Estimation de variance non-paramétrique sous une stratification fine : une alternative au regroupement de strates

F. Jay Breidt
Colorado State University

7ème colloque francophone sur les sondages
Ensai, Bruz, France
Novembre 2012

Travail en collaboration avec Ismael Sanchez Borrego et Jean Opsomer. Ce travail de recherche a été partiellement financé par la US National Science Foundation (SES0922142).

Stratification

- En utilisant une information disponible sur la base de sondage, la population U est partitionnée en sous-populations $\{U_i\}$, appelées *strates*
- Dans chaque strate i, on sélectionne un échantillon $s_i \subset U_i$ selon un plan de sondage $p_i(\cdot)$
- Les échantillons sont tirés indépendamment d'une strate à l'autre
- Fréquemment utilisé dans presque toutes les enquêtes

Pourquoi stratifier? #1. Flexibilité du plan de sondage

- Des sous-populations différentes peuvent disposer de différentes informations auxiliaires
  - e.g., base d'unités pour une sous-population, mais seulement une base aréolaire pour une autre sous-population
- Des sous-populations différentes peuvent présenter différents problèmes d'erreurs non dues à l'échantillonnage
  - différents taux de réponse
  - différentes erreurs de mesure
  - ...

Pourquoi stratifier? #2. Simplicité administrative

- La charge de travail peut être naturellement répartie par les limites de strates
- Différentes forces de travail, différentes administrations
  - peut simplifier la logistique
  - peut éviter des déplacements coûteux, etc.

Pourquoi stratifier? #3. Gains de précision

- Si une population hétérogène peut être répartie en sous-populations homogènes, on obtient une meilleure précision
- La stratification améliore souvent également l'approximation normale

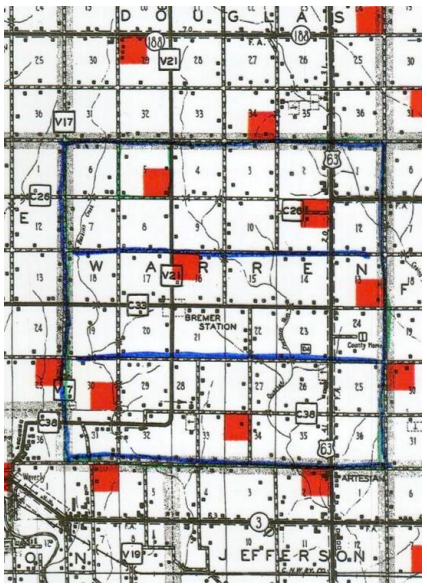
Pourquoi stratifier? #4. Contrôle des domaines d'étude 6

- Taille d'échantillon maîtrisée pour des sous-populations (*domaines*) d'intérêt
  - si la sous-population peut être identifiée dans la base de sondage, alors cette sous-population peut être utilisée comme une strate
  - sinon, la taille d'échantillon dans le domaine est une variable aléatoire et peut être très petite
  - la stratification peut aider à éviter certains de ces problèmes d'estimation sur petits domaines

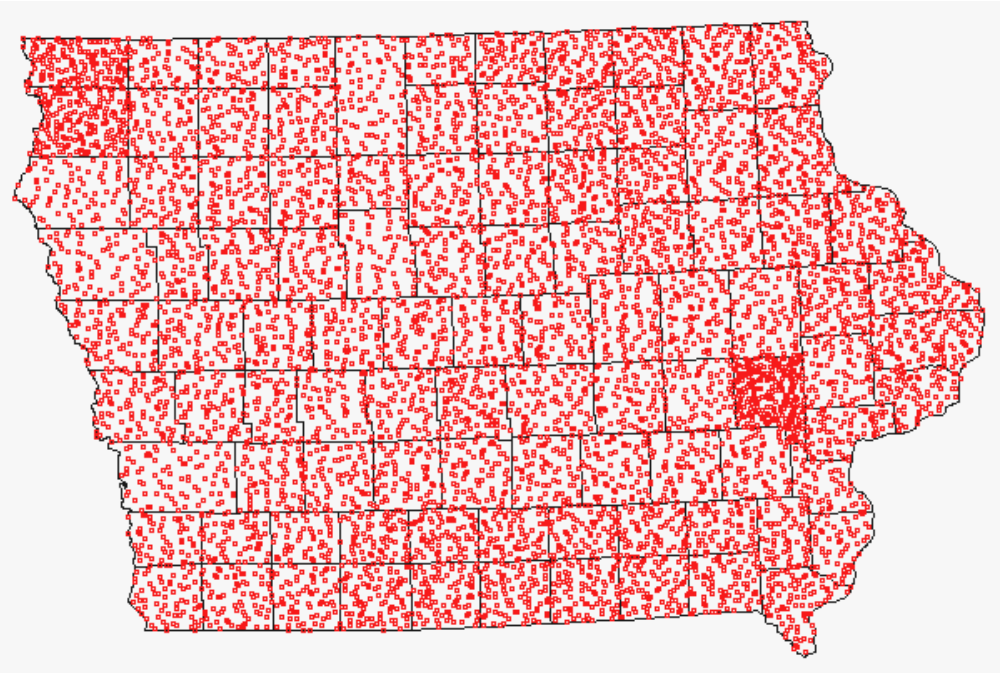
- Les grandes strates (pour #1 la flexibilité du plan ou #2 par simplicité administrative) peuvent être découpées en strates plus petites
- Les points #3 et #4 suggèrent de petites strates :
  - relativement homogènes, améliorant la précision
  - permettent un contrôle des tailles d'échantillon dans les domaines planifiés ou non planifiés, permettant d'éviter des problèmes d'estimation sur petits domaines
- En pratique, cela mène souvent à une stratification fine
  - p unités par strate, avec p petit; H = nombre de strates grand

USDA National Resources Inventory, Iowa

- Stratification fine : deux unités échantillonnées par 1/3 de quartier
- Randomisation contrôlée pour disperser les unités tirées dans les strates



USDA National Resources Inventory, Iowa (2)

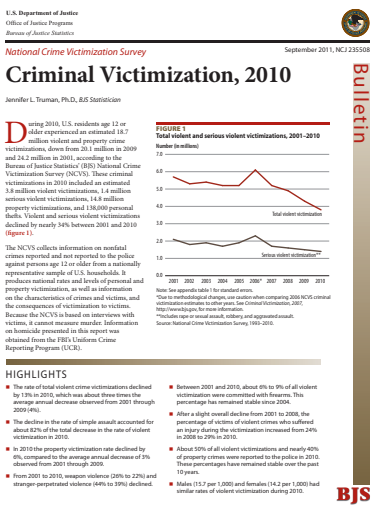


- La Current Population Survey est une enquête mensuelle auprès des ménages conduite par le Bureau of Census pour le Bureau of Labor Statistics.
- C'est la plus importante des enquêtes états-uniennes sur l'emploi :

  ...personnes actives, emploi, chômage, personnes inactives, heures de travail, revenus, et d'autres caractéristiques démographiques et de l'emploi

Tirage multi-degrés avec une unité par strate : NCVS

- Enquête nationale de victimation
  - conduite par le U.S. Census Bureau pour le U.S. Bureau of Justice Statistics



U.S. Department of Justice
Office of Justice Programs
Bureau of Justice Statistics

National Crime Victimization Survey

September 2011, NCJ 235508

Criminal Victimization, 2010

Jennifer L. Truman, Ph.D., BJS Statistician

During 2010, U.S. residents age 12 or older experienced an estimated 18.7 million violent and property crime victimizations, down from 20.1 million in 2009 and 24.2 million in 2001, according to the Bureau of Justice Statistics' (BJS) National Crime Victimization Survey (NCVS). These criminal victimizations in 2010 included an estimated 3.8 million violent victimizations, 1.4 million serious violent victimizations, 14.8 million property victimizations, and 138,000 personal thefts. Violent and serious violent victimizations declined by nearly 34% between 2001 and 2010 (figure 1).

The NCVS collects information on nonfatal crimes reported and not reported to the police against persons age 12 or older from a nationally representative sample of U.S. households. It produces national rates and levels of personal and property victimization, as well as information on the characteristics of crimes and victims, and the consequences of victimization to victims. Because the NCVS is based on interviews with victims, it cannot measure murder. Information on homicide presented in this report was obtained from the FBI's Uniform Crime Reporting Program (UCR).

FIGURE 1
Total violent and serious violent victimizations, 2001–2010

Note: See appendix table 1 for standard errors.
*Due to methodological changes, use caution when comparing 2006 NCVS criminal victimization estimates with estimates from other Criminal Victimization, 2007. http://www.bjs.gov for more information.
**Includes rape or sexual assault, robbery, and aggravated assault.
Source: National Crime Victimization Survey, 1993–2010.

HIGHLIGHTS

- The rate of total violent crime victimizations declined by 13% in 2010, which was about three times the average annual decrease observed from 2001 through 2009 (4%).
- The decline in the rate of simple assault accounted for about 82% of the total decrease in the rate of violent victimization in 2010.
- In 2010 the property victimization rate declined by 6%, compared to the average annual decrease of 5% observed from 2001 through 2009.
- From 2001 to 2010, weapon violence (26% to 22%) and stranger-perpetrated violence (44% to 39%) declined.
- Between 2001 and 2010, about 6% to 9% of all violent victimization were committed with firearms. This percentage has remained stable since 2004.
- After a slight overall decline from 2001 to 2008, the percentage of victims of violent crimes who suffered an injury during the victimization increased from 24% in 2008 to 29% in 2010.
- About 53% of all violent victimizations and nearly 40% of property crimes were reported to the police in 2010. These percentages have remained stable over the past 10 years.
- Males (15.7 per 1,000) and females (14.2 per 1,000) had similar rates of violent victimization during 2010.

Bulletin

BJS

For a list of all publications in this series, go to http://www.bjs.gov/index.cfm?ty=pbse&sid=32.

Quelques notations

- Population finie $U = \{1, 2, \ldots, N\} = \cup_{i=1}^{H} U_i$
- Echantillon probabiliste stratifié $s = \cup_{i=1}^{H} s_i$, $s_i \subset U_i$
- Probabilités d'inclusion $\pi_j = \mathsf{Pr}\,(j \in s)$ et $\pi_{jk} = \mathsf{Pr}\,(j, k \in s)$
- Total en population finie

$$t = \sum_{j \in U} y_j = \sum_{i=1}^{H} \sum_{j \in U_i} y_j = \sum_{i=1}^{H} t_i$$

- Indicatrices d'appartenance à l'échantillon : $I_j = 1$ si $j \in s$, $I_j = 0$ si $j \notin s$

- L'estimateur de Horvitz-Thompson (1952)

$$\hat{t} = \sum_{i=1}^{H} \hat{t}_i = \sum_{i=1}^{H} \sum_{j \in U_i} y_j \frac{I_j}{\pi_j}$$

est non-biaisé pourvu que $\pi_j > 0$ pour tout $j \in U$

- La variance est donnée par

$$\text{Var}\left(\hat{t}\right) = \sum_{i=1}^{H} \text{Var}\left(\hat{t}_i\right) = \sum_{i=1}^{H} V_i$$

avec

$$V_i = \sum_{j,k \in U_i} \sum (\pi_{jk} - \pi_j \pi_k) \frac{y_j y_k}{\pi_j \pi_k}$$

- Si le plan de sondage est mesurable ($\pi_{jk} > 0$ pour tout $j, k \in U$), alors

$$\sum_{i=1}^{H} \widehat{V}_i = \sum_{i=1}^{H} \sum\sum_{j,k \in s_i} \frac{(\pi_{jk} - \pi_j \pi_k)}{\pi_{jk}} \frac{y_j y_k}{\pi_j \pi_k}$$

  est non-biaisé pour Var $\left(\hat{t}\right)$

- Un plan de sondage avec une unité tirée par strate est non-mesurable :

$$\pi_{jk} = \begin{cases} \pi_j, & \text{si } j, k \in U_i,\ j = k; \\ 0, & \text{si } j, k \in U_i,\ j \neq k; \\ \pi_j \pi_k, & \text{sinon} \end{cases}$$

- Supposons que H est pair et définissons

$$c_j(i) = \begin{cases} 1, & \text{si } i, j : i \neq j \text{ dans la même strate regroupée ;} \\ 0, & \text{sinon} \end{cases}$$

- L'estimateur de variance par regroupement des strates est

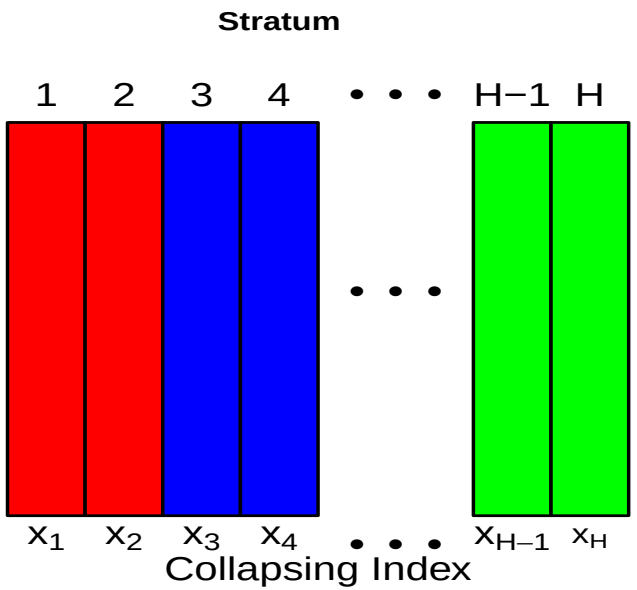
$$\widehat{V}_{coll} = \frac{1}{2} \sum_{i=1}^{H} \left(\hat{t}_i - \sum_{j=1}^{H} c_j(i) \hat{t}_j \right)^2$$

Regroupement de strates en pratique

- Souvent utilisé même quand on sélectionne plus d'une unité dans chaque strate
  - assure l'existence d'estimateurs de variance dans les domaines et améliore leur stabilité
  - Rust et Kalton (1987) comparent des estimateurs de variance construits en regroupant par paires, par triplets ou par groupes plus grands
- Le regroupement utilise souvent un index x_i connu pour chaque strate ; e.g., une localisation spatiale

Regroupement de strates avec un index, x_i

- La fonction de regroupement $c_j(i)$ rassemble "des strates proches" au sens des distances entre les x_i :



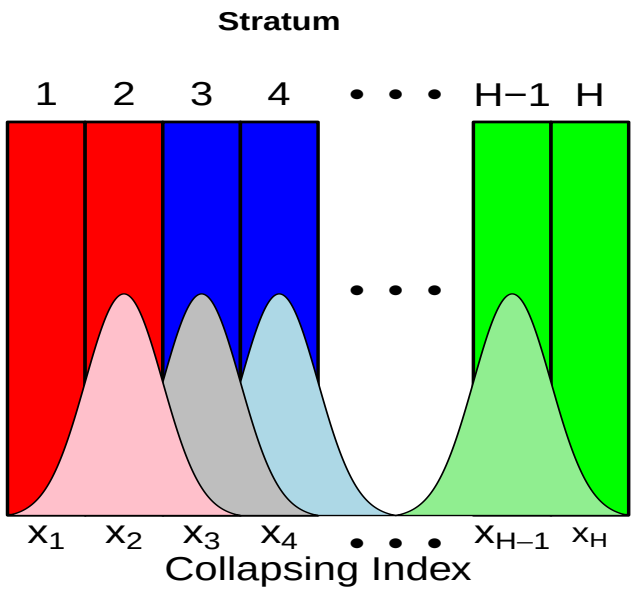
- L'espérance sous le plan vaut

$$\mathsf{E}\left[\widehat{V}_{coll}\right] = \mathsf{Var}\left(\hat{t}\right) + \frac{1}{2}\sum_{i=1}^{H}\left(t_i - \sum_{j=1}^{H} c_j(i)t_j\right)^2 ,$$

  si bien que $\widehat{V}_{coll}$ présente un biais positif
- Le couplage doit être réalisé indépendamment de toute information sur l'échantillon
- Le biais est faible si le couplage est efficace : $t_i \simeq t_j$ quand $c_j(i) = 1$

Une alternative: "regroupement par noyaux"

- Groupes discrets remplaçés par des alternatives continues basées sur des moyennes pondérées par des noyaux :



- On remplace la fonction binaire

$$c_j(i) = \begin{cases} 1, & \text{si } i, j : i \neq j \text{ dans la même strate regroupée;} \\ 0, & \text{sinon} \end{cases}$$

par des poids de noyaux

$$d_j(i) = \frac{K\left(\frac{x_i - x_j}{h}\right)}{\sum_{k=1}^{H} K\left(\frac{x_i - x_k}{h}\right)},$$

  - $K(\cdot)$ noyau borné, symétrique
  - h largeur de la fenêtre

Estimation de variance non-paramétrique, II

- Comme alternative à l'estimateur de variance par regroupement des strates

$$\widehat{V}_{coll} = \frac{1}{2}\sum_{i=1}^{H}\left(\hat{t}_i - \sum_{j=1}^{H} c_j(i)\hat{t}_j\right)^2,$$

considérons l'estimateur de variance "non-paramétrique"

$$\widehat{V}_{ker} = \frac{1}{C_d}\sum_{i=1}^{H}\left(\hat{t}_i - \sum_{j=1}^{H} d_j(i)\hat{t}_j\right)^2$$

- Extension naturelle des estimateurs de variance basés sur des "modèles linéaires" décrits par Hartley et al. (1969), et Isaki (1983)

Plan du reste de l'exposé

- Etude des propriétés analytiques de $\widehat{V}_{ker}$ dans un contexte asymptotique
  - hypothèses sur le plan de sondage, et plans pour lesquels elles sont vérifiées
  - théorême central limite basé sur le plan de sondage
  - modèle non-paramétrique et hypothèses de lissage
  - approximations du biais asymptotique
  - intervalles de confiance
- Comparaison des propriétés empiriques de $\widehat{V}_{coll}$ et $\widehat{V}_{ker}$ par simulations

- Stratification fine : H grand
- Suite de populations finies : quand $H \to \infty$, $N \to \infty$
- Strates de tailles similaires : $N_i \simeq N/H$
- Petites tailles d'échantillon dans les strates : n_i par strate, avec $n_i \simeq p$, fixé
- Taux de sondage :

$$\frac{n_i}{N_i} \simeq \frac{p}{N/H} = O\left(HN^{-1}\right)$$

Hypothèses sur le plan : quand $H \to \infty$ et $N \to \infty$

- Strates de tailles similaires : $\exists\, \delta_U \geq \delta_L > 0$ tels que

$$\delta_L \frac{N}{H} \leq \min_{1 \leq i \leq H} N_i \leq \max_{1 \leq i \leq H} N_i \leq \delta_U \frac{N}{H}.$$

- Variances intra-strates non nulles :

$$\exists\, \nu_L > 0 \text{ tel que } \min_{1 \leq i \leq H} V_i \geq \nu_L N^2 H^{-2}$$

- Plan de sondage probabiliste :

$$\pi_{*H} = \min_{i \in U} \pi_j > 0 \text{ et } \pi_{*H}^{-1} = O(NH^{-1})$$

- Probabilités d'inclusion : pour $m = 1, 2, 3, 4$

$$\max_{(j_1, \ldots, j_m) \in \Delta_m} \pi_{j_1 \cdots j_m} = O\left(H^m N^{-m}\right)$$

où $\Delta_m = m$-uples distincts $(j_1, \ldots, j_m) \in U^m$

- Dans les strates, $\pi_{j_1 \cdots j_m} \equiv 0$ pour $m > n_i$
- Dans les strates, si

$$\max_{(j_1,\ldots,j_m)\in\Delta_m} \pi_{j_1 \cdots j_m} = O\left(H^m N^{-m}\right)$$

  alors c'est vrai également entre les strates par indépendance
- Pour un plan avec remise ou pour un plan de Poisson, si

$$\max_{j_1\in\Delta_1} \pi_{j_1} = O\left(H N^{-1}\right)$$

  est vrai $(m = 1)$, alors c'est vrai également pour $m \geq 2$ par indépendance

- Sans remise avec p éléments par strate et $p \geq 1$ un entier petit, fixé
- La probabilité d'inclusion d'ordre m sur $\Delta_m \cap U_i^m$ est alors

$$\frac{\prod_{j=1}^{m}(p-j+1)}{\prod_{j=1}^{m}(N_i-j+1)} 1_{\{m \leq p\}} = O(H^m N^{-m})$$

si bien que l'hypothèse est vérifiée pour tout m

- Considérons $z_j > 0$ une mesure de taille connue pour chaque élément
- Soit $\bar{z}_{U_i} = N_i^{-1} \sum_{j \in U_i} z_j$ avec $\max_i \bar{z}_{U_i}^{-1} = O(1)$
- Pour $j \in U_i$,

$$\pi_j = \frac{pz_j}{(HN^{-1}N_i)\bar{z}_{U_i}} HN^{-1} = O\left(HN^{-1}\right)$$

  si bien que l'hypothèse est vérifiée pour $m = 1$ $(\Delta_1 = U)$
- L'hypothèse tient pour $m = 2, 3, 4$ pour des plans avec une unité tirée par strate

Méthode de Brewer pour deux éléments par strate

- Brewer (1963):
  - probabilités prop. à la taille, tient pour $m = 1$
  - deux éléments tirés par strate, tient pour $m = 3, 4$
- Pour $m = 2$,

$$\begin{aligned}\pi_{jk} &= H^2 N^{-2} \frac{2 z_j z_k}{H^2 N^{-2} N_i^2 \bar{z}_{U_i} C_{zi}} \frac{\bar{z}_{U_i} - (z_j + z_k) N_i^{-1}}{\left(\bar{z}_{U_i} - 2 z_j N_i^{-1}\right)\left(\bar{z}_{U_i} - 2 z_k N_i^{-1}\right)} \\ &= O\left(H^2 N^{-2}\right)\end{aligned}$$

à la condition que

$$C_{zi}^{-1} = \left(\frac{1}{N_i} \sum_{j \in U_i} \frac{z_j (\bar{z}_{U_i} - z_j N_i^{-1})}{\bar{z}_{U_i} (\bar{z}_{U_i} - 2 z_j N_i^{-1})} \right)^{-1} = O(1)$$

Quelques conséquences des hypothèses

- Faible dépendance du plan :

$$\begin{aligned}
\max_{(j,k)\in\Delta_2} \left|\mathsf{E}\left[(I_j-\pi_j)^3(I_k-\pi_k)\right]\right| &= O\left(H^2N^{-2}\right)\\
\max_{(j,k)\in\Delta_2} \left|\mathsf{E}\left[(I_j-\pi_j)^2(I_k-\pi_k)^2\right]\right| &= O\left(H^2N^{-2}\right)\\
\max_{(j,k,\ell)\in\Delta_3} \left|\mathsf{E}\left[(I_j-\pi_j)^2(I_k-\pi_k)(I_\ell-\pi_\ell)\right]\right| &= O\left(H^3N^{-3}\right)\\
\max_{(j,k,\ell,m)\in\Delta_4} \left|\mathsf{E}\left[(I_j-\pi_j)(I_k-\pi_k)(I_\ell-\pi_\ell)(I_m-\pi_m)\right]\right| &= O\left(H^4N^{-4}\right)
\end{aligned}$$

- Bon comportement des moments d'ordre 4 et des variances sous le plan :

$$\max_{1\le i\le H}\mathsf{E}\left[\left(\hat{t}_i-t_i\right)^4\right] = O(N^4H^{-4}) \text{ et } \max_{1\le i,i'\le H}\mathsf{E}\left[\left(\hat{t}_i-\hat{t}_{i'}\right)^4\right] = O(N^4H^{-4})$$

$$\nu_L\frac{N^2}{H^2} \le \min_{1\le i\le H} V_i \le \max_{1\le i\le H} V_i \le \nu_U\frac{N^2}{H^2} \text{ et } \nu_L\frac{N^2}{H} \le \mathsf{Var}\left(\hat{t}\right) \le \nu_U\frac{N^2}{H}$$

en supposant de plus que

$$M_{*H} = \max_{1\le i\le H} HN^{-1}\sum_{j\in U_i} y_j^4 = O(1)$$

Théorème central limite de Lyapunov

- Pour une suite de populations et de plans stratifiés satisfaisant

$$M_{*H} = \max_{1 \le i \le H} H N^{-1} \sum_{j \in U_i} y_j^4 = O(1)$$

et les hypothèses précédentes sur le plan de sondage,

$$\frac{\sum_{i=1}^{H} \left(\hat{t}_i - t_i \right)}{\sqrt{\text{Var} \left(\sum_{i=1}^{H} \hat{t}_i \right)}} \xrightarrow{\mathcal{L}} N(0,1) \text{ quand } H \to \infty$$

– semblable à Krewski et Rao (1981), Bickel et Freedman (1984)

- Les totaux par strate suivent le modéle :

$$t_i = \sum_{j \in U_i} y_j = m(x_i) + \varepsilon_i,$$

où $m(\cdot)$ possède deux dérivées continues

- Les valeurs d'index $\{x_i\}_{i=1}^{H}$ ont un support compact $[a, b]$ et une "densité sous le plan" $f_x(x) > 0$ continument différentiable, et telle que si $\int_a^b |q(x)| f_x(x)\, dx < \infty$, alors quand $H \to \infty$

$$H^{-1} \sum_{i=1}^{H} q(x_i) \to \int_a^b q(x) f_x(x)\, dx$$

Hypothèses de lissage

- La fonction noyau $K(\cdot)$ est symétrique, continue et bornée et possède un support compact
- La largeur de fenêtre h satisfait $h \to 0$ et $Hh^2 \to \infty$
- Sous les hypothèses précédentes,

$$\sum_{k=1}^{H} K\left(\frac{x_i - x_k}{h}\right)^r = H\,h\,f_x(x_i)\int K(z)^r dz\,(1 + O(h^2 + 1/H))$$

  par des arguments d'approximation standard pour des noyaux
- On applique ceci à

$$\mathsf{E}\left[\hat{V}_{ker}\right] = C_d^{-1}\left\{\sum_{i=1}^{H} V_i\left(1 - 2d_i(i) + \sum_{j=1}^{H} d_j^2(i)\right) + \sum_{i=1}^{H}\left(\sum_{j=1}^{H} d_j(i)(t_i - t_j)\right)^2\right\}$$

Le biais n'est pas négligeable!

- Notons que $\mathsf{Var}\left(\hat{t}\right) = O\left(N^2 H^{-1}\right)$:

$$\mathsf{E}\left[\widehat{V}_{coll}\right] = \mathsf{Var}\left(\hat{t}\right) + \frac{1}{2}\sum_{i=1}^{H}\left(\sum_{j=1}^{H} c_j(i)\left(t_i - t_j\right)\right)^2$$

$$\mathsf{E}\left[\widehat{V}_{ker}\right] = \mathsf{Var}\left(\hat{t}\right) + \frac{1}{C_d}\sum_{i=1}^{H}\left(\sum_{j=1}^{H} d_j(i)(\varepsilon_i - \varepsilon_j)\right)^2 + O\left(\frac{N^2}{H}\left(h^2 + \frac{1}{Hh}\right)\right)$$

- $\widehat{V}_{ker}$ peut être beaucoup moins biaisé si $m(x_i)$ n'est pas constant et explique une grande partie de la variation des t_i

Convergence en moyenne quadratique de $\widehat{V}_{ker}$

- Rappelons que $\mathsf{Var}\left(\hat{t}\right) = O\left(N^2 H^{-1}\right)$
- La variance sous le plan de $\widehat{V}_{ker}$ est $\mathsf{Var}\left(\widehat{V}_{ker}\right) = o(N^4 H^{-2})$, de sorte que

$$\frac{\widehat{V}_{ker} - \mathsf{E}\left[\widehat{V}_{ker}\right]}{\mathsf{Var}\left(\hat{t}\right)} \to 0$$

en moyenne quadratique quand $H \to \infty$

Intervalles de confiance utilisant $\widehat{V}_{ker}$

- Pour une suite de populations et de plans de sondage stratifiés satisfaisant les conditions ci-dessus,

$$\widehat{V}_{ker}^{-1/2} \sum_{i=1}^{H} \left(\hat{t}_i - t_i\right) \xrightarrow{\mathcal{L}} N\left(0, \tau^{-2}\right) \text{ quand } H \to \infty,$$

$$\text{où } \tau^2 = 1 + \lim_{H\to\infty} \frac{\sum_{i=1}^{H} \left(\sum_{j=1}^{H} d_j(i)(\varepsilon_i - \varepsilon_j)\right)^2}{\sum_{i=1}^{H} V_i} \geq 1.$$

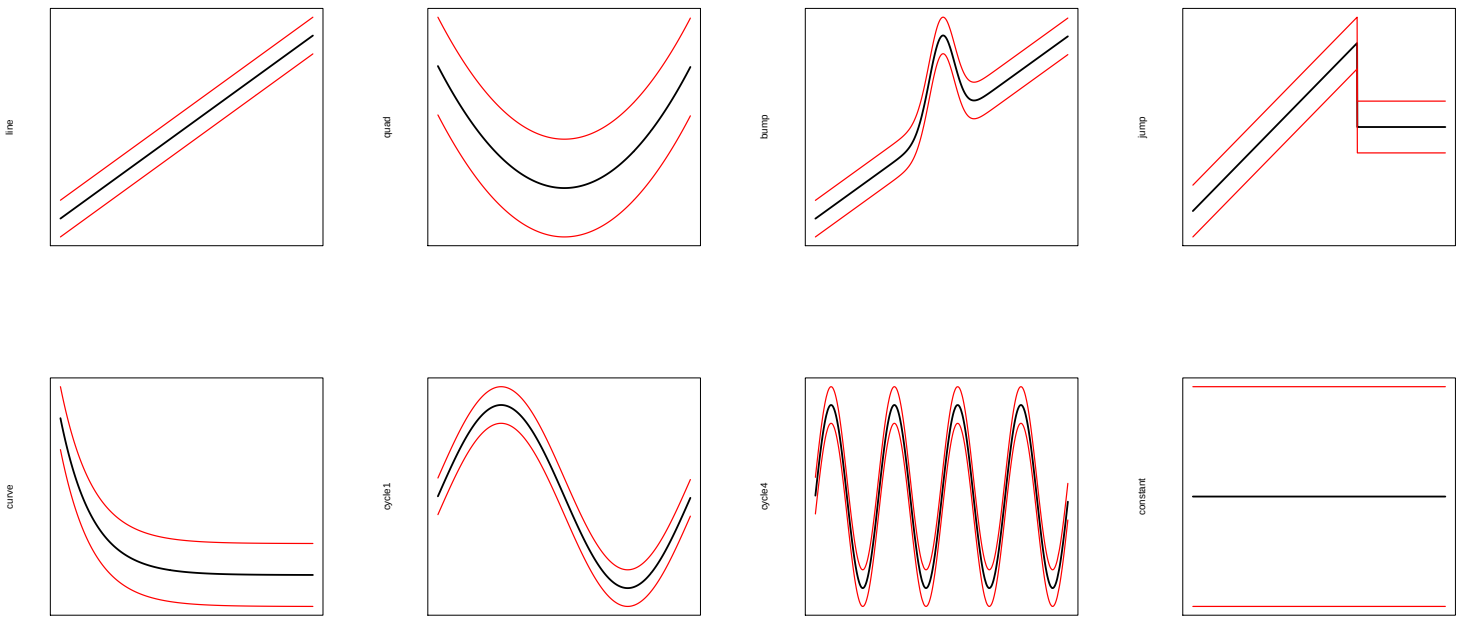
- L'intervalle de confiance pour t donné par

$$\left(\hat{t} + \Phi^{-1}(\alpha/2)\widehat{V}_{ker}^{1/2}, \hat{t} + \Phi^{-1}(1-\alpha/2)\widehat{V}_{ker}^{1/2}\right)$$

possède un taux de couverture asymptotique d'au moins $(1-\alpha)100\%$.

Étude par simulations

- $y_j = m_*(x_i) + \sigma e_j$ pour $j \in U_i$; $\{e_j\}$ iid $\mathsf{N}(0, \sigma^2)$
  $t_i = \frac{N}{H} m_*(x_i) + \sigma \sum_{j \in U_i} e_j = m(x_i) + \epsilon_i$



Biais exact de $\widehat{V}_{coll}$ et $\widehat{V}_{ker}$ pour $\sigma = 0$, $N_i \equiv 60$

- Les variables de réponse sont constantes dans les strates, si bien que $V_i \equiv 0$ et $\mathrm{Var}\left(\hat{t}\right) = 0$ *pour n'importe quel plan de sondage*

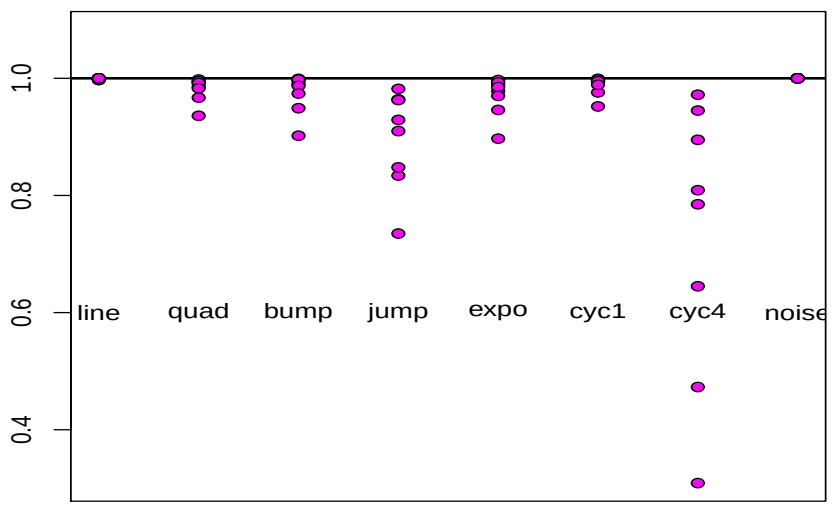
$$\mathrm{E}\left[\widehat{V}_{coll}\right] = \mathrm{Var}\left(\hat{t}\right) + \frac{1}{2}\sum_{i=1}^{H}\left(\sum_{j=1}^{H} c_j(i)(t_i - t_j)\right)^2$$

$$\mathrm{E}\left[\widehat{V}_{ker}\right] = C_d^{-1}\left\{\sum_{i=1}^{H} V_i\left(1 - 2d_i(i) + \sum_{j=1}^{H} d_j^2(i)\right) + \sum_{i=1}^{H}\left(\sum_{j=1}^{H} d_j(i)(t_i - t_j)\right)^2\right\}$$

Scenario		line	quad	bump	jump	expo	cycle1	cycle4
$H = 50$	$\widehat{V}_{coll}$	144.00	767.69	801.22	239.75	666.49	712.49	11176.46
$h = 0.025$	$\widehat{V}_{ker}$	3.13	47.37	53.96	1349.38	89.86	34.18	1287.39
$H = 100$	$\widehat{V}_{coll}$	72.00	383.96	397.28	3825.88	311.05	355.19	5677.42
$h = 0.015$	$\widehat{V}_{ker}$	0.89	13.84	7.36	1287.48	26.92	9.22	249.57
$H = 200$	$\widehat{V}_{coll}$	36.00	192.00	196.30	56.25	149.84	177.64	2849.94
$h = 0.0055$	$\widehat{V}_{ker}$	0.16	2.51	0.96	1244.74	4.97	1.63	39.75

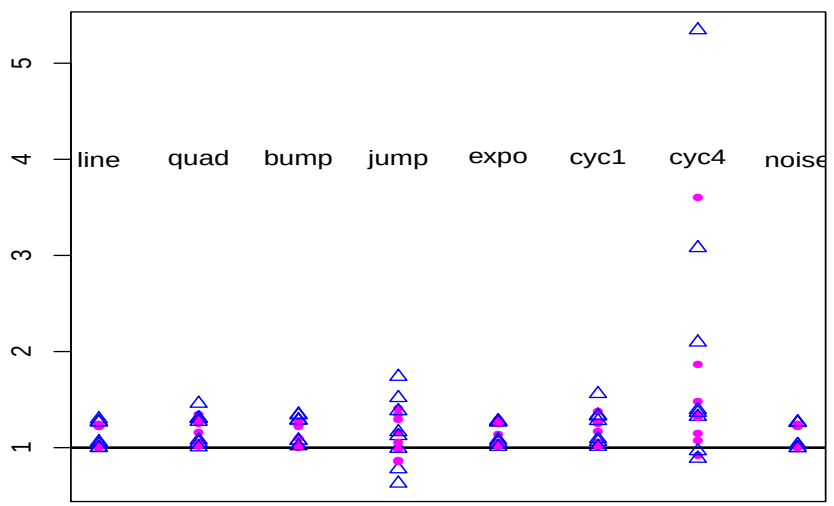
Ratio de l'espérance asymptotique et de l'espérance exacte de $\widehat{V}_{ker}$

- $H = 50$; deux largeurs de fenêtre; deux niveaux de bruit; $n_i \equiv 1$ ou $n_i \equiv 2$
- E $\left[\widehat{V}_{ker}\right] \simeq \frac{1}{C_d}\sum_{i=1}^{H} V_i + \frac{1}{C_d}\sum_{i=1}^{H}\left(\sum_{j=1}^{H} d_j(i)(\varepsilon_i - \varepsilon_j)\right)^2$ en supprimant les termes en $O\left(\frac{N^2}{H}\left(h^2 + \frac{1}{Hh}\right)\right)$



Racine carrée du ratio des MSE de $\widehat{V}_{coll}$ et $\widehat{V}_{ker}$

- $H = 50, 100$; 2 largeurs h; 2 niveaux de bruit; $n_i \equiv 1$ ou 2 (bleu)
- $\left\{\text{MSE}\left(\widehat{V}_{coll}\right)/\text{MSE}\left(\widehat{V}_{ker}\right)\right\}^{1/2}$: ≥ 1 en faveur du noyau



- Dans presque tous les cas, les intervalles de couverture nominale 95% avaient une couverture entre 94% et 96%
  - quelques exceptions pour `jump` et `cycle4`
  - pour les exceptions, sur-couverture (comme prévu)
- Longueur moyenne des intervalles remarquablement consistante dans les différents cas
- La variabilité des longueurs des intervalles favorise $\widehat{V}_{ker}$
  - $\widehat{V}_{coll}$ donne la même longueur moyenne et un taux de couverture correct
  - mais les intervalles sont parfois beaucoup trop restreints et parfois beaucoup trop larges selon les échantillons

Résumé

- "Regroupement par noyaux" : alternative viable à l'estimation de variance classique par regroupement des strates
- Cadre asymptotique avec des hypothèses vérifiées pour une gamme raisonnable de plans de sondage
- Le biais de $\widehat{V}_{ker}$ est du même ordre asymptotiquement que celui de $\widehat{V}_{coll}$
  - mais le biais de $\widehat{V}_{ker}$ peut être beaucoup plus faible s'il existe une relation fonctionnelle lisse entre les totaux par strate et l'"index de regroupement"
- Les résultats des simulations appuient la théorie asymptotique

Quelques références

- Bickel, P. J. and D. A. Freedman (1984). Asymptotic normality and the bootstrap in stratified sampling. *Annals of Statistics 12*, 470–482.
- Breidt, F. J. and J. D. Opsomer (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics 28*, 1026–1053.
- Brewer, K. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics 5(1)*, 5–13.
- Hartley, H., J. Rao, and G. Kiefer (1969). Variance estimation with one unit per stratum. *Journal of the American Statistical Association*, 841–851.
- Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*, 663–685.
- Isaki, C. (1983). Variance estimation using auxiliary information. *Journal of the American Statistical Association*, 117–123.
- Krewski, D. and J. N. K. Rao (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics 9*, 1010–1019.
- Rust, K. and G. Kalton (1987). Strategies for collapsing strata for variance estimation. *Journal of Official Statistics 3(1)*, 69–81.