Neuchâtel, Octobre 2001

La technique de la double enquête pour la statistique du transport routier de marchandise

Paul-André Salamin

Service de méthodes statistiques
Dienst Statistische Methoden
Servizio metodi statistici

Pour tout renseignement, contacter

Paul-André Salamin

Office fédéral de la statistique
Service de méthodes statistiques

Espace de l'Europe 10, CH-2010 Neuchâtel
Tel. :(+41 32) 71 36098
Fax :(+41 32) 71 36093
Bestellnummer : 338-0006
© BFS/OFS

L'objectif de l'Enquête sur le Transport de Marchandises (ETM) de l'Office fédéral de la statistique est de décrire les prestations et la structure du transport de marchandises. Lors de l'ETM de 1993, on avait utilisé la technique de la double enquête pour effectuer une correction pour la non-réponse. Dans cette approche, une partie de l'échantillon reçoit un questionnaire simplifié, et l'autre partie de l'échantillon reçoit un questionnaire détaillé.

Les travaux décrits dans ce rapport ont été effectués dans la phase de planification de l'ETM de 1998. On veut déterminer, sur la base des données de l'ETM93, la proportion de questionnaires simplifiés pour l'ETM98. Cette étude a été effectuée en collaboration avec Louis Flury, de la Section des transports (VERK), en particulier en ce qui concerne les calculs de la section 5. Mes remerciements vont aussi à Monique Graf, du Service de méthodes statistiques (METH), pour sa relecture attentive de ce rapport.

Table des matières

1 Introduction **1**

2 Notations **1**

3 Estimateur d'un total **2**

4 Variance de l'estimateur d'un total **2**

5 Résultats **4**

A Tirage initial **7**

B Echantillon des questionnaires simplifiés **7**
B.1 Variance de $\sum_h \bar{Y}_h \hat{N}_h(R)$ 8
B.2 Estimation de κ_1 8

C Echantillon des questionnaires détaillés **9**
C.1 Variance de $\sum_h N_h \bar{y}_{R_h}$ 9
C.2 Estimation de κ_2 10

D Covariance **10**

1 Introduction

L'objectif de l'Enquête sur le Transport de Marchandises (ETM) est de décrire les prestations et la structure du transport de marchandises. Les principales variables relevées par l'ETM sont les tonnes transbordées (t), les tonnes-kilomètres effectuées (tkm), ce qui donne une mesure de l'utilité des transports, et les kilomètres parcourus par les véhicules (km). Les rapports de ces quantités renseignent sur la structure et l'efficacité des transports. Dans la plupart des analyses, les prestations ont encore besoin d'être différenciées selon des critères géographiques, économiques et techniques.

Pour l'enquête sur le transport de marchandises 1993 (ETM93), l'année a été divisée en 4 trimestres et, au sein de chaque trimestre, on a tiré des échantillons aléatoires simples de jours d'enquête, de tailles 5, 6, 6, et 5. Pour chacun des 22 jours d'enquête, le parc des véhicules est divisé en 7 strates.

Lors de l'ETM de 1993, on a utilisé la technique de la double enquête pour effectuer une correction pour la non-réponse. Dans chacune des strates, on a tiré deux échantillons aléatoires simples disjoints S_1 et S_2 de véhicules. Les unités de l'échantillon S_1 ont reçu un questionnaire simplifié, dans lequel on ne relève que le kilométrage journalier (km), et les unités de S_2 ont reçu un questionnaire détaillé. Lors de l'ETM93, il est apparu que le taux de réponse dans l'échantillon des questionnaires détaillés S_2 dépendait du nombre de kilomètres parcourus (km). Pour tenir compte de cet effet, les échantillons ont été post-stratifiés en 4 classes selon la variable km : $km = 0$, $0 < km < 100$, $100 \leq km < 250$, $km \geq 250$. Comme la post-stratification est basée sur une variable d'enquête, les tailles des post-strates sont inconnues. Elles ont été estimées à partir de l'échantillon des questionnaires simplifés S_1, pour lequel le taux de réponse ne dépendait pas de km.

Ce document décrit une étude dont l'objet était de déterminer, sur la base des données de l'ETM93, la proportion de questionnaires simplifiés pour l'ETM de 1998 (ETM98). Dans la section 3 on décrit l'estimateur du total qui est utilisé dans l'ETM. Dans la section 4, on calcule la variance théorique de cet estimateur. Enfin, dans la section 5, on détermine sur la base des données de l'ETM93 la proportion de questionnaires simplifés pour l'ETM98.

2 Notations

Soit U une strate de taille N dans laquelle on a tiré deux échantillons aléatoires simples disjoints S_1 et S_2 de tailles n_1 et n_2. Les échantillons nets sont $R_1 \subseteq S_1$ et $R_2 \subseteq S_2$, de tailles m_1 et m_2. Les unités de S_1 ont reçu un *questionnaire simplifié* et les unités de S_2 ont reçu un *questionnaire détaillé*. Pour une variable y, on note $Y = \sum_U y_k$ le total de y dans U, $\bar{Y} = \sum_U y_k/N$ la moyenne de y dans U et

$$S_{yU}^2 = \frac{1}{N-1} \sum_U (y_k - \bar{Y})^2$$

la variance de y dans U. Pour une post-stratification $U = \bigcup_{h=1}^{H} U_h$ de la strate U en H post-strates, on note N_h la taille de la post-strate h, $W_h = N_h/N$ la taille relative de la post-strate h par rapport à la taille de la strate, $\bar{Y}_h$ la moyenne de y dans la post-strate h et $S_{yU_h}^2$ la variance de y dans la post-strate h. On note $R_{1h} = R_1 \cap U_h$ la partie de l'échantillon R_1 qui tombe dans la post-strate h, m_{1h} la taille de R_{1h}, $\bar{y}_{R_1}$ la moyenne de y

dans R_1, $\bar{y}_{R_{1h}}$ la moyenne de y dans R_{1h} et $S^2_{yR_{1h}}$ la variance de y dans R_{1h}. Les définitions pour R_2 sont analogues. Finalement, on note θ le taux de réponse pour l'échantillon des unités ayant reçu un questionnaire simplifié et θ_h le taux de réponse dans la post-strate h pour l'échantillon des unités ayant reçu un questionnaire détaillé.

3 Estimateur d'un total

On considère l'estimation d'un total $Y = \sum_U y_k$, où la variable y_k est disponible pour les unités $k \in R_2$. Les pondérations initiales dans R_1 et R_2 sont données par l'inverse des probabilités d'inclusion multiplié par l'inverse d'un estimateur des taux de réponse dans S_1 et S_2 :

$$w_{1k} = \frac{N}{n_1}\frac{n_1}{m_1} = \frac{N}{m_1} \text{ et } w_{2k} = \frac{N}{n_2}\frac{n_2}{m_2} = \frac{N}{m_2}.$$

Soit $U = \bigcup_h U_h$ une post-stratification de la strate U. Si les tailles N_h des post-strates sont connues, l'estimateur de post-stratification de Y est donné par

$$\hat{Y}_{post} = \sum_h N_h \frac{\sum_{R_{2h}} w_{2k} y_k}{\sum_{R_{2h}} w_{2k}} = \sum_h N_h \bar{y}_{R_{2h}}.$$

Pour l'ETM93, les tailles des post-strates étaient inconnues et elles ont été estimées à partir de l'échantillon R_1 des unités ayant répondu au questionnaire simplifié ([Hulliger 1993], formule (7)) :

$$\hat{N}_h(R_1) = \sum_{R_{1h}} w_{1k} = \frac{N}{m_1} m_{1h}.$$

On obtient ainsi l'estimateur d'un total Y ([Hulliger 1993], section 4.1)

$$\hat{Y} = \sum_h \hat{N}_h(R_1)\bar{y}_{R_{2h}} = \sum_h (N\frac{m_{1h}}{m_1})(\frac{1}{m_{2h}}\sum_{R_{2h}} y_k). \tag{1}$$

4 Variance de l'estimateur d'un total

Un développement de Taylor de premier ordre de (1) dans un voisinage de $(N_h, \bar{Y}_h)$ donne

$$\hat{Y} \approx \sum_h \left\{ N_h\bar{Y}_h + \bar{Y}_h(\hat{N}_h(R_1) - N_h) + N_h\left(\bar{y}_{R_{2h}} - \bar{Y}_h\right)\right\}.$$

On montre dans la section D de l'appendice que la covariance entre $\sum_h \bar{Y}_h\hat{N}_h(R_1)$ et $\sum_h N_h\bar{y}_{R_{2h}}$ est égale à 0. On peut alors calculer approximativement la variance de $\hat{Y}$ comme

$$\text{var}(\hat{Y}) \approx \text{var}(\sum_h \bar{Y}_h\hat{N}_h(R_1)) + \text{var}(\sum_h N_h\bar{y}_{R_{2h}}).$$

On montre dans les sections B et C de l'appendice que

$$\begin{aligned}\text{var}(\sum_h \bar{Y}_h \hat{N}_h(R_1)) &\approx N^2(1-\frac{n_1\theta}{N})\frac{1}{n_1\theta}\sum_h W_h(\bar{Y}_h-\bar{Y})^2,\\ \text{var}(\sum_h N_h \bar{y}_{R_{2h}}) &\approx N^2\sum_h W_h(1-\frac{n_2\theta_h}{N})\frac{1}{n_2\theta_h}S^2_{yU_h}.\end{aligned}$$

La variance de l'estimateur (1) est donc donnée approximativement par

$$\text{var}(\hat{Y}) \approx N^2\sum_h W_h\left\{(1-\frac{n_1\theta}{N})\frac{1}{n_1\theta}(\bar{Y}_h-\bar{Y})^2+(1-\frac{n_2\theta_h}{N})\frac{1}{n_2\theta_h}S^2_{yU_h}\right\}$$

et, si on néglige les corrections de population finie,

$$\text{var}(\hat{Y}) \approx N^2\sum_h W_h\left\{\frac{1}{n_1\theta}(\bar{Y}_h-\bar{Y})^2+\frac{1}{n_2\theta_h}S^2_{yU_h}\right\}. \qquad (2)$$

Pour mieux comprendre la structure de (2), il est instructif de considérer le cas d'une post-stratification sans non-réponse : $\theta=\theta_1=\ldots=\theta_H=1$. Alors

$$\text{var}(\hat{Y}) \approx N^2\left\{\frac{1}{n_1}\sum_h W_h(\bar{Y}_h-\bar{Y})^2+\frac{1}{n_2}\sum_h W_h S^2_{yU_h}\right\}.$$

La variance de la variable y dans la strate U peut s'exprimer comme $S^2_{yU} \approx SS_b + SS_w$ où

$$SS_b = \sum_h W_h S^2_{yU_h}$$

est la dispersion de y à l'intérieur des post-strates et

$$SS_b = \sum_h W_h(\bar{Y}_h-\bar{Y})^2$$

est la dispersion de y entre les post-strates. Avec ces définitions, on a

$$\text{var}(\hat{Y}) \approx N^2\left\{\frac{1}{n_1}SS_b+\frac{1}{n_2}SS_w\right\}.$$

Maintenant, N^2SS_w/n_2 est essentiellement la variance de $\hat{Y}_{post}$ et on a donc

$$\text{var}(\hat{Y}) \approx \text{var}(\hat{Y}_{post})\left\{1+\frac{n_2}{n_1}\frac{SS_b}{SS_w}\right\}.$$

On peut interpréter le facteur $(n_2/n_1)(SS_b/SS_w)$ comme la perte de précision due à l'estimation de la taille des post-strates. Pour une post-stratification classique, plus SS_b est grand par rapport à SS_w, plus le gain de précision est important. Pour l'estimateur (1), la situation est inversée. On peut tout au plus atténuer la perte de précision en prenant $n_1 > n_2$, c'est-à-dire estimant la taille des post-strates le plus précisément possible.

Si on revient au cas général, on peut mettre la variance (2) sous la forme

$$\text{var}(\hat{Y}) \approx N^2 \frac{1}{n} f(\alpha, \kappa_1, \kappa_2)$$

où $0 < \alpha < 1$ est la proportion de questionnaires simplifiés, c'est-à-dire $n_1 = \alpha n$ et $n_2 = (1-\alpha)n$ avec $n = n_1 + n_2$, et

$$f(\alpha, \kappa_1, \kappa_2) = \frac{\kappa_1}{\alpha} + \frac{\kappa_2}{1-\alpha}, \tag{3}$$

avec

$$\kappa_1 = \frac{1}{\theta} \sum_h W_h (\bar{Y}_h - \bar{Y})^2 \text{ et } \kappa_2 = \sum_h \frac{1}{\theta_h} W_h S^2_{yU_h}.$$

5 Résultats

On décrit dans cette section l'estimation de l'approximation (2) de $\text{var}(\hat{Y})$ sur la base des données de l'ETM93, où $\hat{Y}$ est l'estimateur (1) du nombre de kilomètres parcourus (variable km). Pour chacun des 22 jours d'enquête, le parc de véhicules a été divisé en 7 strates. On a combiné les strates 1, 2 et 3 correspondant aux véhicules légers, et les strates 4, 5, 6 et 7 correspondant aux véhicules lourds. On a ainsi obtenu 44 strates dans lesquelles les paramètres κ_1 et κ_2 ont été estimés. Dans chacune des 44 strates, les échantillons des questionnaires simplifiés et des questionnaires détaillés ont été post-stratifiés en 4 classes selon la variable km ($km = 0$, $0 < km < 100$, $100 \leq km < 250$, $km \geq 250$). Dans chacune des 44 strates, les paramètre κ_1 et κ_2 sont estimés par

$$\hat{\kappa}_1 = \frac{1}{\hat{\theta}} \sum_h \hat{W}_h (\bar{y}^2_{R_{1h}} - \bar{y}^2_{R_1})^2 \text{ et } \hat{\kappa}_2 = \sum_h \hat{W}_h \frac{1}{\hat{\theta}_h} S^2_{yR_{2h}},$$

où $\hat{\theta} = 0.837$ est un estimateur du taux de réponse global dans l'échantillon des unités ayant reçu un questionnaire simplifié, $\hat{W}_h = m_{1h}/m_1$ et $\hat{\theta}_h = m_{2h}/\hat{n}_{2h}$ avec $\hat{n}_{2h} = n_2 (m_{1h}/m_1) = n_2 \hat{W}_h$. Le choix de ces estimateurs de κ_1 et de κ_2 est justifié dans les sections B et C de l'appendice. Une représentation des 44 estimations de $\hat{\kappa}_1$ et de $\hat{\kappa}_2$, voir la figure 1, montre que $\hat{\kappa}_2$ est environ 4 fois plus petit que $\hat{\kappa}_1$. Une régression par l'origine de $\hat{\kappa}_2$ sur $\hat{\kappa}_1$ donne une estimation de la pente $\hat{\beta} = 0.252(0.008)$.

Comme $\hat{\kappa}_2 \approx \hat{\beta}\hat{\kappa}_1$, on peut mettre (3) sous la forme

$$f(\alpha, \kappa_1, \kappa_2) \approx \hat{\kappa}_1 \left(\frac{1}{\alpha} + \frac{\hat{\beta}}{1-\alpha} \right).$$

La proportion de questionnaires simplifiés qui minimise f est donnée par $\alpha_{opt} = 0.666$. Dans ce cas, on observe une perte de précision de 73% par rapport à une post-stratification classique. Une proportion de questionnaires simplifés de 67% a été jugée pour des raisons pratiques trop élevée. La proportion choisie se situe entre 30% et 40%. Pour $\alpha \in (0.3, 0.4)$, on a des pertes de précision par rapport à la proportion optimale $\alpha_{opt} = 0.67$ entre 14% et 28%.

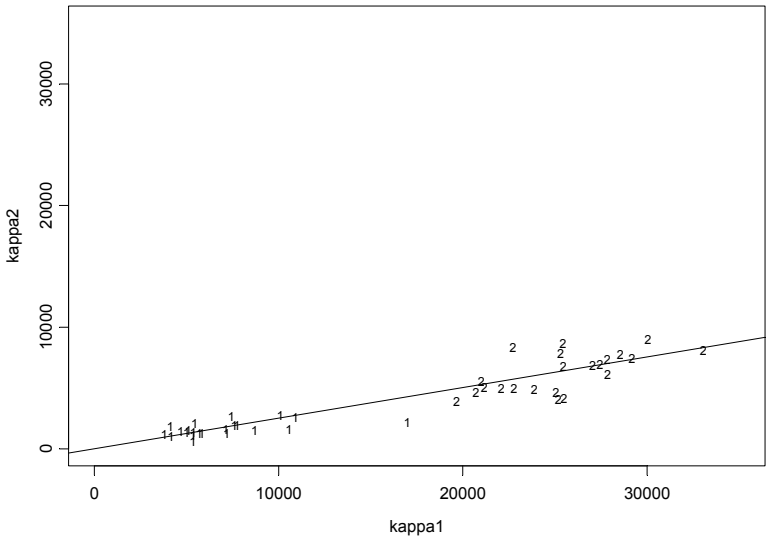


FIG. 1: Estimation de κ_1 et de κ_2 pour 44 strates de la STM93. 1 : véhicules légers, 2 : véhicules lourds.

Appendice

A Tirage initial

La méthode de tirage de l'échantillon initial est

$$U \xrightarrow{EAS} \begin{cases} S_1 \\ U \setminus S_1 \end{cases} \xrightarrow{EAS} S_2$$

où U est une population de taille N, S_1 est un échantillon aléatoire simple de taille n_1 tiré dans U – les questionnaires *simplifiés* – et S_2 est un échantillon aléatoire simple de taille n_2 tiré dans $U \setminus S_1$ – les questionnaires *détaillés* –. Par définition de la méthode de tirage,

$$P(S_2 \mid S_1) = \begin{cases} 1/\binom{N-n_1}{n_2} & \text{si } S_2 \subseteq U \setminus S_1 \\ 0 & \text{autrement} \end{cases}$$

ce qui permet de montrer que S_1 et S_2 sont des échantillons aléatoires simples de U.

B Echantillon des questionnaires simplifiés

Pour l'échantillon des questionnaires simplifiés, on a un sondage en deux phases effectué selon le modèle

$$U \xrightarrow{EAS} S \xrightarrow{\theta} R \tag{4}$$

Lors de la première phase, on tire un échantillon aléatoire simple S de taille n dans une population U de taille N. L'échantillon de deuxième phase R est un échantillon de Bernoulli de paramètre θ tiré dans S. On note m la taille de l'échantillon R. Par définition du processus d'échantillonnage,

$$P(R \mid S) = \begin{cases} \theta^m (1-\theta)^{n-m} & \text{si } R \subseteq S \\ 0 & \text{autrement} \end{cases}$$

On considère aussi une post-stratification de la population : $U = \bigcup_h U_h$, $N_h = |U_h|$. On note m_h la taille de l'échantillon des répondants dans la post-strate h. Sous le modèle (4), m est de loi binomiale $B(n, \theta)$ et, conditionnellement à m, les tailles $m_1, \ldots, m_H$ sont de loi hypergéométrique multiple $\mathcal{H}(N, N_1, \ldots, N_H, m)$:

$$P(m_1, \ldots, m_H \mid m) = \frac{\prod_h \binom{N_h}{m_h}}{\binom{N}{m}}. \tag{5}$$

Conditionnellement à m, les moments des variables indicatrices sont les mêmes que pour un sondage aléatoire simple de taille m. Conditionnellement à $(m_1, \ldots, m_H)$, les moments des variables indicatrices sont les mêmes que pour un sondage stratifié de tailles $(m_1, \ldots, m_H)$.

B.1 Variance de $\sum_h \bar{Y}_h \hat{N}_h(R)$

On montre que

$$\operatorname{var}(\sum_h \bar{Y}_h \hat{N}_h(R)) = N^2 \operatorname{var}(\sum_h \bar{Y}_h \frac{m_h}{m}) \approx N^2(1 - \frac{n\theta}{N})\frac{1}{n\theta}\sum_h W_h(\bar{Y}_h - \bar{Y})^2.$$

En effet, par (5),

$$\operatorname{var}(\sum_h \bar{Y}_h \frac{m_h}{m} \mid m) = \frac{1}{m}\frac{N-m}{N-1}\left\{\sum_h \bar{Y}_h^2 W_h(1-W_h) - \sum_{h\neq i}\bar{Y}_h\bar{Y}_i W_h W_i\right\}.$$

Comme l'espérence conditionnelle de $(m_h/m)\bar{Y}_h$ ne dépend pas de m et puisque

$$\sum_h \bar{Y}_h^2 W_h(1-W_h) - \sum_{h\neq i}\bar{Y}_h\bar{Y}_i W_h W_i = \sum_h W_h(\bar{Y}_h - \bar{Y})^2,$$

on peut écrire

$$N^2 \operatorname{var}(\sum_h \bar{Y}_h \frac{m_h}{m}) = N^2 \frac{1}{N-1}\mathrm{E}\,(\frac{N}{m} - 1)\sum_h W_h(\bar{Y}_h - \bar{Y})^2.$$

Si on fait l'approximation[1] $\mathrm{E}\,(\frac{1}{m}) \approx \frac{1}{\mathrm{E}\,m}$, alors

$$\frac{1}{N-1}\mathrm{E}\,(\frac{N}{m} - 1) \approx \frac{N}{N-1}\frac{1}{\mathrm{E}\,m}(1 - \frac{\mathrm{E}\,m}{N}) \approx \frac{1}{\mathrm{E}\,m}(1 - \frac{\mathrm{E}\,m}{N})$$

et donc

$$N^2 \operatorname{var}(\sum_h \bar{Y}_h \frac{m_h}{m}) \approx N^2 \frac{1}{\mathrm{E}\,m}(1 - \frac{\mathrm{E}\,m}{N})\sum_h W_h(\bar{Y}_h - \bar{Y})^2.$$

Comme $m \sim B(n, \theta)$, on obtient finalement le résultat cherché.

B.2 Estimation de κ_1

Un estimateur de κ_1

$$\kappa_1 = \sum_h W_h \frac{1}{\theta}(\bar{Y}_h - \bar{Y})^2$$

est donné par

$$\hat{\kappa}_1 = \frac{1}{\hat{\theta}}\sum_h \hat{W}_h(\bar{y}_{R_h}^2 - \bar{y}_R^2)^2,$$

où $\hat{\theta}$ est un estimateur du θ et $\hat{W}_h = m_h/m$. Cet estimateur est sans biais en première approximation. En effet, par (5),

$$\mathrm{E}\,(\hat{W}_h \mid m) = \frac{1}{m}\mathrm{E}\,(m_h \mid m) = \frac{N_h}{N} = W_h.$$

Par les propriétés des variables indicatrices, conditionnellement à $(m_1, \ldots, m_H)$ ou à m, on obtient $\mathrm{E}\,(\bar{y}_{R_h} \mid m_1, \ldots, m_H) = \bar{Y}_h$ et $\mathrm{E}\,(\bar{y}_R \mid m) = \bar{Y}$.

[1] Pour être tout-à-fait correct, on devrait prendre l'espérance par rapport à la distribution de m conditionnellement à $m \geq 1$. Pour les tailles d'échantillons de la STM, la probabilité que $m = 0$ est négligeable.

C Echantillon des questionnaires détaillés

Pour l'échantillon des questionnaires détaillés, on a un sondage en deux phases effectué selon le modèle

$$U \quad \stackrel{EAS}{\longrightarrow} \quad S,\, S_h \quad \stackrel{\theta_h}{\longrightarrow} \quad R_h \tag{6}$$

Lors de la première phase, on tire un un échantillon aléatoire simple S de taille n dans une population U de taille N. La population est post-stratifiée : $U = \bigcup_h U_h$, $N_h = |U_h|$. On note $S_h = S \cap U_h$ la partie de l'échantillon qui tombe dans la post-strate U_h et n_h la taille de S_h. L'échantillon de deuxième phase est $R = \bigcup_h R_h$, où R_h est un échantillon de Bernoulli de paramètre θ_h tiré dans S_h. On note m_h la taille de l'échantillon R_h et $m = \sum_h m_h$ la taille de R. Par définition du processus d'échantillonnage,

$$P(R \mid S) = \begin{cases} \prod_h \theta_h^{m_h}(1-\theta_h)^{n_h - m_h} & \text{si } R \subseteq S \\ 0 & \text{autrement} \end{cases}$$

Sous le modèle (6), les tailles $\mathbf{n} = (n_1, \ldots, n_H)$ sont de loi hypergéométrique multiple $\mathcal{H}(N, N_1, \ldots, N_H, n)$ et, conditionellement à $\mathbf{n}$, les tailles $\mathbf{m} = (m_1, \ldots, m_H)$ sont indépendantes de loi binomiale $B(n_h, \theta_h)$. Conditionnellement à $\mathbf{n}$ et $\mathbf{m}$, les moments des variables indicatrices sont les mêmes que pour un sondage stratifié de tailles $\mathbf{m}$.

C.1 Variance de $\sum_h N_h \bar{y}_{R_h}$

On montre que

$$\operatorname{var}\Big(\sum_h N_h \bar{y}_{R_h}\Big) = \operatorname{var}\Big(\sum_h \frac{N_h}{m_h} \sum_{R_h} y_k\Big) \approx N^2 \sum_h W_h \Big(1 - \frac{n\theta_h}{N}\Big) \frac{1}{n\theta_h} S^2_{yU_h}.$$

En effet, conditionnellement à $\mathbf{n}$ et $\mathbf{m}$, les probabilités d'inclusion sont les mêmes que pour un sondage stratifié. Alors

$$\operatorname{var}\Big(\sum_h \frac{N_h}{m_h} \sum_{R_h} y_k \mid \mathbf{n}, \mathbf{m}\Big) = \sum_h N_h^2 \Big(1 - \frac{m_h}{N_h}\Big) \frac{1}{m_h} S^2_{yU_h}.$$

Comme l'espérance conditionnelle de $\bar{y}_{R_h}$ ne dépend pas de $\mathbf{n}$ et $\mathbf{m}$, on a

$$\operatorname{var}\Big(\sum_h \frac{N_h}{m_h} \sum_{R_h} y_k\Big) = \sum_h N_h^2 \left\{ \mathrm{E}\,\Big(\frac{1}{m_h}\Big) - \frac{1}{N_h} \right\} S^2_{yU_h}.$$

Si on fait l'approximation[2] $\mathrm{E}\,(\frac{1}{m_h}) \approx \frac{1}{\mathrm{E}\, m_h}$ on obtient

$$\operatorname{var}\Big(\sum_h \frac{N_h}{m_h} \sum_{R_h} y_k\Big) \approx \sum_h N_h^2 \Big(1 - \frac{\mathrm{E}\, m_h}{N_h}\Big) \frac{1}{\mathrm{E}\, m_h} S^2_{yU_h}.$$

Maintenant, $m_h \mid n_h \sim B(n_h, \theta_h)$ et n_h est de loi hypergéométrique. Il suit

$$\mathrm{E}\,(m_h) = \mathrm{E}\,(\mathrm{E}\,(m_h \mid n_h)) = \theta_h \mathrm{E}\,(n_h) = \theta_h n W_h$$

ce qui donne le résultat cherché.

[2]Pour être tout-à-fait correct, on devrait prendre l'espérance par rapport à la distribution de m_h conditionnellement à $m_h \geq 1$. Pour les tailles d'échantillons de la STM, la probabilité que $m_h = 0$ est négligeable.

C.2 Estimation de κ_2

Un estimateur de κ_2

$$\kappa_2 = \sum_h W_h \frac{1}{\theta_h} S^2_{yU_h}$$

est donné par

$$\hat{\kappa}_2 = \sum_h \hat{W}_h \frac{1}{\hat{\theta}_h} S^2_{yR_{2h}},$$

où $\hat{W}_h = m_{1h}/m$ et $\hat{\theta}_h = m_{2h}/\hat{n}_{2h}$ avec $\hat{n}_{2h} = n_2 \left(m_{1h}/m_1\right) = n_2 \hat{W}_h$. Cet estimateur est sans biais en première approximation. En effet,

$$\mathrm{E}\left(\hat{\theta}_h\right) \approx \frac{\mathrm{E}\left(m_{2h}\right)}{\mathrm{E}\left(\hat{n}_{2h}\right)} = \frac{\theta_h n_2 W_h}{n_2 W_h} = \theta_h.$$

Par les propriétés des variables indicatrices conditionnellement à $\mathbf{n}$ et $\mathbf{m}$, on obtient $\mathrm{E}\left(S^2_{yR_{2h}}\right) = S^2_{yU_h}$.

D Covariance

On montre que

$$\mathrm{cov}(\sum_h \bar{Y}_h \hat{N}_h(R_1), \sum_h N_h \bar{y}_{R_{2h}}) = 0.$$

On a ici un modèle de tirage en trois phases :

$$U \quad \stackrel{EAS}{\longrightarrow} \quad \left\{ \begin{array}{lllll} S_1 & \longrightarrow & S_1 & \stackrel{\theta_1}{\longrightarrow} & R_1 \\ U \setminus S_1 & \stackrel{EAS}{\longrightarrow} & S_2, S_{2h} & \stackrel{\theta_{2h}}{\longrightarrow} & R_{2h} \end{array} \right.$$

Comme l'incorporation du modèle de non-réponse ne présente pas de difficultés majeures, et afin d'alléger un peu les calculs, on ne tient pas compte dans ce qui suit de la troisième phase d'échantillonnage. La difficulté est de devoir traiter simultanément les échantillons des phases 1 et 2. Afin de simplifier la présentation d'un résultat intermédiaire important, on ignore pour commencer la post-stratification de l'échantillon de deuxième phase. On a donc le modèle de la section A, c'est-à-dire

$$U \quad \stackrel{EAS}{\longrightarrow} \quad \left\{ \begin{array}{ll} S_1 & \\ U \setminus S_1 & \stackrel{EAS}{\longrightarrow} \quad S_2 \end{array} \right.$$

On peut prendre comme espace des réalisations $\Omega = \mathcal{P}_{n_1}(U) \times \mathcal{P}_{n_2}(U)$, où $\mathcal{P}_n(U)$ dénote l'ensemble des échantillons de U de taille n, avec la probabilité

$$P(S_1, S_2) = \left\{ \begin{array}{ll} 1/\binom{N}{n_1}\binom{N-n_1}{n_2} & \text{si } S_1 \cap S_2 = \emptyset \\ 0 & \text{autrement} \end{array} \right.$$

On définit pour chaque unité k de la population U une variable indicatrice I_k sur $\mathcal{P}(U)$, l'ensemble des parties de U, par

$$I_k : \mathcal{P}(U) \longrightarrow \{0,1\}, \quad S \longmapsto I_k(S) = \begin{cases} 1 & \text{si } k \in S \\ 0 & \text{si } k \notin S \end{cases}$$

Ceci permet de définire pour chaque $k \in U$ une variable indicatrice multivariée $\mathbf{I}_k$ sur Ω par

$$\mathbf{I}_k : \Omega \longrightarrow \{0,1\}^2, \quad (S_1, S_2) \longmapsto \begin{bmatrix} I_{1k}(S_1) \\ I_{2k}(S_2) \end{bmatrix} := \begin{bmatrix} I_k(S_1) \\ I_k(S_2) \end{bmatrix}$$

On définit aussi les vecteurs $\mathbf{I}_1 = (I_{11}, \ldots, I_{1k}, \ldots, I_{1N})^T$ et $\mathbf{I}_2 = (I_{21}, \ldots, I_{2k}, \ldots, I_{2N})^T$. On montre que

$$\mathrm{cov}(\mathbf{I}_1, \mathbf{I}_2) = -\frac{N}{N-1} f_1 f_2 \left(\mathrm{Id}_N - 1_N (1_N^T 1_N)^{-1} 1_N^T\right)$$

où $f_1 = n_1/N$ et $f_2 = n_2/N$ sont les taux de sondages pour les échantillons S_1 et S_2, Id_N est la matrice identité de dimension $N \times N$ et $1_N = (1, \ldots, 1)^T$ est un vecteur de dimension $N \times 1$. La matrice $C = \mathrm{Id}_N - 1_N (1_N^T 1_N)^{-1} 1_N^T$ est la projection orthogonale sur $1_N^\perp$. On a $C^2 = C$, $C^T = C$ et $C 1_N = 0$.

Pour $\mathrm{cov}(I_{1k}, I_{2k})$ on doit calculer $\mathrm{E}\,(I_{1k} I_{2k}) = P(\{(S_1, S_2); S_1 \ni k, S_2 \ni k\})$. Comme $P(S_1, S_2) = 0$ si $S_1 \cap S_2 \neq \emptyset$, on a $\mathrm{E}\,(I_{1k} I_{2k}) = 0$ et donc $\mathrm{cov}(I_{1k}, I_{2k}) = -f_1 f_2$. Pour $\mathrm{cov}(I_{1k}, I_{2l})$, $k \neq l$, on doit calculer $\mathrm{E}\,(I_{1k} I_{2l}) = P(\{(S_1, S_2); S_1 \ni k, S_2 \ni l\})$. P est non nulle sur l'ensemble $\{S_1 \ni k, S_2 \ni l, S_1 \cap S_2 = \emptyset\}$ que l'on peut écrire comme

$$[\{k\} \cup \mathcal{P}_{n_1 - 1}(U \setminus \{k, l\})] \times [\{l\} \cup \mathcal{P}_{n_2 - 1}(U \setminus (S_1 \cup \{l\}))],$$

ce qui permet immédiatement de trouver sa taille :

$$|\{S_1 \ni k, S_2 \ni l, S_1 \cap S_2 = \emptyset\}| = \binom{N-2}{n_1 - 1} \binom{N - n_1 - 1}{n_2 - 1}.$$

Nous avons alors

$$\mathrm{E}\,(I_{1k} I_{2l}) = \frac{\binom{N-2}{n_1-1}\binom{N-n_1-1}{n_2-1}}{\binom{N}{n_1}\binom{N-n_1}{n_2}} = \frac{N}{N-1} f_1 f_2$$

d'où $\mathrm{cov}(I_{1k}, I_{2l}) = \frac{1}{N-1} f_1 f_2$. Alors, pour tout $k, l \in U$,

$$\mathrm{cov}(I_{1k}, I_{2l}) = -\frac{N\delta_{kl} - 1}{N-1} f_1 f_2 = -\frac{N}{N-1}\left(\delta_{kl} - \frac{1}{N}\right) f_1 f_2.$$

Finalement, il faut remarquer que $\delta_{kl} - 1/N$ correspond à l'élément kl de la matrice de projection

$$C = \mathrm{Id}_N - \frac{1}{N} 1_N 1_N^T = \mathrm{Id}_N - 1_N (1_N^T 1_N)^{-1} 1_N^T.$$

On considère maintenant une post-stratification sur S_2 :

$$U \quad \stackrel{EAS}{\longrightarrow} \quad \left\{ \begin{array}{ll} S_1 & \\ U \setminus S_1 & \stackrel{EAS}{\longrightarrow} \quad S_2, S_{2h} \end{array} \right.$$

L'estimateur (1) prend la forme

$$\hat{Y} = \sum_h \hat{N}_h (\frac{1}{n_{2h}} \sum_{S_{2h}} y_k) = \sum_h \hat{N}_h \bar{y}_{S_{2h}}$$

où $\hat{N}_h = (N/n_1) n_{1h}$. On doit calculer

$$\operatorname{cov}(\sum_h \bar{Y}_h \hat{N}_h, \sum_h N_h \bar{y}_{S_{2h}}) = \sum_{h,i} \bar{Y}_h N_i \operatorname{cov}(\hat{N}_h, \bar{y}_{S_{2i}})$$

Soient $\mathbf{I}_{h1}$ et $\mathbf{I}_{h2}$, $h = 1, \ldots, H$, les vecteurs de variables indicatrices dans les post-strates. Ainsi, $\mathbf{I}_{h1}$ est un vecteur de dimension $N_h \times 1$ dont les composantes sont I_{1k} pour $k \in U_h$. La définition de $\mathbf{I}_{h2}$ est similaire. De même, on définit les vecteurs y_h, $h = 1, \ldots, H$, de la variable dans les post-strates. On a alors

$$\operatorname{cov}(\hat{N}_h, \bar{y}_{S_{2i}}) = \operatorname{cov}(\frac{N}{n_1} 1_{N_h}^T \mathbf{I}_{h1}, \frac{1}{n_{2i}} y_i^T \mathbf{I}_{i2})$$

et

$$\operatorname{cov}(\sum_h \bar{Y}_h \hat{N}_h, \sum_h N_h \bar{y}_{S_{2h}} \mid \mathbf{n}_2) = \sum_{h,i} \bar{Y}_h N_i \frac{N}{n_1 n_{2i}} 1_{N_h}^T \operatorname{cov}(\mathbf{I}_{h1}, \mathbf{I}_{i2} \mid \mathbf{n}_2) y_i$$

où $\mathbf{n}_2 = (n_{21}, \ldots, n_{2H})$. Pour calculer $\operatorname{cov}(\mathbf{I}_{h1}, \mathbf{I}_{i2} \mid \mathbf{n}_2)$, le plus simple est de considérer le modèle où l'on post-stratifie les échantillons S_1 et S_2 :

$$U \quad \stackrel{EAS}{\longrightarrow} \quad \left\{ \begin{array}{lll} S_1 & \longrightarrow & S_1, S_{1h} \\ U \setminus S_1 & \stackrel{EAS}{\longrightarrow} & S_2, S_{2h} \end{array} \right.$$

On montre par un calcul analogue à celui que l'on peut faire en l'absence de post-stratification que

$$\operatorname{cov}(I_{h1}, I_{i2} \mid \mathbf{n}_1, \mathbf{n}_2) = \left\{ \begin{array}{ll} -\dfrac{N_h}{N_h - 1} \dfrac{n_{1h}}{N_h} \dfrac{n_{2h}}{N_h} C_h & \text{si } h = i \\ 0 & \text{si } h \neq i \end{array} \right.$$

où $\mathbf{n}_1 = (n_{11}, \ldots, n_{1H})$ et C_h est la matrice de projection $C_h = \mathrm{Id}_{N_h} - 1_{N_h} (1_{N_h}^T 1_{N_h})^{-1} 1_{N_h}^T$. On peut alors calculer $\operatorname{cov}(\mathbf{I}_{h1}, \mathbf{I}_{i2} \mid \mathbf{n}_2)$ par

$$\operatorname{cov}(I_{h1}, I_{i2} \mid \mathbf{n}_2) = \mathrm{E}\left(\operatorname{cov}(I_{h1}, I_{i2} \mid \mathbf{n}_1, \mathbf{n}_2) \mid \mathbf{n}_2\right) + \operatorname{cov}(\mathrm{E}\left(I_{h1} \mid \mathbf{n}_1, \mathbf{n}_2\right), \mathrm{E}\left(I_{i2} \mid \mathbf{n}_1, \mathbf{n}_2\right) \mid \mathbf{n}_2).$$

Mais puisque

$$\operatorname{cov}(\mathrm{E}\left(I_{h1} \mid \mathbf{n}_1, \mathbf{n}_2\right), \mathrm{E}\left(I_{i2} \mid \mathbf{n}_1, \mathbf{n}_2\right) \mid \mathbf{n}_2) = \operatorname{cov}(\frac{n_{1h}}{N_h}, \frac{n_{2h}}{N_h} \mid \mathbf{n}_2) = n_{2h} \operatorname{cov}(\frac{n_{1h}}{N_h}, \frac{1}{N_h} \mid \mathbf{n}_2) = 0$$

on a en fait

$$\mathrm{cov}(I_{h1}, I_{i2} \mid \mathbf{n}_2) = \mathrm{E}\left(\mathrm{cov}(I_{h1}, I_{i2} \mid \mathbf{n}_1, \mathbf{n}_2) \mid \mathbf{n}_2\right).$$

On obtient donc

$$\mathrm{cov}(\sum_h \bar{Y}_h \hat{N}_h, \sum_h N_h \bar{y}_{S_{2h}} \mid \mathbf{n}_2) = -\frac{N}{n_1} \sum_h \bar{Y}_h \frac{\mathrm{E}\left(n_{1h} \mid \mathbf{n}_2\right)}{N_h - 1} \left(1_{N_h}^T C_h y_h\right) = 0$$

puisque $C_h 1_{N_h} = 0$.

Références

[Hulliger 1993] Hulliger, B. (1993). Gütertransportstatistik 1993 : Schätzverfahren mit Kompensation der Antwortausfälle. Rapport de méthode du Service de méthodes statistiques, Office fédéral de la statistique, Neuchâtel.