

Séminaire en ligne sur les sondages - Groupe enquêtes - SFdS

# Statistique robuste et sondages

Anne Ruiz-Gazen

Jeudi 14 Septembre 2023

# Introduction

## Statistique robuste

Etude du comportement des méthodes statistiques en présence d'**observations atypiques**<sup>1</sup>.

## Sondages

Ensemble de méthodes pour mise en œuvre d'**enquêtes** et traitement de données issues d'enquêtes.

Quelques réflexions sur articulation entre les deux domaines et apport de la statistique robuste aux sondages.

---

1. observations avec comportement différent de majorité des données.

# Introduction

## Mon parcours recherche académique

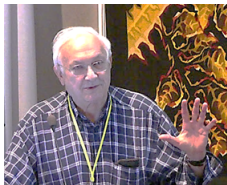
- **Statistique robuste** depuis 1989. Détection d'anomalies pour données complexes (multivariées, fonctionnelles, composition).
- **Sondages** depuis 1999. Plans complexes, estimation de variance, estimation robuste.



Y. Aragon



C. Goga



J.-C. Deville



D. Haziza



J.-F. Beaumont

# Plan de la présentation

1. Introduction
2. Statistique robuste
3. Statistique robuste et sondages
4. Conclusion

## 1. Introduction

## 2. Statistique robuste

- Introduction et objectifs
- Modèle contaminé et mesures de robustesse
- Grandes familles d'estimateurs robustes
- Détection d'observations atypiques
- Synthèse

## 3. Statistique robuste et sondages

## 4. Conclusion

# Introduction à la statistique robuste

École dont les précurseurs sont J. Tukey, P. Huber, F. Hampel et V. Yohai (fin années 60, début années 70).

## Objectifs :

- Cadre théorique pour étudier le comportement des méthodes statistiques (estimateurs, tests) lorsque majorité données observées suit un modèle paramétrique tandis que minorité ne suit pas ce modèle<sup>2</sup>

⇒ **outils de mesure de robustesse.**

- La plupart des méthodes classiques étant peu robustes  
⇒ **méthodes robustes,**  
méthodes sur lesquelles les observations atypiques ont **peu d'influence.**

---

2. Dans ce contexte, on parle de **données contaminées** et d'**observations atypiques.**

# Introduction à la statistique robuste

La théorie (complexe) développée consiste à se placer au **voisinage modèle paramétrique supposé** et à étudier propriétés estimateurs (ou tests) dans ce voisinage.

Estimateurs définis comme fonctions de lois de probabilités et voisinage défini à partir de distance entre lois de probabilités.

Pour plus de détails, voir ouvrages Huber et Ronchetti 2009, Hampel et al. 1986 et Maronna et al. 2019, tous intitulés “Robust Statistics”.

# Introduction à la statistique robuste

Dans la suite, présentation version **empirique** (concepts définis à partir des données et pas sur lois de probabilités) et **exemple simple** du modèle Gaussien et du problème d'estimation de la moyenne.

## Exemple du modèle Gaussien de variance connue

Loi Normale de moyenne  $\theta$  et de variance 1 où on cherche à estimer la moyenne  $\theta$ .



# Introduction à la statistique robuste

Échantillon (i.i.d.)  $Y_1, \dots, Y_n$  de loi  $\mathcal{N}(\theta, 1)$ .

Moyenne et médiane : estimateurs de  $\theta$

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \quad \text{et} \quad \text{med}(Y) = \text{median}_{i=1, \dots, n}(Y_i)$$

Si données issues du modèle Gaussien, moyenne et médiane estimateurs sans biais de  $\theta$  et  $\bar{Y}$  est plus efficace que  $\text{med}(Y)$ .

Plus précisément, la variance de la moyenne est environ 2/3 de la variance asymptotique de la médiane.

**Que se passe-t-il si données contaminées par des observations atypiques ?**

## 1. Introduction

## 2. Statistique robuste

- Introduction et objectifs
- **Modèle contaminé et mesures de robustesse**
- Grandes familles d'estimateurs robustes
- Détection d'observations atypiques
- Synthèse

## 3. Statistique robuste et sondages

## 4. Conclusion

# Modèle contaminé

## Modèle de mélange

$$(1 - \varepsilon)\mathcal{N}(\theta, 1) + \varepsilon G \quad \text{avec } \varepsilon < 50\%$$

Mélange du modèle supposé en proportion  $1 - \varepsilon$  et d'une loi contaminante  $G$  en proportion  $\varepsilon$ .

L'objectif est toujours d'estimer  $\theta$ .

La statistique robuste cherche à mesurer impact contamination du modèle  $\mathcal{N}(\theta, 1)$  sur estimateur  $\hat{\theta}$  de  $\theta$ .

## Mesurer la robustesse d'un estimateur $\hat{\theta}$

Pour étudier l'influence de données contaminées sur un estimateur  $\hat{\theta}$ , calcul différence :

$$\hat{\theta}(\text{échantillon contaminé}) - \hat{\theta}(\text{échantillon non contaminé})$$

Une façon de contaminer un échantillon : remplacer dans l'échantillon observé  $y_1, \dots, y_n$ , l'observation  $y_n$  par une valeur arbitraire notée  $y^*$ .

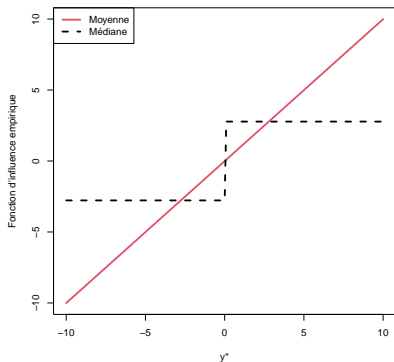
Après calcul différence précédente et standardisation, on obtient une fonction de  $y^*$  appelée **courbe de sensibilité** de Tukey ou **fonction d'influence empirique**.

Fonction d'influence empirique

$$\text{IF}_{\hat{\theta}}(y^*) = \frac{\hat{\theta}(y_1, \dots, y_{n-1}, y^*) - \hat{\theta}(y_1, \dots, y_n)}{\frac{1}{n}}.$$

## Exemple

On génère échantillon de loi  $\mathcal{N}(0, 1)$ , on remplace l'une des observations par  $y^*$  qui varie de -10 à 10.



Pour la moyenne,  $IF_{\bar{y}}(y^*) = y^* - \bar{y}$  pas bornée alors que pour la médiane,  $IF_{\text{med}(Y)}$  bornée. On dit que la médiane est **B-robuste**.

## Mesurer la robustesse d'un estimateur $\hat{\theta}$

Fonction d'influence empirique :

$$\frac{1}{n} \text{IF}_{\hat{\theta}}(y^*) = \hat{\theta}(y_1, \dots, y_{n-1}, y^*) - \hat{\theta}(y_1, \dots, y_n).$$

Au niveau de l'échantillon des variables aléatoires de loi  $F_{\theta}$ , on peut considérer l'espérance de la fonction d'influence empirique.

$$\frac{1}{n} \text{EIF}_{\hat{\theta}}(y^*) = \mathbf{E}(\hat{\theta}(Y_1, \dots, Y_{n-1}, y^*)) - \mathbf{E}(\hat{\theta}(Y_1, \dots, Y_n))$$

qui s'écrit

Espérance de la fonction d'influence empirique

$$\frac{1}{n} \text{EIF}_{\hat{\theta}}(y^*) = \mathbf{E}(\hat{\theta}(Y_1, \dots, Y_n) | Y_n = y^*) - \theta$$

si  $\hat{\theta}$  est un estimateur sans biais de  $\theta$ .

# Mesurer la robustesse d'un estimateur $\hat{\theta}$

Fonction d'influence : **contamination infinitésimale** des données (1 observation sur  $n$  soit  $\varepsilon = 1/n$  dans le modèle de mélange).

Autres mesures de robustesse : **courbe de biais asymptotique maximum** et **point de rupture** prenant en compte proportion  $0 < \varepsilon < 1/2$  de contamination qui ne tend pas vers 0 quand taille de l'échantillon tend vers l'infini et en considérant une loi contaminante ( $G$ ) qui est la pire possible pour l'estimateur.

# Limitation moyenne et médiane

Pour modèle Gaussien,

- La moyenne est efficace mais pas robuste,
- La médiane est robuste mais peu efficace.

Existe-t-il des **estimateurs efficaces** si données sont **non-contaminées ET robustes à la contamination** ?



## 1. Introduction

## 2. Statistique robuste

- Introduction et objectifs
- Modèle contaminé et mesures de robustesse
- **Grandes familles d'estimateurs robustes**
- Détection d'observations atypiques
- Synthèse

## 3. Statistique robuste et sondages

## 4. Conclusion

## Deux grandes familles

- **Les M-estimateurs** : “M” pour Maximum de vraisemblance (cas particulier).
  - La moyenne  $\bar{Y}$  minimise  $\sum_{i=1}^n (Y_i - \theta)^2$ .
  - La médiane  $\text{med}(Y)$  minimise  $\sum_{i=1}^n |Y_i - \theta|$ .
  - Un M-estimateur obtenu en minimisant  $\sum_{i=1}^n \rho(Y_i - \theta)$  pour une certaine fonction  $\rho$ .
- **Les L-estimateurs** : “L” pour Linéaire.  
Combinaisons linéaires de statistique d'ordre.
  - La médiane et la moyenne.
  - La moyenne tronquée.
  - La moyenne winsorisée.

## M-estimateurs

Pour une fonction  $\rho$  donnée et un échantillon  $Y_1, \dots, Y_n$ , le M-estimateur  $\hat{\theta}_n$  minimise

$$\sum_{i=1}^n \rho(Y_i - \theta) \quad \text{en } \theta.$$

Si  $\rho$  dérivable et  $\psi = \rho'$ ,  $\hat{\theta}_n$  est solution de

$$\sum_{i=1}^n \psi(Y_i - \theta) = 0$$

Se réécrit

$$\sum_{i=1}^n w_i (Y_i - \theta) = 0 \quad \text{avec} \quad w_i = \frac{\psi(Y_i - \theta)}{Y_i - \theta},$$

et conduit à

$$\theta = \frac{\sum_{i=1}^n w_i Y_i}{\sum_{i=1}^n w_i}.$$

# M-estimateurs

## Définition

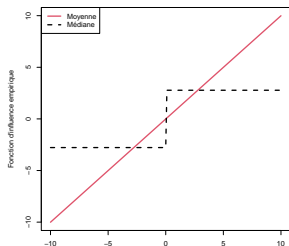
$$\hat{\theta}_n = \frac{\sum_{i=1}^n \hat{w}_i Y_i}{\sum_{i=1}^n \hat{w}_i} \quad \text{avec} \quad \hat{w}_i = \frac{\psi(Y_i - \hat{\theta}_n)}{Y_i - \hat{\theta}_n},$$

## Propriété

La fonction d'influence d'un M-estimateur est proportionnelle à  $\psi$ .

Pour  $\bar{Y}$ ,  $\rho(x) = x^2$ .

Pour  $\text{med}(Y)$ ,  $\rho(x) = |x|$ .



# M-estimateurs

## M-estimateur de Huber

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & \text{si } |x| \leq c \\ c|x| - \frac{1}{2}c^2 & \text{si } |x| > c \end{cases}$$

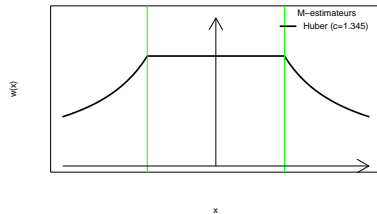
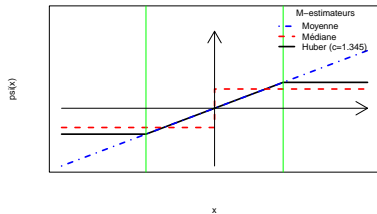
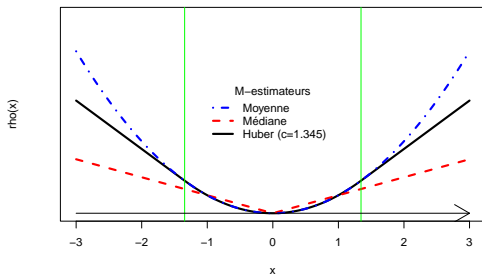
Estimateur B-robuste.

Si  $c = 1.345$ , 5% perte efficacité par rapport à la moyenne en utilisant M-estimateur de Huber pour modèle Gaussien.

### Remarque :

- Le M-estimateur qui minimise le biais asymptotique maximum est la médiane.
- Le M-estimateur qui minimise la **variance** asymptotique maximum est l'estimateur de Huber.

# M-estimateurs



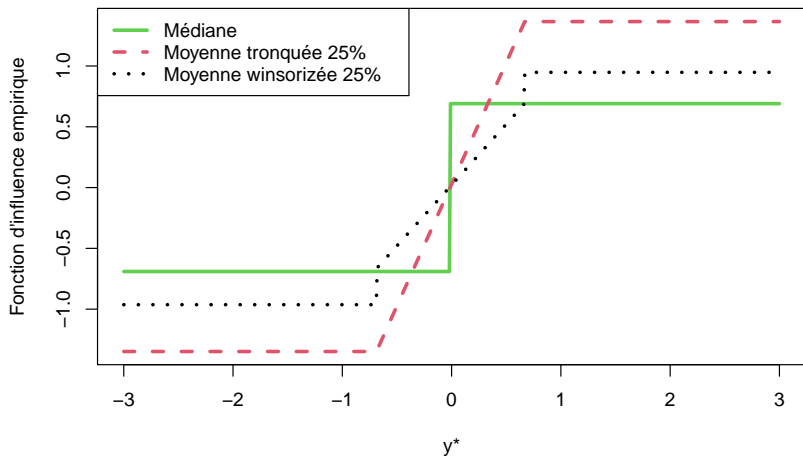
# L-estimateurs

Combinaisons linéaires de statistique d'ordre. On ordonne les observations de la plus petite à la plus grande. Soient  $0 \leq \alpha < 0.5$  et  $m = \lfloor (n - 1)\alpha \rfloor$ .

- **Moyenne tronquée** : on enlève les  $m$  observations les plus petites et les  $m$  observations les plus grandes et on calcule la moyenne des  $n - 2m$  observations restantes.
- **Moyenne winsorisée** : on remplace les  $m$  observations les plus petites par l'observations  $(m + 1)$ ième et les  $m$  observations les plus grandes par l'observation  $(n - m)$ ième.

# L-estimateurs

Dès que  $\alpha > 0$ , moyennes tronquées et winsorisées sont B-robustes.





# L-estimateurs

**Remarque** : Courbe de biais asymptotique maximum étudiées pour les moyennes tronquées mais pas pour les moyennes winsorisées.

Moyennes tronquées davantage étudiées car en fait variance asymptotique d'une moyenne tronquée est une variance winsorisée.

Pour le modèle Gaussien, la perte d'efficacité est de 17% pour une moyenne tronquée à 25% (soit 50% d'observations pas prises en compte dans la moyenne) et de 5% pour une moyenne tronquée à 8.5%.

## 1. Introduction

## 2. Statistique robuste

- Introduction et objectifs
- Modèle contaminé et mesures de robustesse
- Grandes familles d'estimateurs robustes
- **Détection d'observations atypiques**
- Synthèse

## 3. Statistique robuste et sondages

## 4. Conclusion

# Détection d'observations atypiques

En statistique robuste, détection d'anomalies souvent considérée comme “sous-produit” de l'estimation robuste.

**Exemple** : M-estimateur dans modèle Gaussien, observations associées à des poids faibles peuvent être considérées comme atypiques.

Des règles de détection d'anomalie qui n'utilisent pas d'estimateur robuste sont considérées comme “dangereuses” car elles peuvent souffrir de l'**effet de masque**.

**Exemple** : règle du  $k$ -sigma qui consiste à décider qu'une observation est atypique si elle est à plus de  $k$  écarts-type de la moyenne peut souffrir de cet effet.

## Petit exemple d'effet de masque

14 observations et 2 observations atypiques

-0.9, 1.1, 1.3, 1.4, 1.9, 2.0, 2.5, 2.6, 2.8, 2.9, 3.0, 3.1,  
50.9, 99.1

Règle des 3-sigma avec la moyenne et l'écart-type :

0.41, 0.41, 0.40, 0.40, 0.38, 0.37, 0.36, 0.35, 0.35, 0.34, 0.34, 0.34,  
1.36, 3.08

Seule la dernière observation est déclarée atypique. Avec la médiane et la médiane de l'écart absolu à la médiane :

1.86, 1.63, 1.41, 1.29, 0.73, 0.62, 0.06, 0.06, 0.28, 0.39, 0.51, 0.62,  
54.35, 108.54

## 1. Introduction

## 2. Statistique robuste

- Introduction et objectifs
- Modèle contaminé et mesures de robustesse
- Grandes familles d'estimateurs robustes
- Détection d'observations atypiques
- Synthèse

## 3. Statistique robuste et sondages

## 4. Conclusion

# Synthèse

La statistique robuste classique permet de :

- définir des outils pour mesurer la robustesse de méthodes statistiques au voisinage d'un modèle paramétrique,
- définir de nouveaux estimateurs robustes et étudier leurs propriétés,
- définir des méthodes de détection d'anomalies.

Les méthodes préconisées consistent en général à **pondérer plus faiblement** certaines observations de manière à **ne pas perdre en efficacité** tout en étant **peu influencées** par observations atypiques.

## 1. Introduction

## 2. Statistique robuste

## 3. Statistique robuste et sondages

- Introduction
- Contexte des sondages
- Adaptations de la fonction d'influence
- Estimateurs robustes

## 4. Conclusion

# Adaptations à prendre en compte

- Détection des anomalies en amont de l'estimation
- Population finie vs. population infinie
- Plan de sondage avec poids de sondages à prendre en compte vs. pas de plan de sondage.

Adaptation difficile et souvent dans des cas particuliers .



## Chercheurs intéressés par les deux domaines

- R. Chambers,
- B. Hülliger (ancien doctorant de F. Hampel),
- M. Hidioglou, L.-P. Rivest,
- J.-F. Beaumont, D. Haziza, C. Favre-Martinoz, T. Deroyon,
- E. Ronchetti, A. Welsh,
- M. Templ, V. Todorov, P. Filzmoser, A. Alfons,
- M. Salibian-Barreira,
- ...

Adaptation difficile et souvent dans des cas particuliers

## 1. Introduction

## 2. Statistique robuste

## 3. Statistique robuste et sondages

- Introduction
- **Contexte des sondages**
- Adaptations de la fonction d'influence
- Estimateurs robustes

## 4. Conclusion

# Contexte des sondages

Les offices et instituts nationaux doivent fournir :

- ① des données de qualité,
- ② des estimations fiables (ponctuelles ou par intervalles de confiance) de paramètres de population finie.

Concernant les observations atypiques, objectifs :

- ① détecter et corriger les données erronées,
- ② à partir des données corrigées, améliorer les méthodes d'estimation en réduisant l'influence de valeurs atypiques sur l'erreur quadratique moyenne des estimateurs.

# Contexte des sondages

- 1 La recherche et la correction de données erronées se déroulent au moment de l'étape de **vérification** (“editing”) des données, en général en amont de l'estimation (De Waal 2009).
- 2 Les observations **influentes (non erronées)** sont prises en compte au moment de l'estimation en remplaçant estimateurs classiques tel que Horvitz-Thompson par estimateurs robustes. (Beaumont et Rivest 2009).

Les méthodes statistiques pour chacune des 2 étapes sont différentes.

# Contexte des sondages

- 1 **Étape de vérification** : identification champs erronés puis remplacement par valeurs imputées (correctes ou “moins erronées”).

Très nombreuses méthodes d'editing dont beaucoup sont basées sur des règles d'identifications (“edits”) déterministes.

Statistique robuste classique et surtout **méthodes de détection d'anomalies** qui en découlent jouent un rôle lorsqu'on s'intéresse à des observations non détectées par règles d'identification mais qui diffèrent des autres observations.

## Contexte des sondages

- ② **Étape d'estimation** : on suppose données corrigées et donc modèle de contamination de la statistique robuste n'est plus pertinent.

Si on s'intéresse à l'estimation d'un total, pas question de pondérer faiblement certaines observations sous prétexte qu'elles prennent de grandes valeurs car introduction d'un biais.

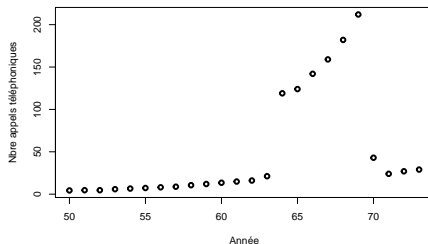
L'objectif reste d'estimer le total de l'**ensemble** de la population (pas le total d'une partie "non-contaminée").

**Exemple** : dans enquêtes entreprises, problème de l'impact des grandes valeurs sur l'erreur d'échantillonnage pris en compte au niveau du plan de sondages (stratifié avec une strate recensement). Mais parfois observations pas allouées à la bonne strate ("saut de strates").

# Contexte des sondages

En statistique robuste classique on ne distingue pas ces 2 étapes.

**Exemple** : Rousseeuw et Leroy 1987



Problème d'erreur dans les unités de mesure bien connu et traité par l'editing en sondages.

# Atypiques représentatifs et non-représentatifs

**Remarque** : Chambers 1986 distingue

- les observations atypiques représentatives : observations de l'échantillon qui **représentent** des observations qui ne sont pas dans l'échantillon, ou qui ne sont pas uniques.
- les observations atypiques non-représentatives : autres observations.

On retrouve la dichotomie précédente.

- Atypiques **non représentatifs détectés et imputés** au moment de la **vérification** (editing) des données.
- Atypiques **représentatifs détectés et influence limitée** au moment de l'estimation des paramètres.



## 1. Introduction

## 2. Statistique robuste

## 3. Statistique robuste et sondages

- Introduction
- Contexte des sondages
- Adaptations de la fonction d'influence
- Estimateurs robustes

## 4. Conclusion

# Plusieurs adaptations de la fonction d'influence

- **Approche de Gwet et Rivest 1992.** Cas particulier d'un estimateur par le ratio et d'un plan aléatoire simple sans remise. Fonction d'influence basée sur régression en statistique classique et utilisée pour dériver variance asymptotique.
- **Approche de Hulliger 1995.** Cas particulier d'un plan proportionnel à la taille d'une variable auxiliaire. Paramètre de population finie exprimé comme une fonction de la variable auxiliaire, estimateur Horvitz-Thompson exprimé comme un estimateur moindres carrés pondérés.

## Plusieurs adaptations de la fonction d'influence

- **Approche de Deville 1999.** Linéarisation d'estimateurs complexes pour paramètres exprimés comme fonctionnelles d'une mesure sur population finie. Mesure contaminée par une Dirac en proportion epsilon. La contamination ne prend pas en compte le plan de sondages. Voir aussi Goga, Deville et Ruiz-Gazen 2009.
- **Approche du biais conditionnel :** Moreno-Rebollo, Muñoz-Reyes et Muñoz-Pichardo 1999 et Beaumont, Haziza et Ruiz-Gazen 2013.

## Contexte d'intérêt

**Population finie**  $U = 1, \dots, k, \dots, N$ , variable d'intérêt  $y (> 0)$ ,  
paramètres d'intérêt : total inconnu

$$t_y = \sum_{i \in U} y_i.$$

Inférence **basée sur le plan de sondage** (probabiliste  $p$ ). Estimateur  
de Horvitz-Thompson de  $t_y$  :

$$\hat{t}_\pi = \sum_{i \in s} d_i y_i,$$

où  $d_i = 1/\pi_i$  poids de sondage et  $\pi_i > 0$  probabilité d'inclusion du  
premier ordre ( $\pi_{ij}$  deuxième ordre).

## Contexte d'intérêt

On est à l'étape de l'estimation (erreurs corrigées) et on cherche

- à **mesurer l'influence des observations sur estimateur Horvitz-Thompson,**
- à **proposer des estimateurs robustes.**

Méthodes "inspirées" de la statistique robuste mais très différentes.

# Adaptation de la fonction d'influence

En population infinie :

Espérance de la fonction d'influence empirique

$$\frac{1}{n} \text{EIF}_{\hat{\theta}}(y_i) = \mathbf{E}(\hat{\theta}(Y_1, \dots, Y_n) | Y_n = y_i) - \theta$$

si  $\hat{\theta}$  est un estimateur sans biais de  $\theta$ .

En sondages :

Biais conditionnel

$$B_i^{\hat{\theta}} = \mathbf{E}_p(\hat{\theta}(s) | i \in s) - \theta.$$

Moreno-Rebollo, Muñoz-Reyes et Muñoz-Pichardo 1999

# Propriétés biais conditionnel

## Biais conditionnel de l'estimateur Horvitz-Thompson

$$B_i^\pi = \mathbf{E}_p(\hat{t}_\pi | i \in s) - t_y = (d_i - 1)y_i + \sum_{\substack{j \in U \\ j \neq i}} \left( \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) y_j$$

- Si  $\pi_i = 1$  alors  $B_i^\pi = 0$ .
- **Plan de Poisson :**

$$B_i^\pi = (d_i - 1)y_i,$$

calculable.

- **Plan aléatoire simple sans remise :**

$$B_i^\pi = \frac{N}{N-1}(d_i - 1)(y_i - \bar{Y}) \text{ avec } \bar{Y} = t_y/N,$$

$$\text{estimé par } \hat{B}_i^\pi = \frac{N}{N-1}(d_i - 1)(y_i - \bar{y}_\pi).$$

- **Plan stratifié aléatoire simple sans remise :**  $U = \cup_{h=1}^H U_h$ ,  
strate  $U_h$  de taille  $N_h$  et échantillon de taille  $n_h$ ,  
 $t_{yh} = \sum_{i \in U_h} y_i$ .

$$B_i^\pi = \frac{N_h}{N_h - 1}(d_i - 1)(y_i - \bar{Y}_h) \text{ avec } \bar{Y}_h = t_{yh}/N_h \text{ et } i \in U_h,$$

$$\text{estimé par } \hat{B}_i^\pi = \frac{n_h}{n_h - 1}(d_i - 1)(y_i - \bar{y}_h) \text{ avec } \bar{y}_h = \sum_{i \in s_h} y_i/n_h.$$

- **Plans complexes :** estimation par bootstrap (Beaumont, Bocci et St-Louis 2021).



## 1. Introduction

## 2. Statistique robuste

## 3. Statistique robuste et sondages

- Introduction
- Contexte des sondages
- Adaptations de la fonction d'influence
- Estimateurs robustes

## 4. Conclusion

# Estimateurs robustes

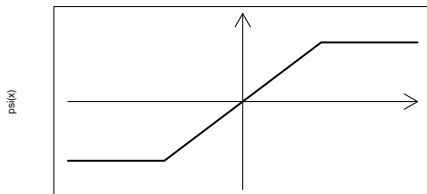
$\hat{t}_\pi$  estimateur Horvitz-Thompson de biais conditionnel  $\hat{B}_i^\pi$ .

Beaumont, Haziza et Ruiz-Gazen 2013

$$\hat{t}_\pi^R = \hat{t}_\pi - \sum_s \hat{B}_i^\pi + \sum_s \psi(\hat{B}_i^\pi)$$

avec fonction  $\psi$  de Huber

$$\psi(x) = \begin{cases} c & \text{si } x > c \\ x & \text{si } |x| \leq c \\ -c & \text{si } x < -c \end{cases}$$



La constante  $c$  peut être précisée en utilisant un argument minimax.

## Estimateurs robustes

$\hat{t}_\pi$  estimateur Horvitz-Thompson de biais conditionnel  $\hat{B}_i^\pi$ .

Estimateur minimax (Beaumont, Haziza et Ruiz-Gazen 2013)

$$\hat{t}_\pi^R = \hat{t}_\pi - \gamma_s$$

avec  $\gamma_s$  qui minimise

$$\max\{|\hat{B}_i^R|; i \in s\} \text{ où } \hat{B}_i^R \text{ biais conditionnel de } \hat{t}_\pi^R.$$

Formule estimateur minimax

$$\hat{t}_\pi^R = \hat{t}_\pi - \frac{1}{2} \left( \hat{B}_{\min}^\pi + \hat{B}_{\max}^\pi \right)$$

avec  $\hat{B}_{\min}^\pi = \min(B_i^\pi; i \in s)$  et  $\hat{B}_{\max}^\pi = \max(B_i^\pi; i \in s)$

## Lien avec la winsorization

Favre-Martinoz, Haziza et Beaumont 2015 montrent que l'estimateur minimax peut s'écrire comme un estimateur winsorisé standard mais aussi comme un estimateur winsorisé de type Dalén-Tambay.

Deroyon et Favre-Martinoz 2018 comparent l'estimateur minimax à l'estimateur de Kokic et Bell (Kokic et Bell 1994)

1. Introduction
2. Statistique robuste
3. Statistique robuste et sondages
4. Conclusion

# Synthèse

- La statistique robuste en population infinie utilise les concepts de fonctions d'influence (bornée) et de courbe de biais (point de rupture élevé).
- En population finie, avec inférence sous le plan et en présence d'atypiques représentatifs, fonction d'influence généralisée au biais conditionnel.
- La statistique robuste en population infinie s'intéresse aux classes de M et L estimateurs.
- En population finie, certains concepts sont repris comme la winsorization et la fonction de Huber mais estimateurs très différents. Dérivation d'estimateurs robustes en modifiant estimateur Horvitz-Thompson pour diminuer le biais conditionnel maximum par exemple.

# Nombreux autres développements








## Autour du biais conditionnel :

- Paramètres non-linéaires (Beaumont, Bocci et St-Louis 2021 )
- Estimation sur domaines (Favre-Martinoz, Haziza et Beaumont 2021)
- Données fonctionnelles (Cardot, De Moliner et Goga 2020)
- Plans en deux phases avec application à la non-réponse (Favre-Martinoz, Haziza et J.-F. Beaumont 2016)
- Estimation sur petits domaines (Jiongo, Haziza et Duchesne 2013)
- ...









Merci de votre attention !









# Littérature I

-  Beaumont, J-F, C. Bocci et M. St-Louis (2021). « Bootstrap Estimation of the Conditional Bias for Measuring Influence in Complex Surveys ». In : *Journal of Survey Statistics and Methodology*.
-  Beaumont, J-F, D. Haziza et A. Ruiz-Gazen (2013). « A unified approach to robust estimation in finite population sampling ». In : *Biometrika* 100.3, p. 555-569.
-  Beaumont, J-F et L-P Rivest (2009). « Dealing with outliers in survey data ». In : *Handbook of Statistics*. T. 29. Elsevier, p. 247-279.
-  Cardot, H., A. De Moliner et C. Goga (2020). « Conditional bias robust estimation of the total of curve data by sampling in a finite population : an illustration on electricity load curves ». In : *Journal of Survey Statistics and Methodology* 8.3, p. 453-482.
-  Chambers, R.L. (1986). « Outlier robust finite population estimation ». In : *Journal of the American Statistical Association* 81.396, p. 1063-1069.
-  De Waal, T. (2009). « Statistical data editing ». In : *Handbook of Statistics*. T. 29. Elsevier, p. 187-214.
-  Deroyon, T. et C. Favre-Martinoz (2018). « Comparison of the conditional bias and Kobic and Bell methods for Poisson and stratified sampling ». In : *Survey Methodology* 44.2, p. 309-338.

# Littérature II

-  Deville, J-C (1999). « Variance estimation for complex statistics and estimators : Linearization and residual techniques ». In : *Survey Methodology* 25.2, p. 193-203.
-  Favre-Martinoz, C., D. Haziza et J-F Beaumont (2015). « A method of determining the winsorization threshold, with an application to domain estimation ». In : *Survey Methodology* 41.1, p. 57-77.
-  — (2021). « Efficient nonparametric estimation for skewed distributions ». In : *Canadian Journal of Statistics* 49.2, p. 471-496.
-  — (2016). « Robust Inference in Two-Phase Sampling Designs with Application to Unit Nonresponse ». In : *Scandinavian Journal of Statistics* 43.4, p. 1019-1034.
-  Goga, C., J-C Deville et A. Ruiz-Gazen (2009). « Use of functionals in linearization and composite estimation with application to two-sample survey data ». In : *Biometrika* 96.3, p. 691-709.
-  Gwet, J-P et L-P Rivest (1992). « Outlier resistant alternatives to the ratio estimator ». In : *Journal of the American Statistical Association* 87.420, p. 1174-1182.
-  Hampel, F.R et al. (1986). *Robust Statistics : The Approach Based on Influence Functions*. Wiley Series in Probability and Statistics. Wiley.
-  Huber, P.J. et E.M. Ronchetti (2009). *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley.

# Littérature III

-  Hulliger, B. (1995). « Outlier robust Horvitz-Thompson estimators ». In : *Survey Methodology* 21.1, p. 79-87.
-  Jiongo, V.D., D. Haziza et P. Duchesne (2013). « Controlling the bias of robust small-area estimators ». In : *Biometrika* 100.4, p. 843-858.
-  Kokic, PN et PA Bell (1994). « Optimal winsorizing cutoffs for a stratified finite population estimator ». In : *Journal of Official Statistics* 10, p. 419-419.
-  Maronna, R.A. et al. (2019). *Robust Statistics : Theory and Methods (with R)*. Wiley Series in Probability and Statistics. Wiley.
-  Moreno-Rebollo, J.L., A. Muñoz-Reyes et J. Muñoz-Pichardo (1999). « Miscellanea. Influence diagnostic in survey sampling : conditional bias ». In : *Biometrika* 86.4, p. 923-928.
-  Rousseeuw, P.J. et A. Leroy (1987). *Robust Regression and Outlier Detection*. Wiley Series in Probability and Statistics. Wiley.