

MAT 3777

Échantillonnage et sondages

Chapitre 8

Sujets choisis

P. Boily (uOttawa)

Session d'hiver – 2022

P. Boily (uOttawa)

Aperçu

8.1 – Échantillonnage avec probabilité proportionnelle à la taille (p.3)

- Méthodes de sélection d'un PPT avec remise (p.5)
- Estimation des paramètres (p.9)

8.2 – Échantillonnage à plusieurs degrés (p.16)

- Échantillonnage aléatoire simple à deux degrés (p.18)
- Estimation des paramètres (p.20)

8.3 – Échantillonnage à plusieurs phases (p.28)

- Échantillonnage aléatoire simple à deux phases (p.29)
- Estimation de la variance d'échantillonnage (p.34)

8.4 – Méli-mélo (p.38)

- Effet de plan (p.39)
- Ajustement pour la non-réponse (p.41)
- Estimation de la taille d'une population (p.46)
- Réponse aléatoire (p.53)
- Échantillonnage de Bernoulli (p.60)

8.1 – Échantillonnage avec probabilité proportionnelle à la taille

En pratique, la **taille** (que cela soit une caractéristique physique ou non) des unités d'échantillonnage est souvent très **variable** – un EAS n'est pas toujours efficace puisqu'il ne tient pas compte de l'**importance que peuvent avoir les unités plus grandes** de la population.

On peut parfois mettre à profit des **renseignements supplémentaires sur la taille des unités** afin de sélectionner un échantillon donnant un estimateur plus précis des paramètres d'intérêt.

Une façon possible de s'y prendre: **assigner des probabilités de sélection (potentiellement) inégales** aux différentes unités.

Exemple: en général, plus la superficie d'un pays est élevé, plus sa population l'est aussi ($\rho = 0.46$).

Si on cherche à estimer la population de la planète, il pourrait être souhaitable d'adopter un système d'échantillonnage dans lequel la probabilité de sélection d'un pays est **proportionnelle à sa superficie** – dans un EAS, il est fort probable que ni la **Chine**, ni l'**Inde** ne soient sélectionnées, ce qui entraîne une sous-estimation du total recherché.

Si la variable d'intérêt est liée (plus ou moins) à la taille de l'unité, on peut assigner une **probabilité de sélection proportionnelle à la taille de l'unité (PPT)**.

Dans un PPT, les unités prélevées au préalable peuvent être **remises** dans la population, permettant la **sélection multiple d'une même unité**.

8.1.1 – Méthodes de sélection d'un PPT avec remise

Nous allons considérer deux méthodes de sélection d'un échantillon PPT:

- **PPT selon l'approche des totaux cumulés**, et
- **PPT selon l'approche de Lahiri**.

Dans les deux cas, la procédure de sélection de l'échantillon PPT consiste à associer à chaque unité une **étendue de nombres** (ce sont souvent des **entiers**, mais ce n'est pas nécessaire), liée à la **taille de l'unité**, et à prélever les unités qui correspondent à des nombres choisis **au hasard** dans l'ensemble de nombres associés à la population **dans son entiereté**.

Avec la **méthode des totaux cumulés**, la **taille** de la i –ème unité (dans une population contenant N unités) est dénotée par x_i , $1 \leq i \leq N$.

On associe ensuite une **étendue** à chaque unité de la manière suivante:

Unité	Étendue
1	1 à x_1
2	$x_1 + 1$ à $x_1 + x_2$
3	$x_1 + x_2 + 1$ à $x_1 + x_2 + x_3$
\vdots	\vdots
$N - 1$	$x_1 + \cdots + x_{N-2} + 1$ à $x_1 + \cdots + x_{N-2} + x_{N-1}$
N	$x_1 + \cdots + x_{N-1} + 1$ à $x_1 + \cdots + x_{N-1} + x_N$

Finalement, on prélève un échantillon PPT en choisissant n entiers **au hasard** entre 1 et $X = x_1 + \cdots + x_{N-1} + x_N$ (**avec remise**) et en sélectionnant les unités **associées à ces entiers**.

Exemple: Dans un village, il y a 8 vergers, contenant respectivement un certain nombre de pommiers. Un échantillon de $n = 3$ vergers est prélevé (avec remise), de manière proportionnelle au nombre de pommiers.


# série i	Taille x_i	Taille cumulée	Étendue associée
1	50	50	1 – 50
2	30	80	51 – 80
3	25	105	81 – 105
4	40	145	106 – 145
5	26	171	146 – 171
6	44	215	172 – 215
7	20	235	216 – 235
8	35	270	236 – 270


On choisit $n = 3$ entiers au hasard entre 1 et 270: 108, 140, et 201, par exemple. Les unités associées sont la 4^{ième}, la 4^{ième}, et la 6^{ième}.

Avec la **méthode de Lahiri**, on dénote toujours la taille d'une unité par x_i , $1 \leq i \leq N$, mais sans avoir **à calculer et reporter les totaux cumulés successifs** (ce qui peut s'avérer fastidieux, même avec un ordinateur).

La méthode consiste à sélectionner un couple (i, j) d'entiers au hasard, où $1 \leq i \leq N$ et $1 \leq j \leq M = \max\{x_i | 1 \leq i \leq N\}$.

Si $j \leq x_i$, la i -ème unité est ajoutée à l'échantillon. Sinon, on rejette la paire (i, j) et on continue jusqu'à ce que n unités aient été choisies.

 Il y a d'autre façon de s'y prendre; ce qui importe, c'est d'avoir un **mécanisme pour sélectionner un échantillon PPT**.

 Il est préférable de prélever sans remise, mais l'échantillonnage avec remise offre une approximation raisonnable si $\frac{n}{N}$ est "**suffisamment petit**".

8.1.2 – Estimation des paramètres

Revisitons l'exemple des vergers, dans lequel u_i représente le rendement de tous les pommiers du i -ème verger.

# série i	# pommiers x_i	π_i	Rendement
1	50	50/270	$u_1 = 2250$
2	30	30/270	$u_2 = 1080$
3	25	25/270	$u_3 = 1300$
4	40	40/270	$u_4 = 1400$
5	26	26/270	$u_5 = 1196$
6	44	44/270	$u_6 = 1716$
7	20	20/270	$u_7 = 820$
8	35	35/270	$u_8 = 1680$

On s'intéresse à la production **totale** de pommes du village, $\tau = 11,442$.

Puisqu'**en principe**, un verger qui contient plus de pommiers devrait produire plus de pommes, on prélève un échantillon PPT (avec remise) de $n = 3$ unités, où le nombre de pommiers dans verger représente sa taille.

Dans ce qui suit, nous illustrerons les concepts à l'aide de l'échantillon

$$\{y_1 = u_4 = 1400, y_2 = u_4 = 1400, y_3 = u_6 = 1716\}.$$

Si l'échantillon \mathcal{Y} , avec $|\mathcal{Y}| = n$, est prélevé de la population \mathcal{U} à partir d'un plan d'échantillonnage PPT, les unités y_1, \dots, y_n sont **indépendantes** et distribuées selon

$$\begin{array}{c|ccccc} y_i & u_1 & \cdots & u_j & \cdots & u_N \\ \hline p(y_i) & \pi_1 & \cdots & \pi_j & \cdots & \pi_N \end{array}$$

où $0 < \pi_j < 1$ pour tout $1 \leq j \leq N$ et $\pi_1 + \cdots + \pi_N = 1$.

Pour tout $1 \leq i \leq n$, posons $w_i = \frac{u_j}{\pi_j}$, si $y_i = u_j$ pour un $1 \leq j \leq N$. Les **poids** w_i sont également **indépendants** et distribués selon

$$P(y_i = u_j) = P\left(w_i = \frac{u_j}{\pi_j}\right) = \pi_j, \quad 1 \leq i \leq n, \quad 1 \leq j \leq N.$$

On remarque que, pour tout $1 \leq i \leq n$, l'**espérance des poids** équivaut à

$$E(w_i) = \sum_{j=1}^N w_j P(w_i = w_j) = \sum_{j=1}^N \frac{u_j}{\pi_j} \cdot \pi_j = \sum_{j=1}^N u_j = \tau.$$

C'est donc dire que

$$\hat{\tau}_{\text{ppt}} = \overline{w} = \frac{1}{n} \sum_{i=1}^n w_i$$

offre un **estimateur sans biais** du total τ .

La **variance d'échantillonnage** se calcule comme suit:

$$\begin{aligned}
 V(\hat{\tau}_{\text{ppt}}) &= V\left(\frac{1}{n} \sum_{i=1}^n w_i\right) = \underbrace{\frac{1}{n^2} \sum_{i=1}^n V(w_i)}_{\text{ind. des } w_i} = \frac{1}{n^2} \sum_{i=1}^n \left[\sum_{j=1}^N (w_j - \tau)^2 P(w_i = w_j) \right] \\
 &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^N \left(\frac{u_j}{\pi_j} - \tau \right)^2 \pi_j = \frac{1}{n} \sum_{j=1}^N \left(\frac{u_j}{\pi_j} - \tau \right)^2 \pi_j = \frac{1}{n} \sum_{j=1}^N \left(\frac{u_j^2}{\pi_j} - \frac{2\tau u_j}{\pi_j} + \tau^2 \right) \pi_j \\
 &= \frac{1}{n} \left(\sum_{j=1}^N \frac{u_j^2}{\pi_j} - 2\tau \underbrace{\sum_{j=1}^N u_j}_{=\tau} + \tau^2 \underbrace{\sum_{j=1}^N \pi_j}_{=1} \right) = \frac{1}{n} \left(\sum_{j=1}^N \frac{u_j^2}{\pi_j} - \tau^2 \right).
 \end{aligned}$$

En pratique, on ne connaît pas τ , alors on utilise l'estimateur **non-biaisé**

$$\hat{V}(\hat{\tau}_{\text{ppt}}) = \frac{1}{n(n-1)} \left(\sum_{i=1}^n w_i^2 - n\hat{\tau}_{\text{ppt}}^2 \right).$$

Théorème de la limite centrée – PPT

Si n et $N - n$ sont suffisamment élevés, alors

$$\hat{\tau}_{\text{ppt}} \sim_{\text{approx.}} \mathcal{N} \left(\tau, \hat{V}(\hat{\tau}_{\text{ppt}}) \right).$$

La **marge d'erreur sur l'estimation** et l'**intervalle de confiance de τ à environ 95%** sont ainsi

$$\hat{B}_{\tau;\text{ppt}} = 2\sqrt{\hat{V}(\hat{\tau}_{\text{ppt}})} \quad \text{et} \quad \text{IC}_{\text{ppt}}(\tau; 0.95) = \hat{\tau}_{\text{ppt}} \pm \hat{B}_{\tau;\text{ppt}}.$$

Exemple: Dans l'exemple des vergers, nous avons

$$\hat{\tau}_{\text{ppt}} = \frac{1}{3} \left[\underbrace{\frac{1400}{40/270}}_{w_1} + \underbrace{\frac{1400}{40/270}}_{w_2} + \underbrace{\frac{1716}{44/270}}_{w_3} \right] = 9810;$$

$$\hat{V}(\hat{\tau}_{\text{ppt}}) = \frac{1}{3(2)} \left[\left(\underbrace{\frac{1400}{40/270}}_{w_1} \right)^2 + \left(\underbrace{\frac{1400}{40/270}}_{w_2} \right)^2 + \left(\underbrace{\frac{1716}{44/270}}_{w_3} \right)^2 - 3 \cdot \underbrace{9810^2}_{\hat{\tau}_{\text{ppt}}^2} \right]$$

$$= 129,600.$$

Conséquemment, l'intervalle de confiance pour le rendement total des pommiers du village à environ 95% est

$$\text{IC}_{\text{ppt}}(\tau; 0.95) = 9810 \pm 2\sqrt{129,600} \equiv (9090, 10530).$$

La valeur réelle $\tau = 11,442$ **ne se retrouve pas** dans l'intervalle de confiance – pourquoi est-ce le cas? Est-ce problématique?

En général, $V(\hat{\tau}_{\text{ppt}}) \leq V(\hat{\tau}_{\text{EAS}})$. Dans l'exemple des pommiers, on peut montrer que

$$V(\hat{\tau}_{\text{EAS}}) \approx 8^2 \cdot \frac{172981.4375}{3} \left(\frac{8-3}{8-1} \right) = 2,635,907.619, \quad \text{et}$$

$$V(\hat{\tau}_{\text{ppt}}) \approx \frac{1}{3} \left[\frac{2250^2}{50/270} + \cdots + \frac{1680^2}{35/270} - 11,442^2 \right] = 723,912.$$

On peut également donner un estimé de la **moyenne** de population μ à l'aide de

$$\hat{\mu}_{\text{ppt}} = \frac{\hat{\tau}_{\text{ppt}}}{N}, \quad \hat{V}(\hat{\mu}_{\text{ppt}}) = \frac{\hat{V}(\hat{\tau}_{\text{ppt}})}{N^2}, \quad \text{IC}_{\text{ppt}}(\mu; 0.95) = \frac{\text{IC}_{\text{ppt}}(\tau; 0.95)}{N}.$$

8.2 – Échantillonnage à plusieurs degrés

En séparant l'échantillonnage en plusieurs étapes, on peut **réduire les coûts** et **concentrer les opérations logistique autour de points centraux**.

Dans un **échantillonnage à plusieurs degrés** (EnD), on prélève un échantillon d'unités de grande taille (**unités primaires**), puis des sous-unités de ces grandes unités (**unités secondaires**), etc.

Exemple: l'échantillonnage d'une province peut se faire en trois étapes:

1. échantillon de municipalités (**unités primaires**),
2. échantillon de quartiers par municipalités (**unités secondaires**), et
3. échantillon de ménages par quartiers (**unités tertiaires**).

Dans un EnD , l'échantillon est concentré autour de plusieurs **pivots**: dans les études sur le terrain, par exemple, cela à l'avantage de réduire considérablement la surface d'enquête, ce qui aide à **réduire les erreurs non liées à l'échantillonnage** (en plus de **réduire les coûts opérationnels**).

De plus, il arrive souvent que l'on dispose d'informations détaillées pour des **groupes** d'unités d'échantillonnage, mais pas pour des unités **individuelles**: il n'est donc pas nécessaire d'obtenir une base de sondage **complète** (pour **toutes** les unités d'échantillonnage), mais seulement pour celles appartenant aux unités primaires sélectionnés lors du premier tour, par exemple.

On peut utiliser n'importe quelle méthode d'échantillonnage probabiliste à chaque stade, et elles peuvent changer d'un stade à l'autre (un EAS de municipalités, un EAS de quartiers, un SYS de ménages, par exemple).

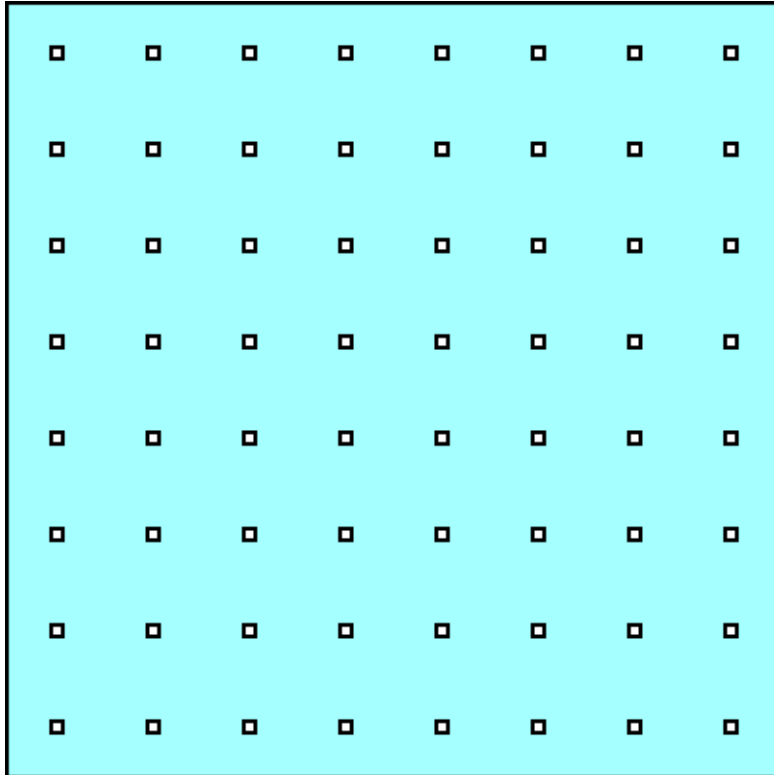
8.2.1 – Échantillonnage aléatoire simple à deux degrés

Si les deux étapes de la sélection se font par EAS, la méthode prend le nom d'**échantillonnage aléatoire simple à deux degrés** (EAS2D).

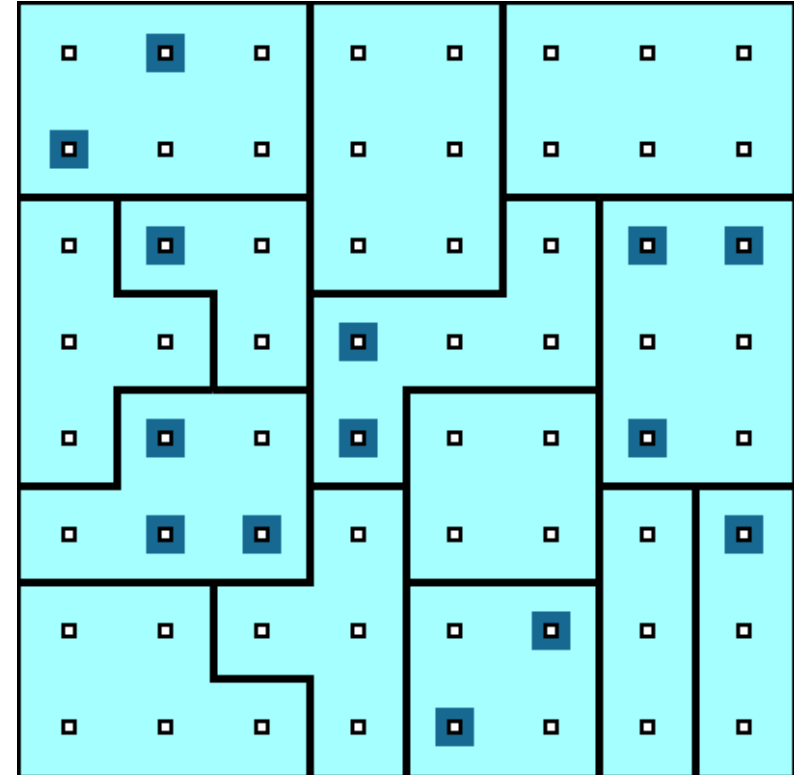
Exemple: on peut estimer la biomasse d'une espèce de plante dans une superficie forestière composée de 40 compartiments (unités primaires) en prélevant un EAS de $m = 8$ compartiments des $M = 40$ compartiments composant la population à l'étude.

Pour chacun de ces compartiments $1 \leq i \leq m$, on prélève ensuite un EAS de n_i parcelles, et on mesure la biomasse en question.

On peut calculer les estimations de la quantité moyenne ou totale de biomasse dans la superficie forestière à l'aide des formules appropriées.



Population



EAS à 2 degrés

8.2.2 – Estimation des paramètres

Soient une population constituée de M unités primaires, et possédant N_ℓ unités secondaires dans la ℓ –ème unité primaire.

Notons par $u_{i,j}$ la valeur de la variable réponse de la j –ième unité du second degré dans la i –ième unité du premier degré.

La **moyenne de la population** est

$$\mu = \frac{\sum_{\ell=1}^M \sum_{j=1}^{N_\ell} u_{\ell,j}}{\sum_{\ell=1}^M N_\ell}.$$

Supposons que l'on prélève un EAS de m unités primaires, et un EAS de n_i unités secondaires dans la i -ème unité primaire.

L'échantillon est donc de taille $n = n_1 + \cdots + n_m$.

On obtient un estimateur non biaisé de μ grâce à l'équation

$$\bar{y}_{\text{EAS2D}} = \frac{1}{m\bar{N}} \sum_{i=1}^m N_i \bar{y}_i = \frac{1}{m\bar{N}} \sum_{i=1}^m \frac{N_i}{n_i} \sum_{k=1}^{n_i} y_{i,k} = \frac{1}{m\bar{N}} \sum_{i=1}^m \sum_{k=1}^{n_i} \frac{M N_i}{m n_i} y_{i,k},$$

où

$$\bar{N} = \frac{1}{M} \sum_{\ell=1}^M N_{\ell} \approx \frac{N_1 + \cdots + N_m}{m}.$$

La variance d'échantillonnage est composée de deux éléments:

- une mesure de la variation **entre les unités du premier degré**, et
- une mesure de la variation **à l'intérieur des unités du premier degré**.

Lorsque $n_i = N_i$, pour tout $1 \leq i \leq N_i$, on fait affaire à un **EPG** et la variance est donnée **uniquement par le premier élément** (cf. chapitre 6).

Dans le cas où $m = M$, on fait affaire à un STR (cf. chapitre 3) et la variance est donnée **uniquement par le second élément**.

Quand $m \neq M$ et $n_i \neq N_i$ pour au moins un i , la variance est une combinaison de ces deux extrêmes: dans ce cas, le second terme représente **la contribution du sous-échantillonnage** (un autre nom pour un EnD).

On peut se servir du **théorème de la variance totale** afin d'estimer la variance d'échantillonnage:

$$\begin{aligned}
 V(\bar{y}_{\text{EAS2D}}) &= E[V(\bar{y}_{\text{EAS2D}} \mid m)] + V(E[\bar{y}_{\text{EAS2D}} \mid m]) \\
 &= \frac{1}{\bar{N}^2} \cdot \frac{\sigma_T^2}{m} \left(\frac{M-m}{M-1} \right) + \frac{1}{mM\bar{N}^2} \sum_{i=1}^m N_i^2 \cdot \frac{\sigma_i^2}{n_i} \left(\frac{N_i - n_i}{N_i - 1} \right) \\
 &\approx \frac{1}{\bar{N}^2} \cdot \frac{s_T^2}{m} \left(1 - \frac{m}{M} \right) + \frac{1}{mM\bar{N}^2} \sum_{i=1}^m N_i^2 \cdot \frac{s_i^2}{n_i} \left(1 - \frac{n_i}{N_i} \right),
 \end{aligned}$$

où

$$s_T^2 = \frac{1}{m-1} \sum_{i=1}^m \left(N_i \bar{y}_i - \bar{N} \bar{y}_{\text{EAS2D}} \right)^2, \quad s_i^2 = \frac{1}{n_i-1} \sum_{k=1}^{n_i} (y_{i,k} - \bar{y}_i)^2.$$

Exemple:

On mesure la biomasse d'une espèce de plante (kg) dans des parcelles de 0.025 ha (unités secondaires) sélectionnées dans $m = 8$ compartiments (unités primaires) choisis au hasard parmi les $M = 40$ compartiments d'une étendue forestière.

Le sommaire des résultats se retrouve dans le tableau suivant:

Comp.	1	2	3	4	5	6	7	8
\bar{y}_i	118	107	109	110	120	95	93	90
s_i^2	436	516	586	456	412	497	755	496
N_i	1760	1975	1615	1785	1775	2050	1680	1865
n_i	9	10	8	9	9	10	8	9

Déterminer des intervalles de confiance (à environ 95%) de la biomasse moyenne par parcelle et par compartiment, et de son total dans la forêt.

Solution: Puisque l'on ne connaît pas \bar{N} , on l'approxime à l'aide de la moyenne

$$\bar{N} \approx \frac{1}{8}(1760 + \cdots + 1865) = 1813.125.$$

On calcule ensuite les totaux dans les unités primaires sélectionnées:

Comp.	1	2	3	4	5	6	7	8
$N_i \bar{y}_i (\times 10^5)$	2.077	2.113	1.760	1.964	2.130	1.946	1.562	1.679

et les estimateurs EAS2D de la moyenne μ , de la moyenne des totaux dans les compartiments, et du total sont:

$$\bar{y}_{\text{EAS2D}} = \frac{1}{8(1813.125)}(2.077 + \cdots + 1.679) \times 10^5 = 105.01;$$

$$\bar{N} \bar{y}_{\text{EAS2D}} = 1813.125 \cdot 105.01 = 190,403.75; \quad \tau_{\text{EAS2D}} = M \cdot \bar{N} \bar{y}_{\text{EAS2D}} = 7,616,150.$$

La variance entre les compartiments (unités primaires) est ainsi:

$$s_T^2 = \frac{1}{8-1} \sum_{i=1}^8 (N_i \bar{y}_i - 190,403.75)^2 = 4.55 \times 10^8$$

Finalement, on calcule la variance à même les compartiments:

Comp.	1	2	3	4	5	6	7	8
$\frac{N_i^2}{\bar{N}^2} \cdot \frac{s_i^2}{n_i} \left(1 - \frac{n_i}{N_i}\right)$	48.2	51.3	72.7	50.4	45.6	49.4	93.9	54.9

La variance d'échantillonnage devient alors

$$\hat{V}(\bar{y}_{\text{EAS2D}}) = \frac{4.55 \times 10^8}{8(1813.125)^2} \left(1 - \frac{8}{40}\right) + \frac{1}{8(40)} (48.2 + \dots + 54.9) = 14.03$$

Les variances des deux autres estimateurs se calculent aisément:

$$\hat{V}(\overline{N}\bar{y}_{\text{EAS2D}}) = \overline{N}^2 \hat{V}(\bar{y}_{\text{EAS2D}}) = (1813.125)^2 \cdot 14.03 = 46,141,324.55;$$

$$\hat{V}(\tau_{\text{EAS2D}}) = M^2 \overline{N}^2 \hat{V}(\bar{y}_{\text{EAS2D}}) = (40)^2 \cdot (1813.125)^2 \cdot 14.03 = 73,826,119,284;$$

tout comme les intervalles de confiance:

$$\text{IC}_{\text{EAS2D}}(\mu; 0.95) : 105.01 \pm 2\sqrt{14.03} \equiv (97.5, 112.5)$$

$$\text{IC}_{\text{EAS2D}}\left(\frac{N_0}{M}\mu; 0.95\right) : 190,403.75 \pm 2\sqrt{46,141,324.55} \equiv (176818, 203989.2312)$$

$$\text{IC}_{\text{EAS2D}}(\tau; 0.95) : 7,616,150 \pm 2\sqrt{73,826,119,284} \equiv (7072730, 8159569)$$

... en supposant bien sûr que le théorème de la limite centrée demeure valide dans le contexte d'un EAS2D.

8.3 – Échantillonnage à plusieurs phases

L'**échantillonnage à plusieurs phases** (EnP) joue un rôle crucial dans plusieurs types d'enquêtes, incluant entre autre les enquêtes à distance, telle que celles menées par **télé-détection**.

Lors de la première phase, on prélève un nombre **élevé** d'unités, mais on ne capte qu'un **petit** nombre de caractéristiques pour chaque unité.

Dans chaque phase successive, on mesure un plus **grand** nombre de caractéristique sur un plus **petit (sous-)**échantillon d'unités.

De cette façon, on arrive à estimer le paramètre visé avec **plus de précision** et à un **plus faible coût**, en étudiant la relation entre les caractéristiques mesurées lors des différentes phases.

8.3.1 – Échantillonnage aléatoire simple à deux phases

Un EnP qui ne comporte que deux phases prend le nom d'**échantillonnage à deux phases** (E2P), ou **échantillonnage double**.

Les E2P sont particulièrement utiles dans une situation où l'énumération de la **caractéristique étudiée** (ou **caractère principal**) est dispendieuse (en \$\$\$ ou en main d'oeuvre), mais dans laquelle on peut aisément observer une **caractéristique auxiliaire** corrélée au caractère principal.

Il est ainsi parfois préférable de prélever, en **première phase**, un EAS de **grande taille** afin d'analyser la variable auxiliaire, ce qui mène à des estimations précises (tout du moins, c'est ce que l'on espère) de τ ou de μ pour cette variable auxiliaire.

Lors de la seconde phase, on choisit un **plus petit** échantillon, généralement un **sous-échantillon de l'échantillon obtenu lors de la première phase**, dans lequel on mesure la **caractéristique principal** et la **variable auxiliaire**.

On obtient ensuite des estimations de la caractéristique principale à l'aide des renseignements obtenus lors de la **première phase**, en utilisant la **méthode du quotient** ou la **méthode de la régression**.

On peut augmenter la précision des estimations finales en incluant **plusieurs variables auxiliaires corrélées**, au lieu d'une seule.

Exemple: Si l'on cherche à estimer le volume total τ d'une forêt, on commence par mesurer la circonférence c_i et la hauteur h_i des arbres i dans un échantillon, puis le volume v_{i_k} des arbres i_k dans un sous-échantillon. On détermine ensuite la relation statistique entre τ_v , τ_c , et τ_h , et voilà!

Le mode d'échantillonnage EnP aide à réduire le **coût des énumérations** et à accroître la **précision des estimations**.

On peut aussi s'en servir afin de **stratifier** une population: un premier échantillon est prélevé en se fondant sur la caractéristique auxiliaire, que l'on utilise pour subdiviser la population en strates dans lesquelles la caractéristique principale est plus ou moins **homogène**.

Tant que les deux caractéristiques sont **corrélées**, on obtient ainsi des estimations précises de la caractéristique principale à partir d'un deuxième échantillon relativement petit.

On peut aussi jumeler le mode $E2P$ avec le mode $E2D$, par exemple (ou avec n'importe quel mode d'échantillonnage).

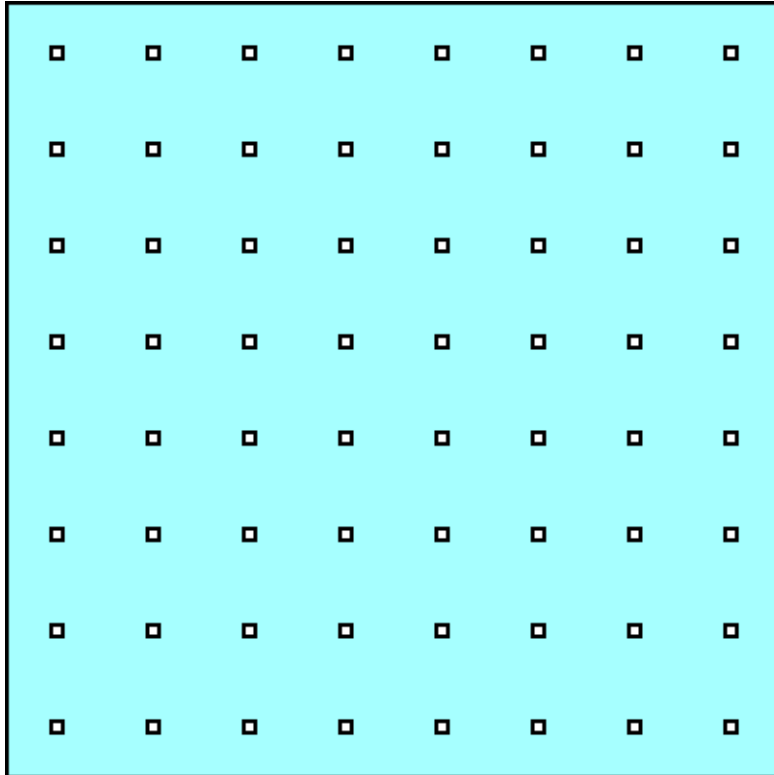
Si les deux étapes de sélection se font par EAS, la méthode prend le nom d'**échantillonnage aléatoire simple à deux phases** (EAS2P).

Lors de la première phase, la population est divisée en unités d'échantillonnage bien définies et on y prélève un EAS \mathcal{Y}_1 de taille n_1 ; on mesure la **variable auxiliaire** x sur toutes les unités de \mathcal{Y}_1 .

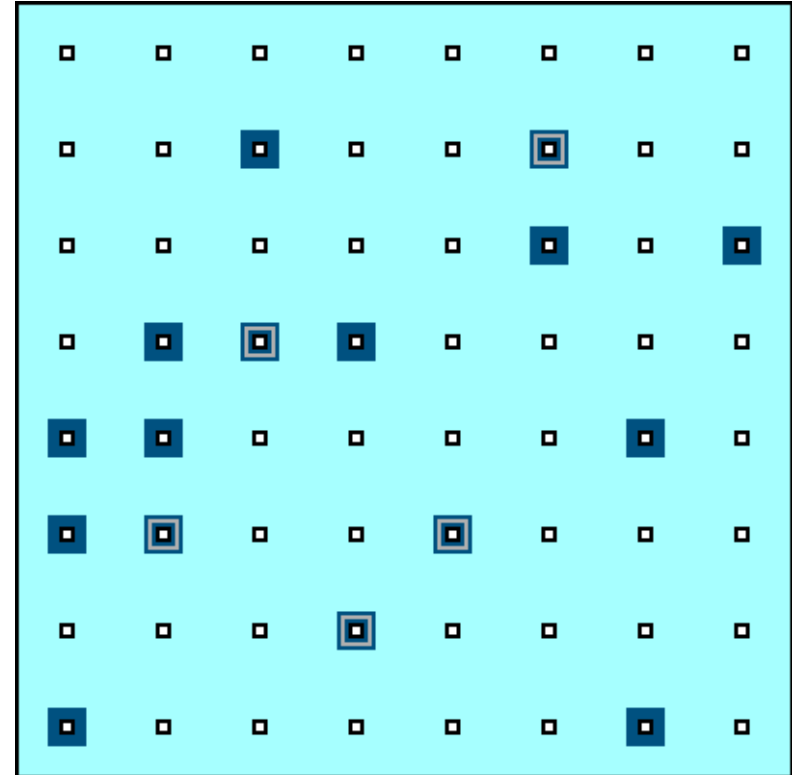
Ensuite, on prélève un sous-EAS $\mathcal{Y}_2 \subseteq \mathcal{Y}_1$ de taille n_2 ; on mesure la **caractéristique principale** y sur toutes les unités de \mathcal{Y}_2 .

On évalue les paramètres $r_{\mathcal{Y}_2}$ ou $b_{\mathcal{Y}_2}$ à partir de \mathcal{Y}_2 (à l'aide de la méthode du quotient ou de la méthode de la régression), ce qui donne

$$\hat{\mu}_{Y;R;EAS2P} = r_{\mathcal{Y}_2} \cdot \bar{x}_{\mathcal{Y}_1} \quad \text{ou} \quad \hat{\mu}_{Y;L;EAS2P} = \bar{y}_{\mathcal{Y}_2} + b_{\mathcal{Y}_2}(\bar{x}_{\mathcal{Y}_1} - \bar{x}_{\mathcal{Y}_2}).$$



Population



EAS à 2 phases

8.3.2 – Estimation de la variance d'échantillonnage

En raison du **double échantillonnage**, on retrouve deux termes qui contribuent à chacune des variances d'échantillonnage des estimateurs (la première lorsque l'on passe de \mathcal{U} à \mathcal{Y}_1 , et la seconde de \mathcal{Y}_1 à \mathcal{Y}_2):

$$\hat{V}(\hat{\mu}_{Y;R;\text{EAS2P}}) = \frac{1}{n_2}(s_Y^2 - 2r_{\mathcal{Y}_2}s_{XY} + (r_{\mathcal{Y}_2})^2s_X^2) + \frac{1}{n_1}(2r_{\mathcal{Y}_2}s_{XY} - (r_{\mathcal{Y}_2})^2s_X^2)$$

$$\hat{V}(\hat{\mu}_{Y;L;\text{EAS2P}}) = \frac{1}{n_2}s_{XY;L}^2 + \frac{1}{n_1}(s_{XY;L}^2 - s_Y^2)$$

où s_Y^2 , s_{XY} et s_X^2 représentent les quantités habituelles (dans \mathcal{Y}_2),

$$r_{\mathcal{Y}_2} = \frac{\bar{y}_{\mathcal{Y}_2}}{\bar{x}_{\mathcal{Y}_2}}, \quad b_{\mathcal{Y}_2} = \frac{s_{XY}}{s_X^2}, \quad \text{et} \quad s_{XY;L}^2 = \frac{n_2 - 1}{n_2 - 2} \cdot \{s_Y^2 - b_{\mathcal{Y}_2}^2 s_X^2\}.$$

Exemple:

On s'intéresse à la biomasse d'une plante quelconque dans une région, divisée en parcelles de 0.025 ha chacune.

En premier lieu, on mesure le nombre de bosquets x par unité dans un EAS \mathcal{Y}_1 de $n_1 = 200$ parcelles. Ensuite, on calcule la biomasse y de la plante en question dans chaque unité d'un sous-EAS \mathcal{Y}_2 de $n_2 = 40$ parcelles:

$$\bar{x}_{\mathcal{Y}_1} = 374.4; \quad \sum_{i=1}^{40} x_i = 15,419; \quad \sum_{i=1}^{40} y_i = 2104;$$

$$\sum_{i=1}^{40} x_i^2 = 7,744,481; \quad \sum_{i=1}^{40} x_i y_i = 960,320; \quad \sum_{i=1}^{40} y_i^2 = 125,346.$$

Donner un I.C. de la biomasse moyenne par parcelle, à environ 95%.

Solution: on calcule les quantités intermédiaires requises:

$$\bar{x}_{Y_2} = \frac{15419}{40} = 385.5; \quad \bar{y}_{Y_2} = \frac{2104}{40} = 52.6; \quad r_{Y_2} = \frac{\bar{y}_{Y_2}}{\bar{x}_{Y_2}} = \frac{52.6}{385.5} = 0.14;$$

$$s_X^2 = \frac{1}{39}[7744481 - 40(385.5)^2] \approx 46175; \quad s_Y^2 = \frac{1}{39}[125346 - 40(52.6)^2] \approx 376$$

$$s_{XY} = \frac{1}{39}[960320 - 40(385.5)(52.6)] \approx 3827.7; \quad b_{Y_2} = \frac{s_{XY}}{s_X^2} = \frac{3827.7}{46175.4} \approx 0.08;$$

$$s_{XY;L}^2 = \frac{39}{38}[376.3 - 0.08^2(46175.4)] \approx 82.9;$$

ce qui donne

$$\hat{\mu}_{Y;R;EAS2P} = 0.14(374.4) \approx 51.1; \quad \hat{\mu}_{Y;L;EAS2P} = 52.6 + 0.08(374.4 - 385.5) \approx 51.7$$

et

$$\hat{V}(\hat{\mu}_{Y;R;\text{EAS2P}}) = \frac{376.3 - 2(0.14)(3827.7) + (0.14)^2 46175.4}{40} \\ + \frac{2(0.14)3827.7 - (0.14)^2 46175.4}{200} \approx 5.67;$$

$$\hat{V}(\hat{\mu}_{Y;L;\text{EAS2P}}) = \frac{82.9}{40} + \frac{82.9 - 376.3}{200} \approx 3.54;$$

d'où

$$\text{IC}_{R;\text{EAS2P}}(\mu_Y; 0.95) = 51.1 \pm 2\sqrt{5.67} \equiv (46.3, 55.8)$$

$$\text{IC}_{L;\text{EAS2P}}(\mu_Y; 0.95) = 51.7 \pm 2\sqrt{3.54} \equiv (47.9, 55.5).$$

8.4 – Méli-mélo

Nous terminons le cours en discutant brièvement de quelques notions qui n'ont pas trouvé de place naturelle dans les sections précédentes:

- les effets de plan;
- l'ajustement pour la non-réponse;
- l'estimation de la taille d'une population,
- la méthode de la réponse aléatoire, et
- l'échantillonnage de Bernoulli.

8.4.1 – Effet de plan

[Adapté de *Méthodes et pratiques d'enquête*, Statistique Canada]

L'**effet de plan** compare la variance des estimateurs entre un plan d'échantillonnage et un EAS. Il s'agit du rapport entre la **variance d'échantillonnage d'un estimateur selon un plan d'échantillonnage donné**, et la **variance d'échantillonnage de l'estimateur d'un EAS** (provenant d'un échantillon de **même taille**).

Cette mesure est souvent appliquée pour comparer l'**efficacité** des estimateurs de divers plans d'échantillonnage. Si le rapport < 1 , le plan d'échantillonnage est plus efficace que l'EAS; s'il est > 1 , il est moins efficace que l'EAS.

Nous avons comparé directement les variances théoriques de plusieurs plan d'échantillonnage aux sections 3.4, 4.5, et 6.4 – typiquement, on calcule l'effet de plan à l'aide des échantillons réalisés.

Les effets du plan d'échantillonnage aident aussi à obtenir des estimations approximatives de la variance pour des plans d'**échantillonnage complexes**.

Si une estimation de l'effet du plan d'échantillonnage est disponible dans une enquête précédente qui a utilisé le même plan d'échantillonnage, elle peut servir à déterminer la **taille de l'échantillon nécessaire de l'enquête**.

8.4.2 – Ajustement pour la non-réponse

[*Ibid.*] Les non-réponses représentent un problème dans **toutes** les enquêtes.

La **non-réponse totale** (lorsque toutes les données ou presque d'une unité échantillonnée sont manquantes) survient lorsque:

- une unité de l'échantillon **refuse de participer** au sondage;
- il est impossible d'**établir le contact avec une unité de l'échantillon**;
- l'unité ne peut être **repérée**, ou encore
- si l'information obtenue est **inutile**.

La façon la plus simple de traiter ces non-réponses est de les ignorer; dans certaines circonstances **exceptionnelles**, des proportions ou des moyennes estimées sans ajustement pour les non-réponses totales sont **plus ou moins identiques** à celles produites en appliquant un ajustement pour les non-réponses.

Si l'on néglige de **compenser** pour les unités non répondantes, les **totaux sont généralement sous-estimés** (e.g. la taille d'une population, le total des revenus ou le total d'acres récoltés).

La façon la plus commune de traiter la non-réponse totale est d'**ajuster les poids de base** en supposant que les unités répondantes représentent les unités répondantes et non répondantes.

(Est-ce un hypothèse vraisemblable, en pratique?)

Si les **non-répondants** sont **équivalents aux répondants** pour les caractéristiques mesurées dans l'enquête, c'est une approche raisonnable.

Les poids de base pour les non-répondants sont ensuite redistribuées entre les répondants, à l'aide d'un **facteur d'ajustement pour les non-réponses** qui est multiplié par la poids de base, afin d'obtenir une pondération ajustée.

Par exemple, si on prélève un EAS de taille $n = 25$ d'une strate de taille $N = 1000$, la **probabilité d'inclusion** de chacune de ces unités et le **poids de base** correspondant sont

$$\pi = \frac{n}{N} = \frac{25}{1000} = 0.025$$
$$w = \frac{1}{\pi} = \frac{1}{0.025} = 40.$$

Chaque unité sélectionnée représente 40 unités dans la strate.

Si nous n'obtenons une réponse que de $n_r = 20$ des $n = 25$ unités sélectionnées, le **facteur d'ajustement pour les non-réponses** et la **pondération ajustée** (pour la non-réponse) deviennent:

$$FANR = \frac{n}{n_r} = \frac{25}{20} = 1.25$$

$$w_{nr} = w \cdot FANR = 1.25(40) = 50;$$

chaque unité répondante représente alors 50 unités dans la strate. C'est avec cette pondération ajustée que l'on travaillerait.

Il va de soit que la pondération ajustée peut varier d'une strate à l'autre, en fonction du plan d'échantillonnage et de la taille de l'échantillon.

Lorsque l'on cherche à déterminer la taille de l'échantillon et sa répartition dans diverses strates, on obtient en pratique la taille de l'**échantillon visé** (on suppose ici que les populations cible et à l'étude coïncident). On peut avoir alors recours à l'**inflation** de la taille de l'échantillon.

Exemple: on détermine que la répartition d'un STR de taille $n = 29$ est $(17, 9, 3)$. Lors d'une étude préalable, on a déterminé que les taux de non-réponse par strate sont de $(16.2\%, 20.8\%, 31.2\%)$. Quelle répartition optimise les chances d'obtenir la répartition visée?

Solution: il suffit de résoudre

$$n_1(1 - 0.162) = 17, \quad n_2(1 - 0.208) = 9, \quad n_3(1 - 0.312) = 3,$$

c'est-à-dire $(n_1, n_2, n_3) = (20.3, 11.3, 4.3) \approx (21, 12, 5)$.

8.4.3 – Estimation de la taille d'une population

Comment s'y prend-on si la taille N de la population \mathcal{U} est inconnue? Lorsque la population est suffisamment large, on peut toujours utiliser l'approximation $N \approx \infty$ dans les formules de variance d'échantillonnage.

Mais c'est parfois le paramètre N qui représente la quantité d'intérêt.

Exemple: combien de billets de 5 dollars, N , y a-t-il en circulation?

On donne un estimé de N à l'aide de la méthode de **remise en circulation**:

1. on capture n_1 billets au hasard (sans remise) dans la population;
2. on les marque et on les remet en circulation;
3. à un moment ultérieur, on capture n_2 billets au hasard (sans remise) dans la population;
4. on compte le nombre X de billets marqués, $0 < X \leq n_2$.

Si on attend assez longtemps (question de laisser les billets marqués se propager dans la population), on obtient

$$\frac{n_1}{N} \approx \frac{X}{n_2}, \quad \text{d'où } \hat{N} = \frac{n_1 n_2}{X},$$

où $X \sim$ loi **hypergéométrique** dont les paramètres sont $n_1, N - n_1, n_2$, et

$$P(X = x) = \frac{\binom{n_1}{x} \binom{N - n_1}{n_2 - x}}{\binom{N}{n_2}}, \quad 0 \leq x \leq n_2$$

$$\mu_X = E[X] = n_2 \underbrace{\left(\frac{n_1}{N} \right)}_p = n_2 p, \quad \sigma_X^2 = V[X] = n_2 p (1 - p) \left(\frac{N - n_2}{N - 1} \right).$$

Si $\frac{n_2}{N} < 0.05$, on peut ignorer le FCPF dans la variance:

$$\sigma_X^2 = V[X] \approx n_2 p(1 - p).$$

On peut maintenant développer des expressions pour $E[\hat{N}]$ et $V[\hat{N}]$, en se servant de la **série de Taylor d'ordre 2 près de** $X \approx \mu_X = n_2 p$:

$$f(X) \approx f(\mu_X) + f'(\mu_X)(X - \mu_X) + \frac{f''(\mu_X)}{2}(X - \mu_X)^2.$$

Si $\hat{N} = f(X) = \frac{n_1 n_2}{X}$, alors

$$\begin{aligned}\hat{N} &\approx \frac{n_1 n_2}{\mu_X} - \frac{n_1 n_2}{\mu_X^2}(X - \mu_X) + \frac{n_1 n_2}{\mu_X^3}(X - \mu_X)^2 \\ &= \frac{n_1}{p} - \frac{n_1}{n_2 p^2}(X - n_2 p) + \frac{n_1}{n_2^2 p^3}(X - n_2 p)^3.\end{aligned}$$

Conséquemment,

$$\begin{aligned} E[\hat{N}] &= E \left[\frac{n_1 n_2}{\mu_X} - \frac{n_1 n_2}{\mu_X^2} (X - \mu_X) + \frac{n_1 n_2}{\mu_X^3} (X - \mu_X)^2 \right] \\ &= E \left[\frac{n_1 n_2}{\mu_X} \right] - E \left[\frac{n_1 n_2}{\mu_X^2} (X - \mu_X) \right] + E \left[\frac{n_1 n_2}{\mu_X^3} (X - \mu_X)^2 \right] \\ &= \frac{n_1 n_2}{\mu_X} - \frac{n_1 n_2}{\mu_X^2} (\underbrace{E[X]}_{\mu_X} - \mu_X) + \frac{n_1 n_2}{\mu_X^3} E[(X - \mu_X)^2] \\ &= \frac{n_1 n_2}{\mu_X} + \frac{n_1 n_2}{\mu_X^3} V[X] \approx \frac{n_1}{p} + \frac{n_1}{n_2^2 p^3} \cdot n_2 p(1 - p) = \frac{n_1}{p} + \frac{n_1}{n_2 p^2} (1 - p) \\ &= \frac{n_1}{p} \left(1 + \frac{1 - p}{n_2 p} \right) = N \left(1 + \frac{1 - p}{n_2 p} \right). \end{aligned}$$

Puisque $\frac{1-p}{n_2p} > 0$, $E[\hat{N}] \neq N$ et l'estimateur \hat{N} est **asymptotiquement non-biaisé** lorsque la taille n_2 du second échantillon augmente.

On obtient un estimateur de la variance en utilisant de la **série de Taylor d'ordre 1 près de** $X \approx \mu_X = n_2p$:

$$\hat{N} \approx \frac{n_1 n_2}{\mu_X} - \frac{n_1 n_2}{\mu_X^2} (X - \mu_X) = \frac{n_1}{p} \left(1 - \frac{X - n_2 p}{n_2 p} \right) = \frac{n_1}{p} \left(2 - \frac{X}{n_2 p} \right).$$

Dans ce cas,

$$\begin{aligned} V[\hat{N}] &\approx V \left[\frac{n_1}{p} \left(2 - \frac{X}{n_2 p} \right) \right] = \frac{n_1^2}{p^2} \cdot V \left[-\frac{X}{n_2 p} \right] = \frac{n_1^2}{n_2^2 p^4} \cdot V[X] \\ &\approx \frac{n_1^2 n_2 p (1-p)}{n_2^2 p^4} = \frac{n_1^2 (1-p)}{n_2 p^3}. \end{aligned}$$

En pratique, on ne connaît pas p ; on utilise alors

$$\hat{V}[\hat{N}] = \frac{n_1^2(1 - \hat{p})}{n_2\hat{p}^3}, \quad \text{où } \hat{p} = \frac{X}{n_2}.$$

Théorème de la limite centrée – taille de la population N

Si n_2 et N sont suffisamment élevés, alors

$$\hat{N} \sim_{\text{approx.}} \mathcal{N} \left(\mathbb{E}[\hat{N}], \hat{V}[\hat{N}] \right) \approx \mathcal{N} \left(\frac{n_1 n_2}{X}, \frac{n_1^2(1 - \hat{p})}{n_2 \hat{p}^3} \right),$$

et l'intervalle de confiance de N à environ **95%** est ainsi

$$\text{IC}(N; 0.95) : \quad \frac{n_1 n_2}{X} \pm 2 \sqrt{\frac{n_1^2(1 - \hat{p})}{n_2 \hat{p}^3}}.$$

Exemple: supposons que $n_1 = 500$ billets aient été capturés et marqués initialement; des $n_2 = 300$ billets recapturés à la 2e étape, $X = 127$ étaient marqués. Donner un intervalle de confiance du nombre total de billets de 5\$ à environ 95%.

Solution: on calcule l'estimé ponctuel à l'aide de $\hat{N} = \frac{500 \cdot 300}{127} \approx 1181.102$.
De plus, $\hat{p} = \frac{X}{n_2} = \frac{127}{300} \approx 0.423$, d'où

$$2\sqrt{\hat{V}(\hat{N})} = 2\sqrt{\frac{500^2 \cdot (1 - 0.42)}{300 \cdot (0.42)^3}} = 159.176,$$

d'où

$$\text{IC}(N; 0.95) : \quad 1181.102 \pm 159.176 \equiv (1021.9, 1340.3).$$

8.4.4 – Réponse aléatoire

Avez-vous déjà triché lors d'un contrôle durant la pandémie?

Avec un “**Oui**”, on peut vraisemblablement conclure que c'est la vérité.

Mais puisqu'il y a un **coût social** associé à une telle réponse, on peut s'attendre à ce que certains tricheurs répondent “**Non**”.

Comment peut-on s'y prendre afin de réduire l'erreur de mesure pour les **questions délicates**?

Première approche: avec de telles questions, la compétence de l'enquêteur.e joue un rôle crucial – il ne faut pas négliger ce volet.

Seconde approche: la technique de la **réponse aléatoire** nécessite l'utilisation de deux questions:

- la question **délicate**, et
- un question **innocente**,

et d'un **mécanisme aléatoire** à **paramètres connus** (pile ou face, etc.).

Le principe est le suivant: la répondante tire à pile ou face (sans annoncer le résultat à l'enquêteur), et elle répond honnêtement à une des 2 questions:

- **“face”**: “Avez-vous déjà triché lors d'un contrôle?”;
- **“pile”**: “Êtes-vous née en janvier?”;

Puisque l'enquêteur ne connaît pas le résultat du tirage au sort, il ne sait pas si la répondante répond à la question délicate ou à la question innocente.

En théorie, l'anonymat assuré par la réponse aléatoire libère les répondants (le coût social est **diminué, voir éliminé**) – conséquemment, on peut s'attendre à une réponse honnête, quelle que soit la question.

 Cette approche ne peut porter fruit que si l'on connaît les probabilités:

- θ d'observer une réponse positive à la question innocente;
- ρ de poser la question délicate, et
- ϕ d'observer une réponse positive, quelle que soit la question.

Soit p la **proportion de réponses positives à la question délicate** (**quantité recherchée**).

Ainsi,

$$\begin{aligned}\phi &= P(\text{réponse positive}) \\ &= \underbrace{P(\text{positive} \mid \text{délicate})}_p \underbrace{P(\text{délicate})}_\rho + \underbrace{P(\text{positive} \mid \text{innocente})}_\theta \underbrace{P(\text{innocente})}_{1-\rho}, \\ &= p\rho + \theta(1 - \rho)\end{aligned}$$

d'où

$$p = \frac{\phi - \theta(1 - \rho)}{\rho}.$$

Si $\hat{\phi}$ représente la proportion de réponses positives dans l'échantillon réalisé, on peut construire l'**estimateur**

$$\hat{p}_{\text{ra}} = \frac{\hat{\phi} - \theta(1 - \rho)}{\rho}, \quad \theta, \rho \text{ des constantes,}$$

dont la variance est

$$V(\hat{p}_{\text{ra}}) = V\left(\frac{\hat{\phi} - \theta(1 - \rho)}{\rho}\right) = V\left(\frac{\hat{\phi}}{\rho}\right) = \frac{1}{\rho^2} \cdot V(\hat{\phi}).$$

Puisque $\hat{\phi}$ est l'estimateur d'une proportion dans une population \mathcal{U} de taille N , obtenu à l'aide d'un EAS de taille n , sa **variance d'échantillonnage** est

$$V(\hat{\phi}) = \frac{\phi(1 - \phi)}{n} \left(\frac{N - n}{N - 1} \right),$$

d'où

$$V(\hat{p}_{ra}) = \frac{1}{\rho^2} \cdot \frac{\phi(1-\phi)}{n} \left(\frac{N-n}{N-1} \right).$$

Puisque ϕ est inconnu en général, on utilise l'estimateur (non-biaisé)

$$\hat{V}(\hat{p}_{ra}) = \frac{1}{\rho^2} \cdot \frac{\hat{\phi}(1-\hat{\phi})}{n-1} \left(1 - \frac{n}{N} \right),$$

et on construit un **intervalle de confiance de p à environ 95%** avec

$$IC_{ra}(p; 0.95) : \quad \hat{p}_{ra} \pm 2\sqrt{\hat{V}(\hat{p}_{ra})}.$$

Le facteur $1/\rho^2$ vient **pénaliser l'incertitude** apportée par la réponse aléatoire – plus ρ est élevé, plus $\hat{V}(\hat{p}_{ra})$ est faible. Mais si ρ est trop élevé, l'anonymat conféré par l'approche s'évapore...

Exemple: on cherche à déterminer l'incidence de tricherie chez les étudiants ($N = 442$) du département de mathématiques et de statistique lors des cours en ligne offerts pendant la pandémie, à l'aide d'un EAS ($n = 65$). On se sert du stratagème décrit dans cette section avec $\rho = 1/2$, et on observe $\theta = \frac{52}{442}$ et $\hat{\phi} = \frac{21}{65}$. Déterminer un intervalle de confiance de la proportion des étudiants qui ont triché pendant la pandémie.

Solution: il suffit de calculer

$$\hat{p}_{ra} = \frac{21/65 - 52/442(1 - 1/2)}{1/2} = 0.53$$

$$\hat{V}(\hat{p}_{ra}) = \frac{1}{1/2^2} \cdot \frac{21/65(1 - 21/65)}{65 - 1} \left(1 - \frac{65}{442}\right) = 0.012,$$

d'où $IC_{ra}(p; 0.95) = 0.53 \pm 2\sqrt{0.012} \equiv (0.31, 0.74)$.

8.4.5 – Échantillonnage de Bernoulli

[Adapté des notes de cours de D. Haziza]

L'**échantillonnage de Bernoulli** (BE) est un plan de sondage à taille **aléatoire** – il est impossible de fixer la **taille de l'échantillon a priori**.

On assigne à chaque unité de la population $\mathcal{U} = \{u_1, \dots, u_N\}$ la même probabilité d'inclusion dans l'échantillon \mathcal{Y} : $\pi_j = \pi \in (0, 1)$, pour tout j .

On dénote la **taille de l'échantillon obtenu** par n_a .

Le plan BE consiste à effectuer N épreuves de Bernoulli indépendantes, chacune avec probabilité de succès π (succès: **unité incluse**; échec: **unité rejetée**).

La probabilité d'obtenir un échantillon \mathcal{Y} de taille n_a devient alors

$$P(|\mathcal{Y}| = n_a) = \pi^{n_a}(1 - \pi)^{N-n_a}.$$

Il y a 2^N échantillons possibles, dont la taille varie de $n_a = 0$ à $n_a = N$.

La taille de l'échantillon suit une **loi binomiale** $n_a \sim B(N, \pi)$:

$$P(n_a = n) = \binom{N}{n} \pi^n (1 - \pi)^{N-n}, \quad E[n_a] = N\pi, \quad V[n_a] = N\pi(1 - \pi).$$

Lorsque N est suffisamment élevé, cette loi est **approximativement normale**; on peut alors construire un **intervalle de confiance de la taille de l'échantillon à environ 95%** à l'aide de

$$\text{IC}(n_a; 0.95) : \quad N\pi \pm 2\sqrt{N\pi(1 - \pi)}.$$

Soit $\pi_{j,k}$ la probabilité d'inclusion des unités $j \neq k$ dans l'échantillon \mathcal{Y} . Puisque les épreuves de Bernoulli sont indépendantes les unes des autres,

$$\pi_{j,k} = P(\{u_j, u_k\} \in \mathcal{Y}) = P(u_j \in \mathcal{Y}) \cdot P(u_k \in \mathcal{Y}) = \pi_j \pi_k = \pi^2.$$

L'estimateur

$$\hat{\tau}_{\text{BE}} = \frac{1}{\pi} \sum_{i=1}^{n_a} y_i$$

est un **estimateur sans biais du total** τ : en effet

$$\mathbb{E}[\hat{\tau}_{\text{BE}}] = \frac{1}{\pi} \mathbb{E}[n_a \bar{y}] = \frac{\mathbb{E}[n_a] \mathbb{E}[\bar{y}]}{\pi} = \frac{N\pi\mu}{\pi} = N\mu = \tau,$$

puisque n_a et \bar{y} sont indépendants.

Dans le même ordre d'idée, la **variance d'échantillonnage de l'estimateur** $\hat{\tau}_{\text{BE}}$ peut être approchée par

$$\hat{V}[\hat{\tau}_{\text{BE}}] = \frac{1}{\pi} \left(\frac{1}{\pi} - 1 \right) \sum_{i=1}^{n_a} y_i^2.$$

Si N et n_a sont suffisamment élevés, le théorème de la limite centrée entre de nouveau en jeu, et on peut construire un **intervalle de confiance de τ à environ 95%** à l'aide de

$$\text{IC}_{\text{BE}}(\tau; 0.95) : \quad \hat{\tau}_{\text{BE}} \pm 2\sqrt{\hat{V}[\hat{\tau}_{\text{BE}}]}.$$

Les estimateurs correspondants pour la moyenne \bar{y}_{BE} et la proportion \hat{p}_{BE} s'obtiennent de la manière habituelle.

Exemple: Une professeure doit corriger 600 copies d'examen. Pour chaque copie, elle lance un dé ne le corrige que s'il montre un 6. À la fin du processus, elle a corrigé 90 copies, desquelles 60 obtiennent une note de passage. Déterminer un IC à 95% du nombre total de succès dans la classe.

Solution: soit $y_i = 1$ si le i -ème examen corrigé est un succès, et $y_i = 0$ autrement. Nous avons $N = 600$, $\pi = 1/6$, $n_a = 90$,

$$\sum_{i=1}^{90} y_i = 60, \quad \sum_{i=1}^{90} y_i^2 = 60, \quad \hat{\tau}_{\text{BE}} = \frac{1}{1/6} \sum_{i=1}^{90} y_i = 6(60) = 360$$

$$\hat{V}[\hat{\tau}_{\text{BE}}] = \frac{1}{1/6} \left(\frac{1}{1/6} - 1 \right) \sum_{i=1}^{90} y_i^2 = 6(5)(60) = 1800$$

$$\implies \text{IC}_{\text{BE}}(\tau; 0.95) = 360 \pm 2\sqrt{1800} \equiv [277, 443].$$