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	0.127	0.685	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.127	0.684	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.127	0.684	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.127	0.684	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.127	0.683	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.127	0.683	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.127	0.683	1.055	1.310	1.697	2.042	2.457	2.750	3.646

10.10 Table : Loi du chi-deux à ν degrés de liberté

Soit $K \sim \chi^2(\nu)$. La table ci-dessous donne, pour un α et un ν choisis, la valeur $k_\alpha(\nu)$ telle que $\mathbb{P}(K \geq k_\alpha(\nu)) = \alpha$.

ν \ α	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.001
1	0.0002	0.001	0.004	0.016	2.71	3.84	5.02	6.63	10.83
2	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21	13.82
3	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34	16.27
4	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28	18.47
5	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09	20.51
6	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81	22.46
7	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	24.32
8	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	26.12
9	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	27.88
10	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	29.59
11	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.73	31.26
12	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	32.91
13	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	34.53
14	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	36.12
15	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	37.70
16	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	39.25
17	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	40.79
18	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	42.31
19	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	43.82
20	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	45.31
21	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	46.80
22	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	48.27
23	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	49.73
24	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	51.18
25	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	52.62
26	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	54.05
27	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	55.48
28	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	56.89
29	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	58.30
30	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	59.70

Index