

MARC-ANDRÉ DUBÉ

# Une introduction à l'échantillonnage

Essai présenté  
à la Faculté des études supérieures de l'Université Laval  
dans le cadre du programme de maîtrise en statistique  
pour l'obtention du grade de Maître ès sciences (M.Sc.)

FACULTÉ DES SCIENCES ET DE GÉNIE  
UNIVERSITÉ LAVAL  
QUÉBEC

Décembre 2005

©Marc-André Dubé, 2005

# Résumé

Cet essai se veut une introduction à une méthode d'échantillonnage appelée "échantillonnage," ou "ranked set sampling" en anglais. La technique est d'abord présentée dans le cas d'une population univariée dont on cherche à estimer l'espérance. On montre que sous des hypothèses assez faibles, la moyenne expérimentale d'un échantillon est sans biais et que sa variance est inférieure à celle de l'estimateur traditionnel basé sur un échantillonnage aléatoire simple de même taille. Une généralisation de la technique d'échantillonnage est ensuite considérée dans le cas d'une population à deux caractères aléatoires. On montre que dans les mêmes conditions que pour le cas univarié, l'efficacité de l'estimateur fondé sur l'échantillon bivarié est supérieure à celle de la moyenne bivariée calculée à partir d'un échantillonnage aléatoire simple.

# Avant-propos

Je désire d'abord remercier M. Christian Genest d'avoir accepté de diriger mes recherches et de les avoir financées à l'aide de fonds octroyés par le Conseil de recherches en sciences naturelles et en génie du Canada, ainsi que par le Fonds québécois de la recherche sur la nature et les technologies.

J'aimerais aussi profiter de l'occasion pour remercier toutes les personnes avec qui j'ai eu beaucoup de plaisir au cours de toutes mes années d'étude. Tout spécialement, j'aimerais remercier Yannick Thiffeault, un ami d'enfance avec qui je suis particulièrement heureux d'avoir pu garder contact jusqu'à aujourd'hui. Ensuite, mon groupe du Collège Shawinigan : Dominic Blais, Simon Durand, Chantal Grenier, David Lafontaine, Stéphane Mailhot, Francis Pronovost et bien d'autres.

J'ai aussi une bonne pensée pour toutes les personnes que j'ai connues à l'Université du Québec à Trois-Rivières, dont Hanni Boulet, Patrice Gagnon, Guillaume Giguère et Jean-Marc Lévesque, qui m'a ensuite rejoint à la maîtrise.

Finalement, je dois remercier du plus profond de mon cœur ma sœur, Véronique, ainsi que mes parents, Céline et René. Sans leur soutien, je n'aurais probablement jamais été en mesure de poursuivre mes études si loin. Je ne pourrai jamais assez les remercier pour tout ce qu'ils ont fait pour moi. Évidemment, j'ai bien apprécié connaître d'autres personnes, mais elles ont joué un rôle un peu moins important dans ma vie !

Malheureusement, je ne peux pas nommer tout le monde, mais sachez que j'ai tout de même une bonne pensée pour vous. Ainsi, toutes les personnes mentionnées ci-dessus ont été, à différents degrés, importantes dans ma vie jusqu'à maintenant. Sachez que je ne l'oublierai jamais.

# Table des matières

Résumé	ii
Avant-Propos	iii
Table des matières	v
Liste des tableaux	vi
Table des figures	vii
<b>1 Introduction</b>	<b>1</b>
<b>2 L'échantillonnage univarié</b>	<b>3</b>
2.1 Notation . . . . .	4
2.2 Résultat préliminaire . . . . .	5
2.3 La moyenne échantillonnale . . . . .	5
2.4 La variance de la moyenne échantillonnale . . . . .	6
2.5 Taille d'échantillon . . . . .	7
2.6 Estimation de la fonction de répartition . . . . .	8
2.7 Procédure de Mann–Whitney . . . . .	9
2.8 Illustration . . . . .	10
<b>3 Échantillonnage bivarié</b>	<b>14</b>
3.1 Notation . . . . .	16
3.2 Résultats préliminaires . . . . .	16
3.3 Les moyennes échantillonnales . . . . .	19
3.4 La variance des moyennes échantillonnales . . . . .	20
3.5 Comparaisons élémentaires . . . . .	22
3.6 Comparaisons plus fines . . . . .	23
3.6.1 Le cas d'indépendance . . . . .	24
3.6.2 Le cas de dépendance parfaite . . . . .	24
3.6.3 Le cas de dépendance linéaire . . . . .	24
3.7 Cas de la loi normale bivariée . . . . .	25

3.8 Illustration . . . . .	28
4 Conclusion	30

# Liste des tableaux

2.1	Calculs intermédiaires nécessaires à la détermination de la loi de la statistique $U_{ETU}$ lorsque deux échantillons de taille 2 à un cycle chacun sont tirés d'une loi uniforme. . . . .	11
2.2	Loi de la statistique $U_{ETU}$ lorsque deux échantillons de taille 2 à un cycle chacun sont tirés d'une loi uniforme. . . . .	12

# Table des figures

3.1	Graphe de $\text{eff}_\rho(\hat{\mu}_{ETB} \hat{\mu}_{EAS})$ en fonction de $\rho$ dans le cas normal bivarié, lorsque $k = 2$ . . . . .	29
-----	--	----

# Chapitre 1

## Introduction

L'échantillonnage joue un rôle central en statistique et les méthodes d'échantillonnage se sont beaucoup développées au fil du temps. Parmi celles-ci, la plus connue et la plus simple est sans aucun doute l'échantillonnage aléatoire simple. Dans un échantillon de ce genre, toutes les unités de la population ont la même probabilité d'être choisies. Bien que l'augmentation de la taille d'un échantillon aléatoire simple permet généralement d'accroître la précision de l'inférence, ce mode de collecte de données n'est pas toujours le plus approprié, particulièrement lorsque l'on s'intéresse à des sous-populations.

Lorsqu'une analyse plus fine de sous-populations est envisagée, on fait typiquement appel à des méthodes d'échantillonnage plus structurées que l'échantillonnage aléatoire simple. C'est le cas de l'échantillonnage stratifié ou par grappes, par exemple. Les méthodes traditionnelles d'échantillonnage ont toutefois un point en commun : la nécessité de mesurer la ou les caractéristique(s) d'intérêt sur chacune des unités sélectionnées.

Cet essai porte sur une méthode de collecte de données qui se distingue des techniques classiques sur ce point. Appelée "ranked set sampling" par McIntyre (1952), cette stratégie d'échantillonnage consiste à recueillir, selon un certain protocole, un nombre d'observations beaucoup plus grand que celles sur lesquelles des mesures seront éventuellement prises. Ces observations, regroupées en échantillons aléatoires simples, sont alors triées au sein de chaque groupe, mais sans avoir recours à un instrument de mesure. En supposant que cette opération puisse être effectuée correctement (et à faible coût), on est alors en mesure d'extraire de chaque groupe une statistique d'ordre, sur laquelle la ou les caractéristique(s) d'intérêt sera (seront) mesurée(s).

Puisque cette technique fait appel à la fois à de l'échantillonnage et à un tri, le terme "échan**tr**illonnage" sera utilisé ici pour la décrire. Comme McIntyre (1952) l'a fait valoir, cette technique est avantageuse dans les situations où la prise de mesures sur les observations s'avère difficile, coûteuse ou destructive. De plus, comme nous le verrons, la structure supplémentaire apportée par le fait que les mesures sont prises sur des statistiques d'ordre est susceptible de mener, à taille comparable, à des estimations plus précises que par l'échantillonnage aléatoire simple. Toutefois, il faut pour cela que le tri des observations dont sont extraites les statistiques d'ordre puisse se faire sans erreur.

Une introduction à l'échantrillonnage univarié sera présentée au chapitre 2. Le protocole d'échantrillonnage y sera précisé et quelques travaux classiques portant sur cette technique seront relatés. En particulier, une comparaison y sera faite de l'efficacité d'une estimation de la moyenne d'une population, selon que les données ont été recueillies par échantillonnage aléatoire simple ou par échantrillonnage.

Le chapitre 3 présentera pour sa part une récente généralisation de l'échantrillonnage au cas bivarié. Ce chapitre s'appuiera largement sur un article de Al-Saleh & Zheng (2002).

Une courte conclusion sera donnée au chapitre 4.

# Chapitre 2

## L'échantrillonnage univarié

L'échantrillonnage se distingue des autres méthodes de cueillette de données en ceci qu'il nécessite la sélection d'un plus grand nombre d'unités qu'il y en aura éventuellement dans l'ensemble final.

Comment procède-t-on à la sélection d'un échantillon de taille  $k$ ? Dans un premier temps, on doit extraire de la population à étudier  $k$  échantillons aléatoires simples de taille  $k$  chacun. Au sein de chacun de ces échantillons, il faut ensuite trier les  $k$  observations en ordre croissant. Ce tri doit pouvoir se faire sans erreur, mais sans pour autant déterminer la valeur précise des variables. On peut recourir pour ce faire à des comparaisons visuelles, à l'opinion d'un expert ou à toute autre procédure permettant d'ordonner les unités autrement qu'en mesurant quoi que ce soit. Nous supposons partout dans la suite que ce classement peut être effectué *sans erreur*.

Une fois les  $k$  échantillons triés, chacun d'entre eux fournira une et une seule unité à l'échantillon. Ainsi, la première observation retenue sera celle qui a été jugée la plus petite dans le premier échantillon. Le deuxième élément sera la deuxième plus petite observation du deuxième échantillon et ainsi de suite, jusqu'à la  $k^e$  unité, qui sera la plus grande observation du  $k^e$  échantillon.

Les mesures qui seront éventuellement utilisées aux fins d'inférence statistique ne seront prises que sur les  $k$  observations ainsi sélectionnées. Les autres unités ne joueront aucun rôle par la suite. Elles n'auront servi en fin de compte qu'à extraire de chaque échantillon *une statistique d'ordre*, laquelle contient toutefois implicitement une certaine quantité d'information concernant les unités non mesurées.

## 2.1 Notation

Soit  $F$  la fonction de répartition de la population d'intérêt, dont on suppose qu'elle admet une densité  $f$ . Soient en outre  $\mu \in \mathbb{R}$  et  $\sigma^2 > 0$  l'espérance et la variance de cette population. Admettons que l'on cherche à estimer  $\mu$  par échantrillonnage.

Par construction, un échantrillon univarié (ETU) est formé de variables aléatoires  $X_{(1)}^*, \dots, X_{(k)}^*$  telles que  $X_{(i)}^*$  a la même distribution que la  $i^e$  statistique d'ordre d'un échantillon aléatoire de taille  $k$  de loi  $F$ . Bien qu'ils ne soient pas identiquement distribués, les éléments de l'échantrillon sont mutuellement indépendants, puisqu'il s'agit de statistiques d'ordre extraites d'échantillons distincts.

Il est important de noter ici que bien que  $X_{(1)}^*$  représente la plus petite observation du premier échantillon, et donc que sa distribution est celle du minimum d'un échantillon de taille  $k$ , cette variable aléatoire ne constitue pas nécessairement le plus petit élément de l'échantrillon. De même,  $X_{(2)}^*$  n'est pas assurément le deuxième plus petit et ainsi de suite.

Les variables aléatoires  $X_{(1)}^*, \dots, X_{(k)}^*$  composant un échantrillon constituent collectivement ce que l'on appelle traditionnellement un *cycle*. En répétant le cycle à  $m$  reprises indépendantes, on obtient alors un échantrillon de taille  $km$  observations. Les éléments de l'échantrillon seront alors notés  $X_{(1)j}^*, \dots, X_{(k)j}^*$ , où  $j = 1, \dots, m$ . La constitution d'un tel échantrillon nécessite donc la sélection (et le tri) de  $k^2$  observations pour chacun des  $m$  cycles.

Pour la suite, notons par  $X_1, \dots, X_k$ , les éléments d'un échantillon aléatoire simple (EAS) et par  $X_{(1)} < \dots < X_{(k)}$  les statistiques d'ordre qui y sont associées. Les fonctions de densité d'un échantillon aléatoire simple et d'un échantrillon sont respectivement données par

$$g_{EAS}(x_1, \dots, x_k) = \prod_{i=1}^k f(x_i).$$

et

$$g_{ETU}(x_1, \dots, x_k) = \prod_{i=1}^k f_i(x_i),$$

où

$$f_i(x) = \frac{k!}{(i-1)!(k-1)!} \{F(x)\}^{i-1} \{1-F(x)\}^{k-i} f(x) \quad (2.1)$$

est la fonction de densité de la  $i^e$  statistique d'ordre d'un EAS de taille  $k$  issu de la loi  $F$ .

C'est la structure supplémentaire provenant de l'ordonnement et de l'indépendance des statistiques d'ordre  $X_{(1)}^*, \dots, X_{(k)}^*$  qui permet aux procédures basées sur l'échantrillonnage d'être plus efficaces que celles basées sur l'échantillonnage aléatoire simple effectué avec le même nombre de données mesurées. En revanche, cette même structure rend le développement théorique des propriétés de l'échantrillonnage plus difficile que celles de l'échantillonnage aléatoire simple.

## 2.2 Résultat préliminaire

Le résultat suivant est utilisé à plusieurs reprises dans la suite.

**Lemme 2.1** *Soient  $X_{(1)} < \dots < X_{(k)}$  les statistiques d'ordre associées à un échantillon aléatoire  $X_1, \dots, X_k$  de densité  $f$ . Soit  $f_i$  la densité de  $X_{(i)}$ , telle que définie en (2.1). Alors*

$$\sum_{i=1}^k f_i(x) = kf(x), \quad x \in \mathbb{R}.$$

La démonstration de ce résultat est très simple. En effet, pour une valeur de  $x$  donnée et pour  $p = F(x)$ , on a par définition

$$\begin{aligned} \sum_{i=1}^k f_i(x) &= kf(x) \sum_{i=1}^k \binom{k-1}{i-1} \{F(x)\}^{i-1} \{1-F(x)\}^{k-i} \\ &= kf(x) \sum_{j=0}^{k-1} \binom{k-1}{j} p^j (1-p)^{k-1-j} \\ &= kf(x), \end{aligned}$$

puisque la somme représente la probabilité qu'une variable binomiale de paramètres  $k-1$  et  $p$  prenne une valeur quelconque dans l'ensemble  $\{0, \dots, k-1\}$ .

## 2.3 La moyenne échantrillonnale

Soit

$$\bar{X} = \frac{1}{k} (X_1 + \dots + X_k)$$

la moyenne expérimentale d'un EAS. Il est bien connu que  $\bar{X}$  est un estimateur sans biais de la moyenne  $\mu$  de la population totale. Autrement dit,

$$E(\bar{X}) = \mu.$$

Mais en est-il de même pour

$$\bar{X}^* = \frac{1}{k} (X_1^* + \dots + X_k^*),$$

la moyenne échantillonnale?

La réponse est oui, et le résultat découle immédiate du Lemme 2.1. En effet, on a

$$\begin{aligned} E(\bar{X}^*) &= \frac{1}{k} \sum_{i=1}^k E(X_{(i)}^*) = \frac{1}{k} \sum_{i=1}^k \int x f_i(x) dx \\ &= \frac{1}{k} \int x \left\{ \sum_{i=1}^k f_i(x) \right\} dx \\ &= \frac{1}{k} \int x \{k f(x)\} dx = \mu, \end{aligned}$$

ce qui permet de conclure.

## 2.4 La variance de la moyenne échantillonnale

Nous allons maintenant nous intéresser à la variance de la moyenne échantillonnale. Posons d'abord  $\mu_i = E(X_{(i)}^*)$  et  $\sigma_i^2 = \text{var}(X_{(i)}^*)$  pour tout  $i \in \{1, \dots, k\}$ .

Puisque les variables  $X_{(1)}^*, \dots, X_{(k)}^*$  sont mutuellement indépendantes, nous avons alors

$$\begin{aligned} \text{var}(\bar{X}^*) &= \frac{1}{k^2} \sum_{i=1}^k \sigma_i^2 \\ &= \frac{1}{k^2} \sum_{i=1}^k \int (x - \mu_i)^2 f_i(x) dx \\ &= \frac{1}{k^2} \sum_{i=1}^k \int \{(x - \mu)^2 + 2(\mu - \mu_i)(x - \mu) + (\mu - \mu_i)^2\} f_i(x) dx. \end{aligned}$$

Or, en vertu de Lemma 2.1,

$$\sum_{i=1}^k \int (x - \mu)^2 f_i(x) dx = \int (x - \mu)^2 k f(x) dx = k \text{var}(X) = k\sigma^2$$

et

$$2 \sum_{i=1}^k (\mu - \mu_i) \int (x - \mu) f_i(x) dx = -2 \sum_{i=1}^k (\mu - \mu_i)^2,$$

de sorte qu'au total,

$$\frac{1}{k^2} \sum_{i=1}^k \sigma_i^2 = \frac{1}{k^2} \left\{ k\sigma^2 - \sum_{i=1}^k (\mu - \mu_i)^2 \right\} = \frac{\sigma^2}{k} - \frac{1}{k^2} \sum_{i=1}^k (\mu - \mu_i)^2 \quad (2.2)$$

et donc

$$\text{var}(\bar{X}^*) \leq \frac{\sigma^2}{k} = \text{var}(\bar{X}).$$

Cette inégalité revient à dire que l'estimateur de la moyenne fondé sur un échantrillon est plus précis que celui qui est déduit d'un échantillon aléatoire simple. Bien que la démonstration que nous venons de donner ne vaut que si le tri des observations se fait sans erreur, le résultat reste vrai à moins que les rangs servant à l'échantrillonnage aient été attribués aléatoirement. Wolfe (2004), qui souligne ce fait, mentionne qu'en cas de tri aléatoire, on aura  $\mu_1 = \dots = \mu_k = \mu$ , ce qui conduira alors à l'égalité des variances.

Takahasi & Wakimoto (1968) ont aussi démontré que

$$1 \leq \text{eff}(\bar{X}^*|\bar{X}) \equiv \frac{\text{var}(\bar{X})}{\text{var}(\bar{X}^*)} \leq \frac{1}{2}(m+1),$$

où la borne supérieure est atteinte si et seulement si nous sommes en présence de la loi uniforme. Comme Dell & Clutter (1972) l'ont fait remarquer, la valeur de  $\text{eff}(\bar{X}^*|\bar{X})$  est d'ailleurs proche du maximum pour un grand nombre de lois unimodales.

## 2.5 Taille d'échantrillon

Une question très importante que se pose toute personne voulant faire une étude statistique concerne le choix de la taille de l'échantillon. L'échantrillonnage n'y échappe pas. Bien qu'il soit impossible de répondre à cette question de façon tout à fait générale, deux éléments devraient être pris en compte au moment de prendre une décision à ce sujet.

D'une part, comme chaque donnée mesurée amène une information supplémentaire du fait de son rang parmi les  $k$  unités de son échantillon, il est évident que plus  $k$  est

élevé, plus nous obtiendrons d'information additionnelle si tous les rangs sont attribués de façon exacte.

D'autre part, plus la taille est grande, plus il sera difficile d'établir un ordre qui, faut-il le rappeler, est déterminé sans prendre une seule mesure sur les observations. Ainsi, le risque d'erreur augmente à mesure que croît la taille de l'échantillon. En somme, il faut être capable de trouver un certain équilibre entre ces deux aspects du problème.

Par ailleurs, les contraintes monétaires entrent aussi en ligne de compte lorsqu'il faut déterminer la taille d'un échantillon ou d'un échantillon. Afin de pouvoir fixer la taille optimale, il faudra donc être capable d'estimer les probabilités de commettre des erreurs dans les rangs ainsi que d'avoir une bonne idée de l'impact de ces erreurs potentielles sur les procédures statistiques qui seront utilisées ultérieurement.

Enfin, notons qu'il est parfois avantageux de recourir à une forme d'échantillonnage non équilibré, c'est-à-dire dans laquelle chacune des statistiques d'ordre n'apparaît pas obligatoirement une seule fois dans l'échantillon  $X_{(1)}^*, \dots, X_{(k)}^*$ . En effet, prenons le cas où nous avons une distribution unimodale et symétrique autour de la médiane. Supposons que nous désirons faire de l'inférence sur cette médiane à l'aide d'un échantillon de taille  $k$  impaire. Dans cette situation, il serait adéquat de prendre la médiane  $X_{(k+1)/2}$  de chacun des  $k$  échantillons afin de bâtir l'échantillon en question. Toutefois, cette problématique ne sera pas considérée plus avant dans cet essai.

## 2.6 Estimation de la fonction de répartition

Stokes & Sager (1988) ont montré comment estimer une fonction de répartition à partir de l'information supplémentaire qu'apporte l'échantillonnage. Supposons que nous disposons de  $X_{(1)j}^*, \dots, X_{(k)j}^*$  pour tout  $j \in \{1, \dots, m\}$ . En d'autres termes, supposons que nous avons en main un échantillon de taille  $k$  et de  $m$  cycles tiré d'une population de loi  $F$ .

L'estimation proposée par Stokes & Sager (1988) est donnée par

$$F^*(t) = \frac{1}{mk} \sum_{i=1}^k \sum_{j=1}^m I_{(-\infty, t)}(X_{(i)j}^*).$$

Ces auteurs ont montré que  $F^*$  est une estimation non biaisée de  $F$ . De plus, ils ont établi que si  $\hat{F}$  est la fonction de répartition empirique d'un échantillon aléatoire simple

de taille  $mk$ , alors

$$\text{var}\{F^*(t)\} \leq \text{var}\{\hat{F}(t)\}$$

pour tout  $t \in \mathbb{R}$ . En d'autres termes, l'estimation de  $F$  extraite d'un échantrillon est à la fois sans biais et plus précise, en moyenne, que celle déduite d'un échantillon aléatoire simple par la méthode classique.

## 2.7 Procédure de Mann–Whitney

Bohn & Wolfe (1992) ont exploité les travaux de Stokes & Sager (1988) dans le but de développer l'équivalent de la procédure de Mann–Whitney dans le cas de deux échantrillons.

Soit  $X_{(1)j}^*, \dots, X_{(k)j}^*$  un premier échantrillon de taille  $k$  à  $m$  cycles, où  $j \in \{1, \dots, m\}$ . Soit aussi  $Y_{(1)t}^*, \dots, Y_{(q)t}^*$  un deuxième échantrillon de taille  $q$  à  $n$  cycles, où  $t \in \{1, \dots, n\}$ . Comme dans le cadre classique de développement du test de Mann–Whitney, supposons que les deux échantrillons proviennent de populations continues indépendantes ayant pour fonctions de répartition  $F$  et  $G$  identiques, à une translation près. En d'autres mots, supposons qu'il existe  $\Delta \in \mathbb{R}$  tel que

$$G(t) = F(t - \Delta), \quad t \in \mathbb{R}.$$

Dénotons respectivement par  $F_{m,k}^*$  et  $G_{n,q}^*$  les fonctions de répartition empiriques des échantrillons  $X$  et  $Y$ , telles que définies par Stokes & Stager (1988). Finalement, posons

$$\Psi(t) = \begin{cases} 1 & \text{si } t \geq 0; \\ 0 & \text{sinon.} \end{cases}$$

La version de la statistique de Mann–Whitney pour deux échantrillons univariés est alors donnée par

$$\begin{aligned} U_{ETU} &= \#\{X^* \leq Y^*\} = \sum_{s=1}^q \sum_{t=1}^n \sum_{i=1}^k \sum_{j=1}^m \Psi(Y_{(s)t}^* - X_{(i)j}^*) \\ &= mnkq \int_{-\infty}^{\infty} F_{m,k}^*(t) dG_{n,q}^*(t). \end{aligned}$$

Sous l'hypothèse  $H_0 : \Delta = 0$ , les variables aléatoires  $X$  et  $Y$  sont indépendantes, mais contrairement au cas de l'échantillonnage aléatoire simple, elles ne sont pas identiquement distribuées. Pour cette raison, nous devons calculer la probabilité de chacune

des  $(mk + nq)!$  permutations possibles et, ensuite, les combiner afin d'obtenir la distribution de  $U_{ETU}$  sous l'hypothèse nulle.

## 2.8 Illustration

Soient  $X$  et  $Y$ , deux échantrillons de même taille à un cycle chacun, c'est-à-dire tels que  $k = q = 2$  et  $m = n = 1$ . L'évaluation des probabilités de toutes les combinaisons possibles sous l'hypothèse nulle nécessite donc  $4! = 24$  calculs différents, lesquels dépendent tous de la densité conjointe des variables  $X_{(1)1}, X_{(2)1}, Y_{(1)1}$  et  $Y_{(2)1}$ .

En supposant comme toujours l'absence d'erreurs dans le tri des observations, les variables  $X_{(1)1}, X_{(2)1}, Y_{(1)1}$  et  $Y_{(2)1}$  sont des statistiques d'ordre indépendantes, dont la fonction de densité conjointe est donnée par

$$g_{ETU}(x_{(1)}, x_{(2)}, y_{(1)}, y_{(2)}) =$$

$$\left[ \prod_{i=1}^2 \frac{2!}{(i-1)!(2-i)!} \{F(x_{(i)})\}^{i-1} \{1 - F(x_{(i)})\}^{2-i} f(x_{(i)}) \right] \\ \times \left[ \prod_{s=1}^2 \frac{2!}{(s-1)!(2-s)!} \{F(y_{(s)})\}^{s-1} \cdot \{1 - F(y_{(s)})\}^{2-s} f(y_{(s)}) \right].$$

Cette dernière expression peut se simplifier comme suit :

$$g_{ETU}(x_{(1)}, x_{(2)}, y_{(1)}, y_{(2)}) =$$

$$16 \{1 - F(x_{(1)})\} F(x_{(2)}) \{1 - F(y_{(1)})\} F(y_{(2)}) \prod_{i=1}^2 f(x_{(i)}) \prod_{s=1}^2 f(y_{(s)}).$$

À titre d'illustration, supposons que  $F$  soit la fonction de répartition d'une variable aléatoire uniforme sur l'intervalle  $(0, 1)$ . On trouve alors

$$g_{ETU}(x_{(1)}, x_{(2)}, y_{(1)}, y_{(2)}) = 16(1 - x_{(1)})x_{(2)}(1 - y_{(1)})y_{(2)}.$$

En supposant par exemple que

$$x_{(1)} < x_{(2)} < y_{(1)} < y_{(2)},$$

on trouve  $U_{ETU} = 4$ . La probabilité de cet événement est donnée par

$$\int_0^1 \int_{x_1}^1 \int_{x_2}^1 \int_{y_1}^1 16(1-x_1)x_2(1-y_1)y_2 dy_2 dy_1 dx_2 dx_1 = \frac{137}{2520}$$

sous l'hypothèse nulle. Le traitement des 23 autres cas se fait de manière semblable. Le tableau 2.1 résume ces calculs.

Comme nous pouvons le constater à la lecture du tableau 2.1, les probabilités sont égales par groupe de 4. Si nous remarquons bien, ces probabilités sont celles où les

TAB. 2.1 – Calculs intermédiaires nécessaires à la détermination de la loi de la statistique  $U_{ETU}$  lorsque deux échantrillons de taille 2 à un cycle chacun sont tirés d'une loi uniforme.

Combinaison	Prob. sous $H_0$	$U_{ETU}$
$x_{(1)} < x_{(2)} < y_{(1)} < y_{(2)}$	$\frac{137}{2520}$	4
$x_{(1)} < x_{(2)} < y_{(2)} < y_{(1)}$	$\frac{7}{360}$	4
$x_{(1)} < y_{(1)} < x_{(2)} < y_{(2)}$	$\frac{41}{280}$	3
$x_{(1)} < y_{(1)} < y_{(2)} < x_{(2)}$	$\frac{41}{280}$	2
$x_{(1)} < y_{(2)} < x_{(2)} < y_{(1)}$	$\frac{7}{360}$	3
$x_{(1)} < y_{(2)} < y_{(1)} < x_{(2)}$	$\frac{137}{2520}$	2
$x_{(2)} < x_{(1)} < y_{(1)} < y_{(2)}$	$\frac{7}{360}$	4
$x_{(2)} < x_{(1)} < y_{(2)} < y_{(1)}$	$\frac{17}{2520}$	4
$x_{(2)} < y_{(1)} < x_{(1)} < y_{(2)}$	$\frac{7}{360}$	3
$x_{(2)} < y_{(1)} < y_{(2)} < x_{(1)}$	$\frac{17}{2520}$	2
$x_{(2)} < y_{(2)} < x_{(1)} < y_{(1)}$	$\frac{1}{280}$	3
$x_{(2)} < y_{(2)} < y_{(1)} < x_{(1)}$	$\frac{1}{280}$	2
$y_{(1)} < x_{(1)} < x_{(2)} < y_{(2)}$	$\frac{41}{280}$	2
$y_{(1)} < x_{(1)} < y_{(2)} < x_{(2)}$	$\frac{41}{280}$	1
$y_{(1)} < x_{(2)} < x_{(1)} < y_{(2)}$	$\frac{137}{2520}$	2
$y_{(1)} < x_{(2)} < y_{(2)} < x_{(1)}$	$\frac{7}{360}$	1
$y_{(1)} < y_{(2)} < x_{(1)} < x_{(2)}$	$\frac{137}{2520}$	0
$y_{(1)} < y_{(2)} < x_{(2)} < x_{(1)}$	$\frac{7}{360}$	0
$y_{(2)} < x_{(1)} < x_{(2)} < y_{(1)}$	$\frac{17}{2520}$	2
$y_{(2)} < x_{(1)} < y_{(1)} < x_{(2)}$	$\frac{7}{360}$	1
$y_{(2)} < x_{(2)} < x_{(1)} < y_{(1)}$	$\frac{1}{280}$	2
$y_{(2)} < x_{(2)} < y_{(1)} < x_{(1)}$	$\frac{1}{280}$	1
$y_{(2)} < y_{(1)} < x_{(1)} < x_{(2)}$	$\frac{7}{360}$	0
$y_{(2)} < y_{(1)} < x_{(2)} < x_{(1)}$	$\frac{17}{2520}$	0

statistiques d'ordre sont dans le même ordre, qu'il s'agisse de la variable  $X$  ou  $Y$ . En fait, ce résultat était à prévoir, car sous  $H_0$ ,  $X$  et  $Y$  sont identiques. À l'aide de ce tableau, nous sommes maintenant en mesure d'établir la loi de  $U_{ETU}$  sous  $H_0$ . Celle-ci est précisée dans le tableau 2.2.

TAB. 2.2 – Loi de la statistique  $U_{ETU}$  lorsque deux échantillons de taille 2 à un cycle chacun sont tirés d'une loi uniforme.

Prob. de $U_{ETU}$ sous $H_0$
$P(U_{ETU} = 0) = \frac{1}{10}$
$P(U_{ETU} = 1) = \frac{17}{90}$
$P(U_{ETU} = 2) = \frac{19}{45}$
$P(U_{ETU} = 3) = \frac{17}{90}$
$P(U_{ETU} = 4) = \frac{1}{10}$

En consultant le tableau 2.2, on observe que la distribution de  $U_{ETU}$  sous  $H_0$  est symétrique par rapport à sa moyenne. Plus précisément, on a ici

$$E(U_{ETU}) = \frac{mnkq}{2} = 2.$$

Toutefois, cette remarque tient pour tous les  $m, n, k, q$  possibles.

**Résultat** Soient  $N = m + n$  et  $\lambda = \lim_{N \rightarrow \infty} m/N$ . Si

$$0 < \lambda < 1 \quad \text{et} \quad \lim_{N \rightarrow \infty} \frac{N}{m^2 n^2} \text{var}(U_{ETU}) > 0,$$

alors

$$\frac{\sqrt{N}}{mn} \{U_{ETU} - E(U_{ETU})\} \rightsquigarrow \mathcal{N}(0, \sigma_\infty^2).$$

Une formule pour le calcul de  $\sigma_\infty^2$  a été donnée par Bohn & Wolfe (1992). Ces auteurs ont également démontré que sous l'hypothèse nulle  $H_0 : \delta = 0$ , ni  $E(U_{ETU}) = mnkq/2$  ni la variance asymptotique  $\sigma_\infty^2$  ne dépendent de la fonction de répartition  $F$ .

Le résultat précédent peut être utilisé afin d'approximer les valeurs critiques pour le test basé sur l'échantrillonnage univarié de  $H_0 : \delta = 0$  pour des valeurs données de  $k$  et de  $q$ . Prenons par exemple  $m = n$ , de sorte que  $\lambda = 1/2$ . Supposons de plus que  $k = q = 2$ . D'après les travaux de Bohn & Wolfe (1992), nous avons alors  $\sigma_\infty^2 = 16/9$ . Ainsi, la distribution asymptotique sous l'hypothèse nulle de

$$\frac{\sqrt{N}}{mn} \{U_{ETU} - E_0(U_{ETU})\} = \sqrt{2n} \left( \frac{U_{ETU}}{n^2} - 2 \right)$$

est  $\mathcal{N}(0, 16/9)$ . De là, nous pouvons déduire que

$$P \left[ \left\{ \sqrt{2n} \left( \frac{U_{ETU}}{n^2} - 2 \right) \right\} \geq z_{(\alpha)} \right] \approx \alpha,$$

où  $z_{(\alpha)}$  est le  $\alpha^e$  percentile de la loi normale centrée réduite. Dans l'exemple ci-dessus, le  $\alpha^e$  percentile de la distribution sous  $H_0$  est donc donné par

$$(n^{3/2}/\sqrt{2})z_\alpha + 2n^2.$$

# Chapitre 3

## Échantrillonnage bivarié

Al-Saleh & Zheng (2002) ont récemment proposé une généralisation de la méthode d'échantrillonnage de McIntyre (1952) à l'estimation de deux caractéristiques simultanées. De manière à expliquer leur approche le plus simplement possible, nous allons nous limiter ici au cas de  $k = 2$  unités. Dans cette situation, les paires de rangs possibles des deux caractéristiques sont

$$(1, 1), \quad (1, 2), \quad (2, 1), \quad (2, 2).$$

Instinctivement, nous pourrions penser procéder de la même manière que nous le faisons dans l'échantrillonnage univarié, c'est-à-dire sélectionner la paire  $(1, 1)$  à partir d'un premier échantillon aléatoire simple, la paire  $(1, 2)$  d'un deuxième échantillon, et ainsi de suite. Ces paires constituant notre échantrillon, nous pourrions alors les utiliser pour prendre les mesures nécessaires.

Le principal écueil lié à cette approche vient du fait que la recherche de représentants de chacune des quatre paires pourrait s'avérer laborieuse en pratique. En effet, dans le cas où les deux caractéristiques sont fortement corrélées positivement, les paires  $(1, 2)$  et  $(2, 1)$  pourraient s'avérer rares. De plus, le seul fait de juger le rang de deux caractéristiques en même temps rend cette méthode très fastidieuse.

Pour pallier ce problème, Al-Saleh & Zheng (2002) ont proposé une approche différente, qui prend appui dans le fait que l'on peut simuler une observation aléatoire d'une population bivariée en générant d'abord une observation de la loi marginale de la première variable, puis en générant une observation de la seconde à partir de sa loi conditionnelle sachant la valeur de la première variable.

Dans le cas particulier où  $k = 2$ , l'approche proposée par Al-Saleh & Zheng (2002) consiste à prélever huit échantillons aléatoires de taille 2 de la population. Ces huit échantillons sont ensuite divisés en quatre groupes de deux échantillons chacun. La paire (1, 1) sera alors extraite du premier groupe, la paire (1, 2) sera extraite du deuxième groupe, et ainsi et suite.

Pour obtenir la paire (1, 1), on procède d'abord à un tri de chacun des deux échantillons du premier groupe et on ne retient que la paire correspondant au minimum de la première caractéristique. Ceci nous laisse donc deux paires d'observations. L'élément (1, 1) de l'échantrillon bivarié est alors celle de ces deux paires pour laquelle on juge que la deuxième caractéristique est minimale.

La procédure à suivre pour l'obtention des paires (1, 2), (2, 1) et (2, 2) est semblable. Pour plus de clarté, voici, étape par étape, la procédure à suivre afin d'obtenir un échantrillon bivarié (ETB) :

1. Pour une taille  $k$ , nous avons besoin d'un échantillon de  $k^4$  unités de la population visée.
2. Nous divisons aléatoirement ces  $k^4$  unités en  $k^2$  groupes de  $k^2$  unités chacun. Chaque groupe a ainsi la forme d'une matrice carrée à  $k$  rangées et  $k$  colonnes.
3. Dans le premier groupe, nous identifions la valeur minimum de la première caractéristique pour chacune des  $k$  rangées.
4. Parmi chaque minimum ainsi obtenu, nous choisissons la paire ayant la valeur minimale de la seconde caractéristique. Nous obtenons ainsi la paire (1, 1), notre premier élément de l'échantrillon bivarié.
5. Ensuite, nous répétons les étapes 3 et 4 dans le deuxième groupe, à la différence que nous prenons la paire ayant le deuxième minimum pour la deuxième caractéristique. Ceci conduit au choix de l'élément (1, 2) de l'échantrillon.
6. Nous continuons ce processus jusqu'à ce que nous ayons extrait la paire  $(k, k)$  du dernier groupe.

En procédant de cette manière, nous pouvons obtenir un échantrillon bivarié de taille  $k^2$ . Comme pour l'échantrillonnage univarié, même si nous utilisons seulement  $k^2$  des  $k^4$  unités, toutes les unités apportent de l'information sur les  $k^2$  unités qui seront éventuellement mesurées.

### 3.1 Notation

Supposons que nous ayons un échantillon aléatoire de  $k^2$  groupes carrés de taille  $k^2$  chacun. Les éléments de chaque groupe sont supposés avoir été divisés aléatoirement en  $k$  ensembles de taille  $k$ .

Notons les valeurs des deux caractéristiques des éléments dans le  $n^e$  groupe par

$$\left\{ (X_{ij}^{(n)}, Y_{ij}^{(n)}), \quad i = 1, \dots, k, \quad j = 1, \dots, k, \right\}, \quad n = 1, \dots, k^2.$$

Ici,  $X_{ij}^{(n)}$  représente la valeur de la première caractéristique pour le  $j^e$  élément de la  $i^e$  rangée du  $n^e$  groupe. De même,  $Y_{ij}^{(n)}$  représente la valeur de la deuxième caractéristique pour le  $j^e$  élément de la  $i^e$  rangée dans le  $n^e$  groupe.

Enfin, introduisons les notations supplémentaires suivantes, pour tous  $i \in \{1, \dots, k\}$ ,  $j \in \{1, \dots, k\}$  et  $n \in \{(j-1)k+1, \dots, jk\}$  :

$X_{i(j)}^{(n)}$  : le  $j^e$  élément le plus petit parmi  $X_{i1}^{(n)}, \dots, X_{ik}^{(n)}$  ;

$Y_{i[j]}^{(n)}$  : la valeur correspondante de la variable  $Y$  ;

$Y_{(i)[j]}^{(n)}$  : le  $i^e$  élément le plus petit parmi les  $Y_{1[j]}^{(n)}, \dots, Y_{k[j]}^{(n)}$  ;

$X_{[i](j)}^{(n)}$ , la valeur correspondante de la variable  $X$ .

Ainsi, un échantillon bivarié est constitué de  $k^2$  paires

$$(X_{[i](j)}^{(n)}, Y_{(i)[j]}^{(n)}), \quad i = 1, \dots, k, \quad j = 1, \dots, k, \quad n = (j-1)k + i.$$

Il est important pour la suite de garder à l'esprit que ces dernières variables sont indépendantes, quoique non identiquement distribuées.

### 3.2 Résultats préliminaires

Nous présentons ici deux résultats qui jouent un rôle important dans la suite. La démonstration du premier est immédiate.

**Lemme 3.1** Soit  $(X_1, Y_1), \dots, (X_k, Y_k)$  un échantillon aléatoire de densité  $f_{X,Y}$ . Notons respectivement par  $f_X$  et par  $f_{Y|X}$  la densité de  $X$  et la densité conditionnelle de  $Y$  sachant  $X$ . Appelons  $Y_{[i]}$  la variable concomitante de la statistique d'ordre  $X_{(i)}$  pour tout  $i \in \{1, \dots, k\}$ . Alors la densité conditionnelle de  $Y_{[i]}$  sachant que  $X_{(i)} = x$  est donnée par

$$f_{Y_{[i]}|X_{(i)}}(y|X_{(i)} = x) = f_{Y|X}(y|x), \quad i \in \{1, \dots, k\}.$$

Notons par

$$f_{X_{i(j)}, Y_{i[j]}} \quad \text{et} \quad f_{X_{[i](j)}, Y_{(i)[j]}}(x, y)$$

les densités respectives des vecteurs aléatoires

$$(X_{i(j)}^{(n)}, Y_{i[j]}^{(n)}) \quad \text{et} \quad (X_{[i](j)}^{(n)}, Y_{(i)[j]}^{(n)}).$$

Les densités marginales et conditionnelles sont notées de manière similaire.

Remarquons que par construction, les paires  $(X_{i(j)}^{(n)}, Y_{i[j]}^{(n)})$  sont indépendantes et de même loi peu importe les valeurs de  $i$  et de  $n$ . D'après le Lemme 3.1, on sait en outre que

$$f_{X_{i(j)}, Y_{i[j]}}(x, y) = f_{X_{(j)}}(x) f_{Y|X}(y|x),$$

où  $f_{X_{(j)}}(x)$  est la densité de la  $j^{\text{e}}$  statistique d'ordre d'un échantillon aléatoire de taille  $k$  de la variable  $X$ . En sommant sur  $i$  et sur  $j$  de part et d'autre de la dernière équation, nous obtenons

$$\sum_{j=1}^k \sum_{i=1}^k f_{X_{i(j)}, Y_{i[j]}}(x, y) = \sum_{j=1}^k \sum_{i=1}^k f_{X_{(j)}}(x) f_{Y|X}(y|x) = k^2 f_{X,Y}(x, y).$$

Cette dernière égalité est vraie en vertu du Lemme 2.1. Par conséquent, la moyenne des densités associées aux paires  $(X_{i(j)}^{(n)}, Y_{i[j]}^{(n)})$  est donc égale à  $f_{X,Y}$ .

De façon semblable, nous avons

$$\begin{aligned} f_{X_{[i](j)}, Y_{(i)[j]}}(x, y) &= f_{Y_{(i)[j]}}(y) f_{X_{[i](j)}|Y_{(i)[j]}}(x|y) \\ &= f_{Y_{(i)[j]}}(y) f_{X_{(j)}|Y_{[j]}}(x|y) \end{aligned}$$

car  $X_{[i](j)}$  est concomitante de  $Y_{(i)[j]}$ . En appliquant le Lemme 3.1, on trouve alors

$$\begin{aligned} f_{X_{[i](j)}, Y_{(i)[j]}}(x, y) &= f_{Y_{(i)[j]}}(y) \frac{f_{X_{(j)}, Y_{[j]}}(x, y)}{f_{Y_{[j]}}(y)} \\ &= f_{Y_{(i)[j]}}(y) \frac{f_{X_{(j)}}(x) f_{Y|X}(y|x)}{f_{Y_{[j]}}(y)} \end{aligned}$$

puisque  $f_{Y_{[j]}}(y) = f_{Y_{[j]}}(y)$ .

En sommant sur  $i$  de part et d'autre de l'identité, on obtient

$$\begin{aligned} \sum_{i=1}^k f_{X_{[i](j)}, Y_{(i)[j]}}(x, y) &= f_{X_{(j)}}(x) f_{Y|X}(y|x) \sum_{i=1}^k \frac{f_{Y_{(i)[j]}}(y)}{f_{Y_{[j]}}(y)} \\ &= k f_{X_{(j)}}(x) f_{Y|X}(y|x), \end{aligned}$$

par application du Lemme 2.1.

En intégrant par rapport à  $y$  de chaque côté, on voit alors que

$$\sum_{i=1}^k f_{X_{[i](j)}}(x) = k f_{X_{(j)}}(x).$$

Si au lieu d'intégrer on somme plutôt par rapport à  $j$ , une nouvelle invocation du Lemme 2.1 permet de conclure que

$$\sum_{j=1}^k \sum_{i=1}^k f_{X_{[i](j)}, Y_{(i)[j]}}(x, y) = k f_{Y|X}(y|x) \sum_{j=1}^k f_{X_{(j)}}(x) = k^2 f_{X,Y}(x, y).$$

Ces faits sont énoncés formellement ci-dessous, de même qu'une de leurs conséquences immédiates.

**Lemme 3.2** *Soit  $(X, Y)$  une paire de variables aléatoires ayant pour densité  $f_{X,Y}$ . Pour tous  $i \in \{1, \dots, k\}$ ,  $j \in \{1, \dots, k\}$  et  $n = (j-1)k + i$ , soit  $f_{X_{[i](j)}, Y_{(i)[j]}}$  la densité de  $(X_{[i](j)}^{(n)}, Y_{(i)[j]}^{(n)})$ . Alors*

$$f_{X,Y}(x, y) = \frac{1}{k^2} \sum_{j=1}^k \sum_{i=1}^k f_{X_{[i](j)}, Y_{(i)[j]}}(x, y).$$

De plus,

$$f_{X_{(j)}}(x) = \frac{1}{k} \sum_{i=1}^k f_{X_{[i](j)}}(x) \quad \text{et} \quad f_X(x) = \frac{1}{k^2} \sum_{j=1}^k \sum_{i=1}^k f_{X_{[i](j)}}(x).$$

### Remarques

(i) Si  $X$  et  $Y$  sont indépendants, alors

$$(X_{[i](j)}^{(n)}, Y_{(i)[j]}^{(n)}) \quad \text{et} \quad (X_{i(j)}^{(n)}, Y_{(i)j}^{(n)})$$

sont de même loi et les valeurs de  $X_{i(j)}^{(n)}$  et de  $Y_{(i)j}^{(n)}$  constituent des échantrillons univariés à  $n$  cycles de taille  $k$  respectivement construits à partir des densités  $f_X$  et  $f_Y$ . Dans ce cas, l'échantrillonnage bivarié est équivalent à deux échantrillons univariés de taille  $k^2$  chacun.

(ii) Si  $X$  et  $Y$  sont parfaitement corrélés, de sorte que  $\rho(X, Y) = 1$ , alors

$$(X_{[i](j)}^{(n)}, Y_{(i)[j]}^{(n)}) \quad \text{et} \quad (X_{(i)(j)}^{(n)}, Y_{(i)(j)}^{(n)})$$

sont de même loi, de sorte que

$$\{X_{(i)(j)}^{(n)}, i = 1, \dots, k\} \quad \text{et} \quad \{Y_{(i)(j)}^{(n)}, j = 1, \dots, k\}$$

se réduisent à des échantrillons univariés à un cycle de taille  $k$  respectivement obtenus à partir de  $f_{X_{(i)}}$  et  $f_{Y_{(i)}}$ , où  $i \in \{1, \dots, k\}$  et  $n = (j - 1)k + i$ .

### 3.3 Les moyennes échantrillonnales

Dénotons par

$$\left\{ (X_{[i](j)}^{(n)}, Y_{(i)[j]}^{(n)}), \quad i = 1, \dots, k, \quad j = 1, \dots, k, \quad n = (j - 1)k + i \right\}$$

un échantrillon bivarié de taille  $k^2$  prélevé à partir d'une population bivariée ayant pour densité  $f_{X,Y}$ . On suppose comme précédemment que les rangs ne sont sujets à aucune erreur de jugement. Dénotons respectivement par  $\mu$  et par  $\sigma_X^2$  l'espérance et la variance de  $X$ . De la même manière, soient  $\theta$  et  $\sigma_Y^2$  l'espérance et la variance de  $Y$ . Enfin, soit  $\rho$  le coefficient de corrélation entre  $X$  et  $Y$ .

Supposons que l'on veuille estimer  $\mu$  et  $\theta$ . Appelons  $\hat{\mu}_{ETB}$  la moyenne échantrillonnale de

$$\{X_{[i](j)}^{(n)}, i = 1, \dots, k, j = 1, \dots, k\}$$

et  $\hat{\theta}_{ETB}$  la moyenne échantrillonnale de

$$\{Y_{(i)[j]}^{(n)}, i = 1, \dots, k, j = 1, \dots, k\}.$$

Posons aussi

$$\mu_{[i](j)} = E(X_{[i](j)}^{(n)}), \quad \theta_{(i)[j]} = E(Y_{(i)[j]}^{(n)})$$

et

$$\sigma_{[i](j)}^2 = \text{var}(X_{[i](j)}^{(n)}).$$

En vertu du Lemme 3.2, nous avons

$$\begin{aligned} \frac{1}{k^2} \sum_{j=1}^k \sum_{i=1}^k \mathbb{E}(X_{[i](j)}^{(n)}) &= \frac{1}{k^2} \sum_{j=1}^k \sum_{i=1}^k \mu_{[i](j)} \\ &= \frac{1}{k^2} \sum_{j=1}^k \sum_{i=1}^k \int_{-\infty}^{\infty} x f_{X_{[i](j)}}(x) dx \\ &= \int_{-\infty}^{\infty} x f_X(x) dx = \mu. \end{aligned}$$

De la même manière, on vérifie que

$$\frac{1}{k^2} \sum_{j=1}^k \sum_{i=1}^k \mathbb{E}(Y_{(i)[j]}^{(n)}) = \frac{1}{k^2} \sum_{j=1}^k \sum_{i=1}^k \theta_{(i)[j]}^{(n)} = \theta.$$

Ainsi,  $\hat{\mu}_{ETB}$  et  $\hat{\theta}_{ETB}$  sont respectivement des estimateurs sans biais de  $\mu$  et de  $\theta$ , comme nous l'avons déjà montré au chapitre 2 dans le cas de l'échantrillonnage univarié.

### 3.4 La variance des moyennes échantrillonnales

Deux expressions différentes seront données ci-dessous pour le calcul de la variance de l'estimateur  $\hat{\mu}_{ETB}$  de  $\mu = \mathbb{E}(X)$ . Des formules semblables peuvent être obtenues de façon semblable pour la variance de la moyenne échantrillonnale  $\hat{\theta}_{ETB}$ .

Notons d'abord que par une application du Lemme 3.2, on a

$$\begin{aligned} \sigma_X^2 &= \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx = \frac{1}{k^2} \sum_{i=1}^k \sum_{j=1}^k \int_{-\infty}^{\infty} (x - \mu)^2 f_{X_{[i](j)}}(x) dx \\ &= \frac{1}{k^2} \sum_{i=1}^k \sum_{j=1}^k \int_{-\infty}^{\infty} (x - \mu_{[i](j)} + \mu_{[i](j)} - \mu)^2 f_{X_{[i](j)}}(x) dx \\ &= \frac{1}{k^2} \sum_{i=1}^k \sum_{j=1}^k \int_{-\infty}^{\infty} (x - \mu_{[i](j)})^2 f_{X_{[i](j)}}(x) dx \\ &\quad + \frac{1}{k^2} \sum_{i=1}^k \sum_{j=1}^k \int_{-\infty}^{\infty} (\mu_{[i](j)} - \mu)^2 f_{X_{[i](j)}}(x) dx, \end{aligned}$$

puisque le terme croisé s'annule du fait que par définition,

$$\int (x - \mu_{[i](j)}) f_{[i](j)}(x) dx = 0$$

pour tous  $i, j \in \{1, \dots, k\}$ . Il s'ensuit que

$$\sigma_X^2 = \frac{1}{k^2} \sum_{j=1}^k \sum_{i=1}^k \sigma_{[i](j)}^2 + \frac{1}{k^2} \sum_{j=1}^k \sum_{i=1}^k (\mu_{[i](j)} - \mu)^2$$

et que par conséquent,

$$\begin{aligned} \text{var}(\hat{\mu}_{ETB}) &= \text{var}\left(\frac{1}{k^2} \sum_{i=1}^k \sum_{j=1}^k X_{[i](j)}^{(n)}\right) = \frac{1}{k^4} \sum_{i=1}^k \sum_{j=1}^k \sigma_{[i](j)}^2 \\ &= \frac{\sigma_X^2}{k^2} - \frac{1}{k^4} \sum_{j=1}^k \sum_{i=1}^k (\mu_{[i](j)} - \mu)^2. \end{aligned} \quad (3.1)$$

Une deuxième expression pour la variance de  $\hat{\mu}_{ETB}$  fait intervenir l'espérance et la variance des variables  $X_{i(j)}$ , notées

$$\mu_j = \mathbb{E}(X_{i(j)}), \quad \sigma_j^2 = \text{var}(X_{i(j)})$$

pour tous  $i, j \in \{1, \dots, k\}$ . Cette formule, semblable à celle déjà donnée dans le cas univarié, s'obtient en deux temps.

Notons d'abord que

$$\begin{aligned} \text{var}(\hat{\mu}_{ETB}) &= \frac{1}{k^4} \sum_{i=1}^k \sum_{j=1}^k \text{var}(X_{[i](j)}^{(n)}) \\ &= \frac{1}{k^4} \sum_{i=1}^k \sum_{j=1}^k \int (x - \mu_{[i](j)})^2 f_{X_{[i](j)}}(x) dx \\ &= \frac{1}{k^4} \sum_{i=1}^k \sum_{j=1}^k \int (x - \mu_j + \mu_j - \mu_{[i](j)})^2 f_{X_{[i](j)}}(x) dx \\ &= \frac{1}{k^4} \sum_{i=1}^k \sum_{j=1}^k \int (x - \mu_j)^2 f_{X_{[i](j)}}(x) dx \\ &\quad + \frac{2}{k^4} \sum_{i=1}^k \sum_{j=1}^k (\mu_j - \mu_{[i](j)}) \int (x - \mu_j) f_{X_{[i](j)}}(x) dx \\ &\quad + \frac{1}{k^4} \sum_{i=1}^k \sum_{j=1}^k (\mu_j - \mu_{[i](j)})^2 \int f_{X_{[i](j)}}(x) dx \end{aligned}$$

et donc que

$$\begin{aligned} \text{var}(\hat{\mu}_{ETB}) &= \frac{1}{k^4} \sum_{i=1}^k \sum_{j=1}^k \int (x - \mu_j)^2 f_{X_{[i](j)}}(x) dx \\ &\quad - \frac{1}{k^4} \sum_{i=1}^k \sum_{j=1}^k (\mu_j - \mu_{[i](j)})^2. \end{aligned} \quad (3.2)$$

Or, d'après le Lemme 3.2, on sait que

$$\sum_{i=1}^k f_{X_{[i](j)}} = k f_{X_{(j)}}.$$

Par substitution dans le premier terme du membre de droite de la formule (3.2), on trouve alors

$$\begin{aligned} \text{var}(\hat{\mu}_{ETB}) &= \frac{1}{k^3} \sum_{j=1}^k \int (x - \mu_j)^2 f_{X_{(j)}}(x) dx - \frac{1}{k^4} \sum_{i=1}^k \sum_{j=1}^k (\mu_{[i](j)} - \mu_j)^2 \\ &= \frac{1}{k^3} \sum_{j=1}^k \sigma_j^2 - \frac{1}{k^4} \sum_{i=1}^k \sum_{j=1}^k (\mu_{[i](j)} - \mu_j)^2. \end{aligned} \quad (3.3)$$

### 3.5 Comparaisons élémentaires

Pour juger de l'efficacité de l'estimation de  $\mu = E(X)$  par échantrillonnage bivarié, on peut comparer la variance de  $\hat{\mu}_{ETB}$  à deux autres estimations fondées sur des efforts de mesure semblables, à savoir :

- a) l'estimateur  $\hat{\mu}_{EAS}$  fondé sur un échantillon aléatoire simple de taille  $k^2$  ;
- b) l'estimateur  $\hat{\mu}_{ETU}$  fondé sur un échantrillon à  $k$  cycles de taille  $k$  chacun.

Les variances de ces deux compétiteurs sont les suivantes :

$$\begin{aligned} \text{var}(\hat{\mu}_{EAS}) &= \sigma_X^2 / k^2, \\ \text{var}(\hat{\mu}_{ETU}) &= \frac{1}{k^4} \sum_{i=1}^k \sum_{j=1}^k \text{var}(X_{i(j)}) = \frac{1}{k^3} \sum_{j=1}^k \sigma_j^2. \end{aligned}$$

Au vu de l'équation (3.1), l'efficacité relative de  $\hat{\mu}_{ETB}$  par rapport à  $\hat{\mu}_{EAS}$  est donc donnée par

$$\text{eff}(\hat{\mu}_{ETB}|\hat{\mu}_{EAS}) = \frac{\text{var}(\hat{\mu}_{EAS})}{\text{var}(\hat{\mu}_{ETB})} = \frac{\sigma_X^2/k^2}{\frac{\sigma_X^2}{k^2} - \frac{1}{k^4} \sum_{j=1}^k \sum_{i=1}^k (\mu_{[i](j)} - \mu)^2}. \quad (3.4)$$

Le rapport des variances étant supérieur à 1, on voit qu'en général,  $\hat{\mu}_{ETB}$  est plus efficace que  $\hat{\mu}_{EAS}$  pour un même effort d'échantillonnage, soit  $k^2$ . Il en va évidemment de même pour  $\hat{\theta}_{ETB}$  par rapport à  $\hat{\theta}_{EAS}$ .

En faisant appel à la formule (3.3), on voit en outre que

$$\text{eff}(\hat{\mu}_{ETB}|\hat{\mu}_{ETU}) = \frac{\text{var}(\hat{\mu}_{ETU})}{\text{var}(\hat{\mu}_{ETB})} = \frac{\frac{1}{k^3} \sum_{j=1}^k \sigma_j^2}{\frac{1}{k^3} \sum_{j=1}^k \sigma_j^2 - \frac{1}{k^4} \sum_{i=1}^k \sum_{j=1}^k (\mu_{[i](j)} - \mu_j)^2}. \quad (3.5)$$

Une fois de plus, le rapport est donc supérieur à 1, ce qui montre qu'une estimation de  $\mu$  fondée sur un échantillon bivarié est préférable à une estimation fondée sur un échantillon univarié à  $k$  cycles.

### 3.6 Comparaisons plus fines

Nous allons maintenant affiner nos comparaisons d'efficacité entre les estimateurs  $\hat{\mu}_{ETB}$  et  $\hat{\mu}_{ETU}$  en considérant trois scénarios de dépendance particuliers. Avant de procéder, remarquons qu'en vertu de l'identité (2.2), on a

$$\text{var}(\hat{\mu}_{ETU}) = \frac{1}{k^3} \sum_{j=1}^k \sigma_j^2 = \frac{\sigma_X^2}{k^2} - \frac{1}{k^3} \sum_{j=1}^k (\mu_j - \mu)^2.$$

Étant donné l'identité (3.3), on a donc aussi

$$\text{var}(\hat{\mu}_{ETB}) = \frac{\sigma_X^2}{k^2} - \frac{1}{k^3} \sum_{j=1}^k (\mu_j - \mu)^2 - \frac{1}{k^4} \sum_{i=1}^k \sum_{j=1}^k (\mu_{[i](j)} - \mu_j)^2. \quad (3.6)$$

### 3.6.1 Le cas d'indépendance

Si les variables  $X$  et  $Y$  sont indépendantes, alors  $\mu_{[i](j)} = \mu_j$  pour tout  $i \in \{1, \dots, k\}$ . De toute évidence, le troisième membre du terme de droite de l'équation (3.6) est alors identiquement égal à zéro.

Par conséquent, on a

$$\text{eff}_0(\hat{\mu}_{ETB} | \hat{\mu}_{ETU}) = 1,$$

ce qui signifie que l'échantrillonnage bivarié est aussi efficace que l'échantrillonnage univarié lorsque  $X$  et  $Y$  sont indépendants. Sous ces conditions, l'échantrillonnage bivarié peut tout de même s'avérer plus avantageux que l'approche univariée dans certaines circonstances. Cette façon de procéder pourrait être plus économique, par exemple, lorsque la prise de mesure revêt un caractère destructeur.

### 3.6.2 Le cas de dépendance parfaite

Si les variables  $X$  et  $Y$  sont parfaitement corrélées positivement ( $\rho = 1$ ), alors

$$(X_{[i](j)}^{(n)}, Y_{(i)[j]}^{(n)}) \quad \text{et} \quad (X_{(i)(j)}^{(n)}, Y_{(i)(j)}^{(n)})$$

sont de même loi, de sorte que  $\mu_{[i](j)} \equiv \mu_{(i)(j)}$  pour tous  $i, j \in \{1, \dots, k\}$ . Il découle alors de l'équation (3.5) que

$$\text{eff}_1(\hat{\mu}_{ETB} | \hat{\mu}_{ETU}) = \frac{\frac{1}{k^3} \sum_{j=1}^k \sigma_j^2}{\frac{1}{k^3} \sum_{j=1}^k \sigma_j^2 - \frac{1}{k^4} \sum_{i=1}^k \sum_{j=1}^k (\mu_{(i)(j)} - \mu_j)^2} \quad (3.7)$$

est supérieur à 1.

### 3.6.3 Le cas de dépendance linéaire

À l'instar de Stokes (1977), qui fait cette hypothèse dans le cas univarié, supposons de façon plus générale que la régression de  $X$  en  $Y$  soit linéaire, comme ce serait le cas par exemple si la paire  $(X, Y)$  obéissait à une loi normale ou à une loi de Pareto

bivariée. Cette relation de linéarité n'étant pas affectée par des opérations de tri et de classement, on a alors

$$\mu_{[i](j)} = \mu_j + \rho(\mu_{(i)(j)} - \mu_j)$$

pour un certain coefficient de corrélation  $\rho \in [-1, 1]$ . Noter qu'en particulier on retrouve les relations  $\mu_{[i](j)} = \mu_j$  et  $\mu_{[i](j)} = \mu_{(i)(j)}$  correspondant respectivement à l'indépendance ( $\rho = 0$ ) et à la dépendance positive parfaite ( $\rho = 1$ ).

Sous ce modèle particulier de dépendance, l'équation (3.5) devient

$$\text{eff}_\rho(\hat{\mu}_{ETB}|\hat{\mu}_{ETU}) = \frac{\frac{1}{k^3} \sum_{j=1}^k \sigma_j^2}{\frac{1}{k^3} \sum_{j=1}^k \sigma_j^2 - \frac{\rho^2}{k^4} \sum_{i=1}^k \sum_{j=1}^k (\mu_{(i)(j)} - \mu_j)^2}.$$

Or à la lumière de l'identité (3.7), on sait que

$$\frac{\frac{1}{k^4} \sum_{i=1}^k \sum_{j=1}^k (\mu_{(i)(j)} - \mu_j)^2}{\frac{1}{k^3} \sum_{j=1}^k \sigma_j^2} = 1 - \frac{1}{\text{eff}_1(\hat{\mu}_{ETB}|\hat{\mu}_{ETU})}.$$

En exploitant ce fait, il est donc possible d'exprimer la formule d'efficacité générale comme suit, en fonction du paramètre  $\rho$  et de l'efficacité en  $\rho = 1$  :

$$\Phi(\rho) = \text{eff}_\rho(\hat{\mu}_{ETB}|\hat{\mu}_{ETU}) = \frac{\text{eff}_1(\hat{\mu}_{ETB}|\hat{\mu}_{ETU})}{\text{eff}_1(\hat{\mu}_{ETB}|\hat{\mu}_{ETU}) - \rho^2 \{\text{eff}_1(\hat{\mu}_{ETB}|\hat{\mu}_{ETU}) - 1\}}. \quad (3.8)$$

De plus, puisque la même équation (3.7) entraîne que

$$\text{eff}_1(\hat{\mu}_{ETB}|\hat{\mu}_{ETU}) \geq 1,$$

on déduit facilement de (3.8) que  $\Phi(\rho)$  est une fonction croissante de  $|\rho|$  dont le minimum est atteint en  $\Phi(0) = 1$ , en accord avec le résultat déjà énoncé dans la sous-section 3.6.1. Un exemple de calcul explicite de la fonction  $\Phi(\rho)$  sera présenté ci-dessous.

### 3.7 Cas de la loi normale bivariée

Afin d'illustrer l'efficacité de l'estimation par échantrillonnage bivarié, nous nous penchons dans cette section sur le cas où la population est normale bivariée, c'est-à-

dire où

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_2 \left[ \begin{pmatrix} \mu \\ \theta \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \right].$$

Comme précédemment, concentrons-nous sur l'estimation de  $\mu$  et comparons la performance de l'estimateur  $\hat{\mu}_{ETB}$  à celle de l'estimateur  $\hat{\mu}_{EAS}$  correspondant à un échantillon aléatoire de taille  $k^2$ .

Partons du fait que

$$\begin{aligned} \text{eff}_\rho(\hat{\mu}_{ETB}|\hat{\mu}_{EAS}) &= \text{eff}_\rho(\hat{\mu}_{ETB}|\hat{\mu}_{ETU}) \times \text{eff}(\hat{\mu}_{ETU}|\hat{\mu}_{EAS}) \\ &= \Phi(\rho) \times \text{eff}(\hat{\mu}_{ETU}|\hat{\mu}_{EAS}). \end{aligned}$$

Puisque le modèle normal répond aux hypothèses ayant conduit à l'identité (3.8), on sait que  $1/\Phi(\rho)$  est une fonction quadratique de  $\rho$ , dont le comportement dépend exclusivement de  $\Phi(1)$ . L'efficacité relative de l'estimateur  $\hat{\mu}_{ETB}$  ne dépend donc que de cette constante et de  $\text{eff}(\hat{\mu}_{ETU}|\hat{\mu}_{EAS})$ .

Or, compte tenu de la formule (2.2), on sait déjà que

$$\text{eff}(\hat{\mu}_{ETU}|\hat{\mu}_{EAS}) = \frac{\sigma_X^2/k}{\frac{\sigma_X^2}{k} - \frac{1}{k^2} \sum_{j=1}^k (\mu_j - \mu)^2} = \frac{1}{1 - \frac{1}{k} \sum_{j=1}^k \xi_j^2},$$

où pour tout  $j \in \{1, \dots, k\}$ ,

$$\xi_{(j)} = \text{E} \left( \frac{X_{(j)} - \mu}{\sigma_X} \right) = \frac{\mu_j - \mu}{\sigma_X}$$

est l'espérance de la  $j^{\text{e}}$  statistique d'ordre d'une variable normale centrée réduite, aussi appelée  $j^{\text{e}}$  *rankit*. Ces constantes, indépendantes de  $\mu$  et de  $\sigma_X$ , sont tabulées dans le livre de David (1981) et, de toute façon, très faciles à déterminer par intégration numérique. Il ne reste donc qu'à déterminer la valeur de  $\text{eff}_1(\hat{\mu}_{ETB}|\hat{\mu}_{ETU})$  dans le cas normal.

Avant de ce faire, notons qu'en général,

$$\begin{aligned} \sum_{i=1}^k \sum_{j=1}^k (\mu_{[i](j)} - \mu_j)^2 &= \sum_{i=1}^k \sum_{j=1}^k (\mu_{[i](j)} - \mu + \mu - \mu_j)^2 \\ &= \sum_{i=1}^k \sum_{j=1}^k \{(\mu_{[i](j)} - \mu)^2 + (\mu - \mu_j)^2 + 2(\mu_{[i](j)} - \mu)(\mu - \mu_j)\}, \\ &= \sum_{i=1}^k \sum_{j=1}^k (\mu_{[i](j)} - \mu)^2 - k \sum_{j=1}^k (\mu - \mu_j)^2, \end{aligned}$$

puisque en vertu du Lemme 3.2, on a

$$\sum_{i=1}^k \mu_{[i](j)} = \sum_{i=1}^k \int x f_{X_{[i](j)}}(x) dx = k \int x f_{X_{(j)}}(x) dx = k\mu_j$$

et donc pour tout  $j \in \{1, \dots, k\}$ ,

$$\sum_{i=1}^k (\mu_{[i](j)} - \mu) = k(\mu_j - \mu).$$

Par suite, une formule équivalente à (3.6) est donnée par

$$\begin{aligned} \text{var}(\hat{\mu}_{ETB}) &= \frac{\sigma_X^2}{k^2} - \frac{1}{k^3} \sum_{j=1}^k (\mu_j - \mu)^2 - \frac{1}{k^4} \sum_{i=1}^k \sum_{j=1}^k (\mu_{[i](j)} - \mu_j)^2 \\ &= \frac{\sigma_X^2}{k^2} - \frac{1}{k^4} \sum_{i=1}^k \sum_{j=1}^k (\mu_{[i](j)} - \mu)^2, \end{aligned}$$

alors que

$$\text{var}(\hat{\mu}_{ETU}) = \frac{1}{k^3} \sum_{j=1}^k \sigma_j^2 = \frac{\sigma_X^2}{k^2} - \frac{1}{k^3} \sum_{j=1}^k (\mu_j - \mu)^2.$$

On conclut donc que

$$\begin{aligned} \text{eff}_1(\hat{\mu}_{ETB} | \hat{\mu}_{ETU}) &= \frac{\frac{\sigma_X^2}{k^2} - \frac{1}{k^3} \sum_{j=1}^k (\mu_j - \mu)^2}{\frac{\sigma_X^2}{k^2} - \frac{1}{k^4} \sum_{i=1}^k \sum_{j=1}^k (\mu_{[i](j)} - \mu)^2} \\ &= \frac{1 - \frac{1}{k} \sum_{j=1}^k \xi_j^2}{1 - \frac{1}{k^2} \sum_{i=1}^k \sum_{j=1}^k \xi_{[i](j)}^2}, \end{aligned}$$

où par définition,

$$\xi_{[i](j)} = \frac{\mu_{[i](j)} - \mu}{\sigma_X}, \quad i, j \in \{1, \dots, k\}.$$

Ces constantes sont, elles aussi, tabulées dans le manuel de David (1981). Tout est donc en place pour donner un exemple de calcul.

### 3.8 Illustration

Pour conclure ce chapitre, nous présentons ici le calcul de

$$\text{eff}_\rho(\hat{\mu}_{ETB}|\hat{\mu}_{EAS}) = \Phi(\rho) \times \text{eff}(\hat{\mu}_{ETU}|\hat{\mu}_{EAS})$$

dans le cadre de la loi normale bivariée, dans le cas spécial où  $k = 2$ . D'après David (1981, pp. 60–62), on a

$$\xi_{(2)} = -\xi_{(1)} = 0.5642,$$

de sorte que

$$\text{eff}(\hat{\mu}_{ETU}|\hat{\mu}_{EAS}) = \frac{1}{1 - 0.5642^2} = 1.467.$$

De plus, on trouve

$$\xi_{(2)(2)} = -\xi_{(1)(1)} = 1.0294, \quad \xi_{(2)(1)} = -\xi_{(1)(2)} = 0.0990.$$

Il s'ensuit que

$$\text{eff}_1(\hat{\mu}_{ETB}|\hat{\mu}_{ETU}) = 1.465$$

et donc que

$$\Phi(\rho) = \frac{1.465}{1.465 - \rho^2(1.465 - 1)}.$$

Par conséquent, on trouve

$$\text{eff}_\rho(\hat{\mu}_{ETB}|\hat{\mu}_{EAS}) = \frac{1.467}{1 - 0.317\rho^2}.$$

Cette fonction, qui vaut au maximum 2.148 en  $\rho = 1$  est tracée à la figure 3.1.

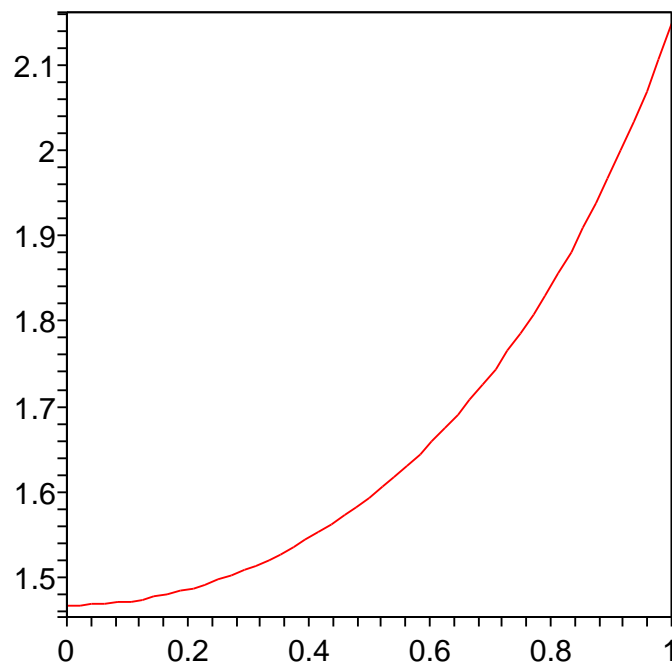


FIG. 3.1 – Graphe de  $\text{eff}_\rho(\hat{\mu}_{ETB} | \hat{\mu}_{EAS})$  en fonction de  $\rho$  dans le cas normal bivarié, lorsque  $k = 2$

# Chapitre 4

## Conclusion

Cet essai se voulait une introduction aux méthodes d'échantillonnage univarié et bivarié, ainsi qu'à certaines questions portant sur l'inférence pour ce type de données.

Au chapitre 2, nous avons traité de l'échantillonnage univarié. Nous avons vu ce qu'est un échantillon et comment s'y prendre afin d'en obtenir un à l'aide des données que nous avons en main. De plus, nous avons pu vérifier que l'estimateur de la moyenne échantillonnale est plus précis que celui de la moyenne échantillonnale. Nous avons ensuite montré qu'il existe un test de Mann–Whitney qui peut s'appliquer à l'échantillonnage.

Le chapitre 3 a été consacré à l'échantillonnage bivarié. Nous y avons expliqué la méthode à suivre pour construire ce type d'échantillon. Différentes efficacités relatives ont été présentées afin de faire ressortir les avantages de procéder à une inférence à partir de telles données.

Les notions et résultats exposés ici ne peuvent prétendre à l'exhaustivité. Le lecteur intéressé à prolonger sa réflexion sur l'échantillonnage est invité à consulter les articles et ouvrages répertoriés dans la bibliographie annotée publiée par Patil, Sinha & Taillie (1999).

En terminant, mentionnons qu'il n'existe pas, à notre connaissance, de travaux portant sur l'échantillonnage multivarié. Il y a donc là une possibilité de développement intéressante, mais certes non élémentaire, compte tenu de la complexité des modalités d'échantillonnage bivarié!

# Bibliographie

- [1] Al-Saleh, M. F. & Zheng, G. (2002). Estimation of bivariate characteristics using ranked set sampling. *Australian and New Zealand Journal of Statistics*, 44, 221–232.
- [2] Bohn, L. L. & Wolfe, D. A. (1992). Nonparametric two-sample procedures for ranked-set samples data. *Journal of the American Statistical Association*, 87, 552–561.
- [3] David, H. A. (1981). *Order Statistics*, deuxième édition. Wiley, New York.
- [4] Dell, T. R. and Clutter, J. L. (1972). Ranked set sampling theory with order statistics background. *Biometrics*, 28, 545–555.
- [5] McIntyre, G. A. (1952). A method for unbiased selective sampling, using ranked sets. *Australian Journal of Agricultural Research*, 3, 385–390.
- [6] Patil G. P., Sinha A. K. & Taillie C. (1999). Ranked set sampling : A bibliography. *Environmental and Ecological Statistics*, 6, 91–98.
- [7] Stokes, S. L. (1977). Ranked set sampling with concomitant variables. *Communications in Statistics—Theory and Methods*, 6, 1207–1211.
- [8] Stokes, S. L. & Sager, T. W. (1988). Characterization of a ranked-set sample with application to estimating distribution functions. *Journal of the American Statistical Association*, 83, 374–381.
- [9] Takahasi, K. & Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics*, 20, 1–31.
- [10] Wolfe, D. A. (2004). Ranked set sampling : An approach to more efficient data collection. *Statistical Science*, 19, 636–643.