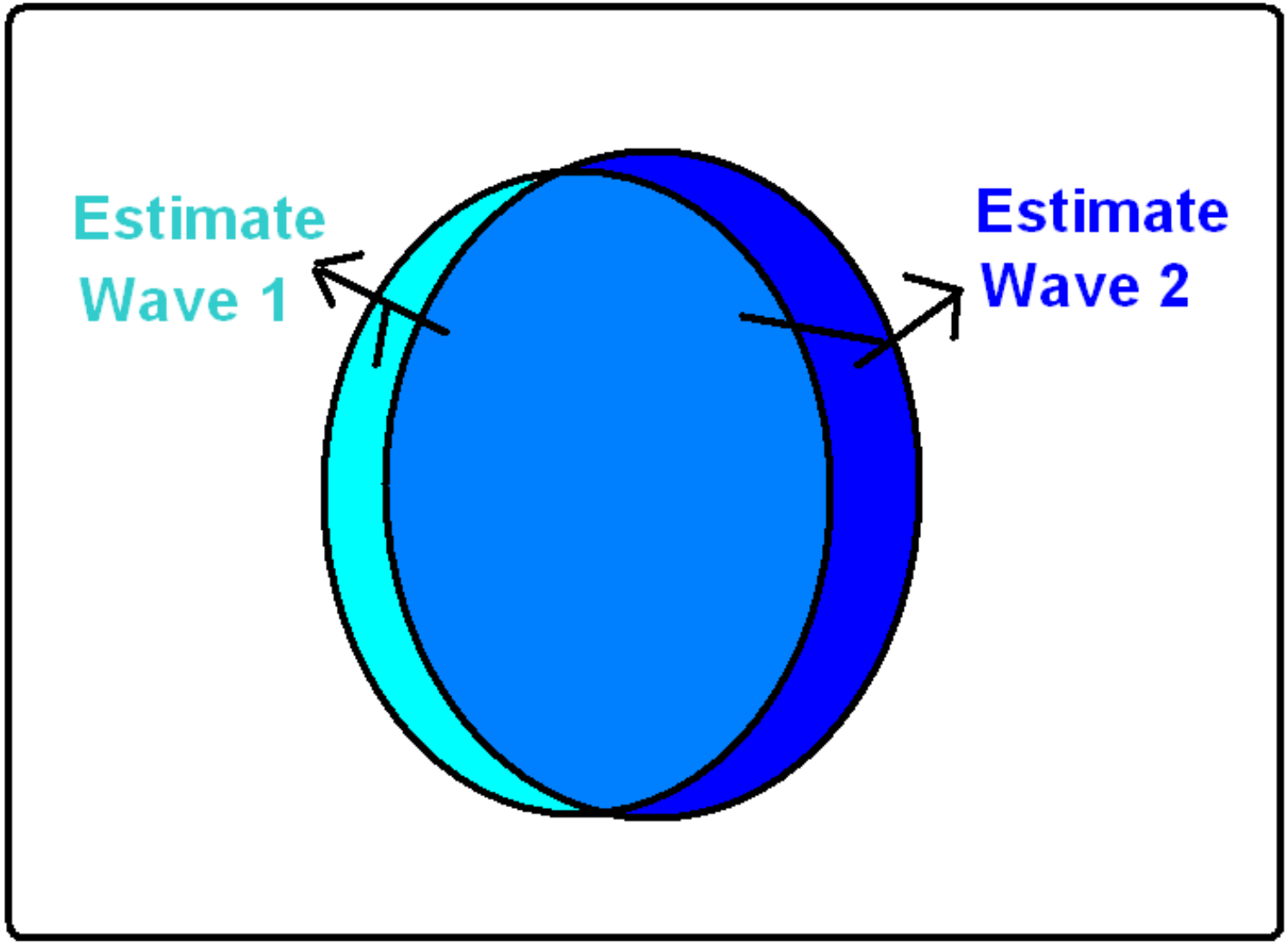
ESTIMATION DE LA VARIANCE POUR UN ESTIMATEUR IMPUTE DU CHANGEMENT TEMPOREL

Yves G. Berger et Emilio L. Escobar

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Paris, Janvier 2012

ÉCHANTILLONNAGE RÉPÉTÉ



ÉCHANTILLONNAGE RÉPÉTÉ

Vague 1	Vague 2
$y_{1;i}$	$y_{2;i}$
100	-
500	-
400	-
300	500
500	400
200	300
400	500
300	400
200	400
-	200
-	400
-	800

ÉCHANTILLONNAGE RÉPÉTÉ …

Pourquoi?

- Éviter que des unités soient interrogées pendent un long laps de temps
- Actualiser l'échantillon
- Courant dans les enquêtes longitudinales (EPA)

CHANGEMENT TEMPOREL

Vague 1

$$\tau_1 = \sum_{i \in U} y_{1;i}$$

Vague 2

$$\tau_2 = \sum_{i \in U} y_{2;i}$$

➔ **Changement =** $\Delta = \tau_2 - \tau_1$

➔ **Estimateur du changement =** $\hat{\Delta} = \hat{\tau}_2 - \hat{\tau}_1$

➔ **Variance de l'estimateur du changement :**

$$\mathrm{var}(\hat{\Delta}) = \mathrm{var}(\hat{\tau}_1) + \mathrm{var}(\hat{\tau}_2) - 2\,\mathrm{cov}(\hat{\tau}_1, \hat{\tau}_2)$$

ESTIMATION DE LA COVARIANCE

- Nous proposons d'utiliser **une approche multivariée (Berger & Priam 2010)**

- **Simulations** basées sur l'enquête sur la population active suédoise:

  Andersson et al. (2011) ont montré que cette approche donne des estimations plus exactes que les estimateurs standards de covariance (Tam, 1984; Qualité & Tillé, 2008).

ESTIMATEUR DE LA COVARIANCE

Illustration:

Covariance entre **deux estimateurs HT de totaux**

$$\hat{\tau}_1 = \sum_{i \in s_1} \frac{y_{1;i}}{\pi_{1;i}} \qquad \text{and} \qquad \hat{\tau}_2 = \sum_{i \in s_2} \frac{y_{2;i}}{\pi_{2;i}}$$

Vague 1 Vague 2

APPROCHE PAR LA REGRESSION MULTIVARIÉE

Les données

Variables dépendentes		Variables de sondage		
$\frac{y_{1;i}}{\pi_{1;i}}$	$\frac{y_{2;i}}{\pi_{2;i}}$	$z_{1;i}$	$z_{2;i}$	$z_{1;i} \times z_{2;i}$
1	0	1	0	0
4	0	1	0	0
3	4	1	1	1
2	1	1	1	1
6	7	1	1	1
5	4	1	1	1
0	2	0	1	0
0	8	0	1	0
$\hat{\tau}_1 = 26$	$\hat{\tau}_2 = 30$	$n_1 = 7$	$n_2 = 7$	$n_{12} = 4$

APPROCHE PAR LA REGRESSION MULTIVARIÉE …

- Considérons la **régression multivariée** $\breve{\mathbf{Y}} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
  (modèle de régression généralisé)

$$\begin{pmatrix} y_{1;i} / \pi_{1;i} \\ y_{2;i} / \pi_{2;i} \end{pmatrix} = \begin{pmatrix} \beta_{11} z_{1;i} + \beta_{21} z_{2;i} + \beta_{121}\ z_{1;i} \times z_{2;i} \\ \beta_{12} z_{1;i} + \beta_{22} z_{2;i} + \beta_{122}\ z_{1;i} \times z_{2;i} \end{pmatrix} + \boldsymbol{\varepsilon}_i \qquad \begin{array}{l} E(\boldsymbol{\varepsilon}_i) = \mathbf{0} \\ \mathrm{var}(\boldsymbol{\varepsilon}_i) = \mathbf{S} \end{array}$$

INTERACTIONS

➔ **Matrice de covariance Résiduelle :**

$$\begin{pmatrix} \hat{S}_{11} & \hat{S}_{12} \\ \hat{S}_{21} & \hat{S}_{22} \end{pmatrix} \rightarrow \hat{corr} = \frac{\hat{S}_{12}}{\sqrt{\hat{S}_{11}\hat{S}_{22}}}$$

➔ $\hat{\mathrm{cov}}(\hat{\tau}_1, \hat{\tau}_2) = \hat{corr} \times \sqrt{\hat{\mathrm{var}}_d(\hat{\tau}_1)\,\hat{\mathrm{var}}_d(\hat{\tau}_2)}$

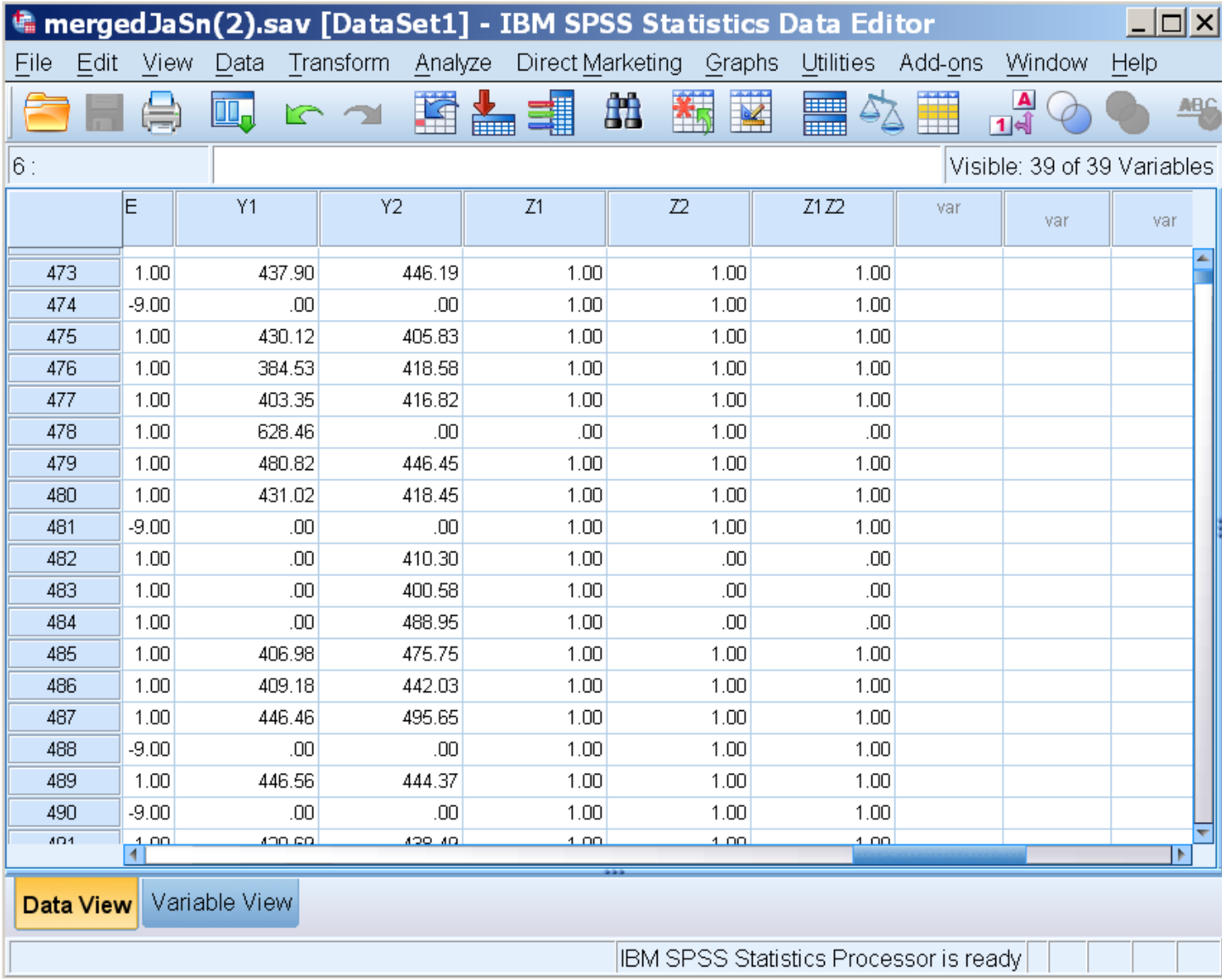
Approximativement sans biais lorsque la fraction d'échantillonnage est négligeable (Berger & Priam, 2010)

EXEMPLE British Labour Force Survey SPSS

Vague1:
Juin – Aout '98

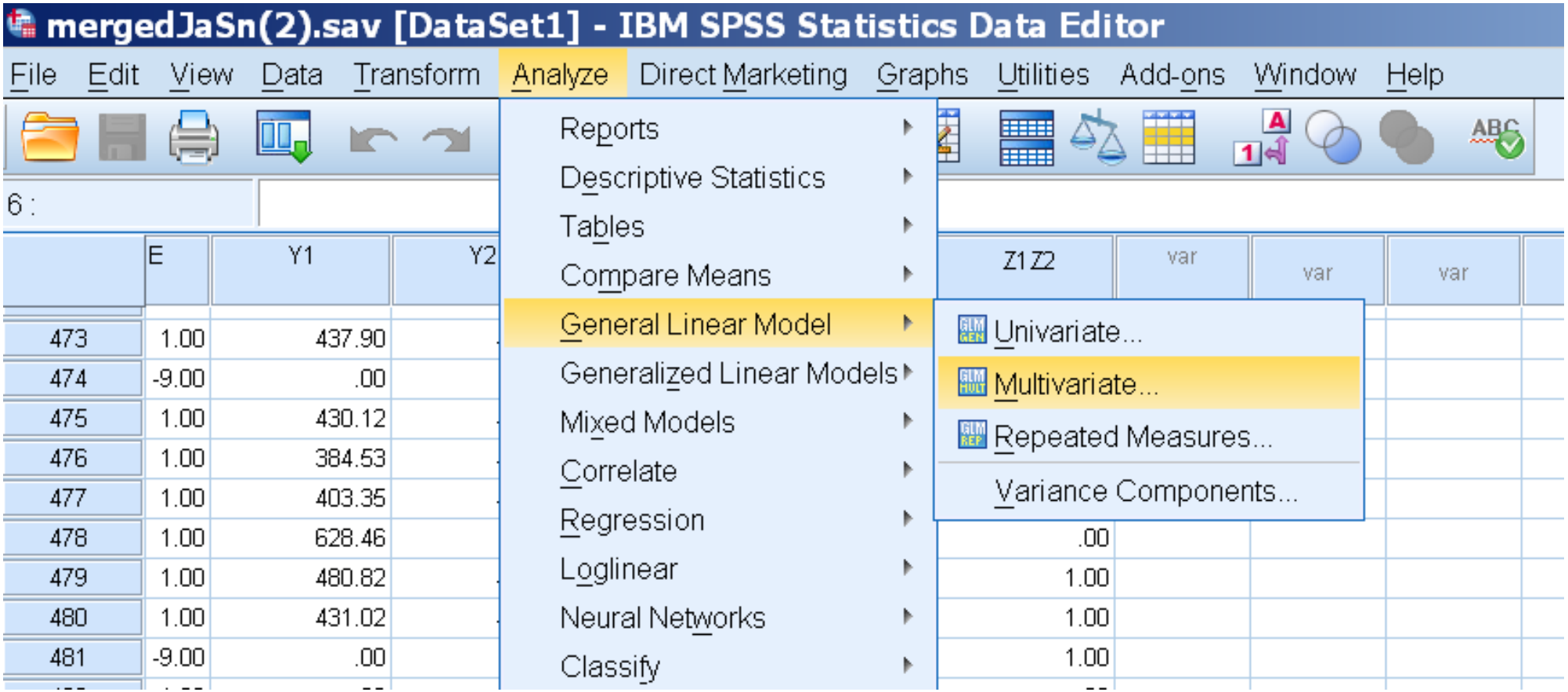
Vague 2
Sept – Nov '98

Changement du nombre d'employés



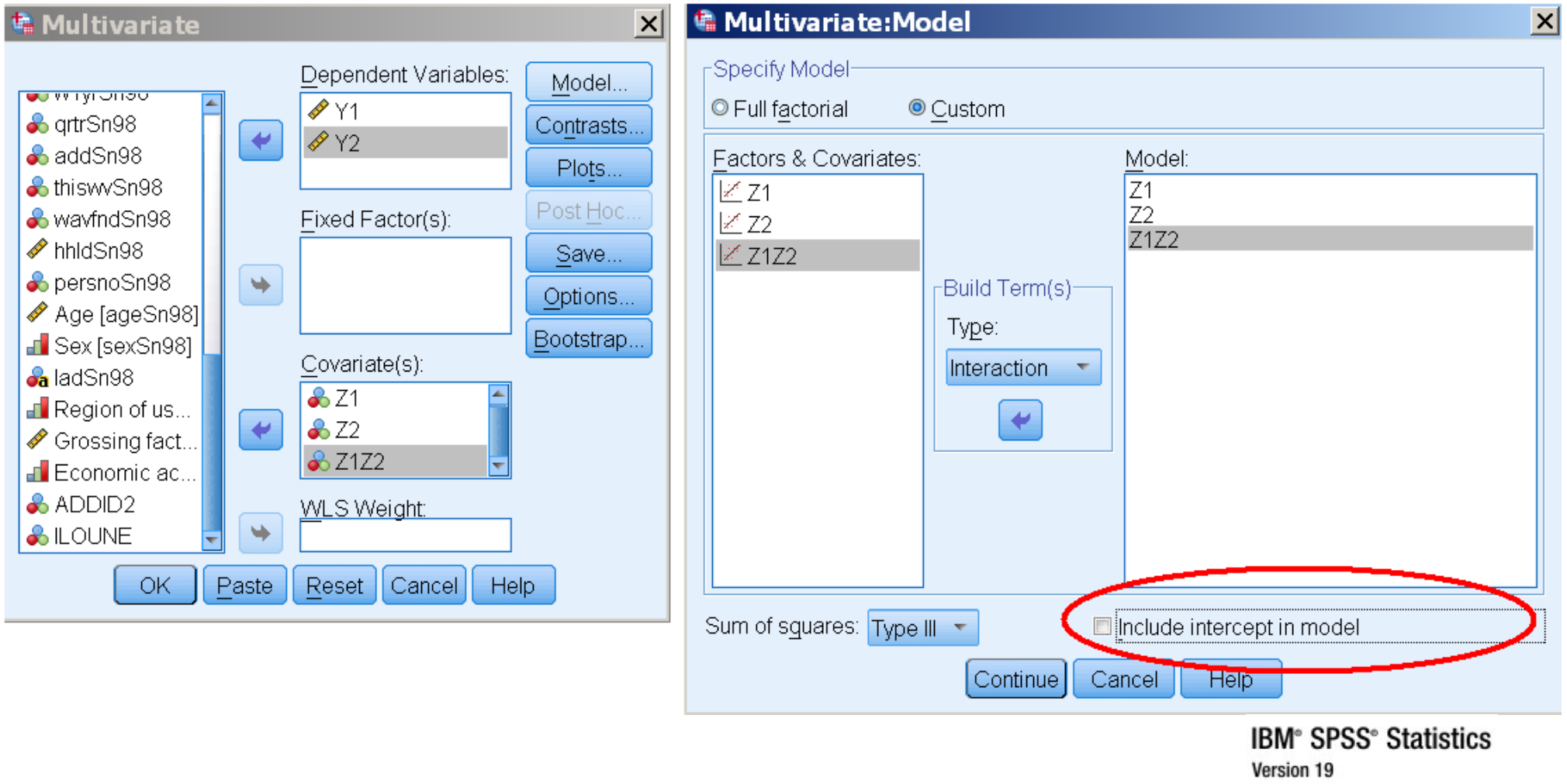
	E	Y1	Y2	Z1	Z2	Z1Z2	var	var	var
473	1.00	437.90	446.19	1.00	1.00	1.00			
474	-9.00	.00	.00	1.00	1.00	1.00			
475	1.00	430.12	405.83	1.00	1.00	1.00			
476	1.00	384.53	418.58	1.00	1.00	1.00			
477	1.00	403.35	416.82	1.00	1.00	1.00			
478	1.00	628.46	.00	.00	1.00	.00			
479	1.00	480.82	446.45	1.00	1.00	1.00			
480	1.00	431.02	418.45	1.00	1.00	1.00			
481	-9.00	.00	.00	1.00	1.00	1.00			
482	1.00	.00	410.30	1.00	.00	.00			
483	1.00	.00	400.58	1.00	.00	.00			
484	1.00	.00	488.95	1.00	.00	.00			
485	1.00	406.98	475.75	1.00	1.00	1.00			
486	1.00	409.18	442.03	1.00	1.00	1.00			
487	1.00	446.46	495.65	1.00	1.00	1.00			
488	-9.00	.00	.00	1.00	1.00	1.00			
489	1.00	446.56	444.37	1.00	1.00	1.00			
490	-9.00	.00	.00	1.00	1.00	1.00			

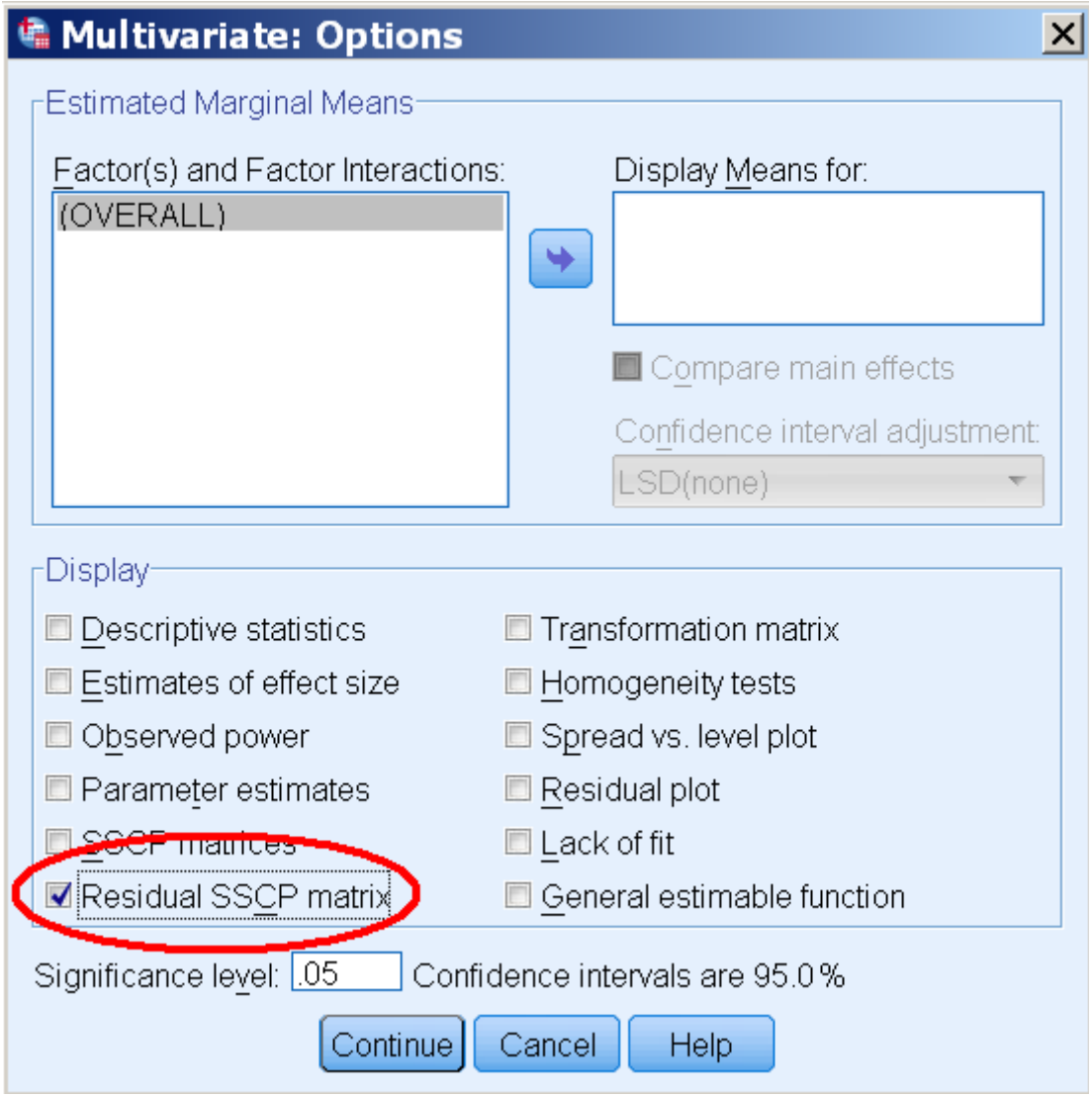
EXEMPLE British Labour Force Survey SPSS



IBM® SPSS® Statistics
Version 19

EXEMPLE British Labour Force Survey SPSS





IBM® SPSS® Statistics
Version 19

EXEMPLE British Labour Force Survey SPSS

Residual SSCP Matrix

		Y1	Y2
Sum-of-Squares and Cross-Products	Y1	4.873E9	4.775E9
	Y2	4.775E9	4.855E9
Covariance	Y1	45962.091	45037.857
	Y2	45037.857	45792.157
Correlation	Y1	1.000	.982
	Y2	.982	1.000

Based on Type III Sum of Squares

IBM® SPSS® Statistics
Version 19

➔ **CORRELATION = 0.982**

$$\text{Variance du changement} = \text{var}_1 + \text{var}_2 - 2\,\text{correlation}\sqrt{\text{var}_1 \times \text{var}_2}$$

Remarque: Les variances peuvent être calculées avec une régression simple

AUTRES LOGICIELS

SAS: procedure `REG`

STATA: La sortie `e(Signma)` de la fonction `mvregress()`

R: `estVar(lm(formula=Y~-1+Z1*Z2))`

LA STRATIFICATION

IDÉE : Ajouter des variable dans $\breve{\mathbf{Y}} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

➔ 2 variables qualitatives:

- **Variable Qualitative 1**: Stratification vague 1
  ➔ H variables dichotomique

- **Variable Qualitative 2**: Stratification vague 2
  ➔ H variables dichotomique

➔ **+ Interactions** au seins des strates

➔ $\breve{\mathbf{Y}} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ ➔ **CORRELATION**

NONRESPONSE

Vague 1	Vague 2
$y_{1;i}$	$y_{2;i}$
???	-
500	-
???	-
300	500
500	???
200	300
???	???
300	400
200	400
-	200
-	???
-	800

Mécanisme de non-réponse:

Probabilités de réponse inconnue à la vague 1 et 2

Uniforme au seins de classes

Remarque: il ne sera pas nécessaire d'estimer ces probabilités

IMPUTATION HOTDECK ALEATOIRE

Wave 1	Wave 2
$y_{1;i}$	$y_{2;i}$
300	-
500	-
200	-
300	500
500	**300**
200	300
500	**200**
300	400
200	400
-	200
-	**800**
-	800

Donneurs sélectionnés de manière aléatoire avec des probabilités inégales proportionnelles aux poids

$$w_{1;i} = 1 / \pi_{1;i}$$

Remarques:

- Dans des classes
- Les unités qui tournent ne sont pas imputées

ESTIMATEUR IMPUTÉ DU CHANGEMENT

$$y_{1;i}^{*} = \begin{cases} y_{1;i} & \text{si } y_{1;i} \text{ pas manquant} \\ \text{valeur imputée} & \text{si } y_{1;i} \text{ manquant} \end{cases}$$

$$y_{2;i}^{*} = \begin{cases} y_{2;i} & \text{si } y_{2;i} \text{ pas manquant} \\ \text{valeur imputée} & \text{si } y_{2;i} \text{ manquant} \end{cases}$$

Vague 1

$$\hat{\tau}_1^{*} = \sum_{i \in s_1} \frac{y_{1;i}^{*}}{\pi_{1;i}}$$

Vague 2

$$\hat{\tau}_2^{*} = \sum_{i \in s_1} \frac{y_{2;i}^{*}}{\pi_{2;i}}$$

➔ **Estimateur de Changement =** $\hat{\Delta}^{*} = \hat{\tau}_2^{*} - \hat{\tau}_1^{*}$

VARIANCE DE L'ESTIMATEUR IMPUTÉ DU CHANGEMENT

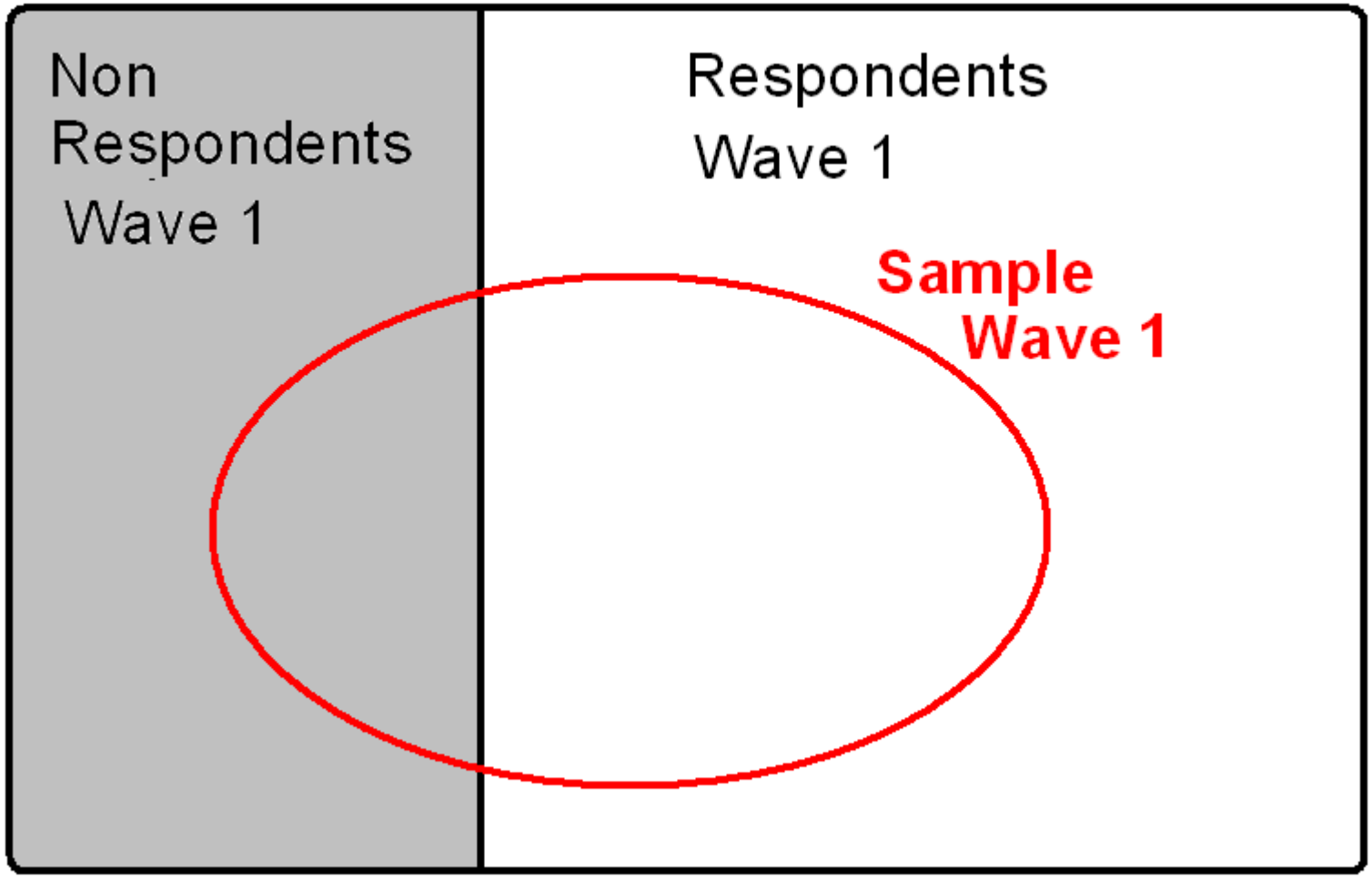
Traiter les valeurs imputées comme si elles etaient **réelles** peut conduire à **une sous-estimation sérieuse** de la variance lorsque la proportion de valeurs manquantes n'est pas petite (Rao et Shao, 1992; Särndal, 1992).

Nous voulons tenir compte de

- Effet de la non-réponse
- Effet de l'imputation aléatoire
- Effet de la rotation
- Effet de la stratification
- Effet des probabilités inégales

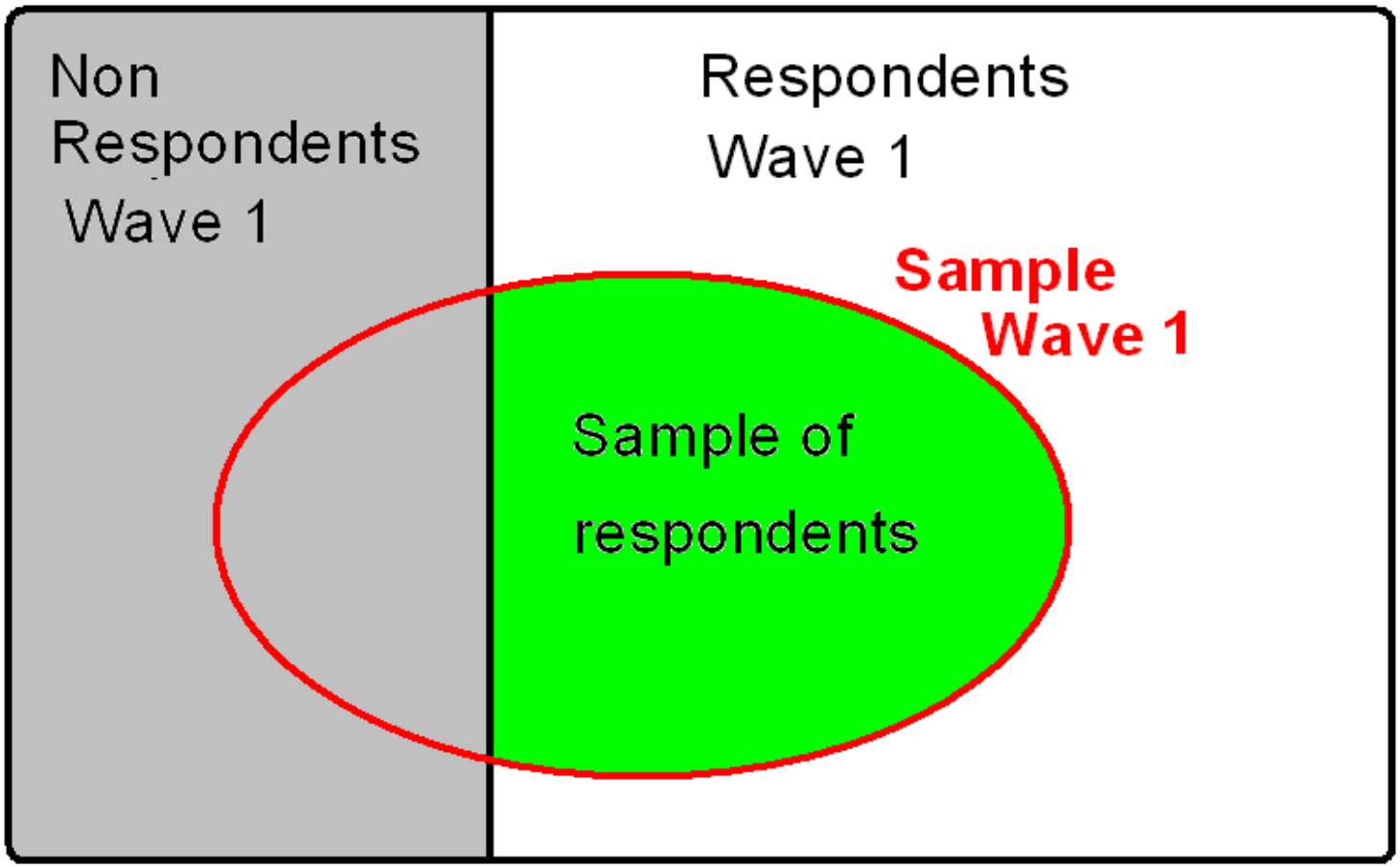
APPROCHE POUR LA NON-REPONSE

APPROCHE RENVERSÉE DE FAY'S (1991)



APPROCHE POUR LA NON-REPONSE

APPROCHE RENVERSÉE DE FAY'S (1991)



Non réponse → Échantillon → Imputation aléatoire (Hotdeck)

APPROCHE RENVERSÉE DE FAY'S (1991)

- **Non réponse survient avant la sélection de l'échantillon**

  ➔ les probabilités de réponse sont des paramètres de la population qui ne dépendent pas de l'échantillon.

- **Estimation de la variance simple et plus robuste**, car il n'est pas nécessaire d'estimer les probabilités de réponse (lorsque la fraction d'échantillonnage est négligeable)

VARIANCE DE L'ESTIMATEUR IMPUTÉ DU CHANGEMENT

Non réponse → Echantillon → Imputation aléatoire (Hotdeck)

Trois termes pour la variance.

$$\mathrm{var}(\hat{\Delta}^*) = \mathrm{E}_R\{\mathrm{var}_S[\tilde{\Delta} \mid R]\} + \mathrm{E}_R\{\mathrm{E}_S[\mathrm{var}_I(\hat{\Delta}^* \mid S,R) \mid R]\} + \mathrm{var}_R\{\mathrm{E}_S[\tilde{\Delta} \mid R]\},$$

avec $\tilde{\Delta} = \mathrm{E}_I(\hat{\Delta}^* \mid S,R)$

Espérance	Variance	Lié au
$\mathrm{E}_R(\cdot)$	$\mathrm{var}_R(\cdot)$	mécanisme de réponse
$\mathrm{E}_S(\cdot \mid R)$	$\mathrm{var}_S(\cdot \mid R)$	Échantillonnage
$\mathrm{E}_I(\cdot \mid S,R)$	$\mathrm{var}_I(\cdot \mid S,R)$	Imputation aléatoire (Hotdeck)

VARIANCE DE L'ESTIMATEUR IMPUTÉ DU CHANGEMENT …

$$\mathrm{var}(\hat{\Delta}^*) = \mathrm{E}_R\{\mathrm{var}_S[\tilde{\Delta} \mid R]\} + \mathrm{E}_R\{\mathrm{E}_S[\mathrm{var}_I(\hat{\Delta}^* \mid S,R) \mid R]\} + \mathrm{var}_R\{\mathrm{E}_S[\tilde{\Delta} \mid R]\},$$

Négligeable

$$\mathrm{var}(\hat{\Delta}^*) \approx \mathrm{E}_R\{\mathrm{var}_S[\tilde{\Delta} \mid R]\} + \mathrm{E}_R\{\mathrm{E}_S[\mathrm{var}_I(\hat{\Delta}^* \mid S,R) \mid R]\},$$

Unbiased predictor

$$\hat{\mathrm{var}}(\hat{\Delta}^*) \approx \hat{\mathrm{var}}_S(\tilde{\Delta} \mid R) + \hat{\mathrm{var}}_I(\hat{\Delta}^* \mid S,R)$$

Facile à estimer

➔ **Non Biaisé**

VARIANCE DE L'ESTIMATEUR IMPUTÉ DU CHANGEMENT ...

Estimation du premier terme

$$\hat{\text{var}}(\hat{\Delta}^*) \approx \hat{\text{var}}_S(\tilde{\Delta} \mid R) + \hat{\text{var}}_I(\hat{\Delta}^* \mid S, R)$$

$$\tilde{\Delta} = \text{E}_I(\hat{\Delta}^* \mid S, R) \mid R) = \hat{N}_2 \frac{\hat{\tau}_{2;r}}{\hat{N}_{2;r}} - \hat{N}_2 \frac{\hat{\tau}_{1;r}}{\hat{N}_{1;r}} = f(\hat{\boldsymbol{\tau}})$$

Fonction de 6 totaux $\hat{\boldsymbol{\tau}} = (\hat{N}_1, \hat{N}_2, \hat{\tau}_1^r, \hat{\tau}_2^r, \hat{N}_1^r, \hat{N}_2^r)'$

VARIANCE DE L'ESTIMATEUR IMPUTÉ DU CHANGEMENT ...

Comme $\widetilde{\Delta} = f(\hat{\boldsymbol{\tau}})$

➔ $\mathrm{var}_d(\widetilde{\Delta} \mid R) \approx \mathbf{grad}(\boldsymbol{\tau})' \mathbf{var}(\hat{\boldsymbol{\tau}} \mid R)\ \mathbf{grad}(\boldsymbol{\tau})$

$\mathbf{var}(\hat{\boldsymbol{\tau}} \mid R)$ = Matrice de covariance entre totaux de la vague 1 et 2.

= Matrice de covariance étant **donné les répondants (= échantillon fixe)**

VARIANCE DE L'ESTIMATEUR IMPUTÉ DU CHANGEMENT ...

Approche multivariée généralisée à plusieurs totaux :

➔ $\breve{\mathbf{Y}} = \mathbf{Z}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ ➔ $\hat{\mathrm{var}}(\hat{\boldsymbol{\tau}} \mid R)$

$$\hat{\mathrm{var}}_d(\tilde{\Delta} \mid R) \approx \mathbf{grad}(\hat{\boldsymbol{\tau}})' \hat{\mathrm{var}}(\hat{\boldsymbol{\tau}} \mid R)\ \mathbf{grad}(\hat{\boldsymbol{\tau}})$$

➔ Estimateur final $\hat{\mathrm{var}}(\hat{\Delta}^*) \approx \hat{\mathrm{var}}_S(\tilde{\Delta} \mid R) + \hat{\mathrm{var}}_I(\hat{\Delta}^* \mid S, R)$

CONCLUSION

- Nous utilisons l'**approche inversée** afin de prendre en compte les effets de l'imputation et la non-réponse.
- Dépend d'une **matrice de covariance entre des totaux** évalués à des vagues différentes.
- Nous utilisons une approche de **régression multivariée** pour estimer les corrélations entre les totaux.
- Les **interactions prennent en compte de la rotation**.
- **Stratification:** ajouter des variables d'échantillonnages supplémentaires

CONCLUSION …

- **Ce n'est pas une approche basée sur un modèle!!!!**

Le modèle multivariée n'a pas besoin d'ajuster les données!

« Projection dans l'espace engendré par les variables d'echantilonnages »

- L'estimateur de variance proposé repose sur l'hypothèse que les fractions d'échantillonnage sont négligeables

- **L'estimateur proposé ne dépend pas des probabilités de réponse.**

- **L' estimateur de variance est sans biais.**

Merci