



Avant de tirer l'échantillon on peut diviser la population en ensembles mutuellement exclusifs appelés strates et échantillonner chaque strate indépendamment. C'est ce qu'on appelle un plan de sondage stratifié. On peut créer des strates à des fins *administratives* ou à des fins *statistiques*. Dans les enquêtes canadiennes il faut produire des estimations provinciales avec une bonne précision. Il est donc avantageux de stratifier par province et de fixer un objectif de précision

provincial. En pratique ceci amène souvent à utiliser des fractions de sondage plus grandes dans les petites provinces. La stratification par province est un exemple de stratification à des fins administratives. En général si on doit produire une estimation pour un domaine particulier de la population il est intéressant de considérer ce domaine comme une strate et de fixer a priori sa taille d'échantillon. Dans un plan aléatoire simple sans remise, la taille d'échantillon pour un domaine est une variable aléatoire qui peut parfois prendre des petites valeurs.

Exemple: stratification à des fins statistiques



- La population est l'ensemble des producteurs de sirop d'érable d'une région du Québec;
- X =nb d'entailles que le producteur prévoit faire;
- Y =production de sirop d'érable en hectolitre;
- Objectif: obtenir une estimation de T_y afin de déterminer le prix de vente du sirop.

Dans des enquêtes de type « entreprise » les unités sont des établissements de tailles variables. Il y a donc de grosses unités qui contribuent de façon importante au total de la variable d'intérêt. La taille des établissements est souvent utilisée comme variable de stratification (X). On fait en sorte que la fraction de sondage augmente avec la taille des

unités. Parfois, pour les grosses unités, on utilise même des strates à tirage complet avec une fraction de sondage égale à 1. C'est une stratification à des fins statistiques qui cherche à obtenir un estimateur avec une petite variance pour une taille d'échantillon raisonnable.

Exemple: stratification à des fins statistiques

- Enquête mensuelle sur le commerce de détail de Statistique Canada;
- Stratification administrative par province et par secteur d'activités;
- Population = ensemble des entreprises dans une strate administrative (plus de 300 strates);
- X = vente annuelle estimée avec des données fiscales
- Y = vente mensuelle réelle;
- Objectif: Estimer les ventes totales dans une strate administrative.

Si on dispose d'une variable de stratification x associée à la variable d'intérêt y on peut former les strates à partir de bornes $b_0 = \min x < b_1 < \dots < b_H = \max x + I$ de la façon suivante. La strate h contient toutes les unités pour lesquelles $x \in [b_{h-1}, b_h)$. Le choix des bornes b_h a été beaucoup discuté dans la littérature.

Population et plan de sondage : Dans un plan stratifié aléatoire, la sélection des unités à l'intérieur des strates se fait selon un plan aléatoire simple sans remise. On utilise l'indice « h », $h=1, \dots, H$ pour représenter la strate et on note N_h , la taille de la strate h ; la taille de la population est $N=\sum N_h$. Les valeurs de la variable d'intérêt y dans la population sont $\{y_{hj} : h=1, \dots, H \text{ et } j=1, \dots, N_h\}$.

La moyenne et la variance de y dans la strate h sont

$$\bar{y}_{hU} = \sum_{j=1}^{N_h} y_{hj} / N_h \text{ et } S_h^2 = \sum_{j=1}^{N_h} (y_{hj} - \bar{y}_{hU})^2 / (N_h - 1)$$

Notons que $\bar{y}_U = \sum \frac{N_h}{N} \bar{y}_{hU}$.

Si on appelle n_h la taille d'échantillon dans la strate h , le plan de sondage s'écrit

$$p(S) = \frac{1}{\prod_{h=1}^H \binom{N_h}{n_h}} \text{ si } S \text{ contient exactement } n_h \text{ unités de la strate } h \text{ et } 0 \text{ sinon.}$$

Un plan de sondage stratifié est constitué de plans de sondage aléatoires simples indépendants dans les H strates. Ainsi les probabilités de sélection simple et conjointe dans la strate h sont :

$$\pi_{hi} = \frac{n_h}{N_h} \text{ et } \pi_{hij} = \frac{n_h(n_h - 1)}{N_h(N_h - 1)}$$

La probabilité de sélection conjointe d'une unité i de la strate h et d'une unité k de la strate ℓ est $\pi_h \times \pi_\ell$.

Échantillon : $y_{hj} : h = 1, \dots, H$ et $j \in \mathcal{S}_h$ $n = \sum n_h$ est la taille d'échantillon totale.

Estimateur de \bar{y}_U : $\bar{y}_{str} = \sum \frac{N_h}{N} \bar{y}_{hs}$ et $\bar{y}_{hs} = \sum_{\mathcal{S}_h} \frac{y_{hj}}{n_h}$ et $\text{Var}(\bar{y}_{str}) = \sum \frac{N_h^2}{N^2} \frac{1-f_h}{n_h} S_h^2$

où $f_h = n_h / N_h$ est la fraction de sondage dans la strate h . L'estimateur de variance

$$v(\bar{y}_{str}) = \sum \frac{N_h^2}{N^2} \frac{1-f_h}{n_h} S_h^2$$

où $s_h^2 = \sum_{j \in \mathcal{S}_h} (y_{hj} - \bar{y}_{hs})^2 / (n_h - 1)$ est la variance échantillonnale dans la strate h .

Le poids de sondage d'une unité i de la strate h est $w_{hi} = N_h / n_h$. On vérifie que l'estimation du total de y dans la population s'écrit $\hat{T}_y = \sum_{(h,i) \in \mathcal{S}} w_{hi} y_{hi}$. L'estimateur de variance s'écrit

$$v(\hat{T}_y) = \sum_{h=1}^H (1-f_h) \sum_{i \in \mathcal{S}_h} \frac{n_h}{n_h - 1} \left(w_{hi} y_{hi} - \sum_{j \in \mathcal{S}_h} \frac{w_{hj} y_{hj}}{n_h} \right)^2$$

Le théorème de la limite centrale s'applique dans chaque strate et un intervalle de confiance à $100(1-\alpha)\%$ pour le total de y dans la population est

$$\hat{T}_y \pm z_{1-\alpha/2} \sqrt{v(\hat{T}_y)}$$

On construit de la même manière un intervalle de confiance pour la moyenne.

Allocation

L'allocation est la méthode utilisée pour répartir les n unités de l'échantillon dans les H strates. Pour ce faire on utilise des fonctions a_h telles que $\sum a_h = 1$ et on prend $n_h = a_h n$ pour $h=1, \dots, H$. Voici quelques règles d'allocation que l'on retrouve en pratique :

1- Égale $a_h = 1/H$ pour $h=1, \dots, H$

2- Proportionnelle $a_h = N_h/N$ pour $h=1, \dots, H$

3-Puissance $a_h = (N_h \bar{y}_{hU})^p / \sum (N_\ell \bar{y}_{\ell U})^p$ pour $h=1, \dots, H$ où p est dans $(0,1)$

4-Optimale $a_h = (N_h S_h) / \sum (N_\ell S_\ell)$ pour $h=1, \dots, H$.

Sous allocation proportionnelle les poids d'échantillonnage sont de N/n pour toutes les unités de l'échantillon. Un tel plan de sondage est dit *autopondéré*.

Les règles d'allocation 3 et 4 font intervenir la variable y qui est inconnue. On peut utiliser ces méthodes en utilisant une variable de stratification x connue pour toutes les unités de population. On peut également utiliser les données d'une enquête passée pour calculer les tailles d'échantillon pour une nouvelle enquête.

Pourquoi stratifier à des fins statistiques?

Supposons que les fractions de sondage f_h sont à toutes fins pratiques nulles et comparons les variances des estimateurs de la moyenne obtenus selon un plan de sondage aléatoire simple et un plan stratifié sous allocation proportionnelle ($n_h = n \times N_h / N$).

En utilisant la décomposition de la somme de carrés en ANOVA on a :

$$\begin{aligned}(N-1)S^2 &= \sum_{h,i} (y_{hi} - \bar{y}_U)^2 = \sum_{h=1}^H N_h (\bar{y}_{hU} - \bar{y}_U)^2 + \sum_{h,i} (y_{hi} - \bar{y}_{hU})^2 \\ &= \sum_{h=1}^H N_h (\bar{y}_{hU} - \bar{y}_U)^2 + \sum_{h=1}^H (N_h - 1)S_h^2\end{aligned}$$

On peut réécrire la variance de \bar{y}_s , l'estimateur obtenu avec un plan aléatoire simple, comme

$$\text{Var}(\bar{y}_s) = \frac{S^2}{n} = \frac{1}{n} \left\{ \sum_{h=1}^H \frac{N_h (\bar{y}_{hU} - \bar{y}_U)^2}{N-1} + \sum_{h=1}^H \frac{(N_h - 1)S_h^2}{N-1} \right\} \approx \frac{1}{n} \sum_{h=1}^H \frac{N_h (\bar{y}_{hU} - \bar{y}_U)^2}{N-1} + \frac{1}{n} \sum_{h=1}^H \frac{N_h S_h^2}{N}$$

La variance de l'estimateur obtenu avec un plan stratifié sous allocation proportionnelle est

$$\text{Var}(\bar{y}_{str}) = \sum \frac{N_h^2}{N^2} \frac{S_h^2}{n_h} = \frac{1}{n} \sum \frac{N_h}{N} S_h^2 \text{ et}$$

Donc $\text{Var}(\bar{y}_{str}) \leq \text{Var}(\bar{y}_s)$; le plan stratifié est beaucoup plus précis qu'un plan aléatoire simple lorsque les différences de moyennes entre les strates contribuent beaucoup à la variabilité totale de y .

Allocation optimale (ou de Neyman)

Si on connaît les variances $\{S_h^2\}$ de y dans les H strates quelle est l'allocation $\{a_h : \sum a_h = 1\}$, qui minimise la variance de \bar{y}_{str} ?

Réponse : il faut prendre l'allocation optimale, $a_h = (N_h S_h) / \sum (N_l S_l)$.

Démonstration : On va utiliser l'inégalité de Cauchy-Schwarz (CS) qui dit que si $\{a_i\}$ et $\{b_i\}$

sont deux ensembles de n nombres alors $\sum_{i=1}^n a_i b_i \leq \sqrt{\sum_{i=1}^n a_i^2} \times \sqrt{\sum_{i=1}^n b_i^2}$ avec égalité si $a_i = c b_i$ pour

tout i où c est une constante.

On a $\text{Var}(\bar{y}_{str}) = \frac{1}{nN^2} \sum N_h^2 \frac{S_h^2}{a_h} - \sum \frac{N_h}{N^2} S_h^2$ en utilisant CS on obtient la borne inférieure

suivante pour le premier terme,

$$\frac{1}{nN^2} \left[\sqrt{\sum N_h^2 \frac{S_h^2}{a_h}} \right]^2 = \frac{1}{nN^2} \left[\sqrt{\sum N_h^2 \frac{S_h^2}{a_h}} \times \sqrt{\sum a_h} \right]^2 \geq \frac{1}{nN^2} \sum N_h S_h.$$

avec égalité si $N_h \frac{S_h}{\sqrt{a_h}} = c \sqrt{a_h}$. Ainsi la variance est minimale pour l'allocation optimale

$a_h = \frac{N_h S_h}{\sum N_l S_l}$. Si toutes les variances sont égales l'allocation proportionnelle est optimale.

En général il faut suréchantillonner les strates dans lesquelles y est plus variable.

EXEMPLE 1 : Pour déterminer le pourcentage d'employés favorables à une nouvelle politique, une compagnie a tiré un échantillon stratifié proportionnel de taille $n=300$ auprès des employés de ses 3 usines. Estimer la proportion des employés favorables et sa variance. Dire si un plan aléatoire simple sans remise aurait été acceptable pour cette enquête.

Tableau 1 : Organisation des calculs

Usine	N_h	N_h/N	n_h	# favor.	p_h
1	1490	0.509	153	75	0.490
2	987	0.337	101	40	0.396
3	453	0.155	46	10	0.217
tot	2930		300		
Estimation de la proportion de succès				0.416	
Estimation de l'erreur-type				0.02658	
Erreur-type sous un plan aléatoire simple					0.02701
Allocation optimale			Erreur-type minimale		
strate	n_h		0.02653		
1	158				
2	102				
3	40				

$$\hat{p}_{str} = \sum_{h=1}^3 \frac{N_h}{N} p_h$$

$$= 0.509 \times 0.49 + \dots + 0.155 \times 0.217$$

$$= 0.416$$

Formule de variance

$$\text{Var}(\hat{p}_{str}) = \sum_{h=1}^3 \frac{N_h^2}{N^2} (1 - f_h) \frac{p_h(1 - p_h)}{n_h - 1}$$

Estimation de la variance sous un plan aléatoire simple

$$\text{Var}(\hat{p}) = \frac{1 - f}{n - 1} \hat{p}(1 - \hat{p})$$

Allocation optimale :

$$n_h = n \frac{N_h ((p_h(1 - p_h)))^{1/2}}{\sum N_l ((p_l(1 - p_l)))^{1/2}}$$

Conclusion: Pour des variables dichotomiques les gains de précision obtenus avec un plan stratifié sont souvent faibles.

EXEMPLE 2 : Un échantillon stratifié proportionnel de 250 producteurs bovins a été tiré parmi les 1456 d'une liste fournie par un syndicat de producteurs; y = nombre de têtes de bétail (en milliers). Estimer la taille moyenne d'un troupeau et sa variance. Peut-on augmenter la précision en changeant la règle d'allocation?

Tableau 2 : Estimations et erreurs types pour différents scénarios d'échantillonnage.

strate	N_h	N_h/N	n_h	\bar{y}_{hs}	s_h^2
1	625	0.43	107	7.50	13.45
2	418	0.29	72	15.60	75.32
3	255	0.18	44	29.33	230.45
4	158	0.11	27	55.90	898.45
tot	1456		250		
Estimation de la taille moy.			18.90	et du total	27 520
Estimation de l'erreur-type			0.74	et du CV	3.9%

Notez que la taille moyenne du troupeau croît avec la strate. La variance est aussi croissante. L'allocation proportionnelle est loin d'être optimale!

L'estimation de la taille moyenne d'un troupeau est :

$$\bar{y}_{str} = \sum_h \frac{N_h}{N} \bar{y}_h = 0.43 \times 7.50 + \dots + 0.11 \times 55.90 = 18.90$$

L'estimation de variances est

$$v(\bar{y}_{str}) = \sum_h \left(\frac{N_h}{N} \right)^2 \frac{1-f_h}{n_h} s_h^2 = 0.43^2 \times \frac{(1-.17)}{107} \times 13.45 + \dots + 0.11^2 \times \frac{(1-.17)}{27} \times 898.45 = 0.74^2$$

Le CV est de $0.74/18.90=3.9\%$.

On va déterminer les tailles d'échantillon pour une prochaine enquête. Les variances changent beaucoup d'une strate à l'autre et une allocation optimale pourrait permettre des gains de précision important. L'allocation optimale pour la strate 4 est de

$$n_h^a = 250 \frac{158 \times \sqrt{898.45}}{625 \times \sqrt{13.45} + 418 \times \sqrt{75.32} + 255 \times \sqrt{230.45} + 158 \times \sqrt{898.45}} = 250 \times 0.33 = 82,$$

par rapport à l'allocation proportionnelle qui est de 27.

Tableau 3 : Calcul de tailles d'échantillon selon deux méthodes d'allocation.

Estimation d'erreur-type sous 2 scénarios d'échantillonnage					
Allocation optimale			All. puissance avec p=.7		
strate	n_h^a	E-T min	strate	n_h^a	E-T min
1	39	0.5336	1	48	0.5352
2	62		2	61	
3	67		3	67	
4	82		4	75	
total	250		total	251	

L'allocation optimale permet de faire des gains de précision importants. Attention, les tailles n_h^a obtenues sont sujettes à des erreurs d'échantillonnage. L'allocation de puissance donne des résultats semblables; puisque les n_h^a dépendent seulement des moyennes on peut présumer que les erreurs d'échantillonnage sont plus faibles que pour l'allocation optimale.

POST-STRATIFICATION

Le tableau 4 donne la répartition âge sexe pour la région métropolitaine de Québec. On peut utiliser cette information pour traiter les données d'une enquête faite dans la région de Québec

Tableau 4. Répartition âge-sexe pour la région métropolitaine de Québec selon le site web de Statistique Canada.

Figure Caractéristiques selon l'âge	Québec (RMR)			Québec		
	Total	Sexe masculin	Sexe féminin	Total	Sexe masculin	Sexe féminin
Population totale ⁴	715 515	345 075	370 440	7 546 130	3 687 695	3 858 440
0 à 4 ans	32 875	16 720	16 160	375 270	191 565	183 710
5 à 9 ans	34 155	17 480	16 680	398 980	203 985	195 000
10 à 14 ans	41 310	20 875	20 435	478 255	243 595	234 655
15 à 19 ans	41 530	20 855	20 675	475 005	242 185	232 820
20 à 24 ans	48 805	24 380	24 425	472 170	238 440	233 730
25 à 29 ans	52 020	26 245	25 780	492 870	245 335	247 540
30 à 34 ans	44 180	22 460	21 715	467 325	232 800	234 525
35 à 39 ans	44 990	22 565	22 430	502 300	250 340	251 960
40 à 44 ans	56 615	28 005	28 605	619 120	308 570	310 550
45 à 49 ans	60 820	29 670	31 145	644 040	318 145	325 895
50 à 54 ans	58 415	27 935	30 485	588 085	289 780	298 300
55 à 59 ans	53 520	25 900	27 620	524 350	257 790	266 560
60 à 64 ans	43 230	20 640	22 590	428 070	208 805	219 270
65 à 69 ans	29 890	13 945	15 950	315 560	150 165	165 395
70 à 74 ans	25 100	10 985	14 120	268 145	121 940	146 205
75 à 79 ans	20 550	8 130	12 415	220 530	92 485	128 045
80 à 84 ans	15 240	5 335	9 905	156 775	58 075	98 695
85 ans et plus	12 270	2 965	9 305	119 285	33 695	85 585

Mise en situation : Une enquête a été réalisée auprès des résidents (agés de 15 ans et plus) de la région de Québec. On a noté l'âge et le sexe des répondants qui ont été répartis en 10 groupes selon le sexe (H ou F) et l'âge (15-24, 25-44, 45-64, 65-74, 75 et +).

Grâce aux données du recensement de Statistique Canada, on connaît la taille de ces 10 groupes ou strates. Par exemple chez les 15-24 il y a 45 100 femmes et 45 155 hommes. Les tailles d'échantillon n_h dans ces 10 strates n'ont pas été fixées a priori. Peut-on quand même traiter les données comme si elles venaient d'un plan stratifié, en considérant que les tailles d'échantillon dans les strates sont fixes?

La réponse est oui, on peut faire comme si le plan était stratifié.

Même si les tailles d'échantillons sont aléatoires, on peut conditionner sur les valeurs obtenues et le plan d'échantillonnage devient alors un plan stratifié où l'allocation est proche de l'allocation proportionnelle. Formellement si S est un échantillon tiré selon un plan aléatoire simple et si $\{n_h : h=1, \dots, H\}$ sont les tailles d'échantillon observées dans les H strates on a le résultat suivant :

$$p(S | n_1, \dots, n_H) = \frac{1}{\prod_{h=1}^H \binom{N_h}{n_h}} \quad \text{si } S \text{ contient exactement } n_h \text{ unités de la strate } h \text{ et } 0 \text{ sinon.}$$

Si on conditionne sur les tailles d'échantillon observées dans chacune des strates on obtient un plan stratifié standard et les formules d'estimation pour un plan stratifié s'appliquent.

NON-RÉPONSE

Il y a non-réponse lorsque certaines des personnes de l'échantillon ne remplissent pas le questionnaire. Elles refusent peut-être de répondre aux questions pour des motifs personnels. La non-réponse peut également être causée par une incapacité à rejoindre les personnes de l'échantillon; malgré de nombreux appels l'interviewer n'arrive pas à entrer en contact avec des personnes de l'échantillon. Ce type de non-réponse est dit *complète* ou *totale* car la personne échantillonnée ne fournit aucune information pour l'enquête. On parle de non-réponse *partielle* lorsque le répondant fournit des réponses à seulement une partie du questionnaire. On s'intéresse ici à la non-réponse totale.

On dit que la non-réponse est *ignorable* si la probabilité de répondre est la même pour toutes les unités de la population. Si la non-réponse est ignorable, un ensemble de n_r répondants obtenu à l'aide d'un échantillon aléatoire simple de personnes a les mêmes caractéristiques qu'un échantillon aléatoire simple de n_r personnes. Il permet de calculer des estimations non biaisées des caractéristiques de la population.

Lorsque la probabilité de répondre varie dans la population on parle de non-réponse *non-ignorable*. La non-réponse est non ignorable lorsque la proportion de jeunes dans l'échantillon de répondants est, par exemple, inférieure à celle de la population. Les jeunes sont plus difficiles à joindre. Ils sont souvent sous-représentés dans des échantillons. Lorsque la non-réponse est non-ignorable l'échantillon des répondants n'est pas représentatif

de la population. Calculer des estimations à partir des seuls répondants risque de donner des résultats biaisés. La post-stratification permet de corriger certains biais associés à une non-réponse non ignorable. En pratique la non-réponse est souvent non-ignorable et il faut en tenir compte lors du traitement des données.

EXEMPLE : CORRECTION DU BIAIS DE NON-RÉPONSE GRÂCE À LA POST-STRATIFICATION

Dans une enquête réalisée en 1998 auprès des résidents de la région de Québec de 15 ans et plus on s'intéresse à la variable y , le nombre d'années de scolarité (données fictives). On dispose de 108 répondants qui ont été classés selon les catégories âge-sexe dans le tableau 5 et des données de recensement données au tableau 6.

Tableau 5. Nombre de répondants (n_h) et leur scolarité moyenne (\bar{y}_h) par catégorie âge-sexe.

(n_h, \bar{y}_h)	Classe d'âge				
Sexe	15 - 24	25 - 44	45 - 64	65 - 74	75 +
Homme	(10,10)	(15,12)	(10,11)	(8,9)	(7,7)
Femme	(8,11)	(19,14)	(15,12)	(9,9)	(8,6)

Tableau 6. Données du recensement pour la région de Québec.

N_h	15 - 24	25 - 44	45 - 64	65 - 74	75 +
Hommes	47170	110735	76430	19510	9875
Femmes	46745	112675	82095	26940	21855
tot	554030				

Chez les 25-44 le taux de participation est de $34/(110735+112675)=0.000152187$ alors que chez les 75+ il est de $15/(9875+21855)=0.000472739$, environ quatre fois plus élevé. Il semble donc que l'échantillon est biaisé car les jeunes sont sous représentés. La non-réponse est donc non-ignorable.

L'estimation du nombre moyen d'années de scolarité obtenue en traitant les données selon un plan aléatoire simple est $\bar{y}_s = \sum n_h \bar{y}_h / \sum n_h = 10.77$. Puisque les jeunes sont sous-représentées et qu'ils sont plus scolarisées, cette estimation est sans doute biaisée.

On post-stratifie avec les données de Statistique Canada et on obtient l'estimation suivante :

$$\bar{y}_{str} = \sum N_h \bar{y}_h / \sum N_h = 11.44.$$

une augmentation de 6%. La post-stratification permet de corriger un biais associé à une sous-représentation des jeunes. D'autres biais, par exemple celui associé à une sous-représentation des immigrants, peuvent encore être présent après avoir post-stratifié.

On peut calculer une estimation de variance à l'aide des variances intra-strate suivantes

var	15 - 24	25 - 44	45 - 64	65 - 74	75 +
Hommes	9	11	13	8	7
Femmes	12	12	9	6	3

On obtient

$$v(\bar{y}_{str}) = \sum N_h^2 s_h^2 / (N^2 n_h) = 0.338^2 .$$

Ainsi un intervalle de confiance à 95% pour la scolarité moyenne est

$$11.44 \pm 1.96 \times 0.338 = (10.78, 12.10)$$

L'estimation brute sans la post-stratification de 10.77 est sans doute biaisée car elle n'est pas dans l'intervalle de confiance.

EXEMPLE : Extrait du sondage discuté au Ch.2 paru dans le Soleil le 30 décembre 2010 sur l'artiste préféré pour le festival d'été.



La post-stratification selon les données âge-sexe de Statistique Canada est une bonne pratique statistique qui est couramment utilisée.

Dans l'extrait du Soleil ci-contre, la post-stratification est utilisée pour faire la compilation des données. Noter que la marge d'erreur est calculée comme dans un plan aléatoire simple cependant. En effet, $1.96 / (2 \times 425^{1/2}) = 4.75\%$.