# Théorie des sondages : cours 4
# Techniques de linéarisation et de ré-échantillonnage

Camelia GOGA

IMB, Université de Bourgogne
e-mail : camelia.goga@u-bourgogne.fr

Master Besançon-2010

# Chapitre 4 : Techniques de linéarisation et de ré-échantillonnage

**Objectif** : estimer un paramètre qui est une fonction non-linéaire de totaux, $\Phi = \Phi(t_Y, t_X, \ldots)$, et ensuite calculer ou approcher sa variance. On considère dans la suite que $\Phi = \Phi(t_Y, t_X)$.

**Exemple** :
- la moyenne $\overline{y}_U = \sum_U y_k / N$ quand $N$ est inconnu ;
- le ratio (situation très fréquente) $R = \frac{\sum_U y_k}{\sum_U x_k}$;
- un coefficient d'une régression $B = \frac{\sum_U x_k y_k}{\sum_U x_k^2}$ ...

**Question : Comment construire un estimateur ?**

**La réponse est simple** : on écrit notre paramètre comme une fonction de totaux et ensuite, chaque total est remplacé par son estimateur HT, $\widehat{\Phi} = \Phi(\hat{t}_{y\pi}, \hat{t}_{x\pi})$ appelé estimateur par substitution.

Remarque : nous allons utiliser les même poids $1/\pi_k$ quel que soit le total qu'on veut estimer puisque les $\pi_k$ ne dépendent pas de la variable d'intérêt.

- $\overline{y}_U = \frac{\sum_U y_k}{\sum_U 1}$ est estimé par $\widehat{\overline{y}}_U = \frac{\sum_s y_k/\pi_k}{\sum_s 1/\pi_k}$ (l'estimateur de Hajek),
- $\hat{R} = \dfrac{\sum_s y_k/\pi_k}{\sum_s x_k/\pi_k}$
- $\hat{B} = \dfrac{\sum_s x_k y_k/\pi_k}{\sum_s x_k^2/\pi_k}$ ...

Les estimateurs obtenus ne sont plus sans biais **mais pour $n$ grand, le biais est négligeable**.

**Difficulté** : il n'existe pas de formule générale pour la variance comme dans le cas linéaire alors,

nous allons essayer de se ramener au cas linéaire par des techniques de linéarisation : un développement de Taylor de $\Phi$ dans $(t_y, t_x)$.

## Résultat

*Supposons que $\Phi$ est différentiable, alors*

$$\begin{aligned}
\Phi(\hat{t}_{y\pi}, \hat{t}_{x\pi}) - \Phi(t_y, t_x) &\simeq (\hat{t}_{y\pi} - t_y)\alpha_1 + (\hat{t}_{x\pi} - t_x)\alpha_2 \\
\alpha_1 &= \left.\frac{\partial \Phi(v_1, v_2)}{\partial v_1}\right|_{(v_1,v_2)=(t_y,t_x)} \\
\alpha_2 &= \left.\frac{\partial \Phi(v_1, v_2)}{\partial v_2}\right|_{(v_1,v_2)=(t_y,t_x)}
\end{aligned}$$

On peut écrire sous une forme équivalente :

$$\begin{aligned}
\hat{\Phi} - \Phi &\simeq \sum_s \frac{u_k}{\pi_k} - \sum_U u_k \\
u_k &= \alpha_1 y_k + \alpha_2 x_k, \quad k \in U, \quad \text{la variable linéarisée de } \Phi
\end{aligned}$$

c'est à dire,

notre estimateur est approché par l'estimateur de HT pour le total de $u_k$.

Il résulte (sous des hypothèses supplémentaires) que pour $n$ grand,

1. le biais de $\hat{\Phi}$ est négligeable, $B(\hat{\Phi}) \simeq 0$ ;
2. $V(\hat{\Phi}) \simeq EQR(\hat{\Phi}) \simeq V(\sum_s \frac{u_k}{\pi_k}) = \sum_U \sum_U \Delta_{kl} \frac{u_k}{\pi_k} \frac{u_l}{\pi_l}$

**Problème** : les $u_k$ sont inconnus puisque $\alpha_1$ et $\alpha_2$ sont inconnus, par conséquent on les estime par

$$\begin{aligned} \hat{u}_k &= \hat{\alpha}_1 y_k + \hat{\alpha}_2 x_k \\ &= y_k \left.\frac{\partial f(v_1, v_2)}{\partial v_1}\right|_{(v_1,v_2)=(\hat{t}_{y\pi},\hat{t}_{x\pi})} + x_k \left.\frac{\partial f(v_1, v_2)}{\partial v_2}\right|_{(v_1,v_2)=(\hat{t}_{y\pi},\hat{t}_{x\pi})} \end{aligned}$$

3. L'estimateur de la variance est $\widehat{V}(\hat{\Phi}) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{\hat{u}_k}{\pi_k} \frac{\hat{u}_l}{\pi_l}$.

# Intervalles de confiance asymptotique normaux

- On suppose que $n$ est assez grand pour avoir l'hypothèse de normalité. On a

$$\hat{\Phi} - \Phi \simeq \sum_s \frac{u_k}{\pi_k} - \sum_U u_k$$

- On suppose aussi que l'estimateur de la variance $\widehat{V}(\hat{\Phi})$ est convergent pour $V(\hat{\Phi})$.
- Alors, l'intervalle de confiance asymptotique de $\Phi$ à 95% est

$$\mathrm{IC}_{95\%}(\Phi) = [\hat{\Phi} - z_{\alpha/2}\sqrt{\widehat{V}(\hat{\Phi})}, \hat{\Phi} + z_{\alpha/2}\sqrt{\widehat{V}(\hat{\Phi})}]$$

# Techniques de ré-échantillonnage

Plusieurs : bootstrap, jacknife, ...
On présente dans la suite la méthode du **bootstrap**.

Cette technique a été introduite dans la statistique classique (Efron, 1979) d'où la difficulté lors de son application dans des populations finies sans la condition de *iid* des observations (de date assez récente, Gross 1980).

# Le bootstrap et son application au calcul des intervalles de confiance

Soit $\theta$ un paramètre inconnu et $\hat{\theta}_n$ un estimateur.

- Le bootstrap est une méthode de réechantillonnage qui permet de calculer le biais, la variance de l'estimateur $\hat{\theta}$ sans faire aucune hypothèse sur la loi de $\hat{\theta}$;
- Cette méthode permet également de déterminer un intervalle de confiance pour $\theta$ sans calculer forcement la variance de $\hat{\theta}$;
- Utile quad la taille de l'échantillon n'est pas assez grande pour pouvoir utiliser le théorème central limite.

Elle donne directement une estimation de la variance.

**La méthode de bootstrap pour un plan aléatoire simple sans remise** :

- soit $\widehat{\Phi}$ l'estimateur de $\Phi$ et $\pi_k = \frac{n}{N}$ les probas d'inclusion ;
- à l'aide de l'échantillon $s$, une population "artificielle" $U^*$ est créée en dupliquant $N/n$ (supposé entier) fois chaque unité $k \in s$ ;

- dans cette nouvelle population $U^*$, nous allons tirer $A$ échantillons indépendants et selon le même plan de sondage de $s$ dans $U$.

$$\begin{array}{ccccc} \text{l'échantillon} & s_1^* & \text{donne une estimation} & \widehat{\Phi}_1^* \\ & & \dots & \\ \text{l'échantillon} & s_A^* & \text{donne une estimation} & \widehat{\Phi}_A^* \end{array}$$

- la distribution de $\widehat{\Phi}_1^*, \dots, \widehat{\Phi}_A^*$ est considérée comme une estimation de la distribution de $\widehat{\Phi}$ et $V(\widehat{\Phi})$ est estimée par

$$\hat{V}_{BS} = \frac{1}{A-1} \sum_{a=1}^{A} (\widehat{\Phi}_a^* - \overline{\widehat{\Phi}^*})^2$$

$$\overline{\widehat{\Phi}^*} = \frac{1}{A} \sum_{a=1}^{A} \widehat{\Phi}_a^*.$$

## Intervalles de confiance par le bootstrap : la méthode de *percentiles*

- Soient $\theta_L$, resp. $\theta_U$, le quantile d'ordre $(\alpha/2)\%$ et resp. $(1-\alpha/2)\%$ de la série de réplications $\tilde{\theta}_n^{(1)}, \dots, \tilde{\theta}_n^{(B)}$.

- Un intervalle de confiance de $\theta$ avec une confiance $(1-\alpha)\%$ est obtenu en prenant l'intervalle

$$[\theta_L, \theta_U]$$