Chapitre 6 - Échantillonnage par grappes

47. Un producteur dispose ses conserves de soupe dans des boîtes contenant 24 conserves, en suivant l'ordre dans lequel elles sont produites. Le poids de l'emballage est une caractéristique essentielle: s'il est trop faible, le producteur enfreint la loi sur les poids et mesures et s'expose donc à des poursuites; s'il est trop élevé, le producteur encourt des frais supplémentaires (pour la soupe additionnelle et pour les difficultés à placer les couvercles sur les conserves). Le contrôle de qualité de la chaîne de production consiste en un EAS de n boîtes; les 24 conserves de chaque boîte choisie sont ensuite ouvertes et le poids de la soupe (en grammes) dans chaque conserve est mesuré. Les données résultantes sont présentées ci-dessous, dans l'ordre dans lequel les cartons ont été retirés de la chaîne de production:

Boîte	**Poids moyen**	**Écart-type**	**Boîte**	**Poids moyen**	**Écart-type**
1	340.6	0.27	6	340.5	0.61
2	340.8	0.39	7	340.2	0.34
3	340.5	0.24	8	340.3	0.50
4	340.1	0.43	9	340.0	0.22
5	340.8	0.34	10	339.7	0.44

(a) Déterminer un I.C. (à environ 95%) pour le poids moyen d'une conserve de soupe.

(b) On suggère qu'une alternative à la prise d'un EAS de boîtes et à l'examen de toutes les conserves dans les boîtes choisies aurait été de sélectionner un EAS de conserves. Discuter brièvement des avantages et des inconvénients de cette suggestion.

(c) À l'aide d'un diagramme de dispersion des poids moyens des conserves dans une boîte en fonction de l'ordre dans lequel les boîtes sont choisies, discuter brièvement de l'état du contrôle statistique du processus de remplissage des conserves.

(d) Si l'on utilise l'estimateur $\overline{y}_G$, combien de boîtes devraient être échantillonnées afin de donner un estimé du poids moyen de la soupe dans toutes les conserves du cycle de production avec une marge d'erreur sur l'estimation de 0.2g? [Il faudra d'abord dériver une formule pour déterminer la taille de l'échantillon.]

Solution:

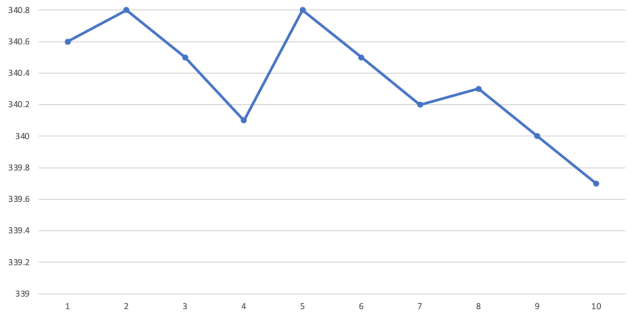
(a) L'intervalle de confiance de μ à environ 95% est

$$\overline{y}_G \pm 2\sqrt{\hat{V}\left(\overline{y}_G\right)} \approx \frac{1}{10}\sum_{i=1}^{10}\overline{y}_i \pm 2\sqrt{\frac{s_G^2}{10}}, \quad s_G^2 = \frac{\sum \overline{y}_i^2 - 10\overline{y}_G^2}{9}$$

On peut calculer que $\overline{y}_G = 340.35$, $s_G^2 = 0.1272$, et $B \approx 0.2256$, d'où l'intervalle de confiance est 340.35 ± 0.2256.

(b) Du point de vue physique, l'approche par grappes est beaucoup plus rapide que l'approche par échantillon aléatoire simple. Cependant, les boîtes de soupe peuvent avoir tendance à être plus homogènes au sein d'un carton que dans l'ensemble de la population (voir le diagramme de dispersion de la partie (c)), de sorte que l'échantillonnage aléatoire simple donnerait une estimation plus efficace. Mais cela ne semble pas du tout pratique, et peut-être même inutile, puisque nous ne savons pas ce qu'il advient des autres boîtes de soupe qui appartenaient à un carton contenant une unité sélectionnée dans un EAS.

(c) Voici le graphique du poids moyenne en fonction de l'ordre dans lequel les boîtes sont choisies.



Remarquez la nette tendance à la baisse dans la queue à la droite. Comme l'échantillon de cartons a été choisi au hasard, il semble bien que quelque chose ne tourne pas rond dans la machine.

(d) Nous ne connaissons pas le nombre de boîtes N; on suppose alors que $\frac{N-n}{N-1} \approx 1$. La marge d'erreur sur l'estimation dans ce cas est

$$B = 2\sqrt{\frac{\sigma_G^2}{n}} \implies n = 4\frac{\sigma_G^2}{B^2}.$$

Dans notre cas, nous allons utiliser $s_G^2 = 0.1272$ et $B = 0.2$, d'où $n \approx 4 \cdot \frac{0.1272}{(0.2)^2} = 12.72$; 13 boîtes devraient suffire. ■

48. Considèrons une population répartie en M grappes, toutes de taille n. La variable d'intérêt est Y. Un EAS de m grappes est prélevé; soit $\overline{y}_G$ l'estimateur EPG de la moyenne de population μ_Y. Pour $1 \leq j \leq M$, posons σ_j^2 la variance de Y dans la grappe j (par rapport à la moyenne de grappe μ_j). Soient $\overline{\sigma^2}$ la moyenne des σ_j^2, et σ^2 la variance de Y dans la population (par rapport à la moyenne μ_Y). Si M est suffisamment grand, montrer que

$$V(\overline{y}_G) \approx \frac{\sigma^2 - \overline{\sigma^2}}{m}\Big(1 - \frac{m}{M}\Big).$$

Indice: commencer par montrer que

$$\sigma^2 = \frac{1}{Mn}\sum_{j=1}^{M}\sum_{k=1}^{n}(Y_{j,k}-\mu)^2 = \frac{1}{Mn}\left\{\sum_{j=1}^{M}\sum_{k=1}^{n}(Y_{j,k}-\mu_j)^2 + n\sum_{j=1}^{M}(\mu_j-\mu)^2\right\}.$$

Démonstration: Pour tout $1 \leq j \leq M$, soit $\mu_j = \frac{1}{n}(Y_{j,1} + \cdots + Y_{j,n})$. On utilise l'indice:

$$\begin{aligned}
\sum_{j=1}^{M}\sum_{k=1}^{n}(Y_{j,k}-\mu)^2 &= \sum_{j=1}^{M}\sum_{k=1}^{n}(Y_{j,k}-\mu_j+\mu_j-\mu)^2 \\
&= \sum_{j=1}^{M}\sum_{k=1}^{n}\left\{(Y_{j,k}-\mu_j)^2 + 2(Y_{j,k}-\mu_j)(\mu_j-\mu) + (\mu_j-\mu)^2\right\} \\
&= \sum_{j=1}^{M}\sum_{k=1}^{n}(Y_{j,k}-\mu_j)^2 + 2\sum_{j=1}^{M}\sum_{k=1}^{n}(Y_{j,k}-\mu_j)(\mu_j-\mu) + \sum_{j=1}^{M}\sum_{k=1}^{n}(\mu_j-\mu)^2 \\
&= \sum_{j=1}^{M}\sum_{k=1}^{n}(Y_{j,k}-\mu_j)^2 + 2\sum_{j=1}^{M}(\mu_j-\mu)\left\{\sum_{k=1}^{n}(Y_{j,k}-\mu_j)\right\} + \sum_{k=1}^{n}\sum_{j=1}^{M}(\mu_j-\mu)^2 \\
&= \sum_{j=1}^{M}\sum_{k=1}^{n}(Y_{j,k}-\mu_j)^2 + 2\sum_{j=1}^{M}(\mu_j-\mu)\left\{\sum_{k=1}^{n}Y_{j,k} - n\mu_j\right\} + n\sum_{j=1}^{M}(\mu_j-\mu)^2 \\
&= \sum_{j=1}^{M}\sum_{k=1}^{n}(Y_{j,k}-\mu_j)^2 + 2\sum_{j=1}^{M}(\mu_j-\mu)\underbrace{\{n\mu_j - n\mu_j\}}_{=0} + n\sum_{j=1}^{M}(\mu_j-\mu)^2.
\end{aligned}$$

Ainsi,

$$\begin{aligned}
\sigma^2 &= \frac{1}{Mn}\sum_{j=1}^{M}\sum_{k=1}^{n}(Y_{j,k}-\mu)^2 = \frac{1}{Mn}\left\{\sum_{j=1}^{M}\sum_{k=1}^{n}(Y_{j,k}-\mu_j)^2 + n\sum_{j=1}^{M}(\mu_j-\mu)^2\right\} \\
&= \frac{1}{M}\sum_{j=1}^{M}\underbrace{\left\{\frac{1}{n}\sum_{k=1}^{n}(Y_{j,k}-\mu_j)^2\right\}}_{=\sigma_j^2} + \frac{1}{M}\sum_{j=1}^{M}(\mu_j-\mu)^2 = \underbrace{\frac{1}{M}\sum_{j=1}^{M}\sigma_j^2}_{=\overline{\sigma^2}} + \frac{1}{M}\sum_{j=1}^{M}(\mu_j-\mu)^2 = \overline{\sigma^2} + \frac{1}{M}\sum_{j=1}^{M}(\mu_j-\mu)^2.
\end{aligned}$$

Mais nous avons déjà vu que

$$V(\overline{y}_G) = \frac{\sigma_G^2}{m}\left(\frac{M-m}{M-1}\right) = \frac{1}{m}\cdot\frac{1}{M}\sum_{j=1}^{M}(\mu_j-\mu)^2\left(\frac{M-m}{M-1}\right) \approx \frac{1}{m}\cdot\frac{1}{M}\sum_{j=1}^{M}(\mu_j-\mu)^2\left(1-\frac{m}{M}\right),$$

d'où

$$V(\overline{y}_G) \approx \frac{\sigma^2 - \overline{\sigma^2}}{m}\left(1-\frac{m}{M}\right),$$

ce qui termine la démonstration. ■

49. Une entreprise souhaite donner un estimé du montant total des comptes débiteurs dûs par ses clients. Ces clients, ainsi que le montant qu'ils doivent, sont répertoriés par ordre alphabétique dans un grand livre de 5001 pages. Chaque page comporte 40 noms différents, à l'exception de la dernière, qui n'en comporte que 3, et qui est donc exclue de la base de sondage; par conséquent, on considère qu'il n'y a que $N = 200,000$ clients. On utilise un EPG afin de donner un estimé du montant total des comptes à recevoir, une grappe étant définie comme une paire de pages se faisant face. Ainsi, chaque grappe contient 80 noms. Un EAS de dix grappes a été sélectionné au hasard, et le montant moyen dû (en dollars) pour les 80 clients de chaque grappe a été déterminé. Donner des I.C. pour la moyenne et pour le montant total dû par les clients pour l'échantillon suivant, à 95% près.

Grappe	**Somme dûe (moyenne)**	**Grappe**	**Somme dûe (moyenne)**
1	174	6	157
2	162	7	132
3	141	8	169
4	129	9	155
5	138	10	163

Solution: Nous avons: $m = 10$, $M = 5000/2$, $n = 80$, $N = 200000$. Les sommes intermédiaires sont

$$\sum_{i=1}^{10} \overline{m}_i = 1520 \quad \text{et} \quad \sum_{i=1}^{10} \overline{y}_i^2 = 233314.$$

L'estimateur $\overline{y}_G$ est alors

$$\overline{y}_G = \frac{1}{m}\sum_{i=1}^{m} \overline{y}_i = \frac{1520}{10} = 152,$$

et

$$s_G^2 = \frac{1}{m-1}\left(\sum_{i=1}^{m} \overline{y}_i^2 - m\overline{y}_G^2\right) = \frac{1}{9}(233314 - 10(152)^2) = 252.67.$$

La variance d'échantillonnage de $\overline{y}_G$ est ainsi

$$\hat{V}(\overline{y}_G) = \frac{s_G^2}{m}\left(1 - \frac{m}{M}\right) = \frac{252.67}{10}\left(1 - \frac{10}{2500}\right) = 25.17,$$

d'où $2\sqrt{\hat{V}(\overline{y}_G)} = 10.03$. Les intervalles de confiance à environ 95% pour la moyenne et le total sont ainsi:

$$152 \pm 10.03 \equiv (141.97, 162.03) \quad \text{et} \quad 200000(152 \pm 10.03) \equiv (28.39M, 32.41M),$$

respectivement. ■

50. Les responsables d'un parc souhaitent connaître le nombre total de visiteurs annuels. Un échantillon de 5 semaines a été choisi au hasard, et le nombre de visiteurs quotidiens a été répertorié pour chacun des jours. Les observations sont présentées dans le tableau ci-dessous.

Semaine i	**Lun**	**Mar**	**Mer**	**Jeu**	**Ven**	**Sam**	**Dim**
1	208	194	125	130	180	200	310
2	130	120	123	105	111	111	113
3	200	150	130	190	177	150	140
4	114	132	107	121	130	160	170
5	200	107	101	98	103	111	137

Déterminer des I.C. pour le nombre moyen de visiteurs quotidiens et le nombre total de visiteurs annuels, à environ 95%. Combien de quartiers devrait-on prélevé afin d'obtenir une marge d'erreur sur l'estimation $B = 5$ et $B = 2000$, respectivement, pour la moyenne quotidienne et le total annuel?

Solution: Ici, nous utilisons $m = 5$, $n = 7$, et $M = 52$. La somme des visites quotidiennes pour les 35 journées de l'échantillon est 5088, d'où

$$\overline{y}_G = \frac{5088}{5(7)} = 145.3714.$$

La moyenne dans chacune des grappes est

$$\overline{y}_1 = 192.4286,\ \overline{y}_2 = 116.1429,\ \overline{y}_3 = 162.4286,\ \overline{y}_4 = 133.4286,\ \overline{y}_5 = 122.4286.$$

La variance empirique de ces moyennes de grappes est ainsi

$$s_G^2 = \frac{1}{5-1}\sum_{k=1}^{5}(\overline{y}_k - \overline{y}_G)^2 = 1007.159,$$

et la variance d'echantillonnage de $\overline{y}_G$ est

$$\hat{V}(\overline{y}_G) = \frac{s_G^2}{m}\left(1 - \frac{m}{M}\right) = \frac{1007.159}{5}\left(1 - \frac{5}{52}\right) = 182.0634,$$

d'où la marge d'erreur sur l'estimation est $B = 2\sqrt{\hat{V}(\overline{y}_G)} = 2\sqrt{182.0634} = 26.98617$. Les intervalles de confiance à environ 95% pour la moyenne quotidienne et le total annuel sont:

$$\mu : 145.4 \pm 27.0 \equiv (118.3853, 172.3576) \quad \text{et} \quad \tau : 7(52)(145.4 \pm 27.0) \equiv (43092.23, 62738.17).$$

Si on cherche à approximer μ avec une marge d'erreur sur l'estimation de $B = 5$, on utilise $\sigma_G^2 = 1007.159$ et on obtient

$$m = \frac{M\sigma_G^2}{(M-1)B^2/4 + \sigma_G^2} = \frac{52(1007.159)}{(52-1)5^2/4 + 1007.159} = 39.49914 \approx 40;$$

pour le total et $B = 2000$, cela devient

$$m = \frac{M\sigma_E^2}{(M-1)B^2/(4N^2) + \sigma_E^2} = \frac{52 \cdot 7^2(1007.159)}{(52-1)2000^2/(4 \cdot 52^2) + 7^2 \cdot 1007.159} = 37.6217 \approx 38.$$

Un échantillon de grappes de taille $m = 40$ ferait l'affaire pour les deux. ■

51. Une chaîne de télévision locale souhaite donner un estimé de la proportion d'électeurs favorables à la candidate A lors d'une élection municipale. Il s'avère trop coûteux de sélectionner et d'interviewer un EAS d'électeurs, c'est pourquoi la chaîne a opté pour un EPG, en utilisant les quartiers comme grappes. Un EAS de 9 quartiers est sélectionné parmi les 503 circonscriptions de la ville. La chaîne de télévision souhaite réaliser l'estimation le jour de l'élection, mais avant que les résultats finaux ne soient comptabilisés. Des reporters sont alors envoyés dans les bureaux de vote de chaque quartier sélectionné afin obtenir les informations pertinentes, présentées ci-dessous.

Quartier i	1	2	3	4	5	6	7	8	9
# Électeurs x_i	1290	1171	1170	1066	840	843	1893	971	1942
Favorisant la candidate y_i	680	596	631	487	475	321	1143	542	1187

Déterminer un I.C. pour la proportion d'électeurs de la ville qui favorisent le candidat A, à environ 95%. Combien de quartiers devrait-on prélevé afin d'obtenir une marge d'erreur sur l'estimation $B = 0.03$?

Solution: On a $m = 9$, $M = 503$, N **inconnu**, et

$$\sum_{i=1}^{9} y_i = 6062, \quad \sum_{i=1}^{9} x_i = 11186, \quad \sum_{i=1}^{9} y_i^2 = 4790794, \quad \sum_{i=1}^{9} x_i y_i = 8497266, \quad \sum_{i=1}^{9} x_i^2 = 15254500,$$

d'où

$$\hat{p}_G = \frac{6062}{11186} = 0.542, \quad \overline{n} = \frac{11186}{9} = 1242.89, \quad s_p^2 = \frac{1}{9-1}\sum_{i=1}^{9}(y_i^2 + \hat{p}_G x_i^2 - 2\hat{p}_G x_i y_i) = 7626.75;$$

la variance d'échantillonnage est

$$\hat{V}(\hat{p}_G) = \frac{1}{\overline{n}^2} \cdot \frac{s_P^2}{m}\left(1 - \frac{m}{M}\right) = \frac{1}{1242.89^2} \cdot \frac{7262,75}{9}\left(1 - \frac{9}{503}\right) = 0.0005387548.$$

L'erreur est ainsi $B = 2\sqrt{0.0005387548} = 0.04642218$, et l'intervalle de confiance recherché est $0.542 \pm 0.046 \equiv (0.496, 0.588)$.

Pour obtenir une marge d'erreur sur l'estimation $B = 0.03$, on suppose que $\sigma_P^2 = 7626.75$. Posons $D = B^2\overline{n}^2/4 = (0.03)^2(1242.80)^2/4 = 347.57$ (puisque N est inconnu, on ne peut pas déterminer $\overline{N}$), d'où

$$m = \frac{M\sigma_P^2}{(M-1)D + \sigma_P^2} = \frac{503(7626.75)}{(503-1)(347.57) + 7626.75} = 21.06595 \approx 22$$

serait requis pour obtenir la marge d'erreur requise. ■

56. Une grande entreprise est divisée en 11 départements. Le nombre d'employés dans chaque département est indiqué ci-dessous:

Département	A	B	C	D	E	F	G	H	I	J	K
# Employés	230	110	25	322	17	65	63	210	77	12	45

Dans le cadre d'une enquête d'opinion réalisée auprès des employés, on utilise un EPG afin d'étudier des départements entiers. Un EAS de $m = 5$ départements est sélectionné. On s'intéresse notamment à l'opinion des employés sur la façon dont la direction communique ses objectifs. Cette réponse est mesurée pour chaque employé en combinant les scores de trois questions, chacune mesurée selon une échelle de Likert à 5 niveaux. Plus les scores sont élevés, plus l'évaluation de l'employé sur la façon dont la direction communique ses objectifs est positive. Les données ci-dessous sont des résumés pour chaque département sélectionné.

Département	A	E	F	H	I
Moyenne Likert	3.62	4.24	4.07	3.36	3.81

(a) Préparer un diagramme de dispersion du score moyen en fonction du nombre d'employés dans chaque département sélectionné. Une relation semble-t-elle exister? Si oui, comment l'expliquer?

(b) En utilisant l'estimateur $\overline{y}_G$, calculez un I.C. du score moyen de tous les employées de l'entreprise, à environ 95%.

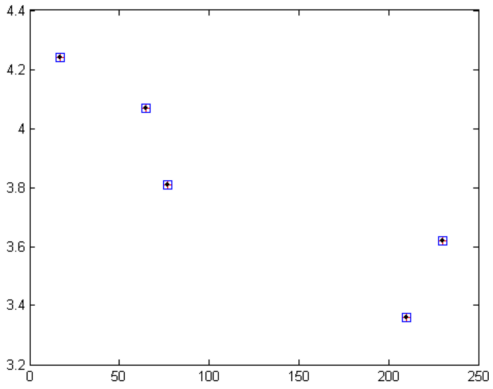
(c) Nous avons introduit l'estimateur $M\overline{y}_T$ afin de déterminer le total dans la population. En divisant cet estimateur par N, on obtient un estimateur pour la moyenne dans la population. En utilisant l'estimateur $\frac{M}{N}\overline{y}_T$, déterminer un I.C. pour le score moyen de tous les employés de cette entreprise, à environ 95%.

(d) Pourquoi l'I.C. obtenu en (b) est-il plus étroit que celui obtenu en (c)?

(e) Nous avons vu que $\overline{y}_G$ est un estimateur biasé de μ lorsque les grappes sont de tailles différentes. Montrer que $\frac{M}{N}\overline{y}_T$ est un estimateur non-biaisé de μ.

Solution:

(a) Le diagramme de dispersion se retrouve ici.



D'après le petit échantillon dont nous disposons, il semblerait que plus un département est petit, plus les employés ont tendance à être satisfaits de la manière dont la direction communique ses objectifs. Cela n'est pas tout à fait surprenant, car il doit être plus facile de les communiquer à des groupes plus petits, ne serait-ce que parce qu'une plus grande proportion d'employés peut être informée des objectifs en personne.

(b) On peut montrer que $\overline{y}_G = 3.61970$ et que la variance d'échantillonnage correspondante est $\hat{V}(\overline{y}_G) = 0.0099$; la marge d'erreur sur l'estimation est donc $B = 2\sqrt{0.0099} = 0.199$ et l'intervalle de confiance de la moyenne à environ 95% est 3.62 ± 0.20. (Ça, c'est si on se sert de $\overline{N}$ dans la formule; si on se sert de $\overline{n}$, à la place, on obtient 3.62 ± 0.18.)

(c) On peut montrer que $\frac{N}{M}\overline{y}_T = \frac{4770.04}{1176} = 4.06$ et que la marge d'erreur sur l'estimation est $B = \frac{2332.39}{1176} = 1.98$; l'intervalle de confiance de la moyenne à environ 95% dans ce cas est 4.06 ± 1.98.

(d) L'erreur sur la marge d'estimation de $\overline{y}_G$ est beaucoup plus faible que celle pour $\frac{M}{N}\overline{y}_T$, ce qui n'est pas surprenant puisque nous utilisons davantage d'informations auxiliaires pour calculer la moyenne dans le premier cas. Lorsque nous utilisons l'estimation $\frac{M}{N}\overline{y}_T$, nous sommes à la merci de la taille des grappes de l'échantillon : si trop de "grandes" grappes sont sélectionnées, l'estimation $\frac{M}{N}\overline{y}_T$ sera plus élevée que μ, alors que si trop de petites grappes sont sélectionnées, l'estimation $\frac{M}{N}\overline{y}_T$ sera plus petite que μ. Ce potentiel de variabilité élevée se reflète dans la plus grande variance/erreur d'estimation pour $\frac{M}{N}\overline{y}_T$.

(e) Soit N le nombre total d'employés, μ le score moyen par employé, M le nombre de grappes et τ le total des scores de tous les employés. Alors $\mu = \frac{\tau}{N}$. Puisque l'estimateur $M\overline{y}_T$ est basé sur le prélèvement d'un EAS de taille m sur les M totaux des grappes, nous avons que $E(M\overline{y}_T) = \tau$, de sorte que

$$E\left(\frac{M}{N}\overline{y}_T\right) = \frac{1}{N}E(M\overline{y}_T) = \frac{1}{N}\tau = \mu.$$

■