# TD Données Manquantes dans les enquêtes

Ensai
Année 2025-2026

## Exercice 1

L'objectif de l'exercice est d'étudier l'impact de la non-prise en compte de la non-réponse totale pour l'estimation d'un total, et d'étudier une méthode de correction simple dans le cas d'un mécanisme uniforme (ou MCAR).

Nous sélectionnons dans une population $U$ de taille $N$ un échantillon $S$ avec des probabilités d'inclusion $\pi_k > 0$. En raison de la non-réponse totale, nous n'observons qu'un sous-échantillon $S_r$ de répondants. Nous supposons que le mécanisme de réponse est MCAR, et que les individus répondent indépendamment.

Pour une variable d'intérêt $y$ à **valeurs positives**, il est proposé d'utiliser l'estimateur

$$\hat{t}_y \quad = \quad \sum_{k \in S_r} \frac{y_k}{\pi_k}.$$

1) Montrer que

$$E\left(\hat{t}_y \middle| S\right) = p \sum_{k \in S} \frac{y_k}{\pi_k},$$

et en déduire le biais relatif (non conditionnel) de $\hat{t}_y$ en fonction de la probabilité $p$.
2) Quel est le signe de ce biais ? Ce résultat était-il prévisible ?
3) En utilisant l'équation de décomposition de la variance

$$V(\hat{t}_y) = V\left\{E(\hat{t}_y|S)\right\} + E\left\{V(\hat{t}_y|S)\right\},$$

montrer que la variance de cet estimateur peut s'écrire sous la forme

$$V(\hat{t}_y) \quad = \quad p^2 \sum_{k,l \in U} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} (\pi_{kl} - \pi_k \pi_l) + p(1-p) \sum_{k \in U} \frac{y_k^2}{\pi_k}.$$

A quoi correspondent chacun des deux termes dans la décomposition précédente?

Nous considérons maintenant l'estimateur

$$\hat{t}_{yr} = \frac{\hat{N}_\pi}{\hat{N}} \hat{t}_y \quad \text{avec} \quad \hat{N} = \sum_{k \in S_r} \frac{1}{\pi_k} \text{ et } \hat{N}_\pi = \sum_{k \in S} \frac{1}{\pi_k}.$$

Nous admettons l'approximation (voir annexe A)

$$\begin{aligned} \hat{t}_{yr} - \hat{t}_{y\pi} \quad &\simeq \quad \frac{1}{p} \left( \hat{t}_y - \frac{\hat{t}_{y\pi}}{\hat{N}_\pi} \times \hat{N} \right), \qquad (1) \\ \text{avec } \hat{t}_{y\pi} \quad &= \quad \sum_{k \in S} \frac{y_k}{\pi_k}. \end{aligned}$$

4) Montrer que $\hat{t}_{yr}$ est approximativement sans biais pour $t_y$.
5) Quel estimateur des probabilités de réponse est utilisé pour produire $\hat{t}_{yr}$ ?

# Exercice 2 (adapté de l'examen 2018-2019)

Dans une commune de $N = 10\,000$ habitants, nous sélectionnons un échantillon $S$ de taille $n = 500$ par sondage aléatoire simple. Parmi les personnes interrogées, 300 possèdent une voiture. Nous notons $\pi_k$ la probabilité d'inclusion de $k$ dans $S$.

1) Donner un estimateur du nombre de personnes possédant une voiture.
2) Donner un intervalle de confiance à 95% pour ce paramètre.

Nous faisons passer un questionnaire sur les habitudes d'utilisation de la voiture auprès d'un sous-échantillon $S_2$. Ce sous-échantillon est obtenu par tirage de Poisson dans $S$, avec une probabilité de tirage 1/4 si l'individu a moins de 60 ans, et 1/2 si l'individu a 60 ans ou plus. Nous notons $p_k$ la probabilité de sélection de l'unité $k$ dans $S_2$, conditionnelle à l'échantillon $S$.

3) Comment s'appelle le plan de sondage conduisant à la sélection de $S_2$?
4) Donner les poids d'extrapolation de l'estimateur par expansion $\hat{t}_{ye}$.
5) En utilisant l'équation de décomposition de la variance

$$V(\hat{t}_{ye}) = VE\left\{\hat{t}_{ye}|S_1\right\} + EV\left\{\hat{t}_{ye}|S_1\right\},$$

montrer que

$$V(\hat{t}_{ye}) = \underbrace{N^2\frac{1-f}{n}S_y^2}_{\text{variance sous un SRS}} + E\left\{\sum_{k\in S}\left(\frac{y_k}{\pi_k}\right)^2\frac{1-p_k}{p_k}\right\}.$$

6) Soit $s_y^2$ la dispersion dans l'échantillon $S$. Nous admettons l'identité

$$s_y^2 = \frac{1}{n-1}\sum_{k\in S}(y_k-\bar{y})^2 = \frac{1}{2n(n-1)}\sum_{k\neq l\in S}(y_k-y_l)^2.$$

Montrer que

$$\hat{V}_{1p}\left(\hat{t}_{ye}\right) = N^2\frac{1-f}{n}\times\frac{1}{2n(n-1)}\sum_{k\neq l\in S_2}\frac{(y_k-y_l)^2}{p_kp_l}$$

est un estimateur sans biais de la variance de première phase.

7) Montrer que

$$\hat{V}_{2p}\left(\hat{t}_{ye}\right) = \frac{N^2}{n}\times\frac{1}{n}\sum_{k\in S_2}(y_k)^2\frac{1-p_k}{(p_k)^2}$$

est un estimateur sans biais de la variance de seconde phase.

Dans l'échantillon $S_2$, 10 personnes de moins de 60 ans pratiquent le covoiturage, et 5 personnes de 60 ans et plus pratiquent le covoiturage.

8) Donner une estimation du nombre total de personnes pratiquant le covoiturage.

# Rappel pour les exercices 3 et 4

Si l'échantillon $S$ souffre de non-réponse totale, nous pouvons estimer le total $t_y$ en utilisant l'estimateur corrigé de la non-réponse totale

$$\hat{t}_{yr} = \sum_{k \in S_r} \frac{y_k}{\pi_k \hat{p}_k}.$$

Pour cet estimateur, nous pouvons utiliser l'estimateur de variance :

$$v(\hat{t}_{yr}) = v_p(\hat{t}_{yr}) + v_{nr}(\hat{t}_{yr}), \tag{2}$$

où le premier terme de (2) est un estimateur de la variance d'échantillonnage, et le second terme est un estimateur de la variance due à la non-réponse (cf diapo 96 du poly).

Dans le cas particulier où le plan de sondage est un sondage aléatoire simple stratifié, nous avons (cf diapo 103 du poly) :

$$\begin{aligned} v_p(\hat{t}_{yr}) &= \sum_{h=1}^{H} (N_h)^2 \frac{1-f_h}{n_h} s^2_{yhr}, \qquad (3) \\ \text{avec } s^2_{yhr} &= \frac{\sum_{k \in S_{hr}} \frac{1}{\hat{p}_k} (y_k - \bar{y}_{hr})^2}{\sum_{k \in S_{hr}} \frac{1}{\hat{p}_k}} \\ \text{et } \bar{y}_{hr} &= \frac{\sum_{k \in S_{hr}} \frac{y_k}{\hat{p}_k}}{\sum_{k \in S_{hr}} \frac{1}{\hat{p}_k}}. \end{aligned}$$

Dans le cas particulier où la non-réponse est corrigée selon la méthode des GHR, nous avons (cf diapo 101 du poly) :

$$v_{nr}(\hat{t}_{yr}) = \sum_{c=1}^{C} \frac{1-\hat{p}_c}{(\hat{p}_c)^2} \sum_{k \in S_{rc}} \left( \frac{y_k}{\pi_k} - \frac{1}{n_{rc}} \sum_{l \in S_{rc}} \frac{y_l}{\pi_l} \right)^2 \tag{4}$$

# Exercice 3 (examen 2017-2018)

Nous souhaitons estimer le nombre total $t_y$ d'élèves attachés de 2ème année qui portent des lunettes. Parmi les $N = 50$ attachés de la promotion, nous sélectionnons un échantillon de $n = 25$ individus par sondage aléatoire simple. Parmi ces 25 individus, 20 acceptent de répondre. Parmi ces 20 répondants, 8 portent des lunettes.
Nous supposerons que les individus répondent indépendamment les uns des autres, et que le mécanisme de réponse est MCAR.

1) Donner l'estimateur corrigé de la non-réponse totale de $t_y$, et le calculer.
2) En utilisant la formule (3), montrer que l'estimateur de la variance due à l'échantillonnage peut se réécrire sous la forme

$$v_p(\hat{t}_{yr}) = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) s^2_{yr},$$

$$\text{avec } s^2_{yr} = \frac{1}{n_r} \sum_{k \in S_r} (y_k - \bar{y}_r)^2 \quad \text{et} \quad \bar{y}_r = \frac{1}{n_r} \sum_{k \in S_r} y_k.$$

3) En utilisant la formule (4), montrer que l'estimateur de la variance due à la non-réponse peut se réécrire sous la forme

$$\hat{V}_{nr}(\hat{t}_{yr}) = N^2 \left(\frac{1}{n_r} - \frac{1}{n}\right) s^2_{yr}.$$

4) En utilisant les deux questions précédentes, montrer qu'un estimateur de variance global est donné par

$$\hat{V}(\hat{t}_{yr}) = N^2 \left(\frac{1}{n_r} - \frac{1}{N}\right) s^2_{yr}.$$

A quel plan de sondage correspond cette formule d'estimation de variance?
5) En déduire un intervalle de confiance à 95% pour $t_y$, et le calculer.

Après relance, nous obtenons la réponse des 5 non-répondants initiaux : parmi ceux-ci, 3 portent des lunettes.
6) Proposer un nouvel estimateur de $t_y$, et le calculer.
7) Donner un estimateur de variance sans biais et le calculer.
8) En déduire un intervalle de confiance à 95% pour $t_y$, et le calculer.

# Exercice 4 (extrait de l'examen 2017-2018)

Nous voulons estimer le chiffre d'affaires total dans une population de 1 000 entreprises, découpée en une strate $U_1$ de $N_1 = 400$ entreprises de plus de 50 salariés, et en une strate $U_2$ de $N_2 = 600$ entreprises de moins de 50 salariés. Nous sélectionnons un échantillon $S_1$ de taille $n_1 = 60$ dans $U_1$, et un échantillon $S_2$ de taille $n_2 = 40$ dans $U_2$ par **sondage aléatoire simple stratifié**.

Parmi les entreprises de $S_1$, $n_{r1} = 50$ acceptent de répondre et parmi celles de $S_2$, $n_{r2} = 20$ acceptent de répondre. D'après un expert, les entreprises ont répondu indépendamment les unes des autres et le mécanisme de non-réponse peut être considéré comme **homogène au sein des strates**. Sur les sous-échantillons $S_{r1}$ et $S_{r2}$, nous obtenons les résultats suivants

$$\bar{y}_{r1} = \frac{1}{n_{r1}} \sum_{k \in S_{r1}} y_k = 1.80 \quad \text{et} \quad \frac{1}{n_{r1} - 1} \sum_{k \in S_{r1}} (y_k - \bar{y}_{r1})^2 = 1.79,$$
$$\bar{y}_{r2} = \frac{1}{n_{r2}} \sum_{k \in S_{r2}} y_k = 0.40 \quad \text{et} \quad \frac{1}{n_{r2} - 1} \sum_{k \in S_{r2}} (y_k - \bar{y}_{r2})^2 = 0.50.$$

avec $y_k$ le chiffre d'affaires en millions d'euros.

1) Donner l'estimateur corrigé de la NR totale $\hat{t}_{yr}$ du chiffre d'affaires.
2) Donner un estimateur, noté $v_{NR}(\hat{t}_{yr})$, sans biais de la variance de $\hat{t}_{yr}$ due à la non-réponse, et le calculer.

Vous avez la possibilité de faire de la relance auprès des non-répondants selon deux stratégies, équivalentes en termes de coût :

- Stratégie 1 : faire de la relance auprès des entreprises de $S_1$ uniquement, jusqu'à ce qu'elles répondent toutes.
- Stratégie 2 : faire de la relance auprès des entreprises de $S_2$ uniquement, jusqu'à ce qu'elles répondent toutes.

3) Laquelle des deux stratégies choisissez-vous? Justifiez **quantitativement** votre réponse.

# A Justification de l'approximation (1)

Nous pouvons obtenir cette approximation par un développement de Taylor conditionnel à l'échantillon $S$. Soit $\mathbf{y}_k = (y_k, 1)^\top$, de sorte que $\hat{t}_\mathbf{y} = \left(\hat{t}_y, \hat{N}\right)^\top$. Nous notons $\tilde{t}_\mathbf{y} = E\left(\hat{t}_\mathbf{y} \,|S\right)$, et d'après les résultats de la question 1 :

$$\tilde{t}_\mathbf{y} = p\left(\hat{t}_{y\pi}, \hat{N}_\pi\right)^\top.$$

Soit $f(\cdot)$ une fonction différentiable au voisinage de $\tilde{t}_\mathbf{y}$. Alors en appliquant les résultats vus en cours pour la linéarisation

$$f\left(\hat{t}_\mathbf{y}\right) - f\left(\tilde{t}_\mathbf{y}\right) = f'(\tilde{t}_\mathbf{y}) \times \left(\hat{t}_\mathbf{y} - \tilde{t}_\mathbf{y}\right) + o_p(n^{-1/2}). \tag{5}$$

Nous appliquons ce résultat à la fonction ratio $f(u,v) \mapsto \frac{u}{v}$, pour laquelle

$$f'(u,v) = \left(\frac{1}{v}, -\frac{u}{v^2}\right)^\top. \tag{6}$$

En utilisant les équations (5) et (6), nous obtenons

$$\begin{aligned} f\left(\hat{t}_\mathbf{y}\right) - f\left(\tilde{t}_\mathbf{y}\right) &= \frac{\hat{t}_y}{\hat{N}} - \frac{\hat{t}_{y\pi}}{\hat{N}_\pi} \\ &\simeq \left\{\frac{1}{p\hat{N}_\pi}, -\frac{p\hat{t}_{y\pi}}{\left(p\hat{N}_\pi\right)^2}\right\}^\top \times \begin{pmatrix} \hat{t}_y - p\hat{t}_{y\pi} \\ \hat{N} - p\hat{N}_\pi \end{pmatrix} \\ &= \frac{\hat{t}_y - p\hat{t}_{y\pi}}{p\hat{N}_\pi} - \frac{\hat{t}_{y\pi}}{\hat{N}_\pi} \times \frac{\hat{N} - p\hat{N}_\pi}{p\hat{N}_\pi} = \frac{\hat{t}_y - \frac{\hat{t}_{y\pi}}{\hat{N}_\pi} \times \hat{N}}{p\hat{N}_\pi}. \end{aligned}$$

Finalement

$$\begin{aligned} \hat{N}_\pi \frac{\hat{t}_y}{\hat{N}} - \hat{t}_{y\pi} &= \hat{N}_\pi \left(\frac{\hat{t}_y}{\hat{N}} - \frac{\hat{t}_{y\pi}}{\hat{N}_\pi}\right) \\ &\simeq \hat{N}_\pi \left(\frac{\hat{t}_y - \frac{\hat{t}_{y\pi}}{\hat{N}_\pi} \times \hat{N}}{p\hat{N}_\pi}\right) = \frac{1}{p}\left(\hat{t}_y - \frac{\hat{t}_{y\pi}}{\hat{N}_\pi} \times \hat{N}\right). \end{aligned}$$