Échantillonnage

Chapitre 14 : La non-réponse et le redressement

Jaouad Madkour

30 octobre 2026

Plan du chapitre

Ceux qui ne répondent pas **2**
- Deux sortes de non-réponse . . . 2
- Le taux de réponse . . . 2
- Qui répond ? . . . 3
- Trois mécanismes . . . 3

Le biais de non-réponse **3**
- La moyenne des répondants . . . 3
- Sur la population du cours . . . 4
- Dans un échantillon . . . 4
- Le biais ne diminue pas avec n . . . 5
- Un intervalle de confiance trompeur . . . 6

Repondérer par classes **7**
- Une seconde phase de tirage . . . 7
- Le milieu comme classe de réponse . . . 7
- Ce que la correction répare . . . 8
- Choisir les classes . . . 9
- Sur l'échantillon fil rouge . . . 9

Imputer les valeurs manquantes **9**
- Compléter le fichier . . . 9
- La moyenne imputée écrase la dispersion . . . 10
- Le hot deck garde la dispersion . . . 11
- Dix mille enquêtes à valeurs manquantes . . . 12

Prévenir et corriger en pratique **12**
- Prévenir vaut mieux que corriger . . . 12
- Relancer un sous-échantillon de non-répondants . . . 12
- Le bilan sur l'échantillon fil rouge . . . 13

S'entraîner **14**
- Exercice 1 : une enquête sur les revenus . . . 14
- Exercice 1 : solution . . . 14

Exercice 2 : les commerces de Tétouan 14
Exercice 2 : solution 15
Exercice 3 : deux dépenses manquantes 15
Exercice 3 : solution 15

Ce qu'il faut retenir **16**
Le résumé du chapitre 16

Ceux qui ne répondent pas

Deux sortes de non-réponse

Totale ou partielle

- **Non-réponse totale** : le ménage tiré ne fournit rien. Refus, absence répétée, logement introuvable : il manque une **ligne** entière du fichier.
- **Non-réponse partielle** : le ménage répond, mais pas à toutes les questions. Il manque une **case**, souvent celle du revenu ou de la dépense.

On traite la première en **repondérant** les répondants, la seconde en **imputant** une valeur à la case vide.

Une erreur que la taille ne réduit pas

L'introduction l'a rangée parmi les erreurs qui ne se calculent pas. Le Literary Digest avait reçu 2,3 millions de réponses, un quart des bulletins envoyés : tirer deux fois plus de ménages donne deux fois plus de répondants, qui ressemblent **toujours aussi peu** aux absents.

Le taux de réponse

Une population coupée en deux

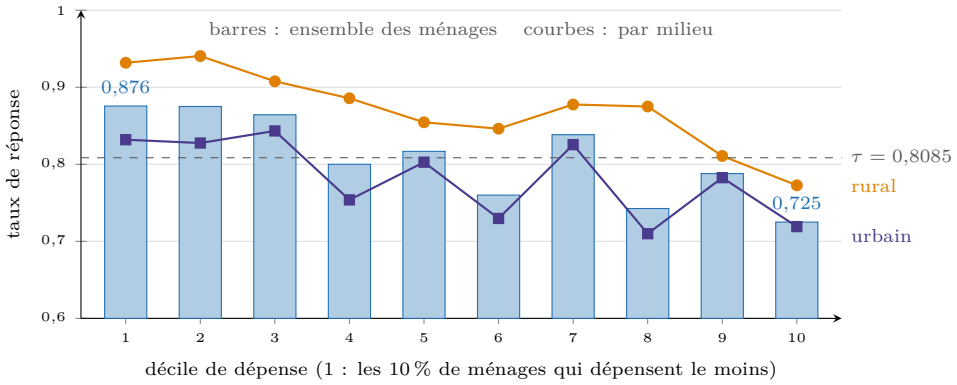
On suppose que chaque ménage, s'il était tiré, répondrait ou non. La population se partage en N_R **répondants** et $N - N_R$ **non-répondants**, de dépenses moyennes $\overline{Y}_R$ et $\overline{Y}_{NR}$. Le **taux de réponse** est $\tau = N_R/N$.

La colonne repond du fichier des ménages

Milieu	Ménages	Répondants	Taux de réponse
Urbain	1 438	1 119	0,778
Rural	562	498	0,886
Ensemble	2 000	1 617	$\tau = 0{,}8085$

Un institut ne sait jamais qui refuserait : il ne voit que les ménages tirés qui ont refusé. La population du cours le dit pour chacun, et permet de **mesurer** le biais.

Qui répond ?



Les ménages qui dépensent le plus répondent le moins

Le taux de réponse passe de 0,876 dans le premier décile de dépense à 0,725 dans le dernier. Dans chaque milieu, il tend à baisser quand la dépense augmente ; à dépense égale, les ruraux répondent davantage.

Trois mécanismes

Pourquoi ne répond-on pas ?

1. **Au hasard complet** : la réponse ne dépend de rien. Les répondants forment un sous-échantillon aléatoire : moins de précision, pas de biais.
2. **Au hasard conditionnel** : elle dépend de caractères connus pour tous les ménages tirés, comme le milieu ou la taille ; à caractères égaux, elle se fait au hasard. On peut la corriger.
3. **Non ignorable** : elle dépend de la variable étudiée elle-même, même à caractères égaux. Aucune correction n'est complète.

La population du cours relève du troisième cas

En ville, le taux passe de 0,837 dans le cinquième des ménages qui dépensent le moins à 0,747 dans celui qui dépense le plus ; à la campagne, de 0,930 à 0,812. À milieu fixé, la réponse dépend encore de la **dépense** : c'est elle qu'on cherche, et c'est elle qui fait refuser.

Le biais de non-réponse

La moyenne des répondants

Le biais de la moyenne des répondants

La moyenne générale mélange les deux groupes : $\overline{Y} = \tau\,\overline{Y}_R + (1-\tau)\,\overline{Y}_{NR}$. Qui n'interroge que les répondants estime $\overline{Y}_R$ au lieu de $\overline{Y}$, avec le biais

$$B = \overline{Y}_R - \overline{Y} = (1-\tau)\,(\overline{Y}_R - \overline{Y}_{NR})$$

La démonstration en une ligne

$\overline{Y}_R - \overline{Y} = \overline{Y}_R - \tau\,\overline{Y}_R - (1-\tau)\,\overline{Y}_{NR} = (1-\tau)(\overline{Y}_R - \overline{Y}_{NR})$.
Le biais est le produit de **deux** facteurs : la part des non-répondants, que l'on connaît, et l'écart entre répondants et non-répondants, que l'on ne connaît **jamais**.

Sur la population du cours

Un biais de 219 dirhams

Groupe	Ménages	Dépense moyenne (dh)
Répondants	1 617	$\overline{Y}_R = 7\,692{,}00$
Non-répondants	383	$\overline{Y}_{NR} = 8\,837{,}86$
Ensemble	2 000	$\overline{Y} = 7\,911{,}44$

$B = 0{,}1915 \times (7\,692{,}00 - 8\,837{,}86) = 0{,}1915 \times (-1\,145{,}86) = -219{,}43$ dh, soit $-2{,}77\,\%$.

Le taux de réponse ne suffit pas à juger

Avec $\tau = 0{,}95$, un écart de 4 389 dh donnerait le même biais. Avec $\tau = 0{,}90$ et le même écart de 1 146 dh, le biais tomberait à -115 dh. Un bon taux limite le biais **possible** ; il ne dit rien de l'écart.

Dans un échantillon

Les répondants d'un sondage aléatoire simple

On tire n ménages ; n_r répondent, de moyenne $\overline{y}_r$. À n_r fixé, ils forment un sondage aléatoire simple **parmi les répondants** de la population, d'où

$$E(\overline{y}_r) = \overline{Y}_R \qquad\qquad V(\overline{y}_r) \approx (1-f)\,\frac{S_R^2}{n\,\tau}$$

S_R^2 : variance de la dépense des répondants. Le biais de $\overline{y}_r$ est donc B.

Deux cents ménages tirés

En moyenne $200 \times 0{,}8085 = 161{,}7$ répondants. Avec $S_R = 3\,902{,}44$ dh, l'erreur-type vaut $\sqrt{0{,}9 \times 3\,902{,}44^2/161{,}7} = 291{,}1$ dh, contre 267,3 si tous répondaient. Erreur quadratique moyenne : $\sqrt{291{,}14^2 + 219{,}43^2} = 364{,}6$ dh.
La non-réponse coûte donc deux fois. En **précision** : tirer n/τ ménages y remédie (chapitre 5). En **biais** : aucun nombre de ménages n'y remédie.

Démonstration : Les répondants d'un sondage simple, biais et erreur quadratique

Sachant le nombre de répondants. Le fait de répondre est un caractère fixe de chaque ménage : la population contient N_R répondants. Soit n_r le nombre de répondants parmi les n ménages tirés. Pour tout ensemble A de n_r répondants, le nombre d'échantillons S tels que

$S \cap R = A$ est le nombre de façons de compléter par $n - n_r$ **non-répondants**, le même pour tous les A : sachant n_r, les répondants tirés forment un sondage aléatoire simple de n_r unités parmi N_R. Donc, par le chapitre 3,

$$E(\overline{y}_r \mid n_r) = \overline{Y}_R, \qquad V(\overline{y}_r \mid n_r) = \Big(\frac{1}{n_r} - \frac{1}{N_R}\Big) S_R^2.$$

Le biais ne dépend pas de n. $E(\overline{y}_r) = \overline{Y}_R$, donc le biais de $\overline{y}_r$ pour $\overline{Y}$ est $B = \overline{Y}_R - \overline{Y}$, **le même** quel que soit le nombre de ménages tirés.

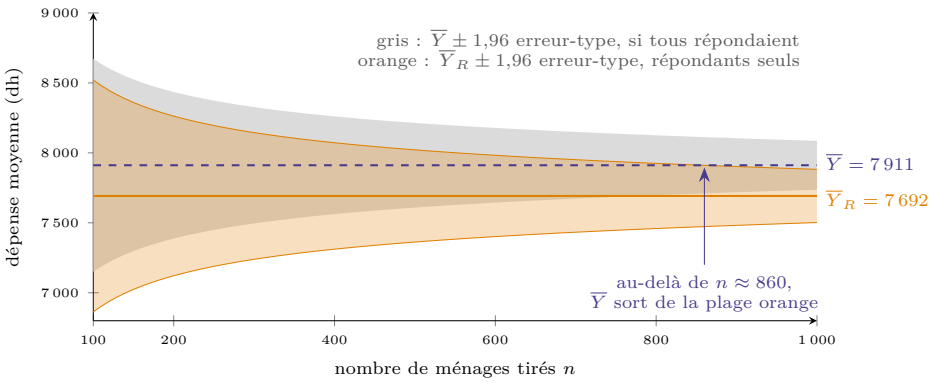
La variance. Avec $n_r \approx n\,\tau$ et $N_R = \tau\,N$, $\dfrac{1}{n\,\tau} - \dfrac{1}{\tau\,N} = \dfrac{1-f}{n\,\tau}$: $V(\overline{y}_r) \approx (1-f)\,\dfrac{S_R^2}{n\,\tau}$.

L'erreur quadratique moyenne. $E\big[(\overline{y}_r - \overline{Y})^2\big] = V(\overline{y}_r) + B^2$. Quand n croît, le premier terme tend vers 0 ; le second reste.

Ce que cela dit. Pour 200 ménages tirés, 161,7 répondants en moyenne : l'erreur-type $\sqrt{0{,}9 \times 3\,902{,}44^2/161{,}7} = 291{,}1$ dh dépasse celle d'un sondage où tous répondent (267,3) ; l'erreur quadratique moyenne vaut $\sqrt{291{,}1^2 + 219{,}4^2} = 364{,}6$ dh. Sur la mini-population, avec A, B, C répondants, l'énumération des échantillons de deux et trois ménages, regroupés selon n_r, retrouve $E(\overline{y}_r \mid n_r) = 17/3 = \overline{Y}_R$ et la variance de la formule, alors que $\overline{Y} = 8$.

■

Le biais ne diminue pas avec n



La plage se resserre autour de la mauvaise valeur

95 % des moyennes de répondants tombent dans la plage orange. Elle se resserre quand n augmente, mais autour de $\overline{Y}_R = 7\,692$. Au-delà de 860 ménages environ, $\overline{Y}$ en sort : un intervalle centré sur une estimation **typique** ne le contient plus.

Un intervalle de confiance trompeur

Dix mille sondages aléatoires simples pour chaque taille

Ménages tirés n	Répondants	Biais	Écart-type	Couverture
100	80,9	−222,6	421,3	0,894
200	161,7	−220,9	292,1	0,864
500	404,3	−219,3	168,0	0,752
1 000	808,3	−220,9	97,5	0,452

Biais et écart-type de $\overline{y}_r$ en dh. Couverture de l'intervalle à 95 % calculé comme si les répondants étaient l'échantillon : $\overline{y}_r \pm 1{,}96\,\sqrt{(1-n_r/N)\,s_r^2/n_r}$.

Plus de ménages, plus d'assurance, pas plus de vérité

Le biais reste vers −220 dh, l'écart-type fond : avec 1 000 ménages, **plus de la moitié** des intervalles manquent $\overline{Y}$. Les grandes enquêtes sont les plus exposées : la non-réponse y pèse plus que le hasard du tirage.

Démonstration : Un intervalle centré sur la mauvaise valeur

La couverture. Pour n grand, $\overline{y}_r$ est approximativement normale, d'espérance $\overline{Y}_R$ et d'écart-type $\sigma_n = \sqrt{(1-f)\,S_R^2/(n\,\tau)}$. L'intervalle usuel $\overline{y}_r \pm 1{,}96\,\hat{\sigma}_n$, avec $\hat{\sigma}_n \approx \sigma_n$, contient $\overline{Y}$ si $|\overline{y}_r - \overline{Y}| \le 1{,}96\,\sigma_n$. Écrivons $\overline{y}_r - \overline{Y} = B + \sigma_n\,Z$, avec Z normale centrée réduite, et $b = B/\sigma_n$: la probabilité de couvrir est

$$P\big(|b+Z| \le 1{,}96\big) = \Phi(1{,}96-b) - \Phi(-1{,}96-b).$$

Elle tend vers 0. B est fixe, σ_n décroît comme $1/\sqrt{n}$: $|b|$ croît comme $\sqrt{n}$ et la probabilité tend vers 0. Pour $b=0$ (pas de biais), on retrouve $\Phi(1{,}96) - \Phi(-1{,}96) = 0{,}95$.

Le seuil. $\overline{Y}$ sort de la plage centrale à 95 % des valeurs de $\overline{y}_r$ dès que $|B| > 1{,}96\,\sigma_n$, c'est-à-dire pour $(1-f)\,S_R^2/(n\,\tau) < \big(B/1{,}96\big)^2$.

Ce que cela dit. Avec $B = -219{,}43$, $S_R = 3\,902{,}44$ et $\tau = 0{,}8085$, ce seuil est atteint à $n = 859$ ménages tirés : au-delà, un intervalle centré sur une estimation **typique** ne contient plus $\overline{Y}$. L'approximation normale donne des couvertures de 0,92 ; 0,89 ; 0,76 et 0,45 pour $n = 100$, 200, 500 et 1 000, proches des 0,894 ; 0,864 ; 0,752 et 0,452 simulés : avec 1 000 ménages, $b = -2{,}3$ et plus de la moitié des intervalles manquent la vraie valeur.

■

Repondérer par classes

Une seconde phase de tirage

Le poids de non-réponse

On partage les ménages tirés en **classes** c, connues pour tous (le milieu, par exemple). Dans la classe c, r_c des n_c ménages tirés répondent : on estime leur probabilité de répondre par $\hat{\tau}_c = r_c/n_c$, et chaque répondant reçoit le poids

$$w_k = \frac{d_k}{\hat{\tau}_c} = d_k \times \frac{n_c}{r_c}$$

En sondage aléatoire simple, la moyenne repondérée vaut $\overline{y}_c = \sum_c \frac{n_c}{n}\,\overline{y}_{r,c}$.

Les répondants d'une classe parlent pour ses absents

La non-réponse est traitée comme un **second tirage**, de probabilité τ_c, que l'enquêteur n'a pas choisi. Si, dans chaque classe, les répondants ressemblent aux non-répondants (réponse **au hasard conditionnel**), la correction supprime le biais.

Le milieu comme classe de réponse

Repondérer par le milieu

Milieu	Part W_c	Taux τ_c	Répondants	Non-répondants
Urbain	0,719	0,778	8 249,12	9 063,48
Rural	0,281	0,886	6 440,16	7 713,28

Dépenses moyennes $\overline{Y}_{R,c}$ et $\overline{Y}_{NR,c}$, en dh. Sur de grands échantillons, $\overline{y}_c$ s'approche de $0{,}719 \times 8\,249{,}12 + 0{,}281 \times 6\,440{,}16 = 7\,740{,}8$ dh : il reste un biais de $-170{,}6$ dh, contre $-219{,}4$ dh sans correction.

Le biais résiduel

C'est la formule du biais, appliquée dans chaque classe :

$$\sum_c W_c\,(1-\tau_c)\,(\overline{Y}_{R,c} - \overline{Y}_{NR,c}) = -129{,}9 - 40{,}7 = -170{,}6 \text{ dh}$$

Démonstration : Repondérer par classes, poids et biais résiduel

Les poids. Dans la classe c, n_c ménages sont tirés et r_c répondent. Un répondant reçoit $w_k = d\,n_c/r_c$ avec $d = N/n$. La somme des poids vaut $\sum_c r_c \dfrac{d\,n_c}{r_c} = d\sum_c n_c = d\,n = N$, et la moyenne repondérée est

$$\frac{1}{N}\sum_{k\in S_r} w_k\,y_k = \frac{d}{N}\sum_c \frac{n_c}{r_c}\,r_c\,\overline{y}_{r,c} = \sum_c \frac{n_c}{n}\,\overline{y}_{r,c}.$$

Ce qu'elle estime. Sachant (n_c, r_c), les répondants de la classe c forment un sondage

simple parmi les $N_{R,c}$ répondants de la classe (même argument que pour l'ensemble) : $E(\overline{y}_{r,c}) = \overline{Y}_{R,c}$. Avec $E(n_c) = n\,W_c$, la moyenne repondérée a pour espérance, à la limite,

$$\sum_c W_c\,\overline{Y}_{R,c}.$$

Le biais résiduel. Dans la classe c, $\overline{Y}_c = \tau_c\,\overline{Y}_{R,c} + (1-\tau_c)\,\overline{Y}_{NR,c}$, donc $\overline{Y}_{R,c} - \overline{Y}_c = (1-\tau_c)(\overline{Y}_{R,c} - \overline{Y}_{NR,c})$. Comme $\overline{Y} = \sum_c W_c\,\overline{Y}_c$,

$$B_{\text{res}} = \sum_c W_c\,\overline{Y}_{R,c} - \overline{Y} = \sum_c W_c\,(1-\tau_c)(\overline{Y}_{R,c} - \overline{Y}_{NR,c}).$$

C'est la formule du biais, appliquée classe par classe : il s'annule si, dans chaque classe, répondants et non-répondants ont la même dépense moyenne.

Ce que cela dit. Pour le milieu, $0{,}719 \times 8\,249{,}12 + 0{,}281 \times 6\,440{,}16 = 7\,740{,}8$: biais de $-170{,}6$ dh, contre $-219{,}4$ sans correction, soit 48,8 dh (22 %) de gagnés. Sur le fil rouge, les poids 12,77 et 11,43 somment à $1\,520 + 480 = 2\,000$ et la moyenne repondérée est $0{,}76 \times 7\,593{,}6 + 0{,}24 \times 7\,192{,}4 = 7\,497{,}3$ dh. Sur une population de huit unités et deux classes, l'énumération des 56 échantillons de cinq unités confirme $E(\overline{y}_c \mid n_c, r_c) = \sum_c \frac{n_c}{n}\,\overline{Y}_{R,c}$ pour chaque composition.

■

Ce que la correction répare

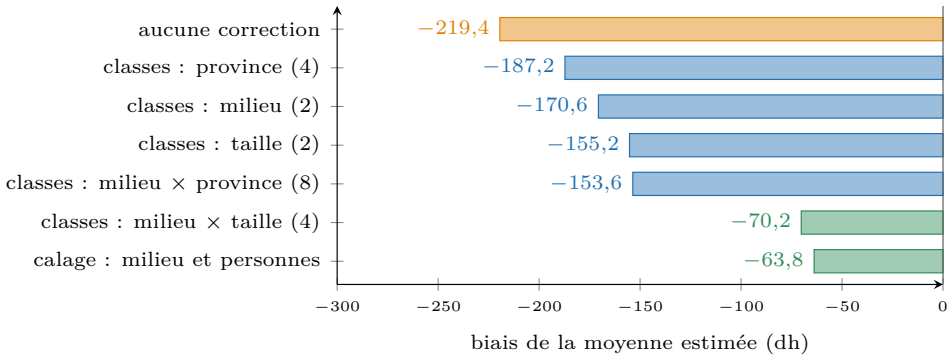
Elle répare la composition

Les urbains répondent moins : ils ne font que 69,2 % des répondants, contre 71,9 % de la population. Rendre à chaque milieu son poids corrige ce déséquilibre, soit 48,8 dh sur 219,4 : 22 % du biais.

Elle ne répare pas les écarts dans chaque classe

Dans chaque milieu, les non-répondants dépensent plus que les répondants : 814 dh de plus en ville, 1 273 dh à la campagne. La repondération suppose cet écart nul. La réponse dépendant de la dépense elle-même, **aucune** classe ne le supprimera entièrement : la non-réponse n'est pas ignorable.

Choisir les classes



De bonnes classes prédisent la dépense

Province et milieu corrigent peu : ils expliquent mal la dépense. La taille du ménage, liée à la dépense, fait mieux ; croisée avec le milieu, elle ramène le biais à −70 dh. Le calage du chapitre 13 sur le milieu et le nombre de personnes donne −64 dh.

Sur l'échantillon fil rouge

161 répondants sur 200 ménages tirés

Milieu	Tirés n_c	Répondants r_c	Poids	Moyenne (dh)
Urbain	152	119	12,77	7 593,6
Rural	48	42	11,43	7 192,4
Ensemble	200	161		7 488,9

Poids $10\,n_c/r_c$, de somme $1\,520 + 480 = 2\,000 = N$. Moyenne repondérée : $0{,}76 \times 7\,593{,}6 + 0{,}24 \times 7\,192{,}4 = 7\,497{,}3$ dh.

Une correction presque nulle sur cet échantillon

Les urbains font 76 % des tirés et déjà 73,9 % des répondants : la correction gagne 8 dh. Or les 39 non-répondants tirés dépensent 9 304,87 dh en moyenne : les 200 réponses auraient donné 7 843,05 dh, 354 dh de plus.

Imputer les valeurs manquantes

Compléter le fichier

Trois façons d'imputer

Le ménage k a répondu à tout, sauf à la question sur la dépense. On remplace la valeur manquante par :

1. la **moyenne** des répondants à la question, dans sa classe ;
2. la valeur d'un **donneur** tiré au hasard parmi ces répondants : le **hot deck** ;
3. une valeur **prédite** par une variable connue, par exemple $\widehat{R}\,x_k$, où x_k est la taille du

ménage (chapitre 12).

Le ménage E ne déclare pas sa dépense

A, B, C, D déclarent 4, 6, 7, 9 milliers de dh ; E, six personnes, ne déclare rien (en vérité : 14).

- Moyenne : on impute 6,5 ; moyenne du fichier complété : 6,5.
- Hot deck : on impute 4, 6, 7 ou 9 ; moyenne complétée : 6, 6,4, 6,6 ou 7.
- Ratio : $\widehat{R} = 26/12$ milliers de dh par personne, on impute $6 \times 26/12 = 13$; moyenne complétée : 7,8, pour $\overline{Y} = 8$.

La moyenne imputée écrase la dispersion

Une variance trop petite

Les valeurs imputées tombent **sur** la moyenne : elles n'ajoutent rien à la somme des carrés des écarts, mais augmentent le nombre d'observations. Avec n_r valeurs observées sur n :

$$s_c^2 = \frac{n_r - 1}{n - 1}\, s_r^2 \qquad \text{ici } s_r^2 = \frac{13}{3} = 4{,}33 \text{ et } s_c^2 = \frac{13}{4} = 3{,}25$$

Une précision imaginaire

Traiter le fichier complété comme n observations donne $(1 - f)\, s_c^2/n$ au lieu de $(1 - f)\, s_r^2/n_r$: la variance est sous-estimée deux fois, par un s^2 trop petit et par un n trop grand. Imputer ne crée **pas** d'information : n lignes complètes ne valent que n_r réponses.

Démonstration : L'imputation par la moyenne écrase la dispersion

La moyenne ne change pas. Les $n - n_r$ valeurs imputées valent $\overline{y}_r$: la moyenne du fichier complété est $\dfrac{n_r\, \overline{y}_r + (n - n_r)\, \overline{y}_r}{n} = \overline{y}_r$.

La somme des carrés ne change pas. Une valeur imputée est à l'écart $\overline{y}_r - \overline{y}_r = 0$ de la moyenne : elle ajoute 0 à la somme des carrés des écarts, qui reste $(n_r - 1)\, s_r^2$. Mais le diviseur passe de $n_r - 1$ à $n - 1$:

$$s_c^2 = \frac{(n_r - 1)\, s_r^2}{n - 1} < s_r^2.$$

Le bilan. La variance « calculée » du fichier complété traité comme n observations est $(1 - f)\, s_c^2/n$, au lieu de $(1 - f)\, s_r^2/n_r$: le rapport est $\dfrac{(n_r - 1)\, n_r}{(n - 1)\, n} < 1$. La précision affichée est **imaginaire**.

Ce que cela dit. Pour A, B, C, D déclarant 4, 6, 7, 9 et E manquant : la somme des carrés vaut 13, $s_r^2 = 13/3$, $s_c^2 = 13/4 = \frac{3}{4}\, s_r^2$. Pour les huit ménages de l'exercice 3, $s_c^2 = \frac{5}{7}\, s_r^2$ (3,33 contre 4,67). Avec 160 valeurs sur 200, le rapport des variances est $\frac{159 \times 160}{199 \times 200} = 0{,}64$, celui des erreurs-types 0,80 ; les simulations du cours donnent 238,1/301,1 = 0,79 pour la moyenne imputée.

■

Le hot deck garde la dispersion

Le prix du tirage du donneur

Avec m valeurs imputées, chacune tirée avec remise parmi les n_r donneurs, le hot deck ajoute à la variance de la moyenne complétée

$$\frac{m\,\sigma_r^2}{n^2} \qquad \sigma_r^2 \text{ : variance des donneurs, divisée par } n_r$$

Pour E : $\sigma_r^2 = 13/4 = 3{,}25$ et $1 \times 3{,}25/25 = 0{,}13$, la variance des quatre moyennes 6 ; 6,4 ; 6,6 ; 7.

Des valeurs réelles, mais un aléa de plus

Le hot deck impute des valeurs **observées** : le fichier complété garde la forme et la dispersion de la distribution, utiles pour une médiane ou un taux de pauvreté. Comme la moyenne imputée, il donne 6,5 en espérance, pour $\overline{Y} = 8$: aucune imputation ne corrige le biais si E se tait **parce qu'**il dépense beaucoup.

Démonstration : Le hot deck, le prix du tirage du donneur

Un tirage de plus. Avec $m = n - n_r$ valeurs manquantes, chacune est remplacée par la valeur d'un donneur tiré au hasard, avec remise, parmi les n_r répondants. Chaque valeur imputée v_i a donc pour espérance $\overline{y}_r$ et pour variance $\sigma_r^2 = \frac{1}{n_r}\sum_j (y_j - \overline{y}_r)^2$, et les v_i sont indépendantes.

La moyenne complétée. Sachant les valeurs observées, $\overline{y}_c = \dfrac{n_r\,\overline{y}_r + \sum_{i=1}^{m} v_i}{n}$ a pour espérance $\overline{y}_r$ et pour variance

$$\frac{m\,\sigma_r^2}{n^2}.$$

Ce terme s'ajoute à la variance du tirage de l'échantillon : c'est le prix d'avoir tiré les donneurs. Les valeurs imputées sont des valeurs **observées** : la forme de la distribution est préservée.

Pas de correction du biais. En espérance, la moyenne complétée est $\overline{y}_r$, comme avec la moyenne imputée : si les non-répondants diffèrent des répondants, le biais subsiste.

Ce que cela dit. Pour E : $\sigma_r^2 = 13/4 = 3{,}25$, et $1 \times 3{,}25/25 = 0{,}13$ est la variance des quatre moyennes complétées 6 ; 6,4 ; 6,6 ; 7, d'espérance 6,5 pour $\overline{Y} = 8$. Pour l'exercice 3, les $6^2 = 36$ fichiers complétés ont pour espérance 7,33 et pour variance $2 \times 3{,}89/64 = 0{,}12$, comme l'énumération de ces 36 fichiers.

■

Dix mille enquêtes à valeurs manquantes

Deux cents ménages, quarante dépenses manquantes au hasard

Imputation	Écart-type réel	Erreur-type calculée	Couverture
Aucune : 160 valeurs seules	306,5	301,1	0,943
Par la moyenne	306,5	238,1	0,875
Hot deck	329,6	266,0	0,885
Par le ratio sur la taille	294,4	260,7	0,916

Erreur-type calculée : formule du sondage aléatoire simple, appliquée aux 160 valeurs renseignées (première ligne) ou aux 200 lignes du fichier complété.

Imputer, puis compter juste

Les valeurs manquent au hasard complet : aucune méthode n'est biaisée. Mais compter les valeurs imputées comme observées donne des intervalles **trop étroits** : 87,5 % de couverture au lieu de 95 % avec la moyenne. Le ratio, qui exploite la taille, est le plus précis.

Prévenir et corriger en pratique

Prévenir vaut mieux que corriger

Pendant la collecte

1. **Relancer** : revenir plusieurs fois, à des jours et des heures différents ; un absent n'est pas un refus.
2. **Former les enquêteurs** : présenter l'enquête, expliquer le secret statistique, savoir revenir.
3. **Soigner le questionnaire** : court et clair ; pour la dépense ou le revenu, proposer des **tranches** à qui refuse de donner un montant.
4. **Suivre les taux de réponse** par enquêteur et par zone pendant la collecte, pour réagir à temps.

Ne pas remplacer un absent par son voisin

Interroger à sa place un voisin qui accepte, c'est faire un hot deck sans le dire : le remplaçant est un **répondant**. Le fichier paraît complet, le taux de réponse affiché est faux, et le biais reste entier.

Relancer un sous-échantillon de non-répondants

Aller chercher l'information chez les absents

Parmi les $n - n_r$ non-répondants, on en tire au hasard un sur k, que l'on relance avec des moyens renforcés jusqu'à obtenir leur réponse. Si $\overline{y}'_{nr}$ est leur moyenne :

$$\widehat{\overline{Y}} = \frac{n_r}{n}\,\overline{y}_r + \frac{n - n_r}{n}\,\overline{y}'_{nr}$$

Sans biais si tous les relancés répondent : chacun représente k non-répondants de

l'échantillon.

Un tiers des non-répondants du fil rouge

On relance 13 des 39 non-répondants ($k = 3$), tirés au hasard (graine 2026) ; ils dépensent en moyenne 8 253,08 dh. L'estimation passe de 7 488,94 dh (répondants seuls) à

$$\widehat{\overline{Y}} = \frac{161}{200} \times 7\,488{,}94 + \frac{39}{200} \times 8\,253{,}08 = 7\,637{,}95 \text{ dh}$$

Démonstration : Relancer un sous-échantillon de non-répondants

L'idée. Parmi les $m = n - n_r$ non-répondants de l'échantillon, on en tire un sous-échantillon aléatoire simple de m' unités, dont la taille ne dépend que de m, et l'on obtient leurs réponses. Soit $\overline{y}_{nr}$ la moyenne des m non-répondants de l'échantillon, **que l'on n'observe pas**, et $\overline{y}'_{nr}$ celle des relancés.

Sans biais, en deux temps. Sachant l'échantillon S, les relancés forment un sondage simple parmi les m non-répondants : $E(\overline{y}'_{nr} \mid S) = \overline{y}_{nr}$. Donc

$$E(\widehat{\overline{Y}} \mid S) = \frac{n_r}{n}\,\overline{y}_r + \frac{n - n_r}{n}\,\overline{y}_{nr} = \overline{y},$$

la moyenne des n ménages tirés, comme s'ils avaient tous répondu. Avec $E(\overline{y}) = \overline{Y}$, on obtient $E(\widehat{\overline{Y}}) = \overline{Y}$: l'estimateur est sans biais **si tous les relancés répondent**. Chaque relancé représente m/m' non-répondants de l'échantillon.

Le prix. Le tirage des relancés ajoute une variance, d'autant plus grande que le sous-échantillon est petit ; mais le biais de non-réponse a disparu.

Ce que cela dit. Sur le fil rouge, $\frac{161\times 7\,488{,}94 + 39\times 9\,304{,}87}{200} = 7\,843{,}05$ dh : c'est la moyenne que les 200 réponses auraient donnée. Avec seulement 13 relancés, de moyenne 8 253,08, l'estimation est 7 637,95 dh. Sur une population de sept unités, l'énumération des 35 échantillons de quatre unités, avec relance de la moitié des non-répondants tirés (arrondie par excès), donne $E(\widehat{\overline{Y}}) = \overline{Y}$ exactement.

■

Le bilan sur l'échantillon fil rouge

Six estimations de la dépense moyenne

Estimation	Valeur (dh)	Écart à $\overline{Y}$
Si les 200 ménages avaient répondu	7 843,05	−68,4
Répondants seuls	7 488,94	−422,5
Classes : milieu	7 497,32	−414,1
Classes : milieu croisé avec la taille	7 596,00	−315,4
Calage : milieu et nombre de personnes	7 684,91	−226,5
Relance d'un tiers des non-répondants	7 637,95	−273,5

Ce que dit le bilan

Toutes les corrections vont dans le bon sens, aucune n'atteint 7 843. Repondérer et caler **supposent** les absents semblables aux répondants ; seule la relance les **observe**, sans biais mais sur 13 ménages.

S'entraîner

Exercice 1 : une enquête sur les revenus

Un taux de réponse de 70 pour cent

Une enquête sur le revenu annuel des ménages obtient un taux de réponse $\tau = 0{,}7$. Les répondants gagnent en moyenne 48 000 dh ; une source administrative révèle que les non-répondants gagnent 60 000 dh. L'écart-type des revenus des répondants est $S_R = 20\,000$ dh ; la population est grande.

1. Calculer $\overline{Y}$ et le biais de la moyenne des répondants.
2. On tire 600 ménages ; 420 répondent, de revenu moyen 48 000 dh. Calculer l'erreur-type de $\overline{y}_r$ et l'intervalle à 95 %. Contient-il $\overline{Y}$? Reprendre avec 1 200 ménages tirés.
3. Avec $\tau = 0{,}9$, quel écart entre répondants et non-répondants donnerait le même biais ?

Exercice 1 : solution

Le biais

1. $\overline{Y} = 0{,}7 \times 48\,000 + 0{,}3 \times 60\,000 = 51\,600$ dh. $B = 0{,}3 \times (48\,000 - 60\,000) = -3\,600$ dh, soit $-6{,}98$ %.

Deux intervalles faux

2. $20\,000/\sqrt{420} = 975{,}9$ dh ; $48\,000 \pm 1\,912{,}8$, soit $[46\,087\,;49\,913]$: 51 600 n'y est pas. Avec 1 200 ménages, 840 répondants : 690,1 dh et $[46\,647{,}5\,;49\,352{,}5]$, plus étroit et toujours faux. L'erreur quadratique moyenne ne passe que de 3 730 à 3 666 dh : le biais domine.
3. $3\,600/0{,}1 = 36\,000$ dh : un taux élevé rend un gros biais improbable, pas impossible.

Exercice 2 : les commerces de Tétouan

Repondérer par zone

On tire 300 commerces parmi 3 000 par sondage aléatoire simple : 180 dans la médina, 120 dans la ville nouvelle. Répondent 117 commerces de la médina, de chiffre d'affaires mensuel moyen 38 milliers de dh, et 108 de la ville nouvelle, de moyenne 62.

1. Calculer les taux de réponse par zone et le taux global.
2. Calculer la moyenne des 225 répondants.
3. Calculer le poids de non-réponse de chaque zone ; vérifier que les poids somment à 3 000. Calculer la moyenne repondérée.
4. Sous quelle hypothèse cette correction supprime-t-elle le biais ?

Exercice 2 : solution

Taux et poids

1. Médina : $117/180 = 0{,}65$; ville nouvelle : $108/120 = 0{,}90$; global : $225/300 = 0{,}75$.
2. $(117 \times 38 + 108 \times 62)/225 = 11\,142/225 = 49{,}52$ milliers de dh.
3. Poids : $10 \times 180/117 = 15{,}38$ et $10 \times 120/108 = 11{,}11$; $117 \times 15{,}38 + 108 \times 11{,}11 = 1\,800 + 1\,200 = 3\,000$. Moyenne repondérée : $0{,}6 \times 38 + 0{,}4 \times 62 = 47{,}6$.

Lecture et hypothèse

La médina fait 60 % des tirés mais 52 % des répondants : la moyenne brute la sous-représente, et dépasse la moyenne repondérée de 1,92 millier de dh (4 %).

4. Si, dans chaque zone, les commerces qui refusent ont en moyenne le même chiffre d'affaires que ceux qui répondent.

Exercice 3 : deux dépenses manquantes

Imputer de trois façons

Huit ménages de 2, 3, 3, 4, 4, 5, 6 et 7 personnes. Les six premiers déclarent 4, 6, 7, 8, 9 et 10 milliers de dh ; les deux derniers ne déclarent pas leur dépense.

1. Imputer par la moyenne : moyenne du fichier complété, s_r^2 et s_c^2.
2. Imputer par hot deck (donneurs tirés avec remise) : combien de fichiers complétés possibles ? Espérance et variance de la moyenne complétée ?
3. Imputer par le ratio sur la taille. Quelle imputation est la plus crédible ici ?

Exercice 3 : solution

Moyenne et hot deck

1. $\overline{y}_r = 44/6 = 7{,}33$, imputé deux fois : moyenne 7,33. Somme des carrés 23,33 : $s_r^2 = 23{,}33/5 = 4{,}67$ et $s_c^2 = 23{,}33/7 = 3{,}33 = \frac{5}{7}\, s_r^2$.
2. $6^2 = 36$ fichiers ; espérance 7,33 ; variance $2\,\sigma_r^2/8^2 = 2 \times 3{,}89/64 = 0{,}12$, avec $\sigma_r^2 = 23{,}33/6$.

Ratio

3. $\widehat{R} = 44/21 = 2{,}095$; on impute $6 \times 2{,}095 = 12{,}57$ et $7 \times 2{,}095 = 14{,}67$; moyenne complétée 8,90. Les deux ménages muets sont les plus grands (6,5 personnes contre 3,5), et leur imputer la moyenne des petits les sous-estime. Le ratio est plus crédible, si la réponse ne dépend que de la taille.

Ce qu'il faut retenir

Le résumé du chapitre

Sept points

1. Une ligne manque : on **repondère** ; une case manque : on **impute**.
2. Biais de la moyenne des répondants : $B = (1-\tau)(\overline{Y}_R - \overline{Y}_{NR})$, ici $-219{,}4$ dh.
3. Ce biais ne diminue pas avec n : les grandes enquêtes donnent des intervalles faux.
4. Repondérer par classes, $w_k = d_k\, n_c/r_c$, corrige la composition : $-170{,}6$ dh avec le milieu.
5. Correction complète si l'on répond au hasard dans chaque classe ; jamais si la réponse dépend de y.
6. Imputer complète le fichier, pas l'information : la variance naïve est trop petite.
7. Prévenir et relancer valent mieux que corriger.

La suite

Le chapitre 15 met en œuvre les plans et les estimateurs du cours sous R et sous Python : tirages, poids, variances, post-stratification et calage.