Annales d'examens de Théorie des Sondages

Guillaume Chauvet
Ensai

Examen 2023-2024

Exercice 1 (10 points)

Nous nous intéressons à l'estimation de la proportion d'individus utilisant les transports publics dans une commune contenant $N = 10\ 000$ individus.
Nous sélectionnons un échantillon S de taille n par sondage aléatoire simple sans remise. Nous savons que la proportion dans une commune comparable vaut 15%.

1) Définir l'unité statistique, la variable d'intérêt, le paramètre d'intérêt.
2) Donner l'estimateur $\hat{p}$ du paramètre, et un estimateur de variance sans biais en fonction de $\hat{p}$.
3) Quelle est la taille d'échantillon minimum que nous devons sélectionner pour que la proportion soit connue à 10% près ?

4) Nous sélectionnons un échantillon de taille $n = 200$, dans lequel 40 personne utilisent les transports en commun. Donner une estimation sans biais de la proportion de personnes utilisant les transports en commun, et une estimation du coefficient de variation.

5) Nous nous intéressons à la proportion d'hommes utilisant les transports en commun. Ecrire le paramètre comme une fonction de totaux, en spécifiant les variables d'intérêt utilisées (pas d'application numérique demandée).

Correction Exercice 1

1) Unité statistique : un individu.
Variable d'intérêt : y_k égale à 1 si k prend les transports publics, et 0 sinon.
Paramètre d'intérêt : la proportion $P = \frac{1}{N}\sum_{k\in U} y_k$.

2) Sous un sondage aléatoire simple :

$$\hat{p} = \frac{1}{n}\sum_{k\in S} y_k \quad \text{et} \quad \hat{V}_{HT}(\hat{p}) = \frac{1-f}{n-1}\hat{p}(1-\hat{p}).$$

3) En utilisant la formule vue en cours (qui peut également être rapidement redémontrée), nous utilisons

$$n \geq \frac{1}{\frac{1}{N} + \frac{N-1}{N}\left[\frac{0.10}{1.96}\right]^2 \frac{P}{1-P}}. \tag{1}$$

Avec $N = 10\ 000$, et en utilisant l'approximation $P = 0.15$, nous obtenons $n \geq 1\ 788$.

4) Nous obtenons successivement

$$\begin{aligned} \hat{p} &= \frac{1}{n}\sum_{k\in S} y_k = 0.20, \\ \hat{V}_{HT}(\hat{p}) &= \frac{1-f}{n-1}\hat{p}(1-\hat{p}) = 7.9\ 10^{-4}, \\ C\hat{V}_{HT}(\hat{p}) &= \frac{\sqrt{\hat{V}_{HT}(\hat{p})}}{\hat{p}} = 14\%. \end{aligned}$$

5) Le paramètre peut s'écrire sous la forme

$$R = \frac{t_z}{t_x},$$

où x_k vaut 1 si k est un homme et 0 sinon, et z_k vaut 1 si k est un homme et utilise les transports en commun, et 0 sinon.

Exercice 2 (10 points)

Nous souhaitons sélectionner un échantillon de taille $n = 4$ dans une population U de 8 entreprises, dont le nombre d'employés est donné dans le tableau suivant :

Entreprise k	1	2	3	4	5	6	7	8
Nombre d'employés x_k	400	200	100	100	60	60	40	40
Nombre aléatoire u_k	0.31	0.21	0.76	0.93	0.98	0.24	0.46	0.98

1) Donner les probabilités d'inclusion d'ordre 1 proportionnelles au nombre d'employés.
2) Dans la population triée comme indiqué dans le tableau, utiliser le nombre aléatoire $u = 0.68$ pour sélectionner un échantillon selon un tirage systématique.
3) Dans la population, utiliser les nombres aléatoires u_k du tableau (générés indépendamment selon une loi uniforme) pour sélectionner un échantillon selon un tirage de Poisson.

Nous supposons que l'échantillon, sélectionné selon un tirage de Poisson (avec des nombres aléatoires différents de ceux de la question 3), est $s = \{1, 2, 4, 5\}$. Nous obtenons les valeurs suivantes sur s pour le chiffres d'affaires (en millions d'euros) :

$$y_1 = 10, \qquad y_2 = 8, \qquad y_4 = 2, \qquad y_5 = 2.$$

4) Donner l'estimateur de Horvitz-Thompson de t_y, et calculer sa valeur.
5) Donner l'estimateur par substitution de μ_y, et calculer sa valeur.
6) Donner l'estimateur par le ratio de t_y, en utilisant la variable auxiliaire $z_k = 1$, et calculer sa valeur.

Correction Exercice 2

1) Après recalcul, nous obtenons :

Entreprise k	1	2	3	4	5	6	7	8
Nombre d'employés x_k	400	200	100	100	60	60	40	40
Probabilité d'inclusion π_k	1	1	0.5	0.5	0.3	0.3	0.2	0.2
Nombre aléatoire u_k	0.31	0.21	0.76	0.93	0.98	0.24	0.46	0.98

2) En utilisant $u = 0.68$, nous obtenons $s = \{1, 2, 4, 7\}$.
3) En utilisant les nombres aléatoires du tableau, nous obtenons $s = \{1, 2, 6\}$.

4)5)6) Nous obtenons successivement :

$$\begin{aligned}
\hat{t}_{y\pi} &= \sum_{k\in s} \frac{y_k}{\pi_k} \simeq 28.67, \\
\hat{N}_\pi &= \sum_{k\in s} \frac{1}{\pi_k} \simeq 7.33, \\
\hat{\mu}_{y\pi} &= \frac{\hat{t}_{y\pi}}{\hat{N}_\pi} \simeq 3.91, \\
\hat{t}_{yR} &= N\frac{\hat{t}_{y\pi}}{\hat{N}_\pi} \simeq 31.27.
\end{aligned}$$

Exercice 3 (14 points)

Rappel : modèle linéaire simple
Dans le cas d'un vecteur de $q = 2$ variables auxiliaires $\mathbf{x}_k = (1, x_{1k})^\top$ et avec homoscédasticité, le modèle de travail s'écrit :

$$y_k = \beta_0 + \beta_1 x_{1k} + \epsilon_k \text{ avec } \left\{ \begin{array}{ll} E_m(\epsilon_k) & = 0, \\ V_m(\epsilon_k) & = \sigma^2. \end{array} \right. \quad (2)$$

L'estimateur des MCG calculé sur la population donne :

$$\begin{aligned}
B_{1,MCG} &= \frac{\sum_{k\in U}(x_{1k} - \mu_{x1})(y_k - \mu_y)}{\sum_{k\in U}(x_{1k} - \mu_{x1})^2} = \frac{S_{xy}}{S_x^2}, \\
B_{0,MCG} &= \mu_y - B_{1,MCG} \times \mu_x, \\
E_k &= (y_k - \mu_y) - \frac{S_{xy}}{S_x^2}(x_{1k} - \mu_{x1}).
\end{aligned}$$

Nous nous intéressons à une population de $N = 600$ fermes laitières, pour lesquelles nous voulons estimer le volume total de lait produit pendant l'année 2023. Cette population contient $N_1 = 250$ fermes de petite taille, $N_2 = 300$ fermes de taille moyenne, et $N_3 = 50$ fermes de grande taille. Nous disposons d'une information auxiliaire, sous la forme de la variable x_k donnant le volume de lait produit par une ferme k en 2022 (en millions de litres). Le tableau

ci-dessous donne pour chaque strate le volume moyen de lait μ_{xh} en 2022, et la dispersion du volume de lait S^2_{xh}.

Strate	Taille N_h	μ_{xh}	S^2_{xh}
U_1	250	0.25	0.16
U_2	300	0.35	0.25
U_3	50	1.50	2.25

Nous souhaitons sélectionner un échantillon de $n = 80$ fermes.
1) Donner les tailles d'échantillon obtenues par strate avec une allocation proportionnelle.
2) Donner les tailles d'échantillon obtenues par strate avec une allocation optimale pour x_k.
3) Donner les tailles d'échantillon obtenues par strate avec une strate U_3 exhaustive, et une allocation proportionnelle dans les deux autres strates.

Nous sélectionnons finalement un échantillon avec l'allocation indiquée dans le tableau suivant. Le tableau contient également les informations collectées dans l'échantillon, avec y_k le volume de lait produit par une ferme k en 2023 (en millions de litres).

Strate	Taille d'échant. n_h	$\bar{x}_h$	$\bar{y}_h$	s^2_{yh}	s^2_{xh}	s_{xyh}
U_1	10	0.22	0.26	0.20	0.14	0.13
U_2	20	0.38	0.37	0.40	0.30	0.30
U_3	50	1.50	1.35	2.00	2.25	1.80

4) Donner les probabilités d'inclusion des unités et leur poids de sondage.
5) Donner l'estimateur de Horvitz-Thompson du total t_y ainsi qu'un intervalle de confiance à 95 %.

Nous réalisons un redressement **par la régression linéaire simple à l'intérieur de chaque strate**.

6) Ecrire le modèle de travail correspondant au redressement réalisé.
7) Montrer que l'estimateur GREG du total de la strate h se réécrit sous la forme

$$\hat{t}_{yh,greg} \quad = \quad N_h \left\{ \bar{y}_h + \frac{s_{xyh}}{s^2_{xh}} (\mu_{xh} - \bar{x}_h) \right\}.$$

En déduire la valeur de l'estimateur redressé du total t_y.

Correction Exercice 3

1) Nous obtenons

$$n_1 = 33.33, \quad n_2 = 40, \quad n_1 = 6.67,$$

soit après arrondi

$$n_1 = 33, \quad n_2 = 40, \quad n_1 = 7.$$

2) Nous obtenons

$$n_1 = 24.62, \quad n_2 = 36.92, \quad n_1 = 18.46,$$

soit après arrondi

$$n_1 = 25, \quad n_2 = 37, \quad n_1 = 18.$$

3) Nous obtenons

$$n_1 = 13.64, \quad n_2 = 16.36, \quad n_1 = 50,$$

soit après arrondi

$$n_1 = 14, \quad n_2 = 16, \quad n_1 = 50.$$

4) Nous obtenons

$$\pi_k = \begin{cases} \frac{1}{25} & \text{pour } k \in S_1, \\ \frac{1}{15} & \text{pour } k \in S_2, \\ 1 & \text{pour } k \in S_3, \end{cases} \quad \text{et} \quad d_k = \begin{cases} 25 & \text{pour } k \in S_1, \\ 15 & \text{pour } k \in S_2, \\ 1 & \text{pour } k \in S_3. \end{cases}$$

5) Nous obtenons

$$\begin{aligned} \hat{t}_{y\pi} &= \sum_{h=1}^{3} N_h \bar{y}_h = 243.5 \\ \hat{V}_{HT}(\hat{t}_{y\pi}) &= \sum_{h=1}^{3} (N_h)^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) s_{yh}^2 = 2\ 880 \\ \widehat{IC}(t_y) &= \left[\hat{t}_{y\pi} \pm 1.96 \sqrt{\hat{V}_{HT}(\hat{t}_{y\pi})} \right] = [243.5 \pm 105.2] = [138\ ,\ 349]\,. \end{aligned}$$

6) Le modèle de travail est

$$y_k \;=\; \beta_{0h} + \beta_{1h} x_k + \epsilon_k \text{ avec } \left\{ \begin{array}{ll} E_m(\epsilon_k) & = 0, \\ V_m(\epsilon_k) & = \sigma_h^2, \end{array} \right. \quad \text{pour } k \in U_h. \qquad (3)$$

7) Dans chaque strate, l'estimateur GREG s'écrit

$$\begin{aligned} \hat{t}_{yh,greg} &= \hat{t}_{yh,\pi} + \hat{\mathbf{b}}_{h,\pi}^{\top} \left[t_{\mathbf{x},h} - \hat{t}_{\mathbf{x}h\pi} \right] \\ &= N_h \bar{y}_h + \frac{s_{xyh}}{s_{xh}^2} N_h (\bar{x}_h - \mu_{xh}) + \hat{b}_{0h,\pi} (N_h - N_h) \\ &= N_h \left\{ \bar{y}_h + \frac{s_{xyh}}{s_{xh}^2} (\bar{x}_h - \mu_{xh}) \right\}. \end{aligned}$$

Nous obtenons

$$\hat{t}_{yh,greg} \;=\; \left\{ \begin{array}{cl} 57.83 & \text{pour } k \in S_1, \\ 120.15 & \text{pour } k \in S_2, \\ 67.5 & \text{pour } k \in S_3, \end{array} \right.$$

soit $\hat{t}_{y,greg} \simeq 245.47$.

8) La strate 3 est exhaustive, et les deux estimateurs (Horvitz-Thompson ou GREG) dans cette strate coincident avec le vrai total, et présentent une variance nulle. Dans les deux autres strates, le gain de variance en utilisant l'estimateur GREG est donné par le facteur $1 - \rho_h^2$, avec ρ_h la corrélation dans la strate entre les variables x_k et y_k.
Cette corrélation peut être estimée approximativement sans biais par

$$\hat{\rho}_h \;=\; \frac{s_{xyh}}{\sqrt{s_{xh}^2 \times s_{yh}^2}} = \left\{ \begin{array}{ll} 0.80 & \text{pour } k \in S_1, \\ 0.88 & \text{pour } k \in S_2. \end{array} \right.$$

La variance va donc être diminuée de 64 % environ dans la strate 1, et de 77 % environ dans la strate 2.

Exercice 4 (6 points)

Nous nous intéressons à l'estimation du paramètre

$$\theta_1 = \frac{\sqrt{\theta_2}}{\sqrt{N} \mu_y} \quad \text{avec} \quad \theta_2 = \sum_{k \in U} (y_k - \mu_y)^2,$$

et μ_y la moyenne simple sur la population de la variable y_k.

1) Montrer que la linéarisée de θ_2 vaut

$$u_k(\theta_2) \quad = \quad (y_k - \mu_y)^2.$$

2) En utilisant les règles de calcul de la linéarisation, en déduire que

$$\frac{u_k(\theta_1)}{\theta_1} \quad = \quad \frac{1}{2}\frac{(y_k - \mu_y)^2}{\theta_2} + \frac{1}{2N} - \frac{y_k}{t_y}.$$

Correction Exercice 4

1) Nous avons vu en Td que

$$\theta_2 \quad = \quad \sum_{k \in U} (y_k)^2 - \frac{(t_y)^2}{N}.$$

D'où, en passant au logarithme le deuxième terme du membre de droite et en utilisant les règles de calcul de la linéarisée

$$\begin{aligned} u_k(\theta_2) &= (y_k)^2 - \frac{(t_y)^2}{N}\left(2\frac{y_k}{t_y} - \frac{1}{N}\right) \\ &= (y_k)^2 - 2y_k\mu_y + (\mu_y)^2 \\ &= (y_k - \mu_y)^2. \end{aligned}$$

2) En passant au logarithme, nous obtenons

$$\ln(\theta_1) \quad = \quad \frac{1}{2}\ln(\theta_2) + \frac{1}{2}\ln(N) - \ln(t_y),$$

puis

$$\begin{aligned} \frac{u_k(\theta_1)}{\theta_1} &= \frac{1}{2}\frac{u_k(\theta_2)}{\theta_2} + \frac{1}{2}\frac{u_k(N)}{N} - \frac{u_k(t_y)}{t_y} \\ &= \frac{1}{2}\frac{(y_k - \mu_y)^2}{\theta_2} + \frac{1}{2N} - \frac{y_k}{t_y}. \end{aligned}$$

Examen 2022-2023

Exercice 1 (6 points)

Nous considérons une population U de taille $N = 10$, pour laquelle nous disposons des valeurs de deux variables auxiliaires x et z, données dans le tableau suivant. La dernière ligne contient la moyenne dans la population de ces deux variables.

k	x_k	z_k
1	6	12
2	6	12
3	6	4
4	4	4
5	4	2
6	4	2
7	4	1
8	2	1
9	2	1
10	2	1
	$\mu_x = 4$	$\mu_z = 4$

1) Donner les probabilités d'inclusion correspondant à un sondage aléatoire simple dans U de taille $n = 4$.
Réponse : nous obtenons $\pi_k = n/N = 0.4$
2) Donner les probabilités d'inclusion proportionnelles à la variable x pour un tirage de taille $n = 4$.
Réponse : nous utilisons $\pi_k = nx_k/t_x$**, soit** $\pi_1 = \pi_2 = \pi_3 = 0.6$**,** $\pi_4 = \pi_5 = \pi_6 = \pi_7 = 0.4$ **et** $\pi_8 = \pi_9 = \pi_{10} = 0.2$**.**
3) Donner les probabilités d'inclusion proportionnelles à la variable z pour un tirage de taille $n = 4$.
Réponse : en utilisant la formule $\pi_k = nz_k/t_z$ **nous obtenons après recalcul** $\pi_1 = \pi_2 = 1$**,** $\pi_3 = \pi_4 = 0.5$**,** $\pi_5 = \pi_6 = 0.25$ **et** $\pi_7 = \pi_8 = \pi_9 = \pi_{10} = 0.125$**.**

Nous utilisons maintenant un découpage de la population en deux strates. La table suivante donne la moyenne et la dispersion des variables x et z dans chacune de ces deux strates. Nous considérons plusieurs stratégies de sondage aléatoire simple stratifié.

k	x_k	z_k	Strate
1	6	12	U_1
2	6	12	U_1
3	6	4	U_1
4	4	4	U_1
	$\mu_{x1} = 5.5$	$\mu_{z1} = 8$	
	$S^2_{x1} = 1$	$S^2_{z1} = 21.333$	
5	4	2	U_2
6	4	2	U_2
7	4	1	U_2
8	2	1	U_2
9	2	1	U_2
10	2	1	U_2
	$\mu_{x2} = 3$	$\mu_{z2} = 1.333$	
	$S^2_{x2} = 1.2$	$S^2_{z2} = 0.267$	

4) Donner l'allocation correspondant à une allocation proportionnelle pour une taille d'échantillon $n = 6$.
Réponse : nous utilisons $n_h = n\ N_h/N$**, soit** $n_1 = 2$ **et** $n_2 = 4$ **après arrondi.**
5) Donner l'allocation correspondant à une allocation optimale sur la variable x pour une taille d'échantillon $n = 6$. Donner les probabilités d'inclusion dans chaque strate.
Réponse : nous utilisons $n_h = n\frac{N_h S_{xh}}{\sum_{j=1}^{2} N_j S_{xj}}$**, soit** $n_1 = 2$ **et** $n_2 = 4$ **après arrondi.**
6) Donner l'allocation correspondant à une allocation optimale sur la variable z pour une taille d'échantillon $n = 6$. Donner les probabilités d'inclusion dans chaque strate.
Réponse : nous utilisons $n_h = n\frac{N_h S_{zh}}{\sum_{j=1}^{2} N_j S_{zj}}$**, soit après recalcul** $n_1 = 4$ **et** $n_2 = 2$**.**

Exercice 2 (6 points)

Dans une population U de taille $N = 1\ 000$, un échantillon S a été sélectionné **selon un plan de Poisson avec des probabilités d'inclusion égales** $\pi_{1k} = \pi_1 = \frac{45}{1\ 000}$. L'échantillon sélectionné contient 50 individus.

1) Donner la taille moyenne d'échantillon n_1 attendue avec le plan de sondage utilisé. Expliquer brièvement pourquoi l'échantillon sélectionné peut avoir une taille différente.
Réponse : la taille moyenne d'échantillon vaut $n_1 = \sum_{k\in U} \pi_{1k} = 45$**. La taille effective d'échantillon (qui vaut 50 dans cet exemple) peut être différente car le plan de Poisson est de taille aléatoire.**

Pour une variable d'intérêt y, nous obtenons les résultats suivants dans l'échantillon :

$$\sum_{k\in s} y_k = 163 \quad \text{et} \quad \sum_{k\in s} y_k^2 = 591. \tag{4}$$

2) Donner un estimateur sans biais du total t_y, et calculer sa valeur sur l'échantillon.
Réponse : nous utilisons $\hat{t}_{y\pi} = \sum_{k\in S} \frac{y_k}{\pi_{1k}} = \frac{1}{\pi_1} \sum_{k\in S} y_k \simeq 3\ 622$**.**
3) Donner un estimateur de variance sans biais associé, et calculer sa valeur sur l'échantillon. En déduire un intervalle de confiance à 95 % pour t_y, ainsi que le coefficient de variation estimé associé.
Réponse : nous utilisons $\hat{V}(\hat{t}_{y\pi}) = \frac{1-\pi_1}{(\pi_1)^2} \sum_{k\in S} y_k^2 \simeq 278\ 719$**. D'où l'intervalle de confiance** $IC = [3\ 622 \pm 1\ 035] = [2\ 587, 4\ 657]$**, et le coefficient de variation estimé** $\widehat{CV} = 14.6\%$.

Nous souhaitons maintenant sélectionner un nouvel échantillon S_2 dans U, toujours selon un plan de Poisson à probabilités égales, mais avec un nouveau jeu de probabilités d'inclusion $\pi_{2k} = \pi_2 = \frac{n_2}{1\ 000}$, où n_2 est à déterminer.

4) Montrer que, pour que le total t_y soit connu à plus ou moins 800 avec un niveau de confiance de 95 %, les probabilités d'inclusion doivent vérifier

$$\frac{1-\pi_2}{\pi_2} \sum_{k\in U} y_k^2 \ \leq \ \left(\frac{800}{1.96}\right)^2. \tag{5}$$

Réponse : la formule précédente s'obtient à partir de l'inégalité $1.96\sqrt{V(\hat{t}_{y\pi})} \leq 800$**, et en utilisant la formule de variance d'un tirage de Poisson**.
5) En déduire que la taille moyenne d'échantillon à sélectionner n_2 doit vérifier

$$n_2 \geq \frac{N}{1 + \left(\frac{800}{1.96}\right)^2 \times \frac{1}{\sum_{k \in U} y_k^2}}. \quad (6)$$

Réponse : la formule précédente s'obtient à partir de l'identité $\frac{1-\pi_2}{\pi_2} = \frac{N}{n_2} - 1$.

6) En utilisant les données déjà collectées, proposer un estimateur de n_2 et calculer sa valeur.
Réponse : dans l'équation précédente, nous pouvons estimer $\sum_{k \in U} y_k^2$ **par** $\sum_{k \in S_1} \frac{y_k^2}{\pi_{1k}} = 13\ 133$**. Nous obtenons** $n_2 \geq 74$.

Exercice 3 (8 points)

Cet exercice est constitué de 4 parties indépendantes. Chaque partie est notée sur 2 points.

Partie 1

Nous considérons une population U de taille $N = 10$, pour laquelle les probabilités d'inclusion sont données dans le tableau suivant :

k	π_k
1	0.6
2	0.6
3	0.6
4	0.4
5	0.4
6	0.4
7	0.4
8	0.2
9	0.2
10	0.2

1) Dans la population U triée comme indiqué dans le tableau, tirer un échantillon selon un plan de sondage systématique en utilisant le nombre uniforme $u = 0.30$.
Réponse : nous obtenons $s = \{1, 3, 5, 7\}$**.**
2) Donner la formule de l'estimateur de Horvitz-Thompson de la taille de la population N, et calculer sa valeur sur l'échantillon sélectionné.
Réponse : nous obtenons $\hat{N} = \sum_{k \in s} \frac{y_k}{\pi_k} = 8.33$**.**

Partie 2

Dans une population U de taille $N = 10$, nous avons sélectionné un échantillon de taille $n = 4$. La table suivante donne pour chaque unité de l'échantillon sa probabilité d'inclusion π_k, la valeur d'une variable d'intérêt y_k, la valeur d'une variable auxiliaire z_k, et la modalité d'appartenance à une post-strate U_1 ou U_2 :

k	π_k	y_k	z_k	Post-strate
1	0.6	5	12	U_1
3	0.6	5	4	U_1
5	0.4	3	2	U_2
7	0.4	1	1	U_2

1) Le total de la variable z_k dans la population vaut $t_z = 40$. Donner l'estimateur par le ratio du total de y, et calculer sa valeur.
Réponse : nous obtenons $\hat{t}_{y\pi} = 26.67$ **et** $\hat{t}_{z\pi} = 34.17$**, soit** $\hat{t}_{yR} = 31.22$
2) La post-strate U_1 contient $N_1 = 4$ individus, et la post-strate U_2 contient $N_2 = 6$ individus. Donner l'estimateur post-stratifié du total de y, et calculer sa valeur.
Réponse : nous obtenons $\hat{N}_1 = 3.33$ **et** $\hat{t}_{y1\pi} = 16.67$**,** $\hat{N}_2 = 5$ **et** $\hat{t}_{y2\pi} = 10$**, puis** $\hat{t}_{y,post} = 32.02$**.**

Partie 3

Un échantillon S est sélectionné dans une population U de taille N selon un sondage aléatoire simple sans remise de taille n. Pour une variable auxiliaire x_k, les totaux $t_x = \sum_{k \in U} x_k$ et $t_{x2} = \sum_{k \in U} x_k^2$ sont connus. La taille de la population N est également connue.

Pour une variable d'intérêt y_k, il est proposé d'utiliser l'estimateur

$$\begin{aligned}\hat{t}_{yw} &= \frac{N}{n}\sum_{k\in S} y_k + \hat{b}^\top \left[\begin{pmatrix} N \\ t_x \\ t_{x2} \end{pmatrix} - \frac{N}{n}\sum_{k\in S} \begin{pmatrix} 1 \\ x_k \\ x_k^2 \end{pmatrix} \right] \\ \text{avec } \hat{b} &= \left[\sum_{k\in S} \frac{1}{x_k} \begin{pmatrix} 1 \\ x_k \\ x_k^2 \end{pmatrix} \begin{pmatrix} 1 \\ x_k \\ x_k^2 \end{pmatrix}^\top \right]^{-1} \sum_{k\in S} \frac{1}{x_k} \begin{pmatrix} 1 \\ x_k \\ x_k^2 \end{pmatrix} y_k.\end{aligned}$$

1) S'agit-il d'un cas particulier de l'estimateur par la régression généralisée ? Ecrire le modèle de travail, sans oublier la structure de variance-covariance.
Réponse : le modèle de travail est $m : y_k = \alpha_0 + \alpha_1 x_k + \alpha_2 x_k^2 + \epsilon_k$**, avec** $E_m(\epsilon_k) = 0$ **et** $V_m(\epsilon_k) = \sigma^2 x_k$.
2) Donner la forme de l'estimateur par la régression généralisée si le modèle de travail est :

$$y_k = \alpha_0 + \alpha_2 x_k^2 + \epsilon_k \text{ avec } \left\{ \begin{array}{rcl} E_m(\epsilon_k) &=& 0, \\ V_m(\epsilon_k) &=& \sigma^2 \end{array} \right.$$

Réponse : nous obtenons l'estimateur :

$$\begin{aligned}\hat{t}_{yw} &= \frac{N}{n}\sum_{k\in S} y_k + \hat{b}^\top \left[\begin{pmatrix} N \\ t_{x2} \end{pmatrix} - \frac{N}{n}\sum_{k\in S} \begin{pmatrix} 1 \\ x_k^2 \end{pmatrix} \right] \\ \textbf{avec } \hat{b} &= \left[\sum_{k\in S} \begin{pmatrix} 1 \\ x_k^2 \end{pmatrix} \begin{pmatrix} 1 \\ x_k^2 \end{pmatrix}^\top \right]^{-1} \sum_{k\in S} \begin{pmatrix} 1 \\ x_k^2 \end{pmatrix} y_k.\end{aligned}$$

Partie 4

Un échantillon S est sélectionné dans une population U avec des probabilités d'inclusion $\pi_k > 0$. Nous nous nous intéréressons à l'estimation de la moyenne quadratique

$$\theta = \sqrt{\frac{1}{N}\sum_{k\in U} y_k^2}.$$

1) Donner l'estimateur par substitution de θ.
Réponse : nous utilisons l'estimateur

$$\hat{\theta} = \sqrt{\frac{\sum_{k \in S} \frac{y_k^2}{\pi_k}}{\sum_{k \in S} \frac{1}{\pi_k}}}$$

.
2) Donner la variable linéarisée de θ.
Réponse : en passant au logarithme, nous obtenons $u_k(\theta) = \frac{1}{2}\theta \left(\frac{y_k^2}{\sum_{l \in U} y_l^2} - \frac{1}{N} \right)$**.**

Examen 2021-2022

Exercice 1 (QCM : 4 points)

Pour chaque question, vous indiquerez dans votre copie le numéro des affirmations qui vous semblent justes. Aucune justification n'est demandée.
Barème : pour une question comprenant x affirmations justes, $1/x$ point par réponse juste, -0.5 point par réponse fausse. Les questions sans réponse ne seront pas notées.
Exemples : pour une question comprenant 1 affirmation juste, 1 point pour la réponse juste, -0.5 point par réponse fausse ; pour une question comprenant 2 affirmations justes, 0.5 point par réponse juste, -0.5 point par réponse fausse.

Question 1

La population U contient $N = 5$ unités. Ci-dessous les valeurs de deux variables d'intérêt et d'une variable de probabilités d'inclusion :

k	1	2	3	4	5
y_{1k}	20	30	40	50	60
y_{2k}	60	50	40	30	20
π_k	0.6	0.5	0.4	0.3	0.2

On utilise l'estimateur de Horvitz-Thompson pour estimer $t_{y1} = \sum_{k\in U} y_{1k}$ et $t_{y2} = \sum_{k\in U} y_{2k}$, avec deux stratégies possibles :

1. Sondage aléatoire simple sans remise de taille $n = 2$
2. Tirage de taille fixe à probabilités d'inclusion π_k

Quelles affirmations sont correctes :

1. Le tirage à probabilités inégales est préférable pour estimer t_{y1} et t_{y2}.
2. Le SRS est préférable pour estimer t_{y1} et t_{y2}.
3. Le tirage à probabilités inégales est préférable pour estimer t_{y2}.
4. Le SRS est préférable pour estimer t_{y1}.

Question 2

La population U contient $N = 5$ unités. Ci-dessous les valeurs d'une variable d'intérêt et d'une variable de probabilités d'inclusion :

k	1	2	3	4	5
y_k	60	50	40	30	20
π_k	0.8	0.7	0.5	0	0

Pour estimer le total $t_y = \sum_{k \in U} y_k$, on tire un échantillon avec les probabilités d'inclusion π_k et on utilise l'estimateur de Horvitz-Thompson. Le biais de cet estimateur est égal à :

1. 0,
2. 200,
3. 50,
4. -50.

Question 3

La population U contient $N = 5$ unités. Ci-dessous les valeurs d'une variable d'intérêt et d'une variable de probabilités d'inclusion :

k	1	2	3	4	5
y_k	60	50	40	30	30
π_k	0.8	0.4	0.3	0.3	0.2

On tire un échantillon selon un plan de Poisson. Pour estimer la moyenne $\mu_y = N^{-1} \sum_{k \in U} y_{1k}$, on propose d'utiliser :

— soit l'estimateur $\hat{\mu}_{y\pi} = \hat{t}_{y\pi}/N$,

— soit l'estimateur $\tilde{\mu}_y = \hat{t}_{y\pi}/\hat{N}_\pi$ avec $\hat{N}_\pi = \sum_{k \in S} \frac{1}{\pi_k}$.

Quelles affirmations sont vraies :

1. L'estimateur $\hat{\mu}_{y\pi}$ est exactement sans biais.
2. $\hat{\mu}_{y\pi}$ est l'estimateur par substitution.
3. L'estimateur $\tilde{\mu}_y$ est exactement sans biais.
4. La variance de $\hat{\mu}_{y\pi}$ est plus grande que celle de $\tilde{\mu}_y$.

Question 4

Quelles affirmations sont exactes :

1. L'estimateur par la régression généralisée est un cas particulier de l'estimateur par calage.
2. L'estimateur par substitution est un cas particulier d'estimateur par le ratio.
3. L'estimateur par le ratio est un cas particulier d'estimateur par la régression généralisée.
4. L'estimateur post-stratifié est un cas particulier de l'estimateur des moindres carrés généralisé.

Correction Exercice 1

Question 1 : réponses 3 et 4. Le tirage à probabilités inégales est préférable pour y_2 qui est positivement corrélée avec π_k. Le SRS est préférable pour y_1, qui est négativement corrélée avec π_k.

Question 2 : réponse 4. Le biais est donné par $-\sum_{k\in U;\pi_k=0} y_k$.

Question 3 : réponses 1 et 4. L'estimateur $\tilde{\mu}_y$ est l'estimateur par substitution, qui est seulement approximativement sans biais. En revanche, sa variance est plus petite que celle de $\hat{\mu}_{y\pi}$ pour un plan de taille aléatoire.

Question 4 : réponses 1 et 3. L'estimateur par substitution est un estimateur général pour une fonction de totaux, et n'est pas un cas particulier d'estimateur par le ratio. L'estimateur des moindres carrés généralisés est utilisé pour estimer un coefficient de régression, et n'a donc rien à voir avec l'estimateur post-stratifié.

Exercice 2 (8 points)

On considère la population des $N = 67$ principaux aéroports français en 2020. On veut sélectionner un échantillon de 25 aéroports dans cette population pour estimer :

— le nombre total t_y de passagers en 2020, avec y_k le nombre de passagers dans l'aéroport k en 2020,

— le nombre total t_z de passagers en transit en 2020, avec z_k le nombre de passagers en transit dans l'aéroport k en 2020.

On note x_k le nombre de passagers dans l'aéroport k en 2019. On connaît les informations suivantes :

— le nombre total de passagers en 2019 vaut $t_x = \sum_{k \in U} x_k = 20\,000\,000$,
— la dispersion du nombre de passagers en 2019 vaut

$$S_x^2 = \frac{1}{N-1} \sum_{k \in U} (x_k - \mu_x)^2 = 1.06\ 10^{14}.$$

1) Quel nombre d'aéroports doit-on échantillonner par sondage aléatoire simple pour que t_x soit estimé à 10 % près (avec un niveau de confiance de 95 %) en utilisant l'estimateur de Horvitz-Thompson (HT) ?

2) Pour un sondage aléatoire simple de taille $n = 25$, donner la variance et le coefficient de variation de l'estimateur de HT de t_x.

La population des $N = 67$ aéroports est découpée en trois strates :

— La strate U_1 des $N_1 = 11$ aéroports ayant accueilli plus de 2 000 000 de passagers en 2019,
— La strate U_2 des $N_2 = 30$ aéroports ayant accueilli entre 100 000 et 2 000 000 de passagers en 2019,
— La strate U_3 des $N_3 = 26$ aéroports ayant accueilli moins de 100 000 passagers en 2019.

On connaît également pour chaque strate U_h la dispersion exacte S_{xh}^2 du nombre de passagers :

$$S_{x1}^2 = 4.49\ 10^{14}, \quad S_{x2}^2 = 2.48\ 10^{11}, \quad S_{x3}^2 = 1.01\ 10^{9}$$

3) Donner l'allocation proportionnelle pour un sondage aléatoire simple stratifié de taille $n = 25$.

4) Donner la variance de l'estimateur de HT de t_x sous cette allocation. En utilisant la question 2, donner l'effet de plan (design-effect).

5) Donner l'allocation optimale pour l'estimation de t_x, pour un sondage aléatoire simple stratifié de taille $n = 25$.

6) Donner la variance de l'estimateur de HT de t_x sous cette allocation. En utilisant la question 2, donner l'effet de plan (design-effect). Expliquer brièvement ce résultat.

On retient finalement l'allocation suivante : on tire $n_1 = 11$ aéroports dans la strate 1, $n_2 = 12$ aéroports dans la strate 2 et $n_3 = 2$ aéroports dans la strate 3. On obtient les résultats suivants dans l'échantillon pour la variable y_k

$$\bar{y}_1 = \frac{1}{n_1}\sum_{k\in S_1} y_k = 5\ 170\ 000 \quad \text{et} \quad s^2_{y1} = \frac{1}{n_1}\sum_{k\in S_1}(y_k - \bar{y}_1)^2 = 3.92\ 10^{13},$$
$$\bar{y}_2 = 214\ 000 \quad \text{et} \quad s^2_{y2} = 4.95\ 10^{10},$$
$$\bar{y}_3 = 19\ 200 \quad \text{et} \quad s^2_{y3} = 6.29\ 10^{8}.$$

On obtient les résultats suivants dans l'échantillon pour la variable z_k

$$\bar{z}_1 = \frac{1}{n_1}\sum_{k\in S_1} z_k = 6\ 160 \quad \text{et} \quad s^2_{z1} = \frac{1}{n_1}\sum_{k\in S_1}(z_k - \bar{z}_1)^2 = 2.85\ 10^{7},$$
$$\bar{z}_2 = 266 \quad \text{et} \quad s^2_{z2} = 2.54\ 10^{5},$$
$$\bar{z}_3 = 27.5 \quad \text{et} \quad s^2_{z3} = 1.51\ 10^{3}.$$

7) Donner l'estimateur de HT de t_y, calculer sa valeur et donner un intervalle de confiance à 95 % .

8) Donner l'estimateur de HT de t_z, calculer sa valeur et donner un intervalle de confiance à 95 % .

Correction Exercice 2

1) Avec un niveau de confiance de 0.95, nous obtenons

$$\left|\frac{\hat{t}_{x\pi} - t_x}{t_x}\right| \leq 0.10 \quad \Leftrightarrow \quad 1.96\sqrt{V(\hat{t}_{x\pi})} \leq 0.10\ t_x$$
$$\Leftrightarrow \quad 1.96\sqrt{N^2\left(\frac{1}{n} - \frac{1}{N}\right)S_x^2} \leq 0.10\ t_x$$
$$\Leftrightarrow \quad n \geq \frac{1}{\frac{1}{N} + \left(\frac{0.10\mu_x}{1.96 S_x}\right)^2}.$$

On obtient $n \geq 66.03$. Cela signifie qu'il faut faire un recensement pour obtenir le niveau de précision voulu avec un sondage aléatoire simple.

2) Avec un sondage aléatoire simple de taille $n = 25$, nous obtenons

$$V(\hat{t}_{x\pi}) \quad = \quad N^2 \left(\frac{1}{n} - \frac{1}{N} \right) S_x^2 = 1.19\ 10^{16},$$

soit le coefficient de variation

$$CV(\hat{t}_{x\pi}) \quad = \quad \frac{\sqrt{V(\hat{t}_{x\pi})}}{t_x} = 545\%.$$

C'est un coefficient de variation très gros, qui signifie que l'estimation avec un sondage aléatoire simple de taille $n = 25$ est impossible.

3) L'allocation proportionnelle est donnée par

$$n_h \quad = \quad n \frac{N_h}{N},$$

et conduit après arrondi à

$$n_1 = 4, \quad n_2 = 11, \quad n_3 = 10.$$

4) Sous cette allocation, nous obtenons

$$V(\hat{t}_{x\pi}) \quad = \quad \sum_{h=1}^{3} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{xh}^2 = 8.66\ 10^{15}.$$

On obtient donc un effet de plan de :

$$DEFF \quad = \quad \frac{V_{STSRS}(\hat{t}_{x\pi})}{V_{SRS}(\hat{t}_{x\pi})} = \frac{8.66\ 10^{15}}{1.19\ 10^{16}} = 0.73$$

La stratification à allocation proportionnelle permet de réduire la variance de 27% par rapport à un sondage aléatoire simple. On se débarasse de la variabilité inter. Ce n'est pas négligeable, mais la variance reste trop forte pour fournir une estimation fiable.

5) L'allocation optimale est donnée par

$$n_h \;=\; n\frac{N_h S_{xh}}{\sum_{j=1}^{3} N_j S_{xj}} \text{ pour } h = 1, ..., 3.$$

et conduit à

$$n_1 = 23.42, \quad n_2 = 1.50, \quad n_3 = 0.08.$$

Comme $n_1 > N_1$, on procède à un recensement dans U_1 ($n_1 = N_1 = 11$), et on recalcule l'allocation dans les deux autres strates selon la formule

$$n_h \;=\; (n-11)\frac{N_h S_{xh}}{\sum_{j=2}^{3} N_j S_{xj}} \text{ pour } h = 2, 3.$$

Après arrondi, on obtient

$$n_2 = 13 \quad \text{et} \quad n_3 = 1.$$

6) Sous cette allocation, nous obtenons

$$V(\hat{t}_{x\pi}) \;=\; \sum_{h=1}^{3} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{xh}^2 = 1.04\ 10^{13}.$$

On obtient donc un effet de plan de :

$$DEFF \;=\; \frac{V_{STSRS}(\hat{t}_{x\pi})}{V_{SRS}(\hat{t}_{x\pi})} = \frac{1.04\ 10^{13}}{1.19\ 10^{16}} = 0.0009$$

La réduction de variance est énorme par rapport à un sondage aléatoire simple, et permet de produire une estimation acceptable. Les unités de la strate 1 sont responsables de la plus grande partie de la variabilité intra. Effectuer un recensement de cette strate permet de gommer complètement cette variabilité.

7) Nous obtenons

$$\hat{t}_{y\pi} \;=\; \sum_{h=1}^{3} N_h \bar{y}_h = 63\ 800\ 000,$$

et l'estimateur de variance

$$\hat{V}(\hat{t}_{y\pi}) \;=\; \sum_{h=1}^{3} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) s_{yh}^2 = 2.42\ 10^{12}.$$

On en déduit l'intervalle de confiance

$$IC_{0.95}(t_y) \;=\; [63\ 800\ 000 \pm 3\ 000\ 000] = [60\ 800\ 000, 66\ 800\ 000]\,.$$

8) Nous obtenons

$$\hat{t}_{z\pi} \;=\; \sum_{h=1}^{3} N_h \bar{z}_h = 76\ 500,$$

et l'estimateur de variance

$$\hat{V}(\hat{t}_{z\pi}) \;=\; \sum_{h=1}^{3} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) s_{zh}^2 = 1.19\ 10^{7}.$$

On en déduit l'intervalle de confiance

$$IC_{0.95}(t_z) \;=\; [76\ 500 \pm 6\ 800] = [69\ 700, 83\ 200]\,.$$

Exercice 3 (4 points)

On considère la population des $N = 20$ aéroports français ayant accueilli entre 300 000 et 2 000 000 de passagers en 2019. La liste est donnée dans le tableau suivant, avec :

- le nombre x_k de passagers dans l'aéroport k en 2019,
- le nombre z_k de passagers en transit dans l'aéroport k en 2020,
- un nombre aléatoire u_k généré selon une loi uniforme.

Aéroport k	Nom	x_k	z_k	u_k	π_k	I_k
1	MONTPELLIER-MEDITERRANEE	2 000 000	300	0.67		
2	BASTIA-PORETTA	1 600 000	1 400	0.02		
3	AJACCIO-NAPOLEON-BONAPARTE	1 500 000	1 300	0.37		
4	STRASBOURG-ENTZHEIM	1 300 000	1 300	0.09		
5	BREST-BRETAGNE	1 200 000	1 800	0.11		
6	BIARRITZ-PAYS-BASQUE	1 100 000	200	0.60		
7	RENNES-ST-JACQUES	900 000	200	0.28		
8	FIGARI-SUD-CORSE	700 000	2 900	0.50		
9	PAU-PYRENEES	600 000	0	0.73		
10	TOULON-HYERES	500 000	0	0.86		
11	PERPIGNAN-RIVESALTES	500 000	0	0.64		
12	TARBES-LOURDES-PYRENEES	500 000	0	0.88		
13	CLERMONT-FERRAND-AUVERGNE	400 000	100	0.38		
14	CARCASSONNE-SALVAZA	400 000	0	0.04		
15	CALVI-STE-CATHERINE	300 000	200	0.23		
16	GRENOBLE-ALPES-ISERE	300 000	300	0.84		
17	CAEN-CARPIQUET	300 000	100	0.44		
18	LIMOGES-BELLEGARDE	300 000	500	0.99		
19	BERGERAC-DORDOGNE-PERIGORD	300 000	0	0.86		
20	BEZIERS-VIAS	300 000	0	0.04		

1) Donner les probabilités d'inclusion π_k pour un tirage de taille $n = 8$, à probabilités proportionnelles au nombre de passagers en 2019. Compléter la colonne correspondante du tableau, en arrondissant à deux décimales.

2) En utilisant les nombres aléatoires u_k, sélectionner un échantillon selon un plan de Poisson à probabilités d'inclusion π_k. Indiquer dans la colonne correspondante du tableau les indicateurs de sélection I_k obtenus.

3) En utilisant l'échantillon sélectionné, calculer l'estimateur de Horvitz-Thompson $\hat{t}_{z\pi}$ du nombre total de passagers en transit en 2020, et donner sa valeur.

4) En utilisant l'échantillon sélectionné, donner un estimateur de variance sans biais pour $\hat{t}_{z\pi}$, et calculer sa valeur. En déduire le coefficient de variation estimé de $\hat{t}_{z\pi}$.

Correction Exercice 3

1) On calcule les probabilités d'inclusion selon la formule

$$\pi_k \;=\; n\frac{x_k}{t_x} = 8 \times \frac{x_k}{15\ 000\ 000} \text{ pour } k \in U.$$

Pour l'unité 1 ayant la plus grande valeur de x_k, on obtient $\pi_1 = 16/15 > 1$. On fixe donc $\pi_1 = 1$, et on recalcule les autres probabilités selon la formule

$$\pi_k \;=\; (n-1)\frac{x_k}{t_x - x_1} = 7 \times \frac{x_k}{13\ 000\ 000} \text{ pour } k \in U \setminus \{1\}.$$

On obtient les probabilités données dans le tableau.

2) Pour un tirage de Poisson, on retient les unités telles que $u_k \leq \pi_k$. On obtient $s = \{1, 2, 3, 4, 5, 7, 14, 20\}$, voir le tableau.

3) Avec l'échantillon $S = \{1, 2, 3, 4, 5, 7, 14, 20\}$, nous obtenons l'estimation :

$$\hat{t}_{z\pi} \;=\; \sum_{k \in S} \frac{z_k}{\pi_k} = 8\ 600.$$

4) L'estimateur de variance de Horvitz-Thompson est donné par

$$\hat{V}(\hat{t}_{z\pi}) \;=\; \sum_{k \in S} \left(\frac{z_k}{\pi_k}\right)^2 (1 - \pi_k) = 4.67\ 10^6,$$

soit un coefficient de variation estimé de

$$C\hat{V}(\hat{t}_{z\pi}) \quad = \quad \frac{\sqrt{\hat{V}(\hat{t}_{z\pi})}}{\hat{t}_{z\pi}} = 25\%.$$

Aéroport k	Nom	x_k	z_k	u_k	π_k	I_k
1	MONTPELLIER-MEDITERRANEE	2 000 000	300	0.67	1.00	1
2	BASTIA-PORETTA	1 600 000	1 400	0.02	0.86	1
3	AJACCIO-NAPOLEON-BONAPARTE	1 500 000	1 300	0.37	0.81	1
4	STRASBOURG-ENTZHEIM	1 300 000	1 300	0.09	0.70	1
5	BREST-BRETAGNE	1 200 000	1 800	0.11	0.65	1
6	BIARRITZ-PAYS-BASQUE	1 100 000	200	0.60	0.59	0
7	RENNES-ST-JACQUES	900 000	200	0.28	0.48	1
8	FIGARI-SUD-CORSE	700 000	2 900	0.50	0.38	0
9	PAU-PYRENEES	600 000	0	0.73	0.32	0
10	TOULON-HYERES	500 000	0	0.86	0.27	0
11	PERPIGNAN-RIVESALTES	500 000	0	0.64	0.27	0
12	TARBES-LOURDES-PYRENEES	500 000	0	0.88	0.27	0
13	CLERMONT-FERRAND-AUVERGNE	400 000	100	0.38	0.22	0
14	CARCASSONNE-SALVAZA	400 000	0	0.04	0.22	1
15	CALVI-STE-CATHERINE	300 000	200	0.23	0.16	0
16	GRENOBLE-ALPES-ISERE	300 000	300	0.84	0.16	0
17	CAEN-CARPIQUET	300 000	100	0.44	0.16	0
18	LIMOGES-BELLEGARDE	300 000	500	0.99	0.16	0
19	BERGERAC-DORDOGNE-PERIGORD	300 000	0	0.86	0.16	0
20	BEZIERS-VIAS	300 000	0	0.04	0.16	1

Exercice 4 (4 points)

On souhaite estimer le total t_y d'une variable d'intérêt y_k, dans une population de taille $N = 1,000,000$ dans laquelle on a sélectionné un échantillon selon un sondage aléatoire simple de taille $n = 1,000$.

On connaît la moyenne $\mu_x = 15$ d'une variable auxiliaire x_k, et on relève également les informations suivantes sur l'échantillon :

$$\bar{x} = 14, \quad s_x^2 = 25, \quad \bar{y} = 10, \quad s_y^2 = 20, \quad s_{xy} = \frac{1}{n-1}\sum_{k\in s}(x_k - \bar{x})(y_k - \bar{y}) = 15.$$

1) Calculer la valeur de l'estimateur de Horvitz-Thompson de t_y, et de son estimateur sans biais de variance.
2) Calculer la valeur de l'estimateur par le ratio de t_y, et de son estimateur approximativement sans biais de variance.

On admet que l'estimateur GREG utilisant le vecteur de variables auxiliaires $\mathbf{x}_k = (1, x_k)^T$ et la variable de pondération $q_k = 1$ se réécrit sous la forme

$$\hat{t}_{y,greg} = \hat{t}_{y\pi} + \frac{s_{xy}}{s_x^2}(t_x - \hat{t}_{x\pi}),$$

et que sa variance peut être estimée par

$$\hat{V}(\hat{t}_{y,greg}) = N^2 \frac{1-f}{n}(1-\rho^2)s_y^2$$

avec ρ le coefficient de corrélation entre x_k et y_k calculé sur l'échantillon.

3) Calculer la valeur de l'estimateur GREG de t_y, et de son estimateur de variance.

4) Quel estimateur proposez vous de retenir ?

Correction Exercice 4

1) Nous obtenons

$$\hat{t}_{y\pi} = N\bar{y} = 10^7 \quad \text{et} \quad \hat{V}(\hat{t}_{y\pi}) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)s_y^2 = 2.00\ 10^{10}.$$

2) Nous obtenons

$$\begin{aligned} \hat{t}_{yR} &= t_x \frac{\bar{y}}{\bar{x}} = 1.07\ 10^7, \\ \hat{V}(\hat{t}_{yR}) &= N^2\left(\frac{1}{n} - \frac{1}{N}\right)(s_y^2 - 2\hat{R}s_{xy} + \hat{R}^2 s_x^2) = 1.13\ 10^{10}. \end{aligned}$$

3) Nous obtenons

$$\begin{aligned}
\hat{t}_{y,greg} &= \hat{t}_{y\pi} + \frac{s_{xy}}{s_x^2}(t_x - \hat{t}_{x\pi}) = 1.06\ 10^7, \\
\hat{V}(\hat{t}_{y,greg}) &= N^2 \frac{1-f}{n}(1-\rho^2)s_y^2 = 1.10\ 10^{10} \text{ avec } \rho = \frac{s_{xy}}{\sqrt{s_x^2 \times s_y^2}} = 0.67.
\end{aligned}$$

4) On en conclut que l'estimateur par la régression est légèrement préférable. L'estimateur par le ratio est quasiment équivalent en termes de précision, ce qui est dû à une relation fortement linéaire entre y_k et x_k.

Examen 2017-2018

Exercice 1 (10 points)

On s'intéresse aux mails arrivant sur le serveur d'une école. On souhaite répondre aux deux questions suivantes :

- Parmi les mails entrants, quel est le nombre t_y de spams ?
- Parmi les mails non spams, quel est le pourcentage R de mails à caractère professionnel ?

On sélectionne un échantillon de la façon suivante. Chaque fois qu'un mail arrive, on le sélectionne avec une probabilité 1/100 dans l'échantillon, **indépendamment des mails déjà arrivés**. Parmi les 200 000 mails qui arrivent en une année dans l'école, on obtient ainsi un échantillon de 2 100 mails. On obtient sur cet échantillon les résultats suivants :

- Parmi les 2 100 mails tirés, 1 800 sont des spams.
- Parmi les 300 mails restants, 250 sont à caractère professionnel.

1) Quel est le plan de sondage utilisé ? Est-il de taille fixe ?
2) Donner l'estimateur de Horvitz-Thompson $\hat{t}_{y\pi}$ du nombre total de spams t_y, et calculer sa valeur.
3) Donner un estimateur de variance sans biais pour $\hat{t}_{y\pi}$. En déduire le coefficient de variation estimé de $\hat{t}_{y\pi}$, et calculer sa valeur.

4) Ecrire le pourcentage R en fonction des totaux de deux variables, que l'on explicitera.
5) Donner l'estimateur par substitution de R, noté $\hat{R}_\pi$ et calculer sa valeur.
6) Donner un estimateur de variance approximativement sans biais pour $\hat{R}_\pi$.
7) En déduire le coefficient de variation estimé de $\hat{R}_\pi$, et calculer sa valeur.

8) Comparer les coefficients de variation obtenus aux questions 3) et 7). Ce résultat est-il surprenant ?

Correction Exercice 1

1) Plan de Poisson à probabilités égales, la taille d'échantillon est donc aléatoire.
2) $\hat{t}_{y\pi} = (1/\pi) \sum_{k \in S} y_k = 180\ 000$.
3) $v(\hat{t}_{y\pi}) = (1-\pi)/(\pi^2) \sum_{k \in S} y_k^2 = 1.78\ 10^7$.

On obtient $cv(\hat{t}_{y\pi}) \simeq 2.3\%$.
4) On peut écrire $R = t_x/t_z$, avec $x_k = 1(k$ non spam et professionnel) et $z_k = 1(k$ non spam).
5) $\hat{R}_\pi = (\sum_{k\in S} x_k)/(\sum_{k\in S} z_k) = 83.3\%$.
6)

$$\begin{aligned} v(\hat{R}_\pi) &= \frac{1-\pi}{\pi^2} \sum_{k\in S} \left(\frac{x_k - \hat{R}_\pi z_k}{\hat{t}_{z\pi}} \right)^2 \\ &= \frac{1-\pi}{(\sum_{k\in S} z_k)^2} \left\{ \sum_{k\in S} x_k^2 + (\hat{R}_\pi)^2 \sum_{k\in S} z_k \right. \\ &\quad \left. -2\hat{R}_\pi \sum_{k\in S} x_k \right\} \simeq 4.6\ 10^{-4}. \end{aligned}$$

7) On obtient $cv(\hat{R}_\pi) \simeq 2.6\%$.
8) Les CV sont comparables alors que la seconde estimation est réalisée sur un sous-ensemble beaucoup plus petit. Cela vient du fait qu'avec un plan de taille aléatoire comme le tirage de Poisson, les proportions sont estimées de façon plus précise que les totaux.

Exercice 2 (10 points)

Dans une population U d'arbres, on s'intéresse à la relation entre le volume de l'arbre (en mètres cubes) noté y_k, et son diamètre à la base (mesuré en mètres) noté x_k. On suppose que les valeurs de y_k dans l'ensemble de la population U sont générées selon le modèle de superpopulation

$$m : y_k = \beta x_k + \sigma\epsilon_k, \text{ avec } \begin{cases} E_m(\epsilon_k) = 0, \\ V_m(\epsilon_k) = 1, \\ Cov_m(\epsilon_k, \epsilon_l) = 0 \text{ pour } k \neq l. \end{cases}$$

1) En supposant que l'on ait accès aux données sur l'ensemble de la population U, donner l'estimateur des MCO de β, noté B. Vous montrerez qu'il s'écrit en fonction des totaux des variables $z_{1k} = x_k y_k$ et $z_{2k} = x_k^2$.

En pratique, on sélectionne dans la population U un échantillon S avec des probabilités d'inclusion $\pi_k > 0$.

2) Donner l'estimateur par substitution de B basé sur l'échantillon S, que l'on notera $\hat{b}$.
3) Donner sa variable linéarisée, ainsi que sa variable linéarisée estimée.

Parmi les $N = 12\ 000$ arbres de U, l'échantillon S est de taille $n = 200$ et sélectionné selon la méthode de sélection-rejet. On obtient les résultats suivants sur l'échantillon :

$$\bar{z}_1 = \frac{1}{n}\sum_{k\in S} z_{1k} = 0.026 \qquad \bar{z}_2 = \frac{1}{n}\sum_{k\in S} z_{2k} = 0.126$$

$$s^2_{z1} = \frac{1}{n}\sum_{k\in S}(z_{1k} - \bar{z}_1)^2 = 0.0014 \qquad s^2_{z2} = \frac{1}{n}\sum_{k\in S}(z_{2k} - \bar{z}_2)^2 = 0.012$$

$$s_{z1z2} = \frac{1}{n}\sum_{k\in S}(z_{1k} - \bar{z}_1)(z_{2k} - \bar{z}_2) = 0.0039$$

4) A quel plan de sondage correspond la méthode de sélection-rejet ?
5) Calculer la valeur de $\hat{b}$ sur cet échantillon.
6) Donner un estimateur de variance pour $\hat{b}$, et calculer sa valeur.
7) En déduire un intervalle de confiance à 95 % pour B.

Correction Exercice 2

1) En appliquant l'estimateur des MCO $B = (X^\top X)^{-1}(X^\top Y)$, avec $X = (x_1, \ldots, x_N)^\top$ et $Y = (y_1, \ldots, y_N)^\top$, on obtient $B = t_{z1}/t_{z2}$.
2) En remplaçant les totaux par leurs estimateurs de Horvitz-Thompson, on obtient $\hat{b} = \hat{t}_{z1\pi}/\hat{t}_{z2\pi}$.
3) $u_k = (1/t_{z2}) * (z_{1k} - Bz_{2k})$ et $\hat{u}_k = (1/\hat{t}_{z2\pi}) * (z_{1k} - \hat{b}z_{2k})$.
4) Il s'agit d'un SRS.
5) $\hat{b} = \bar{z}_1/\bar{z}_2 \simeq 0.21$.
6) $v(\hat{b}) = \frac{1-f}{n(\bar{z}_2)^2}\left\{s^2_{z1} + \hat{b}^2 s^2_{z2} - 2\hat{b}s_{z1z2}\right\} \simeq 9.3\ 10^{-5}$.
7) $IC = [0.21 \pm 0.02]$.

Exercice 3 (12 points)

On considère une région agricole contenant $N = 2\ 000$ fermes découpées en 2 strates : 1 400 fermes de moins de 160 hectares (U_1) et 600 fermes de plus de 160 hectares (U_2). On note x_k la surface totale de l'exploitation k et y_k

sa surface totale en céréales. On cherche à estimer le total t_y en utilisant un sondage aléatoire simple stratifié de taille $n = 200$.

1) Quelle taille d'échantillon doit-on sélectionner dans chaque strate avec une allocation proportionnelle ?
2) On suppose que la dispersion de la variable y dans la strate U_2 est 4 fois plus grande que la dispersion dans la strate U_1. Quelle taille d'échantillon doit-on sélectionner dans chaque strate avec une allocation optimale pour la variable y_k ?

On choisit finalement $n_1 = 140$ et $n_2 = 60$, et on obtient les résultats suivants :

Strate	N_h	n_h	$\bar{y}_h$	$\bar{x}_h$	s^2_{yh}	s^2_{xh}	s_{xyh}
1	1 400	140	1 425	6 225	50 160	668 745	154 830
2	600	60	1 300	6 125	89 730	1 677 200	339 100

3) Donner l'estimateur de Horvitz-Thompson du total t_y, et calculer sa valeur.
4) Donner le coefficient de variation estimé, et calculer sa valeur.
5) Quelle taille d'échantillon aurait-il fallu sélectionner pour que le CV de cet estimateur soit inférieur à 3% ?

On suppose maintenant que le total t_x est connu, et vaut 280 000.

6) L'estimateur par le ratio paraît-il approprié ici ? Justifier brièvement.
7) Donner l'estimateur par le ratio du total t_y, et calculer sa valeur.
8) Donner le coefficient de variation estimé, et calculer sa valeur.
9) Quelle taille d'échantillon aurait-il fallu sélectionner pour que le CV de cet estimateur soit inférieur à 3% ?

10) Quelles autres variables auxiliaires aurait-on dû inclure dans les variables de calage ?

Correction Exercice 3

1) On obtient $n_1 = 140$ et $n_2 = 60$.
2) On obtient après arrondi $n_1 = 108$ et $n_2 = 92$.

3) $\hat{t}_{y\pi} = \sum_{h=1}^{H} N_h \bar{y}_h = 2\ 775\ 000$.
4) $cv(\hat{t}_{y\pi}) \simeq 1.2\%$.
5) En utilisant la formule approché de variance pour un STSRS à allocation proportionnelle :

$$n \geq \left[\frac{1}{N} + \frac{(0.03\mu_y)^2}{S^2_{y,intra}} \right]^{-1} \simeq 36$$

après estimation des paramètres manquants.
6) La surface cultivée en céréales devrait être approximativement proportionnelle à la surface totale, donc l'estimateur par le ratio paraît approprié.
7) $\hat{t}_{yR} = 62\ 712$.
8)

$$\begin{aligned} v(\hat{t}_{yR}) &= \sum_{h=1}^{H} N^2 \frac{(1-f_h)}{n_h (\hat{t}_{x\pi})^2} \left\{ s^2_{yh} + \hat{R}^2 s^2_{xh} - 2\hat{R} s_{xyh} \right\} \\ &\simeq 2.99\ 10^8. \end{aligned}$$

On obtient $cv(\hat{t}_{yR}) \simeq 28\%$.
9) Avec un raisonnement analogue à celui de Q5, on obtient

$$n \geq \left[\frac{1}{N} + \frac{(0.03\mu_y)^2}{S^2_{y-Rx,intra}} \right]^{-1} \simeq 1\ 808.$$

10) On aurait dû inclure les variables de stratification dans le calage.

Exercice 4 (8 points)

On souhaite estimer le poids total de maquereau pêché par bateau en Bretagne, un jour donné. Comme on ne dispose pas d'une liste de bateaux, on procède en deux temps :

- Parmi les 50 ports bretons, on sélectionne un échantillon de 20 ports avec des probabilités proportionnelles à leur nombre de bateaux. On suppose que 40 ports contiennent exactement 50 bateaux chacun, et que 10 ports contiennent exactement 200 bateaux chacun.
- Dans chaque port sélectionné, on tire un échantillon de 10 bateaux. Pour chacun d'entre eux, on obtient le poids total de maquereau pêché (en kg).

1) Quel plan de sondage est utilisé ici ?

2) Donner les probabilités d'inclusion des ports.
3) Donner les probabilités d'inclusion conditionnelles des bateaux, en fonction de la taille de leur port.
4) En déduire les probabilités d'inclusion finales des bateaux.

Dans l'échantillon des $n = 200$ bateaux sélectionnés, on a pêché un total de 7 500 kg de maquereau.

5) En déduire une estimation du poids total de maquereau pêché en Bretagne.

Un autre statisticien propose de sélectionner un échantillon de 10 ports avec des probabilités proportionnelles à leur nombre de bateaux, puis de sélectionner un échantillon de 20 bateaux dans chaque port.

6) Cette stratégie vous paraît-elle plus efficace que la première (justifier brièvement) ?

Correction Exercice 4

1) Plan à 2 degrés (autopondéré).
2) $\pi_{Ii} = 20 \times \frac{N_i}{4\ 000}$, soit $\pi_{Ii} = 0.25$ ou $\pi_{Ii} = 1$.
3)4) $\pi_{k|i} = 10/N_i$, et $\pi_k = 200/4\ 000 = 1/20$.
5) $\hat{t}_{y\pi} = 150$ tonnes.
6) Cette stratégie sera moins efficace, dans un tirage à deux degrés il vaut mieux tirer un plus grand nombre d'UP.

Examen 2016-2017

Exercice 1 (5 points)

Un directeur de cirque veut estimer le poids total de ses $N = 8$ éléphants, en vue d'acheter un nouveau camion. Comme le pesage d'un éléphant est une opération compliquée, il ne peut se permettre de sélectionner qu'un échantillon de $n = 5$ éléphants. Il dispose pour cela d'une information auxiliaire, sous la forme du poids x_k de l'éléphant une année auparavant.

	Eléphants adultes				Eléphanteaux			
k	1	2	3	4	5	6	7	8
x_k	3000	3000	2500	2500	300	300	200	200

1) Donner les probabilités d'inclusion des éléphants, pour un tirage à probabilités proportionnelles à x_k.

2) Sélectionner l'échantillon selon un tirage systématique, en utilisant l'ordre donné dans le tableau et $u = 0.32$ comme nombre aléatoire.

Après pesée, on observe que le poids de chaque éléphant adulte sélectionné a augmenté de 300 kilos, et que le poids de chaque éléphanteau sélectionné a augmenté de 200 kilos.

3) Donner l'estimateur de Horvitz-Thompson du poids total t_y, avec y_k le poids actuel de l'éléphant k, et calculer sa valeur.

4) Donner l'estimateur de Horvitz-Thompson du poids total des éléphants adultes, et calculer sa valeur. Quelle est la variance de cet estimateur ?

5) Donner l'estimateur post-stratifié de t_y, en utilisant le découpage de la population selon l'âge, et calculer sa valeur.

Correction Exercice 1

1) On obtient $\pi_1 = \pi_2 = \pi_3 = \pi_4 = 1$, $\pi_5 = \pi_6 = 0.3$, $\pi_7 = \pi_8 = 0.2$.

2) On obtient $s = \{1, 2, 3, 4, 6\}$.

3) $\hat{t}_{y\pi} = 2 \times 3300 + 2 \times 2800 + 500/0.3 \simeq 13\ 867$.

4) $\hat{t}_{y1\pi} = 2 \times 3300 + 2 \times 2800 = 12\ 200$. La variance est nulle, car tous les éléphants adultes ont un $\pi_k = 1$.

5)

$$\begin{aligned} \hat{t}_{ypost} &= \frac{N_1}{\hat{N}_1}\hat{t}_{y1\pi} + \frac{N_2}{\hat{N}_2}\hat{t}_{y2\pi} \\ &= t_{y1} + \frac{4}{1/0.3} \times \frac{500}{0.3} = 14\ 200. \end{aligned}$$

Problème (15 points)

On souhaite estimer le volume total de bois (en mètres cubes) disponibles dans une forêt. On dispose pour cela d'un échantillon S de $n = 236$ arbres, sélectionnés selon un **plan de Poisson** à probabilités égales $\pi_k = 1/50$.

Partie 1

1) Pour un échantillon sélectionné selon un plan de Poisson à probabilités égales $\pi_k = \pi$, montrer que l'estimateur de Horvitz-Thompson d'un total peut se réécrire

$$\hat{t}_{y\pi} = \frac{1}{\pi}\sum_{k\in S} y_k.$$

2) Pour un échantillon sélectionné selon un plan de Poisson à probabilités égales $\pi_k = \pi$, montrer que l'estimateur sans biais de variance de $\hat{t}_{y\pi}$ peut se réécrire

$$\hat{V}(\hat{t}_{y\pi}) = \frac{1-\pi}{\pi^2}\sum_{k\in S} y_k^2.$$

3) A partir de l'échantillon d'arbres sélectionnés, donner une estimation du nombre total d'arbres dans la forêt.

Les $n = 236$ arbres ont été abattus, ce qui a permis de mesurer pour chaque arbre $k \in S$ son diamètre à la base (x_k mesuré en mètres) et son volume (y_k

mesuré en mètres cubes). On relève les informations suivantes sur l'échantillon :

$$\sum_{k\in S} x_k = 77.11, \quad \sum_{k\in S} y_k = 13.31, \quad \sum_{k\in S} y_k^2 = 1.52.$$

4) Donner l'estimateur de Horvitz-Thompson $\hat{t}_{y\pi}$ du volume total de bois.
5) Calculer le coefficient de variation (estimé) associé à l'estimation du volume total de bois.

6) Donner l'estimateur par substitution $\tilde{\mu}_y$ du volume moyen par arbre μ_y, et calculer sa valeur.
7) Donner un estimateur de variance approximativement sans biais pour $\tilde{\mu}_y$, et le calculer. En déduire un coefficient de variation estimé pour $\tilde{\mu}_y$.

Partie 2

On dispose de l'information auxiliaire suivante : il y a au total $N = 12\ 000$ arbres dans la forêt, et le diamètre moyen des arbres de la forêt vaut

$$\mu_x = \frac{1}{N}\sum_{k\in U} x_k = 0.32.$$

Dans cette partie, on considère un estimateur par la régression généralisée, donné par la formule suivante :

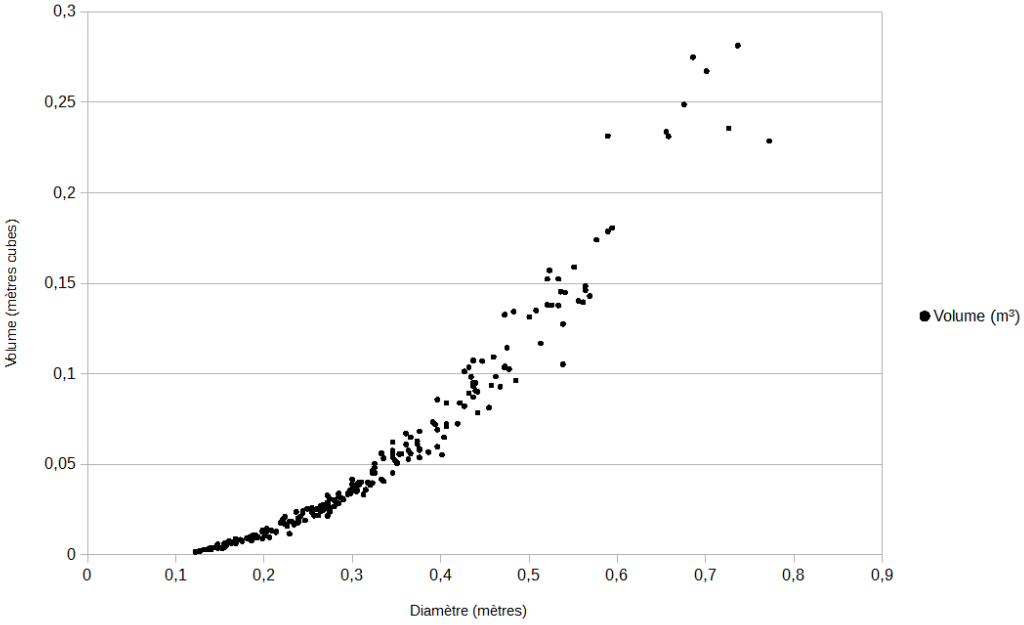
$$\begin{aligned} \hat{t}_{y,greg1} &= \hat{t}_{y\pi} + \hat{\mathbf{b}}_\pi^\top \left[t_{\mathbf{x}} - \hat{t}_{\mathbf{x}\pi}\right] \\ \text{avec} \qquad & \hat{\mathbf{b}}_\pi = \left[\sum_{k\in S} d_k \mathbf{x}_k \mathbf{x}_k^\top\right]^{-1} \sum_{k\in S} d_k \mathbf{x}_k y_k \\ \text{et} \qquad & \mathbf{x}_k = (1, x_k)^\top. \end{aligned}$$

8) Ecrire le modèle de travail associé à cet estimateur par la régression. Expliciter en particulier la structure de variance-covariance supposée.

On note $\hat{\mathbf{b}}_\pi = (\hat{a}, \hat{b})^\top$. Sur la base des données collectées, on obtient :

$$\begin{aligned} \hat{a} &= -0.073, \\ \hat{b} &= 0.397, \\ \sum_{k\in S} e_k^2 &= 0.062 \text{ avec } e_k = y_k - \hat{a} - \hat{b}x_k. \end{aligned}$$

FIGURE 1 – Volume en fonction du diamètre à la base



9) Calculer la valeur de l'estimateur par la régression $\hat{t}_{y,greg1}$.

10) Donner un estimateur de variance approximativement sans biais pour $\hat{t}_{y,greg1}$, et le calculer. En déduire un coefficient de variation estimé pour $\hat{t}_{y,greg1}$.

On représente dans la Figure 1 la relation observée sur l'échantillon entre le volume y_k et le diamètre x_k.

11) Compte-tenu de la Figure 1, le modèle de travail associé à $\hat{t}_{y,greg1}$ semble t-il approprié ? Que pourrait-on faire pour améliorer ce modèle ?

Partie 3

On dispose de l'information auxiliaire suivante :

$$t_{x2} = \sum_{k \in U} x_k^2 = 1\ 400.$$

Dans cette partie, on utilise le modèle de travail

$$m : y_k = \beta x_k^2 + \sigma \epsilon_k, \text{ avec } \begin{array}{c} E_m(\epsilon_k) = 0, \\ V_m(\epsilon_k) = x_k, \\ Cov_m(\epsilon_k, \epsilon_l) = 0 \text{ pour } k \neq l. \end{array}$$

12) Montrer que sous ce modèle de travail, l'estimateur par la régression généralisé est donné par

$$\begin{aligned} \hat{t}_{y,greg2} &= \hat{t}_{y\pi} + \frac{\sum_{k\in S} x_k y_k}{\sum_{k\in S} x_k^3} \left(t_{x2} - \hat{t}_{x2\pi}\right), \\ \text{avec} \qquad & \hat{t}_{x2\pi} = \frac{1}{\pi} \sum_{k\in S} x_k^2. \end{aligned}$$

Sur la base des données collectées, on obtient :

$$\sum_{k\in S} x_k^2 = 29.70, \quad \sum_{k\in S} x_k^3 = 13.14, \quad \sum_{k\in S} x_k^4 = 6.49,$$
$$\sum_{k\in S} x_k y_k = 6.15, \quad \sum_{k\in S} x_k^2 y_k = 3.11.$$

13) Calculer la valeur de l'estimateur par la régression $\hat{t}_{y,greg2}$.

14) Donner un estimateur de variance approximativement sans biais pour $\hat{t}_{y,greg2}$, et le calculer. En déduire un coefficient de variation estimé pour $\hat{t}_{y,greg2}$.

15) Comparer les résultats des questions 10 et 14. Conclusion ?

Correction Problème

Partie 1

3) $\hat{N}_\pi = n(S)/\pi = 236 \times 50 = 11\ 800$.

4) $\hat{t}_{y\pi} = (1/\pi) \sum_{k\in S} y_k \simeq 665.50$.

5) $\hat{V}(\hat{t}_{y\pi}) = (1-\pi)/(\pi^2) \sum_{k\in S} y_k^2 = 3\ 724$.
$\hat{CV} = \sqrt{3\ 724}/665.50 \simeq 9.2\%$.

6) $\tilde{\mu}_y = \hat{t}_{y\pi}/\hat{N}_\pi = 0.056$.

7) $\hat{V}(\tilde{\mu}_y) = (1-\pi)/(\pi^2)\sum_{k\in S} u_k^2$ avec $u_k = (1/\hat{N}_\pi) \times (y_k - \tilde{\mu}_y)$. On obtient $\hat{V}(\tilde{\mu}_y) \simeq 1.35\ 10^{-5}$ et $\hat{CV} \simeq 6.5\%$.

Partie 2

8) Le modèle de travail associé est

$$m : y_k = \beta^\top \mathbf{x}_k + \sigma\epsilon_k$$

avec $E_m(\epsilon_k) = 0$, $V_m(\epsilon_k) = 1$, $Cov_m(\epsilon_k, \epsilon_l) = 0$.

9)

$$\begin{aligned} \hat{t}_{y,greg1} &= \hat{t}_{y\pi} + \hat{a}(N - \hat{N}_\pi) + \hat{b}(t_x - \hat{t}_{x\pi}) \\ &\simeq 644.91. \end{aligned}$$

10) On remplace dans $\hat{V}(\hat{t}_{y\pi})$ la variable y_k par e_k, ce qui donne

$$\begin{aligned} \hat{V}(\hat{t}_{y,greg1}) &= (1-\pi)/(\pi^2)\sum_{k\in S} e_k^2 = 151.90, \\ \hat{CV} &= \sqrt{151.90}/644.91 \simeq 1.9\% \end{aligned}$$

11) On constate sur le graphique que la relation entre y_k et x_k est quadratique et non linéaire. D'autre part, la variance n'est pas constante mais augmente avec x_k.

Partie 3

12) On a $\mathbf{x}_k = x_k^2$ et

$$\begin{aligned} \hat{b}_\pi &= \left(\sum_{k\in S} d_k \frac{(x_k^2)^2}{\sigma^2 x_k}\right)^{-1} \sum_{k\in S} d_k \frac{x_k^2 y_k}{\sigma^2 x_k} \\ &= \frac{\sum_{k\in S} x_k y_k}{\sum_{k\in S} x_k^3}. \end{aligned}$$

On obtient $\hat{t}_{y,greg2} = \hat{t}_{y\pi} + \hat{b}_\pi(t_{x2} - \hat{t}_{x2\pi})$.

13) On obtient

$$\begin{aligned}\hat{t}_{y,greg2} &= 665.50 + \frac{6.15}{13.14}(1\ 400 - 50 \times 29.70) \\ &\simeq 625.72.\end{aligned}$$

14) Les résidus sont donnés par $e_k = y_k - \hat{b}_\pi x_k^2$, ce qui entraîne

$$\begin{aligned}\sum_{k\in S} e_k^2 &= \sum_{k\in S}(y_k - \hat{b}_\pi x_k^2)^2 \\ &= \sum_{k\in S} y_k^2 + (\hat{b}_\pi)^2 \sum_{k\in S} x_k^4 - 2\hat{b}_\pi \sum_{k\in S} x_k^2 y_k \\ &= 1.52 + (0.47)^2 \times 6.49 - 2 \times 0.47 \times 3.11 \\ &\simeq 0.031.\end{aligned}$$

On obtient

$$\begin{aligned}\hat{V}(\hat{t}_{y,greg2}) &= (1-\pi)/(\pi^2) \sum_{k\in S} e_k^2 = 74.73, \\ C\hat{V} &= \sqrt{74.73}/625.93 \simeq 1.4\%\end{aligned}$$

15) Le second estimateur GREG est plus efficace que le premier. Ce résultat est logique, car il utilise un modèle de travail qui représente mieux la relation entre le volume y_k et le diamètre à la base x_k.