Module 6 : Contrôle de la qualité des données
Vidéo 4 sur 6 : Surveillance des données substantielles

Sharan Sharma

MANNHEIM
BUSINESS SCHOOL

Sources de données possibles pour surveiller les enquêteurs

1. Coordonnées GPS
2. Entretiens de contrôle
3. Écoutes
4. Répondants mystère
5. Données substantielles
6. Paradonnées
7. Enregistrements audio (CARI)

5) Surveillance à partir des données : surveillance des données substantielles réelles

- Est-ce que les données collectées présentent des tendances *inhabituels* ?
  - Moyenne
  - Écart type (« répartition des données »)
  - Relations entre les variables, par ex. régression (utile mais peu utilisée dans le contrôle qualité)

- Idée de base : Quel est le résultat de la comparaison entre les différents groupes ?
  - Les « groupes » sont souvent des enquêteurs du contrôle qualité, mais il peut aussi s'agir de superviseurs, d'agences de terrain, de régions géographiques, de moments de la journée, de périodes sur le terrain (tendances), etc.

- Quelques méthodes basées sur l'analyse des données substantielles :

A. Statistiques descriptives

- Dans quelle mesure les données collectées par un enquêteur s'écarte de la moyenne globale/des données attendues ?

- Murphy et al. (2004) ont calculé une note pour chaque enquêteur à partir des écarts dans les taux d'usage de drogues par strate d'âge, de sexe et d'origine ethnique.
  - Les écarts de réponse pour trois falsificateurs connus étaient les plus élevés parmi le groupe de dix enquêteurs.

Exemple de l'enquête National Survey on Drug Use and Health, États-Unis, 2002.

Appels de vérification : 2 premiers non-entretiens et 2 premiers entretiens de chaque quartile + au moins 5 % des filtrages de chaque enquêteur + au moins 15 % des entretiens complets de chaque enquêteur

- Toute situation suspecte, par ex. numéro manquant/refusé → vérification plus approfondie

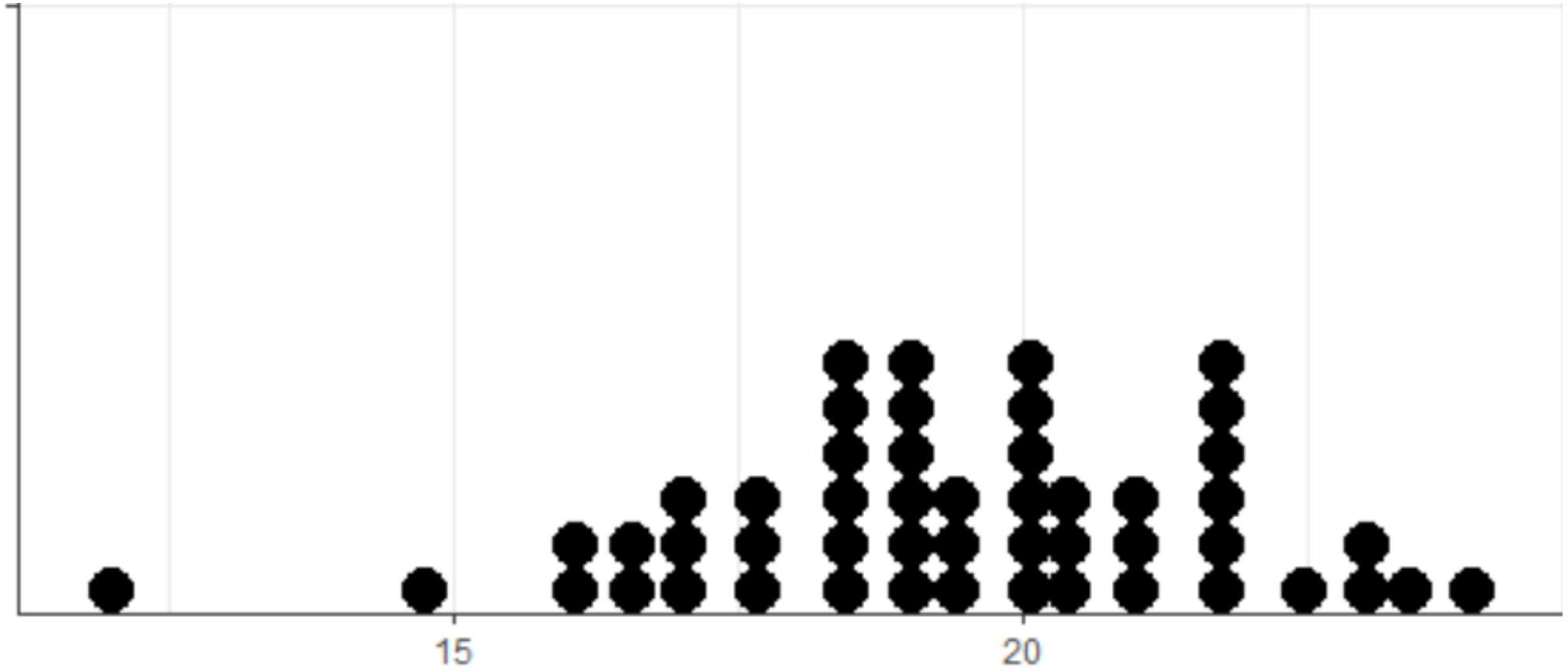
Interviewer / Cases	Number of Interviews	Lifetime Use (%)				
		Cigarettes	Alcohol	Marijuana	Cocaine	Heroin
All valid cases	1,188	56.9	71.3	45.1	15.8	1.8
Falsifier 1's fraudulent cases	92	26.1 [a]	59.8 [a]	35.8	20.6	1.1
Falsifier 2's fraudulent cases	119	48.7	50.4 [a]	24.4 [a]	7.6 [a]	4.2 [a]
Falsifier 3's fraudulent cases	77	29.9 [a]	37.6 [a]	22.1 [a]	1.3 [a]	0.0

[a] rate significantly different from valid case rate at $p<.05$ level

Source : Murphy et al, 2004

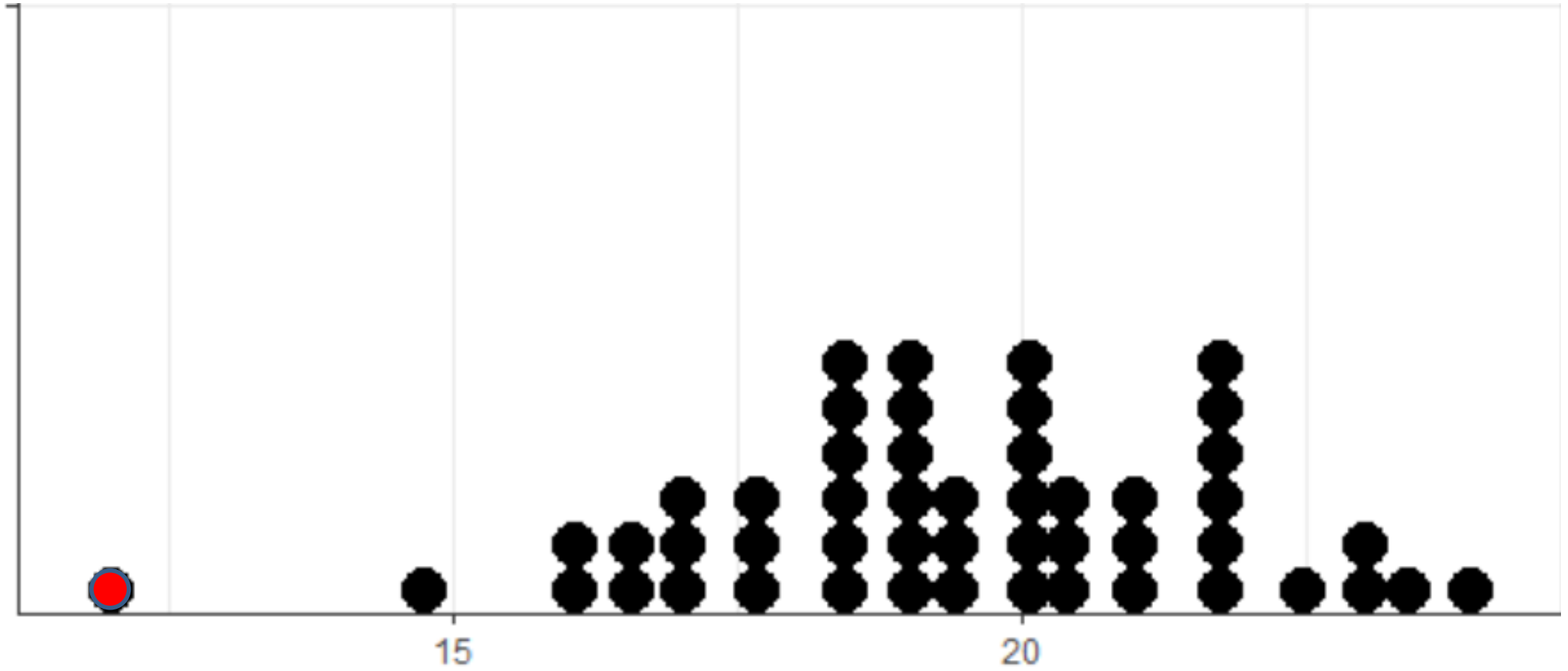
Que pensez-vous de cette situation ?

(Chaque point représente la moyenne pour la charge de travail d'un enquêteur)



Kilogrammes de riz consommés le mois dernier

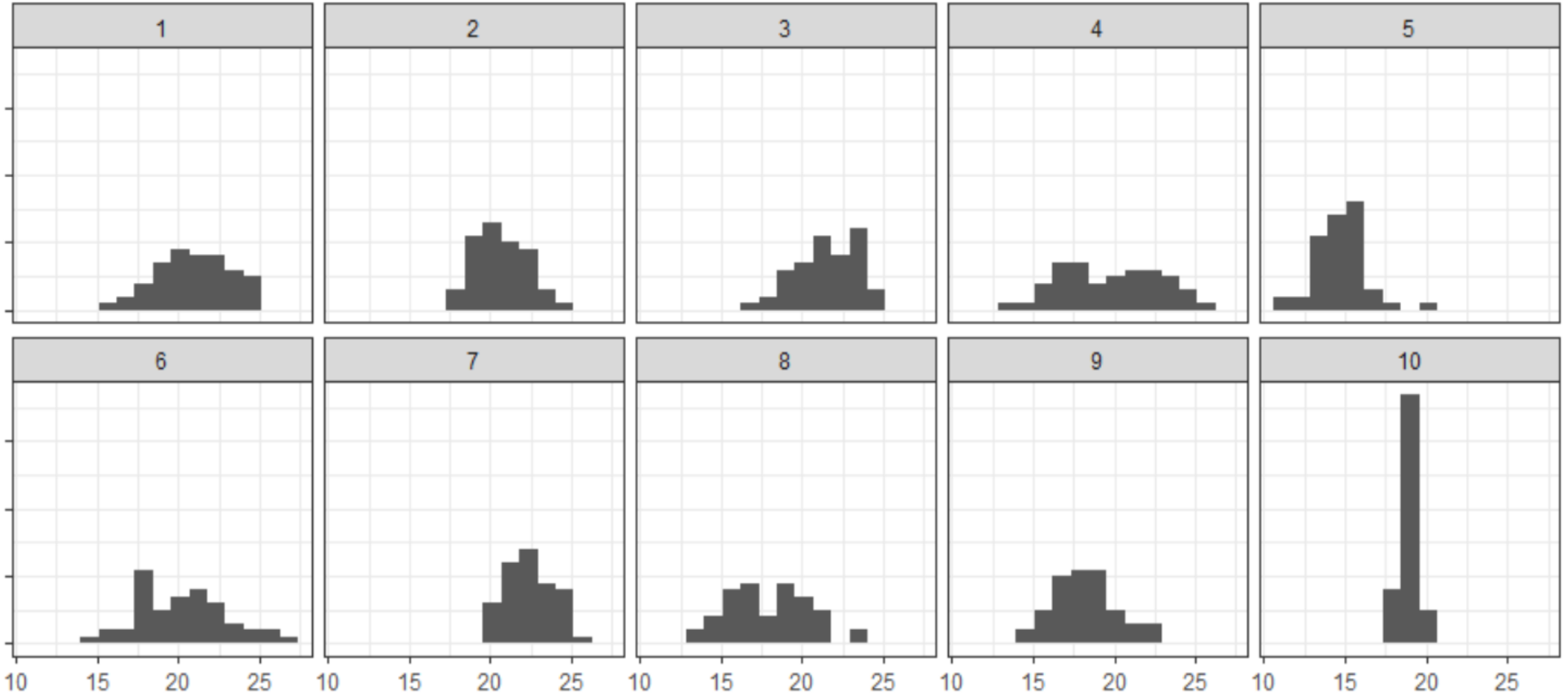
Erreurs de saisie des données ?



Kilogrammes de riz consommés le mois dernier

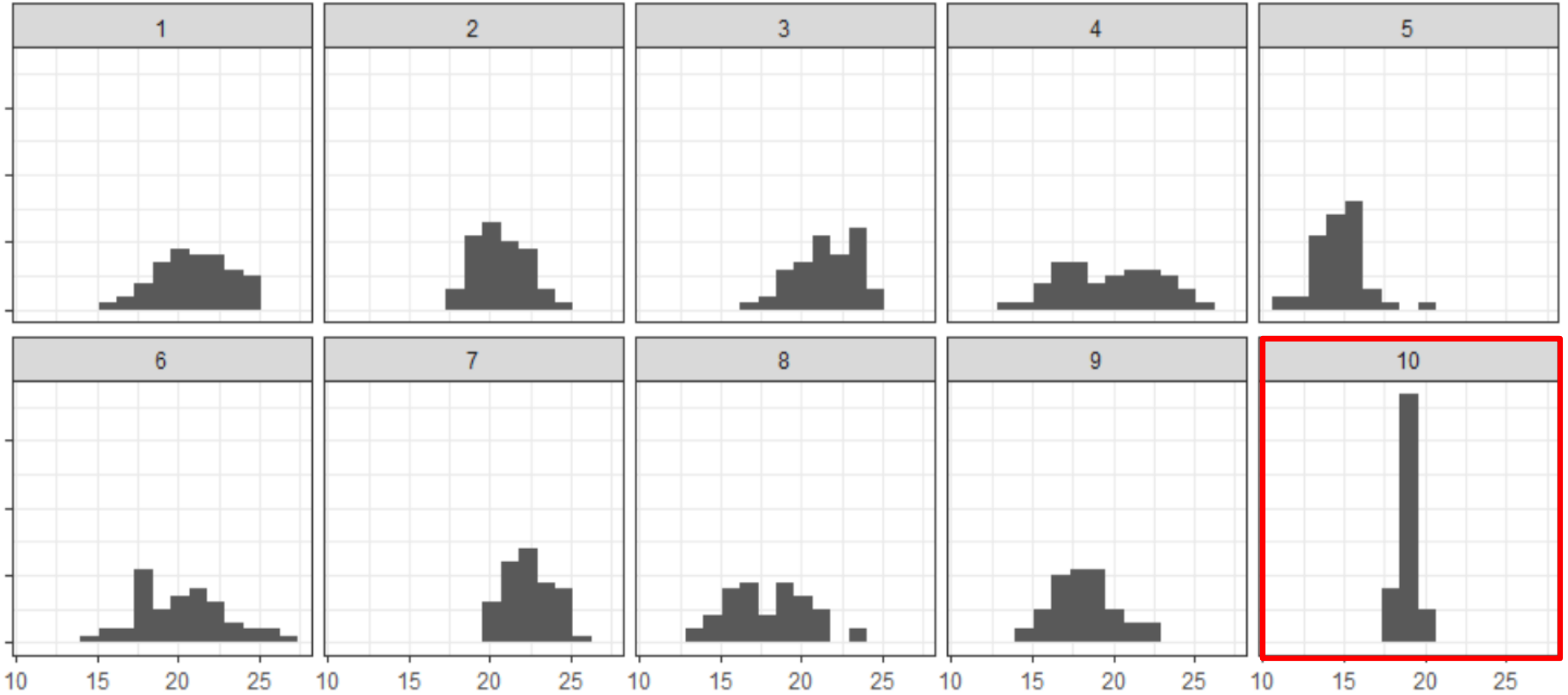
Que pensez-vous de cette situation ?

(Répartition des réponses pour 10 enquêteurs, un pour chaque panel)



Kilogrammes de riz consommés le mois dernier

Se donne du mal pour éviter d'être détecté ?



Kilogrammes de riz consommés le mois dernier

B. Lissage

- Tendance à employer une catégorie de réponse identique pour tous les éléments d'une série de questions. Par ex. obtenir dix réponses « D'accord » à une série de dix questions d'opinion avec une échelle de cinq réponses possibles allant de « tout à fait d'accord » à « pas du tout d'accord »

- Les enquêtes par téléphone sont plus exposées à ce type de comportement que les enquêtes en face à face [Holbrook et al (2003)]

- Pour une série de questions binaires, calculer le pourcentage de Oui (ou de Non) pour chaque enquêteur. Comparez les résultats des enquêteurs avec cet indicateur.
  - Si possible, formulez les questions de façon à ce que l'on attende un mélange de réponses Oui/Non.
  - Kim et al (2018) répertorient différents moyens de mesurer le lissage

C. Arrondi

« Quel est votre revenu avant impôt pour 2019 ? »

Réponses collectées par l'enquêteur 1 (en USD) : 40 235, 110 683, 23 568, 15 000, 89 675, ...

Réponses collectées par l'enquêteur 2 (en USD) : 40 000, 100 000, 20 000, 15 000, 89 000, ….

- On parle aussi de « regroupement » pour les variables comme l'âge.

D. Réponses aux questions sur la liste du ménage

- L'obtention de la liste des membres du ménage et leurs caractéristiques représente souvent la première composante (vitale).

- Dans les pays où les ménages sont prolifiques, les enquêteurs peuvent avoir du mal à inclure tout le monde.
  - Il peut y avoir un problème de définition → « Mon mari travaille à la ville et ne revient que tous les 15 jours. » Faut-il l'inclure ?
  - Il peut s'agir de sous-déclarations volontaires → de nombreux questionnaires comportent des questions identiques pour tous les membres du ménage (niveau d'éducation, par ex.). Certains enquêteurs peuvent être tentés d'exclure les membres requérant l'administration d'un instrument spécifique, par ex. un questionnaire spécial « femme éligible ».

E. Réponses aux questions de filtrage

- Gardez un œil sur le taux de filtrage des ménages

Tourangeau et al (2012) montrent que les enquêteurs dont le taux de filtrage est plus élevé ont tendance à trouver moins de ménages avec des membres éligibles.

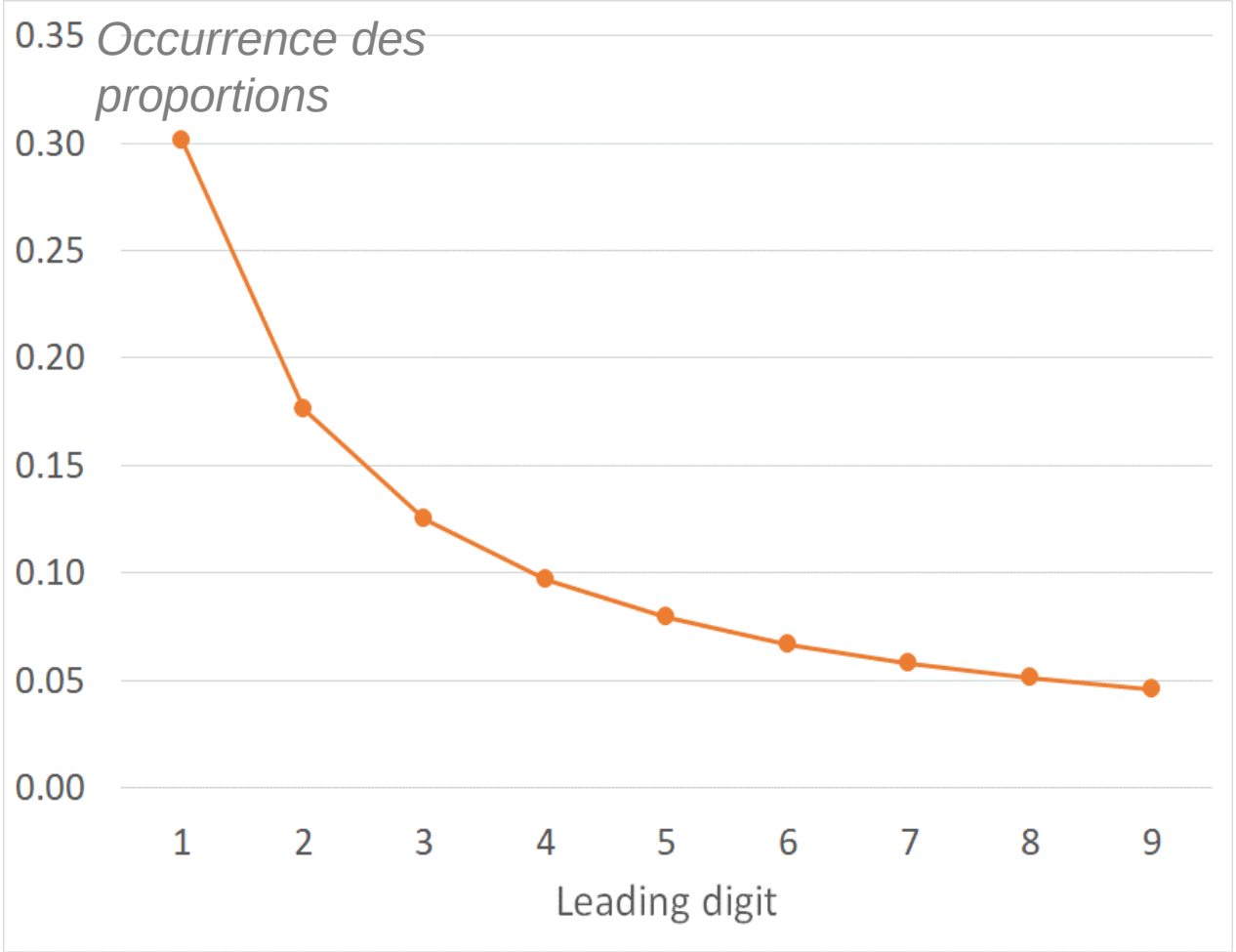
- Au niveau de la question : combien d'enquêteurs obtiennent une réponse négative aux questions avec un saut ?

Par ex. un « Non » à la question « Possédez-vous des terres agricoles » entraîne le saut d'une série de 25 questions sur la propriété et la location de terres agricoles, les cultures pratiquées, etc.

F. Loi de Benford

- Il apparaît que les premiers chiffres d'une forte proportion de nombres de la vraie vie ne sont pas distribués uniformément : le chiffre 1 (ou 2, 3 ... 9) n'apparaît pas 11,1 % (1/9) du temps, mais suit une distribution décroissante.

- C'est bon à savoir, car un enquêteur falsificateur s'attendra généralement à une distribution uniforme.



Loi de Benford...

- Swanson et al (2003) ont étudié 734 684 dépenses enregistrées dans l'enquête trimestrielle Consumer Expenditure Survey (2000). La distribution correspond assez bien à la distribution attendue.

Leading Digit (d)	Reported Expenditures Number	Percent (SE)		Benford's Law $\log_{10}\left(\frac{d+1}{d}\right)\times 100\%$
1	223,776	30.5	(.063)	30.1
2	141,992	19.3	(.053)	17.6
3	90,589	12.3	(.045)	12.5
4	66,266	9.0	(.040)	9.7
5	76,473	10.4	(.044)	7.9
6	50,024	6.8	(.034)	6.7
7	35,019	4.8	(.029)	5.8
8	32,294	4.4	(.028)	5.1
9	18,251	2.5	(.021)	4.6
Total	734,684	100.0		100.0

- Ils ont mené cette analyse pour chaque enquêteur...

Leading Digit (d)	CEQ's Nationwide Distribution (n=734,684)	A Typical FR (θ = 10.39) (n=1,143)	An Unusual FR (θ = 102.43) (n=1,132)
1	30.5	31.4	28.9
2	19.3	19.7	18.0
3	12.3	11.6	8.1
4	9.0	9.5	8.5
5	10.4	8.3	17.2
6	6.8	6.4	10.5
7	4.8	4.7	4.2
8	4.4	5.2	3.2
9	2.5	3.2	1.3
Total	100.0	100.0	100.0

FR → Field representative, représentant sur le terrain

Source : Swanson et al (2003)

Loi de Benford...

Certaines réussites dans le monde des enquêtes, par ex. Schräpler (2010), montrent son efficacité pour repérer les enquêteurs falsificateurs. Mais elle n'est pas très utilisée. Cela est dû en partie au fait qu'elle ne s'applique qu'à des données spécifiques :

- Comportant uniquement des valeurs positives avec une distribution unimodale.
- Présentant une inclinaison positive.
- N'intégrant pas de limite maximum.
- Données brutes, dans le sens où elles ne sont pas basées sur des moyennes ou autres synthèses.

G. Doublons

- Peuvent apparaître pour des raisons techniques, par ex. quand un enquêteur essaie d'envoyer des données alors qu'Internet est coupé. Envoie une deuxième fois alors que le premier envoi a fonctionné.

- Peuvent aussi être dus à des falsifications au niveau de la gestion administrative :
  - Blasius et Thiessen (2012) : 36 questions de l'enquête World Values Survey (2005 - 2008). Un pays présentait 25 % de cas de doublons ! À l'exception d'un pays : « ... pour les autres pays, la seule explication plausible est que la taille des échantillons a été augmentée par copier-coller »

- Les données falsifiées ne sont pas forcément des doublons totalement identiques : Robbins (2015) montre sur 1 200 entretiens, 178 sont identiques à plus de 80 % dans la 3^{e} vague de l'enquête Arab Barometer. Un enquêteur avait à lui seul un taux de correspondance supérieur à 80 % pour 122 de ses 123 entretiens.

H. Tableaux de contrôle terrain

- Calcul de plusieurs synthèses de groupe à partir des données collectées

- Généralement actualisés chaque semaine/toutes les deux semaines, ou quand un certain nombre d'entretiens (par ex. un lot de 200 entretiens) ont été réalisés.
  - Peuvent produire des indicateurs à partir des données incrémentielles permettant des comparaisons avec les données précédentes. Fournissent une tendance pour les indicateurs.

- Voir ICF Macro (2009) et Dupuis (2018) pour des exemples complets.

Par ex. l'âge des femmes est-il ramené artificiellement en-dessous de 15 ans pour éviter les entretiens ?

Table FC.5a: Female Age Displacement

Number of all women ages 12-17 years listed in the household roster by single years of age and age ratios, by team.

Team	Age of women in years (N)						Age ratio (15/14)	Extended age ratio (15+16)/(13+14)
	12	13	14	15	16	17		
Team 1								
Team 2								
Team 3								
Team 4								
Team 5								
All teams								

Source : Dupuis et al (2018)

Question clé pour de nombreuses vérifications basées sur les données substantielles : le phénomène est-il imputable à l'enquêteur ou au répondant ?

- Vérification rapide : en admettant une variation suffisante des caractéristiques des répondants, synthétisez les mesures de lissage, d'arrondi, etc. par enquêteur et par rang.

- Contrôlez les caractéristiques des répondants dans les analyses statistiques...

Ajustements en fonction des profils de répondant...

- Non-réponse : West et Groves (2013) évaluent le taux de coopération par enquêteur après ajustement en fonction de la difficulté prévue des cas.

- Erreur de mesure : Sharma et Elliott (2019) ont recours à des modèles multiniveaux d'ajustement en fonction des caractéristiques des répondants pour détecter les enquêteurs susceptibles de pratiquer la falsification.

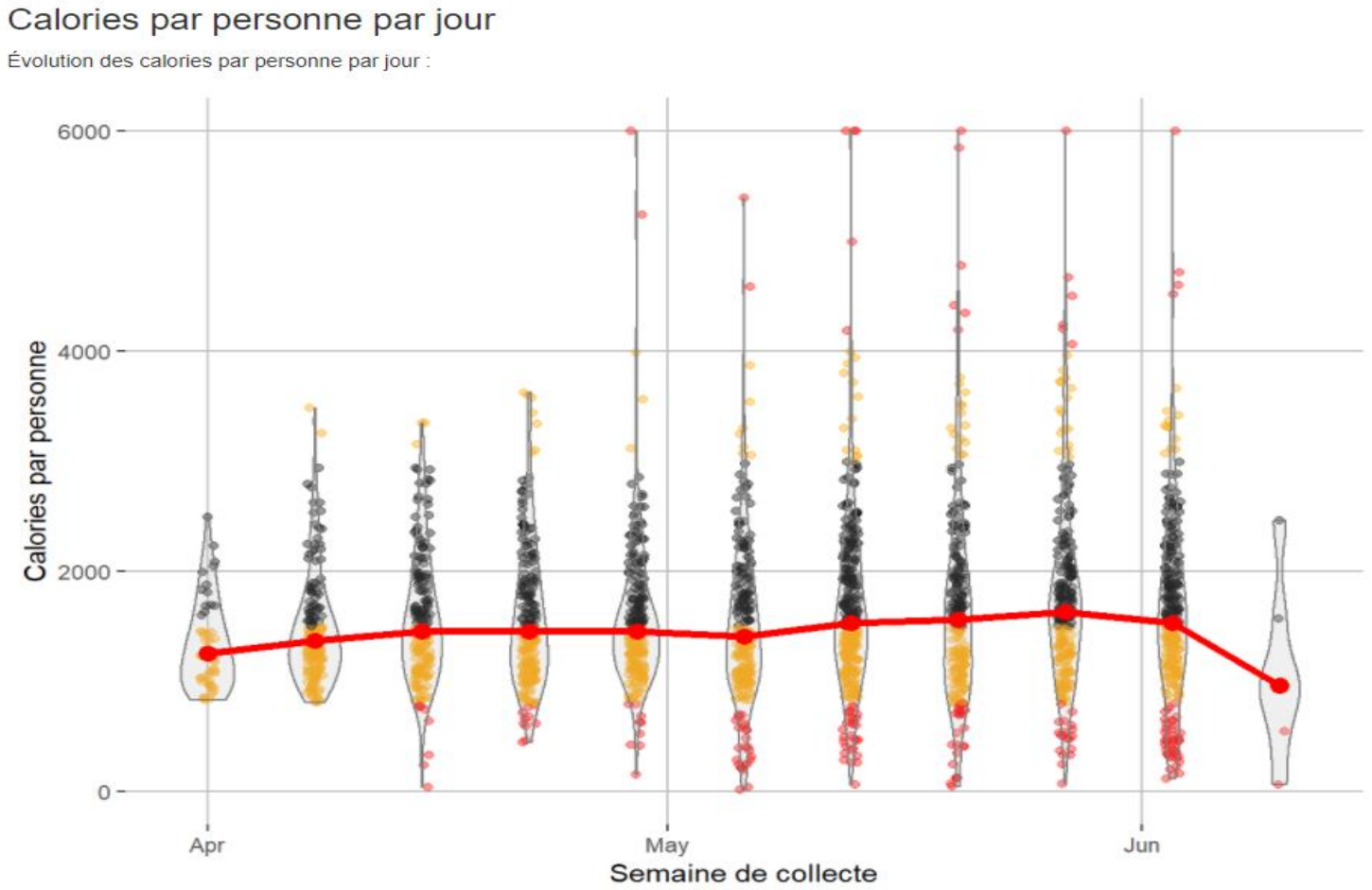
I. Tendances

Pourquoi est-ce important ?

- Le comportement des enquêteurs évolue dans le temps [Olson et Smyth (2020)]
  - Le comportement des enquêteurs est également lié à la durée des entretiens.

- Alors que les enquêteurs deviennent plus efficaces avec le temps, la fatigue peut également entrer en jeu. Avec l'expérience, les enquêteurs peuvent aussi avoir recours à des tactiques comme sauter des questions.

I. Tendances...



Consommation hebdomadaire de nourriture combinée aux calories.

Noir -> OK
Jaune -> relativement faible/élevée
Rouge -> hors limites

Source : https://github.com/arthur-shaw/ehcvm-rapport-hebdomadaire
Enquête Harmonisée sur les Conditions de Vie des Ménages (EHCVM) 2018-19

J. Données manquantes

Contrôle particulièrement utile à mener pour :

- Les questions sensibles
- Les questions nécessitant plus que l’approfondissement habituel
- Les questions nécessitant un surcroît de travail...

Données GPS employées pour des mesures de terrain précises. Surveiller les pourcentages manquants

Parcel GPS area measurement, by team					
			Why not measured (%)		
Team	Parcels (N)	Measured (%)	Refused	Not accessible	Other
Overall	—	—	—	—	—
Team 1	—	—	—	—	—
Team 2	—	—	—	—	—
…	—	—	—	—	—
Team N	—	—	—	—	—

Source : Projet de tableau de contrôle terrain pour l'initiative 50x30.

https://www.50x2030.org/

Sources de données possibles pour surveiller les enquêteurs

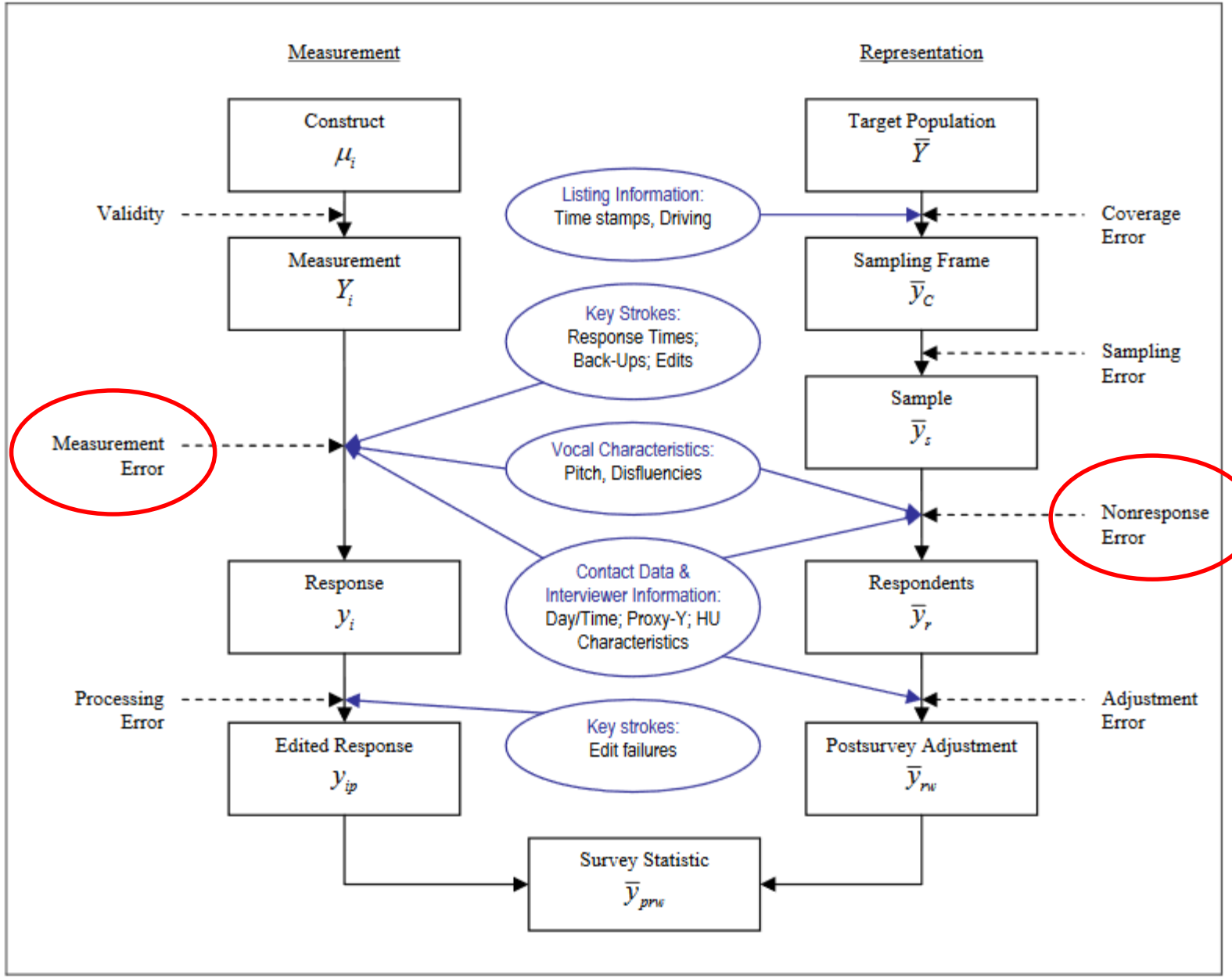
1. Coordonnées GPS
2. Entretiens de contrôle
3. Écoutes
4. Répondants mystère
5. Données substantielles
6. Paradonnées
7. Enregistrements audio (CARI)

6) Paradonnées

- Présentées au module 4 : données auxiliaires collectées au cours d'une enquête et décrivant le processus de collecte des données.

- Le logiciel d'enquête (par ex. Survey Solutions, BLAISE) enregistre les saisies sur le clavier effectuées par l'enquêteur au cours de l'entretien, ainsi que leur horodatage. Ces enregistrements sont ensuite traités pour analyse.

- Les paradonnées seront traitées plus en détail dans la vidéo 5 ; Kreuter (2013) donne beaucoup de détails.



Utilisations des paradonnées pour différentes composantes du cadre d'erreur totale d'enquête.

Source : *Kreuter et Casas-Cordero (2010)*

Sources de données possibles pour surveiller les enquêteurs

1. Coordonnées GPS
2. Entretiens de contrôle
3. Écoutes
4. Répondants mystère
5. Données substantielles
6. Paradonnées
7. Enregistrements audio (CARI)

7) Entretien audio enregistré assisté par ordinateur (CARI)

- Les entretiens assistés par ordinateur modernes permettent d'effectuer des enregistrements numériques des entretiens ou de segments d'entretiens.
- Discret.
- On peut s'appuyer sur les données/paradonnées substantielles pour sélectionner les cas à écouter.
- Informe sur l'interaction entre l'enquêteur et le répondant en direct.
- La vidéo 5 traitera l'utilisation de CARI plus en détails.

FIN DE LA VIDÉO 4