Section 10

Étude de cas 3 : échantillonnage à deux phases et erreur de mesure

Lectures : Breslow et Chatterjee (1999), *Applied Statistics* 48(4) : 457–468 ; Wu (2010), p. 101 ; Tsiatis (2006), ch. 7 et 8 ; notes des sections 8 et 9

Julien Miron – notes rédigées à partir du matériel de Lajmi Lakhal-Chaieb et Alexandre Bureau

Objectifs de la séance

1. Décrire un devis à deux phases et l'interpréter comme un problème de données manquantes *par devis*, dont le mécanisme est connu.
2. Construire et comparer les trois estimateurs étudiés par Breslow et Chatterjee : vraisemblance pondérée (Horvitz-Thompson), pseudo-vraisemblance (Breslow-Cain) et maximum de vraisemblance semi-paramétrique.
3. Dériver la variance sandwich de l'estimateur pondéré en tenant compte du caractère aléatoire des effectifs de phase 1.
4. Relier le devis à l'erreur de mesure (variable de stratification comme substitut imparfait) et à la pondération par probabilité inverse de la section 9.
5. Préparer l'étude de cas 3 (document de consignes de l'étude de cas 3).

1 Mise en contexte : la tumeur de Wilms

La tumeur de Wilms (néphroblastome) est le cancer du rein le plus fréquent chez l'enfant. Le *National Wilms Tumor Study Group* (NWTS) a mené une série d'essais cliniques nord-américains. Deux facteurs pronostiques dominent la rechute : le **stade** de la maladie (I à IV) et l'**histologie** de la tumeur, favorable (FH) ou défavorable (UH, notamment l'anaplasie). L'histologie est évaluée par le pathologiste de l'hôpital participant (*histologie institutionnelle*) et, de façon plus fiable mais plus coûteuse, par le laboratoire central de pathologie de l'étude (*histologie centrale*).

Breslow et Chatterjee exploitent une situation rare : dans les essais NWTS-3 et NWTS-4, l'histologie centrale est disponible pour *toute* la cohorte ($N = 4088$ enfants, 571 rechutes). On peut donc simuler des sous-échantillonnages réalistes, où l'histologie centrale ne serait mesurée que chez quelques centaines d'enfants, et comparer les estimations obtenues à celles de la cohorte complète, qui servent d'étalon.

Le croisement des deux histologies dans la cohorte montre que l'histologie institutionnelle est un *substitut imparfait* de l'histologie centrale :

	Histologie centrale		
Histologie institutionnelle	FH	UH	Total
FH	3547	130	3677
UH	76	335	411
Total	3623	465	4088

La sensibilité de l'histologie institutionnelle pour détecter une histologie défavorable est $335/465 = 0{,}72$ et sa spécificité est $3547/3623 = 0{,}98$. Remplacer l'histologie centrale par l'histologie institutionnelle dans un modèle logistique de la rechute atténue l'effet estimé (section 7) : c'est l'effet classique d'une erreur de classification non différentielle.

Remarque 1.1. Les fichiers fournis pour l'étude de cas ne contiennent que le stade, l'histologie institutionnelle, l'histologie centrale et la rechute. L'article considère aussi d'autres covariables (âge au diagnostic, notamment) ; voir l'article pour la spécification exacte des modèles de ses tableaux.

2 Le devis à deux phases

2.1 Définition et notation

Définition 2.1 (Devis à deux phases)**.** — **Phase 1.** Une cohorte de N sujets i.i.d. est observée pour la réponse $Y \in \{0,1\}$ et pour des variables S peu coûteuses, discrètes, qui définissent J strates ($S \in \{1,\ldots,J\}$). On note N_{ij} le nombre de sujets de phase 1 dans la cellule $(Y = i, S = j)$, $i \in \{0,1\}$.

— **Phase 2.** Dans chaque cellule (i,j), on tire sans remise un échantillon aléatoire simple de taille $n_{ij} \leq N_{ij}$, fixée par le devis. Chez ces sujets, on mesure des variables coûteuses X.

On note $R_k = 1$ si le sujet k est sélectionné en phase 2, et $\pi_{ij} = n_{ij}/N_{ij}$ la probabilité de sélection (conditionnelle à la phase 1) dans la cellule (i,j).

Dans l'exemple : Y = rechute ; S = stade (4 niveaux) × histologie institutionnelle (2 niveaux), soit $J = 8$ strates ; X = histologie centrale. La figure 1 schématise le devis.

2.2 Trois devis de phase 2

Avec les mêmes N_{ij}, on peut répartir un budget de $n = \sum_{ij} n_{ij}$ mesures de diverses façons.

— **Devis cas-témoins** : tous les cas ($n_{1j} = N_{1j}$) et un échantillon aléatoire de témoins, sans tenir compte de S ; les n_{0j} sont alors aléatoires si l'on tire les témoins globalement, ou proportionnels aux N_{0j} si l'on stratifie.

— **Devis équilibré** : on vise des effectifs n_{ij} aussi égaux que possible dans les $2J$ cellules, en prenant tous les sujets des cellules trop petites. On suréchantillonne ainsi les combinaisons rares (par exemple les enfants avec histologie institutionnelle défavorable), qui sont les plus informatives pour les effets d'histologie et les interactions avec le stade.

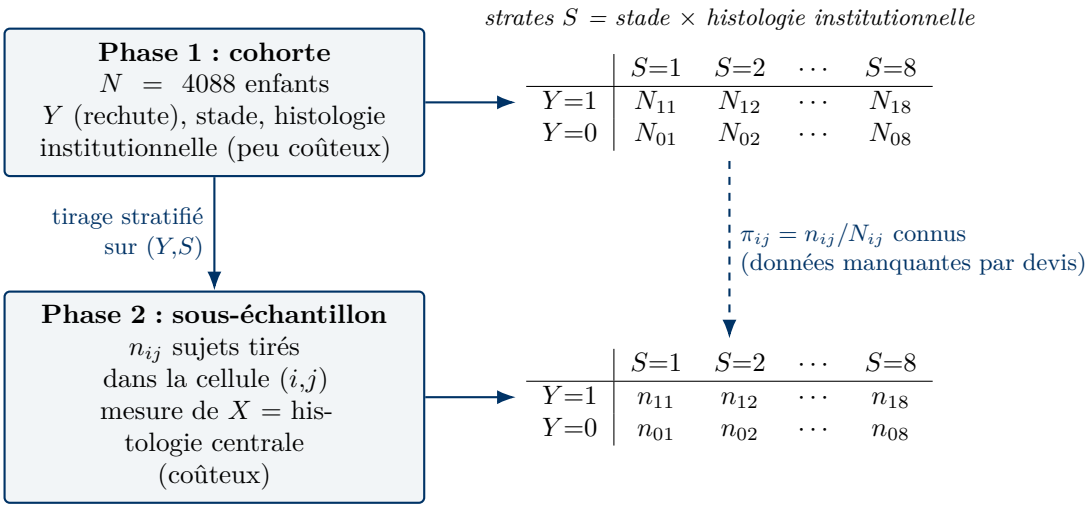


FIGURE 1 – Schéma d'un devis à deux phases stratifié sur la réponse Y et sur les variables de phase 1 S. Les sujets non retenus en phase 2 ont X manquant, avec une probabilité de sélection connue qui ne dépend que de (Y,S).

— **Cohorte complète** : $n_{ij} = N_{ij}$; référence (étalon), inaccessible en pratique quand X est coûteux.

Dans les fichiers de l'étude de cas, les échantillons cas-témoins (`NWTSc.dat`) et équilibré (`nwtsb.dat`) comptent chacun $1142 = 571 + 571$ sujets. Dans l'échantillon équilibré fourni, on retrouve tous les cas, tous les sujets à histologie institutionnelle défavorable, et 316 témoins à histologie institutionnelle favorable répartis entre les quatre stades.

2.3 Un problème de données manquantes par devis

Du point de vue des sections 8 et 9, X est manquant chez les $N - n$ sujets non retenus. Le mécanisme est

$$\mathbb{P}(R = 1 \mid Y, S, X) = \mathbb{P}(R = 1 \mid Y, S) = \pi_{YS},$$

qui ne dépend que de quantités toujours observées : les données sont **manquantes au hasard (MAR) par construction**, et les probabilités de sélection sont **connues** (fixées par l'enquêteur), ce qui est exceptionnel en données manquantes. Toutes les méthodes de la séance tirent parti de cette structure. En revanche, le mécanisme dépend de Y : une analyse des seuls cas complets, qui ignorerait le devis, serait biaisée pour tout coefficient autre que les rapports de cotes de X dans un devis cas-témoins pur.

2.4 Le modèle d'intérêt

L'objet d'intérêt est le vecteur β du modèle logistique

$$\mathbb{P}(Y = 1 \mid X = x) = p(x; \beta) = \frac{\exp(x^\top \beta)}{1 + \exp(x^\top \beta)}, \tag{1}$$

où x contient les variables de phase 2 (histologie centrale) et éventuellement des variables de phase 1 (stade). Lorsque les variables de S ne figurent pas toutes dans x, on suppose

$$\mathbb{P}(Y = 1 \mid X, S) = \mathbb{P}(Y = 1 \mid X), \tag{2}$$

c'est-à-dire que S est un *substitut* de X : une fois l'histologie centrale connue, l'histologie institutionnelle n'apporte plus d'information sur la rechute (erreur de mesure non différentielle). Dans la cohorte NWTS, un test du rapport des vraisemblances de l'ajout de l'histologie institutionnelle et de toutes ses interactions au modèle stade × histologie centrale donne $\chi^2_8 = 3{,}03$ ($p = 0{,}93$) : l'hypothèse est plausible.

À retenir

- Phase 1 : (Y, S) chez tous ; phase 2 : X chez un sous-échantillon stratifié sur (Y, S), avec $\pi_{ij} = n_{ij}/N_{ij}$.
- X est manquant au hasard *par devis*, avec des probabilités de sélection *connues*.
- On estime un modèle logistique pour $\mathbb{P}(Y = 1 \mid X)$, en supposant au besoin que S est un substitut non différentiel de X.

3 Estimateur par vraisemblance pondérée (Horvitz-Thompson)

3.1 Construction

Si X était connu chez tous, on résoudrait l'équation du score de la cohorte

$$U_N(\beta) = \sum_{k=1}^{N} u_k(\beta), \qquad u_k(\beta) = x_k\{y_k - p(x_k; \beta)\}.$$

Le principe de Horvitz et Thompson remplace cette somme sur la cohorte par une somme sur la phase 2 où chaque sujet compte pour l'inverse de sa probabilité de sélection :

$$U_W(\beta) = \sum_{k=1}^{N} \frac{R_k}{\pi_{y_k s_k}} u_k(\beta) = \sum_{i,j} \frac{N_{ij}}{n_{ij}} \sum_{k \in \text{phase 2},\, (y_k, s_k) = (i,j)} u_k(\beta). \tag{3}$$

L'estimateur $\hat{\beta}_W$ résout $U_W(\beta) = 0$. C'est une régression logistique ordinaire sur les données de phase 2 avec des poids $w_k = N_{ij}/n_{ij}$ (Flanders et Greenland, 1991 ; Zhao et Lipsitz, 1992). Il est sans biais pour l'équation de la cohorte : conditionnellement à la phase 1, $\mathrm{E}[R_k \mid \text{phase 1}] = \pi_{y_k s_k}$, donc

$$\mathrm{E}\big[U_W(\beta) \mid \text{phase 1}\big] = U_N(\beta) \quad \Longrightarrow \quad \mathrm{E}[U_W(\beta_0)] = \mathrm{E}[U_N(\beta_0)] = 0.$$

Propriétés. L'estimateur pondéré est simple, se calcule avec n'importe quel logiciel de régression logistique acceptant des poids, et vise le paramètre que l'on estimerait dans la cohorte complète (le paramètre de « recensement ») *même si le modèle* (1) *est mal spécifié* et même si l'hypothèse de substitut (2) est fausse : il n'utilise que le fait que la phase 2 est une image réduite, sans biais, de la cohorte. Sa faiblesse est son efficacité : les sujets non sélectionnés n'interviennent que par les poids, et l'information que la phase 1 contient sur les coefficients des variables de S (ici le stade) est largement perdue. Les poids très variables (cellules très sous-échantillonnées) gonflent la variance.

3.2 Variance : la viande du sandwich

Le développement habituel d'un estimateur M (section 1) donne

$$\hat\beta_W - \beta_0 \approx \mathcal{I}_W^{-1}\, U_W(\beta_0), \qquad \mathcal{I}_W = -\frac{\partial U_W}{\partial \beta^\top} = \sum_k \frac{R_k}{\pi_{y_k s_k}} p_k(1-p_k)\, x_k x_k^\top,$$

d'où $\mathrm{Var}(\hat\beta_W) \approx \mathcal{I}_W^{-1}\,\mathrm{Var}\{U_W(\beta_0)\}\mathcal{I}_W^{-1}$. Toute la difficulté est dans la « viande » $\mathrm{Var}\{U_W(\beta_0)\}$. Le point clé (question 1 de l'étude de cas) est que **les N_{ij} sont aléatoires** : la phase 1 est elle-même un échantillon de la population.

Proposition 3.1. *Supposons les n_{ij} fixés étant donné la phase 1, avec tirage aléatoire simple sans remise dans chaque cellule. Notons $\Sigma = \mathrm{E}[u(\beta_0)u(\beta_0)^\top]$ et, pour la cellule $c=(i,j)$, S_c^2 la matrice de covariance (de population finie) des u_k parmi les N_c sujets de phase 1 de la cellule. Alors*

$$\mathrm{Var}\{U_W(\beta_0)\} = N\,\Sigma + \mathrm{E}\Big[\sum_c \frac{N_c(N_c-n_c)}{n_c}\, S_c^2\Big].$$

Démonstration. Par la loi de la variance totale, en conditionnant sur la phase 1 (qui fixe les N_c, les u_k des N sujets et donc les S_c^2) :

$$\mathrm{Var}(U_W) = \mathrm{Var}\{\mathrm{E}(U_W \mid \text{ph. 1})\} + \mathrm{E}\{\mathrm{Var}(U_W \mid \text{ph. 1})\}.$$

Premier terme. $\mathrm{E}(U_W \mid \text{ph. 1}) = U_N = \sum_{k=1}^N u_k$ est une somme de N vecteurs i.i.d. centrés : sa variance est $N\Sigma$. C'est la variance qu'aurait le score si l'on observait toute la cohorte.

Second terme. Conditionnellement à la phase 1, les tirages dans les différentes cellules sont indépendants, et $U_W = \sum_c N_c\,\bar u_c$, où $\bar u_c$ est la moyenne des u_k dans un échantillon aléatoire simple de taille n_c tiré sans remise parmi N_c. Le résultat classique de l'échantillonnage aléatoire simple donne $\mathrm{Var}(\bar u_c \mid \text{ph. 1}) = (1-n_c/N_c)S_c^2/n_c$, d'où

$$\mathrm{Var}(U_W \mid \text{ph. 1}) = \sum_c N_c^2\Big(1-\frac{n_c}{N_c}\Big)\frac{S_c^2}{n_c} = \sum_c \frac{N_c(N_c-n_c)}{n_c}\, S_c^2. \qquad \square$$

La viande est donc la somme d'une composante de **phase 1** (la variabilité qu'aurait l'analyse de la cohorte complète) et d'une composante de **phase 2** (le prix du sous-échantillonnage), qui s'annule dans les cellules entièrement échantillonnées ($n_c = N_c$). On l'estime en remplaçant $N\Sigma$ par l'estimateur de Horvitz-Thompson $\sum_k (R_k/\pi_k)\,\hat u_k\hat u_k^\top$ et S_c^2 par la covariance empirique des $\hat u_k$ de la cellule c en phase 2 :

$$\hat B = \sum_{k:\,R_k=1} w_k\,\hat u_k\hat u_k^\top + \sum_c \frac{N_c(N_c-n_c)}{n_c}\,\hat S_c^2, \qquad \widehat{\mathrm{Var}}(\hat\beta_W) = \hat{\mathcal{I}}_W^{-1}\hat B\,\hat{\mathcal{I}}_W^{-1}. \tag{4}$$

C'est, à la notation près, la structure de la formule (3) de l'article (à confronter avec le texte lors de la lecture).

Comparaison avec l'estimateur sandwich « naïf ». Un logiciel qui traite les observations pondérées comme indépendantes calcule la viande $\sum_k w_k^2\hat u_k\hat u_k^\top$. En décomposant les sommes par cellule ($\sum_{k\in c}\hat u_k\hat u_k^\top \approx n_c(\hat S_c^2 + \bar u_c\bar u_c^\top)$), on trouve

$$\sum_k w_k^2\hat u_k\hat u_k^\top - \hat B \approx \sum_c N_c\Big(\frac{N_c}{n_c}-1\Big)\bar u_c\bar u_c^\top \;\succeq\; 0.$$

L'estimateur naïf est *conservateur* : il ignore que les effectifs de chaque cellule sont fixés par le devis, ce qui élimine la variabilité entre cellules (c'est le gain de la stratification). L'écart est important pour les coefficients des variables de stratification (ici le stade et l'ordonnée à l'origine), et faible pour ceux des variables de phase 2.

Remarque 3.2 (Quand les n_{ij} ne sont pas fixés). La proposition 3.1 suppose les n_{ij} fixés. Si l'on fixe seulement le nombre total de témoins (devis cas-témoins non stratifié), les n_{0j} deviennent aléatoires et des termes s'ajoutent. À l'autre extrême, si chaque sujet est retenu indépendamment avec probabilité π_{ij} (échantillonnage de Bernoulli), on a $\mathrm{Var}(U_W \mid \text{ph. } 1) = \sum_k (1-\pi_k)/\pi_k \, u_k u_k^\top$ et la viande totale vaut $\mathrm{E}[\sum_k u_k u_k^\top/\pi_k]$, exactement ce qu'estime le sandwich naïf : celui-ci est alors correct.

À retenir

- Vraisemblance pondérée : régression logistique sur la phase 2 avec poids N_{ij}/n_{ij} ; sans biais pour le score de la cohorte.
- Robuste (vise le paramètre de recensement même sous mauvaise spécification), mais peu efficace pour les variables de phase 1.
- Viande = phase 1 ($N\Sigma$) + phase 2 ($\sum_c N_c(N_c - n_c)S_c^2/n_c$) ; le sandwich naïf est conservateur quand les n_{ij} sont fixés.

4 Pseudo-vraisemblance (Breslow-Cain)

4.1 La vraisemblance conditionnelle à la sélection

L'idée de Breslow et Cain (1988) est de modéliser la loi de Y *chez les sujets de phase 2*, en conditionnant sur le fait d'avoir été sélectionné. Par la formule de Bayes, et puisque la sélection ne dépend que de (Y,S),

$$\mathbb{P}(Y=1 \mid X=x, S=j, R=1) = \frac{\pi_{1j}\, p(x;\beta)}{\pi_{1j}\, p(x;\beta) + \pi_{0j}\{1-p(x;\beta)\}},$$

où l'on a utilisé l'hypothèse (2). En passant à la cote,

$$\operatorname{logit} \mathbb{P}(Y=1 \mid x, j, R=1) = x^\top\beta + \log\frac{\pi_{1j}}{\pi_{0j}} = x^\top\beta + \log\frac{n_{1j}N_{0j}}{n_{0j}N_{1j}}. \tag{5}$$

Dans l'échantillon de phase 2, le modèle reste logistique avec les mêmes coefficients, à un *décalage* (*offset*) connu près, propre à chaque strate. On reconnaît la généralisation du résultat de Prentice et Pyke (1979) pour les études cas-témoins, où le décalage ne touche que l'ordonnée à l'origine.

4.2 Pourquoi « pseudo » ?

Écrivons la vraisemblance des données sous la forme d'un produit de deux facteurs (notes H26) :

$$L_1 = \prod_{i,j} \mathbb{P}(Y=i \mid S=j)^{N_{ij}}, \qquad L_2 = \prod_{k:\, R_k=1} \mathbb{P}(Y=y_k \mid X=x_k, S=s_k, R=1).$$

L_1 est la vraisemblance multinomiale des effectifs de phase 1. La vraie contribution de la phase 2 à la vraisemblance des données observées est *rétrospective* : $\prod_k \mathbb{P}(X=x_k \mid Y=y_k, S=s_k)$, qui

fait intervenir la loi inconnue de X sachant S (paramètre de nuisance de dimension infinie). Par la formule de Bayes,

$$\mathbb{P}(Y \mid X, S, R=1) = \frac{\mathbb{P}(X \mid Y, S, R=1)\,\mathbb{P}(Y \mid S, R=1)}{\mathbb{P}(X \mid S, R=1)},$$

de sorte que L_2 n'est pas proportionnelle à ce produit rétrospectif : le dénominateur dépend de la loi de X. La pseudo-vraisemblance remplace la contribution rétrospective par la vraisemblance *prospective* conditionnelle L_2, qui ne dépend de la loi de X qu'à travers le décalage (5). Les probabilités de sélection $\pi_{ij} = \mathbb{P}(R=1 \mid Y=i, S=j)$ qui y figurent ne dépendent pas de β ; on les fixe à leurs valeurs observées n_{ij}/N_{ij}, et la loi de X sachant S n'est pas modélisée. Maximiser L_2 ainsi construit n'est pas maximiser la vraisemblance complète : c'est une *pseudo-vraisemblance*, dans laquelle on substitue des estimations de phase 1 aux quantités inconnues. (Les notes H26 décrivaient L_2 comme proportionnelle au produit rétrospectif ; ce n'est vrai qu'à un facteur près qui dépend de la loi de X.)

Calcul. $\hat{\beta}_{PL}$ s'obtient par une régression logistique ordinaire des données de phase 2, avec la variable de décalage $\log\{n_{1j}N_{0j}/(n_{0j}N_{1j})\}$.

Variance. La matrice de covariance donnée par `glm` pour ce modèle (l'inverse de l'information de L_2) n'est pas la bonne : elle ignore l'information de L_1 et la variabilité des N_{ij}. Breslow et Cain (1988) donnent une correction qui soustrait de $\mathcal{I}_2^{-1}$ un terme lié à l'information de phase 1 sur les $\mathbb{P}(Y \mid S)$; la fonction `tps` fournit une version fondée sur le modèle et une version empirique (sandwich). Nous ne détaillons pas la formule (voir l'article et Breslow et Cain, 1988).

Propriétés. La pseudo-vraisemblance exige que le modèle soit bien spécifié et que l'hypothèse de substitut tienne. Elle récupère une bonne partie de l'information de phase 1 et est en général beaucoup plus efficace que la vraisemblance pondérée, souvent proche du maximum de vraisemblance.

5 Maximum de vraisemblance semi-paramétrique

5.1 La vraisemblance complète

Avec F_j la loi de X dans la strate $S=j$, la vraisemblance des données observées (phase 1 et phase 2) est, à des facteurs indépendants des paramètres près,

$$L(\beta, F) = \prod_{k:\, R_k=1} p(x_k;\beta)^{y_k}\{1-p(x_k;\beta)\}^{1-y_k}\,\mathrm{d}F_{s_k}(x_k) \times \prod_{k:\, R_k=0} \int p(x;\beta)^{y_k}\{1-p(x;\beta)\}^{1-y_k}\,\mathrm{d}F_{s_k}(x). \tag{6}$$

Les sujets de phase 2 contribuent leur loi conjointe (Y,X) sachant S ; les autres contribuent la loi marginale de Y sachant S, obtenue en intégrant sur la loi de la variable manquante (voir l'encadré « Niveau doctorat »). Les probabilités de sélection n'apparaissent pas : le mécanisme est ignorable (MAR et paramètres distincts).

5.2 Pourquoi « non paramétrique » ?

On ne veut pas postuler de forme pour F_j : c'est un paramètre nuisible de dimension infinie, et le modèle est semi-paramétrique. Le maximum de vraisemblance non paramétrique de F_j place sa masse sur les valeurs de X observées en phase 2 dans la strate j. Le problème devient de dimension finie (mais croissante), et Scott et Wild (1997) ainsi que Breslow et Holubkov (1997) montrent que l'on peut profiler les masses de F_j et ramener le calcul à un système d'équations en β et en $2J$ paramètres auxiliaires seulement ; la fonction `tps` résout ces équations par un algorithme de Newton-Raphson modifié.

Quand X est discret, comme l'histologie centrale, F_j est simplement un vecteur de probabilités par strate et l'on peut maximiser (6) directement (section 7).

Propriétés. L'estimateur est semi-paramétriquement efficace : sa variance est la plus petite parmi les estimateurs réguliers dans le modèle où (1) et (2) sont vrais. En contrepartie, il dépend de ces hypothèses ; sous mauvaise spécification, il ne vise pas le paramètre de recensement, contrairement à l'estimateur pondéré.

5.3 Efficacité relative et choix du devis

Breslow et Chatterjee comparent les trois estimateurs et les devis cas-témoins et équilibré, en prenant l'analyse de la cohorte complète comme étalon. Les messages qualitatifs à retenir sont les suivants (voir les tableaux de l'article pour les valeurs) :

— le maximum de vraisemblance et la pseudo-vraisemblance donnent des résultats très voisins, et nettement plus précis que la vraisemblance pondérée, surtout pour les coefficients des variables de phase 1 (stade) ;

— le devis équilibré améliore beaucoup la précision des effets d'histologie et des interactions stade × histologie par rapport au devis cas-témoins, parce qu'il suréchantillonne les sujets à histologie institutionnelle défavorable, qui sont presque tous UH en histologie centrale ;

— avec un sous-échantillon de 1142 sujets sur 4088, on retrouve une précision proche de celle de la cohorte complète pour plusieurs coefficients.

5.4 Exigence de strates discrètes

Les trois méthodes, telles que formulées, exigent que la phase 1 soit résumée par un nombre fini de cellules (Y,S) : les poids N_{ij}/n_{ij}, les décalages $\log\{n_{1j}N_{0j}/(n_{0j}N_{1j})\}$ et les lois F_j sont définis par cellule. Une variable de phase 1 continue (l'âge) ne peut donc servir à stratifier qu'après discrétisation. En revanche, les covariables X mesurées en phase 2 peuvent être continues, et une variable de phase 1 peut entrer dans le modèle (1) sous forme continue même si la stratification utilise une version catégorisée (au prix de l'hypothèse (2) pour la pseudo-vraisemblance et le maximum de vraisemblance).

À retenir

- Pseudo-vraisemblance : logistique sur la phase 2 avec décalage $\log(\pi_{1j}/\pi_{0j})$; le terme $\mathbb{P}(R=1 \mid Y,S)$ est fixé à n_{ij}/N_{ij}.
- Maximum de vraisemblance : intègre X manquant contre une loi F_j non paramétrique ; efficace sous le modèle.
- Efficacité : ML $\gtrsim$ PL $\gg$ WL pour les variables de phase 1 ; robustesse : WL $>$ PL, ML.
- Toutes exigent des strates discrètes.

6 Liens avec l'erreur de mesure et la pondération par probabilité inverse

6.1 Erreur de mesure et échantillon de validation

Le devis à deux phases est une façon de traiter l'erreur de mesure : la variable de phase 1 (histologie institutionnelle) est une mesure imparfaite de la variable d'intérêt (histologie centrale), et la phase 2 est un *échantillon de validation* où l'on mesure la vraie valeur. L'analyse naïve qui remplace X par son substitut dans toute la cohorte est biaisée (atténuation) ; l'analyse des seuls sujets validés est sans biais mais gaspille l'information de phase 1. Les estimateurs de la séance combinent les deux sources. La fonction `tps` est d'ailleurs documentée comme pouvant traiter « certains problèmes de valeurs manquantes et d'erreur de mesure avec données de validation ». L'hypothèse d'erreur non différentielle correspond exactement à (2).

6.2 Lien avec la pondération par probabilité inverse

L'estimateur pondéré (3) est l'estimateur IPW de la section 9 dans le cas particulier où les probabilités sont connues et ne dépendent que de variables discrètes toujours observées. Deux résultats de la section 9 s'appliquent :

- estimer les probabilités de sélection (même connues) par un modèle qui inclut des variables auxiliaires de phase 1 améliore l'efficacité ; c'est le principe de la *calibration* des poids (Breslow et coll., 2009) ;
- les estimateurs *augmentés* (AIPW) ajoutent un terme d'espérance conditionnelle qui récupère l'information des sujets non sélectionnés (encadré « Niveau doctorat »).

7 Mise en œuvre en R

7.1 Données

Les fichiers de l'étude de cas sont décrits dans le document de consignes. Le fichier `nwtsb_indiv.csv` contient la cohorte entière au niveau individuel (4088 lignes), avec l'histologie centrale (`histol`) renseignée seulement pour les 1142 sujets de l'échantillon équilibré, et la variable `strat8` = stade $+\,4\times$(histologie institutionnelle $-\,1$).

```
dossier <- "étude de cas 3"          # à adapter
coh <- read.table(file.path(dossier, "nwts.dat"), header = TRUE)   # cohorte groupée
ind <- read.csv2(file.path(dossier, "nwtsb_indiv.csv"))           # cohorte
    individuelle
ind$R <- as.integer(!is.na(ind$histol))      # R = 1 : sélectionné en phase 2
N <- with(ind, table(strat8, y))             # effectifs de phase 1
n <- with(subset(ind, R == 1), table(strat8, y))   # effectifs de phase 2
cbind(N, n)
```

```
     0   1   0   1
1 1400 100 117 100
2  841 131  89 131
3  707 115  82 115
4  314  69  28  69
5   79  17  79  17
6   62  35  62  35
7   78  60  78  60
8   36  44  36  44
```

Les deux premières colonnes de chiffres sont N_{0j}, N_{1j}, les deux suivantes n_{0j}, n_{1j}. Seuls les témoins des strates 1 à 4 (histologie institutionnelle favorable) sont sous-échantillonnés, avec des poids de 11,97, 9,45, 8,62 et 11,21 ; tous les autres sujets ont un poids de 1.

7.2 Référence et analyse naïve

Pour illustrer sans reproduire les tableaux de l'article (question 5 de l'étude de cas), on utilise un modèle à effets principaux, $\operatorname{logit} p = \beta_0 + \beta_{\text{stade}} + \beta_H \, \mathbf{1}\{\text{UH centrale}\}$. Dans la cohorte complète :

```
fit_coh <- glm(cbind(case, control) ~ factor(stage) + I(histol - 1),
               family = binomial, data = coh)
fit_nai <- glm(cbind(case, control) ~ factor(stage) + I(instit - 1),
               family = binomial, data = coh)    # substitut à la place de X
tab <- cbind(coef(summary(fit_coh))[, 1:2], coef(summary(fit_nai))[, 1:2])
rownames(tab)[5] <- "histologie"
colnames(tab) <- c("centrale", "ET", "instit.", "ET")
round(tab, 3)
```

```
             centrale    ET instit.    ET
(Intercept)    -2.799 0.102  -2.687 0.099
factor(stage)2  0.795 0.132   0.802 0.130
factor(stage)3  0.911 0.132   0.900 0.130
factor(stage)4  1.325 0.151   1.243 0.149
histologie      1.748 0.111   1.435 0.117
```

Le log du rapport de cotes de l'histologie défavorable passe de 1,75 (histologie centrale) à 1,44 (histologie institutionnelle) : atténuation due à l'erreur de classification.

7.3 Vraisemblance pondérée avec la variance à deux phases

```
ph2 <- subset(ind, R == 1)
ph2$hist <- ph2$histol - 1                    # 1 = histologie centrale défavorable
ph2$w <- N[cbind(ph2$strat8, ph2$y + 1)] / n[cbind(ph2$strat8, ph2$y + 1)]
```

```
fit_wl <- glm(y ~ factor(stage) + hist, family = quasibinomial,
              weights = w, data = ph2)
X <- model.matrix(fit_wl); p <- fitted(fit_wl)
U <- X * (ph2$y - p)                          # contributions au score u_k
A <- crossprod(X * sqrt(ph2$w * p * (1 - p)))  # "pain" : I_W
B_naif <- crossprod(U * ph2$w)                # viande naive : sum w^2 u u'
B_ph1  <- crossprod(U * sqrt(ph2$w))          # phase 1 : sum w u u'
cell <- interaction(ph2$strat8, ph2$y, drop = TRUE)
B_ph2 <- Reduce('+', lapply(split(seq_len(nrow(U)), cell), function(k) {
  Nc <- ph2$w[k[1]] * length(k); nc <- length(k)
  if (nc == Nc) return(0 * B_ph1)              # cellule entièrement échantillonnée
  Nc * (Nc - nc) / nc * cov(U[k, , drop = FALSE])
}))
Ainv <- solve(A)
se <- function(B) sqrt(diag(Ainv %*% B %*% Ainv))
cbind(coef = coef(fit_wl), ET_naif = se(B_naif), ET_2phases = se(B_ph1 + B_ph2))
```

```
                coef ET_naif ET_2phases
(Intercept)    -2.761   0.131      0.106
factor(stage)2  0.762   0.180      0.141
factor(stage)3  0.861   0.178      0.140
factor(stage)4  1.280   0.228      0.158
hist            1.691   0.162      0.159
```

Comme prévu, l'erreur type naïve est presque correcte pour l'histologie (variable de phase 2) mais nettement trop grande pour l'ordonnée à l'origine et le stade (variables de stratification). Le module `survey` implante les plans à deux phases (fonction `twophase`) et reproduit exactement ces erreurs types :

```
library(survey)
ind$cell <- interaction(ind$strat8, ind$y); ind$hist <- ind$histol - 1
des <- twophase(id = list(~1, ~1), strata = list(NULL, ~cell),
                subset = ~as.logical(R), data = ind)
fit_svy <- svyglm(y ~ factor(stage) + hist, design = des, family = quasibinomial)
cbind(coef = coef(fit_svy), ET = SE(fit_svy))
```

```
                coef    ET
(Intercept)    -2.761 0.106
factor(stage)2  0.762 0.141
factor(stage)3  0.861 0.140
factor(stage)4  1.280 0.158
hist            1.691 0.159
```

7.4 Maximum de vraisemblance avec X discret

L'histologie centrale étant dichotomique, F_j se réduit à $\pi_j = \mathbb{P}(X = 1 \mid S = j)$, et la vraisemblance (6) se maximise directement ($5 + 8$ paramètres) :

```
nonsel <- subset(ind, R == 0)          # témoins non sélectionnés (X manquant)
lvrais <- function(par) {
  beta <- par[1:5]; pi_j <- plogis(par[6:13])          # P(X = 1 | S = j)
  eta <- function(d, x) beta[1] + c(0, beta[2:4])[d$stage] + beta[5] * x
  py  <- function(d, x) { q <- plogis(eta(d, x)); ifelse(d$y == 1, q, 1 - q) }
```

```
  px  <- function(d, x) pi_j[d$strat8]^x * (1 - pi_j[d$strat8])^(1 - x)  # pas ifelse()
     : x peut être un scalaire
  l2 <- sum(log(py(ph2, ph2$hist)) + log(px(ph2, ph2$hist)))          # phase 2
  l1 <- sum(log(py(nonsel, 1) * px(nonsel, 1) + py(nonsel, 0) * px(nonsel, 0)))
  -(l1 + l2)
}
opt <- optim(c(coef(fit_wl), rep(-2, 8)), lvrais, method = "BFGS",
             hessian = TRUE, control = list(maxit = 500, reltol = 1e-12))
cbind(coef = opt$par[1:5], ET = sqrt(diag(solve(opt$hessian)))[1:5])
```

```
      coef    ET
[1,] -2.780 0.102
[2,]  0.782 0.134
[3,]  0.870 0.134
[4,]  1.296 0.153
[5,]  1.781 0.124
```

Les erreurs types du maximum de vraisemblance sont pratiquement celles de la cohorte complète pour l'ordonnée à l'origine et le stade, alors que la phase 2 ne compte que 28 % des sujets : l'information de phase 1 sur ces coefficients est entièrement récupérée. Pour l'histologie, l'erreur type (0,124) se situe entre celle de la cohorte (0,111) et celle de l'estimateur pondéré (0,159). Les estimations ponctuelles (1,78 pour le maximum de vraisemblance, 1,69 pour la vraisemblance pondérée) encadrent la valeur de la cohorte (1,75). Les probabilités estimées $\hat{\pi}_j = \hat{\mathbb{P}}(X = 1 \mid S = j)$ valent environ 0,02 à 0,04 dans les strates à histologie institutionnelle favorable et 0,68 à 0,87 dans les autres, ce qui reflète la sensibilité et la spécificité de l'histologie institutionnelle.

7.5 Vérification par simulation

On génère 500 cohortes de $N = 4088$ sujets en tirant le triplet (stade, histologie institutionnelle, histologie centrale) selon les fréquences de la cohorte NWTS et Y selon (a) le modèle à effets principaux ajusté sur la cohorte (modèle bien spécifié) ou (b) le modèle avec interaction stade × histologie (modèle d'analyse mal spécifié). Dans chaque cohorte, on tire un échantillon équilibré de la même façon que dans les données (tous les cas, toutes les strates 5 à 8, et 117, 89, 82 et 28 témoins dans les strates 1 à 4), puis on calcule l'estimateur pondéré avec les deux erreurs types et le maximum de vraisemblance. Dans les deux scénarios, la valeur visée par le modèle à effets principaux dans la population est $\beta_H = 1{,}748$. Le code reprend les blocs ci-dessus dans une fonction ; seule la génération d'une cohorte est montrée.

```
cells <- coh[, 1:3]                                  # 16 combinaisons (stade, instit,
    histol)
prob  <- (coh$case + coh$control) / sum(coh$case + coh$control)
gen   <- glm(cbind(case, control) ~ factor(stage) * factor(histol),   # scénario (b)
             family = binomial, data = coh)
d <- cells[sample(16, 4088, replace = TRUE, prob = prob), ]
d$y <- rbinom(4088, 1, predict(gen, newdata = d, type = "response"))
d$strat8 <- d$stage + 4 * (d$instit - 1)
# ... tirage équilibré, puis vraisemblance pondérée (deux ET) et ML comme plus haut
```

Résultats abrégés (500 répétitions par scénario ; `census` est l'estimation du modèle à effets principaux sur la cohorte simulée complète, `wl_s4` le coefficient du stade IV par vraisemblance pondérée) :

```
Scénario (a) : modèle bien spécifié
        census    wl se_naif se_2ph    ml wl_s4 se_naif_s4 se_2ph_s4
moyenne  1.747 1.765   0.155  0.150 1.750 1.319      0.224     0.159
et       0.112 0.146   0.007  0.009 0.121 0.166      0.008     0.008
Couverture IC 95 % WL, histologie : naif 0.96  deux phases 0.954
Couverture IC 95 % WL, stade IV : naif 0.994  deux phases 0.936

Scénario (b) : interaction stade x histologie omise
        census    wl se_naif se_2ph    ml wl_s4 se_naif_s4 se_2ph_s4
moyenne  1.747 1.753   0.153  0.148 1.759 1.321      0.223     0.156
et       0.116 0.159   0.007  0.008 0.127 0.163      0.008     0.008
Couverture IC 95 % WL, histologie : naif 0.938  deux phases 0.934
Couverture IC 95 % WL, stade IV : naif 0.99  deux phases 0.942
```

— Les deux estimateurs sont pratiquement sans biais pour $\beta_H = 1{,}748$. L'écart type du maximum de vraisemblance (0,121) est proche de celui de l'analyse de la cohorte complète (0,112) ; celui de l'estimateur pondéré est plus grand (0,146) : efficacité relative d'environ $(0{,}121/0{,}146)^2 \approx 0{,}69$.

— L'erreur type à deux phases de l'estimateur pondéré est correcte en moyenne (0,150 contre 0,146 pour l'histologie ; 0,159 contre 0,166 pour le stade IV). L'erreur type naïve est correcte pour l'histologie, mais nettement trop grande pour le stade IV (0,224 ; couverture de 99 %), comme le prévoit la comparaison de la section précédente.

— Dans le scénario (b), l'omission de l'interaction ne change pratiquement rien ici : la mauvaise spécification est modérée. L'argument de robustesse de l'estimateur pondéré (il vise toujours le paramètre de recensement) reste un argument théorique, qui devient important quand le modèle est plus gravement faux ou que les probabilités de sélection varient davantage.

7.6 La fonction `tps` du module `osDesign`

Les trois estimateurs de l'article sont implantés dans la fonction `tps` du module `osDesign`. *Ce module n'était pas disponible dans l'environnement où ces notes ont été compilées : le code ci-dessous n'a pas été exécuté.* Les arguments `nn0` et `nn1` sont les effectifs de phase 1 des témoins et des cas par strate, `group` l'identifiant de strate des lignes de phase 2, et `method` vaut `"WL"`, `"PL"` ou `"ML"`.

```r
library(osDesign)
# Effectifs de phase 1 par strate (strate = stade + 4 * (instit - 1))
coh$strat <- coh$stage + 4 * (coh$instit - 1)
nn0 <- as.vector(tapply(coh$control, coh$strat, sum))
nn1 <- as.vector(tapply(coh$case,    coh$strat, sum))
# Données de phase 2 groupées (échantillon équilibré)
b <- read.table(file.path(dossier, "nwtsb.dat"), header = TRUE)
b$strat <- b$stage + 4 * (b$instit - 1)
for (m in c("WL", "PL", "ML")) {
  fit <- tps(cbind(case, control) ~ factor(stage) + factor(histol), data = b,
             nn0 = nn0, nn1 = nn1, group = b$strat, method = m, cohort = TRUE)
  print(round(summary(fit)$coefficients, 3))
}
```

Pour le modèle à effets principaux, la méthode `"WL"` devrait reproduire les estimations de la vraisemblance pondérée ci-dessus. Avec `cohort = FALSE` (phase 1 cas-témoins), l'ordonnée à l'origine n'est estimable que si l'on fournit la cote marginale `alpha` de la population.

8 Niveau doctorat

Niveau doctorat – Vraisemblance des données observées (Wu, p. 101)

Soit $Z = (Y, S, X)$ les données complètes d'un sujet et R l'indicateur d'observation de X. La vraisemblance des données observées s'obtient en intégrant la densité des données complètes sur les composantes manquantes :

$$f(z_{\text{obs}}, r; \theta, \phi) = \int f(z_{\text{obs}}, z_{\text{mis}}; \theta)\, f(r \mid z_{\text{obs}}, z_{\text{mis}}; \phi)\, \mathrm{d}z_{\text{mis}}.$$

Ici $f(z;\theta) = f(y \mid x;\beta)\, f(x \mid s;F)\, f(s;\eta)$ (sous (2)) et $f(r \mid z) = \pi_{ys}^{r}(1-\pi_{ys})^{1-r}$ ne dépend pas de x (MAR). D'où

$$L = \prod_{k=1}^{N} f(s_k;\eta)\, \pi_{y_k s_k}^{r_k}(1-\pi_{y_k s_k})^{1-r_k} \Big\{ f(y_k \mid x_k;\beta) f(x_k \mid s_k;F) \Big\}^{r_k} \times \Big\{ \int f(y_k \mid x;\beta)\, \mathrm{d}F_{s_k}(x) \Big\}^{1-r_k}.$$

Les facteurs $f(s;\eta)$ et π_{ys} ne dépendent pas de (β,F) : pour l'inférence sur β on peut les omettre (ignorabilité), ce qui donne (6). Avec X discret à valeurs $\{x^{(1)}, \ldots, x^{(M)}\}$, l'intégrale est la somme $\sum_m f(y \mid x^{(m)};\beta) F_s(\{x^{(m)}\})$.

Niveau doctorat – Mécanisme de « coarsening » (Tsiatis, ch. 7 et 8)

Tsiatis formalise les données manquantes par une variable de *coarsening* C et des applications $G_r(Z)$ qui décrivent ce qui est observé quand $C = r$. Ici, deux niveaux : $C = \infty$ (données complètes, $G_\infty(Z) = Z$) et $C = 1$ ($G_1(Z) = (Y,S)$). Les probabilités de coarsening sont

$$\mathbb{P}(C = \infty \mid Z) = \pi_{YS} = \frac{n_{YS}}{N_{YS}}, \qquad \mathbb{P}(C = 1 \mid Z) = 1 - \pi_{YS},$$

et ne dépendent de Z qu'à travers $G_1(Z) = (Y,S)$: c'est la condition de *coarsening at random* (CAR), forme générale de MAR. (À strictement parler, le tirage sans remise à n_{ij} fixés rend les R_k dépendants ; la théorie i.i.d. s'applique au devis de Bernoulli équivalent, et la proposition 3.1 corrige la variance.)

Niveau doctorat – Estimateurs augmentés

Soit $\varphi(Z)$ une fonction d'influence (ou d'estimation) des données complètes, ici $u(\beta) = x\{y - p(x;\beta)\}$. Tsiatis (2006, ch. 7 à 10) montre que, sous CAR, toutes les fonctions d'influence des données observées sont de la forme

$$\frac{R\,\varphi(Z)}{\pi(Y,S)} + \Big\{1 - \frac{R}{\pi(Y,S)}\Big\} h(Y,S),$$

où h est arbitraire ; le choix optimal est $h^*(Y,S) = \mathrm{E}[\varphi(Z) \mid Y,S]$. Deux remarques éclairent l'article.

— Si h est estimée non paramétriquement par la moyenne de φ dans chaque cellule de phase 2, $\bar\varphi_c$, la somme sur la cohorte vaut $\sum_c \{(N_c/n_c) \sum_{k \in c, R_k = 1} \varphi_k + N_c \bar\varphi_c - (N_c/n_c) n_c \bar\varphi_c\} =$

$\sum_c (N_c/n_c) \sum_{k \in c, R_k=1} \varphi_k$: on retrouve exactement l'estimateur pondéré. Avec des strates discrètes et un φ fixé, l'estimateur pondéré est donc déjà « augmenté de façon optimale » dans le modèle non paramétrique pour la loi de Z. Les gains viennent d'information auxiliaire supplémentaire de phase 1 (calibration, Breslow et coll., 2009 ; Lumley, Shaw et Dai, 2011).

— Le gain du maximum de vraisemblance vient d'ailleurs : il exploite la *restriction* imposée par le modèle logistique (1) et l'hypothèse (2), qui relient la loi de X sachant (Y,S) chez les sélectionnés et les non-sélectionnés. Dans ce modèle semi-paramétrique plus petit, le choix de φ lui-même peut être optimisé, et l'estimateur efficace est le maximum de vraisemblance semi-paramétrique (Robins, Rotnitzky et Zhao, 1994 ; Breslow et Holubkov, 1997).

9 Préparation de l'étude de cas

Les questions sont dans le document de consignes de l'étude de cas 3. Repères de lecture : le modèle (1) de l'article et l'hypothèse de substitut correspondent à nos équations (1)–(2) ; la formule (3) (variance de l'estimateur pondéré) suppose les n_{ij} fixés et se démontre comme la proposition 3.1 ; la pseudo-vraisemblance se ramène au décalage (5). Les tableaux 2 et 3 comparent les estimateurs et les devis ; leur reproduction (avec `tps`) fait l'objet de l'étude de cas.

10 Synthèse

Synthèse après la séance – Breslow et Chatterjee (1999)

Objet d'intérêt	Les coefficients β (log des rapports de cotes) du modèle logistique de la rechute en fonction de l'histologie centrale et du stade, $\text{logit}\,\mathbb{P}(Y = 1 \mid X) = x^\top\beta$; secondairement, l'efficacité relative des devis (cas-témoins, équilibré) et des estimateurs.
Nuisance / biais	X (histologie centrale) manquant hors phase 2 ; sélection dépendant de Y et de S (biais d'une analyse naïve des cas complets) ; loi F_j de X sachant S (dimension infinie) ; erreur de classification de l'histologie institutionnelle.
Approche	Données manquantes par devis (MAR, probabilités connues $\pi_{ij} = n_{ij}/N_{ij}$) : pondération par l'inverse de π_{ij} ; conditionnement sur la sélection (décalage) ; vraisemblance complète avec F_j non paramétrique ; strates discrètes ; S substitut non différentiel.
Critère	(1) vraisemblance pondérée (Horvitz-Thompson) ; (2) pseudo-vraisemblance $L_1 \times L_2$ de Breslow-Cain ; (3) vraisemblance semi-paramétrique (« non paramétrique » pour F_j).
Calcul	(1) régression logistique pondérée ; (2) régression logistique avec décalage $\log\{n_{1j}N_{0j}/(n_{0j}N_{1j})\}$; (3) équations de score profilées résolues par Newton-Raphson modifié (`tps`, méthode `"ML"`) ; maximisation directe si X est discret.
Inférence	Normalité asymptotique ; variance sandwich dont la viande additionne une composante de phase 1 et une composante de phase 2 (les N_{ij} sont aléatoires, les n_{ij} fixés) ; variance corrigée de Breslow-Cain pour PL ; inverse de l'information pour ML ; comparaison à l'étalon de la cohorte complète.

Références

— Breslow, N. E. et Cain, K. C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, 75, 11–20.

— Breslow, N. E. et Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Applied Statistics*, 48(4), 457–468.

— Breslow, N. E. et Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *JRSS B*, 59, 447–461.

— Breslow, N. E., Lumley, T., Ballantyne, C. M., Chambless, L. E. et Kulich, M. (2009). Improved Horvitz-Thompson estimation of model parameters from two-phase stratified samples : applications in epidemiology. *Statistics in Biosciences*, 1, 32–49.

— Flanders, W. D. et Greenland, S. (1991). Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine*, 10, 739–747.

— Horvitz, D. G. et Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *JASA*, 47, 663–685.

— Lumley, T. (2010). *Complex Surveys : A Guide to Analysis Using R*. Wiley.

— Lumley, T., Shaw, P. A. et Dai, J. Y. (2011). Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review*, 79, 200–220.

— Prentice, R. L. et Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66, 403–411.

— Robins, J. M., Rotnitzky, A. et Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *JASA*, 89, 846–866.

— Scott, A. J. et Wild, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84, 57–71.

— Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer (ch. 7 à 10).

— Wu, L. (2010). *Mixed Effects Models for Complex Data*. Chapman & Hall/CRC (p. 101).

— Zhao, L. P. et Lipsitz, S. (1992). Designs and analysis of two-stage studies. *Statistics in Medicine*, 11, 769–782.