# Sondages stratifiés pour données fonctionnelles : Allocation optimale et bandes de confiance asymptotiques

Hervé Cardot, Etienne Josserand

# Sondages stratifiés pour données fonctionnelles : Allocation optimale et bandes de confiance asymptotiques

Hervé CARDOT[(1)], Etienne JOSSERAND[(1)]

*(1) Institut de Mathématiques de Bourgogne, Université de Bourgogne, 9 Avenue Alain Savary, BP 47870, 21078 DIJON Cedex, FRANCE.*

**Abstract**

This work aims at performing an estimate of the average based on estimates of Horvitz-Thompson type when data and curves are selected from a strata survey, and also a asymptotic band of confidence. The technique of stratification reduces the estimators variance.

**Résumé**

Ce travail propose une estimation de la moyenne basée sur des estimateurs de type Horvitz-Thompson lorsque les données sont des courbes et issues d'un plan de sondage en strates, ainsi qu'une bande de confiance asymptotique. La technique de stratification permet de réduire la variance des estimateurs.

**Mots clés majeurs** : sondages, analyse des données - data mining, stratification.
**Mots clés** : données fonctionnelles, stratification, Horvitz-Thompson, bandes de confiance, variance.

# 1 Introduction

L'analyse statistique de courbes, ou analyse des données fonctionnelles, est un thème de recherche en statistique en pleine expansion dont les applications concernent de nombreux domaines scientifiques (climatologie, médecine, économie, chimie quantitative, ...). On pourra se reporter à Ramsay & Silverman (2005) pour une revue de différentes méthodes d'analyse illustrées sur des exemples variés. Les outils statistiques mis en oeuvre sont issus de l'analyse fonctionnelle et généralisent les procédures classiques de statistique multivariée.

La manière dont les données sont obtenues est rarement prise en compte dans ces analyses qui supposent (implicitement) que les observations sont indépendantes et identiquement distribuées. Or cette hypothèse n'est pas systématiquement vérifiée, les courbes observées pouvant provenir d'un plan de sondage élaboré par le statisticien. C'est par exemple l'étude de l'évolution de la consommation électrique mesurée à partir d'un échantillon de compteurs tirés selon un plan de sondage complexe (Dessertaine, 2006, Chiky et Hebrail, 2009).

Notre objectif est de proposer des estimateurs de la courbe moyenne dans le cadre des sondages stratifiés. Nous nous efforçons de justifier nos choix de normes en les reliant à des techniques classiques (Cochran 1977) et à la généralisation des intervalles de confiance aux bandes de confiance.

# 2 Données fonctionnelles et sondage

On considère une population $U = \{1, \ldots, k, \ldots, N\}$ de taille finie $N$ et on s'intéresse à une variable fonctionnelle $\mathcal{Y}$ définie pour chaque individu $k$ de la population $U$, où $Y_k = (Y_k(t))_{t\in[0,1]}$ appartient à l'espace des fonctions de carré intégrable $L^2[0,1]$ muni de son produit scalaire usuel noté $\langle ., . \rangle$ et de la norme induite $\|.\|$.

On note $\mu \in L^2[0,1]$, la moyenne des $Y_k$ de la population

$$\mu(t) = \frac{1}{N} \sum_{k\in U} Y_k(t), \quad t \in [0,1] \tag{1}$$

et $\Gamma$, l'opérateur de covariance défini sur $L^2[0,1]$ par

$$\Gamma = \frac{1}{N} \sum_{k\in U} (Y_k - \mu) \otimes (Y_k - \mu) \tag{2}$$

où le produit tensoriel de deux éléments $a$ et $b$ de $L^2[0,1]$ est l'opérateur de rang un tel que $a \otimes b(u) = \langle a, u\rangle b$ pour tout $u$ dans $L^2[0,1]$.

Un échantillon $s$, *i.e.* une partie $s \subset U$, est tiré selon un procédé probabiliste $p(s)$ où $p$ est une loi de probabilité sur l'ensemble de parties possibles de $U$. On note $\pi_k = \mathbb{P}(k \in s)$ pour tous les $k \in U$ et $\pi_{kl} = \mathbb{P}(k \ \& \ l \in s)$ pour tous $k, l \in U$, $k \neq l$ les probabilités d'inclusion du premier et deuxième degré. On suppose par ailleurs que $\pi_k > 0$ et $\pi_{kl} > 0$ : tous les individus et les couples d'individus ont une probabilité non-nulle d'être présents dans l'échantillon. La taille $N$ de la population n'est pas toujours connue *a priori.*

Les estimateurs classiques de type Horvitz-Thompson de $\mu$ et $\Gamma$ sont définis par

$$\widehat{\mu} = \frac{1}{\widehat{N}} \sum_{k\in s} \frac{Y_k}{\pi_k} \tag{3}$$

$$\widehat{\Gamma} = \frac{1}{\widehat{N}} \sum_{k\in s} \frac{Y_k \otimes Y_k}{\pi_k} - \widehat{\mu} \otimes \widehat{\mu} \tag{4}$$

où l'estimateur de la taille $N$ de la population est $\widehat{N} = \sum_{k\in s} \frac{1}{\pi_k}$. (Cardot et al. 2008)

# 3 Estimation de la moyenne et allocation optimale

## 3.1 Stratification

Supposons que $N$ soit connu et que la population $U$ soit partitionnée en $H$ sous-ensembles, $U_h$, $h = 1, \ldots, H$, appelés strates tels que

$$\bigcup_{h=1}^{H} U_h = U \text{ et } U_i \bigcap U_j = \emptyset \text{ pour } i \neq j \tag{5}$$

Le nombre d'éléments de la strate $N_h$ est appelé taille de la strate. On a $\sum_{h=1}^{H} N_h = N$. Les $N_h$ sont supposés connus et constituent l'information auxiliaire disponible sur la population entière.

L'objectif est d'estimer la fonction d'intérêt $\mu$ qui est la moyenne des fonctions prises par le caractère d'intérêt de la population. La moyenne sur $U_h$ est donnée par

$$\mu_h = \frac{1}{N_h} \sum_{k \in U_h} Y_k \tag{6}$$

De plus, on note $\Gamma_h$ l'opérateur de covariance de la state $h$ et $\widetilde{\Gamma}_h$ l'opérateur corrigé

$$\Gamma_h = \frac{1}{N_h} \sum_{k \in U_h} (Y_k - \mu_h) \otimes (Y_k - \mu_h) \text{ et } \widetilde{\Gamma}_h = \frac{N_h}{N_h - 1} \Gamma_h \tag{7}$$

On obtient une formule analogue du cadre de sondage stratifié de variable réelle pour la variance de l'estimateur, c'est à dire :

$$\mathrm{Var}(\widehat{\mu}_{\mathrm{strat}}) = \frac{1}{N^2} \sum_{h=1}^{H} N_h \frac{N_h - n_h}{n_h} \widetilde{\Gamma}_h \tag{8}$$

Les $n_h$ sont choisis de sorte qu'ils minimisent le problème suivant :

$$\mathrm{Trace}\Big(\mathrm{Var}(\widehat{\mu}_{\mathrm{strat}})\Big) = \mathrm{Trace}\Big(\frac{1}{N^2} \sum_{h=1}^{H} N_h \frac{N_h - n_h}{n_h} \widetilde{\Gamma}_h\Big) \tag{9}$$

en $n_1,\ldots,n_h,\ldots,n_H$ sous la contrainte $\sum_{h=1}^{H} n_h = n$, où la norme trace est défini par $\mathrm{Trace}(T) = \sum_j |\langle T e_j, e_j \rangle|$ pour toute base orthonormale $\{e_j\}_{j \geq 1}$ de $L^2[0,1]$.
La minimisation de la trace de l'opérateur de covariance nous permet d'obtenir des solutions explicites et de retrouver des résultats similaires au cas multivarié réel (Cochran 1977).

$$n_h = n \frac{N_h \sqrt{\sum_{k=1}^{K} \mathrm{Trace}(\widetilde{\Gamma}_k)}}{\sum_{h=1}^{H} N_h \sqrt{\sum_{k=1}^{K} \mathrm{Trace}(\widetilde{\Gamma}_k)}} \text{ pour } h = 1, \ldots, H \tag{10}$$

## 3.2 Bande de confiance

Nous souhaiterions obtenir l'équivalent des intervalles de confiance du cadre réel pour l'estimation de la moyenne. Un plan de sondage est normal si l'estimateur d'Horvitz-Thompson vérifie le théorème central limite. C'est le cas par exemple des plans simples et stratifiés. Il est alors possible de montrer que si le plan de sondage est normal, $\sqrt{n} \dfrac{\widehat{\mu}(t) - \mu(t)}{\sqrt{\widehat{\gamma}(t,t)}}$ converge en loi vers $Z = \mathscr{G}(0, \rho)$ où $\gamma(s,t)$ est la covariance empirique et $\rho(s,t) = \gamma(s,t)/\sqrt{\gamma(s,s)\gamma(t,t)}$. Grâce à des résultats de Van der Vaart (1998) on prouve à l'aide d'un critère de tension que la convergence a lieu dans $C([0,1])$ sous certaines contraintes de régularité des trajectoires. De manière similaire à Degras (2009), il est possible d'appliquer un résultat classique de Landau & Shepp (1970) :

$$\lim_{\lambda \to \infty} \lambda^{-2} \log \mathbb{P}\Big\{\sup_{t \in [0;1]} Z(t) > \lambda\Big\} = -\Big(2 \sup_{t \in [0;1]} \rho(t,t)\Big)^{-1} = -2^{-1} \tag{11}$$

Fixons nous un niveau de risque $\alpha$. De la précedente égalité, et du fait que

$$\mathbb{P}\{\sup_{t\in[0;1]}|Z(t)| > \lambda\} \leq 2\mathbb{P}\{\sup_{t\in[0;1]}Z(t) > \lambda\} \tag{12}$$

on obtient une bande de confiance pour $\mu$ :

$$\widehat{\mu}(t) \pm \left(2\log(2/\alpha)\widehat{\gamma}(t,t)/n\right)^{1/2} \tag{13}$$

Ce résultat confirme l'utilisation de la norme trace de la section précédente grâce à la relation $\text{Trace}(\Gamma) = \int \gamma(t,t)\mathrm{d}t$. La stratification permet en effet de minimiser cette quantité.

# 4 Simulations

Nous avons simulé une population de 10000 trajectoires de mouvements browniens sur l'intervalle [0,1], discrétisés en 100 points équidistants. Nous avons ensuite constitué deux fois 500 plans de sondages (aléatoire simple et stratifié), chacun de taille n=100, 500 et 1000. Nous avons ensuite évalué la qualité de l'estimateur de la moyenne en mesurant sa trace et sa norme de Hilbert-Schmidt.

Les simulations confirment la théorie ; à savoir une meilleure estimation de l'opérateur stratifié dans le cas où les variances des strates sont hétérogènes.

## Bibliographie


[1] Cardot, H., Chaouch, M., Goga, C. and C. Labruere (2008). *Functional Principal Components Analysis with Survey Data.* In Functional and Operatorial Statistics, Dabo-Niang, S. and Ferraty, F. (Eds.), Physica-Verlag, Heidelberg, 95-102.

[2] Chiky, R. and Hébrail, G. (2009). *Spatio-temporal sampling of distributed data streams.* J. of Computing Science and Engineering, to appear.

[3] Cochran W.G. (1977). *Sampling techniques.* Wiley series in probability and mathematical statistics-apllied, 3nd ed.

[4] Degras D. (2009). *Nonparametric inference of a trend using functionnal data* Compte Rendu à l'Académie des sciences, série I, à paraître.

[5] Dessertaine A. (2006). Sondage et séries temporelles: une application pour la prévision de la consommation électrique. *38èmes Journées de Statistique*, Clamart, Juin 2006.

[6] Landau H. and Shepp L.A. (1970), *On the supremum of a Gaussian process.* Sankhyã 32 369-378

[7] Ramsay J. O. and Silverman B.W. (2005). *Functional Data Analysis.* Springer-Verlag, 2nd ed.

[8] Van der Vaart A.W. (1998). *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics.