Processus empiriques dans le cadre des sondages

E. Chautru • P. Bertail • S. Clémençon

Télécom ParisTech, UMR CNRS 5141 LTCI, Groupe TSI
Unité Met@risk, INRA
ANSES
Modal'X, Université Paris X
Laboratoire de Statistique du CREST

7^e colloque francophone sur les sondages • 5 novembre 2012

Cadre Général – De la population à l'échantillon

Population

- $\mathcal{P}_N := \{1, \ldots, N\}$ de taille N

Cadre Général – De la population à l'échantillon

Population

- $\mathcal{P}_N := \{1, \ldots, N\}$ de taille N ➡ caractéristiques ?

Cadre Général – De la population à l'échantillon

Population

- $\mathcal{P}_N := \{1, \ldots, N\}$ de taille N ➡ caractéristiques ?
- Variable d'intérêt X : $\begin{pmatrix} \mathcal{P}_N & \longrightarrow & \mathcal{X} \\ i & \longmapsto & X_i \end{pmatrix}$

Cadre Général – De la population à l'échantillon

Population

- $\mathcal{P}_N := \{1, \ldots, N\}$ de taille N ➡ caractéristiques ?
- Variable d'intérêt X : $\begin{pmatrix} \mathcal{P}_N & \longrightarrow & \mathcal{X} \\ i & \longmapsto & X_i \end{pmatrix}$ ➡ évaluer $\overline{X}$, etc.

Cadre Général – De la population à l'échantillon

Population

- $\mathcal{P}_N := \{1, \ldots, N\}$ de taille N ➡ caractéristiques ?
- Variable d'intérêt X : $\begin{pmatrix} \mathcal{P}_N & \longrightarrow & \mathcal{X} \\ i & \longmapsto & X_i \end{pmatrix}$ ➡ évaluer $\overline{X}$, etc.

Modèle de surpopulation

- Hypothèse : X v.a. à valeurs dans $(\mathcal{X}, \|.\|)$ Banach de tribu engendrée $\mathcal{B}_{\mathcal{X}}$ et $X_1, \ldots, X_N \overset{\text{i.i.d.}}{\rightsquigarrow} \mathbf{P} \equiv \mathbf{P}_X$

Cadre Général – De la population à l'échantillon

Population

- $\mathcal{P}_N := \{1, \ldots, N\}$ de taille N ➡ caractéristiques ?
- Variable d'intérêt X : $\begin{pmatrix} \mathcal{P}_N & \longrightarrow & \mathcal{X} \\ i & \longmapsto & X_i \end{pmatrix}$ ➡ évaluer $\overline{X}$, etc.

Modèle de surpopulation

- Hypothèse : X v.a. à valeurs dans $(\mathcal{X}, \|.\|)$ Banach de tribu engendrée $\mathcal{B}_{\mathcal{X}}$ et $X_1, \ldots, X_N \overset{\text{i.i.d.}}{\rightsquigarrow} \mathbf{P} \equiv \mathbf{P}_X$
- Mesure empirique $\mathbf{P}_N := \frac{1}{N} \sum_{i=1}^{N} \delta_{X_i}$ ➡ $\overline{X} = \int_{\mathcal{X}} x \, \mathbf{P}_N(dx)$, etc.

Cadre Général – De la population à l'échantillon

Population

- $\mathcal{P}_N := \{1, \ldots, N\}$ de taille N ➡ caractéristiques ?
- Variable d'intérêt X : $\begin{pmatrix} \mathcal{P}_N & \longrightarrow & \mathcal{X} \\ i & \longmapsto & X_i \end{pmatrix}$ ➡ évaluer $\overline{X}$, etc.

Modèle de surpopulation

- Hypothèse : X v.a. à valeurs dans $(\mathcal{X}, \|.\|)$ Banach de tribu engendrée $\mathcal{B}_{\mathcal{X}}$ et $X_1, \ldots, X_N \overset{\text{i.i.d.}}{\rightsquigarrow} \mathbf{P} \equiv \mathbf{P}_X$
- Mesure empirique $\mathbf{P}_N := \frac{1}{N} \sum_{i=1}^{N} \delta_{X_i}$ ➡ $\overline{X} = \int_{\mathcal{X}} x \, \mathbf{P}_N(dx)$, etc.

Problème

On n'a pas accès à tout $\mathcal{P}_N$ (coût, accessibilité, etc.)

Cadre Général – De la population à l'échantillon

Population

- $\mathcal{P}_N := \{1, \ldots, N\}$ de taille N ➡ caractéristiques ?
- Variable d'intérêt X : $\begin{pmatrix} \mathcal{P}_N & \longrightarrow & \mathcal{X} \\ i & \longmapsto & x_i \end{pmatrix}$ ➡ évaluer $\overline{X}$, etc.

Modèle de surpopulation

- Hypothèse : X v.a. à valeurs dans $(\mathcal{X}, \|.\|)$ Banach de tribu engendrée $\mathcal{B}_{\mathcal{X}}$ et $X_1, \ldots, X_N \overset{\text{i.i.d.}}{\rightsquigarrow} \mathbf{P} \equiv \mathbf{P}_X$
- Mesure empirique $\mathbf{P}_N := \frac{1}{N} \sum_{i=1}^{N} \delta_{X_i}$ ➡ $\overline{X} = \int_{\mathcal{X}} x \, \mathbf{P}_N(dx)$, etc.

Problème

On n'a pas accès à tout $\mathcal{P}_N$ (coût, accessibilité, etc.) ➡ échantillon

Cadre Général – De la population à l'échantillon

Echantillon

- $\mathcal{S} \subsetneq \mathcal{P}_N$ de taille $n < N$

Cadre Général – De la population à l'échantillon

Echantillon

- $\mathcal{S} \subsetneq \mathcal{P}_N$ de taille $n < N$ ➡ tirage aléatoire

Cadre Général – De la population à l'échantillon

Echantillon

- $\mathcal{S} \subsetneq \mathcal{P}_N$ de taille $n < N$ ➡ tirage aléatoire
- Variables d'inclusion : $\epsilon_i := \mathbb{I}\{i \in \mathcal{S}\}$ $\qquad i \in \mathcal{P}_N$

Cadre Général – De la population à l'échantillon

Echantillon

- $\mathcal{S} \subsetneq \mathcal{P}_N$ de taille $n < N$ ➡ tirage aléatoire
- Variables d'inclusion : $\epsilon_i := \mathbb{I}\{i \in \mathcal{S}\}$ $\quad i \in \mathcal{P}_N$
- Probabilités d'inclusion : $\pi_i := \mathbb{P}(\epsilon_i = 1) = \mathbb{E}(\epsilon_i)$ $\quad i \in \mathcal{P}_N$

Cadre Général – De la population à l'échantillon

Echantillon

- $\mathcal{S} \subsetneq \mathcal{P}_N$ de taille $n < N$ ➡ tirage aléatoire
- Variables d'inclusion : $\epsilon_i := \mathbb{I}\{i \in \mathcal{S}\}$ $\quad i \in \mathcal{P}_N$
- Probabilités d'inclusion : $\pi_i := \mathbb{P}(\epsilon_i = 1) = \mathbb{E}(\epsilon_i)$ $\quad i \in \mathcal{P}_N$
- Au second ordre : $\pi_{i,j} := \mathbb{P}(\epsilon_i = 1, \epsilon_j = 1) = \mathbb{E}(\epsilon_i \epsilon_j)$ $\quad (i,j) \in \mathcal{P}_N^2$

Cadre Général – De la population à l'échantillon

Echantillon

- $\mathcal{S} \subsetneq \mathcal{P}_N$ de taille $n < N$ ➡ tirage aléatoire
- Variables d'inclusion : $\epsilon_i := \mathbb{I}\{i \in \mathcal{S}\}$ $\quad i \in \mathcal{P}_N$
- Probabilités d'inclusion : $\pi_i := \mathbb{P}(\epsilon_i = 1) = \mathbb{E}(\epsilon_i)$ $\quad i \in \mathcal{P}_N$
- Au second ordre : $\pi_{i,j} := \mathbb{P}(\epsilon_i = 1, \epsilon_j = 1) = \mathbb{E}(\epsilon_i \epsilon_j)$ $\quad (i,j) \in \mathcal{P}_N^2$

Comment tirer les unités ?

Plan de sondage := loi $\mathbf{p}$ sur $\mathcal{S} \equiv (\epsilon_1, \ldots, \epsilon_N)$

Cadre Général – De la population à l'échantillon

Echantillon

- $\mathcal{S} \subsetneq \mathcal{P}_N$ de taille $n < N$ ➡ tirage aléatoire
- Variables d'inclusion : $\epsilon_i := \mathbb{I}\{i \in \mathcal{S}\}$ $\quad i \in \mathcal{P}_N$
- Probabilités d'inclusion : $\pi_i := \mathbb{P}(\epsilon_i = 1) = \mathbb{E}(\epsilon_i)$ $\quad i \in \mathcal{P}_N$
- Au second ordre : $\pi_{i,j} := \mathbb{P}(\epsilon_i = 1, \epsilon_j = 1) = \mathbb{E}(\epsilon_i \epsilon_j)$ $\quad (i,j) \in \mathcal{P}_N^2$

Comment tirer les unités ?

Plan de sondage := loi $\mathbf{p}$ sur $\mathcal{S} \equiv (\epsilon_1, \ldots, \epsilon_N)$

- SASSR : n fixé et $\mathbb{P}(\mathcal{S} = s) = \mathbb{I}\{\#s = n\}/C_N^n$
- BERN : n aléatoire, $\pi_1 = \ldots = \pi_N = \pi$, et $\mathbb{P}(\mathcal{S} = s) = \pi^n(1-\pi)^{N-n}$

Cadre Général – De l'échantillon à la population

Estimation à partir de l'échantillon

Mesure empirique Horvitz-Thompson : $\mathbf{P}_N^{\pi} := \frac{1}{N} \sum_{i=1}^{N} \frac{\epsilon_i}{\pi_i} \delta_{X_i}$

Cadre Général – De l'échantillon à la population

Estimation à partir de l'échantillon

Mesure empirique Horvitz-Thompson : $\mathbf{P}_N^\pi := \frac{1}{N}\sum_{i=1}^{N} \frac{\epsilon_i}{\pi_i}\delta_{X_i}$

- Estimateur sans biais de $\mathbf{P}_N \,\big|\, (X_1, \ldots, X_N)$

Cadre Général – De l'échantillon à la population

Estimation à partir de l'échantillon

Mesure empirique Horvitz-Thompson : $\mathbf{P}_N^\pi := \frac{1}{N}\sum_{i=1}^{N} \frac{\epsilon_i}{\pi_i}\delta_{X_i}$

- Estimateur sans biais de $\mathbf{P}_N \mid (X_1, \ldots, X_N)$
- Consistence et normalité asymptotiques ($N \to +\infty$) ponctuelles (Hájek, 1964 - Berger, 1998 - Robinson, 1982)

Cadre Général – De l'échantillon à la population

Estimation à partir de l'échantillon

Mesure empirique Horvitz-Thompson : $\mathbf{P}_N^\pi := \frac{1}{N}\sum_{i=1}^{N} \frac{\epsilon_i}{\pi_i}\delta_{X_i}$

- Estimateur sans biais de $\mathbf{P}_N \mid (X_1, \ldots, X_N)$
- Consistence et normalité asymptotiques ($N \to +\infty$) ponctuelles (Hájek, 1964 - Berger, 1998 - Robinson, 1982)
- Convergences fonctionnelles pour les plans stratifiés (Saegusa & Wellner, 2012)

Cadre Général – De l'échantillon à la population

Estimation à partir de l'échantillon

Mesure empirique Horvitz-Thompson : $\mathbf{P}_N^\pi := \frac{1}{N}\sum_{i=1}^{N} \frac{\epsilon_i}{\pi_i}\delta_{X_i}$

- Estimateur sans biais de $\mathbf{P}_N \mid (X_1, \ldots, X_N)$
- Consistence et normalité asymptotiques ($N \to +\infty$) ponctuelles (Hájek, 1964 - Berger, 1998 - Robinson, 1982)
- Convergences fonctionnelles pour les plans stratifiés (Saegusa & Wellner, 2012)

Notre Objectif

Convergence fonctionnelle pour d'autres plans de sondage ?

Plan de l'exposé

1. **Processus Empirique et plan Poissonnien**
   - Plan de sondage Poissonnien
   - Convergence vers un processus gaussien
   - Covariance du processus limite

2. **Du plan Poissonnien au plan Réjectif**
   - Plan de sondage Réjectif
   - Distance au plan Poissonnien
   - Résultat de convergence

Plan de l'exposé

Plan de sondage Poissonnien – Rappels

Définition

$$\epsilon_1 \perp \ldots \perp \epsilon_N \qquad \epsilon_i \rightsquigarrow \mathcal{B}(p_i) \qquad \mathbb{E}(n) = \sum_{i=1}^{N} p_i$$

$$T_N(s) := \mathbb{P}(\mathcal{S} = s) = \prod_{i \in s} p_i \prod_{i \notin s} (1 - p_i)$$

Plan de sondage Poissonnien – Rappels

Définition

$$\epsilon_1 \perp \ldots \perp \epsilon_N \qquad \epsilon_i \rightsquigarrow \mathcal{B}(p_i) \qquad \mathbb{E}(n) = \sum_{i=1}^{N} p_i$$

$$T_N(s) := \mathbb{P}(\mathcal{S} = s) = \prod_{i \in s} p_i \prod_{i \notin s} (1 - p_i)$$

Avantages

- Entièrement caractérisé par les probabilités d'inclusion du 1^{er} ordre
- Indépendance entre les ϵ_i ➡ cadre simple

Plan de sondage Poissonnien – Rappels

Définition

$$\epsilon_1 \perp \ldots \perp \epsilon_N \qquad \epsilon_i \rightsquigarrow \mathcal{B}(p_i) \qquad \mathbb{E}(n) = \sum_{i=1}^{N} p_i$$

$$T_N(s) := \mathbb{P}(\mathcal{S} = s) = \prod_{i \in s} p_i \prod_{i \notin s} (1 - p_i)$$

Avantages

- Entièrement caractérisé par les probabilités d'inclusion du 1^{er} ordre
- Indépendance entre les ϵ_i ➡ cadre simple
- p_i peut dépendre d'une v.a. auxiliaire $W \rightsquigarrow \mathbf{P}_W$ observée sur tout $\mathcal{P}_N$:

$$p_i(w) = \mathbb{E}\left(\epsilon_i \mid W_i = w\right) \equiv p(W_i)$$

Plan de sondage Poissonnien – Processus d'intérêt

Processus d'intérêt $(\mathbb{G}^p_{T_N} f)_{f\in\mathcal{F}}$

$$\mathbb{G}^p_{T_N} f := \sqrt{N}\int f(x)\,(\mathbf{P}^p_N - \mathbf{P}_N)\,(dx) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{\epsilon_i}{p_i} - 1\right) f(X_i)$$

Plan de sondage Poissonnien – Processus d'intérêt

Processus d'intérêt $(\mathbb{G}^p_{T_N} f)_{f\in\mathcal{F}}$

$$\mathbb{G}^p_{T_N} f := \sqrt{N}\int f(x)\,(\mathbf{P}^p_N - \mathbf{P}_N)\,(dx) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{\epsilon_i}{p_i} - 1\right) f(X_i)$$

Hypothèses fondamentales

$\mathcal{H}_1$ $\exists\,\lambda > 0,\ N_0 \in \mathbb{N}^\star$:
$\forall N \geq N_0,\ \forall i \in \mathcal{P}_N,\ p_i > \lambda$

Plan de sondage Poissonnien – Processus d'intérêt

Processus d'intérêt $(\mathbb{G}_{T_N}^p f)_{f\in\mathcal{F}}$

$$\mathbb{G}_{T_N}^p f := \sqrt{N}\int f(x)\,(\mathbf{P}_N^p - \mathbf{P}_N)\,(dx) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{\epsilon_i}{p_i} - 1\right) f(X_i)$$

Hypothèses fondamentales

$\mathcal{H}_1$ $\exists\,\lambda > 0,\ N_0 \in \mathbb{N}^\star$:
$\forall N \geq N_0,\ \forall i \in \mathcal{P}_N,\ p_i > \lambda$

$\mathcal{H}_2$ W non proportionnelle à X

Plan de sondage Poissonnien – Processus d'intérêt

Processus d'intérêt $(\mathbb{G}^p_{T_N} f)_{f\in\mathcal{F}}$

$$\mathbb{G}^p_{T_N} f := \sqrt{N}\int f(x)\,(\mathbf{P}^p_N - \mathbf{P}_N)\,(dx) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{\epsilon_i}{p_i} - 1\right) f(X_i)$$

Hypothèses fondamentales

$\mathcal{H}_1$ $\exists\,\lambda > 0,\ N_0 \in \mathbb{N}^\star$:
$\forall N \geq N_0,\ \forall i \in \mathcal{P}_N,\ p_i > \lambda$

$\mathcal{H}_2$ W non proportionnelle à X

$\mathcal{H}_3$ Faible échangeabilité de
$(\epsilon_i, W_i)_{1\leq i\leq N}$

Plan de sondage Poissonnien – Processus d'intérêt

Processus d'intérêt $(\mathbb{G}_{T_N}^p f)_{f\in\mathcal{F}}$

$$\mathbb{G}_{T_N}^p f := \sqrt{N}\int f(x)\,(\mathbf{P}_N^p - \mathbf{P}_N)\,(dx) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{\epsilon_i}{p_i} - 1\right) f(X_i)$$

Hypothèses fondamentales

$\mathcal{H}_1$ $\exists\,\lambda > 0,\ N_0 \in \mathbb{N}^\star$:
$\forall N \geq N_0,\ \forall i \in \mathcal{P}_N,\ p_i > \lambda$

$\mathcal{H}_2$ W non proportionnelle à X

$\mathcal{H}_3$ Faible échangeabilité de
$(\epsilon_i, W_i)_{1\leq i\leq N}$

$\mathcal{H}_4$ $\mathcal{F} \subset L_2(\mathbf{P}) := \{h : \mathcal{X} \to \mathbb{R},$
h mesurable et $\mathbb{E}_{\mathbf{P}}\left(h^2(X)\right) < \infty\}$

Plan de sondage Poissonnien – Processus d'intérêt

Processus d'intérêt $(\mathbb{G}_{T_N}^p f)_{f\in\mathcal{F}}$

$$\mathbb{G}_{T_N}^p f := \sqrt{N}\int f(x)\,(\mathbf{P}_N^p - \mathbf{P}_N)\,(dx) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{\epsilon_i}{p_i} - 1\right) f(X_i)$$

Hypothèses fondamentales

$\mathcal{H}_1$ $\exists\,\lambda > 0,\ N_0 \in \mathbb{N}^\star$:
$\forall N \geq N_0,\ \forall i \in \mathcal{P}_N,\ p_i > \lambda$

$\mathcal{H}_2$ W non proportionnelle à X

$\mathcal{H}_3$ Faible échangeabilité de
$(\epsilon_i, W_i)_{1\leq i\leq N}$

$\mathcal{H}_4$ $\mathcal{F} \subset L_2(\mathbf{P}) := \{h : \mathcal{X} \to \mathbb{R},$
h mesurable et $\mathbb{E}_{\mathbf{P}}\left(h^2(X)\right) < \infty\}$

$\mathcal{H}_5$ $\exists H : \mathcal{X} \to \mathbb{R}$ mesurable t.q.
$\int H^2(x)\,\mathbf{P}(dx) < \infty$ et
$\forall x \in \mathcal{X}, f \in \mathcal{F}, |f(x)| \leq H(x)$

Plan de sondage Poissonnien – Processus d'intérêt

Processus d'intérêt $(\mathbb{G}_{T_N}^p f)_{f\in\mathcal{F}}$

$$\mathbb{G}_{T_N}^p f := \sqrt{N}\int f(x)\,(\mathbf{P}_N^p - \mathbf{P}_N)\,(dx) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{\epsilon_i}{p_i} - 1\right) f(X_i)$$

Hypothèses fondamentales

$\mathcal{H}_1$ $\exists\,\lambda > 0,\ N_0 \in \mathbb{N}^\star$:
$\forall N \geq N_0,\ \forall i \in \mathcal{P}_N,\ p_i > \lambda$

$\mathcal{H}_2$ W non proportionnelle à X

$\mathcal{H}_3$ Faible échangeabilité de
$(\epsilon_i, W_i)_{1\leq i\leq N}$

$\mathcal{H}_4$ $\mathcal{F} \subset L_2(\mathbf{P}) := \{h : \mathcal{X} \to \mathbb{R},$
h mesurable et $\mathbb{E}_{\mathbf{P}}\left(h^2(X)\right) < \infty\}$

$\mathcal{H}_5$ $\exists H : \mathcal{X} \to \mathbb{R}$ mesurable t.q.
$\int H^2(x)\,\mathbf{P}(dx) < \infty$ et
$\forall x \in \mathcal{X}, f \in \mathcal{F}, |f(x)| \leq H(x)$

Exemple : $\mathcal{F} = \{f_y(x) := \mathbb{I}\{x \leq y\}, (x,y) \in \mathcal{X}^2\}$ ➡ *f.d.r. empirique*

Convergence – Etude d'un processus recentré

Hájek (1964)

$$\tilde{\mathbb{G}}^{p}_{T_N} f = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\epsilon_i - p_i) \left(\frac{f(X_i)}{p_i} - \theta_{N,p} \right), \text{ où}$$

$$\theta_{N,p} = \frac{1}{D_N} \sum_{j=1}^{N} (1 - p_j) f(X_j) \quad \text{et} \quad D_N = \sum_{j=1}^{N} p_j (1 - p_j)$$

Convergence – Etude d'un processus recentré

Hájek (1964)

$$\tilde{\mathbb{G}}^{p}_{T_N} f = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\epsilon_i - p_i) \left(\frac{f(X_i)}{p_i} - \theta_{N,p} \right), \text{ où}$$

$$\theta_{N,p} = \frac{1}{D_N} \sum_{j=1}^{N} (1 - p_j) f(X_j) \quad \text{et} \quad D_N = \sum_{j=1}^{N} p_j (1 - p_j)$$

➥ à $f \in \mathcal{F}$ fixée, normalité asymptotique $\Big| (X_1, \ldots, X_N)$

Convergence – Etude d'un processus recentré

Hájek (1964)

$$\tilde{\mathbb{G}}^{p}_{T_N} f = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\epsilon_i - p_i) \left(\frac{f(X_i)}{p_i} - \theta_{N,p} \right), \text{ où}$$

$$\theta_{N,p} = \frac{1}{D_N} \sum_{j=1}^{N} (1 - p_j) f(X_j) \quad \text{et} \quad D_N = \sum_{j=1}^{N} p_j (1 - p_j)$$

➥ à $f \in \mathcal{F}$ fixée, normalité asymptotique $\Big| (X_1, \ldots, X_N)$

Remarque

n fixé (plan de sondage Réjectif) : $\sum_{i=1}^{N} (\epsilon_i - p_i) = n - n = 0$

➥ on retrouve le processus initial

Convergence – Résultat asymptotique

Van der Vaart & Wellner (1996)

On étudie la collection triangulaire des v.a. indépendantes

$$\left\{ Z_{N,i}(f) := \frac{1}{\sqrt{N}} (\epsilon_i - p_i) \left(\frac{f(X_i)}{p_i} - \theta_{N,p}(f) \right), \ 1 \leq i \leq N \right\}_{f \in \mathcal{F}}$$

Convergence – Résultat asymptotique

Van der Vaart & Wellner (1996)

On étudie la collection triangulaire des v.a. indépendantes

$$\left\{ Z_{N,i}(f) := \frac{1}{\sqrt{N}} (\epsilon_i - p_i) \left(\frac{f(X_i)}{p_i} - \theta_{N,p}(f) \right), \ 1 \leq i \leq N \right\}_{f \in \mathcal{F}}$$

➥ Vérifie les conditions du Théorème 2.11.1 sur les "Triangular arrays"

Convergence – Résultat asymptotique

Van der Vaart & Wellner (1996)

On étudie la collection triangulaire des v.a. indépendantes

$$\left\{ Z_{N,i}(f) := \frac{1}{\sqrt{N}} (\epsilon_i - p_i) \left(\frac{f(X_i)}{p_i} - \theta_{N,p}(f) \right),\ 1 \leq i \leq N \right\}_{f \in \mathcal{F}}$$

➥ Vérifie les conditions du Théorème 2.11.1 sur les "Triangular arrays"

TLC Fonctionnel pour le plan de sondage Poissonnien

Sous les hypothèses $\mathcal{H}_1 - \mathcal{H}_5$ et sous respect de conditions de type Lindeberg-Feller et d'entropie uniforme,

$$\tilde{\mathbb{G}}^p_{T_N} \underset{N \to \infty}{\Rightarrow} \mathbb{G} \text{ faiblement dans } l_\infty(\mathcal{F})$$

$\mathbb{G}$ processus Gaussien de covariance Σ

Covariance du processus limite

Covariance de $(\tilde{\mathbb{G}}^p_{T_N})_{f\in\mathcal{F}} \mid X_1, \ldots, X_N$

$$C_{N,p}(f,g) = \frac{1}{N}\sum_{i=1}^{N} f(X_i)\, g(X_i)\, (1/p_i - 1) - \theta_{N,p}(f)\, \theta_{N,p}(g)\, \frac{D_N}{N}$$

Covariance du processus limite

Covariance de $(\tilde{\mathbb{G}}^p_{T_N})_{f\in\mathcal{F}} \mid X_1,\dots,X_N$

$$C_{N,p}(f,g) = \frac{1}{N}\sum_{i=1}^{N} f(X_i)\, g(X_i)\, (1/p_i - 1) - \theta_{N,p}(f)\,\theta_{N,p}(g)\,\frac{D_N}{N}$$

Covariance limite

Sous les hypothèses $\mathcal{H}_1$–$\mathcal{H}_5$ et si $0 < \int p^2(w)\,\mathbf{P}_W(w) < \infty$,
alors $\forall\,(f,g)\in\mathcal{F}^2$, $C_{N,p}(f,g) \underset{N\to\infty}{\longrightarrow} \Sigma(f,g)$

$$\Sigma(f,g) := \int f(x)\, g(x)\, (1/p(w) - 1)\, \mathbf{P}_{X,W}(dx,dw) - \theta_p(f)\,\theta_p(g)\,D_p$$

$$\theta_p(f) := \int (1-p(w))\, f(x)\, \mathbf{P}_{X,W}(dx,dw)$$

$$D_p := \int p(w)(1-p(w))\, \mathbf{P}_W(dw)$$

Covariance du processus limite

Covariance de $(\tilde{\mathbb{G}}^p_{T_N})_{f\in\mathcal{F}} \mid X_1,\dots,X_N$

$$C_{N,p}(f,g) = \frac{1}{N}\sum_{i=1}^{N} f(X_i)\, g(X_i)\, (1/p_i - 1) - \theta_{N,p}(f)\,\theta_{N,p}(g)\,\frac{D_N}{N}$$

Covariance limite

Sous les hypothèses $\mathcal{H}_1$–$\mathcal{H}_5$ et si $0 < \int p^2(w)\,\mathbf{P}_W(w) < \infty$,
alors $\forall\,(f,g)\in\mathcal{F}^2$, $C_{N,p}(f,g) \underset{N\to\infty}{\longrightarrow} \Sigma(f,g)$

$$\Sigma(f,g) := \int f(x)\, g(x)\, (1/p(w) - 1)\, \mathbf{P}_{X,W}(dx,dw) - \theta_p(f)\,\theta_p(g)\,D_p$$

$$\theta_p(f) := \int (1-p(w))\, f(x)\, \mathbf{P}_{X,W}(dx,dw)$$

$$D_p := \int p(w)(1-p(w))\, \mathbf{P}_W(dw)$$

Covariance du processus limite

Covariance de $(\tilde{\mathbb{G}}^p_{T_N})_{f\in\mathcal{F}} \mid X_1,\dots,X_N$

$$C_{N,p}(f,g) = \frac{1}{N}\sum_{i=1}^{N} f(X_i)\, g(X_i)\, (1/p_i - 1) - \theta_{N,p}(f)\,\theta_{N,p}(g)\,\frac{D_N}{N}$$

Covariance limite

Sous les hypothèses $\mathcal{H}_1$–$\mathcal{H}_5$ et si $0 < \int p^2(w)\,\mathbf{P}_W(w) < \infty$,
alors $\forall\,(f,g)\in\mathcal{F}^2$, $C_{N,p}(f,g) \underset{N\to\infty}{\longrightarrow} \Sigma(f,g)$

$$\Sigma(f,g) := \int f(x)\, g(x)\, (1/p(w) - 1)\, \mathbf{P}_{X,W}(dx,dw) - \theta_p(f)\,\theta_p(g)\,D_p$$

$$\theta_p(f) := \int (1-p(w))\, f(x)\, \mathbf{P}_{X,W}(dx,dw)$$

$$D_p := \int p(w)(1-p(w))\, \mathbf{P}_W(dw)$$

Covariance du processus limite

Covariance de $(\tilde{\mathbb{G}}^p_{T_N})_{f\in\mathcal{F}} \mid X_1,\dots,X_N$

$$C_{N,p}(f,g) = \frac{1}{N}\sum_{i=1}^{N} f(X_i)\, g(X_i)\, (1/p_i - 1) - \theta_{N,p}(f)\,\theta_{N,p}(g)\,\frac{D_N}{N}$$

Covariance limite

Sous les hypothèses $\mathcal{H}_1$–$\mathcal{H}_5$ et si $0 < \int p^2(w)\,\mathbf{P}_W(w) < \infty$,
alors $\forall\,(f,g)\in\mathcal{F}^2$, $C_{N,p}(f,g) \underset{N\to\infty}{\longrightarrow} \Sigma(f,g)$

$$\Sigma(f,g) := \int f(x)\, g(x)\, (1/p(w) - 1)\, \mathbf{P}_{X,W}(dx,dw) - \theta_p(f)\,\theta_p(g)\, D_p$$

$$\theta_p(f) := \int (1-p(w))\, f(x)\, \mathbf{P}_{X,W}(dx,dw)$$

$$D_p := \int p(w)(1-p(w))\, \mathbf{P}_W(dw)$$

Plan de l'exposé

Plan de sondage Réjectif

Définition

n fixé probabilités d'inclusion $\pi_1^R, \ldots, \pi_N^R$

$$R_N := \underset{\mathbf{p}:(\pi_1,\ldots,\pi_N)=(\pi_1^R,\ldots,\pi_N^R)}{\operatorname{argmax}} - \sum_{\{s:\#s=n\}} \mathbf{p}(s) \log \mathbf{p}(s)$$

Plan de sondage Réjectif

Définition

n fixé — probabilités d'inclusion $\pi_1^R, \ldots, \pi_N^R$

$$R_N := \underset{\mathbf{p}:(\pi_1,\ldots,\pi_N)=(\pi_1^R,\ldots,\pi_N^R)}{\operatorname{argmax}} - \sum_{\{s:\#s=n\}} \mathbf{p}(s) \log \mathbf{p}(s)$$

➥ plan d'entropie maximale

Plan de sondage Réjectif

Définition

n fixé — probabilités d'inclusion $\pi_1^R, \ldots, \pi_N^R$

$$R_N := \underset{\mathbf{p}:(\pi_1,\ldots,\pi_N)=(\pi_1^R,\ldots,\pi_N^R)}{\operatorname{argmax}} \; - \sum_{\{s:\#s=n\}} \mathbf{p}(s) \log \mathbf{p}(s)$$

➥ plan d'entropie maximale, ou Poissonnien conditionnel

Plan de sondage Réjectif

Définition

n fixé probabilités d'inclusion $\pi_1^R, \ldots, \pi_N^R$

$$R_N := \underset{\mathbf{p}:(\pi_1,\ldots,\pi_N)=(\pi_1^R,\ldots,\pi_N^R)}{\operatorname{argmax}} - \sum_{\{s:\#s=n\}} \mathbf{p}(s) \log \mathbf{p}(s)$$

➥ plan d'entropie maximale, ou Poissonnien conditionnel

Lien avec le plan Poissonnien : construction

- Tirer $\mathcal{S}$ un selon un plan Poissonnien de probabilités d'inclusions $p_1, \ldots, p_N$ bien choisies, avec $\sum_{i=1}^N p_i = n$
- Si $\#\mathcal{S} = n$ conserver $\mathcal{S}$, sinon recommencer le tirage

Plan de sondage Réjectif

Définition

n fixé — probabilités d'inclusion $\pi_1^R, \ldots, \pi_N^R$

$$R_N := \underset{\mathbf{p}:(\pi_1,\ldots,\pi_N)=(\pi_1^R,\ldots,\pi_N^R)}{\operatorname{argmax}} - \sum_{\{s:\#s=n\}} \mathbf{p}(s) \log \mathbf{p}(s)$$

➥ plan d'entropie maximale, ou Poissonnien conditionnel

Lien avec le plan Poissonnien : construction

- Tirer $\mathcal{S}$ un selon un plan Poissonnien de probabilités d'inclusions $p_1, \ldots, p_N$ bien choisies, avec $\sum_{i=1}^N p_i = n$
- Si $\#\mathcal{S} = n$ conserver $\mathcal{S}$, sinon recommencer le tirage

➥ liens entre $(p_1, \ldots, p_N)$ et $(\pi_1^R, \ldots, \pi_N^R)$ (Hájek, 1964)

Distance au plan Poissonnien

Quelques distances

- Variation totale : $d_{VT}(T_N, R_N) := \sum_{s \subset \mathcal{P}_N} \left| R_N(s) - T_N(s) \right|$
- Entropie : $d_E(T_N, R_N) := \sum_{s \subset \mathcal{P}_N} T_N(s) \log \frac{T_N(s)}{R_N(s)}$
- Bounded Lipschitz : $d_{BL}(fX, fY) := \sup_{b \in BL_1(l_\infty(\mathcal{F}))} \left| \mathbb{E}(b(fX)) - \mathbb{E}(b(fY)) \right|$

Distance au plan Poissonnien

Quelques distances

- Variation totale : $d_{VT}(T_N, R_N) := \sum_{s \subset \mathcal{P}_N} \left| R_N(s) - T_N(s) \right|$
- Entropie : $d_E(T_N, R_N) := \sum_{s \subset \mathcal{P}_N} T_N(s) \log \frac{T_N(s)}{R_N(s)}$
- Bounded Lipschitz : $d_{BL}(fX, fY) := \sup_{b \in BL_1(l_\infty(\mathcal{F}))} \left| \mathbb{E}(b(fX)) - \mathbb{E}(b(fY)) \right|$

Résultats fondamentaux

$$d_{BL}(\tilde{\mathbb{G}}^p_{T_N}, \tilde{\mathbb{G}}^p_{R_N}) \leq d_{VT}(T_N, R_N) \leq d_E(T_N, R_N)$$

Distance au plan Poissonnien

Quelques distances

- Variation totale : $d_{VT}(T_N, R_N) := \sum_{s \subset \mathcal{P}_N} \left| R_N(s) - T_N(s) \right|$
- Entropie : $d_E(T_N, R_N) := \sum_{s \subset \mathcal{P}_N} T_N(s) \log \frac{T_N(s)}{R_N(s)}$
- Bounded Lipschitz : $d_{BL}(fX, fY) := \sup_{b \in BL_1(l_\infty(\mathcal{F}))} \left| \mathbb{E}(b(fX)) - \mathbb{E}(b(fY)) \right|$

Résultats fondamentaux

$$d_{BL}(\tilde{\mathbb{G}}^p_{T_N}, \tilde{\mathbb{G}}^p_{R_N}) \leq d_{VT}(T_N, R_N) \leq d_E(T_N, R_N)$$

Si $(T_N)_{N \geq 1}$ et $(R_N)_{N \geq 1}$ vérifient

$$d_{VT}(T_N, R_N) \underset{N \to \infty}{\longrightarrow} 0 \text{ ou } d_E(T_N, R_N) \underset{N \to \infty}{\longrightarrow} 0$$

et si $\exists\, \mathbb{G}$ processus Gaussien t.q. $d_{BL}(\tilde{\mathbb{G}}^p_{T_N}, \mathbb{G}) \underset{N \to \infty}{\longrightarrow} 0$, alors

$$d_{BL}(\tilde{\mathbb{G}}^p_{R_N}, \mathbb{G}) \leq d_{VT}(T_N, R_N) \leq d_E(T_N, R_N)$$

Convergence – Résultats asymptotiques

Processus empirique du plan Réjectif sous $(p_1, \ldots, p_N)$

- Hájek (1964) ➡ $d_E(T_N, R_N) \underset{N \to \infty}{\longrightarrow} 0$

Convergence – Résultats asymptotiques

Processus empirique du plan Réjectif sous $(p_1, \ldots, p_N)$

- Hájek (1964) ➡ $d_E(T_N, R_N) \underset{N\to\infty}{\longrightarrow} 0$
- TLC Fonctionnel pour le plan Poissonnien ➡ $d_{BL}(\tilde{\mathbb{G}}^p_{T_N}, \mathbb{G}) \underset{N\to\infty}{\longrightarrow} 0$

Convergence – Résultats asymptotiques

Processus empirique du plan Réjectif sous $(p_1, \ldots, p_N)$

- Hájek (1964) ➡ $d_E(T_N, R_N) \underset{N\to\infty}{\longrightarrow} 0$
- TLC Fonctionnel pour le plan Poissonnien ➡ $d_{BL}(\tilde{\mathbb{G}}^p_{T_N}, \mathbb{G}) \underset{N\to\infty}{\longrightarrow} 0$
- A n fixé $\tilde{\mathbb{G}}^p_{T_N} f = \mathbb{G}^p_{R_N} f$

Convergence – Résultats asymptotiques

Processus empirique du plan Réjectif sous $(p_1, \ldots, p_N)$

- Hájek (1964) ➡ $d_E(T_N, R_N) \underset{N\to\infty}{\longrightarrow} 0$
- TLC Fonctionnel pour le plan Poissonnien ➡ $d_{BL}(\tilde{\mathbb{G}}^p_{T_N}, \mathbb{G}) \underset{N\to\infty}{\longrightarrow} 0$
- A n fixé $\tilde{\mathbb{G}}^p_{T_N} f = \mathbb{G}^p_{R_N} f$

Sous les hypothèses $\mathcal{H}_1 - \mathcal{H}_5$ et sous respect de conditions de type Lindeberg-Feller et d'entropie uniforme,

$$\mathbb{G}^p_{R_N} \underset{N\to\infty}{\Rightarrow} \mathbb{G} \quad \text{faiblement dans } l_\infty(\mathcal{F})$$

$\mathbb{G}$ processus Gaussien de covariance Σ

Convergence – Résultats asymptotiques

Processus empirique du plan Réjectif sous $(p_1, \ldots, p_N)$

- Hájek (1964) ➡ $d_E(T_N, R_N) \underset{N\to\infty}{\longrightarrow} 0$
- TLC Fonctionnel pour le plan Poissonnien ➡ $d_{BL}(\tilde{\mathbb{G}}^p_{T_N}, \mathbb{G}) \underset{N\to\infty}{\longrightarrow} 0$
- A n fixé $\tilde{\mathbb{G}}^p_{T_N} f = \mathbb{G}^p_{R_N} f$

Sous les hypothèses $\mathcal{H}_1 - \mathcal{H}_5$ et sous respect de conditions de type Lindeberg-Feller et d'entropie uniforme,

$$\mathbb{G}^p_{R_N} \underset{N\to\infty}{\Rightarrow} \mathbb{G} \text{ faiblement dans } l_\infty(\mathcal{F})$$

$\mathbb{G}$ processus Gaussien de covariance Σ

➥ Reste vrai pour les plans de sondage comme Rao-Sampford (Berger, 1998)

Convergence – Résultats asymptotiques

Processus empirique du plan Réjectif sous $(\pi_1^R, \ldots, \pi_N^R)$

- Hájek (1964), Théorème 5.1 ➡ $(p_i)_{1 \leq i \leq N}$ "proches" des $(\pi_i^R)_{1 \leq i \leq N}$

Convergence – Résultats asymptotiques

Processus empirique du plan Réjectif sous $(\pi_1^R, \ldots, \pi_N^R)$

- Hájek (1964), Théorème 5.1 ➡ $(p_i)_{1 \leq i \leq N}$ "proches" des $(\pi_i^R)_{1 \leq i \leq N}$
- Hypothèse $\mathcal{H}_1$ ➡ $(p_i)_{1 \leq i \leq N}$ et $(\pi_i^R)_{1 \leq i \leq N}$ jamais trop petits

Convergence – Résultats asymptotiques

Processus empirique du plan Réjectif sous $(\pi_1^R, \ldots, \pi_N^R)$

- Hájek (1964), Théorème 5.1 ➡ $(p_i)_{1 \leq i \leq N}$ "proches" des $(\pi_i^R)_{1 \leq i \leq N}$
- Hypothèse $\mathcal{H}_1$ ➡ $(p_i)_{1 \leq i \leq N}$ et $(\pi_i^R)_{1 \leq i \leq N}$ jamais trop petits
- Version fonctionnelle du Théorème de Slutsky & résultat précédent

Convergence – Résultats asymptotiques

Processus empirique du plan Réjectif sous $(\pi_1^R, \ldots, \pi_N^R)$

- Hájek (1964), Théorème 5.1 ➡ $(p_i)_{1\leq i\leq N}$ "proches" des $(\pi_i^R)_{1\leq i\leq N}$
- Hypothèse $\mathcal{H}_1$ ➡ $(p_i)_{1\leq i\leq N}$ et $(\pi_i^R)_{1\leq i\leq N}$ jamais trop petits
- Version fonctionnelle du Théorème de Slutsky & résultat précédent

Sous les hypothèses $\mathcal{H}_1 - \mathcal{H}_5$ et sous respect de conditions de type Lindeberg-Feller et d'entropie uniforme,

$$\mathbb{G}_{R_N}^{\pi^R} \underset{N\to\infty}{\Rightarrow} \mathbb{G} \quad \text{faiblement dans } l_\infty(\mathcal{F})$$

$\mathbb{G}$ processus Gaussien de covariance Σ

Convergence – Résultats asymptotiques

Processus empirique du plan Réjectif sous $(\pi_1^R, \ldots, \pi_N^R)$

- Hájek (1964), Théorème 5.1 ➡ $(p_i)_{1\leq i\leq N}$ "proches" des $(\pi_i^R)_{1\leq i\leq N}$
- Hypothèse $\mathcal{H}_1$ ➡ $(p_i)_{1\leq i\leq N}$ et $(\pi_i^R)_{1\leq i\leq N}$ jamais trop petits
- Version fonctionnelle du Théorème de Slutsky & résultat précédent

Sous les hypothèses $\mathcal{H}_1 - \mathcal{H}_5$ et sous respect de conditions de type Lindeberg-Feller et d'entropie uniforme,

$$\mathbb{G}_{R_N}^{\pi^R} \underset{N\to\infty}{\Rightarrow} \mathbb{G} \quad \text{faiblement dans } l_\infty(\mathcal{F})$$

$\mathbb{G}$ processus Gaussien de covariance Σ

➥ Reste vrai pour les plans de sondage comme Rao-Sampford (Berger, 1998)

En bref...

Résultats principaux

Convergence asymptotique des processus empiriques vers un processus Gaussien pour les plans Poissonnien, Réjectif et assimilés

En bref...

Résultats principaux

Convergence asymptotique des processus empiriques vers un processus Gaussien pour les plans Poissonnien, Réjectif et assimilés

Convergence asymptotique ➨ il faut de grandes populations

En bref...

Résultats principaux

Convergence asymptotique des processus empiriques vers un processus Gaussien pour les plans Poissonnien, Réjectif et assimilés


Convergence asymptotique ➡ il faut de grandes populations


Intervalles de confiance pour la f.d.r. empirique pas “distribution free”

En bref...

Résultats principaux

Convergence asymptotique des processus empiriques vers un processus Gaussien pour les plans Poissonnien, Réjectif et assimilés


Convergence asymptotique ➡ il faut de grandes populations


Intervalles de confiance pour la f.d.r. empirique pas “distribution free”

Richesse de l’approche

- Résultats pour le plan Poissonnien “faciles” à démontrer
- Extension naturelle à tous les plans “proches” du plan Poissonnien

En bref...

Résultats principaux

Convergence asymptotique des processus empiriques vers un processus Gaussien pour les plans Poissonnien, Réjectif et assimilés


Convergence asymptotique ➡ il faut de grandes populations


Intervalles de confiance pour la f.d.r. empirique pas “distribution free”

Richesse de l’approche

- Résultats pour le plan Poissonnien “faciles” à démontrer
- Extension naturelle à tous les plans “proches” du plan Poissonnien
- Convergence asymptotique de l’estimateur de Hill repondéré

Quelques références

- Y.G. Berger, *Rate of convergence to normal distribution for the Horvitz-Thompson estimator*, J. Stat. Plan. Inf **67** (1998), no. 2, 209–226.
- J. Hajek, *On the Convergence of the Horvitz-Thompson Estimator*, The Annals of Mathematical Statistics **35** (1964), no. 4, 1491–1523.
- B. Hill, *A simple approach to inference about the tail of a distribution*, Ann. Statist. **3** (1975), 1163–1174.
- P.M. Robinson, *On the Convergence of the Horvitz-Thompson Estimator*, Australian Journal of Statistics **24** (1982), no. 2, 234–238.
- T. Saegusa and J.A. Wellner, *Weighted likelihood estimation under two-phase sampling*, Preprint available at http ://arxiv.org/abs/1112.4951v1 (2011).
- A. Van der Vaart and J. Wellner, *Weak convergence and empirical processes : with applications to statistics*, Springer, 1996.