



Statistique
Canada

Statistics
Canada

Canada



Statistique Canada
www.statcan.gc.ca



Atelier ELNEJ

Atelier présenté au CIQSS

Par

Yves Lafortune

Vendredi, le 12 février 2010

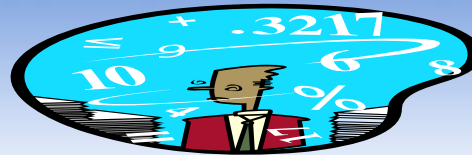


Aperçu de l'atelier

- Survol rapide de l'ELNEJ
- Trois concepts statistiques liés à l'analyse des données de l'ELNEJ:
 - Poids normalisés
 - Non-réponse
 - Regroupement de cohortes
- Un exemple complet:
 - “L'utilisation de l'ordinateur par les adolescents”



Survole rapide de l'ELNEJ





Qu'est-ce que l'ELNEJ?

- Une étude à long terme qui vise principalement à observer le développement et le bien-être des enfants au Canada dans leur cheminement de la naissance à l'âge adulte
- Menée par Statistique Canada et parrainée par Ressources humaines et Développement des Compétences Canada (RHDCC)

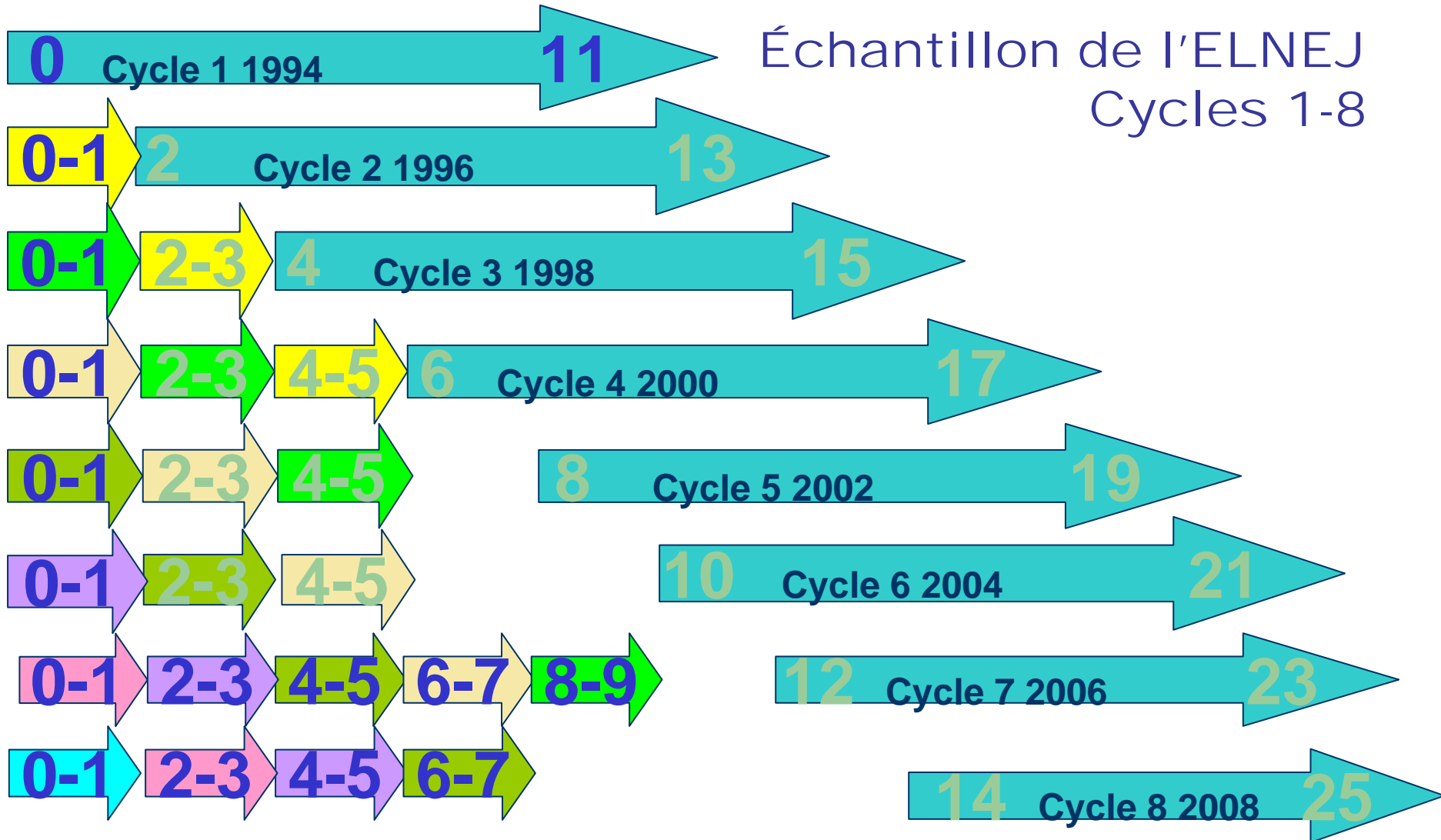
Objectifs de l'ELNEJ

- De déterminer la fréquence de divers facteurs de risque et de protection chez les enfants et les jeunes
- Comprendre comment ces facteurs, tout comme les événements de la vie, influent sur le développement de l'enfance
- Mettre cette information au service de l'élaboration de politiques et de programmes destinés à aider les enfants et les jeunes
- Recueillir des renseignements sur une grande diversité de sujets d'ordre biologique, social et économique
- Recueillir des renseignements sur le milieu où grandit l'enfant, qu'il s'agisse de la famille, des pairs, de l'école ou de la collectivité



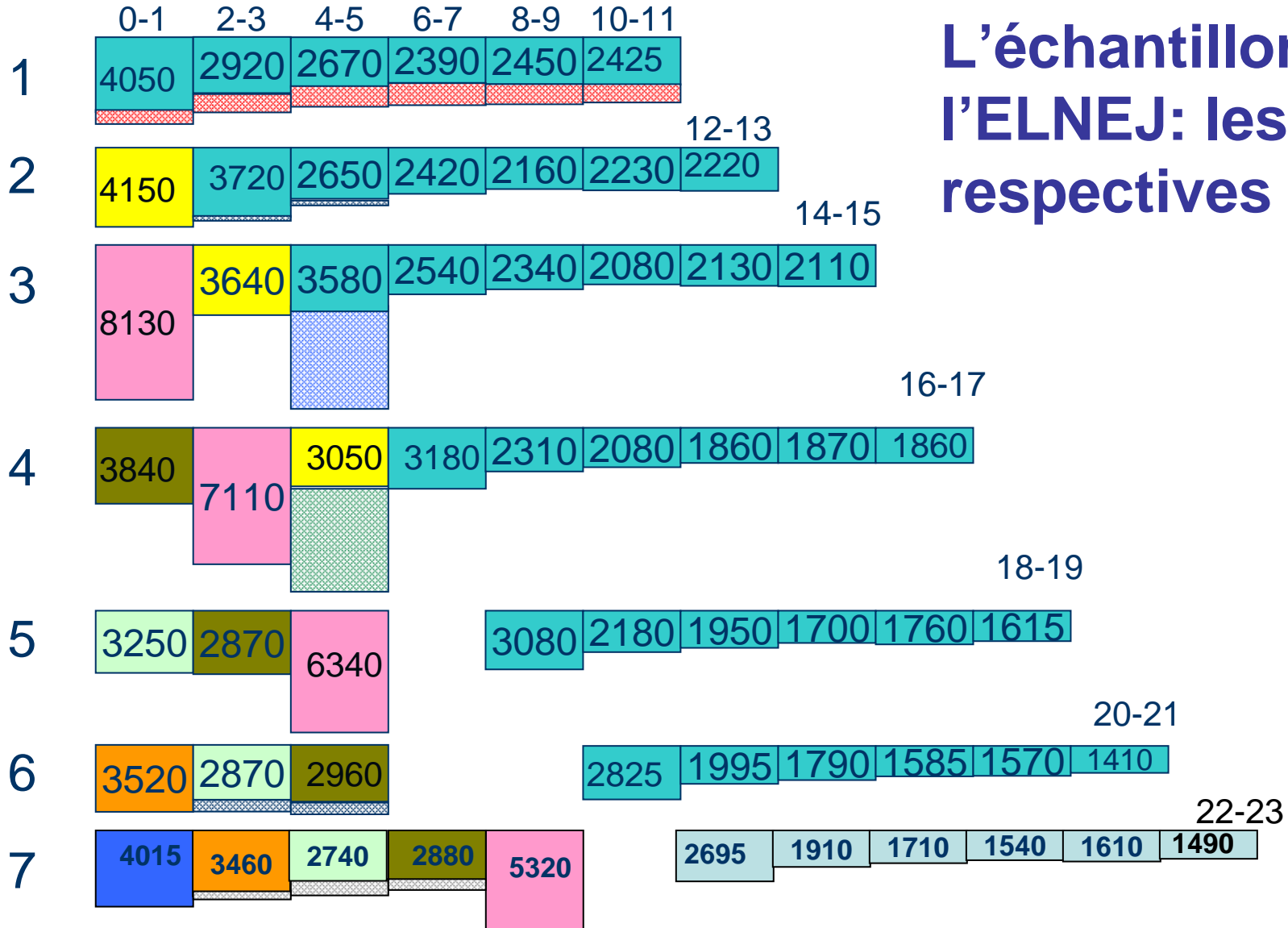
Statut actuel de l'ELNEJ

- Sept cycles de données dans les CDR
- Données du Cycle 8 (2008-09) seront accessibles à compter de l'automne 2010.





L'échantillon de l'ELNEJ: les tailles respectives





Questionnaires

PMR	Enfant/Jeune	Professeur/Directeur (cycles 1-5) / (cycles 1-4)
<ul style="list-style-type: none">• Ménage• Adulte• Enfant	<ul style="list-style-type: none">• Auto-administré• Composante du jeune	<ul style="list-style-type: none">• Maternelle• Primaire
	<ul style="list-style-type: none">• Mesures directes	

Préparation

- **Revoir la documentation de l'ELNEJ**
- **Choisir et travailler avec les données**
- **Identifier les contraintes**



Documentation de l'ELNEJ

■ Guides d'utilisateur

- http://www.statcan.gc.ca/imdb-bmdi/document/4450_D4_T9_V6-fra.pdf

■ Aperçu d'enquête

- http://www.statcan.gc.ca/imdb-bmdi/document/4450_D2_T9_V2-fra.pdf

■ Questionnaires

- http://www.statcan.gc.ca/imdb-bmdi/instrument/4450_Q2_V5-fra.pdf
- http://www.statcan.gc.ca/imdb-bmdi/instrument/4450_Q3_V5-fra.pdf

■ Dictionnaires des données

■ Errata

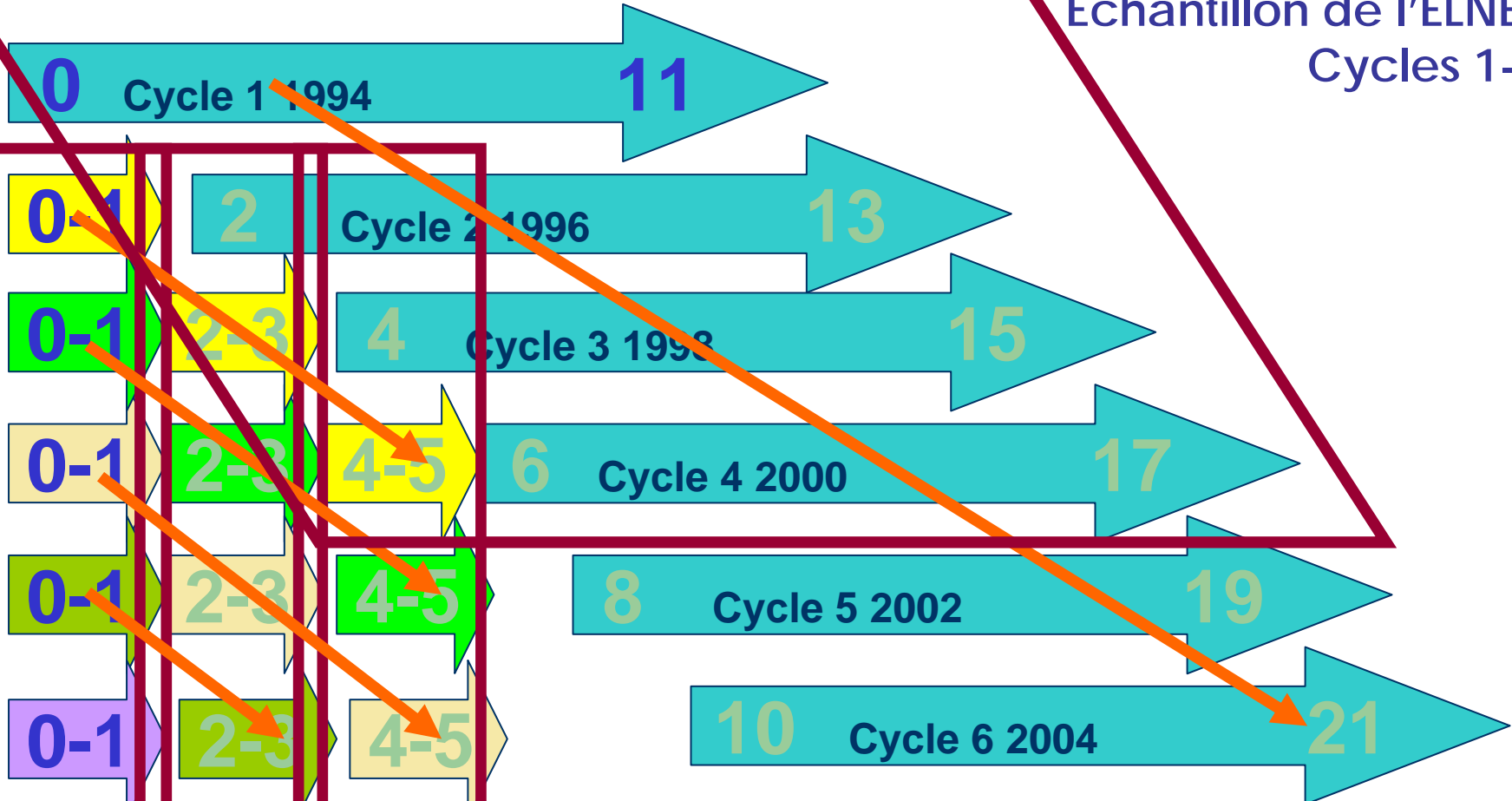
- http://www.statcan.gc.ca/imdb-bmdi/document/4450_D30_T9_V1-fra.pdf

■ Bulletin (Fenêtre sur l'ELNEJ)

- http://www.statcan.gc.ca/imdb-bmdi/document/4450_D3_T9_V1-fra.pdf



Échantillon de l'ELNEJ
Cycles 1-6



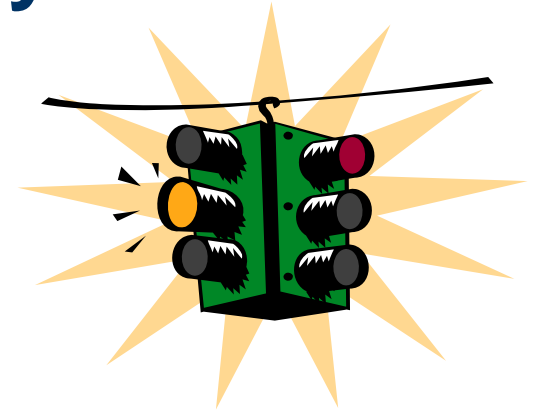
Orange transverse

Données – Cycle 6

- Quatre fichiers de données:
 - Cohorte longitudinale (10-17 ans) – Adulte, Enfant et ménage:
NLSCY2004_C6_LONG_REV_Master.txt
 - Cohorte longitudinale (16-21 ans) – Jeune:
NLSCY2004_C6_YOUTH_Master.txt
 - Fichier ECD file (0-5 ans) :
NLSCY2004_C6_ECD_REV_Master.txt
 - Fichier auto-administré (10-17 ans) :
NLSCY2004_C6_1017_Master.txt
- Fichiers Bootstrap
- Structure des fichiers a changé à travers les cycles. Important de consulter la documentation.

Quelques contraintes à considérer

- **Unité d'analyse**
- **Érosion et non-réponse**
- **Changements au cours des cycles**
 - **Tailles d'échantillon**
 - **Règles de collecte**
 - **Âge**
 - **Contenu (nouvelles questions, changements aux questions...)**





Statistique Canada
www.statcan.gc.ca



Poids normalisés



Poids normalisés: est-ce suffisant?

- Il n'y a pas si longtemps encore, 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.
- 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 microdonnées à grande diffusion (FMGD)** par exemple, ou encore aux situations où **l'analyse ne peut pas être encore réalisée avec un logiciel offrant une approche fondée sur le plan.**

Poids normalisés: est-ce suffisant?

- Afin de s'assurer que les estimations des caractéristiques de la population soient précises sans biais, les poids sont normalisés :



Poids normalisés: est-ce suffisant?

- L'utilisation de procédures de l (SAS, SPSS) peut être surprenants.
- Cela est dû au fait que ces logiciels ne prennent pas en compte les poids au noir.



mal avec certaines procédures spécialisées en sondage. Les résultats pour le moins

et associe la somme des poids à sa disposition.

⇒ une puissance statistique surévaluée!

Poids normalisés: est-ce suffisant?

- Cas classiques:

Test d'indépendance avec PROC FREQ de SAS

Régression logistique avec PROC LOGISTIC

Exemple à partir des données du cycle 6 de l'ELNEJ:

Les différents concepts seront présentés en détail un peu plus tard, mais supposons pour l'instant que nous sommes intéressés à vérifier si l'étendue de l'usage des ordinateurs faits par les adolescents est liée à la situation de travail/études du/des parent(s).

Poids normalisés: est-ce suffisant?

- Résultats:

- La procédure FREQ de SAS avec l'option *chisq* nous donne une valeur de X^2 de 8 929,7088 avec une valeur-p associée inférieure à 0,0001.
- On devrait donc en conclure que la situation de travail/études des parents et l'étendue de l'usage des ordinateurs par les adolescents sont fortement liés.
- Heureusement, avant de proclamer notre découverte à la planète entière, on pouvait remarquer la note suivante:

Effective Sample Size = 1 816 357,2108

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 comme 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 utilisée 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  
  where in_analysis=1;  
/* On suppose ici que les unités participant à l'analyse ont  
été identifiées à l'aide de la variable dichotomique  
in_analysis. */  
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 surestimation 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 la situation, 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 sur l'exemple du lien entre l'usage de l'ordinateur et la situation de travail/études des parents:

Résultats après normalisation:

- SAS: une valeur de X^2 de 25,9481 ($p < 0.0001$).
- On conclurait donc encore une fois que l'usage de l'ordinateur et la situation de travail/études des parents sont fortement liés, et ce, même en adoptant la règle du pouce de 1%.

Poids normalisés: est-ce suffisant?

- Retour sur l'exemple du lien entre l'usage de l'ordinateur et la situation de travail/études des parents:

Résultat avec un logiciel utilisant une approche fondée sur le plan de sondage:

- SUDAAN: une valeur de X^2 de 1,75 ($p=0,1212$).
- On en conclut que le lien n'est pas vraiment significatif finalement.

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 (ou une compensation basée sur des effets de plan). 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 et pourrait même mener à des erreurs (pour l'estimation d'un total par exemple).

Poids normalisés: est-ce suffisant?

- Conclusion:
 - Lorsque l'analyse est possible (une approche fondée sur le plan de sondage a été développée pour le type d'analyse réalisée) et que les données requises sont disponibles (travail à partir d'un CDR, pas à partir d'un FMGD) les poids bootstrap devraient être utilisés.
 - Plusieurs logiciels et ensembles de macros permettent et facilitent l'utilisation des poids bootstrap: SUDAAN, STATA, WesVar, Bootvar...



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

Logiciel	SUDAAN 10	WesVar 5.1	Stata 11	Bootvar (Version SAS 3.1)	SAS 9.2
Méthode d'estimation de la variance	BRR (Bootstrap) Jackknife Linéarisation	BRR (Bootstrap) Jackknife	BRR (Bootstrap) Jackknife Linéarisation	Bootstrap (BRR)	BRR Jackknife Linéarisation
Modélisation					
Régression linéaire	<i>proc regress</i>	Oui	<i>svyreg</i>	<i>%regress</i>	<i>proc surveyreg</i>
Régression logistique	<i>proc logistic (rlogist)</i>	Oui	<i>svylogit</i>	<i>%logreg</i>	<i>proc surveylogistic</i>
Modèles logits generalises	<i>Proc multilog</i>	Oui	<i>svymlog</i>	Non	<i>proc surveylogistic</i>
Modèles des odds proportionnelles	<i>Proc multilog</i>	Non	<i>svyolog</i>	Non	<i>proc surveylogistic</i>
Régression de Poisson et log-linéaire	<i>Proc loglink</i>	Non	<i>svypoiss</i>	Non	Non
Régression probit	Non	Non	<i>svyprobit</i>	Non	<i>proc surveylogistic</i>
Régression probit ordonnée	Non	Non	<i>svyoprobit</i>	Non	<i>proc surveylogistic</i>
Modèles à risques proportionnels	<i>proc survival</i>	Non	Oui	Non	Non
Régression par variables instrumentales	Non	Non	<i>svyivreg</i>	Non	Non
Régression par intervalles	Non	Non	<i>svyintrg</i>	Non	Non
Modèles de Heckman	Non	Non	<i>svyheck</i>	Non	Non
Statistiques descriptives					
Moyennes	<i>Proc descript</i>	Oui	<i>svymean</i>	<i>%ratio</i>	<i>proc surveymeans</i>
Totaux	<i>proc descript</i>	Oui	<i>svytotal</i>	<i>%total</i>	<i>proc surveymeans</i>
Proportions	<i>proc descript</i>	Oui	<i>svyprop</i>	<i>%ratio</i>	<i>proc surveymeans</i>
Ratios	<i>proc ratio</i>	Oui	<i>svyratio</i>	<i>%ratio</i>	<i>proc surveymeans</i>
tests d'indépendance	<i>proc crosstab</i>	Oui	<i>svytab</i>	<i>%chi2</i>	<i>proc surveyfreq</i>
Quantiles	<i>proc descript</i>	Oui	Oui	<i>%prntle</i>	Oui
Valeurs plausibles / Imputation multiple	Certains	Certains	Non	Non	Non



Non-réponse

Comment faire face à la non-réponse
avec l'ELNEJ



Section sur la non-réponse:

- Qu'est-ce que la non-réponse?
- Types de non-réponse
- Pourquoi la non-réponse est-elle un enjeu pour les analystes?
- Non-réponse et l'ELNEJ
- Techniques pour traiter la non-réponse partielle
- Exemple
- Résumé et conclusion

Qu'est-ce que la non-réponse?

- La non-réponse survient lorsque de l'information n'est pas disponible pour des unités échantillonnées.
- Les données peuvent être manquantes pour une ou plusieurs questions, ou encore pour l'ensemble des questions.



Types de non-réponse

- Non-réponse totale

- Non-réponse de vague

- Non-réponse partielle
 - d'item
 - de composante

Types de non-réponse

- Non-réponse totale
 - Aucune information n'est recueillie
 - L'information recueillie est insuffisante pour qu'elle soit jugée utile
- Non-réponse de vague
 - L'information sur un répondant est disponible, mais pas pour tous les cycles (présence de non-réponse totale pour un cycle donné)
ELNEJ: principalement avec la cohorte originale

Types de non-réponse

- **Non-réponse partielle d'item**
 - Certaines questions spécifiques n'ont pas été répondues
 - Certaines questions n'ont pas été demandées, mais elles auraient dû l'être
- **Non-réponse partielle de composante**
 - L'ELNEJ est segmentée en groupes de questions posées à une personne précise (ex.: questionnaire auto-administré)
 - Ce type de non-réponse survient lorsqu'au moins l'une de ces sections du questionnaire est manquante entièrement.



Quelques raisons pour la non-réponse totale ou de vague

- Raisons reliées au processus d'enquête
 - Moment choisi pour effectuer l'enquête
 - Information de mauvaise qualité sur la base de sondage
 - Erreur sur le terrain ou de la part des intervieweurs
 - Population cible et règles de collecte
- Raisons reliées aux circonstances
 - Température
 - Barrière linguistique
 - Difficultés à dépister les individus
- Raisons reliées aux répondants
 - Incapacité de participer
 - Refus de participer



Quelques raisons pour la non-réponse partielle

- Refus
- Ne sait pas
- Questions délicates (recommandations parfois faites de sauter les questions si le répondant semble mal à l'aise)
- Fardeau de réponse et le temps disponible
- Questions omises par erreur

Pourquoi la non-réponse est-elle importante à considérer pour les analystes?

- Le biais est la distorsion systématique d'un résultat statistique due à l'omission d'un facteur lors de sa dérivation.
- Mathématiquement, le biais réfère à la différence entre la valeur espérée de l'estimateur et la vraie valeur du paramètre d'intérêt.

Pourquoi la non-réponse est-elle importante à considérer pour les analystes?

- Biais dû à la non-réponse?
 - Les non-répondants ont souvent des caractéristiques différentes des répondants. Ceci peut résulter en des estimations biaisées si cette situation est ignorée.
- Conclusion: peut mener à rapporter des résultats erronés.

Non-réponse et ELNEJ

- Non-réponse totale et de vague
 - Statistique Canada répondra les répondants afin de représenter aussi les non-répondants (voir le guide d'utilisateur pour plus de détails).
 - *Vous n'avez rien à faire, seulement utiliser les poids.*
- Non-réponse partielle
 - Statistique Canada impute les variables de revenus et quelques autres variables tout dépendant du cycle (voir le guide d'utilisateur).
 - *Le choix du traitement pour les autres variables vous revient...*

Pourquoi est-ce à vous de jouer et non à Statistique Canada?

- Le traitement par Statistique Canada de tous les cas de non-réponse partielle de l'ELNEJ n'est pas la meilleure option, ni pour vous, ni pour Statistique Canada.
 - Trop de variables à traiter (environ 1 500 variables)
 - Cela retarderait la diffusion des données
- Les utilisateurs de l'ELNEJ travaillent avec un sous-ensemble de variables habituellement beaucoup plus petit

Pourquoi est-ce à vous de jouer et non à Statistique Canada?

- Le traitement par Statistique Canada de tous les cas de non-réponse partielle de l'ELNEJ n'est pas la meilleure option, ni pour vous, ni pour Statistique Canada.
 - Le traitement de la NR dépend souvent du contexte
 - Plusieurs stratégies peuvent être valides.
 - Le choix d'une stratégie dépend vraisemblablement du type d'analyse réalisée, des variables impliquées, du domaine d'analyse, des outils disponibles...
 - Connaissant le contexte de leur analyse, les utilisateurs de l'ELNEJ sont mieux placés afin de déterminer la stratégie la plus appropriée pour traiter la NR partielle.

Pourquoi est-ce à vous de jouer et non à Statistique Canada?

- Exemple #1:

- Un de vos collègues a déjà procédé à l'imputation des scores de mathématiques manquants en utilisant le score moyen de l'ensemble des répondants.
- Vous vous intéressez effectivement aux scores de mathématiques (et à leur distribution) des enfants, mais plus précisément aux liens existant entre les scores et le type d'école fréquentée.
- Les résultats pourraient possiblement être fort différents si l'imputation avait été réalisée en tenant compte du type d'école fréquentée ou en préservant la courbe de distribution des scores.

Pourquoi est-ce à vous de jouer et non à Statistique Canada?

- Exemple #2:

- Un analyste est intéressé à mesurer la corrélation entre deux variables (score de dépression de la PMR et score d'anxiété de l'enfant), chacune d'elles comportant une certaine quantité de valeurs manquantes.
- Imputer les variables séparément pourraient possiblement fausser la corrélation entre les deux variables.

Pourquoi est-ce à vous de jouer et non à Statistique Canada?

- Réponses au pourquoi:
 - Ce n'est pas possible d'envisager toutes les utilisations pertinentes et toutes les classifications d'intérêt.
 - Ce n'est pas possible d'envisager tous les groupements de variables d'intérêt.
 - Et même si ce l'était, cela signifierait d'inclure de nombreuses fois la même variable dans le fichier de diffusion, avec des valeurs imputées différentes à chaque fois... Imaginez la taille des fichiers...

Non-réponse partielle et ELNEJ

- Les données manquantes sont codées avec les valeurs suivantes:

Noms	Valeurs
Ne sait pas	7, 97, 997...
Refus	8, 98, 998...
Non déclaré	9, 99, 999...

- Note: Ces cas diffèrent de:

Noms	Valeurs
Sans objet / Enchaînement valide	6, 96, 996...

Sans objet / Enchaînement valide

- Rattachée à l'univers de la question
 - Identifie les personnes pour qui la question ne s'applique pas.
 - Ne pas se fier uniquement au dictionnaire des données, mais aussi au questionnaire, aux fichiers des données diffusées...

- Généralement, ces cas sont exclus de l'analyse dès le début.

Sans objet / Enchaînement valide

- Exemple: Questionnaire auto-administré du cycle 6
 - Question: Au cours des 12 derniers mois, combien de fois avez-vous été bénévole ou avez-vous aidé sans être payé pour le faire?
 - Univers: Répondants de 12 à 15 ans qui ont fait du bénévolat au cours des 12 derniers mois.
 - Fichier contient tous les 10 à 17 ans.
 - Sans objet / Enchaînement valide:
 - 10-11 ans, 16-17 ans
 - 12-15 ans qui n'ont pas fait de bénévolat au cours des 12 derniers mois.



Techniques pour traiter la non-réponse partielle

Quelles sont vos options?

- La question est...
 - Comment faire des inférences en présence de données manquantes pour des unités échantillonnées?
- Points à considérer:
 - Ampleur de la non-réponse
 - Type de données (continues vs catégoriques)
 - Nombre de variables d'intérêt
 - Type d'analyse
 - Contexte de l'analyse

Quelles sont vos options?

- Vous avez essentiellement 5 options:
 - a) Ignorer la non-réponse partielle (utiliser seulement les enregistrements avec une réponse complète)
 - b) Rapporter la non-réponse partielle comme une catégorie
 - c) Établir un profil des non-répondants partiels
 - d) Repondérer les enregistrements répondants pour qu'ils représentent également les non-répondants partiels
 - e) Imputer la non-réponse partielle (remplacer les valeurs manquantes par des valeurs plausibles)

Option a) Ignorer la non-réponse partielle

- Éliminer les enregistrements des non-répondants partiels.
 - Rejet de l'information possiblement utile (questions répondues)
 - Les inférences visant la population entière supposent alors que la répartition des données manquantes est complètement au hasard (MCAR: missing completely at random), c'est-à-dire que la non-réponse ne dépend d'aucune covariable, que les non-répondants sont similaires aux répondants en tout point).
 - Les inférences seront biaisées si cette hypothèse s'avère fausse (ce qui est souvent le cas).

Option a) Ignorer la non-réponse partielle

- Éliminer les enregistrements des non-répondants partiels.
 - Limite les types d'inférences possibles (modèles, moyennes, proportions, mais pas de totaux, du moins pas directement).
 - Option plus viable s'il y a peu de non-réponse.
 - Parfois la seule option lorsque la non-réponse est très faible (il est difficile d'en connaître suffisamment sur les non-répondants pour en établir un profil).

Option a) Ignorer la non-réponse partielle

- Éliminer les enregistrements des non-répondants partiels.
 - On peut faire des inférences sur la sous-population représentée par les répondants
 - Interprétation peut être nébuleuse (concept abstrait): cette sous-population de répondants n'existe probablement pas en réalité
 - Ces inférences sont-elle pertinentes???

Option b) Rappporter la non-réponse partielle comme une catégorie

- Rappporter les valeurs manquantes (non-réponse partielle) comme une catégorie valide dans les tableaux ou les modèles
 - Cela complique l'interprétation des résultats présentés dans les tableaux, mais cela donne par contre une indication de la qualité des données ainsi qu'une idée des répercussions possibles sur les valeurs présentées.

	Faible	Médium	Élevé	Non-Réponse
Exemple 1	10%	30%	50%	10%
Exemple 2	18%	20%	22%	40%



Option b) Rapporter la non-réponse partielle comme une catégorie

- S'applique seulement aux variables catégoriques (ou catégorisées)
- Peut permettre les inférences à la population entière, mais l'interprétation demeure nébuleuse (ex.: sondage sur les intentions de vote)
- Peut réduire la capacité de détecter certains effets, puisque les cas qui sont en réalité associés au même niveau d'une variable se retrouvent séparés dans deux catégories différentes.

Option c) Établir un profil des non-répondants partiels

- Il s'agit en fait de la première étape vers l'application de l'option d).
- Considérer les non-répondants partiels à un ensemble donné de questions comme une sous-population d'intérêt.
- Considérer les répondants complets au même ensemble de questions comme une autre sous-population d'intérêt.

Option c) Établir un profil des non-répondants partiels

- Déterminer comment ces 2 sous-populations diffèrent entre elles par rapport à d'autres variables clés ou reliées (variables pour lesquelles l'information est disponible pour les deux groupes):
 - Variables reliées au statut de réponse
 - Variables reliées aux valeurs de la (des) variable(s) présentant de la non-réponse partielle et que nous désirons traiter (peut être moins pratique si on a plusieurs variables à traiter)

Option c) Établir un profil des non-répondants partiels

- Inclure les résultats principaux de cette analyse de profil dans votre rapport, afin de fournir au lecteur des indications sur les sources possibles de biais.
- Peut-on faire de l'inférence à l'ensemble de la population? Pas vraiment, il faudrait encore le faire de façon conditionnelle:
 - “En l'absence de biais dû à la non-réponse, les résultats seraient les suivants. D'après nos analyses, les sources les plus probables de biais proviennent des éléments suivants:....”
- Comparativement à l'option a), le lecteur est davantage informé sur les facteurs qui pourraient expliquer certains résultats et sur la façon dont les résultats pourraient être appelés à changer.

Option c) Établir un profil des non-répondants partiels

- Il y a un exemple d'analyse où l'on a appliqué cette option dans le guide d'utilisateur du cycle 6:



Option d) Repondération

- Laisser tomber les enregistrements associés à des non-répondants partiels.
- MAIS, ajuster les poids des répondants à la hausse pour tenir compte des non-répondants partiels.
- Généralement, des groupes de repondération basés sur les variables identifiés à l'option c) sont formés.
- Il existe plusieurs techniques (classification croisée, méthode des scores de propension, algorithme CHAID...) pour former les groupes, et certaines contraintes sont habituellement imposées (nombre de groupes, pourcentage minimal de répondants dans chaque groupe...)

Option d) Repondération

- Les poids des répondants dans chaque groupe sont ajustés à la hausse par un facteur correspondant généralement à l'inverse du taux de réponse pondéré à l'intérieur du groupe.
- Exemple:

Groupe (âge par sexe par situation p/r seuil de faible revenu)	Taux de réponse pondéré	Facteur d'ajustement
Fille de 10 ans sous le seuil de faible revenu	0.80	1.25

Option d) Repondération

- Option habituellement utilisée pour traiter la non-réponse totale et la non-réponse de vague (du moins pour l'ELNEJ).
- Option intéressante pour la non-réponse de composante (il manque une section complète du questionnaire, c'est donc un peu similaire à de la non-réponse totale).
- Option un peu moins intéressante pour traiter la non-réponse partielle.
 - Rejette des questions répondues (on garde seulement les 'répondants complets' dans l'analyse)
 - Suppose que les raisons de la non-réponse partielle sont similaires.

Option d) Repondération

- Permet l'inférence à la population entière.
- Nécessite qu'une redistribution similaire des poids soit faite avec chacune des répliques bootstrap.
- Ne requiert aucun autre ajustement supplémentaire pour le calcul approprié d'une variance fondée sur le plan de sondage (« design-based variance »).

Option d) Repondération

- Note concernant les totaux de contrôle:
 - Les poids finaux de l'ELNEJ sont ajustés afin que les comptes de population issus de l'ELNEJ concordent avec les comptes officiels de Statistique Canada (voir 'Stratification a posteriori' dans le Guide de l'utilisateur)
 - Pour chaque combinaison âge-sexe-province.
 - Lorsqu'on repondère pour traiter la non-réponse partielle, les totaux de contrôle ne sont plus respectés.

Option d) Repondération

- Vous tenez absolument à ce que les totaux de contrôle soient respectés?
 - Recalculer les ajustements de poids au niveau âge-sexe-province (refaire la stratification a posteriori)
 - Encore plus de travail...
 - Ou considérer traiter la non-réponse partielle par imputation (prochaine option) plutôt que par repondération.

Option e) Imputation

- Remplacer les valeurs manquantes par des valeurs plausibles
 - Peut être vu comme de l'estimation au niveau micro
 - Il existe plusieurs approches et méthodes
- Permet les inférences à la population entière
- Habituellement utilisée pour la NR partielle d'item (ex: variables de revenus de l'ELNEJ)
 - Plus intéressant que la repondération puisqu'on garde les questions répondues (on ne fait que remplir les trous)

Option e) Imputation

- Peut artificiellement réduire la variance estimée.
- Les approches typiques d'estimation de variance requièrent des ajustements pour calculer une estimation de variance appropriée selon le plan.
- Tout dépendant de l'approche retenue (imputation multiple ou unique), les logiciels communément utilisés offrent ou n'offrent pas à l'utilisateur de faire les ajustements requis.
- Important d'inclure de l'information sur l'imputation avec les résultats (méthodes utilisées, taux d'imputation...).

Une recommandation générale

- On ne peut jamais être certain que le biais dû à la non-réponse a été complètement éliminé par un traitement de la non-réponse:
 - On n'a peut-être pas considéré toutes les variables liées à la non-réponse dans notre ajustement.
 - Certaines des variables peuvent ne même pas être dans le fichier.

Une recommandation générale

- On ne peut jamais être certain que le biais dû à la non-réponse a été complètement éliminé par un traitement de la non-réponse:
 - La non-réponse peut être liée à la variable analysée: on dit alors que les données ne sont pas manquantes au hasard (« Not Missing At Random (NMAR) »)
 - On ne peut pas alors complètement ajuster pour la non-réponse.
 - On peut partiellement ajuster si la non-réponse est aussi associée à certaines variables (ex.: variables socio-démographiques).



Une recommandation générale

- Notre but est de réduire au maximum le biais dû à la non-réponse, (car difficile à éliminer complètement).
- Tout biais qui demeure et qu'on suspecte devrait être rapporté.

Exemple 1: La population et l'échantillon

- Une école de 50 enfants, 80% d'entre eux ont 15 ans.
- Pour une enquête, un échantillon aléatoire simple (EAS) de 5 enfants est sélectionné parmi les étudiants de l'école.



Exemple 1: Les données et la question d'analyse

- Une question de l'enquête demande à chaque enfant de 15 ans si il fume régulièrement la cigarette.
- On veut estimer le nombre d'enfants de 15 ans qui fument.

Enfant	Âge	Sexe	Fumeur	Poids
1	15	F	1	10
2	15	M	0	10
3	15	F	1	10
4	16	M	6	10
5	15	F	9	10



1: signifie que l'enfant fume
0: signifie que l'enfant ne fume pas
6: Sans objet
(ex.: l'enfant n'a pas 15 ans)
9: Non-déclaré

Exemple 1: Option a) Ignorer la non-réponse partielle

- On veut estimer le nombre d'enfants de 15 ans de l'école qui fument
 - Enlever les cas Sans objet / Enchaînement valide (6)
 - Traiter la non-réponse (9)

Option a) Ignorer les non-répondants partiels

Enfant	Âge	Sexe	Fumeur	Poids
1	15	F	1	10
2	15	M	0	10
3	15	F	1	10
4	16	M	6	10
5	15	F	9	10

Pour la sous-population des 30 enfants de l'école représentée par les répondants, le nombre estimé d'enfants de 15 ans qui fument est

20

Exemple 1: Option a) Ignorer la non-réponse partielle

- Pour la sous-population des 30 enfants de l'école représentée par les répondants, le nombre estimé d'enfants de 15 ans de l'école qui fument est 20.
- Puisqu'on estime un total et qu'on ignore la non-réponse partielle, on peut seulement faire des inférences basées sur la sous-population représentée par les répondants.
- Si on estimait une proportion ou une moyenne, on pourrait justifier de faire des inférences sur la population entière si au moins une des conditions suivantes était respectée:
 - Le taux de non-réponse est très faibleOU
 - La répartition des données manquantes est complètement au hasard (Missing Completely At Random (MCAR))

Exemple 1: Option b) Rappporter la non-réponse partielle comme une catégorie

- On veut estimer le nombre d'enfants de 15 ans de l'école qui fument.
 - Enlever les cas Sans objet / Enchaînement valide (6)
 - Traiter la non-réponse partielle (9)

Option b) Rappporter la non-réponse partielle comme une catégorie

La distribution du statut de fumeur parmi les enfants de 15 ans de l'école est ...

Enfant	Âge	Sexe	Fumeur	Poids
1	15	F	1	10
2	15	M	0	10
3	15	F	1	10
4	16	M	6	10
5	15	F	9	10

Fumeur	Non Fumeur	Statut inconnu
20	10	10
50%	25%	25%

Exemple 1: Option c) Établir un profil des non-répondants partiels

- Option c) Établir un profil des non-répondants partiels

Ici, il n'y a qu'un seul non-répondant et tout ce qu'on en connaît, c'est qu'il s'agit d'une fille. Mais, si cette relation était observée à une plus grande échelle, l'analyste pourrait rapporter que 'la non-réponse a été évaluée, et il apparaît qu'elle soit principalement liée au sexe de l'enfant. Les résultats rapportés ne tiennent pas compte de cette différence et pourraient par conséquent inclure un certain biais dû à cette caractéristique de la non-réponse'.

Enfant	Âge	Sexe	Fumeur	Poids
1	15	F	1	10
2	15	M	0	10
3	15	F	1	10
4	16	M	6	10
5	15	F	9	10

Exemple 1: Option d) Repondération

- On veut estimer le nombre d'enfants de 15 ans de l'école qui fument.
 - Enlever les cas Sans objet / Enchaînement valide (6)
 - Traiter la non-réponse partielle (9)

Enfant	Âge	Sexe	Fumeur	Poids
1	15	F	1	10 *30/20
2	15	M	0	10 *10/10
3	15	F	1	10 *30/20
4	16	M	6	10
5	15	F	9	10

Option d): Repondérer les poids des répondants en utilisant les groupes déterminés avec l'option c) (ici, des groupes basés sur le sexe)

Après repondération afin de compenser pour la non-réponse partielle, le nombre estimé d'enfants de 15 ans de l'école qui fument est de

30

Exemple 1: Option e) Imputation

- On veut estimer le nombre d'enfants de 15 ans de l'école qui fument.
 - Enlever* les Sans objet / Enchaînement valide (6)
 - Traiter la non-réponse partielle (9)

Option e) Imputer les valeurs manquantes

- On va imputer selon le sexe (dans ce cas-ci, cela signifie qu'on impute un 1 pour les filles, et un 0 pour les garçons)

Child	Age	Gender	Smoker	Weight
1	15	F	1	10
2	15	M	0	10
3	15	F	1	10
4	16	M	6	10
5	15	F	9 1	10

Après imputation des non-répondants, le nombre estimé d'enfants de 15 ans de l'école qui fument est de

30.

En résumé

- Évaluer l'impact de la non-réponse partielle sur votre analyse.
 - Ampleur de la non-réponse partielle
 - Variables reliées au statut de réponse
- Prendre ensuite les mesures appropriées.
 - Très peu de non-réponse, pas de variables reliées à la non-réponse partielle ou encore des inférences basées sur la sous-population représentée par les répondants
 - a) Vous pouvez considérer de simplement ignorer les non-répondants partiels

En résumé

- Autrement, vos options pour le traitement de la non-réponse partielle sont...
 - b) Rapporter la non-réponse partielle comme une catégorie valide
 - Simple et rapide
 - Peut être difficile à interpréter
 - Ne s'applique pas à toutes les types d'analyse
 - c) Établir un profil de la non-réponse partielle
 - Plus de travail, mais plus informatif
 - Moins de travail que d) ou e), mais ne permet pas vraiment de faire des inférences à l'ensemble de la population d'intérêt.

En résumé

- vos options pour le traitement de la non-réponse partielle sont...
 - d) Repondération
 - Relativement simple, mais plus de travail
 - Met de côté des items répondus
 - Taux de contrôle ne sont plus respectés
 - e) Imputation
 - Tire avantage des items répondus
 - Plus complexe et requiert plus de travail
 - Risque de générer des incohérences au niveau de l'enregistrement
 - Peut fausser l'estimation de variance ou encore requérir des ajustements additionnels

En résumé

- Annexer à votre rapport d'analyse une section qui discute du traitement de la non-réponse partielle:
 - Stratégie de traitement choisie avec justifications (ignorer la non-réponse partielle EST une stratégie)
 - Fait partie du processus scientifique (autres chercheurs peuvent reproduire vos résultats)

En résumé

- Annexer à votre rapport d'analyse une section qui discute du traitement de la non-réponse partielle:
 - Soyez prêt à fournir à un lecteur/jury l'information pertinente:
 - Taux de réponse et taux d'imputation
 - Variables considérées pour le biais potentiel de non-réponse
 - Méthodes utilisées et contraintes pour la création des groupes de repondération ou d'imputation

En résumé

- Est-ce plus de travail de traiter la non-réponse partielle que de l'ignorer?
 - Oui, évidemment... mais les résultats de votre analyse sont plus *fiables* (le biais dû à la non-réponse est considéré et réduit) et plus *pratiques* (les conclusions s'appliquent à l'ensemble de la population d'intérêt)
- Rappelez-vous qu'en ignorant la non-réponse partielle, vos conclusions ne s'appliquent qu'à la sous-population représentée par les répondants (concept abstrait), à moins que la non-réponse partielle soit complètement répartie au hasard.



Regroupement de cohortes au sein de l'ELNEJ

Aperçu de la section

- Pourquoi envisager un regroupement de cohortes?
- Quand est-ce faisable d'envisager un regroupement?
- L'approche la plus commune
 - Approche par 'Pooling'
- Quelques exemples

Pourquoi envisager un regroupement de cohortes?

- La taille d'échantillon fournie par une cohorte individuelle est insuffisante.
 - Estimations ne peuvent pas être publiées, selon les critères de C.V. requis par l'enquête
 - On veut ajouter de la puissance statistique afin de possiblement mieux détecter certains effets.
- Impression que les quantités d'intérêt sont relativement stables dans le temps.

Pourquoi envisager un regroupement de cohortes?

- Exemple:

- Si l'objectif est de décrire certains aspects de la vie des enfants âgés d'un an et vivant au sein de ménages sous le seuil de faibles revenus (disons la proportion d'entre eux avec un faible score de développement moteur et social), alors il est probable que la taille d'échantillon disponible à partir de n'importe laquelle des cohortes individuelles sera trop petite.

Quand est-ce faisable d'envisager un regroupement?

- Il existe un lien entre la population cible du chercheur et les populations d'enquête regroupées:
 - Si la population d'intérêt est l'ensemble des populations d'enquête regroupées.
 - Si chacune des populations d'enquête peut être considérée comme représentant la population d'intérêt
 - Si aucune des populations enquêtées ne sont représentatives, mais qu'une mesure sommaire à partir des populations enquêtées est informative.

Quand est-ce faisable d'envisager un regroupement?

- Si vous croyez que le modèle que vous suggérez aurait pu générer les données de chacune des populations enquêtées
 - Habituellement, le modèle inclura des paramètres spécifiques à chacune des populations finies (du moins initialement).

L'approche la plus commune

- Les étapes de l'approche par 'pooling':
 - 1) Combiner les données des différentes cohortes au sein d'un seul fichier.
 - 2) Créer un poids qui est approprié pour les données combinées, pour la population cible et pour les quantités d'intérêt.
 - 3) Créer de nouveaux poids bootstrap selon la stratégie de pondération retenue à l'étape 2).

L'approche la plus commune

- Les étapes de l'approche par 'pooling':
 - Vérifier les hypothèses que les quantités d'intérêt sont les mêmes pour les différentes cohortes regroupées.
 - Aller de l'avant avec l'estimation et l'inférence à partir des données regroupées en utilisant les techniques qui seraient appropriées pour des données provenant d'un seul échantillon.

L'approche la plus commune

- L'approche par 'pooling':

Remarques:

Le regroupement d'échantillons qui ne sont pas indépendants ajoute des complications additionnelles.

- Si le même enfant (PERSRUK) fait partie de deux échantillons regroupés, alors les échantillons ne sont définitivement pas indépendants.

L'approche la plus commune

- L'approche par 'pooling':

Remarques:

- Il y a de multiples façons de créer une variable de poids et des poids bootstrap dans l'approche par 'pooling'. Dans plusieurs cas, simplement regrouper les poids des cohortes respectives, sans ajustement additionnel, est un bon choix.
 - Parmi les éléments à considérer:
 - Les tailles d'échantillons respectives de chacune des cohortes regroupées
 - Les tailles respectives des populations enquêtées
 - Est-ce que tous les enregistrements de chacune des cohortes sont utilisés ou est-ce que certains cas sont mis de côté?

Quelques exemples

■ Premier exemple:

- Le point d'intérêt de l'analyse sera le sommeil des bébés. Nous limiterons l'analyse aux enfants âgés de moins d'un an pour minimiser l'impact des erreurs de rappel.
- Les questions sur le sommeil sont relativement nouvelles. La plupart n'ont été demandées que depuis le cycle 4, et certaines seulement depuis le cycle 5.
- On devrait consulter la littérature à la recherche de toute forme d'indications d'un changement dans les recommandations faites aux parents en ce qui concerne les positions de sommeil recommandées pour les bébés entre 2002 et 2004.

Quelques exemples

- Premier exemple:

On va définir la population cible comme la population des enfants de moins d'un an nés entre 2002 et 2004.

On pourrait se construire un plus gros échantillon en regroupant les 0 ans du cycle 5 et ceux du cycle 6.

Les tailles d'échantillon de chaque cycle sont très similaires, de même que les tailles des populations enquêtées. Tous les bébés de moins d'un an des 2 cohortes feront partie de l'analyse.

Quelques exemples

- Premier exemple:

Donc, ici, on pourrait simplement utiliser les poids transversaux et les poids bootstrap transversaux associés aux enfants du cycle 5 et 6 qui sont âgés de moins d'un an (deux échantillons indépendants).

Même si on ne retrouve aucune indication de changement dans les recommandations transmises aux parents, il est recommandé de tout de même effectuer des vérifications que les quantités d'intérêt mesurent le même concept dans les deux cohortes regroupées.

Quelques exemples

- Deuxième exemple:

Le point d'intérêt de l'analyse sera les bébés nourris au sein pendant plus de 6 mois, et les moments où certaines étapes de développement sont atteintes. On voudra comparer les bébés nourris au sein pour plus de 6 mois aux autres bébés.

Ici, la population cible sera les enfants qui ont célébré leur 1^{er} anniversaire entre 2000 et 2002. On utilisera les informations fournies par les enfants âgés d'un an entre 2000 et 2002 pour déterminer le groupe d'appartenance (nourri au sein plus de 6 mois ou non), mais on utilisera les données au moment où les enfants sont âgés de 3 ans, afin qu'ils aient le temps d'atteindre diverses étapes de développement importantes. Il est jugé que les recommandations sur les pratiques d'allaitement des bébés étaient relativement stables au cours de la période 1999-2001.

Quelques exemples

- Deuxième exemple:

On va par conséquent combiner les 3 ans du cycle 5 et du cycle 6.

Les tailles d'échantillons sont relativement homogènes, les tailles des populations cibles aussi. Tous les enfants longitudinaux de 3 ans des deux cohortes feront partie de l'analyse.

On pourrait donc simplement utiliser les poids longitudinaux et les poids bootstrap longitudinaux associés à chaque enfant au moment où l'enfant a 3 ans.

Quelques exemples

- Deuxième exemple:

Supposons que la taille d'échantillon résultante était encore insuffisante. Comment pourrait-on obtenir plus d'enfants?

- Prendre des enfants plus jeunes (2 ans) pourrait mener à des problèmes de censure (certains enfants pourraient ne pas encore avoir atteint les étapes de développement d'intérêt.).

Quelques exemples

- Deuxième exemple:

Supposons que la taille d'échantillon résultante était encore insuffisante. Comment pourrait-on obtenir plus d'enfants?

- Prendre les 3 ans du cycle 4 ou encore les 4 ans du cycle 5 pourrait être une option, mais les différences dans les tailles respectives des échantillons auraient probablement nécessité la création de nouveaux poids. Il aurait aussi fallu vérifier la stabilité de l'environnement pour la période précédent 1999 et modifier la définition de la population cible de manière appropriée.



Statistique
Canada

Statistics
Canada

Canada



Statistique Canada
www.statcan.gc.ca

Un exemple complet

(conçu par un méthodologiste...)

Un exemple complet

- Nous avons choisi le sujet de notre prochain article: l'usage des ordinateurs par les adolescents: qui sont les utilisateurs assidus?"
- Nous avons fait une revue de littérature et avons identifié une liste de caractéristiques et de facteurs qui seraient apparemment liés à l'assiduité de l'utilisation de l'ordinateur par les adolescents.
- Nous souhaitons utiliser des données canadiennes pour mettre ces théories/opinions à l'épreuve.

Un exemple complet

- Nous avons identifié l'ELNEJ comme une source potentielle de données canadiennes pour répondre à cette question.
- Nous avons lu la documentation de l'enquête. Nous sommes prêts à aller de l'avant.
- Nous allons maintenant réaliser ensemble certaines des étapes du processus d'analyse.

Un exemple complet

- Voici quelques-uns des objectifs que nous avons en tête pour l'article:
 - Fournir au lecteur un portrait actuel de l'assiduité de l'utilisation de l'ordinateur par les adolescents.
 - Comparer l'usage courant à l'usage passé.
 - Identifier quelques facteurs liés à une utilisation assidue de l'ordinateur par les adolescents.

Un exemple complet

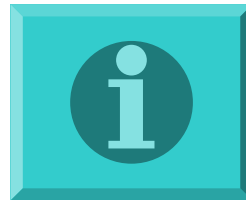
- Au sein du questionnaire auto-administré du cycle 6, une composante demandée des enfants âgés entre 10 et 17 ans, il y a une section sur les activités.
- Une des questions de cette section est la suivante:

FATCeQ21

- En moyenne, combien d'heures par jour passes-tu sur l'ordinateur (à faire des travaux, à jouer à des jeux, à envoyer et à recevoir des messages, à bavarder ou à naviguer sur l'Internet)?

Un exemple complet

- Veuillez noter que le nom de la variable inclut une lettre minuscule 'e' comme 5^e caractère.
 - Cette lettre minuscule réfère au cycle de l'ELNEJ ('e'=5, 'f'=6, 'd'=4, etc.) au cours duquel la variable est apparue pour la 1^{ère} fois ou au cours duquel elle a subi une modification par rapport à une question posée précédemment.



Un exemple complet

- Dans ce cas-ci, en consultant les dictionnaires de données des cycles 4 et 5, on constate qu'il s'agit d'une nouvelle question, utilisée seulement depuis le cycle 5.
- Donc, la première contrainte imposée par les données est qu'il ne sera pas vraiment possible de retourner très loin dans le passé avec les données de l'ELNEJ (du moins pas plus loin qu'en 2002).

Un exemple complet

- Voici les réponses à la question telles que rapportées dans le dictionnaire des données du fichier pour les 10-17 ans du cycle 6:

Valeur	Étiquette	Fréq	Pond.
01	Je n'utilise pas d'ordinateur	54	18,782
02	Moins d'une heure par jour	2,345	722,791
03	1 ou 2 heures par jour	2,177	784,766
04	3 ou 4 heures par jour	764	291,690
05	5 ou 6 heures par jour	180	72,835
06	7 heures ou plus par jour	78	25,978
96	Enchaînement valide	1,331	611,323
99	Non déclaré	1,267	512,195
		=====	=====
		8,196	3,040,360

Univers: Les répondants de 10 à 15 ans

Un exemple complet

- Veuillez noter l'énoncé concernant l'univers de la question, généralement situé sous le tableau des valeurs associées aux réponses à une variable donnée dans le dictionnaire des données. Dans le cas de la variables **FATCeQ21**, l'énoncé affirme que cette question s'adressait aux répondants âgés entre 10 et 15 ans.
- Donc, une 2^e contrainte imposée par les données concerne la définition du terme adolescent. Nous devons nous limiter au mieux à ne considérer que les enfants âgés entre 10 et 15 ans.

Un exemple complet

- À la toute fin du dictionnaire des données du fichier pour les 10-17 ans du Cycle 6, juste avant l'index, il y a une section qui présente les différents poids disponibles pour ces données. Davantage d'information à propos des poids se retrouvent dans le guide d'utilisateur.
- Dans notre cas, il n'y a que 2 types de poids disponibles:
 - FWTCW01L: Poids longitudinal
 - FWTCWd1L: Poids entonnoir longitudinal

Un exemple complet

- Puisqu'un poids transversal n'est pas disponible, mais plutôt seulement des poids longitudinaux, cela signifie qu'il ne sera pas possible de faire un portrait de l'assiduité de l'utilisation de l'ordinateur par les adolescents de 10 à 15 ans en 2004.
- Un poids transversal n'est plus fourni pour la cohorte originale puisqu'il a été jugé que la population des enfants a trop changé depuis 1994 pour que la cohorte originale puisse être considérée représentative des enfants du même âge en 2004.



Un exemple complet

- Par conséquent, toutes les inférences devront être faites par rapport à la population longitudinale.
- Dans notre cas, cela signifie que les conclusions se rapporteront à la population des enfants âgés entre 0 et 5 ans en 1994, mais tels qu'ils sont lorsqu'ils sont âgés entre 10 et 15 ans en 2004.

Un exemple complet

- Retournons maintenant aux réponses à la question, telles que rapportées par le dictionnaire de données du fichier pour les 10-17 ans du cycle 6:

Valeur	Étiquette	Fréq	Pond.
01	Je n'utilise pas d'ordinateur	54	18,782
02	Moins d'une heure par jour	2,345	722,791
03	1 ou 2 heures par jour	2,177	784,766
04	3 ou 4 heures par jour	764	291,690
05	5 ou 6 heures par jour	180	72,835
06	7 heures ou plus par jour	78	25,978
96	Enchaînement valide	1,331	611,323
99	Non déclaré	1,267	512,195
		=====	=====
		8,196	3,040,360

Univers: Les répondants de 10 à 15 ans

Un exemple complet

- On peut noter certains éléments:
 - Puisque le fichier contient de l'information pour les 10-17 ans, et que la question s'adresse uniquement aux 10-15 ans, il y a un certain nombre (1 331) de réponses avec une valeur 'Enchaînement valide (96)'. Ce nombre devrait correspondre au nombre d'enfants âgés de 16 à 17 ans et ces enregistrements devraient être retirés du fichier d'analyse.

Un exemple complet

- On peut noter certains éléments:
 - Si on disposait de plus de temps pour observer les données, on pourrait constater que le nombre (1 585) d'enfants de 16 et 17 ans ne correspond pas au compte (1 331) 'd'enchaînement valide' à cette question. En fait, tous les comptes 'd'enchaînement valide' sont erronés dans le fichier des 10-17 du Cycle 6. Ceci est dû à une erreur de traitement (présente uniquement au Cycle 6). On a assigné une valeur 'Non déclaré' aux cas pour lesquels il n'y avait aucune donnée pour le questionnaire auto-administré, sans tenir compte de l'univers de la question.

Un exemple complet

- On peut noter certains éléments:
 - Par conséquent, lors de l'utilisation du fichier des 10-17 ans du cycle 6, les 'Enchaînement valide' devrait être retiré en se basant sur l'univers de la question, et non à partir des valeurs prises par la variable d'intérêt.
 - Ou encore, utiliser le programme Errata_20080814.sas pour corriger les valeurs et procéder ensuite comme à l'habitude.

Un exemple complet

- Un élément additionnel:
 - On pourrait vouloir également retirer les enfants qui sont longitudinalement dans le champ de l'enquête, mais pour qui aucune donnée n'a pu être recueillie. Ceci inclut les enfants décédés ou ne résidant plus au Canada.
 - Ces enfants peuvent être identifiés à partir de la variable FLWTCD sur le fichier LONG.
 - Dans le cas d'une analyse réalisée à partir du fichier pour les 10-17 ans, ces enregistrements ont déjà été retirés du fichier, donc il n'est pas nécessaire de tenir compte de cet élément additionnel.

Un exemple complet

- Lorsque tous les 'Enchaînement valide' ont été retirés, les réponses sont distribuées de la façon suivante:

Valeur	Étiquette	Fréq	Pond.
01	Je n'utilise pas d'ordinateur	54	18,782
02	Moins d'une heure par jour	2,345	722,791
03	1 ou 2 heures par jour	2,177	784,766
04	3 ou 4 heures par jour	764	291,690
05	5 ou 6 heures par jour	180	72,835
06	7 heures ou plus par jour	78	25,978
99	Non déclaré	1,013	373,322
		=====	=====
		6,611	2,290,163

Un exemple complet

- On peut noter certains éléments:
 - Il y a une certaine quantité de non-réponse partielle (Non déclaré (99)). Elle représente environ 15% des réponses.
 - Nous discuterons de diverses façons de traiter cette situation sous peu.
 - Mais tout d'abord, jetons un coup d'oeil à la distribution non pondérée et pondérée des réponses.



Un exemple complet

Réponses à FATCeQ21	Non pondérée		Pondérée	
	Fréq.	Pour. (%)	Fréq.	Pour. (%)
Je n'utilise pas d'ordinateur	54	0,82	18 782,5	0,82
Moins d'une heure par jour	2 345	35,47	722 790,7	31,56
1 ou 2 heures par jour	2 177	32,93	784 765,5	34,27
3 ou 4 heures par jour	764	11,56	291 689,6	12,74
5 ou 6 heures par jour	180	2,72	72 834,7	3,18
7 heures ou plus par jour	78	1,18	25 978,3	1,13
Non déclaré	1 013	15,32	373 321,8	16,30

Un exemple complet

- Comparons les deux distributions:
 - La pondération modifie la distribution des fréquences. Ceci a pour but de corriger pour la distorsion inhérente à l'échantillon.
 - Notons entre autres que la catégorie modale est maintenant '1 ou 2 heures par jour', et non 'Moins d'une heure par jour'.

Un exemple complet

- Comparons les deux distributions:
 - Notons également que la pondération a eu un certain impact sur la proportion de 'Non déclaré'. Ceci est un signe que la non-réponse est au moins partiellement reliée à certains des aspects du plan de sondage.
 - Parmi les variables à considérer pour la stratégie de traitement de la non-réponse, il serait avisé d'inclure des variables du plan de sondage qui sont également liées à la non-réponse (et/ou aux valeurs de la variable d'intérêt).

Un exemple complet

- Comparons les deux distributions:
 - Qu'est-ce qui pourrait en partie expliquer les différences entre la distribution des réponses au sein de l'échantillon (distribution non pondérée) et celle au sein de la population (distribution pondérée)?
 - Afin de répondre à cette question, il est utile d'explorer les diverses distorsions présentes au sein de l'échantillon et d'essayer d'en trouver une (ou quelques-unes) qui est (sont) reliée(s) à la variable d'intérêt.

Un exemple complet

- Comparons les deux distributions:
 - Les distorsions les plus communes sont:
 - Habituellement, les provinces de l'Atlantique sont sur représentées, de même que le Manitoba et la Saskatchewan. Le Québec et l'Ontario sont sous représentées.
 - Les régions rurales et les petites régions urbaines sont sur représentées. Les RMRs/ARs sont sur représentées.
 - Au sein de la cohorte originale, les enfants les plus âgés sont sous représentés et les enfants les plus jeunes sont sur représentés.

Un exemple complet

- Comparons les deux distributions:
 - Dans le cadre de cette analyse, l'âge étant fortement lié à l'utilisation de l'ordinateur chez les enfants, il est probable que ce soit la source principale de distorsion.
 - Le fait de résider en milieu rural ou urbain est assurément une autre source de distorsion.

Un exemple complet

- On va mettre de côté pour quelques instants le 1^{er} objectif, et se concentrer davantage sur le 3^e, c'est-à-dire d'identifier les facteurs liés à l'assiduité de l'utilisation de l'ordinateur par les adolescents.
 - Nous avons déjà discuté du fait que nous devons nous limiter aux enfants de 10 à 15 ans.
 - Par conséquent, nous devrions rechercher des facteurs (ou des variables liées) qui sont disponibles à un même moment pour tous les enfants.

Un exemple complet

- Il n'est pas impossible de considérer des facteurs qui ne sont pas disponibles à un même moment pour tous les enfants.
- La nature longitudinale de l'enquête fait en sorte qu'il est possible de retourner dans le passé pour obtenir de l'information à un moment différent pour une partie de l'échantillon, mais ce genre d'information présente souvent quelques inconvénients.

Il faudra possiblement considérer des facteurs mesurés de la même façon dans le temps et dont l'effet demeure constant, ou encore construire des modèles plus complexes.

Un exemple complet

- Pour les fins de modélisation, une des premières choses à faire sera de définir ce qu'est un 'utilisateur assidu'.
- Bien sûr, cette définition devra s'inspirer des données disponibles, mais aussi des mesures présentes dans la littérature.
- Les résultats pourraient être liés à la façon dont les concepts ont été définis. Il est donc important de présenter ces définitions clairement et de pouvoir les justifier.

Un exemple complet

- Dans notre cas, nous utiliserons les concepts suivants:
 - Un 'utilisateur assidu' sera défini comme un adolescent utilisant l'ordinateur 3 heures ou plus par jour en moyenne.
 - Un 'utilisateur non assidu' sera un adolescent utilisant l'ordinateur pendant moins de 3 heures par jour en moyenne.
 - Nous mettrons de côté les non-utilisateurs, puisqu'il est fort probable que ces adolescents aient un profil différent des autres adolescents. Lorsque nous aurons terminé de traiter les données pour la non-réponse, nous traiterons ces cas comme des 'Enchaînement valide' et ils seront retirés de l'analyse.

Un exemple complet

- Un travail similaire devra être réalisé avec chacun des facteurs potentiels identifiés par la revue de littérature. Ici, la disponibilité des données, mais aussi l'adéquation du modèle feront partie des éléments à considérer dans la détermination de la façon d'utiliser chacun des facteurs.
- De plus, certains des facteurs potentiels identifiés pourraient ne pas être directement disponibles dans l'enquête. Une option est alors d'essayer d'identifier parmi les variables disponibles une variable qui pourrait servir de 'proxy' (variable de procuration) pour ce facteur non disponible.

Un exemple complet

- Supposons qu'après une revue de littérature exhaustive (incluant Lafortune (2007), Lafortune (2008) et Lafortune (2009)), les facteurs suivants ont été identifiés comme étant potentiellement liés à l'assiduité avec laquelle les adolescents utilisent l'ordinateur:
 - Âge
 - Sexe
 - Disponibilité d'un ordinateur à la maison et/ou à l'école
 - Quantité de temps libre dont les parents disposent pour être avec leurs enfants
 - Surveillance parentale

Un exemple complet

- Parmi les données du cycle 6 de l'ELNEJ, nous disposons de variables pour la plupart de ces facteurs potentiels:
 - **FMMCQ01** (Âge de l'enfant)
 - **FMMCQ02** (Sexe de l'enfant)
 - **FATCeQ22** (Y a-t-il un ordinateur à la maison?)
 - **FPMCCS3** (Score de surveillance parentale)

- Nous aimerions identifier des variables qui pourraient servir de 'proxy' pour les concepts de '**Disponibilité d'un ordinateur à l'école**' et '**Quantité de temps libre dont les parents disposent pour être avec leurs enfants**'.

Un exemple complet

- Les variables de procuration (proxy) ne seront jamais aussi appropriées que les variables qu'elles visent à représenter, et il est important de garder ce point en tête en analysant les résultats.
- Nous utiliserons les variables proxy suivantes pour '**Disponibilité d'un ordinateur à l'école**' et '**Quantité de temps libre dont les parents disposent pour être avec leurs enfants**':
 - **FEDCbQ0** (Quel type d'école l'enfant fréquente-t-il?)
 - **FLFHD49B** (Situation de travail/études de la PMR et de son conjoint)

Un exemple complet

Nom de la variable	Type de variable
Âge de l'enfant	Continue(?), 10-15
Sexe de l'enfant	Catégorique, 2 niveaux
Ordinateur à la maison	Catégorique, 2 niveaux
Surveillance parentale	Continue, 0-20
Type d'école	Catégorique, 7 niveaux
Situation travail/études des parents	Catégorique, 6 niveaux

Plus valeurs manquantes

Un exemple complet

- Une décision sur la façon d'utiliser chacune de ces variables sera requise:
 - Est-ce qu'une variable continue devrait être utilisée telle qu'elle, ou devrait-elle être catégorisée?
 - Est-ce que certaines catégories devraient être jumelées?
 - Comment déterminer les seuils ou les points de séparation?
- Nous ne discuterons pas en détails de ces éléments ici, mais ces éléments devraient vraisemblablement faire partie de toutes les analyses impliquant une modélisation.

Un exemple complet

- Dans notre cas, nous regrouperons certaines catégories à trop faible fréquence de la variable **FEDCbQ0**.

Type d'école fréquentée?

FEDCbQ0	Fréquence	Pourcentage	Fréquence Cumulative	Pourcentage Cumulatif
---------	-----------	-------------	-------------------------	--------------------------

Cette information a dû être supprimée...

Un exemple complet

- La surveillance parentale est souvent liée à certains dénouements chez les enfants. Il y a une corrélation entre le manque de supervision et certains dénouements négatifs.
- Par conséquent, nous regrouperons les enfants les plus à risque en dichotomisant les scores de l'échelle de surveillance parentale (**FPMCCS3**) et en utilisant le 1^{er} décile (c'est-à-dire les 10% les moins surveillés) comme point de séparation.

Un exemple complet

- Lorsque le groupe de variables constituant le noyau a été identifié, l'étendue de la non-réponse peut être évaluée et une stratégie de traitement de la non-réponse peut être choisie.
- Il y a 5 328 enregistrements avec une réponse complète. Certains de ces enregistrements (les non-utilisateurs) seront mis de côté lorsque les non-répondants auront été traités.
- Il y a environ 1300 cas de non-réponse partielle.

Un exemple complet

- La non-réponse partielle est répartie de la façon suivante:
 - 1 013 enregistrements avec aucune information sur l'utilisation de l'ordinateur (et pour la plupart d'entre eux, peu d'autres informations; en fait, pour 618 enregistrements, il n'y a aucune information provenant de l'enfant).
 - Pour 211 autres enregistrements, c'est le score de surveillance parentale qui est manquant (mais l'information sur l'utilisation de l'ordinateur est disponible).

Un exemple complet

- La non-réponse partielle est répartie de la façon suivante:
 - 53 enregistrements ont de l'information manquantes en ce qui concerne le statut de travail/études des parents, mais de l'information pour les autres variables.
 - Certaines informations ont dû être supprimées ici...

Un exemple complet

- La non-réponse affecte toutes les variables d'intérêt pour notre sujet de recherche

Variable	Enregistrements avec une valeur manquante
Utilisation de l'ordinateur	1 013
Surveillance parentale	+ 211
Situation Travail/études	+ 53
Certaines informations ont été supprimées...	
Total	

Un exemple complet

- Option a) Ignorer la non-réponse partielle
 - Ici, on considère seulement les répondants complets à l'ensemble des questions retenues, sans aucun ajustement (5 278 enfants).
 - Les autres enregistrements sont simplement mis de côté.

Un exemple complet

- Option a) Ignorer la non-réponse partielle

Combien d'heures par jour passes-tu sur l'ordinateur?

	Fréquence	Pourcentage	Fréquence cumulative	Pourcentage cumulatif
Moins d'une heure par jour	684 216,4	37,67%	684 216,4	37,67%
1 ou 2 heures par jour	753 541,7	41,49%	1 437 758	79,16%
3 ou 4 heures par jour	283 800,3	15,62%	1 721 558	94,78%
5 ou 6 heures par jour	69 954,27	3,85%	1 791 513	98,63%
7 heures ou plus par jour	24 844,51	1,37%	1 816 357	100,00%

On ne peut pas faire d'inférence sur les totaux directement

Valide seulement si les données sont manquantes complètement au hasard

Un exemple complet

Avec variance telle que mesurée avec des poids normalisés et SAS PROC LOGISTIC

- Option a) Ignorer la non-réponse partielle

Effet		Estimé	Limite de l'intervalle de confiance	
FMMCQ01	10 vs 15	0,209	0,157	0,278
FMMCQ01	11 vs 15	0,415	0,328	0,524
FMMCQ01	12 vs 15	0,545	0,437	0,679
FMMCQ01	13 vs 15	0,554	0,447	0,687
FMMCQ01	14 vs 15	0,880	0,718	1,077
Ordi_maison	0 vs 1	0,139	0,062	0,312
FLFHD49B	1 vs 6	1,033	0,694	1,539
FLFHD49B	2 vs 6	0,982	0,557	1,730
FLFHD49B	3 vs 6	0,984	0,631	1,532
FLFHD49B	4 vs 6	0,756	0,335	1,707
FLFHD49B	5 vs 6	1,500	0,988	2,277
Surv_parent	0 vs 1	1,395	1,118	1,742

Valide seulement si les données sont manquantes complètement au hasard

Un exemple complet

Avec variance telle que mesurée avec des poids bootstrap et SUDAAN PROC RLOGIST

- Option a) Ignorer la non-réponse partielle

Effet		Estimé	Limite de l'intervalle de confiance	
FMMCQ01	10 vs 15	0,209	0,14	0,31
FMMCQ01	11 vs 15	0,415	0,29	0,60
FMMCQ01	12 vs 15	0,545	0,37	0,81
FMMCQ01	13 vs 15	0,554	0,38	0,82
FMMCQ01	14 vs 15	0,880	0,61	1,27
Ordi_maison	0 vs 1	0,139	0,05	0,37
FLFHD49B	1 vs 6	1,033	0,50	2,13
FLFHD49B	2 vs 6	0,982	0,38	2,53
FLFHD49B	3 vs 6	0,984	0,43	2,24
FLFHD49B	4 vs 6	0,756	0,19	2,95
FLFHD49B	5 vs 6	1,500	0,71	3,17
Surv_parent	0 vs 1	1,395	0,98	1,99

Valide seulement si les données sont manquantes complètement au hasard

Un exemple complet

- Option a) Résultats avec les poids des sondage, les poids normalisés et les poids bootstrap

Effets principaux	Valeurs-p avec les poids de sondage et SAS proc logistic	Valeurs-p avec les poids normalisés et SAS proc logistic	Valeurs-p avec les poids bootstrap et SUDAAN
Âge	<0.0001	<0.0001	0.0000
Sexe	<0.0001	0.1801	0.4291
Type d'école	<0.0001	0.0699	0.5495
Ordinateur à la maison	<0.0001	<0.0001	0.0000
Travail/études	<0.0001	0.0014	0.1680
Surveillance parentale	<0.0001	<0.0001	0.0127

Un exemple complet

- Option b) Rapporter la non-réponse partielle comme une catégorie valide
 - Ici, nous pourrions utiliser l'ensemble des enfants admissibles (c'est-à-dire les enfants de 10 à 15 ans, pour un total de 6 611 enfants).

Un exemple complet

- Option b) Rappporter la non-réponse partielle comme une catégorie valide

Combien d'heures par jour passes-tu sur l'ordinateur?

FATCeQ21	Fréquence	Pourcentage	Fréquence cumulative	Pourcentage cumulatif
Je n'utilise pas un ordinateur	18 782,46	0,82%	18 782,46	0,82%
Moins d'une heure par jour	722 790,7	31,56%	741 573,2	32,38%
1 ou 2 heures par jour	784 765,5	34,27%	1 526 339	66,65%
3 ou 4 heures par jour	291 689,6	12,74%	1 818 028	79,38%
5 ou 6 heures par jour	72 834,72	3,18%	1 890 863	82,56%
7 heures ou plus par jour	25 978,27	1,13%	1 916 841	83,70%
Non déclaré	373 321,8	16,30%	2 290 163	100,00%

On doit inclure cette catégorie car les 'Non déclaré' pourraient s'y retrouver

Permet au lecteur d'envisager des scénarios

Difficile d'interpréter les pourcentages

Un exemple complet

- Option b) Rapporter la non-réponse partielle comme une catégorie valide

- On ne peut pas vraiment réaliser la régression logistique telle qu'exécutée en a).

Pour la non-réponse aux variables explicatrices, on pourrait simplement ajouter une catégorie additionnel pour chaque variable (lorsqu'il y a suffisamment de cas).

Mais pour la non-réponse présente au sein de la variable dépendante, il faudrait plutôt recourir à un modèle multi-logistique.

Un exemple complet

- Option c) Établir un profil de la non-réponse partielle
 - Ici, on cherche à en connaître davantage sur les 1 300 non-répondants partiels et possiblement d'identifier certaines de leurs caractéristiques.
 - On devrait potentiellement tenter d'identifier des similitudes/différences entre les différentes classes de non-répondants partiels (aucune donnée transmise par les enfants, pas de réponse à la question d'utilisation de l'ordinateur, autres formes de non-réponse partielle) pour voir si la même stratégie devrait être appliquée à tous.

Un exemple complet

- Option c) Établir un profil de la non-réponse partielle

Tableau par âge

Type de non-réponse	10 ans	11 ans	12 ans	13 ans	14 ans	15 ans
Aucune donnée transmise par l'enfant	16,45%	16,49%	17,04%	14,79%	18,01%	17,22%
Utilisation de l'ordinateur manquante	21,41%	21,13%	14,94%	16,42%	15,85%	10,25%
Autres	26,09%	18,28%	22,19%	13,00%	8,41%	12,02%
Répondants	14,86%	15,81%	16,64%	17,51%	17,60%	17,58%
Total	15,78%	16,25%	16,79%	16,99%	17,23%	16,96%

Un exemple complet

- Option c) Établir un profil de la non-réponse partielle

Tableau par catégories de revenus

Type de non-réponse	Sous le seuil	Sous 1,5 fois le seuil	Au-dessus de 1,5 fois le seuil
Aucune donnée transmise par l'enfant	15,75%	18,86%	65,39%
Utilisation de l'ordinateur manquante	32,65%	19,64%	47,71%
Autres	16,11%	12,33%	71,56%
Répondants	12,55%	16,01%	71,45%
Total	14,08%	16,38%	69,54%

Un exemple complet

- Option c) Établir un profil de la non-réponse partielle

Tableau par Secteur de résidence

Type de non-réponse	Moins urbain	Plus urbain
Aucune donnée transmise par l'enfant	48,75%	51,25%
Utilisation de l'ordinateur manquante	60,64%	39,36%
Autres	55,76%	44,24%
Répondants	47,36%	52,64%
Total	48,52%	51,48%

Un exemple complet

- Option c) Établir un profil de la non-réponse partielle
 - Les facteurs qui semblent les plus liés à la non-réponse partielle sont l'âge, le revenu, et le type de secteur de résidence (plus rural ou plus urbain).
 - Les résultats rapportés à l'option a) pourraient être affectés par les différences dans les profils respectifs (encore plus si les variables identifiées sont également liées à la variable d'intérêt).

Un exemple complet

- Option d) Repondération
 - Ici, nous allons utiliser les variables identifiées en c) pour ajuster à la hausse les poids des répondants, de sorte à ce qu'ils représentent aussi les non-répondants partiels.
 - On aura tout d'abord besoin de créer une variable indicatrice pour identifier facilement les répondants complets (rep_comp).

Un exemple complet

- Option d) Repondération
 - Un code SAS utile pour la repondération:

```
%macro write_adj;  
    %do i=1 %to 1000;  
        ,sum(bsw&i)/sum(bsw&i*rep_comp)  
        as adj&i  
    %end;  
%mend;
```

Un exemple complet

- Option d) Repondération

- Un code SAS utile pour la repondération:

```
%macro write_newboot,  
    %do i=1 %to 1000;  
        ,bsw&i*calculated adj&i as  
        new_bsw&i  
    %end;  
%mend;
```

Un exemple complet

- Option d) Repondération
 - Un code SAS utile pour la repondération:

```
proc sql;  
    create table filenameev2 as  
        select *, sum(fwtcw01l)/sum(fwtcw01l*rep_comp) as  
adjustment %write_adj, fwtcw01l*calculated adjustment as  
new_fwgt %write_newboot  
    from filename  
    group by fmmcq01, catrev, rural_urbain;  
quit;
```

Un exemple complet

- Option d) Repondération
 - Les 36 ajustements créés (par le croisement des variables d'âge, de revenus et de type de secteur de résidence) varient entre 1,1 et 1,6 et le plus petit groupe contient 19 répondants.

Un exemple complet

■ Option d) Repondération

- Un code similaire aurait pu être écrit pour s'assurer que les totaux de contrôle correspondent encore aux comptes de référence.
- Les groupements pour les totaux de contrôle sont définis par le croisement de la province de résidence, du sexe et de l'âge.

Un exemple complet

- Option d) Repondération

Combien d'heures par jour passes-tu sur l'ordinateur?

FATCeQ21	Fréquence	Pourcentage	Fréquence cumulative	Pourcentage cumulatif
Moins d'une heure par jour	867 235,9	38,25%	867 235,9	38,25%
1 ou 2 heures par jour	935 328,1	41,25%	1 802 564	79,50%
3 ou 4 heures par jour	347 297,4	15,32%	2 149 861	94,82%
5 ou 6 heures par jour	86 306,34	3,81%	2 236 168	98,63%
7 heures ou plus par jour	31 166,5	1,37%	2 267 334	100,00%

Un exemple complet

- Option d) Repondération

Effets principaux	Valeurs-p avec les poids de sondage et SAS proc logistic	Valeurs-p avec les poids normalisés et SAS proc logistic	Valeurs-p avec les poids bootstrap et SUDAAN
Âge	<0.0001	<0.0001	0.0000
Sexe	<0.0001	0.1952	0.4378
Type d'école	<0.0001	0.0859	0.5630
Ordinateur à la maison	<0.0001	<0.0001	0.0002
Travail/études	<0.0001	0.0020	0.1945
Surveillance parentale	<0.0001	<0.0001	0.0106

Un exemple complet

■ Option d) Repondération

- Les résultats ont très peu changé... Pourquoi?

Environ 19% de non-réponse partielle au total.

La moitié d'entre elle (aucune donnée enfant) semble plutôt similaire aux répondants... donc, il n'en reste que 10% pour changer les résultats.

Les variables utilisées pour créer les groupes de repondération sont peu corrélées avec la variable d'intérêt: peu de différence dans l'assiduité à utiliser l'ordinateur selon les différentes catégories des variables de repondération.

L'effet de l'âge était déjà en partie intégré par le modèle.

Un exemple complet

- Notons que nous aurions pu opter pour d'autres stratégies de traitement de la non-réponse:
 - Repondérer seulement les cas avec 'aucune donnée enfant' et imputer les autres formes de non-réponse partielle.
 - Repondérer les cas avec 'aucune donnée enfant', imputer les cas sans réponse pour l'utilisation de l'ordinateur et utiliser les autres données manquantes comme une catégorie valide.
 - Imputer toutes les valeurs manquantes.

Un exemple complet

- Nous allons maintenant nous concentrer un peu plus sur le 2^e objectif: la comparaison de l'usage courant et de l'usage passé. À cette fin, nous pourrions utiliser les données du cycle 5 et du cycle 6 d'au moins deux façons:
 - Rapporter sur les changements dans l'assiduité de l'utilisation des ordinateurs par les adolescents.
 - Proposer un modèle qui conviendrait aux assiduités respectives des cycles 5 et 6.

Un exemple complet

- Débutons par la première de ces deux tâches.
 - D'abord, il faudra vérifier la disponibilité et la comparabilité des variables entre les cycles 5 et 6.
 - Ensuite, si c'est possible, il faudra définir les différents concepts de la même façon.
 - Ensuite, il faudra aussi traiter la non-réponse pour chacun des cycles. Notons ici que les groupes de repondération pourraient être totalement différents ou encore exactement les mêmes.

Un exemple complet

- Dans le cas des données du cycle 5 et de notre exemple, nous utiliserons les mêmes groupes de repondération, puisqu'il semble être encore une fois lié au fait de répondre ou de ne pas répondre à notre ensemble de variables d'intérêt.

Distribution au sein des catégories de revenus pour les répondants complets et les non-répondants partiels du cycle 5

Type de réponse	Sous le seuil	Sous 1,5 fois le seuil	Au-dessus de 1,5 fois le seuil
Non-répondants partiels	21,22%	15,74%	63,04%
Répondants complets	13,92%	14,42%	71,66%

Un exemple complet

- Avant de regrouper les données et d'aller de l'avant avec un modèle combiné, il est important de faire des tests pour vérifier l'hypothèse que les quantités d'intérêt sont les mêmes. En complétant la première tâche, on effectue cette vérification.
- L'usage de l'ordinateur a évolué au fil des ans, et il est probable que les réponses puissent en être affectées.

Un exemple complet

- Profil de l'assiduité de l'utilisation de l'ordinateur par les adolescents:

	Cycle 5 (cohorte des 2 à 7 ans en 1994)	Cycle 6 (Cohorte des 0 à 5 ans en 1994)
Utilisateurs non assidus	82,44% (1,02%)	79,50% (0,95%)
Utilisateurs assidus	17,56% (1,02%)	20,50% (0,95%)

Un exemple complet

- Il faut être prudent en effectuant la vérification!
 - Ici, les 2 échantillons sont dépendants. Le test devra tenir compte de ce fait.

Choisissez un test et une méthode qui incorpore une composante de covariance.

Dans ce cas-ci, on conclurait que la différence est significative (valeur- $p=0,0293$).

Un exemple complet

- Voici donc nos principaux résultats:

Lorsque les 0-5 ans de la cohorte de 1994 sont devenus adolescents en 2004, environ 20% d'entre eux sont des utilisateurs assidus de l'ordinateur, c'est-à-dire qu'ils utilisent un ordinateur 3 heures ou plus par jour en moyenne.

Être plus âgé et avoir un ordinateur à la maison augmente la probabilité d'être un utilisateur assidu.

Un exemple complet

- Voici donc nos principaux résultats:

Les adolescents pour qui la surveillance parentale est très faible sont plus susceptibles d'être des utilisateurs assidus.

Les filles sont aussi susceptibles que les garçons d'être des utilisateurs assidus.

Un exemple complet

- Voici donc nos principaux résultats:

Les deux variables utilisés comme variables de procuration ('proxy') pour 'Disponibilité d'un ordinateur à l'école' et 'Quantité de temps libre dont les parents disposent pour être avec leurs enfants' ne sont pas liées à l'assiduité de l'utilisation de l'ordinateur.

Ceci ne signifie pas nécessairement que les facteurs identifiés ne sont pas liés à l'assiduité de l'utilisation.

Un exemple complet

- Voici donc nos principaux résultats:

L'assiduité avec laquelle les ordinateurs sont utilisés par les adolescents semble en progression rapide: il y a 3% plus d'utilisateurs assidus au sein de la cohorte des 0-5 ans de 1994 qu'au sein de la cohorte de 2-7 ans de 1994 (lorsque les 2 groupes sont respectivement âgés entre 10 et 15 ans), soit une hausse de plus de 15%.



Statistique
Canada

Statistics
Canada

Canada



Statistique Canada
www.statcan.gc.ca



De l'aide en développement



De l'aide en développement...

- Nous travaillons à la mise sur pied d'un outil de repondération (présentement encore à l'étape d'un prototype) pour aider à tenir compte de la non-réponse partielle
 - Outil compatible avec Bootvar
 - Outil prendrait la forme d'une macro en SAS

De l'aide en développement...

- L'outil (macro intégrée à Bootvar) permettrait de:
 - Définir le domaine d'intérêt
 - Créer des groupes de repondération pour redistribuer les poids des non-répondants en modélisant la non-réponse à partir d'une liste de variables identifiées comme potentiellement liées.
 - Faire la repondération
 - Utiliser la même macro de Bootvar pour comparer les résultats

De l'aide en développement...

- Définir le domaine d'intérêt :
 - L'utilisateur fournirait les informations suivantes:
 - varinterest -> noms des variables d'intérêt pour l'analyse
 - nointerestvalue -> valeurs qui indiquent les codes de non-réponse pour les variables d'intérêt
 - L'outil créerait une variable binaire appelée « respondent »
 - Si « respondent » est égal à 0, alors le poids associé à cet enregistrement serait redistribué

De l'aide en développement...

- Création des groupes de repondération :
 - L'utilisateur identifierait les variables potentiellement liées à la non-réponse
 - cat_var_list -> noms des variables de catégorie
 - cont_var_list -> noms des variables continues
 - Parmi celles-ci, les variables pertinentes seraient retenues dans un modèle de régression logistique pour séparer le fichier en 10 groupes de repondération selon la méthode des déciles.

De l'aide en développement...

- Repondération :

- En utilisant les groupes de repondération créés à l'étape précédente, l'outil repondèrerait le fichier d'analyse, en créant un nouveau poids final et de nouveaux poids bootstrap.
- Les totaux de contrôle ne seraient pas corrigés.

De l'aide en développement...

- L'utilisateur fournirait le nom de la macro Bootvar réalisant l'analyse souhaitée:
 - `macro_used` -> inclure, sans le signe %, la macro de Bootvar voulue pour faire l'analyse
- L'outil réaliserait ensuite de nouveau l'analyse souhaitée, avec les nouveaux poids obtenus suite à la repondération.
- En un clin d'œil, le chercheur pourrait comparer les résultats des analyses pré-repondération et post-repondération.



Coordonnées de contact

For more
information,
please
contact

Pour plus
d'information,
veuillez
contacter

Yves Lafortune
yves.lafortune@statcan.gc.ca