

# T. D. n° 7

## Estimation par différence et par régression

**Exercice 1.** *Stratification et estimateur par la différence.* **Extrait du livre « Exercices corrigés de méthodes de sondage » de P. Ardilly et de Y. Tillé. D'après l'examen de Janvier 2006, M2-Statistique.**

Soit un plan stratifié composé de  $H$  strates de taille  $N_h$ . L'objectif est d'estimer la moyenne de la population  $\bar{Y}$  d'un caractère  $y$ . Notons  $\bar{X}_h$ ,  $h = 1, \dots, H$  les moyennes dans les strates (dans la population) d'un caractère auxiliaire  $x$ . Les  $\bar{X}_h$  sont supposées connues et on se propose d'estimer  $\bar{Y}$  au moyen de l'estimateur suivant :

$$\widehat{Y}_D = \widehat{Y}_\pi + \bar{X} - \widehat{X}_\pi.$$

Nous réalisons un sondage aléatoire simple dans chaque strate.

1. Montrer que  $\widehat{Y}_D$  estime sans biais la moyenne de la population  $\bar{Y}$  d'un caractère  $y$ .
2. Donner la variance de l'estimateur  $\widehat{Y}_D$ .
3. Quelle est l'allocation optimale des tailles  $n_h$  pour minimiser la variance de l'estimateur  $\widehat{Y}_D$ ? Nous considérons que le coût unitaire d'enquête ne dépend pas de la strate.
4. Dans quel cas favorable l'estimateur  $\widehat{Y}_D$  est-il indiscutablement préférable à l'estimateur  $\widehat{Y}_\pi$ ?

**Exercice 2.** *Comparaison de plusieurs estimateurs.* **D'après une idée du livre « Exercices corrigés de méthodes de sondage » de P. Ardilly et de Y. Tillé.**

Nous souhaitons estimer la moyenne  $\mu_Y$  de la variable  $Y$ , qui est le nombre d'habitants en 2005 en Bretagne. Pour cela, nous tirons un échantillon de 35 villes bretonnes provenant d'une population de 180 villes.

1. Quels sont les estimateurs que vous connaissez qui peuvent répondre à cette question? Citer-les et donner l'expression de leur variance.
2. À partir de vos réflexions, construire l'estimateur par la différence, noté  $\widehat{Y}_D$ .
3. Calculer l'espérance et la variance de l'estimateur  $\widehat{Y}_D$ .
4. **Application numérique.**

L'échantillon de 35 villes bretonnes, en 2005, donne comme moyenne 22 072, 63. D'autre part, non seulement nous connaissons la moyenne d'un échantillon de villes bretonnes, en 1996, qui vaut 21 585, 26, mais aussi la moyenne de

la population entière (les 180 villes) qui est égale à 32 039,66. Nous avons également calculé sur les échantillons les quantités suivantes :

$$s_Y^2 = 4\,131\,789\,466, \quad s_X^2 = 3\,989\,656\,072 \quad \text{et} \quad s_{YX} = 4\,059\,448\,772.$$

Estimer la population moyenne en 2005, en utilisant les différents estimateurs que vous avez cités au préalable et en essayant d'utiliser l'information auxiliaire dont vous disposez. Que remarquez-vous ?

Déterminer un intervalle de confiance à 95% pour la population moyenne basé sur les différents estimateurs que vous avez cités. Comparer-les. Que concluez-vous ?

5. Quelle est la conclusion de cet exercice ? Essayer à partir de vos résultats de dégager une stratégie pour choisir « le bon estimateur ».

**Exercice 3. Taille des pieds. Extrait du livre « Exercices corrigés de méthodes de sondage » de P. Ardilly et de Y. Tillé.**

Le directeur d'une entreprise de confection de chaussures veut estimer la longueur moyenne des pieds droits des hommes adultes d'une ville. Soient  $y$  le caractère « longueur du pied droit » (en centimètres) et  $x$  la taille de l'individu (en centimètres). Le directeur sait en outre par les résultats d'un recensement que la taille moyenne des hommes adultes de cette ville est de 168 cm. Pour estimer la longueur des pieds, le directeur effectue un SAS SR de 100 hommes adultes. Les résultats sont les suivants :

$$\widehat{X} = 169, \quad \widehat{Y} = 24, \quad s_{xy} = 15, \quad s_x = 10, \quad s_y = 2.$$

Sachant que 400 000 hommes adultes vivent dans cette ville,

1. Calculer l'estimateur de Horvitz-Thompson, l'estimateur par le quotient, l'estimateur par la différence et l'estimateur par la régression.
2. Estimer les variances de ces 4 estimateurs.
3. Quel estimateur conseilleriez-vous au directeur ?
4. Exprimer la différence littérale entre la variance estimée de l'estimateur par le quotient et la variance estimée de l'estimateur par la régression, en fonction de  $\widehat{X}$ ,  $\widehat{Y}$ , et de la pente de la régression de  $y$  sur  $x$  calculée sur l'échantillon. Commenter.