Sources d’erreur et biais

Sources d’erreur et biais

- Utilisations des données d’enquête :
  - « Describers » & « Modelers »
- Sources d’erreur
- « Nonsampling errors »
  - Populations d’intérêt
  - Défaut ou excès de couverture
  - Non-réponse
  - Erreur de mesure
- Sources d’erreur et phases d’enquête
- La pointe de l’iceberg … et le reste
- Sources d’erreur, contraintes, mode de réalisation
- L’ «art» du sondeur

Utilisations des données d'enquête : « Describers » & « Modelers »

Différents langages, différentes préoccupations	
« Describers »	« Modelers »
Accent sur l'estimation des caractéristiques d'une population	Accent sur la validation d'hypothèses théoriques
Accent sur l'estimation de moyennes et proportions	Accent sur l'exploration de structures de covariance
Forte attention aux erreurs de non-observation (défauts de couverture, non-réponse)	Forte attention aux erreurs d'observation (questionnaire)

Sources d'erreur {1/3}

- Erreur d'échantillonnage
  - Hétérogénéité des mesures parmi les individus de la population
- Défaut ou excès de couverture
  - Probabilité de sélection nulle ou non connue pour les individus de la population
- Non-réponse
  - Défaut de collecte de toute ou partie de l'information pour certains individus de l'échantillon
- Erreur de mesure
  - Influence de l'enquêteur sur les réponses des personnes interrogées
  - Incapacité (ou manque de volonté) des personnes interrogées à répondre aux questions : mémoire, impréparation, facteurs psychologiques, …
  - Défauts de l'instrument de mesure (questionnaire ou autre)
  - Effets du mode de recueil (face à face, téléphone, auto-administré papier ou Internet)

Sources d'erreur {2/3}

- Ces erreurs peuvent être liées les unes aux autres
  - Eg : Faire du « forcing » pour réduire la non-réponse peut amener à amplifier les erreurs de mesure
- En général, les efforts de modélisation et de mesure sont portés sur l'erreur d'échantillonnage et la non-réponse
- Souvent on ne sait que très peu – et parfois rien du tout - sur les erreurs d'observation et les défauts de couverture
- Or, cela peut s'avérer létal, car ces erreurs - qui ont essentiellement la nature de biais – ne diminuent pas lorsque la taille d'échantillon augmente

Sources d'erreur {3/3}

Moralité

- Les efforts visant à affiner une méthode de tirage ou l'expression d'un estimateur pour obtenir un gain de précision peuvent s'avérer bien illusoires si, par ailleurs, les erreurs d'observation, les défauts de couverture ou la non-réponse sont importants

- Dans une telle situation, une taille d'échantillon très importante ne sera pas non plus de nature à éviter la déroute
  - Lors de la Présidentielle américaine de 1936, le « vote de paille » organisé par le *Literary Digest* - portant sur près de deux millions de lecteurs - donnait une confortable avance à Alfred Landon (54%) … alors que Franklin Roosevelt allait recueillir 61% des suffrages !

« Nonsampling errors » : Populations d'intérêt

- Population objet de l'inférence (population of inference)
  - Ensemble des unités à étudier
- Population cible du sondage (target population)
  - Ensemble des unités étudiées
- Base de sondage (frame population)
  - Liste des unités utilisée pour la sélection de l'échantillon: l'« univers » auquel font référence la plupart des livres de statistique
- Population enquêtable (survey population)
  - Liste des unités accessibles, physiquement et mentalement prêtes à répondre, souhaitant répondre aux questions
  - Il s'agit bien évidemment d'une abstraction, puisque elle ne peut être observée indépendamment des opérations d'échantillonnage elles-mêmes
- Non-réponse
  - divergences entre « frame » et « survey population »
- Erreurs de couverture
  - divergences entre « frame » et « target population»

« Nonsampling errors » : Défaut ou excès de couverture {1/2}

- Ambiguïté du repérage des unités de la population
  - Une base de sondage se doit pour le moins d'être une liste d'identifiants de bonne qualité
- Manque d'exhaustivité
  - Chaque unité faisant partie du champ de l'enquête doit être présente dans la liste des identifiants
- Doubles comptes
  - Aucune unité doit être présente plusieurs fois dans la base (surtout si le nombre de fois n'est pas connu)
- Absence d'informations auxiliaires
  - Leur disponibilité peut être mise à profit pour améliorer soit la méthode de tirage, soit l'estimateur, soit les deux
- Vieillissement de la base elle-même
- Absence ou inaccessibilité de la base de sondage
  - (situation finalement pas si rare!)

« Nonsampling errors » : Défaut ou excès de couverture {2/2}

- L'erreur de couverture est une fonction
  - de la proportion de population non couverte par la base de sondage
  - de la différence dans la valeur de la variable d'intérêt entre « frame » et « target population»

- $Y_c = Y + (N_{nc} / N) * (Y_C - Y_{nc})$

  où Y représente la valeur auprès des N unités de la target population

  Y_c représente la valeur auprès des N_c unités couvertes par la « frame population »

  Y_{nc} représente la valeur auprès des N_{nc} unités non couvertes par la « frame population »

- L'erreur de couverture
  - est liée à la variable d'intérêt
  - n'est pas une propriété de l'échantillon

« Nonsampling errors » : Non-réponse {1/3}

- Comme pour le défaut de couverture dû au manque d'exhaustivité de la base de sondage, la non-réponse
  - nous met dans l'impossibilité d'observer la valeur de la variable d'intérêt
  - engendre un biais non mesurable, puisque l'on ne sait pas si les unités observées sont comparables aux unités non observées

- A différence du défaut de couverture, la non réponse
  - est d'ampleur mesurable, à partir de l'échantillon tiré (taux de non-réponse calculable)
  - peut être complète ou partielle (l'individu sélectionné répond à certaines questions et pas à d'autres)

- En diminuant la taille de l'échantillon, la non-réponse occasionne une perte de précision (quelles que soient les hypothèses formulées sur le profile des non-répondants)

« Nonsampling errors » : Non-réponse {2/3}

- Le taux de non-réponse est souvent interprété comme LA mesure de qualité de l'estimation de la variable d'intérêt
  - or, il ne s'agit que d'une composante de l'erreur et ne peut pas en donner seul la mesure
- L'erreur dû à la non-réponse est une fonction
  - du taux de non-réponse
  - de la différence dans la valeur de la variable d'intérêt entre répondants et non-répondants
- $y_r = y_n + (nr / n) * (y_r – y_{nr})$
- L'erreur de non-réponse
  - est liée à la variable d'intérêt
  - n'est pas une propriété de l'échantillon

« Nonsampling errors » : Non-réponse {3/3}

- Une expression plus complète de la variable d'intérêt estimée devrait être

$$y_r = y_n + (nc / n) * (y_r - y_{nc}) + (ni / n) * (y_r - y_{ni}) + (rf / n) * (y_r - y_{rf})$$

où y_{nc} représente la valeur auprès des nc unités non contacté

y_{ni} représente la valeur auprès des ni unités incapables de fournir une réponse

y_{rf} représente la valeur auprès des rf unités refusant l'interview

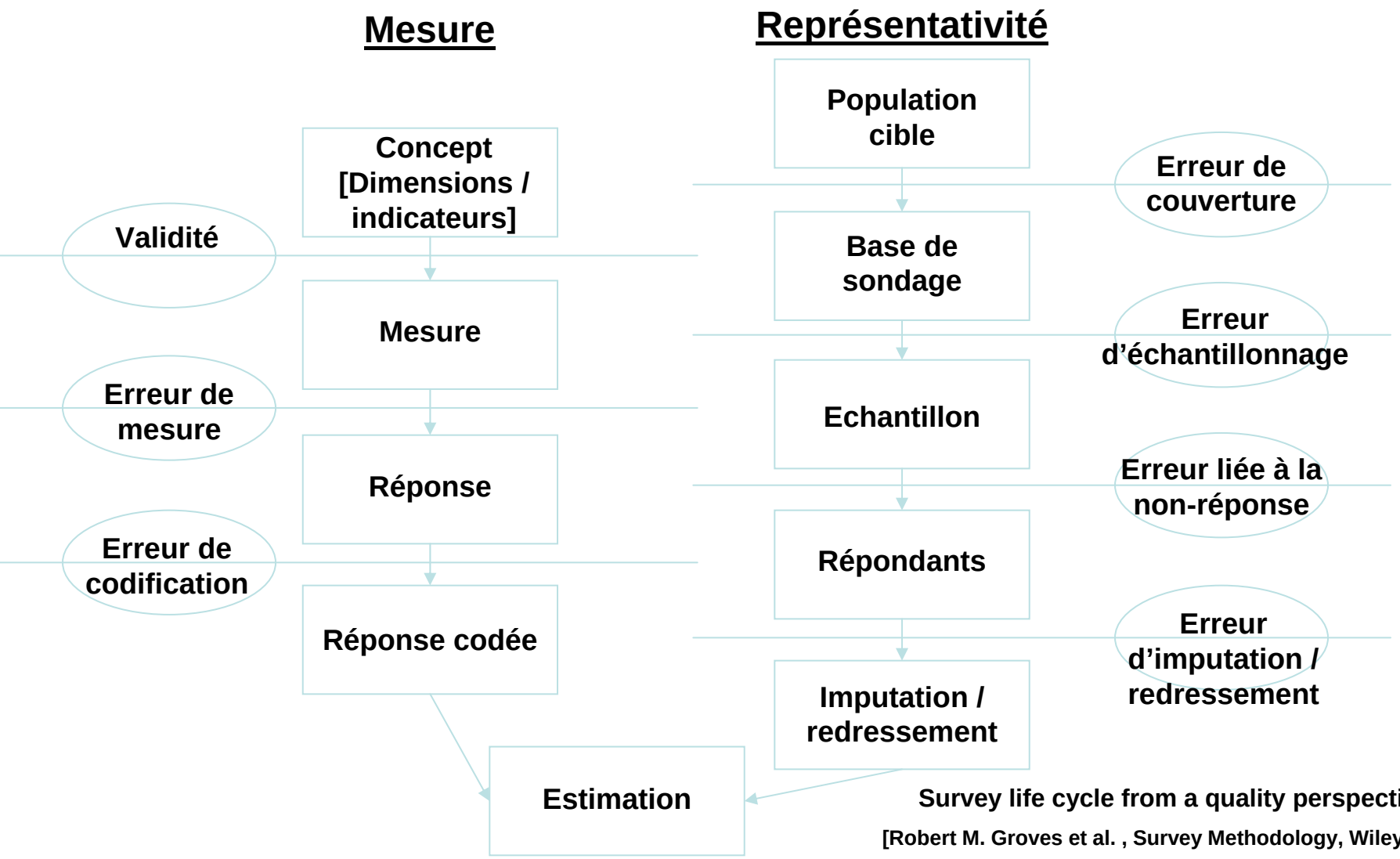
avec nc + ni + rf = nr

« Nonsampling errors » : Erreur de mesure

- Il y a erreur de mesure lorsque la valeur de la variable d'intérêt collectée pour un individu est différente de la vraie valeur attachée à ce même individu. Quelques cas (liste non ordonnée et non exhaustive !) :

- Questions faisant appel à la mémoire des personnes interrogées
- Questions portants sur des sujets sensibles (revenus, comportements sexuels, consommation de drogues, …)
- Mécanismes psychologiques liés à l'interaction enquêteur/enquêté
- Interprétation des réponses de la part de l'enquêteur
- « Suggestions » de l'enquêteur à l'enquêté
- Mauvaise compréhension de la question (surtout en cas de traduction des questions depuis une langue étrangère)
- Formulation de la question, effets d'ordre, …
- Fatigue due à la durée d'interviews
- Autres effets enquêteur : le sexe, l'âge de l'enquêteur, sa façon de se présenter … ne sont pas sans conséquences sur la qualité des réponses obtenues

Sources d'erreur et biais. Les défis à relever



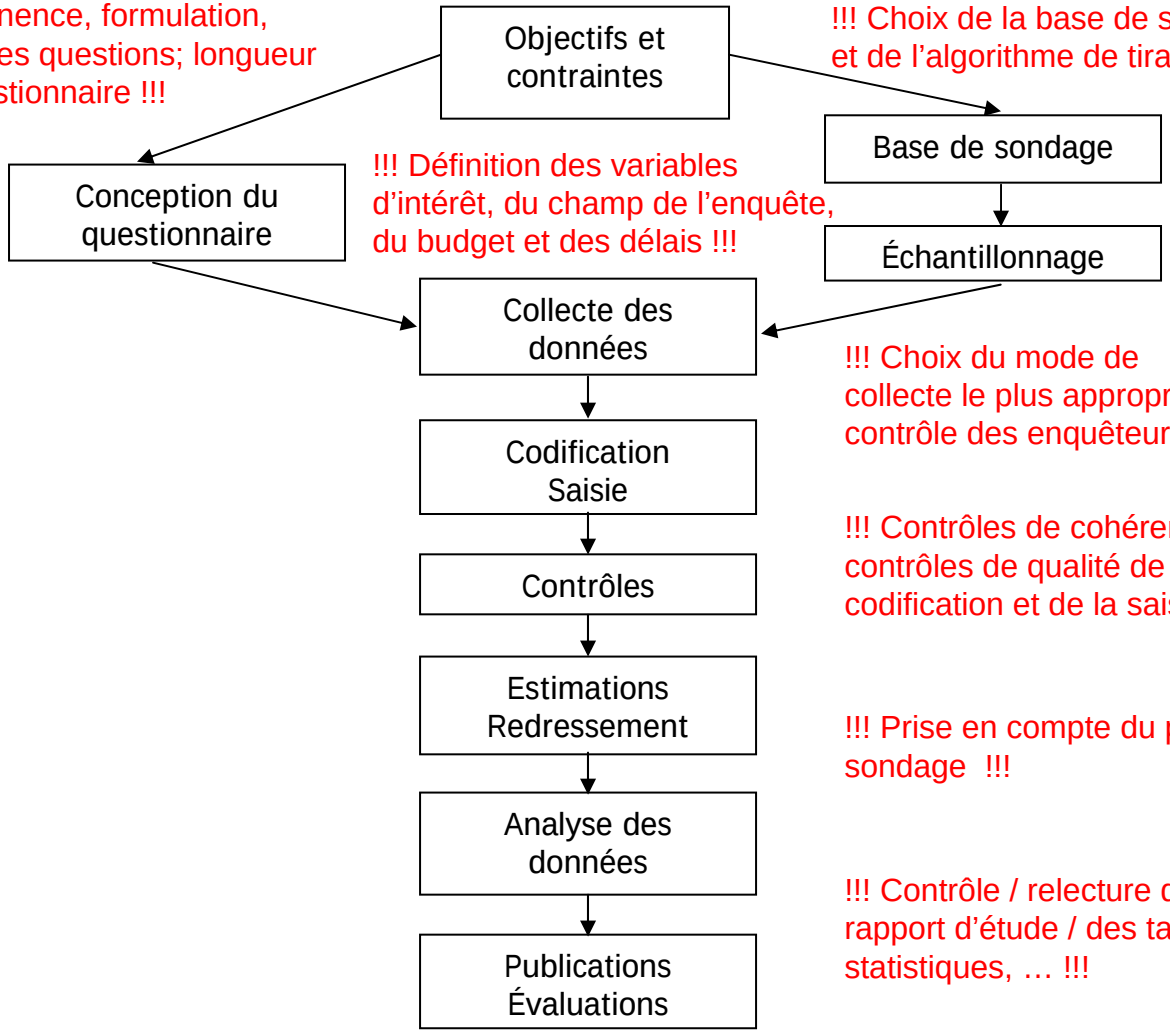
Survey life cycle from a quality perspective

[Robert M. Groves et al. , Survey Methodology, Wiley, 2009]

Sources d'erreur et phases d' enquête

- Couverture
- Non-réponse
- Échantillonnage
- Erreurs de mesure

- Saisie
- Codification
- Traitement
- Présentation

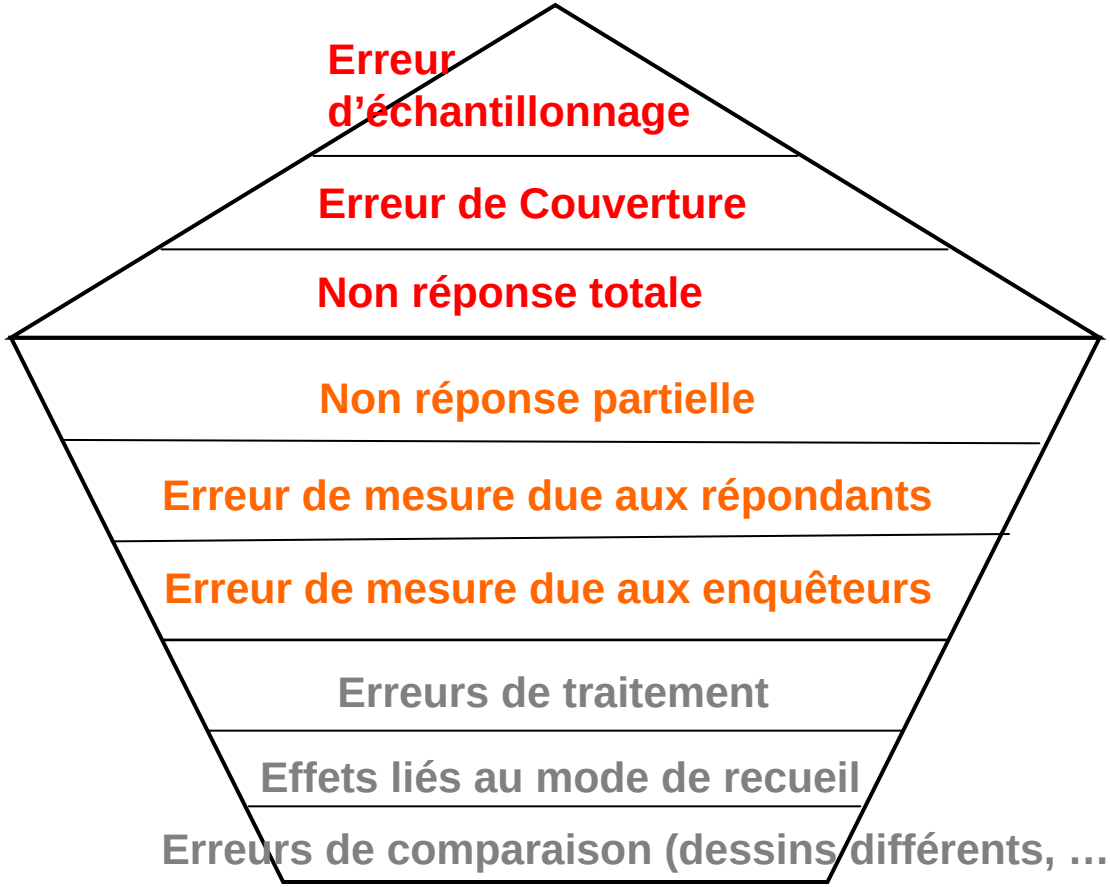


La pointe de l'iceberg … et le reste

Sélection des répondants

Exactitude des réponses

Mode de réalisation



Weisberg, H.F. (2005), The total survey error approach, The University of Chicago, Chicago

Sources d'erreur, contraintes, mode de réalisation

Sources d'erreur et biais

- Echantillonnage
- Couverture
- Non-réponse
- Mesure

Contraintes

- Coûts
- Délais
- Etique

Effets du mode de réalisation de l'enquête

- Questionnaire
- Mode de recueil
- Effets de comparaison (plan d'échantillonnage, temps, …)

L’ «art» du sondeur

- La théorie statistique nous aide à mesurer et à réduire l’erreur d’échantillonnage
- L’«art» du sondeur, praticien d’enquête, consiste à juger de l’importance du non mesurable
- La pratique de cet « art » requière la compréhension
  - des causes qui sont à l’origine des erreurs
  - de leur importance relative
  - des effets générés
  - des coûts relatifs aux efforts de réduction des erreurs
- Juger de l’importance du non mesurable est un « art » qui ne doit pas se transformer en alibi pour arrêter tout effort de modélisation et mesure de l’erreur

Bibliographie

- Lecture minimale
  - Ardilly, P. (*2ème édition actualisée et augmentée,* 2006), Les techniques de sondage, Editions Technip, Paris
    - Chapitre I. Aspects universels, principes de base

- Pour aller plus loin
  - Groves, R.M. (1989), Survey errors and survey costs, Wiley, New York
    - Chapitres I,III,IV,VII
  - Groves, R.M., Fowler Jr., F.J., Couper, M.P., Lepkowski, J.M., Singer, E., Tourangeau, R., (2nd ed. 2009), Survey Methodology, Wiley, New York
    - Chapitre 2
  - Floyd J. Fowler, Jr., (4th ed. 2009), Survey Research Methods, Wiley, New York
    - Chapitre 2
  - Weisberg, H.F. (2005), The total survey error approach, The University of Chicago, Chicago
    - Chapitres 2, 14, 15