Estimation de la variance inconditionnelle dans le cadre d'enquêtes complexes

Yves G. Berger

University of Southampton

12 février 2026

Basé sur Berger (2025) "*International Statistical Review*"

Introduction: exemples

- Données de population : $y_i \sim N(\mu, \sigma)$ $i = 1, \ldots, N$, i.i.d.
- $N = 1000$, population finie
- Échantillon S de taille $n = 200 \Longrightarrow \dfrac{n}{N} = 0.2$
- Nous souhaitons estimer la moyenne μ.
- Considérons différents plans d'échantillonnage.

Échantillonnage aléatoire simple sans remise

- Estimateur

$$\bar{y} := \frac{1}{n} \sum_{i \in S} y_i$$

Échantillonnage aléatoire simple sans remise

- Estimateur

$$\bar{y} := \frac{1}{n} \sum_{i \in S} y_i$$

- Estimateur de variance

$$\widehat{\mathbb{V}}(\bar{y}) := \left(1 - \frac{n}{N}\right) \frac{\widehat{\sigma}^2}{n}, \quad \text{avec } \widehat{\sigma}^2 := \frac{1}{n-1} \sum_{i \in S} \left(y_i - \bar{y}\right)^2$$

Échantillonnage aléatoire simple sans remise

- Estimateur

$$\bar{y} := \frac{1}{n} \sum_{i \in S} y_i$$

- Estimateur de variance

$$\widehat{\mathbb{V}}(\bar{y}) := \left(1 - \frac{n}{N}\right) \frac{\widehat{\sigma}^2}{n}, \quad \text{avec } \widehat{\sigma}^2 := \frac{1}{n-1} \sum_{i \in S} \left(y_i - \bar{y}\right)^2$$

- **$\widehat{\mathbb{V}}(\bar{y})$ sans biais?**

Échantillonnage aléatoire simple sans remise

- Estimateur

$$\bar{y} := \frac{1}{n} \sum_{i \in S} y_i$$

- Estimateur de variance

$$\widehat{\mathbb{V}}(\bar{y}) := \left(1 - \frac{n}{N}\right) \frac{\widehat{\sigma}^2}{n}, \quad \text{avec } \widehat{\sigma}^2 := \frac{1}{n-1} \sum_{i \in S} (y_i - \bar{y})^2$$

- **$\widehat{\mathbb{V}}(\bar{y})$ sans biais?**

$\Longrightarrow \widehat{\mathbb{V}}(\bar{y})$ est **biaisé!** Relative bias $= -20\%$

Échantillonnage aléatoire simple avec remise

- Échantillon $\equiv$ n tirages indépendants
- Estimateur

$$\bar{y} := \frac{1}{n} \sum_{i \in S} y_i$$

- Estimateur de variance

$$\widehat{\mathbb{V}}(\bar{y}) := \frac{\widehat{\sigma}^2}{n}, \quad \text{avec } \widehat{\sigma}^2 := \frac{1}{n-1} \sum_{i \in S} \left(y_i - \bar{y}\right)^2$$

Échantillonnage aléatoire simple avec remise

- Échantillon $\equiv$ n tirages indépendants
- Estimateur

$$\bar{y} := \frac{1}{n} \sum_{i \in S} y_i$$

- Estimateur de variance

$$\widehat{\mathbb{V}}(\bar{y}) := \frac{\widehat{\sigma}^2}{n}, \quad \text{avec } \widehat{\sigma}^2 := \frac{1}{n-1} \sum_{i \in S} \left(y_i - \bar{y}\right)^2$$

- $\widehat{\mathbb{V}}(\bar{y})$ **sans biais?**

Échantillonnage aléatoire simple avec remise

- Échantillon $\equiv$ n tirages indépendants
- Estimateur

$$\bar{y} := \frac{1}{n} \sum_{i \in S} y_i$$

- Estimateur de variance

$$\widehat{\mathbb{V}}(\bar{y}) := \frac{\widehat{\sigma}^2}{n}, \quad \text{avec } \widehat{\sigma}^2 := \frac{1}{n-1} \sum_{i \in S} \left(y_i - \bar{y}\right)^2$$

- **$\widehat{\mathbb{V}}(\bar{y})$ sans biais?**

$\Longrightarrow \widehat{\mathbb{V}}(\bar{y})$ est **biaisé!**

Échantillonnage de Poisson

- Échantillon $\equiv$ sélectionner chaque unité avec une probabilité n/N indépendamment.
- Estimateur

$$\bar{y} := \frac{1}{n} \sum_{i \in S} y_i$$

- Estimateur de variance

$$\widehat{\mathbb{V}}(\bar{y}) := \left(1 - \frac{n}{N}\right) \frac{1}{n^2} \sum_{i \in S} y_i^2$$

Échantillonnage de Poisson

- Échantillon $\equiv$ sélectionner chaque unité avec une probabilité n/N indépendamment.
- Estimateur

$$\bar{y} := \frac{1}{n}\sum_{i\in S} y_i$$

- Estimateur de variance

$$\widehat{\mathbb{V}}(\bar{y}) := \left(1-\frac{n}{N}\right)\frac{1}{n^2}\sum_{i\in S} y_i^2$$

- $\widehat{\mathbb{V}}(\bar{y})$ **sans biais?**

Échantillonnage de Poisson

- Échantillon $\equiv$ sélectionner chaque unité avec une probabilité n/N indépendamment.
- Estimateur

$$\bar{y} := \frac{1}{n}\sum_{i\in S} y_i$$

- Estimateur de variance

$$\widehat{\mathbb{V}}(\bar{y}) := \left(1 - \frac{n}{N}\right)\frac{1}{n^2}\sum_{i\in S} y_i^2$$

- **$\widehat{\mathbb{V}}(\bar{y})$ sans biais?**

$\Longrightarrow \widehat{\mathbb{V}}(\bar{y})$ est **biaisé!**

Échantillonnage aléatoire simple sans remise

- Ignorons la "*Correction pour population finie*" :

$\Longrightarrow$ Estimateur de variance

$$\widehat{\mathbb{V}}(\bar{y}) := \frac{\widehat{\sigma}^2}{n}, \quad \text{c'est-à-dire, même estimateur qu'avec remise}$$

- **Is $\widehat{\mathbb{V}}(\bar{y})$ sans biais?**

Échantillonnage aléatoire simple sans remise

- Ignorons la "*Correction pour population finie*" :

$\Longrightarrow$ Estimateur de variance

$$\widehat{\mathbb{V}}(\bar{y}) := \frac{\widehat{\sigma}^2}{n}, \quad \text{c'est-à-dire, même estimateur qu'avec remise}$$

- **Is $\widehat{\mathbb{V}}(\bar{y})$ sans biais?**

$\Longrightarrow \widehat{\mathbb{V}}(\bar{y})$ est **sans biais!**

Échantillonnage aléatoire simple sans remise

- Ignorons la "*Correction pour population finie*" :

$\Longrightarrow$ Estimateur de variance

$$\widehat{\mathbb{V}}(\bar{y}) := \frac{\widehat{\sigma}^2}{n}, \quad \text{c'est-à-dire, même estimateur qu'avec remise}$$

- **Is $\widehat{\mathbb{V}}(\bar{y})$ sans biais?**

$\Longrightarrow \widehat{\mathbb{V}}(\bar{y})$ est **sans biais!**

- Malgré que
  - Échantillonnage sans remise
  - Nous avons une fraction d'échantillonnage importante ($n/N = 0.2$)
  - La population est finie ($N = 1000$). Cependant, pas de "*correction de population finie*" !

Je reviendrai sur ces exemples plus tard.

Considérons une "*approche inconditionnelle*".

Approche inconditionnelle

- Une **distribution** ξ qui génère les données de population:

$$\boldsymbol{Y} := (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_i, \ldots, \boldsymbol{y}_N)^\top \sim \text{distribution } \xi$$

- $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ ne sont pas nécessairement i.i.d.
- On pourrait qualifier ξ de "*modèle*" au sens général.
- Le **plan d'échantillonnage** $P(s)$ spécifie la distribution de l'échantillon S, qui pourrait être informatif
- **Deux processus aléatoires**:

  Une distribution ξ qui génère $\boldsymbol{Y}$

  Une sélection aléatoire d'un échantillon $S \subset U = \{1, \ldots, N\}$
- Approche hybride basée sur l'échantillonnage et un "*modèle*"

Approche inconditionnelle

- **Deux paramètres cibles possibles** :
  - spécifié par la distribution ξ, par exemple $\boldsymbol{\theta}_0$ est l'espérance de la distribution ξ
  - $\boldsymbol{\theta}_N$ une fonction de $\boldsymbol{Y}$, par exemple $\boldsymbol{\theta}_N = \dfrac{1}{N} \sum_{i \in \mathcal{U}} \boldsymbol{y}_i$
- **Estimation ponctuelle**: Un estimateur $\widehat{\boldsymbol{\theta}}$ sans biais (sous le plan d'échantillonnage) de $\boldsymbol{\theta}_N$ est généralement sans biais pour $\boldsymbol{\theta}_0$:

$$\mathbb{E}_\xi \mathbb{E}_P[\widehat{\boldsymbol{\theta}} \mid \boldsymbol{Y}] = \mathbb{E}_\xi[\boldsymbol{\theta}_N] = \boldsymbol{\theta}_0$$

- **Pour l'estimation de variance**, la distinction entre $\boldsymbol{\theta}_0$ et $\boldsymbol{\theta}_N$ est cruciale

Approche inconditionnelle: $\boldsymbol{\theta}_N$ est la cible

- Soit $\widehat{\boldsymbol{\theta}}$ un estimateur sans biais (sous le plan d'échantillonnage) de $\boldsymbol{\theta}_N$

- La **variance inconditionnelle**:

$$\begin{aligned}
\mathbb{V}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N) &= \mathbb{E}_\xi \mathbb{V}_P(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N \mid \boldsymbol{Y}) + \mathbb{V}_\xi \mathbb{E}_P(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N \mid \boldsymbol{Y}) \\
&= \mathbb{E}_\xi \mathbb{V}_P(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N \mid \boldsymbol{Y}) + 0 \\
&= \mathbb{E}_\xi \mathbb{V}_P(\widehat{\boldsymbol{\theta}} \mid \boldsymbol{Y})
\end{aligned}$$

- **Tout estimateurs de variance sans biais $\widehat{\mathbb{V}}_P(\widehat{\boldsymbol{\theta}} \mid \boldsymbol{Y})$ basé sur le plan d'échantillonnage, est un estimateur sans biais de la variance inconditionnelle**

$$\mathbb{E}_\xi \mathbb{E}_P[\widehat{\mathbb{V}}_P(\widehat{\boldsymbol{\theta}} \mid \boldsymbol{Y}) \mid \boldsymbol{Y}] = \mathbb{E}_\xi \mathbb{V}_P(\widehat{\boldsymbol{\theta}} \mid \boldsymbol{Y}) = \mathbb{V}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N)$$

Approche inconditionnelle: $\boldsymbol{\theta}_N$ est la cible

- Une approche basée sur le plan d'échantillonnage est valable
- On constate que la **distribution/modèle ξ est ignorable**, puisque ξ ne joue aucun rôle dans l'estimation de la variance.
- Notez que les $\boldsymbol{y}_i$ sont des variables aléatoires !

> - La caractéristique principale des approches basées sur le plan d'échantillonnage est le fait que $\boldsymbol{\theta}_N$ est la cible
> - Le fait que les données $\boldsymbol{Y}$ (les $\boldsymbol{y}_i$) soient fixes (pas aléatoires) n'est pas la caractéristique principale des approches basées sur le plan d'échantillonnage !

- Puisque nous estimons $\mathbb{V}_P(\widehat{\boldsymbol{\theta}} \mid \boldsymbol{Y})$, les données $\boldsymbol{Y}$ **peuvent être considérées comme fixes**, du fait du conditionnement. Cela ne signifie pas pour autant que les données $\boldsymbol{Y}$ sont fixes (et non aléatoires) !

Exemple: quantiles

- L'estimation de la variance d'un quantile α implique une linéarisation

$$\widehat{\mathbb{V}}_P\big(\widehat{Y}_\alpha\big) \approx \widehat{\mathbb{V}}_P\Big(\frac{1}{N}\sum_{i\in S}\frac{\widehat{z}_i}{\pi_i}\Big),$$

où $\widehat{z}_i = \dfrac{1}{N\, f(\widehat{Y}_\alpha)}\Big\{\delta(y_i \leqslant \widehat{Y}_\alpha - \alpha\Big\}$

- $f(\cdot)$ est la **densité** de la distribution de y_i.

Densité $\Longrightarrow$ Les y_i sont aléatoires et non constantes.

Approche inconditionnelle: échantillonnage non-informatif

- Échantillonnage non-informatif: $\boldsymbol{Y} \perp\!\!\!\perp S$, c'est à dire ξ et $P(s)$ sont des processus indépendants:

- Soit $\widehat{\boldsymbol{\theta}}$ un estimateur sans biais sous le modèle de $\widetilde{\boldsymbol{\theta}}$, qui pourrait être $\boldsymbol{\theta}_N$ ou $\boldsymbol{\theta}_0$.

- **ξ et P peuvent être inversés**:

$$
\begin{aligned}
\mathbb{V}(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}) &= \mathbb{E}_P \mathbb{V}_\xi(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}} \mid S) + \mathbb{V}_P \mathbb{E}_\xi(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}} \mid S) \\
&= \mathbb{E}_P \mathbb{V}_\xi(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}} \mid S) + 0
\end{aligned}
$$

- **Tout estimateurs de variance sans biais $\widehat{\mathbb{V}}_\xi(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}} \mid S)$ basé sur un modèle, est un estimateur sans biais de la variance inconditionnelle**

$$
\mathbb{E}_P \mathbb{E}_\xi[\widehat{\mathbb{V}}_\xi(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}} \mid S) \mid \boldsymbol{Y}] = \mathbb{E}_P \mathbb{V}_\xi(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}} \mid \boldsymbol{Y}) = \mathbb{V}(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}})
$$

Approche inconditionnelle: échantillonnage non-informatif

- Une approche basée sur un modèle est valable
- On constate que **le plan d'échantillonnage $P(s)$ est ignorable**, car $P(s)$ n'intervient pas dans l'estimation de la variance.
- Notez que l'échantillon reste une variable aléatoire !

> - La **caractéristique principale des approches basées sur des modèles est l'échantillonnage non informatif**.
> - Le fait que les données S ne soient pas aléatoires n'est pas la caractéristique clef des approches basées sur des modèles !

- Puisque nous estimons $\mathbb{V}_\xi(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}} \mid S)$, l'échantillon S peut être **considéré comme fixe**, du fait du conditionnement. Cela ne signifie pas pour autant que l'échantillon S est fixes (et non aléatoires) !

Approche inconditionnelle: échantillonnage non-informatif

- ξ peut impliquer un modèle de régression pour les données $\boldsymbol{Y}$ afin d'estimer $\mathbb{V}_\xi(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}} \mid S)$, par exemple pour l'estimation de petits domaines ; mais le modèle régression n'est pas la caractéristique principale.

- La caractéristique principale de l'approche basée sur un modèle n'est pas la modélisation par régression, mais l'échantillonnage non informatif.

- Les approches basées sur des modèles devraient être appelées "*échantillonnage non informatives*" pour mettre en évidence la caractéristique clef

Approche inconditionnelle: échantillonnage non-informatif

- Sous un échantillonnage non informatif, nous pouvons estimer $\boldsymbol{\theta}_N$ (prédiction) ou $\boldsymbol{\theta}_0$.

- Lorsque $\boldsymbol{\theta}_0$ est la cible, cela ne signifie pas que nous devons envisager des approches basées sur un modèle (non informatives) !

Approche inconditionnelle: échantillonnage informatif

- **Nous ne pouvons pas échanger ξ et P**, c'est-à-dire que nous utilisons $\mathbb{V} \equiv \mathbb{E}_\xi \mathbb{V}_P + \mathbb{E}_\xi \mathbb{V}_P$, et pas $\mathbb{E}_P \mathbb{V}_\xi + \mathbb{E}_P \mathbb{V}_\xi$
- **Lorsque la cible est θ_N**: $\Longrightarrow$ approches basées sur le plan d'échantillonnage

$$\begin{aligned} \mathbb{V}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N) &= \mathbb{E}_\xi \mathbb{V}_P(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N \mid \boldsymbol{Y}) + \mathbb{V}_\xi \mathbb{E}_P(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N \mid \boldsymbol{Y}) \\ &= \mathbb{E}_\xi \mathbb{V}_P(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N \mid \boldsymbol{Y}) + 0 \end{aligned}$$

  Problème résolu lorsque la cible est $\boldsymbol{\theta}_N$, inutile d'en discuter davantage.

- Cibler $\boldsymbol{\theta}_N$ est plus approprié lorsque l'on s'intéresse à une inférence descriptive.
- Inférence analytique : que signifie $\boldsymbol{\theta}_N$? $\boldsymbol{\theta}_0$ est plus logique.

Approche inconditionnelle: échantillonnage informatif

- **Lorsque θ_0 est la cible**:

$$\mathbb{V}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \mathbb{E}_\xi \mathbb{V}_P(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \mid \boldsymbol{Y}) + \mathbb{V}_\xi \mathbb{E}_P(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \mid \boldsymbol{Y})$$

Maintenant $\mathbb{V}_\xi \mathbb{E}_P(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \mid \boldsymbol{Y}) \neq 0$! Les estimateurs de variance traditionnels basés sur le plan d'échantillonnage sont biaisés !

> Contribution : Estimation de la variance sous échantillonnage informatif, lorsque $\boldsymbol{\theta}_0$ est la cible.

- Il existe de nombreuses situations où la cible devrait être $\boldsymbol{\theta}_0$:
  - $\boldsymbol{\theta}_0$ est un paramètre de régression (inférence analytique)
  - Estimation de petits domaines
  - Estimation de quantiles?
  - Lorsque $\boldsymbol{\theta}_0 = \mathbb{E}_\xi \mathbb{E}_P(\widehat{\boldsymbol{\theta}})$. Indice des prix à la consommation?

Exemples de l'introduction

- $y_i \sim N(\mu, \sigma)$
- La cible était $\theta_0 = \mu$, et non $\theta_N = \frac{1}{N} \sum_{i \in \mathcal{U}} y_i$

  $\Longrightarrow$ l'estimateur de **variance sous le plan d'échantillonnage sont biaisés**
- Sous un **échantillonnage aléatoire simple <u>sans remise</u>**:

$$\begin{aligned}
\widehat{\theta} &:= \frac{1}{n} \sum_{i \in S} y_i \quad \text{estimateur sans bias de } \theta_0 = \mu \\
\widehat{\mathbb{V}}(\widehat{\theta}\,) &:= \frac{\widehat{\sigma}^2}{n}, \quad \text{estimateur } \textbf{sans biais} \text{ de } \mathbb{V}(\widehat{\theta}\,) \\
\mathbb{V}(\widehat{\theta}\,) &:= \mathbb{E}_\xi\left[\mathbb{V}_P\left(\widehat{\theta} \mid \boldsymbol{Y}\right)\right] + \mathbb{V}_\xi\left[\mathbb{E}_P\left(\widehat{\theta} \mid \boldsymbol{Y}\right)\right].
\end{aligned}$$

- $\mathbb{V}(\widehat{\theta}\,)$ **a 2 termes et** $\widehat{\mathbb{V}}(\widehat{\theta}\,)$ **a 1 terme ?!?!?**

Échantillonnage aléatoire simple sans remise

- $y_i \sim N(\mu, \sigma)$ i.i.d
- $\mathbb{V}_P(\widehat{\theta} \mid \boldsymbol{Y}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n}$, où $S^2 = \frac{1}{N-1}\sum_{i \in \mathcal{U}}(y_i - \theta_N)^2$
- $\mathbb{E}_\xi \mathbb{V}_P(\widehat{\theta} \mid \boldsymbol{Y}) = \left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n} = \frac{\sigma^2}{n} - \frac{\sigma^2}{N}$, car $\theta_0 = \mu$ est la cible
- $\mathbb{V}_\xi \mathbb{E}_P(\widehat{\theta} \mid \boldsymbol{Y}) = \mathbb{V}_\xi(\theta_N) = \frac{\sigma^2}{N}$

$$\begin{aligned} \mathbb{V}(\widehat{\theta}) &:= \mathbb{E}_\xi\left[\mathbb{V}_P\left(\widehat{\theta} \mid \boldsymbol{Y}\right)\right] + \mathbb{V}_\xi\left[\mathbb{E}_P\left(\widehat{\theta} \mid \boldsymbol{Y}\right)\right] \\ &= \frac{\sigma^2}{n} - \frac{\sigma^2}{N} + \frac{\sigma^2}{N} = \frac{\sigma^2}{n} \end{aligned}$$

- La **correction pour population finie** disparaît à cause de $\mathbb{V}_\xi \mathbb{E}_P(\widehat{\theta} \mid \boldsymbol{Y})$

Échantillonnage aléatoire simple sans remise

- Le terme $(1 - n/N)$ est la principale raison du biais de l'estimateur de variance basé sur le plan
- Le biais est $-\sigma^2 N^{-1} = -\mathbb{V}_\xi \mathbb{E}_P(\widehat{\theta}\,)$ à cause du terme $(1 - n/N)$
- **L'interprétation correcte de $(1 - n/N)$ est**

> Le rôle du terme $(1-n/N)$ est de réduire la variance de $\mathbb{V}_\xi \mathbb{E}_P(\widehat{\theta}\,)$, afin de compenser le fait que la quantité d'intérêt est θ_N, plutôt que θ_0

Correction de population finie?

- Cette correction n'est pas due au fait que N soit fini, car $(1 - n/N)$ doit être remplacé par 1, lorsque θ_0 est le paramètre cible, même lorsque N est fini.
- Cette correction est nécessaire lorsque θ_N est le paramètre cible.
- Qualifier $(1 - n/N)$ de **"correction de population finie" est trompeur**, car cela n'a rien à voir avec le fait que N soit finie.

Correction de population finie?

- Qualifier $(1 - n/N)$ de **"correction de population finie" est trompeur**

> Nous devrions l'appeler "*Correction de population fixe*". Le terme "*fixe*" est utilisé pour souligner que la correction doit être appliquée lorsque θ_N est la cible et que nous devons utiliser une variance conditionnelle étant donné $\boldsymbol{Y}$, en traitant les données $\boldsymbol{Y}$ comme fixes.

- "*Correction pour grande fraction d'échantillonnage*" est aussi adéquat
- "*L'inférence en population finie*" doit être comprise comme le fait que θ_N est la cible et que $\boldsymbol{Y}$ est considéré comme fixe (conditionnement sur $\boldsymbol{Y}$). L'expression "*inférence en population fixe*" est plus appropriée

Généralisation

- Les principes développés jusqu'à présent s'appliquent à des **plans d'échantillonnage plus élaborés**, tels que
  - Échantillonnage à probabilité inégale <u>sans remise</u>
  - Échantillonnage stratifiée
  - Échantillonnage à plusieurs degré
  - Échantillonnage systématique ordonné
- Nous supposons un **échantillonnage informatif**
- **Inférence analytique**, car nous nous intéressons a θ_0
- Extension aux **variables auxiliaires** : voir l'article
- Échantillonnage de Poisson informatif : voir l'article

Échantillonnage à probabilité inégale sans remise

- $(y_i, \pi_i)^\top \sim$ distribution bivariée ξ
- θ_0 est la solution (supposée unique) de

$$\mathbb{E}_\xi\left[g(y,\theta)\right] = 0 \quad \text{iff } \theta = \theta_0\cdot$$

- $\widehat{\theta}$ solution (supposée unique) de l'équivalent empirique

$$\widehat{\Gamma}(\theta) := \sum_{i \in S} \frac{1}{\pi_i} g(y_i, \theta) = 0\cdot$$

- Soit $\varrho_i := g(y_i, \theta_0)$.
- π_i **et** ϱ_i **peuvent être dépendant** (échantillonnage informatif).

Hypothèses

- $N/n < \infty$ est borné, lorsque $n \to \infty$
- $d := \sum_{i \in \mathcal{U}} \pi_i(1-\pi_i) \to \infty$
- $\pi_{ij} = \pi_i \Big[\pi_j \Big\{1 - \big[1+o(1)\big](1-\pi_i)(1-\pi_j)d^{-1}\Big\}\Big]^{1-\delta_{ij}}$,
  où δ_{ij} est le delta de Kronecker

  Entropie élevée. Expression la plus courante pour π_{ij} (Hájek, 1964)
- **Indépendance** $\varrho_i \perp\!\!\!\perp \varrho_j$ et $\pi_i \perp\!\!\!\perp \pi_j$, pour $i \neq j$, où $\varrho_i := g(y_i, \theta_0)$.

  Indépendance entre les ϱ_i, pas entre les y_i !
- $\frac{1}{N} \sum_{i \in \mathcal{U}} \mathbb{V}_\xi(\varrho_i) < \infty$.
- $\pi_{ij} \geqslant \pi_i \pi_j$, a.s. $\forall i, j$.
- Hypothèses asymptotiques standard dans un contexte de population finie

Estimation de la variance

- Considérons **l'estimateur de variance de Hansen & Hurwitz (1943)**:

$$\widehat{\mathbb{V}}[\widehat{\Gamma}(\theta)] = \sum_{i\in S} \frac{1}{\pi_i^2} g(y_i,\theta)^2 - \frac{1}{n}\left\{\sum_{i\in S}\frac{1}{\pi_i} g(y_i,\theta)\right\}^2.$$

- Asymptotiquement **sans biais**:

$$\frac{n}{N^2}\left\{\mathbb{E}_\xi \mathbb{E}_P \widehat{\mathbb{V}}[\widehat{\Gamma}(\theta_0)] - \mathbb{V}[\widehat{\Gamma}(\theta_0)]\right\} = o(1)$$

- Il n'est pas nécessaire d'estimer séparément $\mathbb{E}_\xi \mathbb{V}_P[\widehat{\Gamma}(\theta) \mid \boldsymbol{Y}]$ et $\mathbb{V}_\xi \mathbb{E}_P[\widehat{\Gamma}(\theta) \mid \boldsymbol{Y}]$

  $\Longrightarrow$ Théorème de Taylor:

$$\widehat{\mathbb{V}}(\widehat{\theta}) := \left\{\frac{\partial \widehat{\Gamma}(\widehat{\theta})}{\partial \widehat{\theta}}\right\}^{-2} \widehat{\mathbb{V}}[\widehat{\Gamma}(\widehat{\theta})],$$

L'estimateur de variance de Hansen & Hurwitz (1943)

- L'estimateur de variance de Hansen & Hurwitz (1943) est sans biais sous le plan, lorsque l'échantillon est sélectionné avec remise et que θ_N est la cible.

- Cependant, lorsque θ_0 est la cible, cet estimateur est **biaisé** sous un plan d'échantillonnage **avec remise**

- Cet estimateur n'est pas limité à l'échantillonnage avec remise

- **Il est asymptotiquement sans biais sous échantillonnage sans remise avec de grandes fractions d'échantillonnage, lorsque θ_0 est la cible**

L'estimateur de variance de Hansen & Hurwitz (1943)

- L'absence de FPC et de π_{ij} n'est pas due au fait que n/N est asymptotiquement négligeable, ni à une hypothèse "*avec remise*"

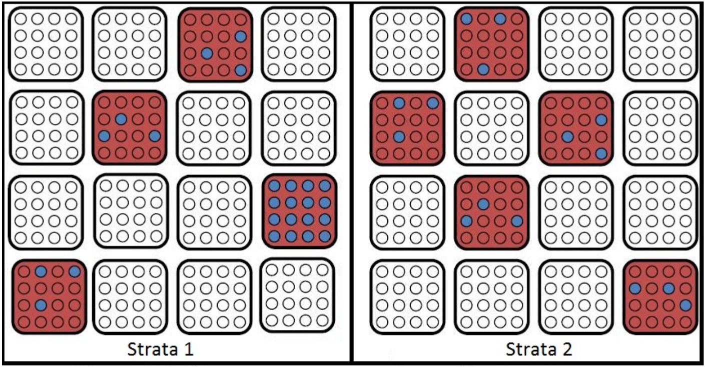
- La variance est asymptotiquement sans biais car θ_0 est la cible

- **Interprétation de π_{ij}**:

> Les π_{ij} sont nécessaires lorsque θ_N est la cible, et non parce que la population est finie

Échantillonnage stratifié à plusieurs degrés

- Un échantillon S de n grappes est sélectionné sans remise avec des probabilités inégales π_i
- Au sein de chaque grappe $\mathcal{U}_i$ échantillonnée ($i \in S$), un échantillon S_i de m_i unités est sélectionné



Échantillonnage stratifié à plusieurs degrés

- $\boldsymbol{\theta}_0 \in \mathbb{R}^{d_\theta}$ est défini par une "*condition de moment multivariée*", c'est-à-dire

$$\mathbb{E}_\xi\{\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{\theta})\} = \mathbf{0}_{d_g}, \quad \text{iff } \boldsymbol{\theta} = \boldsymbol{\theta}_0,$$

- $\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{\theta}) \in \mathbb{R}^{d_g}$ et $\mathbf{0}_d$ est le d-vecteur de 0.

- Le modèle peut être sur-spécifié, c'est-à-dire $d_g \geqslant d_\theta$, afin d'inclure des contraintes de calibration (voir l'article).

- L'estimateur ponctuel $\widehat{\boldsymbol{\theta}}$ est la solution de

$$\widehat{\boldsymbol{\Gamma}}(\boldsymbol{\theta}) := \sum_{i \in S} \frac{1}{\pi_i} \widehat{\boldsymbol{\rho}}_i(\boldsymbol{\theta}) = \mathbf{0}_{d_g},$$

avec

$$\widehat{\boldsymbol{\rho}}_i(\boldsymbol{\theta}) := \sum_{j \in S_i} \frac{1}{\pi_{j|i}} \boldsymbol{g}_{j|i}(\boldsymbol{\theta}), \qquad \boldsymbol{g}_{j|i}(\boldsymbol{\theta}) := \boldsymbol{g}(\boldsymbol{y}_j, \boldsymbol{\theta}), \quad \text{pour } j \in \mathcal{U}_i,$$

Estimateur de variance stratifié "*grappe ultime*" de Hansen & Hurwitz (1943)

$$\widehat{\mathbb{V}}\{\widehat{\boldsymbol{\Gamma}}(\boldsymbol{\theta})\} := \sum_{h=1}^{H} \widehat{\mathbb{V}}\{\widehat{\boldsymbol{\Gamma}}_h(\boldsymbol{\theta})\}, \quad \text{avec } \widehat{\boldsymbol{\Gamma}}_h(\boldsymbol{\theta}) := \sum_{i \in S_h} \frac{1}{\pi_i} \widehat{\boldsymbol{\rho}}_i(\boldsymbol{\theta}),$$

où

$$\widehat{\mathbb{V}}\{\widehat{\boldsymbol{\Gamma}}_h(\boldsymbol{\theta})\} := \sum_{i \in S_h} \frac{1}{\pi_i^2} \widehat{\boldsymbol{\rho}}_i(\boldsymbol{\theta})\, \widehat{\boldsymbol{\rho}}_i(\boldsymbol{\theta})^\top - \frac{1}{n_h} \widehat{\boldsymbol{\Gamma}}_h(\boldsymbol{\theta}) \widehat{\boldsymbol{\Gamma}}_h(\boldsymbol{\theta})^\top,$$

- Asymptotiquement sans biais $\Longrightarrow$ Théorème de Taylor:

$$\widehat{\mathbb{V}}(\widehat{\boldsymbol{\theta}}) := \left\{ \frac{\partial \widehat{\boldsymbol{\Gamma}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \right\}^{-1} \widehat{\mathbb{V}}\{\widehat{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{\theta}})\} \left\{ \frac{\partial \widehat{\boldsymbol{\Gamma}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \right\}^{-1\top}.$$

- **Inutile d'estimer les variances a l'intérieur des grappes** (Gustave (Chevalier & Richer, 2023)), même lorsque n/N est grand

Hypothèse clef pour l'absence de biais asymptotique

- **Indépendance entre les grappes**, c'est-à-dire

$$\boldsymbol{g}_{j|i}(\boldsymbol{\theta}_0) \perp\!\!\!\perp \boldsymbol{g}_{\ell|k}(\boldsymbol{\theta}_0) \text{ et } \pi_i \perp\!\!\!\perp \pi_k,\ \forall\ j \in \mathcal{U}_i,\ \ell \in \mathcal{U}_k \text{ et } i \neq k.$$

avec $\boldsymbol{g}_{j|i}(\boldsymbol{\theta}) := \boldsymbol{g}(\boldsymbol{y}_j, \boldsymbol{\theta})$

- Nous pourrions avoir une **dépendance au sein des grappes**

- Entropie élevée (Hájek, 1964)

- Il n'est pas nécessaire d'utiliser des effets aléatoires ni d'estimer la variance au sein des grappes

Ordered systematic sampling

- S'il y a une tendance, alors les y_i ne peuvent pas être i.i.d.
- L'hypothèse d'entropie élevée est erronée, car certains des π_{ij} sont nuls
- Une solution consiste à sur-spécifier la fonction d'estimation en se basant sur un modèle de tendance, tel que :

$$y_i = \alpha_0 + \ell_i \beta_0 + e_i,$$

où ℓ_i est le label à l'intérieur de U de l'unité $i \in S$

Échantillonnage systématique ordonné

- Une **fonction d'estimation sur-spécifiée** peut être utilisée, c'est-à-dire

$$\boldsymbol{g}(y_i, \boldsymbol{\theta}) = \left\{\varepsilon(y_i, \boldsymbol{\theta}) - \mu + \alpha + \mu_\ell \beta, \varepsilon(y_i, \boldsymbol{\theta}), \ell_i\, \varepsilon(y_i, \boldsymbol{\theta})\right\}^\top,$$

avec
$$\begin{aligned}
\varepsilon(y_i, \boldsymbol{\theta}) &:= y_i - \alpha - \ell_i \beta, \\
\boldsymbol{\theta} &:= (\mu, \alpha, \beta)^\top, \\
\mu_\ell &:= (N+1)/2, \quad \text{la moyenne de la population des } \ell_i, \\
\mu_0 &:= \mathbb{E}_\xi(y), \quad \text{le paramètre cible,} \\
(\alpha_0, \beta_0)^\top & \quad \text{un paramètre de nuisance}
\end{aligned}$$

- **$\boldsymbol{g}(y_i, \theta_0)$ sont i.i.d., malgré le fait que les y_i ne soient pas i.i.d.**

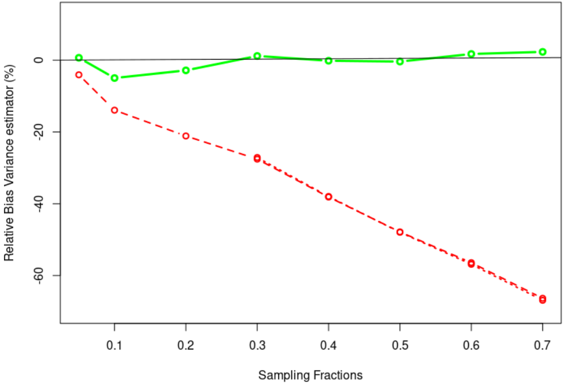
- L'échantillonnage systématique aléatoire et l'échantillonnage systématique ordonné sont équivalents, car i.i.d.
$\Longrightarrow$ Entropie large $\Longrightarrow$ variance de Hansen & Hurwitz (1943)

Numerical example

- $y_i \sim N(10, sd = 2) \Rightarrow \theta_0 = 10$
- $p(S) =$ échantillonnage de Chao
- $n = 100, \qquad 0.05 \leqslant f = \dfrac{n}{N} \leqslant 0.7$
- $0 \leqslant cor(y_i, \pi_i) \leqslant 0.9$
- Population finie $N = n/f$

Biais relatifs

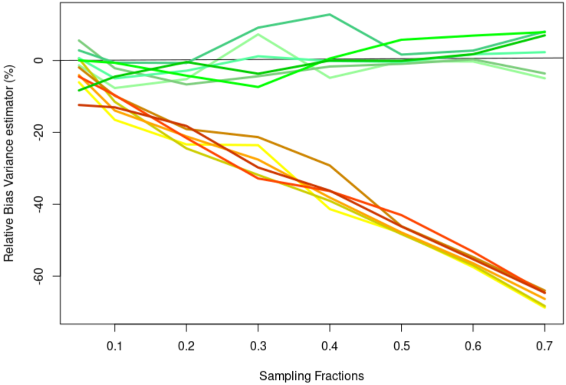
$cor(y_i, \pi_i) = 0.4$



Variance de Hansen & Hurwitz (1943)
Variance de Horvitz & Thompson (1952): $\widehat{\mathbb{V}_P}(\widehat{\theta})$

Biais relatifs pour différentes corrélations

La corrélation n'a aucun effet sur les biais.



Variance de Hansen & Hurwitz (1943)

Variance de Horvitz & Thompson (1952): $\widehat{\mathbb{V}_P}(\widehat{\theta})$

Échantillonnage à deux degrés

- Données générées à partir d'un modèle à effets aléatoires

$$\mathcal{Y}_{j|i} = \alpha_0 + \beta_0\, x_{j|i} + \epsilon_{j|i}, \quad \text{avec } j \in \mathcal{U}_i \text{ et } \alpha_0 = \beta_0 = 1,$$
$$\epsilon_{j|i} = u_i + e_{ji}.$$

- Supposons que nous souhaitions tester $H_0 : \boldsymbol{\theta}_0 = \widetilde{\boldsymbol{\theta}}$ contre $H_a : \boldsymbol{\theta}_0 \neq \widetilde{\boldsymbol{\theta}}$, où $\widetilde{\boldsymbol{\theta}} = (1, 1)^\top$. Niveau $\alpha = 5\%$

- Statistique de pivot:

$$r(\widetilde{\boldsymbol{\theta}}) := \widehat{\boldsymbol{\Gamma}}(\widetilde{\boldsymbol{\theta}})^\top \widehat{\mathbb{V}}\{\widehat{\boldsymbol{\Gamma}}(\widetilde{\boldsymbol{\theta}})\}^{-1} \widehat{\boldsymbol{\Gamma}}(\widetilde{\boldsymbol{\theta}})$$

  tend en distribution vers une distribution χ^2 avec d_θ degrés de liberté, sous H_0.

- Plus simple que les ajustements de Rao & Scott (1987)

Échantillonnage à deux degrés. Tailles observées (%)

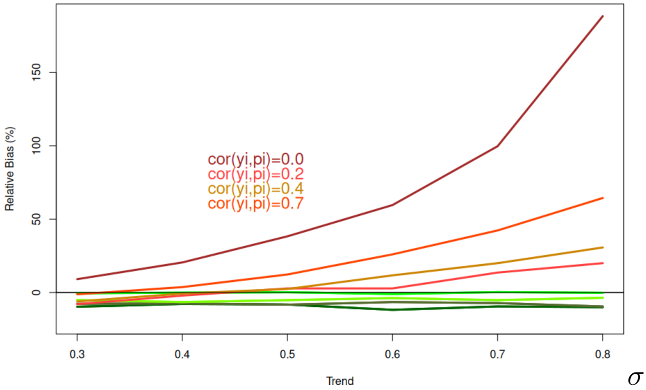
Statistique de test	ρ	$f = n/N$							
		0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7
Proposé	0.0	6.1	6.3	5.5	5.9	4.5	5.5	6.1	5.8
	0.2	5.6	5.3	5.1	5.2	4.4	5.9	5.5	5.4
	0.4	6.6†	4.3	6.3	5.3	4.9	5.6	5.0	5.1
	0.6	5.1	5.5	6.2	6.3	6.0	6.1	5.0	5.1
	0.8	5.6	4.3	5.7	5.6	5.6	5.6	5.4	4.8
	0.9	4.5	4.6	4.3	6.8†	6.2	5.6	4.9	5.5
Rao & Scott 1	0.0	5.6	6.5†	5.2	5.5	4.4	5.5	5.6	5.0
	0.2	5.6	5.3	5.3	4.0	5.9	5.1	6.0	5.4
	0.4	6.6†	5.9	7.1†	7.0†	5.5	8.3†	8.2†	5.7
	0.6	6.5†	7.1†	8.0†	8.3†	7.1†	8.6†	6.9†	9.8†
	0.8	8.1†	6.4†	8.3†	9.2†	7.8†	8.1†	9.6†	8.0†
	0.9	6.1	8.2†	7.4†	8.5†	8.5†	7.8†	8.3†	8.5†
Rao & Scott 2	0.0	5.6	6.5†	5.2	5.5	4.4	5.5	5.6	5.0
	0.2	5.6	5.3	3.6†	3.2†	5.9	5.1	6.0	5.4
	0.4	6.6†	3.5†	7.1†	7.0†	3.1†	8.3†	8.2†	3.4†
	0.6	6.5†	7.1†	8.0†	8.3†	7.1†	8.6†	6.9†	9.8†
	0.8	8.1†	2.4†	8.3†	9.2†	7.8†	8.1†	9.6†	8.0†
	0.9	6.1	8.2†	3.2†	8.5†	8.5†	7.8†	8.3†	8.5†

† Taille significativement différente de 5%: p-valeur < 0.05.

Échantillonnage systématique ordonné

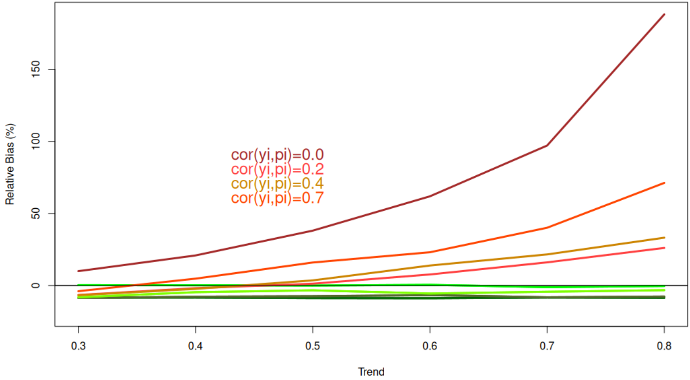
- $y_i = 5 + (i-1)(N-1)^{-1} + \varepsilon_i$, où $\varepsilon_i \sim N(0, \sigma)$ i.i.d.
- C'est-à-dire y_i suit une tendance linéaire.
- La valeur différente de σ permet de contrôler la force de la tendance
- Les corrélations $\rho_{y\pi}$ entre π_i et y_i sont de $0,2$, $0,4$ et $0,7$.
- $n = 500$
- Population de taille $N = \text{round}(n/f)$, où $f = 0.05$, 0.2 et 0.4.

Biais relatifs pour différentes corrélations: $n/N = 0.05$



Variance de Hansen & Hurwitz (1943) surspécifié
Variance de Hansen & Hurwitz (1943)

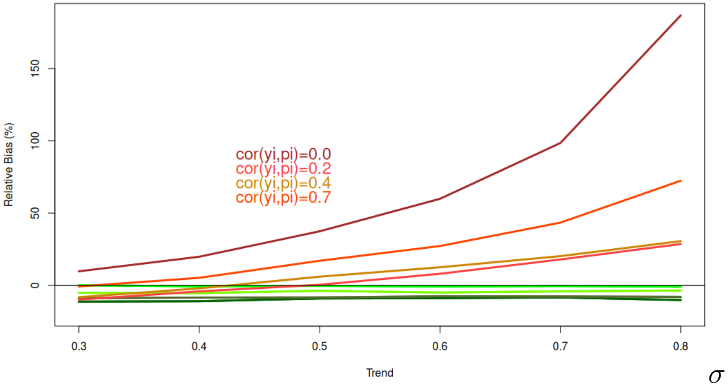
Biais relatifs pour différentes corrélations: $n/N = 0.2$



σ

Variance de Hansen & Hurwitz (1943) surspécifié
Variance de Hansen & Hurwitz (1943)

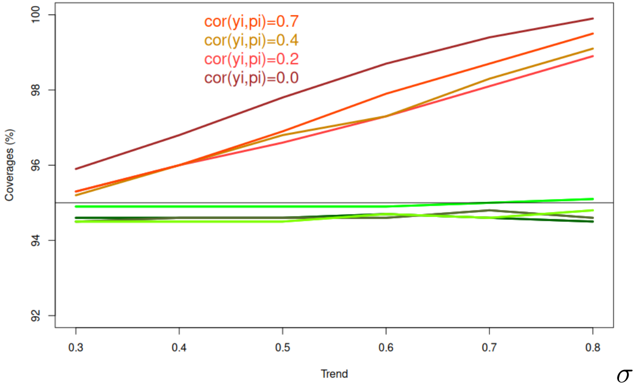
Biais relatifs pour différentes corrélations: $n/N = 0.4$



Variance de Hansen & Hurwitz (1943) surspécifié

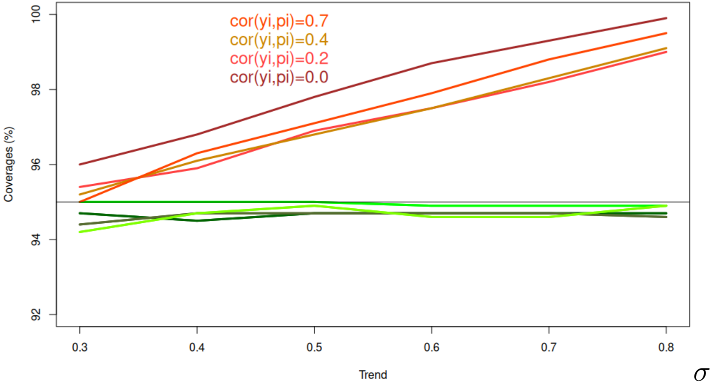
Variance de Hansen & Hurwitz (1943)

Couvertures pour différentes corrélations: $n/N = 0.05$



Variance de Hansen & Hurwitz (1943) surspécifié
Variance de Hansen & Hurwitz (1943)

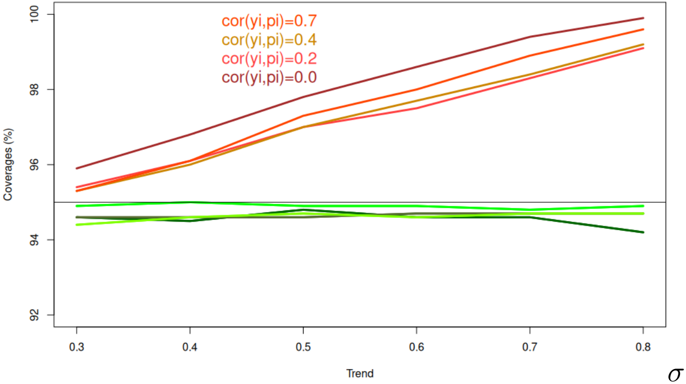
Couvertures pour différentes corrélations: $n/N = 0.2$



Variance de Hansen & Hurwitz (1943) surspécifié

Variance de Hansen & Hurwitz (1943)

Couvertures pour différentes corrélations: $n/N = 0.4$



Variance de Hansen & Hurwitz (1943) surspécifié
Variance de Hansen & Hurwitz (1943)

Conclusion

- Le choix entre les approches basées sur le plan et les approches basées sur les modèles ne doit pas être guidé par
  - Le fait que l'on suppose les y_i fixes ou aléatoires
  - Le fait que l'on traite l'échantillon comme fixe ou aléatoire
- Ces questions sont hors de propos

Conclusion

Deux questions clefs:

- L'échantillonnage est-il informatif ?
- Quel est le paramètre cible ? $\boldsymbol{\theta}_N$ ou $\boldsymbol{\theta}_0$?

Conclusion

Deux questions clefs:

- L'échantillonnage est-il informatif ?
- Quel est le paramètre cible ? $\boldsymbol{\theta}_N$ ou $\boldsymbol{\theta}_0$?

- **Si l'échantillonnage n'est pas informatif**: variance basée sur le plan, si la cible est $\boldsymbol{\theta}_N$ ou variance basée sur le modèle (avec des hypothèses de distribution supplémentaires) si la cible est $\boldsymbol{\theta}_N$ ou $\boldsymbol{\theta}_0$

Conclusion

Deux questions clefs:

- L'échantillonnage est-il informatif ?
- Quel est le paramètre cible ? $\boldsymbol{\theta}_N$ ou $\boldsymbol{\theta}_0$?

- **Si l'échantillonnage n'est pas informatif**: variance basée sur le plan, si la cible est $\boldsymbol{\theta}_N$ ou variance basée sur le modèle (avec des hypothèses de distribution supplémentaires) si la cible est $\boldsymbol{\theta}_N$ ou $\boldsymbol{\theta}_0$
- **Si l'échantillonnage est informatif et la cible est θ_N**
  $\Longrightarrow$ Variance traditionnels basés sur le plan (Gustave (Chevalier & Richer, 2023))

Conclusion

Deux questions clefs:

- L'échantillonnage est-il informatif ?
- Quel est le paramètre cible ? $\boldsymbol{\theta}_N$ ou $\boldsymbol{\theta}_0$?

- **Si l'échantillonnage n'est pas informatif**: variance basée sur le plan, si la cible est $\boldsymbol{\theta}_N$ ou variance basée sur le modèle (avec des hypothèses de distribution supplémentaires) si la cible est $\boldsymbol{\theta}_N$ ou $\boldsymbol{\theta}_0$
- **Si l'échantillonnage est informatif et la cible est $\boldsymbol{\theta}_N$**
  $\Longrightarrow$ Variance traditionnels basés sur le plan (Gustave (Chevalier & Richer, 2023))
- **Si l'échantillonnage est informatif et la cible est $\boldsymbol{\theta}_0$**
  $\Longrightarrow$ **Variance de Hansen & Hurwitz (1943)**
  $\Longrightarrow$ Variance basé sur le plan n'a aucun sens et est beaucoup plus compliquée

Conclusion

Le rôle du terme $(1-n/N)$ est de réduire la variance de $\mathbb{V}_\xi \mathbb{E}_P(\widehat{\theta})$, afin de compenser le fait que la quantité d'intérêt est θ_N, plutôt que θ_0

- Qualifier $(1 - n/N)$ de "*correction pour population finie*" est trompeur, car cela n'a rien à voir avec le fait que la population est finie.
- Nous devrions l'appeler **"correction de population fixe"**. Le mot "*Fixe*" est utilisé pour souligner que la correction doit être utilisée lorsque θ_N est la cible et que nous devons utiliser une variance conditionnelle étant donné $\boldsymbol{Y}$, en traitant les données $\boldsymbol{Y}$ et θ_N comme étant fixes
- Le fait que N soit fini est sans importance

Autres sujets abordés dans l'article

- Variables auxiliaires
- Variance de l'estimateur de régression tenant compte de l'estimation du paramètre de régression
- Échantillonnage informatif de Poisson

References

Berger, Y. G. (2025) Unconditional variance estimation under complex surveys. International Statistical Review, 26pp.

`https://onlinelibrary.wiley.com/doi/10.1111/insr.70001.`

Chevalier, M. & Richer, J.-S. (2023) *gustave: User-Oriented Statistical Toolkit for Analytical Variance Estimation*. URL `https://CRAN.R-project.org/package=gustave`. R package version 0.3.0.

Hájek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 1491–1523. URL `https://doi.org/10.1214/aoms/1177700375`.

Hansen, M. H. & Hurwitz, W. N. (1943) On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, **14**, pp. 333–362.

Horvitz, D. G. & Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685. URL `https://doi.org/10.1080/01621459.1952.10483446`.

Rao, J. N. K. & Scott, A. J. (1987) On simple adjustments to chi-square tests with sample survey data. *The Annals of Statistics*, **15**, 385–397.