Survol de quelques considérations analytiques avec les données de l'ELNEJ

Aperçu de la présentation

0. Changements apportés aux mesures directes

   A. Changements apportés aux scores pour les cycles 6 et 7
   B. Échelle de résolution de problèmes (livret 32)

Aperçu de la présentation

1. Analyse de données
   A. Définition
   B. Objectifs
2. Éléments d'une enquête complexe et leurs impacts sur l'analyse
   A. Stratification
   B. Probabilités inégales de sélection
   C. Échantillonnage en grappes
   D. Calage
   E. Non-réponse

Aperçu de la présentation

3. Recommandation sur l'utilisation des poids et l'estimation de variance
   A. Exemples
   B. Poids normalisés: est-ce suffisant?
   C. Apprivoiser les poids bootstrap
4. Logiciels et bootstrap
   A. Bootvar
   B. WesVar
   C. SUDAAN
   D. STATA

Aperçu de la présentation

5. Quelques problèmes analytiques spécifiques
   A. Situations avec de faibles comptes
   B. Combinaison d'échantillons (pooling of samples)

Changements apportés aux mesures directes

A. Changements apportés aux scores pour les cycles 6 et 7

- Suite à une recommandation résultant d'évaluations diverses (autant de l'ensemble du processus de création, de l'utilisation des scores TRI et de la distinction entre les scores TRI et les scores classiques au sein de l'ELNEJ), il a été décidé de mettre fin à la diffusion de scores TRI à compter du cycle 7.

Changements apportés aux mesures directes

A. Changements apportés aux scores pour les cycles 6 et 7

- Pour faciliter la transition à venir et permettre de procéder à des comparaisons entre les cycles 6 et 7, de nouveaux scores ont été diffusés au cycle 6:
  - Un score classique pour l'exercice de résolution de problèmes (en plus du score TRI habituellement diffusé)
  - un score classique et un score normalisé pour l'évaluation de la connaissance des nombres

Changements apportés aux mesures directes

A. Changements apportés aux scores pour les cycles 6 et 7

- On a aussi publié uniquement des scores classiques pour les deux nouvelles évaluations (évaluation des capacités de lecture (18-19 ans) et évaluation des habiletés au calcul (20-21 ans))

Changements apportés aux mesures directes

B. Échelle de résolution de problèmes (livret 32)

- Pour le cycle 7, deux questions seront retirées de l'exercice de résolution de problèmes (livret 32). L'échelle comptera donc alors 18 questions en tout.
- Il s'agit des 2 premières questions qui sont davantage des questions de calcul.
- La question #2 souffrait en plus d'un problème de traduction. La notation utilisée pour indiquer la division était méconnue de la majorité des étudiants du Québec. <u>Ce détail a été pris en compte pour la création du score TRI, mais pas lors de la création du score classique.</u>

Changements apportés aux mesures directes

B. Échelle de résolution de problèmes (livret 32)

- Un nouveau score classique sera donc créé pour l'exercice de résolution de problèmes du cycle 6.
- Il sera basé uniquement sur les 18 questions qui seront encore présentes au cycle 7, de façon à assurer du même coup une comparabilité.
- Il sera mis à la disponibilité des chercheurs uniquement <u>sur demande</u> (pas de rediffusion)
- Une note à cet effet sera ajoutée au communiqué que l'on fait parvenir aux centres de données de recherche.

Analyse de données

A. Définition

L'analyse de données est un processus de développement de réponses à des questions via l'exploration et l'interprétation de données.

Analyse de données

B. Objectifs

- Descriptifs: Les paramètres d'intérêt sont des caractéristiques de la population finie étudiée.
  - Nombre d'enfants avec un retard de développement moteur et social
  - Proportion d'enfants qui souffrent d'asthme
  - Profil du type de garde utilisée selon la province

Analyse de données

B. Objectifs

- Analytiques: L'objectif est d'observer des relations entre des variables, de trouver des causes possibles qui seraient vérifiées au-delà de la population finie au moment où elle a été échantillonnée.
  - Déterminer les facteurs reliés à l'obésité chez les enfants. Est-ce que le manque de sommeil constitue l'un de ses facteurs?
  - Déterminer les facteurs reliés à l'hyperactivité. Est-ce que le style parental joue un rôle?

Éléments d'une enquête complexe et leurs impacts sur l'analyse

- Les approches appropriées pour l'analyse vont généralement dépendre:
  - Des quantités qu'on cherche à estimer
  - Du cadre inférentiel

Éléments d'une enquête complexe et leurs impacts sur l'analyse

- Pourquoi ne pas tout simplement s'en remettre aux méthodes d'analyse offertes par les logiciels les plus communs (ex.: SAS, SPSS…)?

- Pourquoi se compliquer l'existence avec un poids final et des poids bootstrap?

Éléments d'une enquête complexe et leurs impacts sur l'analyse

- De façon souvent implicite, un certain nombre d'hypothèses sont cachés derrière la façon dont ces logiciels créent les estimations et les inférences:
  - Les variables sont généralement considérées indépendantes et identiquement distribuées (en moyenne et en variance)
  - On considère aussi parfois qu'on a une population infinie de variables à sa disposition obéissant à un certain modèle

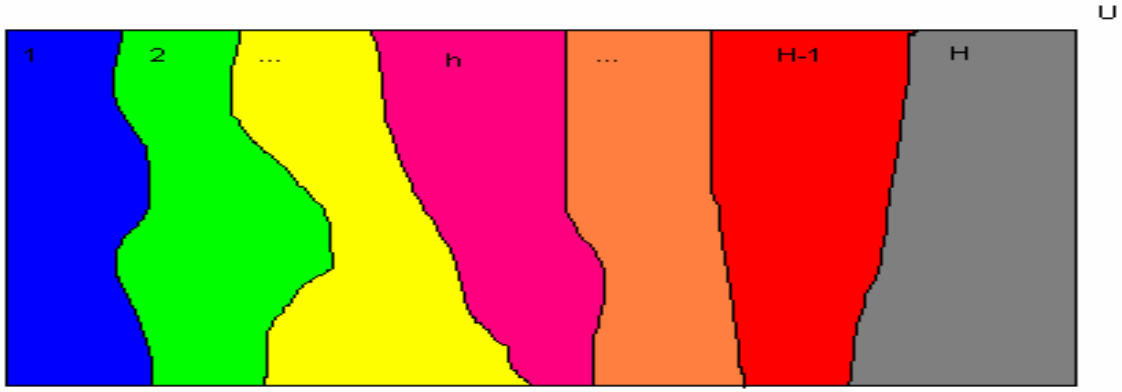
Éléments d'une enquête complexe et leurs impacts sur l'analyse

- Les enquêtes complexes ont des éléments qui violent ces hypothèses.
- Voici un survol rapide des éléments les plus communs.

Éléments d'une enquête complexe et leurs impacts sur l'analyse

A. Stratification

- La stratification est la partition exhaustive de la population en groupes disjoints appelés strates avec sélection indépendante d'échantillons dans chaque strate.



Éléments d'une enquête complexe et leurs impacts sur l'analyse

A. Stratification

- Principales raisons de son utilisation :
  - Représentation de chaque sous-population d'intérêt avec la précision voulue (si la strate est une sous-population).
  - Gains (Pertes) d'efficacité si les strates sont homogènes (hétérogènes) pour les variables étudiées.

Éléments d'une enquête complexe et leurs impacts sur l'analyse

A. Stratification

- Stratification de l'ELNEJ:

  Héritée de celle utilisée par l'EPA

  - Régions économiques et régions d'assurance-emploi
  - Régions rurales, urbaines et éloignées
  - Listes d'appartements et base aérolaire (ordinaire, revenus élevés, faible densité)

Éléments d'une enquête complexe et leurs impacts sur l'analyse

A. Stratification

Stratification de l'ELNEJ:

- Assure une bonne représentation
- Gains en efficacité sont généralement au mieux faibles

Impact sur l'analyse

- Ne pas en tenir compte peut donner lieu à des erreurs-types incorrectes (trop petites ou trop grandes).

Éléments d'une enquête complexe et leurs impacts sur l'analyse

B. Probabilités inégales de sélection

- Afin de s'assurer de pouvoir atteindre plusieurs objectifs différents (estimations nationales de qualité X, estimations provinciales de qualité Y et Z) avec un seul sondage, on a souvent recours à l'utilisation de probabilités inégales de sélection.
- A cela, s'ajoute d'autres facteurs menant à des probabilités inégales de sélection: information auxiliaire disponible, degrés ultérieurs d'échantillonnage…

Éléments d'une enquête complexe et leurs impacts sur l'analyse

B. Probabilités inégales de sélection

- Ceci résulte dans le fait que l'échantillon a souvent une distribution différente de celle de la population pour un certain nombre de caractéristiques.
- Exemple: La proportion d'enfants provenant des Maritimes est nettement plus élevée au sein de l'échantillon (24%) qu'au sein de la population canadienne (6%).

Éléments d'une enquête complexe et leurs impacts sur l'analyse

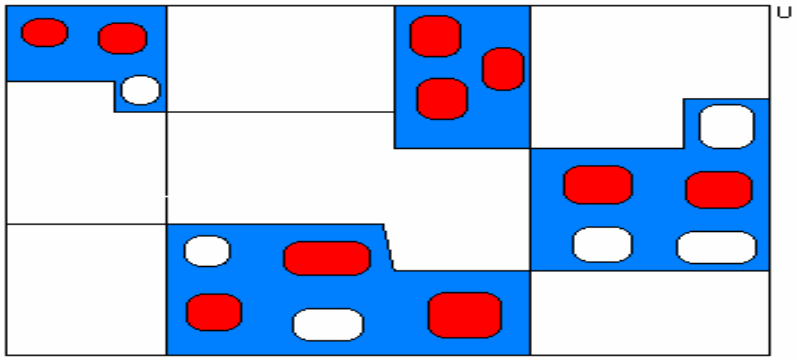
B. Probabilités inégales de sélection

- Impact sur l'analyse:
  - Si l'on ne tient pas compte des poids (c'est-à-dire du nombre d'unités de la population représenté par l'unité échantillonnée), les estimations peuvent être biaisées (si la caractéristique d'intérêt est influencée par la caractéristique dont la distribution a été modifiée).

Éléments d'une enquête complexe et leurs impacts sur l'analyse

c. Échantillonnage en grappes

- Définition : Le processus d'échantillonnage d'« unités de niveau d'agrégation plus élevé » afin d'avoir subséquemment accès aux « unités d'observation » s'appelle échantillonnage en grappes.



Éléments d'une enquête complexe et leurs impacts sur l'analyse

c. Échantillonnage en grappes

- Raison d'utilisation :
  - Il est plus facile de tenir à jour ce genre de liste à un niveau plus élevé d'agrégation (ex.: régions géographiques au lieu d'une liste de logements). Et on a seulement besoin d'obtenir une liste d'unités pour les grappes sélectionnées.
  - Concentre l'échantillon dans des groupes relativement compacts, ce qui réduit les frais de déplacement entre les unités.

Éléments d'une enquête complexe et leurs impacts sur l'analyse

c. Échantillonnage en grappes

- Impact sur l'analyse:
  - Les membres d'une même grappe ont tendance à être plus semblables que des éléments sélectionnés au hasard à partir de la population d'intérêt (exemples : habitant d'un même quartier, membres d'une même famille).
  - Diminution de l'efficacité (comparativement à l'EAS) parce qu'une unité supplémentaire provenant d'une même grappe ne fournit pas autant d'information nouvelle.

Éléments d'une enquête complexe et leurs impacts sur l'analyse

c. Échantillonnage en grappes

- Impact sur l'analyse:
  - La 'vraie' taille d'échantillon disponible est généralement inférieure à la taille obtenue:

    # grappes ≤ taille réelle ≤ # répondants
  - Si l'on ne tient pas compte de la mise en grappes, les erreurs-types présentées seront habituellement beaucoup plus faibles qu'elles ne le devraient. Les intervalles de confiance seront trop étroits et les tests d'hypothèses pourraient mener à des conclusions erronées.

Éléments d'une enquête complexe et leurs impacts sur l'analyse

D. Calage (stratification à posteriori)

- Méthode d'estimation consistant à utiliser des données supplémentaires provenant d'une source indépendante pour augmenter la fiabilité des estimations d'après l'échantillon.

- Raison d'utilisation:
  - Permet d'assurer l'uniformité des estimations produites d'après les diverses enquêtes.

Éléments d'une enquête complexe et leurs impacts sur l'analyse

D. Calage

- Impact sur l'analyse:
  - le calage peut réduire la variance si le modèle est pertinent et si la taille des classes de calage n'est pas trop petite.
- Groupes de calage de l'ELNEJ:

PROVINCE x SEXE x AGE

Éléments d'une enquête complexe et leurs impacts sur l'analyse

E. Non-réponse

- Définition:
  - Le fait de ne pas obtenir l'information demandée auprès d'une unité sélectionnée.
- Les taux de réponse varie selon divers facteurs:
  - Le sujet d'étude et la longueur du questionnaire
  - le mode et la longueur de la fenêtre de collecte
  - …

Éléments d'une enquête complexe et leurs impacts sur l'analyse

E. Non-réponse

- La non-réponse totale est habituellement traitée par repondération des unités répondantes. Ceci contribue encore plus à créer des poids inégaux.
- La non-réponse partielle des unités qui fournissent certaines données est:
  - traitée par imputation (revenu, échelle DMS)
  - laissée intact (avec un code de non-réponse approprié)

Éléments d'une enquête complexe et leurs impacts sur l'analyse

E. Non-réponse

- Impact sur les analyses:
  - Les non-répondants ont souvent des caractéristiques différentes des répondants, ce qui peut résulter en un biais dû à la non-réponse, si ceux-ci sont simplement mis de côté et qu'on base son analyse uniquement sur les répondants.
  - En ce qui concerne la non-réponse totale, les poids finaux et les poids bootstrap distribués avec les fichiers diffusés offrent une protection contre ce genre de biais.

Éléments d'une enquête complexe et leurs impacts sur l'analyse

E. Non-réponse

- Impact sur les analyses:
  - En ce qui concerne la non-réponse partielle, voici quelques solutions possibles:
    - Imputation (difficile d'en tenir compte au moment de l'estimation de la variance)
    - Repondération (Projet 'Online reweighting tool')
    - Analyse des profils respectifs et analyse de fiabilité
    - Utiliser seulement les répondants complets et croire en sa bonne étoile

Recommandation sur l'utilisation des poids et l'estimation de variance

La grande majorité de ces complexités sont reflétées par deux ensembles d'information :

- le poids final
- les poids bootstrap

Recommandation sur l'utilisation des poids et l'estimation de variance

L'utilisation de ces informations, via une approche fondée sur le plan, protège l'analyse contre des biais potentiels (tant au niveau des estimations qu'au niveau des estimations de variance)

Voici quelques exemples, illustrant l'impact de ces différents éléments, basés sur les données transversales des 0-5 ans du cycle 6.

Exemples

- Distribution des enfants par province au sein de l'échantillon et au sein de la population

Province	Sans pondération	Avec Pondération
10	5,9%	1,5%
11	4,7%	0,4%
12	6,7%	2,6%
13	6,7%	2,1%
24	15,1%	22,1%
35	26,1%	40,7%
46	8,1%	3,9%
47	7,7%	3,2%
48	9,9%	11,4%
59	9,1%	12,1%

Exemples

Distribution des enfants par type de régions

Type de régions	Sans pondération	Avec Pondération
RMR	45,9%	65,3%
AR	22,2%	15,9%
Zones fortement influencées	6,7%	5,4%
Zones modérément influencées	10,6%	6,3%
Zones faiblement influencées	13,1%	6,4%
Zones non influencées	1,5%	0,7%

Exemples

Distribution des enfants par sexe

Sexe	Sans pondération	Avec Pondération
Féminin	48,2%	48,8%
Masculin	51,8%	51,2%

Distribution des enfants par sexe pour le Nouveau-Brunswick

Sexe	Sans pondération	Avec Pondération
Féminin	54,1%	48,5%
Masculin	45,9%	51,5%

Exemples

Estimation de la proportion d'enfants âgés d'un an et de l'erreur-type associée à cette estimation

Approche	Proportion	Erreur-type
Sans poids final, sans poids bootstrap	19,6%	0,39%
Avec poids final, sans poids bootstrap	16,2%	0,36%
Avec poids final, avec poids bootstrap	16,2%	0,00%

Exemples

Nombre moyen d'heures par semaine au principal mode de garde selon le type de régions

Type de régions	Sans poids final, sans poids bootstrap	Avec poids final, sans poids bootstrap	Avec poids final, avec poids bootstrap
RMR	26,78h (0,31h)	28,32h (0,31h)	28,32h (0,39h)
AR	25,57h (0,40h)	24,64h (0,40h)	24,64h (0,54h)
Zones fortement influencées	25,80h (0,76h)	23,75h (0,72h)	23,75h (0,92h)
Zones modérément influencées	24,11h (0,61h)	23,04h (0,62h)	23,04h (0,81h)
Zones faiblement influencées	23,86h (0,56h)	22,54h (0,57h)	22,54h (0,85h)
Zones non influencées	24,13h (1,81h)	22,37h (1,77h)	22,37h (2,38h)

Poids normalisés: est-ce suffisant?

- Il n'y a pas si longtemps, la plupart des logiciels statistiques avec une approche fondée sur un modèle n'offraient pas la possibilité de mener une analyse selon une approche fondée sur le plan de sondage (qui tient compte des poids bootstrap).
- On était alors confronté aux choix suivants:
  - Apprendre un nouveau logiciel
  - Programmer ses propres macros
  - Tenter de tirer le maximum de notre logiciel habituel (et accepter la possible présence d'erreurs)

Poids normalisés: est-ce suffisant?

- L'utilisation de poids normalisés est une tentative d'ajustement pour continuer de s'en remettre à son logiciel habituel.
- Les poids normalisés prennent en considération les poids de sondage, mais pas les autres aspects du plan (stratification, échantillonnage en grappes, calibration…). Il s'agit donc d'une modification de l'approche fondée sur un modèle (pour inclure les poids) ou encore d'une application **incomplète** de l'approche fondée sur le plan de sondage.

Poids normalisés: est-ce suffisant?

- Il est recommandé que cette **approche** soit **réservée** aux cas où le plan (les poids bootstrap) n'est pas disponible, comme avec un **fichier de micro-données publiques** par exemple, ou encore aux situations où **l'analyse ne peut être encore conduite avec un logiciel avec une approche fondée sur le plan**.

Poids normalisés: est-ce suffisant?

- Afin de s'assurer que les estimations des
c de l êt
p :ir de es
s ativer ar
r stribu e,
c ion e un
p

Poids normalisés: est-ce suffisant?

- L'utilisation de [illegible] al avec certaines procédures d[illegible] on spécialisés en sondage (SA[illegible] peut causer des résultats pour [illegible] renants.

- Cela est dû au fait que le logiciel associe la somme des poids au nombre d'observations à sa disposition.

$\Rightarrow$ une puissance statistique sur - évaluée!

Poids normalisés: est-ce suffisant?

- Cas classiques:
  - Test d'indépendance avec PROC FREQ de SAS
  - Régression logistique avec PROC LOGISTIC

Exemple provenant des données de l'ENSP:

Vérifions si le trimestre de naissance est lié à l'état matrimonial…

Faut-il être né dans les trois premiers mois de l'année pour que son mariage puisse tenir le coup?

Faut-il être né dans les trois derniers mois de l'année pour demeurer célibataire?

Poids normalisés: est-ce suffisant?

- Résultats:
  - La procédure FREQ de SAS avec l'option *chisq* nous donne une valeur de X^2 de 41 637 avec une valeur-p associée inférieure à 0,0001.
  - On devrait donc en conclure que le trimestre de naissance et l'état matrimonial sont fortement liés.
  - Heureusement, avant de proclamer notre découverte à la planète entière, on pouvait remarquer la note suivante:

    Effective Sample Size = 29 457 681.54

- Comment rectifier le tir? En utilisant les poids normalisés!

Poids normalisés: est-ce suffisant?

- Qu'est-ce qu'un poids normalisé?
  - C'est une version ré-échelonnée du poids final

  - La variable contenant les poids normalisés a la propriété que sa somme donne exactement le nombre d'unités impliquées dans l'analyse. Le nombre effectif d'observations est donc plus près de ce qu'il devrait être.

Poids normalisés: est-ce suffisant?

- Un exemple de normalisation:

Identificateur	Poids Final	Poids Normalisé
1	1,00	0,25
2	3,00	0,75
3	4,00	1,00
4	4,00	1,00
5	6,00	1,50
6	6,00	1,50
Total	24,00	6

Poids normalisés: est-ce suffisant?

- Comment normaliser?
  - Mathématiquement:
    - Il suffit de diviser le poids final de chaque unité utilisé dans l'analyse par la moyenne (non-pondérée) des poids finaux de toutes les unités analysées.

$$w_k^{norm} = \frac{w_k^{final}}{\overline{w}^{final}}$$

    - Dans l'exemple précédent, on a 6 observations et une somme des poids finaux de 24. La moyenne est donc 4. On divise donc chaque poids par 4.

Poids normalisés: est-ce suffisant?

- Comment normaliser?
  - Au niveau informatique:
    - Il est rapide d'utiliser un code similaire au suivant:

```
proc sql;
   create table data2 as
   select *, poidsfinal/mean(poidsfinal) as poidsnorm
   from data;
/* On suppose ici qu'il n'y a aucune non-réponse et
   qu'on procède à une analyse au niveau global. */
quit;
```

Poids normalisés: est-ce suffisant?

- Est-ce suffisant de normaliser?

  Dans le cadre des enquêtes à plan complexe, le nombre effectif d'unités est généralement inférieur au nombre d'observations dans l'échantillon. Ceci est généralement lié aux effets de grappe (corrélation entre les observations d'une même grappe) et parfois aussi à la stratification (stratification non efficace pour assurer une représentativité).

Poids normalisés: est-ce suffisant?

- Est-ce suffisant de normaliser?

  Dans ces cas, la normalisation mène à:

  - Une sur-estimation du nombre effectif d'observations
  - Une sous-estimation de la variabilité
  - Un trop grand nombre de résultats significatifs

Poids normalisés: est-ce suffisant?

- Est-ce suffisant de normaliser?

Pour corriger encore une fois le tir, certains utilisateurs de poids normalisés vont adopter une règle du pouce et recourir à un niveau de signification plus conservateur (1% au lieu de 5%) avant de déclarer un résultat significatif.

Mais cette règle demeure une règle du pouce. Elle est parfois trop sévère, et parfois pas suffisamment…

Poids normalisés: est-ce suffisant?

- Retour à l'exemple du trimestre de naissance et de l'état matrimonial:
  - Résultat après normalisation:
    - SAS: une valeur de X^2 de 24,33 (p=0,0038).
    - On conclurait donc encore une fois que le trimestre de naissance et l'état matrimonial sont liés, et ce même en adoptant la règle du pouce de 1%.
  - Résultat avec un logiciel avec une approche fondée sur le plan de sondage:
    - SUDAAN: une valeur de X^2 de 14,95 (p=0,0955)
    - On conclut à l'indépendance entre le trimestre de naissance et l'état matrimonial... OUF!

Poids normalisés: est-ce suffisant?

- Un autre exemple avec les données de l'ELNEJ:
  - Indépendance entre la sécurité du voisinage et la province de résidence (QC et ONT)
    - Résultat après normalisation:
      - SAS: une valeur de X^2 de 24,94 ($p<0,0001$).
    - Résultat avec un logiciel avec une approche fondée sur le plan de sondage:
      - SUDAAN: une valeur de X^2 de 9,55 ($p=0,0232$)

Poids normalisés: est-ce suffisant?

- Conclusion:
  - Avec des **logiciels utilisant une approche fondée sur un modèle**, **la normalisation est une tentative de récupérer l'usage d'un certain nombre de procédures.**
  - Elle constitue une **application incomplète** de l'approche fondée sur le plan de sondage car elle tient compte des poids, mais pas des autres aspects du plan.

Poids normalisés: est-ce suffisant?

- Conclusion:
  - Elle mène généralement à une **sous-estimation de la variance** des estimations et à un trop grand nombre de résultats significatifs.
  - On adopte très souvent une **règle du pouce** pour tenter de compenser. Cette façon de faire peut être **parfois trop conservatrice, parfois pas suffisamment**.
  - NOTE: Avec des **logiciels utilisant une approche fondée sur le plan de sondage, la normalisation n'est pas requise**.

Apprivoiser le bootstrap

- Avec une approche fondée sur le plan, l'estimation de la variance est l'estimation de la variabilité due aux diverses étapes d'échantillonnage et de pondération (y compris la correction pour la non-réponse et le calage) aboutissant au calcul de l'estimation.

- Afin de calculer des estimations de la variance qui tiennent compte des diverses sources de variabilité, une méthode simple consisterait à tirer un nombre énorme d'échantillons au moyen du même plan de sondage complexe, à suivre les mêmes étapes de pondération et à calculer l'estimation résultante pour chaque échantillon. On pourrait alors calculer une estimation de la variabilité en évaluant la variabilité de ces estimations. Naturellement, il ne s'agit pas d'une option pratique, car elle nécessiterait trop de temps et de ressources.

Apprivoiser le bootstrap

- L'estimation de la variance par rééchantillonnage a pour principe d'utiliser uniquement l'échantillon existant pour construire des échantillons « synthétiques », appelés répliques. Les diverses méthodes (jackknife, bootstrap, répliques équilibrées répétées) se distinguent par la façon dont les répliques sont produites et par le nombre de répliques à produire.
- Pour l'utilisateur, les diverses méthodes sont similaires:
  - On dispose d'un certain nombre de répliques
  - On utilise chacune d'elles pour obtenir une estimation
  - On utilise toutes les estimations pour calculer la variance selon la formule propre à chaque méthode

Apprivoiser le bootstrap

- Bien que le nombre de répliques et la façon dont les répliques sont générées diffèrent, la structure de la formule de variance du bootstrap et des RER est la même.

$$\hat{V}_{BOOT}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_b - \hat{\theta}\right)^2, \quad \hat{V}_{RER}(\hat{\theta}) = \frac{1}{G}\sum_{g=1}^{G}\left(\hat{\theta}_g - \hat{\theta}\right)^2$$

- **Par conséquent, en autant que les poids bootstrap soient calculés à l'extérieur du logiciel, la variance estimée par un logiciel avec l'option RER (BRR) est une estimation appropriée de la variance bootstrap.**

Logiciels et bootstrap

- Plusieurs logiciels d'analyse différents peuvent être utilisés si on choisit d'utiliser l'approche fondée sur le plan (qui tient compte des poids et des poids bootstrap). Notons, entre autres,
  - Bootvar (SAS et SPSS)
  - WesVar 4.2
  - SUDAAN 9.0.1
  - STATA 9.0

Logiciels et bootstrap

- Bootvar 3.1:
  - Il s'agit d'un ensemble de macros SAS/SPSS développés à Statistique Canada pour faciliter l'obtention d'estimations de variance bootstrap pour des données d'enquêtes.

    http://www.statcan.ca/francais/rdc/whatdata_f.htm

    - Pour: Relativement facile à utiliser, support, coût, résultats et flexibilité
    - Contre: Dichotomisation des variables catégorielles, limité quand aux types d'analyse.

Logiciels et bootstrap

- WesVar 4.2:
  - Logiciel pour le calcul d'estimations et des estimations de variance associées à partir de données d'enquêtes qui utilisent une méthode de répliques.
    - Pour: facile d'utilisation (menus, pointe-et-clique), bonne documentation
    - Contre: support faible, limité quand aux types d'analyse, utilisation des résultats, échange entre les logiciels…

Logiciels et bootstrap

- SUDAAN 9.0.1:
  - Logiciel, maintenant exécutable à l'intérieur même de SAS, permettant le calcul d'estimations et des estimations de variance associées à partir de données d'enquête
  - Capable d'accomoder diverses approches pour l'estimation de la variance fondée sur le plan, incluant plusieurs méthodes de répliques (dont RER(BRR))
  - SUDAAN utilise un langage similaire à celui de SAS

Logiciels et bootstrap

- SUDAAN 9.0.1:
  - Pour: Bonne variété de plans de sondage, d'approches à l'estimation de variance, et de type d'analyses, support, une spécialité, pas un à-côté…
  - Contre: Restrictions quant aux types de variables, codage, les fichiers de sortie

Logiciels et bootstrap

- **Exemples de base de procédures SUDAAN: Estimations de moyennes**

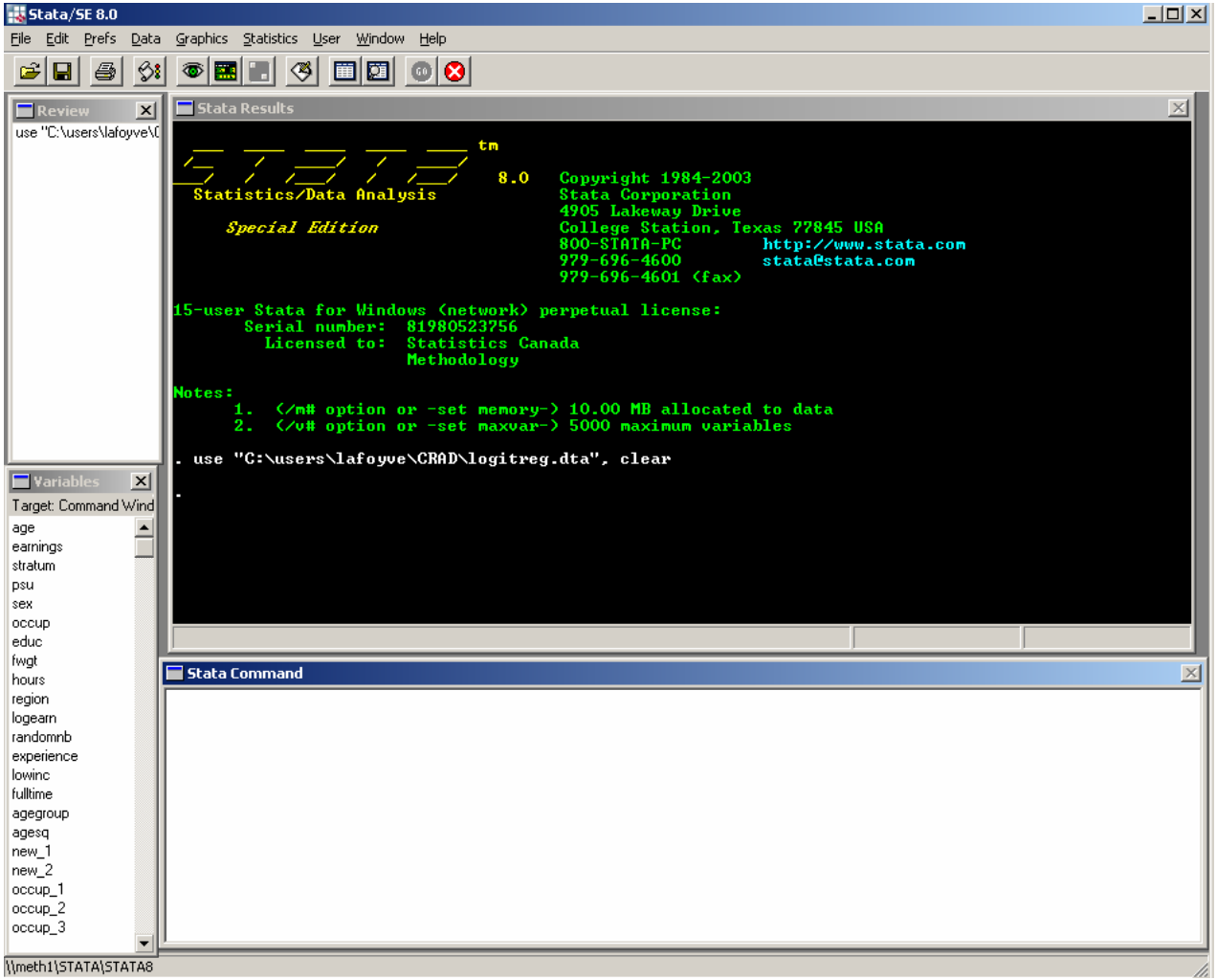
proc means data=combo2c3; weight wt68; class dhc8_sex; var inc8_4; run;	proc descript data=combo2c3 design=brr; weight wt68; repwgt bsw1-bsw500; class dhc8_sex; var inc8_4; run;

Note: SAS n'utilise pas une approche selon le plan dans cet exemple.

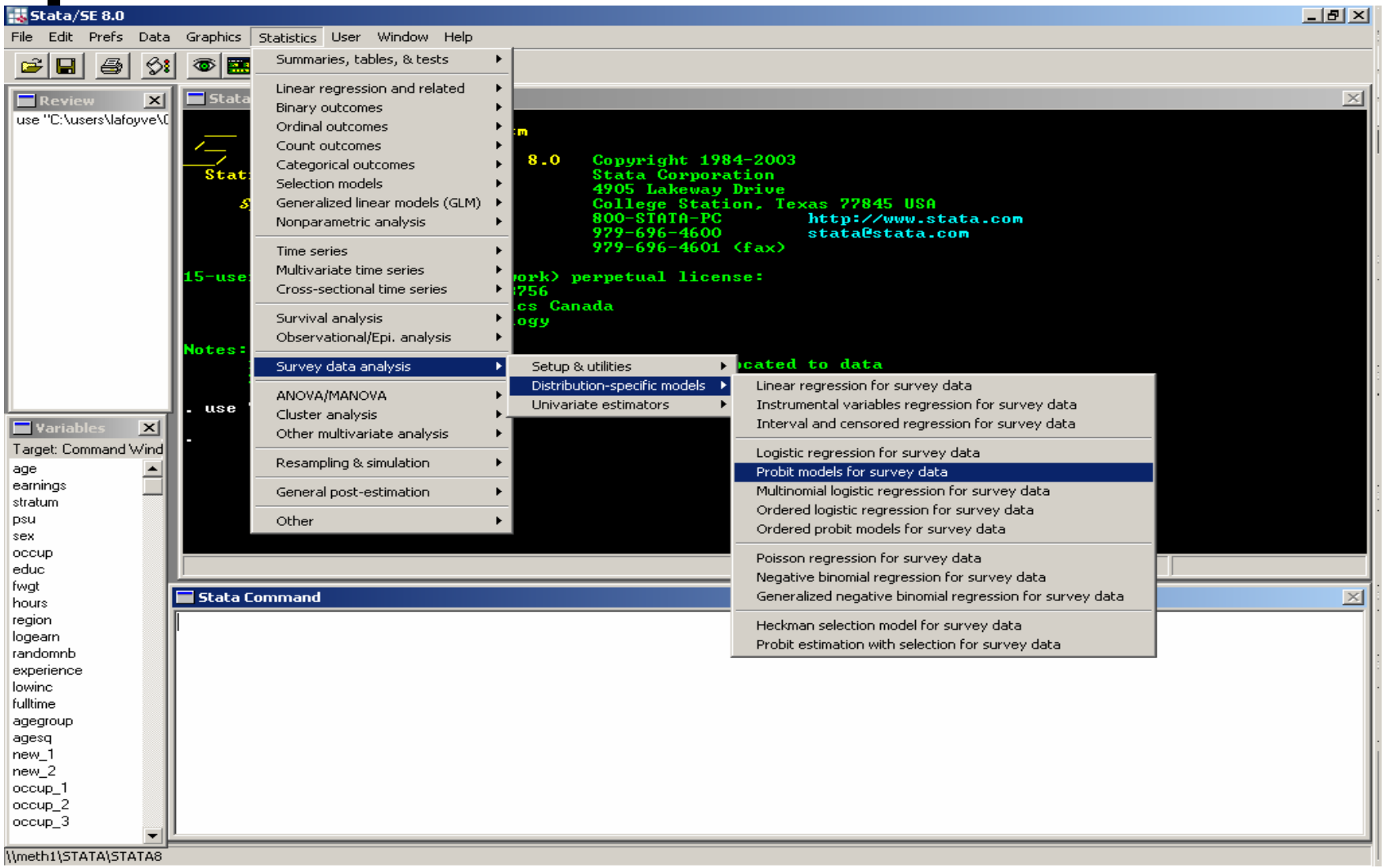
Logiciels et bootstrap

- STATA 9.0:
  - Les commandes SVY de STATA calculent des estimations et les estimations de variance associées à partir de données d'enquête.
  - Applicable depuis la version 9.0 à une approche RÉR (donc bootstrap).
  - Mode interactif (semblable à SPLUS ou R)
  - Possible de se limiter à l'utilisation des menus

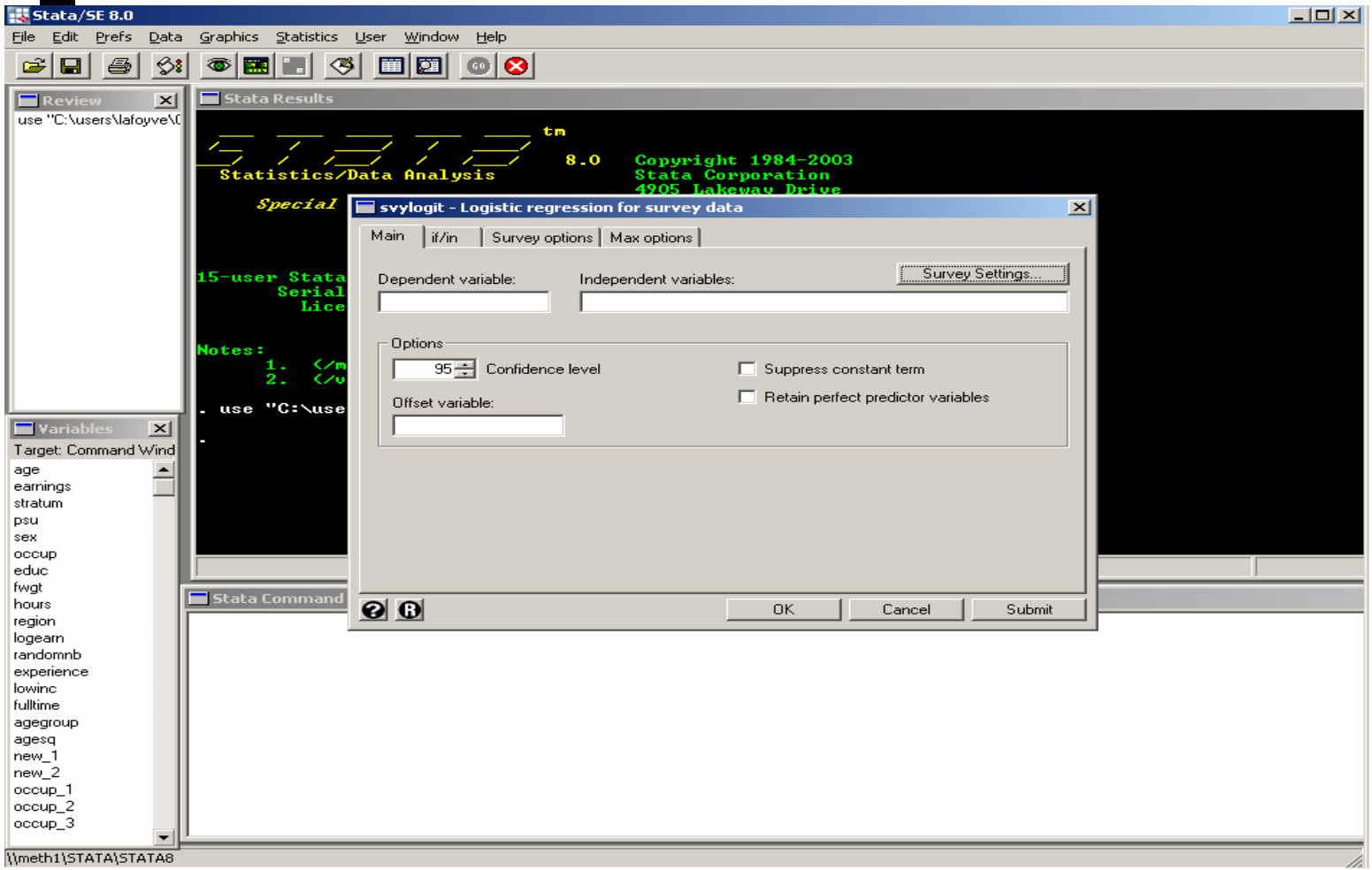
Logiciels et bootstrap



Logiciels et bootstrap



Logiciels et bootstrap



Logiciels et bootstrap

Stata/SE 8.0

File Edit Prefs Data Graphics Statistics User Window Help

Review

use "C:\users\lafoyve\[
svylogit lowinc new_1 o

Variables

Target: Command Wind

age
earnings
stratum
psu
sex
occup
educ
fwgt
hours
region
logearn
randomnb
experience
lowinc
fulltime
agegroup
agesq
new_1
new_2
occup_1
occup_2
occup_3

Stata Results

```
. svylogit lowinc new_1 occup_1 occup_2 occup_3 occup_4 occup_5 educ_1 educ_2 educ_3 region_1 region_
> 3 region_4 fulltime_1 age, ci prob

Survey logistic regression

pweight:  fwgt                                  Number of obs     =     14339
Strata:   stratum                               Number of strata  =       220
PSU:      psu                                   Number of PSUs    =       440
                                                Population size   =  11788719
                                                F(  15,     206)  =    139.81
                                                Prob > F          =    0.0000

------------------------------------------------------------------------------
      lowinc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       new_1 |  -1.285124   .0878031   -14.64   0.000    -1.458167   -1.112081
     occup_1 |  -2.224407   .1712202   -12.99   0.000    -2.561849   -1.886966
     occup_2 |  -1.754152   .1592532   -11.01   0.000    -2.068009   -1.440295
     occup_3 |  -1.123528   .2398181    -4.68   0.000    -1.596163   -.6508935
     occup_4 |     -1.471   .1233217   -11.93   0.000    -1.714043   -1.227957
     occup_5 |  -.2908822   .1079805    -2.69   0.008    -.5036907   -.0780737
      educ_1 |   .6910478   .1563885     4.42   0.000     .3828364    .9992591
      educ_2 |   .4423678   .1406198     3.15   0.002     .1652334    .7195021
      educ_3 |   .1235053   .1369406     0.90   0.368     -.146378    .3933886
    region_1 |   .0762247   .1252711     0.61   0.543    -.1706604    .3231098
    region_2 |   .1346074   .1705749     0.79   0.431    -.2015625    .4707774
    region_3 |   .1269839   .1467931     0.87   0.388    -.1623169    .4162846
    region_4 |   .2620834   .1481378     1.77   0.078    -.0298675    .5540343
--more--
```

Stata Command

\\meth1\STATA\STATA8

Logiciels et bootstrap

- STATA 9.0:
  - Pour: Grande variété de type d'analyses, menus et codage, possibilité de créer ses propres routines, support
  - Contre: Utilisation des résultats et échange avec les autres logiciels

Tableau sommaire des outils d'analyse fondée sur le plan de sondage, disponibles dans quelques logiciels sélectionnés

Logiciel	SUDAAN 9	WesVar 4.2	Stata 9.0	Bootvar	SAS 9.1
Méthode d'estimation de la variance	**BRR (Bootstrap) Jackknife Série de Taylor**	**BRR (Bootstrap) Jackknife**	**BRR (Bootstrap) Jackknife Série de Taylor**	**Bootstrap (BRR)**	**Série de Taylor**
Modélisation					
Régression linéaire	*proc regress*	Oui	*svyreg*	*%regress*	*proc surveyreg*
Régression logistique	*proc logistic* (*rlogist*)	Oui	*svylogit*	*%logreg*	*proc surveylogistic*
Modèles logits generalises	*Proc multilog*	Oui	*svymlog*	Non	*proc surveylogistic*
Modèles des odds proportionnelles	*Proc multilog*	Non	*svyolog*	Non	*proc surveylogistic*
Régression de Poisson et log-linéaire	*Proc loglink*	Non	*svypois*	Non	Non
Régression probit	Non	Non	*svyprobt*	Non	*proc surveylogistic*
Régression probit ordonnée	Non	Non	*svyoprob*	Non	*proc surveylogistic*
Modèles à risques proportionnels	*proc survival*	Non	Non	Non	Non
Régression par variables instrumentales	Non	Non	*svyireg*	Non	Non
Régression par intervalles	Non	Non	*svyintrg*	Non	Non
Modèles de Heckman	Non	Non	*svyheck*	Non	Non
Statistiques descriptives					
Moyennes	*Proc descript*	Oui	*svymean*	*%ratio*	*proc surveymeans*
Totaux	*proc descript*	Oui	*svytotal*	*%total*	*proc surveymeans*
Proportions	*proc descript*	Oui	*svyprop*	*%ratio*	*proc surveymeans*
Ratios	*proc ratio*	Oui	*svyratio*	*%ratio*	*proc surveymeans*
tests d'indépendance	*proc crosstab*	Oui	*svytab*	*%chi2*	*proc surveyfreq*
Quantiles	*proc descript*	Oui	Non	*%prcntle*	Non
Valeurs plausibles / Imputation multiple	Certains	Certains	Non	Non	Non

Quelques problèmes analytiques spécifiques

A. Situations avec de faibles comptes

- Même si la variance due à la sélection de l'échantillon et aux autres étapes de pondération est bien capturée par les poids bootstrap, cela ne veut pas dire pour autant que le bootstrap est une solution magique à tous les problèmes d'analyse.

Quelques problèmes analytiques spécifiques

A. Situations avec de faibles comptes

- Lors d'une analyse, il est possible que, pour un certain nombre de répliques, les poids bootstrap de toutes les unités (ou d'un certain nombre d'unités clés) du domaine d'intérêt soient nuls.
- Ces situations entraînent leur part de complications…

Quelques problèmes analytiques spécifiques

A. Situations avec de faibles comptes

- En effet, si, pour une certaine réplique, les poids bootstrap des observations du domaine d'intérêt sont tous zéros, l'estimateur usuel de la moyenne n'est pas défini car le dénominateur est 0.

$$\hat{\bar{y}}_d = \frac{\sum_{h=1}^{H}\sum_{i=1}^{i_h}\sum_{j=1}^{j_{hi}}\sum_{k=1}^{k_{hij}} w_{hijk} y_{hijk} I_{hijk}}{\sum_{h=1}^{H}\sum_{i=1}^{i_h}\sum_{j=1}^{j_{hi}}\sum_{k=1}^{k_{hij}} w_{hijk} I_{hijk}}$$

Quelques problèmes analytiques spécifiques

A. Situations avec de faibles comptes

- L'estimateur de l'erreur-type devrait être modifié pour tenir compte du nombre réel de répliques utilisées.
- Par exemple, dans le cas du bootstrap, si seulement 490 des 500 répliques permettent d'obtenir une estimation pour le domaine d'intérêt, l'estimateur de la variance devrait être :

$$\hat{V}_{BOOT}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_b - \hat{\theta}\right)^2$$

avec *B=490* et non *B*=500.

Quelques problèmes analytiques spécifiques

A. Situations avec de faibles comptes

Les logiciels diffèrent dans la façon de traiter ces cas:

- **SUDAAN** affiche un message dans le journal et utilise une valeur corrigée de *B*.
- **BootVar** affiche le nombre de répliques utilisées et utilise une valeur corrigée de *B*.
- **WesVar** affiche une valeur manquante comme estimation de variance.
- **STATA** produit une estimation de 0 et utilise toutes les répliques pour calculer la variance (incluant les estimations 0).

Quelques problèmes analytiques spécifiques

A. Situations avec de faibles comptes

- Un problème plus fréquent est celui où un certain nombre d'observations clés ont toutes un poids de zéro pour une réplique donnée.

- Cela peut réduire le rang de la matrice X'WX et causer des problèmes d'estimations lors d'analyses impliquant la modélisation.

Quelques problèmes analytiques spécifiques

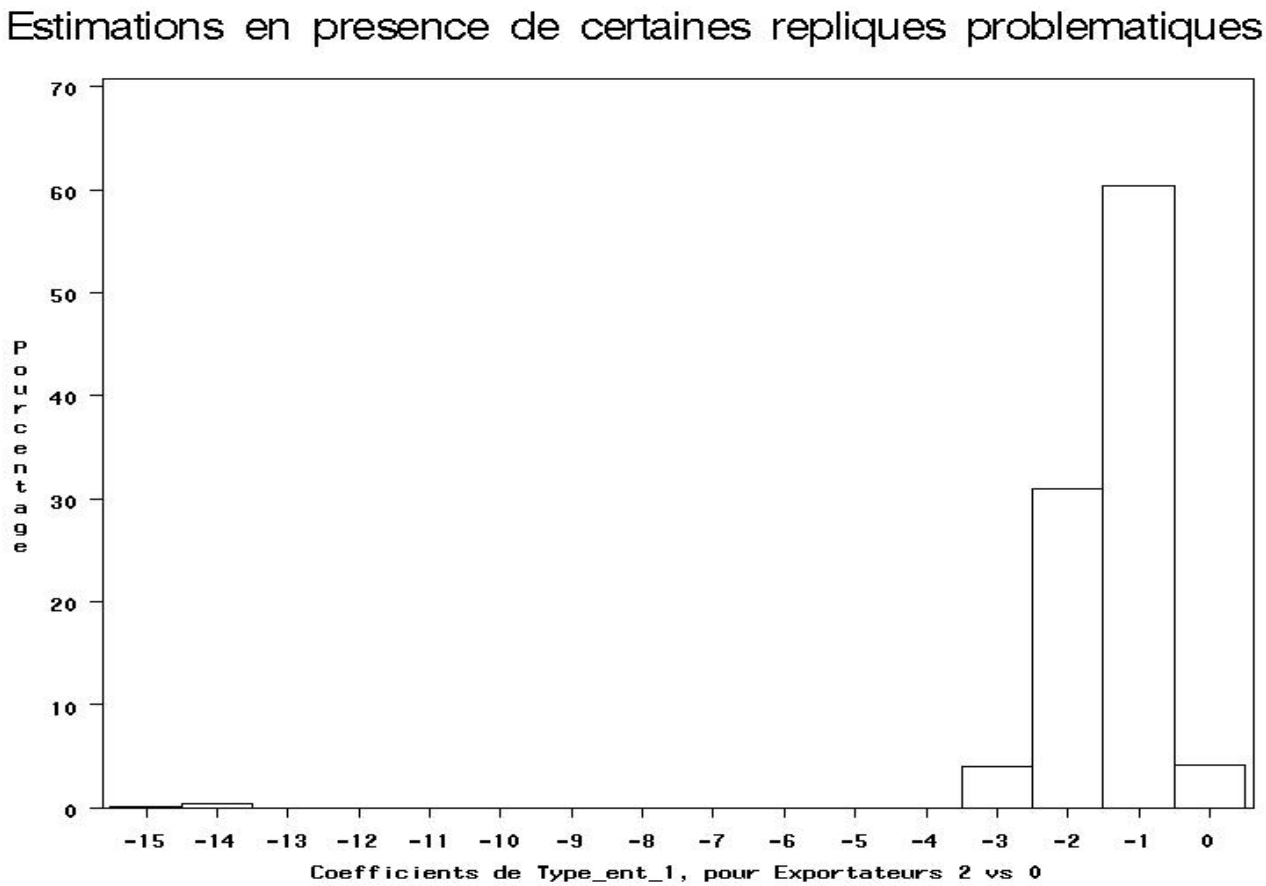
A. Situations avec de faibles comptes

- Exemple:

Type d'expor-tateurs	Comptes non pondérés des observations avec des poids supérieurs à 0 dans l'échantillon		Comptes non pondérés des observations avec des poids supérieurs à 0 pour la réplique #22	
	Type d'entreprises_1		Type d'entreprises_1	
	Oui	Non	Oui	Non
0	20	300	15	190
1	35	200	20	125
2	5	150	0	95

Quelques problèmes analytiques spécifiques

A. Situations avec de faibles comptes



Quelques problèmes analytiques spécifiques

A. Situations avec de faibles comptes

- Exemple:

Erreur-type associée avec toutes les répliques	Erreur-type associée sans les répliques problématiques
1,065	0,550

Dans cet exemple, environ la moitié de la variabilité dans l'estimation du coefficient provenait de 5 des 1000 répliques…

Quelques problèmes analytiques spécifiques

A. Situations avec de faibles comptes

- Des façons possibles d'éviter ce genre de situations:
  - Examiner le journal, pas seulement la sortie

SUDAAN: DATA WARNING:

One or more p-values are approaching 0 or 1 in replicate 22. The data may have singularities for the model you are trying to fit. You may want to remove this replicate from the analysis by removing the corresponding variable from the REPWGT statement.

WesVar: Estimates are converging slowly or are diverging for replicates bsw22, bsw413, bsw571, bsw781, bsw949. This may affect both parameter and standard error estimates. Using a smaller log likelihood convergence criteria may improve parameter convergence. If not, review the model and data to assess whether parameters are finite.

Quelques problèmes analytiques spécifiques

A. Situations avec de faibles comptes

- D'autres façons possibles d'éviter ce genre de situations:
  - Analyser les estimations obtenues à l'aide des répliques (ex.: produire un histogramme) (en autant qu'elles soient disponibles…)
  - Recourir au bootstrap moyen?

Quelques problèmes analytiques spécifiques

B. Combinaison d'échantillons

Puisque la puissance statistique est fonction de la taille de l'échantillon, il est logique de parfois vouloir chercher à maximiser la taille disponible.

Il arrive aussi parfois qu'on veuille plutôt procéder à une analyse avec des mesures répétées sur la même personne.

Dans un cas, comme dans l'autre, on va généralement procéder à une forme de combinaison d'échantillons (sample pooling).

Quelques problèmes analytiques spécifiques

B. Combinaison d'échantillons

Selon que la combinaison implique ou non des échantillons dépendants, la façon de procéder à l'analyse sera légèrement différente.

Si un même individu (PERSRUK) se retrouve dans l'échantillon à plus d'une reprise, on parle de combinaison d'échantillons dépendants. Sinon, on parle de combinaison d'échantillons indépendants.

Quelques problèmes analytiques spécifiques

Étapes pour la combinaison d'échantillons indépendants:

1- Redéfinir la population-cible

2- Utilisez les poids de sondage et les poids bootstrap appropriés pour chaque unité (peut nécessiter des ajustements).

En autant que tous les membres des deux échantillons combinés peuvent être choisis, cette méthode est simple car elle traite chaque année de naissance comme une strate et permet d'utiliser les poids de sondage sans devoir y apporter des ajustements.

Quelques problèmes analytiques spécifiques

Premier exemple:

On combine les répondants transversaux du groupe d'âge 0-1 an de plusieurs cycles (cycle 4 - cycle 6).

On définit la population de référence comme étant les enfants de 0-1 an au sein des populations transversales de 2000, 2002 et 2004.

On utilise simplement les poids de sondage et les poids bootstrap transversaux associés à chaque unité au cycle où l'unité est âgé de 0-1 an.

Quelques problèmes analytiques spécifiques

Deuxième exemple:

On combine les répondants longitudinaux de 4-5 ans au cycle 3 aux répondants longitudinaux du même groupe d'âge au cycle 4.

On définit la population de référence comme étant les enfants de 0-1 an au sein des populations de 1994 et 1996 qui ont atteint l'âge de 4-5 ans.

Quelques problèmes analytiques spécifiques

Deuxième exemple:

On utilise alors les poids de sondage et les poids bootstrap longitudinaux associés à chaque unité au cycle où l'unité est âgé de 4-5 ans.

Quelques problèmes analytiques spécifiques

Troisième exemple:

On combine les répondants transversaux de 0-5 ans du cycle 2 aux répondants transversaux de 0-5 ans du cycle 3 qui n'étaient pas du cycle 2.

Comme les poids transversaux des 5 ans du cycle 3 tiennent compte à la fois de la présence des répondants longitudinaux et de l'échantillon supplémentaire issue du Registre des naissances, on ne peut pas utiliser directement les poids. Il faudrait procéder à un ajustement des poids.

Quelques problèmes analytiques spécifiques

Combinaison d'échantillons dépendants:

Lorsque des échantillons dépendants sont combinés, il est important de tenir compte des structures de corrélation inhérentes aux observations provenant d'un même individu, en plus des autres dépendances inhérentes au plan d'échantillonnage.

Quelques problèmes analytiques spécifiques

Combinaison d'échantillons dépendants:

Il est, entre autres, possible de le faire dans la situation suivante:

Exemple:

On veut construire un modèle marginal pour les 6-10 ans. On utilise les observations répétées provenant des répondants longitudinaux de la cohorte originale.

Quelques problèmes analytiques spécifiques

Combinaison d'échantillons dépendants:

Exemple (suite):

Les répondants de 9-10 ans au cycle 1 contribueront potentiellement une seule fois. Les répondants de 7-8 ans contribueront potentiellement 2 fois, et ceux de 6 ans et moins contribueront potentiellement 3 fois.

Quelques problèmes analytiques spécifiques

Combinaison d'échantillons dépendants:

Exemple (suite):

Il est possible de considérer les observations répétées comme un simple degré additionnelle d'échantillonnage.

Et si <u>tous</u> les événements associés à chaque individu sont sélectionnés, la probabilité de sélection d'un événement est la même que celle de l'individu.

Quelques problèmes analytiques spécifiques

Combinaison d'échantillons dépendants:

Exemple (suite):

Par conséquent, on peut créer un fichier où chaque événement est un enregistrement, et associer à cet événement les poids de sondage et les poids bootstrap de l'individu ayant vécu cet événement.

Ainsi, les structures de corrélations seront capturées par le recours aux poids bootstrap.

Quelques problèmes analytiques spécifiques

Combinaison d'échantillons dépendants:

Exemple (suite):

Enregistrement	Individu	W	x_{kt}	y_{kt}	bs1	bs2
1	1	w_1	x_{11}	y_{11}	$bs1_1$	$bs2_1$
2	1	w_1	x_{12}	y_{12}	$bs1_1$	$bs2_1$
3	2	w_2	x_{21}	y_{21}	$bs1_2$	$bs2_2$
4	3	w_3	x_{31}	y_{31}	$bs1_3$	$bs2_3$

Quelques problèmes analytiques spécifiques

Combinaison d'échantillons dépendants:

Exemple (suite):

En absence présumée de biais de non-réponse, le poids longitudinal du cycle 1 (et ses poids bootstrap correspondant) serait le choix le plus naturel et approprié.

En présence présumée de biais de non-réponse, il semble probablement plus naturel de s'en remettre aux poids entonnoir (funnel weights) du dernier cycle combiné, même si cela pourrait résulter en une certaine perte d'échantillon.

Pour nous contacter…

Les services à la clientèle de la DES:

- Par courriel: ssd@statcan.ca
- Par téléphone: 613-951-3321 ou
  1-800-461-9050