Chapitre 4
Estimation par la différence et par le quotient

4.1 Introduction

Jusqu'ici, nous avons estimé μ par la moyenne échantillonnale $\bar{y}$ (et τ par $T = N\bar{y}$). Ces estimateurs naturels ne sont pas les seuls possibles et dans certaines situations d'autres estimateurs peuvent se révéler plus précis. Dans ce chapitre, nous étudions deux nouveaux estimateurs, l'***estimateur par la différence*** et l'***estimateur par le quotient.*** Ces deux estimateurs ont ceci de commun qu'ils utilisent une information supplémentaire, une deuxième variable *x*, appelée ***variable auxiliaire***, dont les valeurs sont connues pour la population entière. Dans les applications en vérification, par exemple, *x* est typiquement la valeur aux livres, qui peut être erronée, et *y* est la valeur corrigée. Les valeurs de *y* sont observées dans un échantillon de taille *n*, alors que celles de *x* sont connues pour la population entière. Si *x* et *y* sont des variables dépendantes, alors toute information sur *x* est susceptible d'améliorer l'estimation de $\bar{y}$ ou de τ. C'est ce qui fait que les estimateurs présentés ici peuvent être parfois *beaucoup* plus précis que ceux présentés aux chapitres 2 et 3.

***Exemple* 4.1.1** *Différents estimateurs d'une même moyenne*

Le tableau A.14 présente certaines données sur un échantillon de 35 villes québécoises, tirées d'une population de 180 villes. Supposons que nous voulons estimer la moyenne μ_y de la variable *y* = nombre d'habitants en 2001. Nous allons montrer trois façons d'estimer μ_y.

a) *Estimation par la moyenne* Puisque l'échantillon montre que la moyenne échantillonnale est $\bar{y}$ = 22 072,63, ce nombre est notre estimation de μ_y. C'est l'*estimation par la moyenne*.

b) *Estimation par la différence* L'estimation par la moyenne est la seule raisonnable si on n'a aucune autre information pertinente sur la population. Mais supposons que nous connaissons les valeurs de la variable *x* = nombre d'habitants en 1996. Pas seulement la moyenne échantillonnale, qui est $\bar{x}$ = 21 585,26, mais aussi la moyenne de la *population entière* (les 180 villes): μ_x = 32 039,66. On peut faire le raisonnement suivant: d'après l'échantillon, la population des villes s'est accrue de 487,3714 personnes en moyenne [$\bar{y} - \bar{x}$ = 22 072,63-21 585,26 = 487,3714]. Puisqu'elle était de 32 039,66 en 1996, on l'estime maintenant à 32 039,66 + 487,3714 = 32 527. Ainsi l'échantillon a été utilisé ici pour estimer *l'accroissement* de la population, un accroissement qu'on ajoute à la moyenne connue de 1996. L'estimateur utilisé ici est appelé *estimateur par la différence*.

c) *Estimation par le quotient* Une deuxième approche consiste à estimer non pas l'accroissement absolu mais plutôt le *pourcentage* d'accroissement, et d'appliquer celui-ci à la population connue de 1996. Ainsi, l'accroissement de la population depuis 1996 a été, d'après notre échantillon, de 2,257 89 % [($\bar{y}$ - $\bar{x}$)/ $\bar{x}$ = (22 072,63-21 585,26)/21 585,26 = 2,25789 2 %], et nous allons donc appliquer ce taux à la population de 1996, qui est de 32 039,66: 32 039,66×1,0225789 = 32 763. Cet estimateur est appelé *estimateur par le quotient*. $\prec$

Le dernier exemple a exposé trois façons, toutes raisonnables, d'estimer la moyenne d'une population. Elles donnent trois estimations différentes:

- Estimation par la moyenne: 22 073;
- Estimation par la différence: 32 527;
- Estimation par le quotient: 32 763.

Il n'est pas question, dans un contexte donné, d'utiliser trois estimateurs pour un même paramètre. Il va donc falloir examiner les propriétés générales de ces estimateurs afin de se faire une idée des conditions dans lesquelles tel estimateur est préférable à tel autre. Essentiellement, nous examinerons les variances de ces estimateurs.

Notation Les observations de la population sont dénotées par

$$x_1, x_2, \ldots, x_N$$
$$y_1, y_2, \ldots, y_N$$

et ceux de l'échantillon par

$$x_1, x_2, \ldots, x_n$$
$$y_1, y_2, \ldots, y_n$$

Les moyennes des deux variables sont dénotées par μ_x et μ_y pour la population, et par $\bar{x}$ et $\bar{y}$ pour l'échantillon. Le paramètre S et la statistique échantillonnale s sont maintenant indicés par x ou par y. Pour la population, ce sont

$$S_x = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu_x)^2}{N-1}} \quad , \quad S_y = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \mu_y)^2}{N-1}}$$

et les quantités analogues pour l'échantillon sont

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad , \quad s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

La ***covariance ajustée*** entre x et y sera dénotée par S_{xy} pour la population et par s_{xy} pour l'échantillon:

$$S_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N-1} \quad , \quad s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Rappelons la définition du coefficient de corrélation, que nous dénoterons par ρ pour la population et par r pour l'échantillon:

$$\rho = \frac{S_{xy}}{S_x S_y}, \quad r = \frac{s_{xy}}{s_x s_y}$$

Les estimateurs que nous introduisons dans ce chapitre exploitent la dépendance qui peut exister entre x et y, et nous verrons que ces estimateurs sont d'autant meilleurs que la dépendance est forte.

4.2 Estimation par la différence

L'estimateur par la différence est

$$\hat{\mu}_{yd} = \mu_x + (\bar{y} - \bar{x})$$

Ici μ_x dénote la moyenne des x pour la population entière.

Justification de l'estimateur

Il y a deux façons de justifier intuitivement cet estimateur.

Première justification On peut décomposer μ_y en deux parties, l'une connue, l'autre pas :

$$\mu_y = \mu_x + (\mu_y - \mu_x)$$

La première partie, μ_x, est connu, et on n'a pas besoin de l'estimer. La deuxième partie, $\mu_y - \mu_x$, la différence entre la moyenne des y et celle des x, n'est pas connue et doit être estimée. On l'estime, naturellement, par la différence entre les deux moyennes échantillonnales, $\bar{y} - \bar{x}$. C'est ce qui donne l'estimateur par la différence.

Deuxième justification L'estimateur de naturel de μ_y est $\bar{y}$ et a priori c'est l'estimateur privilégié. Dans l'estimateur par la différence, écrit comme

$$\hat{\mu}_{yd} = \bar{y} + (\mu_x - \bar{x}),$$

l'ajout du terme $(\mu_x - \bar{x})$ peut s'interpréter comme un *ajustement* à l'estimateur $\bar{y}$. Grâce à notre information sur la variable x, on peut deviner si, en l'occurrence, $\bar{y}$ a surestimé ou sous-estimé μ_y. Reprenons l'exemple traité au tout début du chapitre. La moyenne échantillonnale actuelle est $\bar{y}$ = 22 073 (nombre moyen d'habitants en 2001). Maintenant, nous *savons* que le nombre moyen d'habitants en 1996 était de 32 040—ce n'est pas une estimation, c'est la vraie moyenne. Mais l'échantillon a donné une moyenne de 21 585 en 1996, une différence de 32 040 – 21858 = 10 182 habitants. Notre échantillon, semble-t-il, contient des villes plus petites que celles de la population, et on devine que $\bar{y}$ sous-estime μ_y aussi. Donc on majore $\bar{y}$ de 10 182. L'estimateur est donc $\bar{y}$ + 10 182 = 32 527.

Écart-type de $\hat{\mu}_{yd}$

L'écart-type de $\hat{\mu}_{yd}$ est

$$\sigma_{\hat{\mu}_{yd}} = \sqrt{1-f}\,\frac{\sqrt{S_y^2 + S_x^2 - 2S_{xy}}}{\sqrt{n}} \quad \text{où et } f = \frac{n}{N}$$

Il y a trois paramètres inconnus dans cette expression : S_y^2 , S_x^2 et S_{xy}. Pour estimer cet écart-type à partir de l'échantillon, il suffit de remplacer ces paramètres par leurs analogues échantillonnaux, . s_y^2 , s_x^2 et s_{xy} On obtient alors un estimateur $\hat{\sigma}_{\hat{\mu}_{yd}}$ défini par:

$$\hat{\sigma}_{\hat{\mu}_{yd}} = \sqrt{1-f}\,\frac{\sqrt{s_y^2 + s_x^2 - 2s_{xy}}}{\sqrt{n}}$$

L'intervalle de confiance approximatif à 95 % est donné par

$$\hat{\mu}_{yd} - 2\,\hat{\sigma}_{\hat{\mu}_{yd}} \le \mu_y \le \hat{\mu}_{yd} + 2\,\hat{\sigma}_{\hat{\mu}_{yd}}$$

***Exemple* 4.2.1** *Estimation par la différence*

Considérons encore l'échantillon de 35 municipalités québécoises au tableau A.14, tiré d'une population de taille 180 dont les données sont dans le tableau A.13.

a) Estimer la population moyenne μ_y en 2001, en utilisant la population de 1996 comme variable auxiliaire, sachant que le nombre moyen d'habitants en 1996 était de 32 039,66.

b) Déterminer un intervalle de confiance pour μ_y basé sur l'estimation par la différence.

Solution

Voici les calculs de base :

$\bar{y}$ = 22072,63; $\bar{x}$ = 21585,26; s_y^2 = 4 131 789 466; s_x^2 = 3 989 656 072; et s_{xy} = 4 059 448 772

a) L'estimation est $\hat{\mu}_{yd} = \mu_x + (\bar{y} - \bar{x})$ = 32039,66 + (22072,63 -21585,26) = 32039,66 + 487,3714 = 32527,03.

b) Estimons maintenant l'écart-type de $\hat{\mu}_{yd}$:

$$\hat{\sigma}_{\hat{\mu}_{yd}} = \sqrt{1-f}\,\frac{\sqrt{s_y^2 + s_x^2 - 2s_{xy}}}{\sqrt{n}} = \sqrt{1-35/180}\,\frac{\sqrt{4131789466 + 3989656072 - 2(4059448772)}}{\sqrt{35}}$$

= 242,1659

L'intervalle de confiance est donc 32527,03 ± 2(242,1659) :

$$32042{,}7 \le \mu_y \le 33011{,}36$$

À titre de comparaison, estimons l'écart-type de $\bar{y}$, l'estimateur classique de μ_y :

$$\hat{\sigma}_{\bar{y}} = \sqrt{1-f}\;\frac{s_y}{\sqrt{n}} = \sqrt{1-\frac{35}{180}}\,\frac{\sqrt{4131789466}}{\sqrt{35}} = 9751{,}756.$$

Cet écart-type est énorme comparé à celui de $\hat{\mu}_{y_d}$. Il montre à quel point $\hat{\mu}_{y_d}$ peut être supérieur à $\bar{y}$ dans certains cas. ≺

4.3 Estimation par le quotient

Désignons par R le quotient

$$R = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} = \frac{\mu_y}{\mu_x}$$

et par $\hat{R}$ son estimateur naturel

$$\hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}}$$

L'***estimateur par le quotient*** de la moyenne μ_y est donné par

$$\hat{\mu}_{yq} = \mu_x \hat{R}$$

où μ_x est la moyenne des x pour la population.

Cet estimateur est légèrement biaisé, mais le biais est négligeable lorsque n est grand. L'écart-type de cet estimateur peut être déduit de celui de $\hat{R}$: nous savons que l'écart-type de $\hat{R}$ est donné approximativement par

$$\sigma_{\hat{R}} \approx \sqrt{1-f}\ \frac{\sqrt{S_y^2 + R^2 S_x^2 - 2RS_{xy}}}{\mu_x \sqrt{n}}$$

Donc l'écart-type de $\hat{\mu}_{yq} = \mu_x \hat{R}$ est à peu près

$$\sigma_{\hat{\mu}_{yq}} = \mu_x \sigma_{\hat{R}} = \sqrt{1-f}\ \frac{\sqrt{S_y^2 + R^2 S_x^2 - 2RS_{xy}}}{\sqrt{n}}.$$

Pour estimer cet écart-type à partir de l'échantillon, on remplace les paramètres inconnus S_y^2 , S_x^2 , S_{xy} et R par leurs analogues échantillonnaux, . s_y^2 , s_x^2 , s_{xy} et $\hat{R}$. On obtient alors un estimateur $\hat{\sigma}_{\hat{\mu}_{yd}}$:

$$\hat{\sigma}_{\hat{\mu}_{yq}} = \mu_x \hat{\sigma}_{\hat{R}} = \sqrt{1-f}\ \frac{\sqrt{s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy}}}{\sqrt{n}}$$

L'intervalle de confiance approximatif à 95 % est donné par

$$\hat{\mu}_{y_q} - 2\,\hat{\sigma}_{\hat{\mu}_{yq}} \le \mu_y \le \hat{\mu}_{y_q} + 2\,\hat{\sigma}_{\hat{\mu}_{yq}}$$

***Exemple* 4.3.1** *Estimation par le quotient*

Considérons l'estimateur par le quotient dans le problème traité à l'exemple 4.2.1. On a

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{22072,63}{21585,26} = 1,022579$$

$$\hat{\mu}_{y_q} = \hat{R}\mu_x = 1,022579(32\,039,66) = 32\,763,08$$

L'estimation de l'écart-type est

$$\hat{\sigma}_{\hat{\mu}_{yq}} = \sqrt{1-35/180}\,\frac{\sqrt{4131789466 + (1{,}022579)^2\,3989656072 - 2(1{,}022579)(4059448772)}}{\sqrt{35}} = 181{,}4353.$$

Il y a peu de différence entre cet estimateur et l'estimateur par la différence, dont l'écart-type est estimé à 242,1659.

Intervalle de confiance :

$$32763 - 2(181{,}4353) \leq \mu_y \leq 32763 + 2(181{,}4353)$$

$$32\,400 \leq \mu \leq 32\,763 \quad \prec$$

4.4 Comparaison des estimateurs

Nous avons déterminé, aux exemples 4.2.1 et 4.3.1, trois estimateurs de μ_y et estimé les écarts-types :

Estimateur par la moyenne :	$\bar{y} = 22\,073$	$\hat{\sigma}_{\bar{y}} = 9752$
Estimateur par la différence :	$\hat{\mu}_{yd} = 32\,527$	$\hat{\sigma}_{\hat{\mu}_{yd}} = 242$
Estimateur par le quotient :	$\hat{\mu}_{yq} = 31\,763$	$\hat{\sigma}_{\hat{\mu}_{yq}} = 181$

On constate que les deux estimateurs qui utilisent la variable auxiliaire sont très nettement meilleurs. Les estimations elles-mêmes sont très différentes : de l'ordre de 22 000 pour l'estimateur par la moyenne; et de l'ordre de 32 000 pour les deux autres estimateurs. S'il fallait décider lequel des estimateurs employer, on dirait 32 000 plutôt que 22 000.

Mais il n'est pas question, en pratique, de calculer plusieurs estimateurs et puis leur écart-type. Il faudrait pouvoir faire un choix d'avance. Est-ce qu'on aurait pu prévoir la supériorité des deux derniers estimateurs, et s'en tenir à l'un ou l'autre de ces deux, sans même considérer le premier? Dans le cas présent, on aurait pu le prévoir. En général, les estimateurs qui font appel à une variable auxiliaire sont avantageux dans la mesure où la variable auxiliaire est corrélée positivement avec la variable d'intérêt. Il est évident que le nombre d'habitants en 1996 est corrélé avec le nombre d'habitants en 2001. C'est donc une information pertinente, et les estimateurs par la différence et par le quotient en tirent profit.

C'est en fonction de ρ, le coefficient de corrélation entre x et y, que nous allons distinguer les diverses situations. Commençons par énoncer mathématiquement les conditions dans lesquels les estimateurs par la différence et par le quotient sont préférables à $\bar{y}$. Nous verrons ensuite dans quelle mesure il est possible de s'inspirer de ces conditions pour faire un choix.:

1. L'estimateur par la différence est plus précis que la moyenne $\bar{y}$ si et seulement si $\rho > \frac{1}{2}\frac{S_x}{S_y}$
2. Cette condition est équivalent à $b_1 > \frac{1}{2}$, où b_1 est la pente de la régression de *y* sur *x*
3. L'estimateur par le quotient est plus précis que l'estimateur par la moyenne si et seulement si $\rho > \frac{1}{2}\frac{S_x/\mu_x}{S_y/\mu_y} = \frac{1}{2}\frac{C_x}{C_y}$. C_X et C_Y sont les coefficients de variation de *x* et de *y*, respectivement.

Quel estimateur employer?

Toutes ces conditions sont mathématiquement intéressantes, mais est-ce qu'elles peuvent aider à faire un choix? La première impression est que non, puisqu'elles s'expriment toutes en fonction de paramètres que nous ne connaissons pas : la coefficient de corrélation, les écarts-types, les coefficients de variation. Mais il y a quand même une importante leçon qui se dégage de ces conditions : si la variable auxiliaire est fortement liée à la variable d'intérêt, il vaut mieux s'en servir, c'est-à-dire, utiliser l'estimateur par la différence ou l'estimateur par le quotient. Cela suffit déjà à traiter d'un grand nombre de cas, car on peut souvent savoir, par la nature des variables *x* et *y*, si leur corrélation est positive et élevée. Dans l'exemple discuté jusqu'ici, c'est très clair : on *sait* que le nombre d'habitants en 1996 (*x*) est fortement corrélé avec le nombre d'habitants en 2001 (*y*) : les villes qui étaient petites en 1996 restent petites et les grandes restent grandes : Causapscal n'atteindra pas la taille de Montréal en 5 ans, ni Montréal celle de Duparquet.

Une des applications importantes de cette technique est la vérification et correction de données, telles celles relatives à des transactions ou comptes divers : x_i est la valeur nominale, ou la valeur aux livres, et y_i est la valeur corrigée. Les valeurs x_i et y_i sont donc égales pour la presque totalité des unités de la population, ce qui donne une corrélation très forte, et l'estimation par le quotient ou par la différence est probablement très efficace.

D'autres cas sont moins évidents. Supposons que *y* est la production de maïs dans les fermes d'un territoires et *x* est le nombre d'acres cultivés dans les fermes. Si les valeurs de *x* sont répertoriées pour toute la population, on peut s'en servir dans l'estimation de μ_y. Devrait-on le faire? C'est bien vrai que la corrélation entre *x* et *y* est positive : les grandes fermes produisent plus de maïs que les petites. Mais il y a des fermes, y compris les grandes et petites, qui cultivent très peu de maïs, ou pas du tout. Donc la corrélation n'est pas forte. Si par contre, la population de fermes est limitée à celles où l'on cultive le maïs, la corrélation est probablement plus forte. Est-elle assez forte? Nous ne le savons pas. C'est là qu'une connaissance du contexte est indispendable : ceux qui travaillent dans le domaine, et qui collectent régulièrement des données sur la production des fermes, finissent par accumuler assez d'information pour prendre une décision éclairée.

La condition 1 énoncée plus haut ($\rho > \frac{1}{2}\frac{S_x}{S_y}$) prend, dans certaines situations, une forme simple qui la rend un peu plus facile à vérifier. Si *x* et *y* s'expriment dans les mêmes unités (par exemple, mesurent la même chose, mais à différents moments), on a $S_y \approx S_x$ et la condition 1 devient $\rho > \frac{1}{2}$. Dans l'exemple de ce chapitre, il est clair (a priori) que cette condition est vérifiée.

La deuxième condition. $b_1 > \frac{1}{2}$, est souvent facile à vérifier. Ce paramètre est le taux d'accroissement de *y* par rapport à *x*. Dans l'exemple de ce chapitre, la condition est vérifiée. Pour avoir une idée de sa valeur, il faudrait estimer une réponse à la question suivante : une ville A qui, en 1996 avec 1 000 habitants de plus qu'une autre ville, B, combien a-t-elle de plus que B et 2001? Il n'est pas nécessaire d'avoir une réponse exacte. Il suffit de pourvoir dire si, en moyenne, la réponse est de l'ordre de 1 000 ($b_1 \cong 1$, très vraisemblable); de *plus* que 1 000 ($b_1 > 1$, possible aussi); ou moins de 500 ($b_1 < ½$, fort peu probable).

Les conditions qui recommandent l'estimateur par la différence normalement recommandent aussi l'estimation par le quotient. La condition $\rho > \frac{1}{2}\frac{C_x}{C_y}$ peut sembler plus complexe que $\rho > \frac{1}{2}\frac{S_x}{S_y}$, mais elle aussi devient $\rho > \frac{1}{2}$ dans des situations comme celle traitée en exemple dans ce chapitre.

Reste à savoir si c'est l'estimateur par la différence qu'il faut employer ou l'estimateur par le quotient. Cette question est plus délicate et dans la plupart des cas il sera simplement impossible de trancher. Nous pouvons quand même faire quelques recommandations. Il faut d'abord remarquer une certaine anomalie à propos de l'estimateur par la différence : il est affecté par l'unité de mesure utilisée pour *x*. Supposons, par exemple, que *y* est le nombre de tonnes de maïs et *x* est le nombre d'acres de culture. L'estimateur de μ_y par la différence est $\hat{\mu}_{yd} = \bar{y} + (\mu_x - \bar{x})$, où $\mu_x - \bar{x}$ est une correction apportée à $\bar{y}$. Supposons, pour fixer les idées, qu'on trouve que $\mu_x - \bar{x} = 12$ acres. Donc on ajoute 12 à $\bar{y}$. Mais si on avait décidé de mesurer la superficie en mètres carrés, on aurait ajouté $\mu_x - \bar{x} = 48\,562{,}32$ m^2 à $\bar{y}$. Ce qui peut aisément se produire, c'est que la condition $\rho > \frac{1}{2}\frac{S_x}{S_y}$, qui favorise l'estimateur par la différence (par opposition à $\bar{y}$), peut être vraie lorsque la surface est mesurée en acres mais pas lorsqu'elle est mesurée en mètres carrés. Un inconvénient, en un sens, mais aussi un certain avantage en souplesse : Si on sait que $\rho > 0$ on peut toujours s'arranger pour que la condition $\rho > \frac{1}{2}\frac{S_x}{S_y}$ soit satisfaite : il suffit de choisir de grosses unités pour *x* (des acres plutôt que des mètres carrés, par

exemple), ce qui aura pour effet de réduire S_x et donc $\frac{1}{2}\frac{S_x}{S_y}$. L'estimateur par le quotient ne dépend pas de l'unité de mesure.

Normalement, on évitera l'estimateur par la différence lorsque *x* et *y* ne s'expriment pas dans les mêmes unités, comme dans l'exemple des fermes et du maïs : la différence $\bar{y} - \bar{x}$ est une différence entre un nombre de tonnes et un nombre d'acres, un non-sens. Par contre, dans le premier exemple, *x* et *y* représentent un nombre d'habitants, la différence $\bar{y} - \bar{x}$ représente un accroissement de population, et l'estimateur par la différence est naturel. On le trouve naturel aussi lorsque la différence représente l'erreur dans la valeur aux livres d'un échantillon de transactions.

Lorsque la différence n'a aucun sens concret, on tend à préférer l'estimateur par le quotient, qui a presque toujours un sens concret. Si *x* est le nombre d'acres de culture et *y* la production de maïs, le quotient *R* représente la production par acre. Une mise en garde, cependant : l'estimateur par le quotient est biaisé. Le biais est négligeable, mais seulement lorsque l'échantillon est grand. Il faut donc éviter de l'employer avec des échantillons petits.

4.5 Résumé

1 Estimateur par la différence : $\hat{\mu}_{yd} = \mu_x + (\bar{y} - \bar{x})$; $\hat{\mu}_{yd}$ est sans biais : $E(\hat{\mu}_{yd}) = \mu_y$

Estimateur de l'écart-type : $\hat{\sigma}_{\hat{\mu}_{yd}} = \sqrt{1-f}\,\frac{\sqrt{s_y^2 + s_x^2 - 2s_{xy}}}{\sqrt{n}}$

2 Estimateur par le quotient : $\hat{\mu}_{yq} = \mu_x \hat{R}$; $\hat{\mu}_{yq}$ est légèrement biaisé $E(\hat{\mu}_{yq}) \neq \mu_y$. Mais le biais est négligeable lorsque l'échantillon est grand.

Estimateur de l'écart-type : $\hat{\sigma}_{\hat{\mu}_{yq}} = \mu_x \hat{\sigma}_{\hat{R}} = \sqrt{1-f}\,\frac{\sqrt{s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy}}}{\sqrt{n}}$.

3 L'intervalle de confiance se calcule comme pour les autres paramètres discutés jusqu'ici :

$$\text{Estimateur} - 2\,\hat{\sigma}_{Estimateur} \leq \text{Paramètre} \leq \text{Estimateur} + 2\,\hat{\sigma}_{Estimateur}$$

4.6 Exercices

4.1 Les 2350 abonnés d'un club vidéo ont dépensé la somme de 60 $ en moyenne le mois dernier. Vous voulez estimer les recettes totales τ pour le mois qui vient de s'écouler, à partir d'un échantillon de taille 20. Pour chaque abonné de l'échantillon, vous prenez note de *x*, les dépenses le mois dernier, et de *y*, les dépenses pour ce mois-ci. Voici les données:

Tableau 4.6.1

x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
58,58	52,00	59,85	53,65	46,53	31,84	50,54	55,12
49,79	53,60	48,57	43,17	58,51	57,16	62,20	52,76
83,50	80,49	61,78	67,26	75,27	62,89	66,07	61,33
53,90	54,22	56,30	60,59	77,08	75,39	50,88	51,21
60,63	71,41	49,90	48,41	79,87	79,85	67,87	70,72

Σx =1217,62, s_x = 11,025795 , Σy = 1183,07, s_y = 12,344788 , s_{xy} = 117,078824.

a) Estimer τ par la méthode de la moyenne $\bar{y}$ et déterminer un intervalle de confiance à 95%.

b) Estimer τ par l'estimateur par différence et déterminer un intervalle de confiance à 95%.

c) Estimer τ par l'estimateur par quotient et déterminer un intervalle de confiance à 95%.

4.2 Vous prélevez un échantillon de 20 comptes d'une population de taille 4 890. Vous connaissez les valeurs aux livres des comptes de la population; leur moyenne est de 80 $. Les valeurs aux livres (*x*) et les valeurs corrigées (*y*) suivent:

Tableau 4.6.2

x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
47,90	47,90	60,08	60,08	63,72	63,72	74,82	74,82
64,41	64,41	69,81	69,81	58,01	58,01	58,28	58,28
65,51	65,51	59,96	59,96	55,59	55,59	63,25	63,25
64,91	64,91	71,27	71,27	62,51	62,51	52,34	52,34
91,55	75,25	79,40	59,70	64,45	71,62	72,75	68,19

Σx =1300,52, s_x = 9,808355 , Σy = 1267,13, s_y = 7,212526 , s_{xy} = 56,474343.

a) Estimer τ par la méthode de la moyenne $\bar{y}$ et déterminer un intervalle de confiance à 95%.

b) Estimer τ par l'estimateur par différence et déterminer un intervalle de confiance à 95%.

c) Estimer τ par l'estimateur par quotient et déterminer un intervalle de confiance à 95%.

4.3 Considérez l'échantillon de 50 paroisses québécoises présenté à l'annexe A.04. La population comprend 210 paroisses et on sait que le nombre total de décès dans la population est de 1585. Estimez le nombre total de naissances

a) par la moyenne b) par le quotient, c) par la différence

Dans chaque cas, déterminer un intervalle de confiance; pour b) et c) utilisez le nombre de décès comme variable auxiliaire.

4.4 Considérer l'échantillon de villes québécoises présentée au tableau A.14. Supposons qu'on estime le nombre de naissances en 1998 en utilisant comme variable auxiliaire le nombre d'habitants en 1996. Montrer que l'estimateur par la différence est moins précis (selon les données de l'échantillon) que l'estimateur par le quotient; mais que si le nombre d'habitants est exprimé en milliers, c'est le contraire. La population est de taille 180. Il est utile de connaître les propriétés suivantes : lorsqu'on divise x par 1 000, l'écart-type s_x et la covariance s_{xy} sont divisés par 1 000.

4.5 Dans chacune des situations suivantes, dites lequel des trois estimateurs de la moyenne μ_y serait, d'après vous le meilleur: est-ce l'estimateur par la moyenne, par la différence ou par le quotient? Justifiez votre réponse (qui ne pourra pas toujours être catégorique).

a) La population est l'ensemble des fermes d'une région; μ_y est la récolte moyenne (en milliers de tonnes) de blé. Quel est le meilleur estimateur si

(i) vous connaissez la superficie cultivée dans chacune des fermes de la population?

(ii) vous connaissez le nombre de personnes dans chacune des fermes de la population?

(iii) vous connaissez la récolte des mêmes fermes l'année dernière?

b) La population est l'ensemble des employés d'une grande compagnie; μ_y est le revenu moyen des employés. Vous connaissez les salaires des employés de la population.

c) La population est l'ensemble des employés d'une grande compagnie; μ_y est le revenu moyen des ménages auxquels ils appartiennent. Vous connaissez le salaire des employés de la population.

d) μ_y est la moyenne des recettes d'une population de petits commerces; vous connaissez le nombre d'employés dans chaque commerce.

e) μ_y est la moyenne des dépenses d'une population de petits commerces; vous connaissez les recettes de chaque commerce.

f) μ_y est le nombre moyen de téléviseurs dans les ménages d'une population. Vous connaissez le nombre de personnes dans chaque ménage de la population.

g) μ_y est le nombre moyen de personnes dans les ménages d'une population. Vous connaissez le nombre de personnes dans les mêmes ménages l'année dernière.

4.6 Considérons la population de $N = 8$ unités pour lesquels sont définies deux variables, x et y, dont les valeurs sont données dans le tableau suivant:

y	8	10	67	44	66	56	89	99
x	3	6	24	27	30	36	51	57

Le but de cet exercice est de comparer les estimateurs de μ_y, dans un échantillon de taille $n = 3$ en supposant que les valeurs de x sont connues pour toutes les unités de la population. Voici les valeurs de l'estimateur $\hat{R}\,\mu_x$ pour les 56 échantillons possibles:

46,6304	49,6641	50,9167	53,6250	55,8649	58,9432	61,8429
47,8636	49,7250	51,6486	53,8370	56,1466	59,1724	62,0921
48,1000	49,7250	51,8523	53,8958	56,2250	59,9444	63,0000
48,4934	49,7946	52,1625	55,0403	56,5057	60,5893	63,9167
48,5063	50,2667	52,2097	55,1087	56,7589	60,8214	64,4583
48,7500	50,3750	52,5549	55,4741	56,8750	61,1351	69,7125
48,7500	50,7672	52,7500	55,5000	57,3529	61,4250	72,3553
49,5625	50,8026	53,6250	55,7143	57,5250	61,5000	75,3409

[*Voici quelques données sur ce tableau: somme* = 3113,1949; *somme des carrés* = 175 204,3705.]

a) Montrez que l'estimateur $\hat{R}\,\mu_x$ est biaisé, comme on le sait, et exprimez une opinion sur l'importance du biais.

b) Calculez l'écart-type (le vrai écart-type, à partir du tableau ci-dessus) de $\hat{R}\,\mu_x$, et comparez avec l'écart-type tel que calculé par la formule approximative. (Rappelez-vous que la formule proposée de $\sigma_{\hat{\mu}_{yq}}$ n'est qu'une approximation et ne donne pas le vrai écart-type).

c) Maintenant calculez, à l'aide des formules, l'écart-type de l'estimateur par la moyenne et l'écart-type de l'estimateur par la différence. Comparez les trois estimateurs.

d) Comparez maintenant les estimateurs par le quotient et par la moyenne sous un autre angle. Voici les valeurs de la moyenne $\bar{y}$ pour les 56 échantillons.

20,66667	39,33333	47,00000	54,33333	57,66667	66,00000	73,66667
24,66667	39,66667	47,00000	54,33333	58,00000	66,33333	74,00000
28,00000	40,00000	47,66667	54,66667	58,33333	66,33333	74,00000
28,33333	40,33333	47,66667	55,00000	58,66667	66,66667	77,33333
35,66667	43,33333	50,33333	55,00000	59,00000	69,66667	77,33333
36,00000	43,66667	51,00000	55,33333	63,00000	70,00000	81,33333
36,66667	44,00000	51,00000	55,33333	63,00000	70,33333	84,66667
39,00000	44,33333	51,66667	55,66667	65,33333	70,66667	85,00000

Calculez, pour les estimateurs par la moyenne et par le quotient, la probabilité de se tromper (*i*) de plus de 10 % dans l'estimation de la moyenne; (*ii*) de plus de 20 % dans l'estimation de la moyenne. Présentez les résultats dans un tableau et commentez.

e) Il existe également une formule de $\hat{\sigma}^2_{\hat{R}\mu_x}$, l'estimateur de $\sigma^2_{\hat{R}\mu_x}$. Voici les 56 valeurs de $\hat{\sigma}^2_{\hat{R}\mu_x}$. Dites ce que vous pensez de cet estimateur à la lumière de ces observations.

5,90393	13,12385	1,66642	46,09245	0,90109	0,03693	72,89673	3,53489
1,49884	67,80659	24,50063	15,08125	0,83958	45,52726	70,96071	3,57461
1,33629	63,41100	17,15109	65,14979	23,12500	69,90413	72,48644	0,90504
2,04815	69,34418	19,38403	58,69427	15,04558	62,35366	66,67237	23,96011
1,17865	14,18958	4,24375	63,63927	16,77419	65,17940	25,08585	24,31291
1,22719	1,88017	4,12862	14,13265	3,33333	69,65833	17,99495	18,22306
45,96901	1,74640	1,15976	0,31664	3,33333	65,14813	18,90625	3,90770

Voici quelques données sur ces 56 nombres: somme = 1 490,58602; somme des carrés = 79 723,73416.