Chapitre 5
Autres modes d'échantillonnage

Jusqu'ici, nous nous sommes concentrés sur l'échantillonnage aléatoire simple, c'est-à-dire, nous avons supposé que le tirage est fait de telle sorte que chaque choix possible de n unités a même probabilité de constituer l'échantillon. Ce mode d'échantillonnage n'est pas toujours adéquat, et il est souvent difficile à appliquer en pratique.

Par exemple, il arrive qu'en plus d'estimer les paramètres d'une population, on veuille estimer ceux de certaines *sous*-populations. Pour cela, il faut que l'échantillon contienne suffisamment d'observations provenant des sous-populations. Par exemple, un échantillon aléatoire simple conçu pour estimer le revenu moyen des ménages de l'île de Montréal, ne contiendra pas toujours assez de ménages d'Outremont pour estimer la moyenne à Outremont.

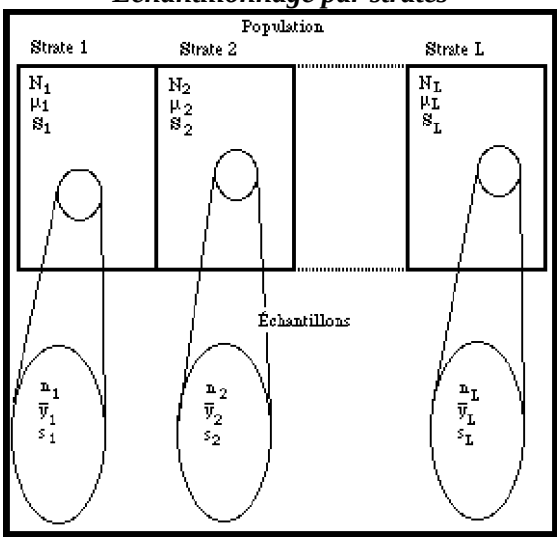
Une autre difficulté avec un échantillonnage aléatoire simple est que pour l'effectuer, il faut disposer d'une liste des unités de la population. Ce n'est souvent pas le cas. Il est rare, par exemple, que vous puissiez dresser une liste des habitants d'un quartier. Mais il se peut que vous puissiez dresser une liste d'adresses ou de numéros de téléphone. Vous allez donc faire un choix d'adresses ou de numéros de téléphone — de ménages, essentiellement — pour *ensuite* choisir des adultes dans les ménages. Il est permis de procéder de cette façon à condition d'ajuster les méthodes d'estimation en conséquence: un estimateur valable dans un échantillon aléatoire simple ne l'est pas nécessairement dans un autre type d'échantillon.

Nous allons décrire ici deux nouveaux modes d'échantillonnage: l'échantillonnage ***stratifié***, et l'***échantillonnage par grappes***. Nous dirons quelques mots également sur l'***échantillonnage systématique***.

5.1 Échantillonnage stratifié

Lorsqu'une population est naturellement partitionnée en sous-populations, il est normal de faire un sondage dans chacune des sous-populations pour ensuite réunir les résultats de façon à estimer les paramètres de la population entière. Par exemple, un sondage auprès des étudiants d'une université peut procéder par faculté: on tire un échantillon dans chaque faculté, puis on réunit les résultats obtenus dans les facultés pour obtenir des estimations globales pour la population entière. Ce mode d'échantillonnage, appelé *échantillonnage par strates*, n'est pas seulement plus aisé à l'exécution: il permet en plus d'améliorer la précision, c'est-à-dire, de diminuer l'écart-type des estimateurs. Mais les estimateurs habituels devront changer, ainsi que les formules pour les écarts-types.

Figure 5.1.1
Échantillonnage par strates



Commençons par illustrer le procédé. Le tableau A.03 présente une population de 210 paroisses québécoises. Cette population a été stratifiée selon la taille (mesurée par le nombre d'habitants en 1996). On a tiré un échantillon aléatoire simple dans chacune des trois strates. Les échantillons, ainsi que certains calculs sont présentés au tableau 5.1.1.

Tableau 5.1.1
Échantillon stratifié de la population de paroisses (*tableau A.03*)

	Strate (*h*)			
	h = 1	*h* = 2	*h* = 3	*h* = 4
	7 10 8 2 7 6 2 6	14 11 14 7 17 19 9	23 36 2 9 24	24 110 17 47 32
Taille de l'échantillon (n_h)	8	7	5	5
Taille de la strate (N_h)	105	63	21	21
Moyenne de l'échantillon ($\bar{y}_h$)	6	13	18,8	46
Écart-type de l'échantillon (s_h)	2,7775	4,2817	13,4052	37,4767
Écart-type de l'estimateur : $\hat{\sigma}_{\bar{y}_h} = \sqrt{1-f_h}\ s_h/\sqrt{n_h}$, $f_h = n_h/N_h$	0,9438	1,5258	5,2329	14,6294
Taille relative de la strate ($W_h = N_h/N$)	0,5	0,3	0,1	0,1

Étant donné que nous avons tiré, dans chaque strate, un échantillon aléatoire simple, nous avons là quatre fois un problème familier. Les quatre strates constituent quatre populations distinctes, de

moyennes μ_1, μ_2, μ_3, et μ_4, disons. Chacun des quatre échantillons fournit une estimation de la moyenne de la strate :

$$\bar{y}_1 = 6 \quad ; \quad \bar{y}_2 = 13 \quad ; \quad \bar{y}_3 = 18{,}8 \quad ; \quad \bar{y}_4 = 46$$

À l'aide des écarts-types corrigés

$$s_1 = 2{,}7775 \quad ; \quad s_2 = 4{,}2817 \quad ; \quad s_3 = 13{,}4052 \quad ; \quad s_4 = 37{,}4767$$

on peut estimer les écarts-types des estimateurs

$$\hat{\sigma}_{\bar{y}_1} = 0{,}9438 \quad ; \quad \hat{\sigma}_{\bar{y}_2} = 1{,}5258 \quad ; \quad \hat{\sigma}_{\bar{y}_3} = 5{,}2329 \quad ; \quad \hat{\sigma}_{\bar{y}_4} = 14{,}6294$$

S'il ne s'agissait que d'estimer les moyennes des strates μ_1, μ_2, μ_3, et μ_4, il n'y aurait rien de nouveau dans ce problème. Mais lorsqu'on procède à un échantillonnage par strates, c'est généralement pour estimer un paramètre relatif à la *population entière*. En particulier μ, la moyenne globale de la population. Comment combiner les estimations $\bar{y}_1, \bar{y}_2, \bar{y}_3, \bar{y}_4$ des moyennes μ_1, μ_2, μ_3, et μ_4 pour obtenir une estimation de μ. Il suffit, pour découvrir la réponse, d'exprimer μ en termes de μ_1, μ_2, μ_3, et μ_4. Il est clair que μ est une moyenne des μ_h; mais ce n'est pas la somme des μ_h divisée par 4 : il faut *pondérer* les μ_h. Pondérer par quoi? Par les tailles des strates *dans la population* — les N_h. Il est commode d'exprimer les taille des strates en tailles relatives, c'est-à-dire, de les diviser par la taille de la population. On obtient ainsi les W_h :

$$W_1 = \frac{N_1}{N} = \frac{105}{210} = 0{,}5 \; ; \; W_2 = \frac{N_2}{N} = \frac{63}{210} = 0{,}3 \; ; \; W_3 = \frac{N_3}{N} = \frac{21}{210} = 0{,}1 \; ; \; W_4 = \frac{N_4}{N} = \frac{21}{210} = 0{,}1$$

La moyenne de la population μ est une moyenne des moyennes des strates μ_h, pondérée par les W_h :

$$\mu = W_1\mu_1 + W_2\mu_2 + W_3\mu_3 + W_4\mu_4$$[1]

Comment estimer μ? Naturellement en remplaçant chaque μ_h ci-dessus par $\bar{y}_h$. On obtient ainsi l'estimateur stratifié, dénoté par $\bar{y}_{st}$:

$$\bar{y}_{st} = W_1\bar{y}_1 + W_2\bar{y}_2 + W_3\bar{y}_3 + W_4\bar{y}_4$$

$$= (0{,}5)(6) + (0{,}3)(13) + (0{,}1)(18{,}8) + (0{,}1)(46) = 13{,}38$$

La variance de $\bar{y}_{st}$ est une combinaison des variances des $\bar{y}_h$, mais ces variances sont multipliées par les W_h^2 plutôt que par les W_h :

[1] Ce calcul est analogue au calcul d'une moyenne à partir d'une distribution, qui consiste à multiplier chaque valeur distincte de la variable par sa fréquence. Les W_h sont essentiellement des fréquences.

$$\sigma^2_{\bar{y}_{st}} = W_1^2\sigma^2_{\bar{y}_1} + W_2^2\sigma^2_{\bar{y}_2} + W_3^2\sigma^2_{\bar{y}_3} + W_4^2\sigma^2_{\bar{y}_4}$$

Et bien sûr, nous estimons cette variance par

$$\hat{\sigma}^2_{\bar{y}_{st}} = W_1^2\hat{\sigma}^2_{\bar{y}_1} + W_2^2\hat{\sigma}^2_{\bar{y}_2} + W_3^2\hat{\sigma}^2_{\bar{y}_3} + W_4^2\hat{\sigma}^2_{\bar{y}_4}$$

Les variances estimées sont :

$$\hat{\sigma}^2_{\bar{y}_1} = \left(1 - \frac{8}{105}\right)\frac{2,7775^2}{8} = 0,6886 \qquad \hat{\sigma}^2_{\bar{y}_2} = \left(1 - \frac{7}{63}\right)\frac{4,2817^2}{7} = 2,3280$$

$$\hat{\sigma}^2_{\bar{y}_3} = \left(1 - \frac{5}{21}\right)\frac{13,4052^2}{5} = 27,3829 \qquad \hat{\sigma}^2_{\bar{y}_4} = \left(1 - \frac{5}{21}\right)\frac{36,4767^2}{5} = 214,0190$$

Nous avons donc

$$\hat{\sigma}^2_{\bar{y}_{st}} = (0,5^2)(0,6886) + (0,3^2)(2,3280) + (0,1^2)(27,3829) + (0,1^2)(214,0190) = 2,7957$$

et $\hat{\sigma}_{\bar{y}_{st}} = \sqrt{2,7957} = 1,6720$.

On détermine l'intervalle de confiance comme d'habitude :

$$\bar{y}_{st} - 2\,\hat{\sigma}_{\bar{y}_{st}} \le \mu \le \bar{y}_{st} + 2\,\hat{\sigma}_{\bar{y}_{st}}$$
$$13,63 - 2(1,6720) \le \mu \le 13,63 + 2(1,6720)$$
$$10,2859 \le \mu \le 16,9741.$$

Les tableaux suivants définissent la notation formellement :

Les unités de la population sont réparties en L *strates*.

Le tableau suivant définit la notation des paramètres:

N = taille de la population entière	μ_h = moyenne de la strate h
N_h = taille de la strate h, $N_1 + \ldots + N_L = N$	μ = moyenne de la population; $\mu = \Sigma_h W_h \mu_h$
$W_h = N_h/N$, la fraction de la population qui appartient à la strate h. $\Sigma W_h = 1$	S_h = écart-type de la strate h

On tire un échantillon aléatoire simple dans chacune des strates.

n = taille de l'ensemble des échantillons	$\bar{y}_h$ = moyenne de l'échantillon tiré dans la strate h
n_h = taille de l'échantillon tiré dans la strate h. $n_1 + \ldots + n_L = n$	$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h$: estimateur de μ
s_h = écart-type (corrigé) de l'échantillon dans la strate h	$\sigma^2_{\bar{y}_h} = (1 - f_h)\dfrac{S_h^2}{n_h}$: variance de $\bar{y}_h$
$\hat{\sigma}^2_{\bar{y}_h} = (1 - f_h)\dfrac{s_h^2}{n_h}$: estimateur de $\sigma^2_{\bar{y}_h}$	$\sigma^2_{\bar{y}_{st}} = \sum_{h=1}^{L} W_h^2 \sigma^2_{\bar{y}_h}$: variance de $\bar{y}_{st}$
$\hat{\sigma}^2_{\bar{y}_{st}} = \sum_{h=1}^{L} W_h^2 \hat{\sigma}^2_{\bar{y}_h}$: estimateur de la variance de $\bar{y}_{st}$	

Intervalle de confiance pour μ

On détermine un intervalle de confiance de niveau 95 % pour μ par la formule

$$\bar{y}_{st} - 2\,\hat{\sigma}_{\bar{y}_{st}} \le \mu \le \bar{y}_{st} + 2\,\hat{\sigma}_{\bar{y}_{st}}$$

***Exemple* 5.1.1** *Estimation et intervalle de confiance - échantillon stratifié*

D'une population répartie en trois strates de tailles 10, 25 et 50, on prélève un échantillon stratifié de 2 observations de la strate 1, 6 de la strate 2 et 12 de la strate 3. On obtient les données suivantes:

Strate 1	30 215	95 684				
Strate 2	5 656	6 532	6 521	6 582	8 457	9 856
Strate 3	214	659	654	658	123	452
	325	445	323	765	325	235

Déterminer

a) une estimation de la moyenne de la population,
b) une estimation du total de la population;
c) une estimation de $\sigma_{\bar{y}_{st}}$;
d) un intervalle de confiance à 95% pour la moyenne de la population.

Solution. $\bar{y}_1$ = 62 949,5, s_1 = 46 293,57; $\bar{y}_2$ = 7 267, s_2 = 1 567,18; $\bar{y}_3$ = 431,5, s_3 = 208,88.

a) L'estimation de la moyenne est $\bar{y}_{st} = W_1\bar{y}_1 + W_2\bar{y}_2 + W_3\bar{y}_3$ = (10/85) 62 949,5 + (25/85) 7 267 + (50/85) 431,5 = 9 797.

b) Si on estime la moyenne à $\bar{y}_{st}$ = 9 797, alors une estimation du total est $N\,\bar{y}_{st}$ = 85(9 797) = 832 745.

c) Les variances estimées des trois moyennes échantillonnales sont

$$\hat{\sigma}^2_{\bar{y}_1} = (1\text{-}f_1)\frac{s_1^2}{n_1} = \left(1 - \frac{2}{10}\right)\frac{(46293{,}57)^2}{2} = 857\,237\,992,$$

$$\hat{\sigma}^2_{\bar{y}_2} = (1\text{-}f_2)\frac{s_2^2}{n_2} = \left(1 - \frac{6}{25}\right)\frac{(1567{,}18)^2}{6} = 311101{,}119, \text{ et}$$

$$\hat{\sigma}^2_{\bar{y}_3} = (1\text{-}f_3)\frac{s_3^2}{n_3} = \left(1 - \frac{12}{50}\right)\frac{(208{,}88)^2}{12} = 2763{,}27364.$$

La variance de l'estimateur de la moyenne est estimée à

$$\hat{\sigma}^2_{\bar{y}_{st}} = \left(\frac{10}{85}\right)^2 (857237992) + \left(\frac{25}{85}\right)^2 (311101{,}119) + \left(\frac{50}{85}\right)^2 (2763{,}27364) \quad = 11\,892\,967,$$

et l'écart-type estimé est $\hat{\sigma}_{\bar{y}_{st}} = \sqrt{11\,892\,967} = 3\,445.$

d) Un intervalle de confiance approximatif à 95% est donné par

$$\bar{y}_{st} \pm 2\,\hat{\sigma}_{\bar{y}_{st}} = 9\,797 \pm 2\,(3\,449) = 9\,797 \pm 6898:$$

On peut affirmer avec à peu près 95% de confiance que $2\,899 \le \mu \le 16\,695$. □

5.2 Allocation des observations

Supposons qu'on ait décidé de la taille totale n d'un échantillon stratifié. Comment répartir ces n observations entre les strates? Nous considérons ici deux types d'allocation: l'allocation *proportionnelle* et l'allocation *optimale*.

Allocation proportionnelle

L'*allocation proportionnelle* est celle qui vient naturellement à l'esprit : on répartit l'échantillon de la même façon que la population. C'est à dire, les n_h sont tels que $n_h/n = N_h/N = W_h$, ou encore

$$\boxed{n_h = nW_h.}$$

Dans l'exemple présenté au début du chapitre, les tailles des 4 strates sont 105, 63, 21, et 21 et les tailles relatives sont $W_1 = 0{,}5$; $W_2 = 0{,}3$; $W_3 = W_4 = 0{,}1$. Supposons qu'on décide de prélever un échantillon de taille $n = 50$. On répartira les 50 comme suit :

$$n_1 = 50W_1 = 50(0{,}5) = 25;\ n_2 = 50W_2 = 50(0{,}3) = 15\ ;\ n_3 = n_4 = 50W_3 = 5$$

Allocation optimale

Bien qu'il soit naturel de répartir les n_h de façon proportionnelle aux N_h (ou aux W_h), ce n'est pas la meilleure allocation possible. L'allocation des effectifs a un effet important dans la précision d'un estimateur. Pour montrer à quel point, supposons connue la population présentée au tableau A.03, une population de 210 paroisses québécoises. Supposons qu'on ait décidé de tirer un échantillon de taille 60 afin d'estimer le nombre moyen de naissances. Les tailles et les écarts-types des strates sont les suivants :

	Strate 1	*Strate* 2	*Strate* 3	*Strate* 4
N_h	105	63	21	21
S_h	6,3526	5,5285	8,6349	34,5275

Nous allons considérer différentes répartitions possibles et calculer l'écart-type de l'estimateur pour chaque répartition, utilisant la formule

$$\sigma_{\bar{y}_{st}} = \sum_{h=1}^{4} W_h^2 \sigma^2_{\bar{y}_h} \ , \text{ où } \sigma^2_{\bar{y}_h} = (1 - f_h)\frac{S_h^2}{n_h},\ f_h = \frac{n_h}{N_h}$$

Répartition	*Écart-type de l'estimateur* ($\sigma_{\bar{y}_{st}}$)
$n_1 = 4$; $n_2 = 25$; $n_3 = 17$; $n_4 = 4$	2,2167
$n_1 = 4$; $n_2 = 21$; $n_3 = 21$; $n_4 = 4$	2,2195
$n_1 = 25$; $n_2 = 15$; $n_3 = 5$; $n_4 = 5$	1,5419
$n_1 = 17$; $n_2 = 9$; $n_3 = 5$; $n_4 = 19$	0,9658
Échantillon alétoire simple de taille 50	1,9599

Pour fins de comparaison, nous avons calculé aussi l'écart-type de $\bar{y}$ pour un échantillon aléatoire simple de taille $n = 50$: $S_y = 15{,}8770$, et $\sigma_{\bar{y}} = \sqrt{1-\frac{50}{210}}\frac{15{,}8770}{\sqrt{50}} = 1{,}9595$. On remarque que, dépendant de l'allocation, l'échantillonnage stratifié peut être plus efficace ou moins efficace qu'un échantillon aléatoire simple. La stratification avec allocation proportionnelle (écart-type de 1,5419) offre généralement de meilleurs résultats que le tirage aléatoire simple, mais ce n'est pas la meilleure allocation. On en voit une dans le tableau pour laquelle l'écart-type est 0,9658. C'est l'allocation optimale.

L'allocation optimale est celle qui minimise l'écart-type pour une taille n fixée. Dire que l'allocation $n_1 = 17$; $n_2 = 9$; $n_3 = 5$; $n_4 = 19$ est optimale, c'est dire qu'il n'y a aucune autre allocation qui donne un écart-type aussi petit que 0,9658. La règle générale est la suivante : les valeurs de $n_1,\ldots, n_L$ qui minimisent la variance $\sigma^2_{\bar{y}_{st}}$ sont les valeurs proportionnelles aux W_hS_h, c'est-à-dire,

$$n_h = \frac{N_hS_h}{\sum_{i=1}^{L} N_iS_i}\, n$$

On constate que le nombre d'observations que l'on doit prendre dans une strate croît avec N_h, ce qui est normal: plus la strate est grande, plus il faut y prendre des observations. Mais on voit que l'écart-type de la strate S_h est aussi un facteur: plus les données de la strate sont dispersées, plus il faut y prélever d'observations.

C'est la formule que nous avons utilisée pour obtenir l'allocation optimale dans l'exemple. Voici les calculs :

	Strate 1	Strate 2	Strate 3	Strate 4	Somme
N_h	105	63	21	21	210
S_h	6,3526	5,5285	8,6349	34,5275	
N_hS_h	667,023	348,2955	181,3329	725,0775	1921,7289
$(N_hS_h/\sum N_iS_i)\times 50$	17,3548	9,062	4,718	18,8652	50

Nous devons arrondir à l'entier le plus proche, ce qui donnerait 17 ,9, 5, et 19, l'allocation optimale. On constate que l'allocation optimale ici est loin d'être proportionnelle. Les strates 3 et 4 sont de même taille, mais les ressources affectées à la 4e strate sont beaucoup plus importantes. C'est que la dispersion

dans cette strate est très grande comparée à celle des autres. C'est de là que la stratification tient son efficacité : elle permet d'allouer des ressources là où on en a le plus besoin, c'est-à-dire, dans les strates, où, à cause d'une forte dispersion, l'estimation est difficile.

Strates recensées

Il peut arriver que la formule pour l'allocation optimale donne pour certaines strates une valeur de n_h supérieure à N_h. Dans ce cas, on prélève toutes les unités des strates en question, et on utilise l'allocation optimale pour les autres strates.

***Exemple* 5.2.1** *Strate recensée*

Les 515 clients d'un compagnie sont classés en trois strates. Déterminer l'allocation optimale pou un échantillon de taille 60. Les données estimées pour les trois strates :

	Strate 1 *(Clients spéciaux)*	*Strate* 2 *(Grossistes)*	*Strate* 3 *(Particuliers)*
N_h	20	95	400
S_h	13250	800	150
$N_h S_h$	265000	76000	60000
$(N_h S_h / \sum N_i S_i) \times 60$	39,65	11,37	8,98

L'allocation exigerait qu'on tire 40 unités de la strate 1 alors qu'elle n'en contient que 20. Donc on alloue 20 à la strate 1, et les 40 qui restent doivent alors être réparties de façon proportionnelle aux $N_h S_h$ 76000 et 60000, soit 22 et 18. □

Paramètres inconnus

Pour déterminer l'allocation proportionnelle, il suffit de connaître les W_h. S'ils ne sont pas connus, et si on ne peut pas les estimer avec confiance, il n'est pas question de faire de l'échantillonnage stratifié: l'estimateur lui-même dépend des W_h. Pour déterminer l'allocation optimale, par contre, il faut aussi connaître les S_h. Or les S_h sont des paramètres de la population et ne sont pas connus. Il n'y a pas de solution générale à ce problème: en pratique, on tente d'une façon ou d'une autre, d'obtenir la meilleure estimation possible des S_h: soit par un échantillonnage conçu à cette fin; soit en se basant sur des données semblables dans d'autres populations; soit en stratifiant par rapport à une autre variable, corrélée à celle qui nous intéresse, et connue pour la population entière.

5.3 Estimation d'une proportion

Un échantillonnage par stratification peut également être employé avec profit pour estimer une proportion p. Aucune théorie nouvelle n'est réellement nécessaire, puisqu'une proportion est essentiellement une moyenne. En effet, la proportion p des unités appartenant à une classe $\mathfrak{C}$ est alors la moyenne μ d'une variable dichotomique y, comme nous l'avons déjà vu. Donc estimer p, c'est estimer une moyenne. Cependant, nous donnons de nouvelles formules ici de façon à profiter de la simplification

qui résulte du caractère dichotomique de la variable. Les proportions des strates seront dénotées par p_h, et la proportion échantillonnale de la strate h par $\hat{p}_h$. L'estimateur de la proportion p est

$$\hat{p}_{st} = \sum_{h=1}^{L} W_h \hat{p}_h$$

L'écart-type de $\hat{p}_{st}$ est

$$\sigma_{\hat{p}_{st}} = \sqrt{\sum_{h=1}^{L} W_h^2 \sigma_{\hat{p}_h}^2} \ ,$$

où

$$\sigma_{\hat{p}_h}^2 = \left(\frac{N_h - n_h}{N_h - 1} \right) \frac{p_h (1 - p_h)}{n_h}$$

Un estimateur de $\sigma_{\hat{p}_{st}}$ est donné par

$$\hat{\sigma}_{\hat{p}_{st}} = \sqrt{\sum_{h=1}^{L} W_h^2 \hat{\sigma}_{\hat{p}_h}^2}$$

où

$$\hat{\sigma}_{\hat{p}_h} = \sqrt{(1 - f_h)} \sqrt{\frac{\hat{p}_h (1 - \hat{p}_h)}{n_h - 1}}$$

Un intervalle de confiance approximatif de niveau 95 % est donné par

$$\hat{p}_{st} - 2\,\hat{\sigma}_{\hat{p}_{st}} \le p \le \hat{p}_{st} + 2\,\hat{\sigma}_{\hat{p}_{st}} .$$

***Exemple* 5.3.1** *Estimation et intervalle de confiance - proportion et effectif*

Pour estimer la proportion d'employés en faveur d'un plan de soins dentaires, on prélève un échantillon aléatoire simple dans chacune des 4 divisions de la compagnie. Les effectifs des 4 divisions sont 4 523; 3 456; 1 300; 1 124, et les tailles des échantillons sont 22; 17; 6 et 5; respectivement. Le nombre de personnes favorables est 10, 5 et 3 dans les trois échantillons, respectivement.

a) Estimez la proportion et le nombre d'employés en faveur d'un plan dentaire dans la compagnie;

b) estimez les écarts-types des estimateurs utilisés en *a*);

c) déterminez un intervalle de confiance pour la proportion et pour le nombre d'employés en faveur d'un plan dentaire.

Solution Les W_h sont $\frac{4523}{10403}$, $\frac{3456}{10403}$, $\frac{1300}{10403}$, $\frac{1124}{10403}$; les $\hat{p}_h$ sont 10/22, 5/17, 3/6, 3/5. Voici les résultats:

W_h	$\hat{p}_h$	$W_h\,\hat{p}_h$	f_h	$\sigma_{\hat{p}_h}^2$	$W_h^2\ \sigma_{\hat{p}_h}^2$
0,43478	0,45455	0,19763	0,00486	0,01175	0,00222
0,33221	0,29412	0,09771	0,00492	0,01291	0,00143
0,12496	0,5	0,06248	0,00462	0,04977	0,00078
0,10805	0,6	0,06483	0,00445	0,05973	0,00070
1		0,42264536			0,00512047

a) Donc $\hat{p}_{st} = \frac{4523}{10403} \times \frac{10}{22} + \frac{3456}{10403} \times \frac{5}{17} + \frac{3456}{10403} \times \frac{5}{17} + \frac{1300}{10403} \times \frac{3}{6} + \frac{1124}{10403} \times \frac{3}{5} = 0{,}42264536$: on estime que 42,26% des employés de la compagnie sont en faveur du plan. $\hat{N}_c = N\,\hat{p}_{st} = 10403 \times 0{,}42264536 = 4397$: il y a 4 397 employés en faveur.

b) Les variances des estimateurs $\hat{p}_h$ sont

$$\hat{\sigma}_{\hat{p}_1}^2 = 0{,}01175;\ \hat{\sigma}_{\hat{p}_2}^2 = 0{,}01291;\ \hat{\sigma}_{\hat{p}_3}^2 = 0{,}04977;\ \hat{\sigma}_{\hat{p}_4}^2 = 0{,}05973.$$

La variance de $\hat{p}_{st}$ est

$(0{,}43478)^2(0{,}01175) + (0{,}33221)^2(0{,}01291) + (0{,}12496)^2(0{,}04977) + (0{,}10805)^2(0{,}05973)$

= 0,00512047

et donc $\hat{\sigma}_{\hat{p}_{st}} = \sqrt{0{,}00512047} = 0{,}0715575$. L'écart-type de $\hat{N}_c$ est $\hat{\sigma}_{\hat{N}_c} = N\hat{\sigma}_{\hat{p}_{st}} = 10\,403\,(0{,}0715575) = 744$.

c) $0{,}4226 - 2\,(0{,}0715575) \le p \le 0{,}4226 + 2\,(0{,}0715575)$, ou $0{,}279445 \le p \le 0{,}565755$: on peut affirmer avec environ 95% de confiance que le pourcentage d'employés en faveur du plan est compris entre 27,9% et 56,6%; on peut donc affirmer que le *nombre* d'employés en faveur est compris entre $(10\,403)(0{,}279445) = 2907$ et $(10\,403)(0{,}565755) = 5\,886$. □

Allocation optimale

L'allocation optimale est donnée par

$$n_h = \frac{W_h\sqrt{N_h p_h(1-p_h)/(N_h-1)}}{\sum_{i=1}^{L} W_i\sqrt{N_i p_i(1-p_i)/(N_i-1)}}\, n$$

Une formule approximative plus simple est la suivante:

$$n_h = \frac{W_h\sqrt{p_h(1-p_h)}}{\sum_{i=1}^{L} W_i\sqrt{p_i(1-p_i)}}\, n$$

Pour appliquer ces formules il faut connaître les valeurs des p_h, ou du moins d'une estimation de ces valeurs. Il peut arriver, en l'absence d'information sur les p_h, qu'on les suppose égaux. Dans ce cas, l'allocation optimale est équivalente à l'allocation proportionnelle.

A propos de la supposition que les p_h sont égaux. En pratique, cette hypothèse ne sera vérifiée qu'approximativement au mieux. Cependant, la formule ci-dessus montre que l'allocation optimale dépend essentiellement des valeurs $p_h(1- p_h)$ et non des p_h eux-mêmes. Or les valeurs du produit $p_h(1- p_h)$ varient peu, même lorsque celles des p_h varient assez. Les seules situations où ceci n'est pas vrai sont celles où les p_h sont très petits ou très grands. Autrement les produits sont en fait à peu près égaux. Ceci signifie, en résumé, qu'à moins que les valeurs des p_h soient extrêmes, l'allocation optimale n'est pas tellement supérieure à l'allocation proportionnelle.

***Exemple* 5.3.2** *Répartition optimale en fonction des proportions p_h des strates*

Les 3 strates d'une population contiennent respectivement 175, 375, et 450 unités.

a) Déterminer la répartition optimale d'un échantillon de taille 100 si (*i*) $p_1 = 0{,}4$, $p_2 = 0{,}5$, $p_3 = 0{,}6$, et si (*ii*) $p_1 = 0{,}05$, $p_2 = 0{,}15$, $p_3 = 0{,}25$; comparer avec la répartition proportionnelle;

b) pour chacun des cas ci-dessus, calculer l'écart-type de $\hat{p}_{st}$.

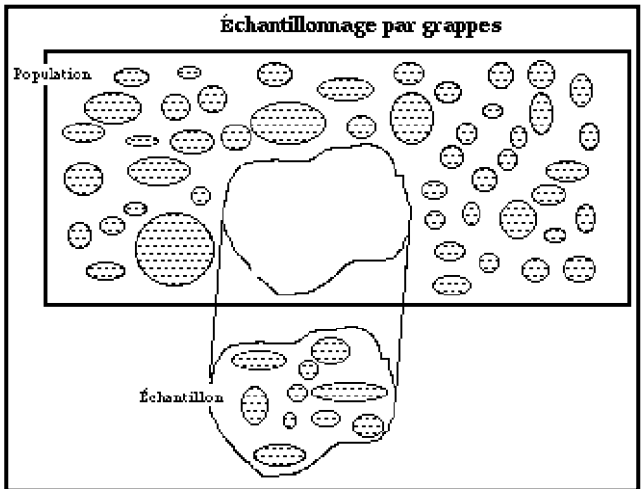
Solution. Les valeurs de W_h sont 0,175, 0,375, 0,450.

a) (*i*) les valeurs de $W_h\sqrt{p_h(1-p_h)}$ sont 0,0857, 0,1875, 0,2205, leur somme est 0,4937, l'échantillon doit être réparti selon les proportions 0,0857/0,4937; 0,1875/0,4937; 0,2205/0,4937, ce qui donne environ n_h = 17, 38 et 45; (*ii*) les n_h cette fois-ci sont 11, 36, 53. On voit bien que la première répartition est essentiellement la répartition proportionnelle.

b) La variance de $\hat{p}_{st}$ dans le cas (*i*) est $[(0{,}175)^2\ 0{,}4(1-0{,}4)/17]\ (1-17/175) + [(0{,}375)^2\ 0{,}5(1-0{,}5)/38]\ (1-38/375) + [(0{,}450)^2\ 0{,}6(1-0{,}6)/45]\ (1-45/450) = 0{,}00219$, alors que dans (*ii*) elle est $(0{,}175)^2\ 0{,}05(1-0{,}05)/11]\ (1-11/175) + [(0{,}375)^2 0{,}15(1-0{,}15)/36]\ (1-36/375) + [(0{,}450)^2\ 0{,}25\ (1-0{,}25)/53]\ (1-53/450) = 0{,}0012$. □

5.4 Échantillonnage par grappes

Figure 5.4.1



Supposons que la population cible d'un sondage est l'ensemble des employés d'une compagnie de vente au détail; et que ceux-ci sont répartis en N succursales. On pourrait songer à considérer les N succursales comme des strates et procéder à un échantillonnage stratifié. Cependant, ceci exigerait qu'on prélève un échantillon dans *chacune* des succursales, ce qui peut être mal commode et coûteux. Pour ces raisons purement pratiques, il est préférable de prélever un échantillon de *succursales plutôt qu'un échantillon de personnes.* Les succursales sont alors appelées des ***grappes*** ou des ***unités primaires***. A l'intérieur de chacune des succursales choisies on interroge tous les employés. Ces derniers sont appelés des ***unités secondaires*** ou des ***sous-unités***.

Plusieurs méthodes ont déjà été proposées pour ce problème. Nous présentons ici, dans leurs grandes lignes, trois des méthodes couramment utilisées. Deux d'entre elles sont basées sur un tirage aléatoire simple de grappes et utilisent un mode d'estimation classique, déjà discuté dans ces notes. La troisième est nouvelle. Elle est basée sur l'échantillonnage avec probabilités inégales.

Tirage aléatoire simple de grappes

On pourrait, à la limite, décider de considérer une grappe comme une unité ordinaire, et prendre pour y la valeur *totale* de la variable d'intérêt. Dans ce cas, toutes les techniques étudiées jusqu'ici peuvent s'appliquer intégralement. Supposons, pour reprendre l'exemple des succursales de compagnie, que nous connaissons les valeurs des paramètres suivants :

N = Nombre de succursales = 180
M = Nombre d'employés dans les succursales = 3 500.

Supposons que nous voulons estimer le nombre moyen de *dépendants* par employé, et que pour ce faire, on prélève un échantillon aléatoire simple de 15 succursales. Voici les données :

Tableau 5.4.1
Échantillon de 15 succursales tirées d'une population de 180 succursales ayant en tout 3 500 employés

i	*Nombre d'employée* (M_i)	*Nombre* total *de dépendants* (y_i)
1	24	34
2	4	7
3	21	27
4	17	21
5	15	19
6	6	11
7	3	3
8	27	38
9	17	24
10	23	20
11	15	15
12	28	33
13	16	15
14	18	25
15	24	30
Somme	**258**	**322**
Moyenne	17,2	21,4667
Écart-type		10,1339

Estimation par la moyenne :

L'estimateur par la moyenne, ainsi que l'estimateur par le quotient que nous introduirons ensuite, sont des estimateurs où l'unité primaire est considérée comme une unité indivisible: les sous-unités n'interviennent qu'à la toute fin. Nous avons donc une population de N unités, à chacune desquelles est associée une valeur y.

On calcule la moyenne et l'écart-type échantillonnal des y_i puis on détermine un intervalle de confiance :

La moyenne $\bar{y}$ = 21,4667 est une estimation du nombre moyen de dépendants *par succursale*. On a alors les estimations suivantes :

Nombre total de dépendants dans la compagnie : $180 \times \bar{y} = 180 \times 21{,}447 = 3864$

Nombre moyen de dépendants par employé : $180 \times \bar{y} / 3\,500 = 3864/3500 = 1{,}104$.

Pour déterminer un intervalle de confiance, commençons par l'estimateur $\bar{y}$:

$$\bar{y} = 21{,}4667 \; ; \; s_y = 10{,}1339 \; ; \; \hat{\sigma}_{\bar{y}} = \sqrt{1-\frac{15}{180}}\,\frac{10{,}1339}{\sqrt{15}} .$$

Intervalle de confiance pour la moyenne par succursale :

$$\bar{y} - 2\,\hat{\sigma}_{\bar{y}} \leq \text{Nombre moyen de dépendants par succursale} \leq \bar{y} + 2\,\hat{\sigma}_{\bar{y}}$$
$$16{,}4563 \leq \text{Nombre moyen de dépendants par succursale} \leq 26{,}4770$$

Si on multiplie par $N = 180$, on obtient un intervalle de confiance pour le nombre *total* de dépendants dans la compagnie :

$$180(16{,}4563) \leq \text{Nombre total de dépendants} \leq 180(26{,}4770)$$
$$2962 \leq \text{Nombre total de dépendants} \leq 4766$$

Enfin, si on divise par $M = 3\,500$, on obtient le nombre moyen de dépendants *par employé* :

$$0{,}8463 \leq \text{Nombre moyen de dépendants par employé} \leq 1{,}3617$$

Estimation par le quotient :

Ici aussi, on procède ici comme si la grappe était une unité ordinaire, mais là on utilise l'estimateur par le quotient, avec le nombre d'employés comme variable auxiliaire. La colonne dénotée par M_i pourrait aussi bien être dénotée par x_i et on peut utiliser les formules connues pour l'estimation par le quotient. Le nombre moyen de dépendants par *employé* est en fait un quotient :

$$R = \frac{\sum_{i=1}^{180} y_i}{M} = \frac{\sum_{i=1}^{180} y_i}{\sum_{i=1}^{180} M_i} .$$

L'estimateur est donc (utilisant la notation x_i à la place de M_i)

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{\sum_{i=1}^{15} y_i}{\sum_{i=1}^{15} x_i} = \frac{322}{258} = 1{,}2481$$

Vous remarquez que cet estimateur est on ne peut plus intuitif : on estime le nombre moyen de dépendants par personnes en divisant le nombre de dépendants par le nombre de personnes.

L'écart-type de $\hat{R}$ est estimé par : $\hat{\sigma}_{\hat{R}} = \frac{\sqrt{1-f}}{\bar{x}}\,\frac{\sqrt{s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R} s_{xy}}}{\sqrt{n}}$. On a

$$s_y^2 = 102{,}6952 \; ; \; s_x^2 = 61{,}8857 \; ; \; s_{xy} = 74{,}4714$$

$$\hat{\sigma}_{\hat{R}} = \frac{\sqrt{1-15/180}}{17{,}2}\,\frac{\sqrt{102{,}6952 + (1{,}2481^2)(61{,}8857) - 2(1{,}2481)(74{,}4714)}}{\sqrt{15}} = 0{,}05222$$

L'intervalle de confiance est

$$\hat{R} - 2\,\hat{\sigma}_{\hat{R}} \leq \text{Nombre moyen de dépendants par employé} \leq \hat{R} - 2\,\hat{\sigma}_{\hat{R}}$$
$$1{,}2481 - 2(0{,}05222) \leq \text{Nombre moyen de dépendants par employé} \leq 1{,}2481 - 2(0{,}05222)$$

$$1{,}1436 \leq \text{ Nombre moyen de dépendants par employé} \leq 1{,}3525.$$

Remarque *La marge d'erreur ici est de 0,1045, alors qu'elle était de 0,2577 pour l'estimateur par la moyenne. L'estimation par le quotient est nettement meilleure, et ce n'est pas étonnant. On se souvient que l'estimateur par le quotient est particulièrement efficace lorsqu'il y a une forte corrélation entre la variable d'intérêt et la variable auxiliaire. Ici, la variable d'intérêt est le nombre de dépendants total dans la grappe; et la variable auxiliaire est le nombre d'employés. Il est évident que plus il y a d'employés, plus il y a de dépendants.*

Échantillonnage avec remise et probabilités proportionnelles aux tailles

Nous étudions maintenant l'une des approches développées spécialement pour l'échantillonnage par grappes. Elle a deux caractéristiques nouvelles:

1) les tirages se font *avec* remise, et
2) à chaque tirage, la probabilité de tirer la grappe *i* est proportionnelle à M_i.

Dans le contexte actuel, ce mode de tirage peut se réaliser de la façon suivante :

- On dresse une liste de tous les *employés* de la compagnie
- Au hasard, on choisit un employé dans la liste, en donnant à chaque employé la même probabilité de sélection, soit $1/M$
- On inclut dans l'échantillon tous les employés de la succursale à laquelle appartient l'employé choisi.
- On recommence *n* fois, toujours à partir de la population entière. Une succursale peut appartenir à l'échantillon plus d'une fois.

Nous allons considérer les mêmes données que celles du tableau 5.4.1, mais nous supposerons que cet échantillon de taille 15 a été tiré avec probabilités proportionnelles aux effectifs des grappes M_i .

La technique d'estimation consiste ici à calculer la *moyenne par employé* dans chaque grappe (succursale) et de considérer ces moyennes comme nos données de bases, à partir desquelles on estimera une moyenne. On montre les calculs au tableau 5.4.2.

L'estimateur de la moyenne par employé est simplement la moyenne des moyennes $\bar{y}_i$. On dénote cet estimateur par $\bar{\bar{y}}$, puisque c'est une moyenne de moyennes :

$$\text{Moyenne par employé : } \bar{\bar{y}} = \frac{\sum_{i=1}^{15} \bar{y}_i}{15} = \frac{19{,}2313727}{15} = 1{,}282$$

On estime l'écart-type de l'estimateur par $\hat{\sigma}_{\bar{\bar{y}}} = \frac{s_{\bar{y}}}{\sqrt{n}}$, où $s_{\bar{y}}$ est l'écart-type des $\bar{y}_i$:

$$\frac{s_{\bar{y}}}{\sqrt{n}} = \frac{0{,}273483}{\sqrt{15}} = 0{,}0706.$$

Finalement, on calcule l'intervalle de confiance :

$$\bar{\bar{y}} - 2\hat{\sigma}_{\bar{\bar{y}}} \leq \text{Moyenne par employé} \leq \bar{\bar{y}} - 2\hat{\sigma}_{\bar{\bar{y}}}$$

$$1{,}1409 \leq \text{Moyenne par employé} \leq 1{,}4233$$

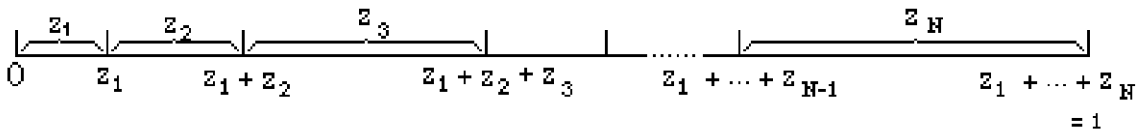
Tableau 5.4.2
Échantillon de 15 succursales tirées d'une population de 180 succursales ayant en tout 3 500 employés

i	*Nombre d'employée* (M_i)	*Nombre* total *de dépendants* (y_i)	*Moyenne par employé* ($\bar{y}_i$)
1	24	34	1,4166667
2	4	7	1,7500000
3	21	27	1,2857143
4	17	21	1,2352941
5	15	19	1,2666667
6	6	11	1,8333333
7	3	3	1,0000000
8	27	38	1,4074074
9	17	24	1,4117647
10	23	20	0,8695652
11	15	15	1,0000000
12	28	33	1,1785714
13	16	15	0,9375000
14	18	25	1,3888889
15	24	30	1,2500000
Total	258	322	19,2313727
Moyenne	17,2	21,4667	1,2820915
Écart-type		10,1339	0,2734830

Comment générer l'échantillon

Supposons que nous disposons des moyens pour générer des nombres de loi uniforme sur l'intervalle [0 ; 1) — des tables de nombres au hasard, des boules numérotées, ou un programme d'ordinateur. Comment utiliser ces facilités pour choisir les unités primaires avec probabilités proportionnelles à leurs tailles? Soit $z_i = M_i / M$, pour $i = 1, 2, \ldots, N$.

Une procédure relativement simple est la suivante. On découpe l'intervalle [0 ; 1) en n sous-intervalles. Le premier est de longueur z_1, le deuxième est de longueur z_2, ainsi de suite, jusqu'au dernier, qui est de longueur z_N. On génère un nombre U de loi uniforme sur l'intervalle [0,1).



Ce nombre appartiendra à l'un des sous-intervalles définis ci-dessus. On choisira la i^e unité primaire si U est dans le i^e intervalle.

***Exemple* 5.4.3** *Choix d'un échantillon à probabilités inégales.*

Supposons que pour tirer un échantillon de taille 4 de la population dans l'exemple 5.4.1, on génère 4 nombres au hasard. On obtient les nombres suivants:

0,0274; 0,5122; 0,6329; 0,8675.

Quelles unités primaires doit-on alors tirer de la population?

Solution. Les z_i et leurs valeurs cumulées sont données dans le tableau suivant

z_i	0,01	0,04	0,07	0,09	0,10	0,11	0,12	0,13	0,15	0,18
z_i *cumulés*	0,01	0,05	0,12	0,21	0,31	0,42	0,54	0,67	0,82	1,00

Les 10 intervalles sont donc [0; 0,01), [0,01; 0,05), [0,05; 0,12), [0,12; 0,21), [0,21; 0,31), [0,31; 0,42), [0,42; 0,54), [0,54; 0,67), [0,67; 0,82), [0,82; 1,00).

Puisque

0,0274 ∈ [0,01; 0,05), *intervalle* 2,
0,5122 ∈ [0,42; 0,54), *intervalle* 7,
0,6329 ∈ [0,54; 0,67), *intervalle* 8,
0,8675 ∈ [0,82; 1,00), *intervalle* 10,

les unités primaires à retirer sont les unités 2, 7, 8 et 10. □

Remarque *Noter que puisqu'on tire avec remise, si deux des nombres au hasard tombent dans le même intervalle, l'unité correspondante est tirée deux fois. Par exemple, si les 4 nombres au hasard sont*

0,0274 ∈ [0,01; 0,05), *intervalle 2,*
0,0482 ∈ [0,42; 0,54), *intervalle 2,*
0,6329 ∈ [0,54; 0,67), *intervalle 8,*
0,8675 ∈ [0,82; 1,00), *intervalle 10,*

les données sont

y_i	110	110	400	510
z_i	0,04	0,04	0,13	0,18

□

5.5 Échantillonnage systématique

L'échantillonnage systématique est un mode d'échantillonnage extrêmement populaire. On suppose d'abord que les unités de la population sont disposées dans un certain ordre. Par exemple, les patients d'un médecin peuvent être représentés par des cartes disposées par ordre alphabétique dans un tiroir; les plants de tomates sont disposées en rangées; les spectateurs d'un film défilent dans un certain ordre à la sortie. On peut donc supposer que les unités sont numérotées de 1 à N. L'échantillon systématique consiste alors à choisir les unités de la population en prenant les unités à des intervalles réguliers.

***Exemple* 5.5.1** Vous voulez prélever un échantillon de patients dans une clinique médicale où les dossiers des patients sont rangés dans des tiroirs en ordre alphabétique. Votre population est de taille 1 000 et vous voulez tirer un échantillon de taille $n = 20$. Au lieu de tirer un échantillon aléatoire simple, vous pouvez décider de choisir systématiquement chaque 50^e dossier dans les tiroirs. Le seul aspect aléatoire du tirage est le choix du premier élément tiré. □

Le principal avantage de l'échantillonnage systématique est une certaine facilité d'exécution. En outre, il semble intuitivement prometteur : il donne l'assurance que chaque partie de la population sera représentée. Il a cependant un inconvénient majeur : il ne permet pas d'estimer l'écart type de l'estimateur. Par conséquent, on ne peut pas se faire une idée de la taille de l'échantillon qu'il faut prélever.

***Exemple* 5.5.2** Vous voulez tirer un échantillon de l'auditoire d'une pièce de théâtre. Vous pouvez vous installer à la porte du théâtre et interroger, disons, chaque dixième spectateur à son arrivée en salle. Remarquez que vous n'avez pas besoin de connaître la taille de la population pour faire ce type d'échantillonnage, et c'est là un autre avantage de l'échantillonnage systématique. □

Exemple 5.5.3 Vous voulez tirer un échantillon de 20 plants dans une rangée de 800 mètres de plants cultivés. Vous pouvez alors tirer un plant à chaque 40 mètres de la rangée. □

Considérons les poids d'une population de 171 oranges cueillies dans un arbre. Les voici, dans l'ordre de la cueillette :

350,430,260,511,492,449,344,406,352,406,294,438,368,492,302,475,458,468,432,245,341,296,510,476,459,515,363,276,468,476,430,287,438,484,477,539,399,470,275,255,422,465,416,410,467,416,459,431,377,351,494,476,410,403,423,432,469,523,476,474,439,455,474,511,426,482,461,431,303,424,509,283,479,513,507,428,451,446,482,481,476,444,525,434,509,398,446,538,413,415,294,426,426,411,475,441,526,443,440,510,431,467,435,444,439,483,463,445,463,444,437,516,441,472,430,456,387,395,406,364,434,410,303,424,448,331,393,509,447,440,439,312,331,328,453,397,543,473,384,401,420,253,246,356,400,431,352,278,404,398,361,287,426,321,337,318,436,346,280,295,444,302,281,376,376,317,266,226,319,240,267.

On décide de tirer et peser chaque 20^{e} orange. Supposons qu'on dispose les unités de la population en colonnes de longueur 20, de la façon suivante:

Ligne									
1	1	21	41	61	81	101	121	141	161
2	2	22	42	62	82	102	122	142	162
3	3	23	43	63	83	103	123	143	163
4	4	24	44	64	84	104	124	144	164
5	5	25	45	65	85	105	125	145	165
6	6	26	46	66	86	106	126	146	166
7	7	27	47	67	87	107	127	147	167
8	8	28	48	68	88	108	128	148	168
9	9	29	49	69	89	109	129	149	169
10	10	30	50	70	90	110	130	150	170
11	11	31	51	71	91	111	131	151	171
12	12	32	52	72	92	112	132	152	
13	13	33	53	73	93	113	133	153	
14	14	34	54	74	94	114	134	154	
15	15	35	55	75	95	115	135	155	
16	16	36	56	76	96	116	136	156	
17	17	37	57	77	97	117	137	157	
18	18	38	58	78	98	118	138	158	
19	19	39	59	79	99	119	139	159	
20	20	40	60	80	100	120	140	160	

Alors chaque ligne représente l'un des échantillons possibles. Pour tirer l'échantillon, on choisira au hasard un nombre entre 1 et 20. Le nombre sur lequel on tombe déterminera la ligne qui constituera

l'échantillon. Ainsi, si on tire le numéro 9, l'échantillon sera composé des unités 9, 29, 49, 69, 89, 109, 129, 149, et 169 — un échantillon de taille 9; si on tombe sur le numéro 18, on prendra les unités 18, 38, 58, 78, 98, 118, 138, et 158 — un échantillon de taille 8. En général, si on tombe sur un numéro de 1 à 11, l'échantillon sera de taille 9; autrement il sera de taille 8. Afin d'examiner les possibilités d'un tel échantillonnage, considérons les salaires moyens de ces échantillons. Les données sont présentées dans le même ordre que dans le tableau précédent. La dernière colonne donne les moyennes des 20 échantillons possibles. La moyenne de la population est 411,42. Les 20 moyennes échantillonnales varient autour de cette valeur. Leur moyenne est 411,74.

Échantillon										***Moyenne $\bar{y}$ de l'échantillon***
1	350	341	422	439	476	431	434	420	444	417,44
2	430	296	465	455	444	467	410	253	302	391,33
3	260	510	416	474	525	435	303	246	281	383,33
4	511	476	410	511	434	444	424	356	376	438,00
5	492	459	467	426	509	439	448	400	376	446,22
6	449	515	416	482	398	483	331	431	317	424,67
7	344	363	459	461	446	463	393	352	266	394,11
8	406	276	431	431	538	445	509	278	226	393,33
9	352	468	377	303	413	463	447	404	319	394,00
10	406	476	351	424	415	444	440	398	240	399,33
11	294	430	494	509	294	437	439	361	267	391,67
12	438	287	476	283	426	516	312	287		378,13
13	368	438	410	479	426	441	331	426		414,88
14	492	484	403	513	411	472	328	321		428,00
15	302	477	423	507	475	430	453	337		425,50
16	475	539	432	428	441	456	397	318		435,75
17	458	399	469	451	526	387	543	436		458,63
18	468	470	523	446	443	395	473	346		445,50
19	432	275	476	482	440	406	384	280		396,88
20	245	255	474	481	510	364	401	295		378,13
Moyenne des moyennes										**411,74**

Il aurait été préférable que la moyenne des $\bar{y}$, $\mu_{\bar{y}}$, soit égale à μ_y, la moyenne de la population, comme elle l'aurait été s'il s'agissait d'un échantillon aléatoire simple: ce serait un signe que la méthode a tendance à estimer juste. Mais la différence entre $\mu_{\bar{y}}$ et μ_y est petite, et c'est ce qui arrivera généralement en pratique. En fait, s'il se trouve que *N* est un multiple de *n*, alors on aura nécessairement $\mu_{\bar{y}} = \mu_y$.

Calculons maintenant l'écart type des $\bar{y}$. Nous obtenons $\sigma_{\bar{y}} = 24{,}17$. Cet écart type est proche de l'écart-type de $\bar{y}$ dans un échantillon aléatoire simple, qui est de 24,32 pour un échantillon de taille 9 et de 25,88 pour un échantillon de taille 8.

Donc dans cet exemple, les propriétés d'un échantillon systématique ont l'air comparables à celles d'un échantillon aléatoire simple. Mais qu'en est-il de l'échantillonnage systématique *en général*? Règle générale, on peut utiliser la moyenne échantillonnale $\bar{y}$ pour estimer la moyenne de la population, tout comme avec un échantillon aléatoire simple. C'est un estimateur « presque » sans biais. C'est au niveau de la précision que l'échantillonnage systématique présente quelques difficultés. Non pas qu'il soit nécessairement imprécis: il est fort probable que dans la plupart des applications, l'échantillonnage systématique soit aussi précis, et peut-être même plus précis, qu'un échantillon aléatoire simple. Mais le problème c'est que dans une situation donnée, il sera impossible d'estimer la précision. Car il n'existe pas d'estimateur de l'écart type. Le calcul $\sqrt{1-f}\ \frac{s}{\sqrt{n}}$ qui dans un échantillon aléatoire simple estime l'écart type de $\bar{y}$, ne donne pas un estimateur valable ici. Donc lorsque pour des raisons pratiques nous faisons de l'échantillonnage systématique, nous devons nous rappeler que nous ne pouvons pas évaluer la qualité de nos estimations.

5.6 Résumé

1 L'estimateur de la moyenne dans un échantillon stratifié est $\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h$; son écart type est $\sigma_{\bar{y}_{st}} = \sqrt{\sum_{h=1}^{L} W_h^2 \sigma_{\bar{y}_h}^2}$ où $\sigma_{\bar{y}_h}^2 = (1-f_h)\frac{S_h^2}{n_h}$ et $f_h = n_h/N_h$.

2 L'écart type de $\bar{y}_{st}$ est minimisé pour un échantillon de taille totale n lorsque les n_h sont proportionnels aux $W_h S_h$.

3 L'estimateur d'une proportion dans un échantillon stratifié est $\hat{p}_{st} = \sum_{h=1}^{L} W_h \hat{p}_h$. Son écart type est estimé par $\hat{\sigma}_{\hat{p}_{st}} = \sqrt{\sum_{h=1}^{L} W_h^2 \hat{\sigma}_{\hat{p}_h}^2}$ où $\hat{\sigma}_{\hat{p}_h} = \sqrt{(1-f_h)}\sqrt{\frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}}$.

4 Échantillonnage par grappes.

Tirage avec probabilités égales, estimateur par la moyenne

Les y_i sont les totaux des grappes, $\bar{y}$ estime la moyenne par grappe, et l'écart-type de $\bar{y}$ est estimé par $\hat{\sigma}_{\bar{y}} = \sqrt{1-\frac{n}{N}}\frac{s_y}{\sqrt{n}}$.

Tirage avec probabilités égales, estimateur par le quotient

On estime la moyenne par unité secondaire par $\hat{R} = \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i}$, qui est tout simplement un estimateur par le quotient avec les M_i pour variable auxiliaire. On estime l'écart-type de $\hat{R}$ par $\hat{\sigma}_{\hat{R}} = \dfrac{\sqrt{1-f}}{\bar{x}} \dfrac{\sqrt{s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R} s_{xy}}}{\sqrt{n}}$, où les x_i sont les M_i.

Tirage avec remise et probabilités proportionnelles aux M_i

On calcule, pour chaque grappe de l'échantillon, la moyenne par unité secondaire $\bar{y}_i$. L'estimateur de la moyenne par unité secondaire est la moyenne de ces $\bar{y}_i$: $\bar{\bar{y}} = \dfrac{\sum_{i=1}^{n} \bar{y}_i}{n}$.

5 La moyenne d'un échantillon systématique est un estimateur presque sans biais de la moyenne de la population. L'échantillonnage systématique est probablement très efficace la plupart du temps, mais il est impossible d'estimer l'écart type de l'estimateur à partir d'un échantillon.

5.7 Exercices

5.1 D'une population formée de 3 strates de 610, 1 670 et 915 unités, on prélève un échantillon stratifié. On obtient les résultats suivants:

Échantillons
Strate 1: 68, 98, 87, 56, 34, 33, 44, 28;
$\Sigma_{i=1}^{n} y_i = 448$; $s_1 = 26{,}23134$

Strate 2: 2, 3, 4, 2, 3, 4, 3, 2, 6, 5, 3, 4, 2, 5, 6, 2, 3, 5, 4, 2, 3, 8.
$\Sigma_{i=1}^{n} y_i = 81$; $s_2 = 1{,}615000$

Strate 3: 687, 675, 237, 99, 123, 456, 231, 324, 543, 654, 345, 234;
$\Sigma_{i=1}^{n} y_i = 4\,608$; $s_3 = 213{,}5926$.

a) Estimer la moyenne de la population
b) Estimer le total de la population
c) Estimer l'écart type de l'estimateur de la moyenne
d) Estimer l'écart type de l'estimateur du total
e) Déterminer un intervalle de confiance à 95% pour la moyenne de la population
f) Déterminer un intervalle de confiance à 95% pour le total de la population.
g) Utiliser les données de l'échantillon pour déterminer l'allocation optimale d'un échantillon de taille 42.
h) Estimer ce qu'aurait été l'écart type de $\bar{y}_{st}$ si l'allocation optimale avait été utilisée.

5.2 Une population est formée de 5 strates comprenant 235, 432, 1 590, 2 300, et 4 321 unités. On estime que les écarts-types S_h sont égaux à 60, 36, 14, 12, et 10.

a) Déterminer l'allocation optimale d'un échantillon de taille 180.
b) Déterminer l'allocation proportionnelle pour une échantillon de taille 180 et comparer l'écart type de $\bar{y}_{st}$ pour l'allocation proportionnelle et l'allocation optimale.

5.3 Déterminer l'allocation optimale d'un échantillon de taille 100 d'une population dont les 3 strates contiennent 30, 300, et 2 000 unités, et les écarts-types sont 150, 38, et 14.

5.4 Une population de 85 comptes est répartie selon le type de client. Les montants de ces comptes sont les suivants:

Strate 1 : Clients industriels

$ 50 212	30 215	12 564	36 598	36 598
36 527	96 532	95 684	69 854	68 594

Strate 2 : Grossistes

3652	6598	6532	5656	6532
6563	6521	6532	6598	6532
3268	8854	6582	8457	6584
9658	6532	9564	9856	6598
9658	6532	2147	3345	5465

Strate 3 : Détaillants

325	695	658	423	214	659	854	632	632	654
985	658	745	698	365	256	985	654	965	965
985	658	321	123	365	965	965	856	452	325
445	323	765	139	239	432	871	347	138	325
762	769	126	247	246	235	345	345	345	298

Quel est l'écart-type de l'estimateur de la moyenne basé sur un échantillon de taille n = 20, réparti de la façon suivante?

a) n_1 = 6, n_2 = 11, n_3 = 3; *b*) n_1 = 10, n_2 = 8, n_3 = 2; *c*) n_1 = 1, n_2 = 4, n_3 = 15.

5.5 Les étudiants d'une université sont répartis en 4 familles ayant 1 230, 3 000, 2 500, et 8 000 étudiants, respectivement. On prélève un échantillon de 25, 61, 51, et 163 étudiants dans les 4 strates pour estimer la proportion p d'étudiants qui ont déjà utilisé la coop. On trouve que les nombres d'étudiants qui l'ont déjà utilisée dans les 4 échantillons sont 20, 43, 46, et 81, respectivement.

a) Estimer la proportion p d'étudiants dans la population qui ont déjà utilisé la coop, ainsi que l'écart type de l'estimateur
b) Déterminer un intervalle de confiance à 95% pour p
c) Estimer le nombre N_c d'étudiants dans la population qui ont déjà utilisé la coop
d) Déterminer un intervalle de confiance à 95% pour N_c
e) Utiliser l'estimation de p obtenue en *a*) pour estimer l'écart type d'un estimateur basé sur un échantillon aléatoire simple de taille 300.
f) Utiliser les résultats de ce sondage pour planifier un prochain sondage éventuel dans la même population: déterminer l'allocation optimale, et dites quelles devront être les tailles des échantillons si on veut que la demi-largeur d'un intervalle de confiance à 95% soit de 0,001.

5.6 D'une population formée de 578 ménages d'un quartier on tire avec remise un échantillon de 20 ménages avec probabilités proportionnelles au nombre de personnes dans le ménage. Le nombre total de personnes dans le quartier est connu: 2 350. Pour chaque ménage, on détermine le montant y_i dépensé par semaine pour dîner et souper dans des restaurants. Le but de l'échantillonnage est d'estimer la valeur totale τ de y pour le quartier. Voici les données de l'échantillon (M_i = nombre de personnes dans la famille, y_i = dépenses hebdomadaires):

M_i	y_i	M_i	y_i	M_i	y_i
4	101,44	5	104,25	4	101,23
4	89,92	2	86,54	5	106,62
5	106,06	4	85,40	5	97,73
4	116,00	4	94,63	3	105,64
5	93,05	3	89,82	4	110,29
4	107,77	4	105,25	4	111,10
4	99,07	4	105,43		

Estimer la moyenne par sous-unité et déterminer l'écart type de l'estimateur,

5.7 D'une population formée de 6 980 propriétés d'une petite ville, on tire avec remise un échantillon de 30 propriétés avec probabilités proportionnelles à la superficie de la propriété. La superficie de chaque terrain du quartier est connue, et la superficie totale est de 79 870 centaines de m^2. Pour chaque propriété, on détermine le montant y_i dépensé par année en engrais. Le but de l'échantillonnage est d'estimer la valeur totale τ de y pour la ville. Voici les données de l'échantillon (M_i = superficie du terrain en centaines de m^2, y_i = dépenses annuelles):

M_i	y_i	M_i	y_i	M_i	y_i
7	84,79	8	96,38	7	99,75
11	104,67	10	105,86	12	98,14
9	99,91	15	105,43	12	94,98
9	88,31	12	99,72	11	95,96
12	107,89	14	105,26	9	91,35
10	96,99	14	98,81	13	101,43
13	112,47	16	114,17	14	109,97
8	81,81	17	115,58	9	100,71
12	101,05	11	93,60	12	97,28
8	93,59	9	98,63	8	92,80

Estimer le total et déterminer l'écart type de l'estimateur.

5.8 D'une population de 99 987 comptes à recevoir, on tire avec remise un échantillon de 35 comptes avec probabilités proportionnelles à l'importance du compte. L'importance du compte est mesurée grossièrement par le nombre d'articles achetés, nombre connu pour chaque client. Le nombre total d'articles achetés est 8 765 432. Pour chaque compte, on détermine le montant dû y_i. Voici les données de l'échantillon (M_i = nombre d'articles achetés, y_i = solde du compte):

M_i	y_i	M_i	y_i	M_i	y_i
111	107,92	133	109,33	81	103,76
69	96,95	107	103,61	91	111,13
102	101,74	114	110,39	115	118,85
110	95,19	139	115,20	50	77,06
120	108,74	51	84,17	91	103,22
106	111,06	127	109,17	82	94,18
61	84,94	90	85,26	69	97,53
85	89,99	65	101,83	116	113,10
97	103,80	109	105,73	101	95,79
98	99,59	75	96,68	89	96,53
105	106,15	47	78,77	59	97,18
70	100,13	74	97,86		

a) Estimer le solde total de tous les comptes, ainsi que l'écart type de l'estimateur.
b) Estimer le solde moyen par compte, ainsi que l'écart type de l'estimateur.

5.9 Considérer la *population* suivante:

M_i	y_i	M_i	y_i	M_i	y_i
1 020	1 032	833	995	882	994
1 179	1 022	989	924	967	975
705	795	1 148	1 138	1 329	1 031
959	951	1 414	1 141	1 183	1 058
1 101	1 027	726	871	815	975
1 028	993	970	936	884	1 012
1 355	116	1 247	1 063	999	959
1 152	980	1 203	1 105	1 206	1 080
793	954	847	958	700	868
952	1 028	710	936	1366	114
860	881	896	915	918	922
315	753	1 260	1 204	658	881
1 041	1 016	1 346	1 112	859	939
464	820	1 184	1 075	990	1 021
820	959	1 269	1 111	1150	1 119
1 272	1 052	905	894	1 226	1 081
1 013	1 048	1 125	1 103		

On tire un échantillon de taille $n = 12$ de la population afin d'estimer la moyenne des y par unité primaire.

a) Calculer l'écart type de l'estimateur basé sur la moyenne arithmétique des unités primaires, dans un échantillon tiré sans remise.
b) Calculer l'écart type de l'estimateur basé sur le quotient, dans un échantillon tiré sans remise.
c) Calculer l'écart type de l'estimateur basé sur un échantillonnage avec remise et probabilités proportionnelles aux M_i.
d) Afin de prélever un échantillon aléatoire avec remise et probabilités proportionnelles aux M_i, on génère 12 nombres au hasard, distribués uniformément sur [0, 1). Les voici:

0,8179;	0,9582;	0,5374;	0,2147;	0,9122;	0,0352;
0,4716;	0,3984;	0,8281;	0,5636;	0,9054;	0,5195.

Quelles sont les unités primaires qu'on doit prélever selon ces nombres aléatoires? (Préservez l'ordre dans lequel les données sont présentées.)

5.10 Considérez la *population* suivante. Il s'agit d'une population de 28 factures pour lesquelles la variable *y* est le montant de la facture et *x* est le *nombre d'articles* achetés. Supposons qu'un vérificateur prélève un échantillon de taille 10 de cette population afin d'estimer la valeur totale τ des factures.

a) Déterminer la variance de l'estimateur $N\bar{y}$.

b) En supposant que le vérificateur est en mesure de déterminer la valeur de *x* pour chaque facture de la population, calculez approximativement la variance (*i*) de l'estimateur par la différence et (*ii*) de l'estimateur par le quotient.

c) Supposons qu'on prélève un échantillon stratifié selon la valeur de *x* (quatre strates, les factures appartenant à une même strate ayant même valeur de *x*). Déterminer la variance de l'estimateur du total pour une allocation à peu près proportionnelle.

Les données suivent:

x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
4	40	3	47	2	34	1	14
4	45	3	52	1	36	1	14
4	56	3	53	1	38	1	16
4	68	2	18	1	10	1	16
3	27	2	24	1	11	1	17
3	35	2	28	1	13	1	20
3	42	2	32	1	14	1	22

$\Sigma x = 57$, $\Sigma y = 842$, $\Sigma x^2 = 151$, $\Sigma y^2 = 32\,092$, $\Sigma xy = 2\,117$

Strate	*x=4*	*x=3*	*x=2*	*x=1*
Σy	209	256	136	241
Σy2	11 385	11 440	3 864	5 403
s^2	154,9167	103,4667	41,2	77,9359

5.11 Considérez la population des maisons présentée au tableau A.07. Supposons qu'une statisticienne s'apprête à prélever un échantillon de cette population afin d'estimer le prix moyen des maisons. Utilisez les données de la population elle-même pour comparer l'efficacité des procédures suivantes:

a) Un échantillon aléatoire simple de taille 50;

b) Un échantillon de taille 50 stratifié selon le secteur avec allocation proportionnelle;

c) Un échantillon de taille 50 stratifié selon le secteur avec allocation optimale;

Résumez vos conclusions.

5.12 Un statisticien tire un échantillon aléatoire simple de 2 localités dans la population décrite au tableau A.07. Il tombe sur Côte-St-Luc et Villeray.

a) Estimer le prix moyen des maisons de la population en utilisant un estimateur par la moyenne et déterminer un intervalle de confiance.

b) Estimer le prix moyen des maisons de la population en utilisant un estimateur par le quotient et déterminer un intervalle de confiance.

5.13 Un statisticien tire un échantillon aléatoire simple de 2 localités avec remise est probabilités proportionnelles aux tailles dans la population décrite au tableau A.07. Il tombe sur St-Henri Sud Ouest. et Ahuntsic. Estimer le prix moyen des maisons de la population et déterminer un intervalle de confiance.

5.14 Considérez la population de professeurs présentée au tableau A.01. Supposons qu'on ait prélevé un échantillon de deux départements, tiré avec probabilités proportionnelles aux tailles (nombre de professeurs) des départements, et qu'on soit tombé sur les départements 5 et 8. Estimez le salaire moyen de la population; estimez l'écart type de l'estimateur; déterminez un intervalle de confiance pour la moyenne de la population.

5.15 Supposons qu'on veuille estimer les recettes totales τ d'une population de 30 magasins à partir d'un échantillon de taille 12.

Magasin	*x*	*y*	*Magasin*	*x*	*y*	*Magasin*	*x*	*y*
1	908	1036	11	724	845	21	625	609
2	865	1037	12	721	745	22	623	830
3	865	951	13	708	907	23	623	725
4	853	963	14	705	833	24	615	624
5	849	926	15	675	897	25	593	724
6	830	787	16	670	762	26	588	683
7	820	945	17	668	740	27	588	816
8	772	865	18	664	840	28	542	732
9	765	878	19	653	694	29	495	414
10	758	792	20	628	682	30	494	602

Pour stratifier la population, on utilise les recettes *x* de l'an dernier: la première strate comprend les 10 magasins ayant les plus fortes recettes; la 2e strate comprend les 10 magasins ayant eu les plus fortes recettes parmi les 20 qui restent; la 3e strate, enfin, comprend les 10 magasins ayant les plus faibles recettes. On détermine ensuite la répartition optimale basée sur *les recettes de l'an dernier*. Considérer la population ci-dessus, pour lesquelles sont données non seulement les recettes x_i de l'an dernier, mais également les recettes y_i (normalement inconnues).Déterminer une stratification à trois strates égales, selon la valeur de *x*, puis

a) Déterminer l'allocation optimale à partir de la variable *x*
b) Déterminer l'écart type de $\bar{y}_{st}$ pour l'allocation déterminée en a)
c) Quelle est la répartition optimale basée sur les *y*?
d) Déterminer l'écart type de $\bar{y}_{st}$ basé sur la répartition proportionnelle. Comparer avec *b*)
e) Déterminer l'écart type d'un échantillon aléatoire simple de taille 12. Commenter la différence avec *b*) et *d*)
f) Quelle est la variance de l'estimateur par le quotient pour un échantillon aléatoire simple de taille 12, utilisant x comme variable auxiliaire?