Chapitre 7 - Échantillonnage systématique

57. On donne un échantillon systématique du nombre de naissances (en milliers) et du taux de natalité (en naissances par 1000 individus) aux États-Unis entre 1950 et 1990.

Année	1950	1955	1960	1965	1970	1975	1980	1985	1990
Naissances	3632	4097	4258	3760	3731	3144	3612	3761	4158
Natalité	24.1	25.0	23.7	19.4	18.4	14.6	15.9	15.8	16.7

(a) Donner un estimé du nombre total de naissances pendant cette période. Trouver une estimation approximative de la variance.

(b) Donner un estimé du taux de natalité moyen pendant cette période et trouver un estimateur approprié de la variance. Cette moyenne est-elle un bon prédicteur du taux de natalité en 1995? Expliquer.

Solution:

(a) La période de 1950 à 1990 recouvre $N = 41$ années. Selon les données disponibles, l'échantillonnage systématique doit être $1-$parmi-5, puisque nous avons des observations tous les 5 ans. Quelle que soit la taille de l'échantillon n, il est impossible que d'avoir $nk = N$, car $k = 5$ ne divise pas $N = 41$. Il y a plusieur possibilités pour se sortir de ce pétrin (ma préférée étant la deuxième). Dénotons le total des naissances au cours de cette période par τ.

i. Si nous insistons pour préserver l'égalité $nk = N$, nous devons utiliser une valeur différente pour N : nous aurons tout simplement des estimations pour une période différente. Pour utiliser toutes les observations de l'échantillon, le choix le plus évident est $N = 45$ et $n = 9$. Dans ce cas, la période couverte est l'une des suivantes

$$1950 - 1994, \quad 1949 - 1993, \quad 1948 - 1992, \quad 1947 - 1991, \quad 1946 - 1990.$$

L'estimation fournie par l'échantillonnage systématique est

$$\hat{\tau}_{\text{SYS}} = N\overline{y}_{\text{SYS}} = \frac{N}{n}\sum_{i=1}^{n} y_i = \frac{45}{9} \cdot 34153 = 170765,$$

avec

$$\hat{V}\left(\hat{\tau}_{\text{SYS}}\right) = N^2\hat{V}\left(\overline{y}_{\text{SYS}}\right) = N^2\left(\frac{N-n}{N}\right)\frac{s^2}{n} \approx 45^2\left(\frac{45-9}{45}\right)\frac{115960}{9} \approx 20872800.$$

L'intervalle de confiance de τ à environ 95% est à peu près $170765 \pm 2\sqrt{20872800} \equiv 170765 \pm 9137.4$ pour une période de 45 années.

ii. Si, en revanche, nous n'insistons que pour préserver l'inégalité $nk \leq N$, nous pouvons utiliser la valeur $N = 41$. Dans ce cas, nous avons toujours $k = 5$ de sorte que $n \leq \frac{N}{k} = \frac{41}{5}$; $n = 8$ sera le choix le plus judicieux. La période couverte est alors $1950 - 1990$, mais nous n'utilisons que les 8 premières observations de l'échantillon (on pourrait aussi le faire pour les 8 dernières observations, mais je commence à en avoir assez de ré-écrire toujours la même chose). L'estimation fournie par l'échantillonnage systématique est

$$\hat{\tau}_{\text{SYS}} = N\overline{y}_{\text{SYS}} = \frac{N}{n}\sum_{i=1}^{n} y_i = \frac{41}{8} \cdot 29995 = 153724,$$

avec

$$\hat{V}\left(\hat{\tau}_{\text{SYS}}\right) = N^2 \hat{V}\left(\overline{y}_{\text{SYS}}\right) = N^2 \left(\frac{N-n}{N}\right) \frac{s^2}{n} \approx 41^2 \left(\frac{41-8}{41}\right) \frac{111322}{8} \approx 18827333.$$

L'intervalle de confiance de τ à environ 95% est à peu près $153724 \pm 2\sqrt{1882733} \equiv 153724 \pm 8678.1$.

iii. Enfin, nous pourrions décider d'utiliser $N = 41$, $k = 5$ et $n = 9$ comme le suggèrent les données, indépendamment du fait que $nk \not\leq N$, c'est-à-dire que l'échantillon n'est qu'un EAS. La période couverte est $1950 - 1990$ et nous utilisons tous les échantillons. L'estimation fournie par l'échantillonnage systématique est

$$\hat{\tau} = N\overline{y} = \frac{N}{n} \sum_{i=1}^{n} y_i = \frac{41}{9} \cdot 34153 = 155586,$$

avec

$$\hat{V}\left(\hat{\tau}\right) = N^2 \hat{V}\left(\overline{y}\right) = N^2 \left(\frac{N-n}{N}\right) \frac{s^2}{n} \approx 41^2 \left(\frac{41-9}{41}\right) \frac{115960}{9} \approx 16904391.$$

L'intervalle de confiance de τ à environ 95% est à peu près $155586 \pm 2\sqrt{16904391} \equiv 155586 \pm 8223.0$.

(b) Nous rencontrons le même problème que ci-dessus. Selon la façon dont nous décidons de traiter N, k et n, il y a au moins 3 possibilités. Dénotons le taux de naissance moyen par μ.

i. $N = 45$, $k = 5$ et $n = 9$. Dans ce cas, la période couverte est l'une des suivantes

$$1950 - 1994, \quad 1949 - 1993, \quad 1948 - 1992, \quad 1947 - 1991, \quad 1946 - 1990.$$

L'estimation fournie par l'échantillonnage systématique est

$$\overline{y}_{\text{SYS}} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{9} \cdot 173.6 = 19.2889,$$

avec

$$\hat{V}\left(\overline{y}_{\text{SYS}}\right) = \left(\frac{N-n}{N}\right) \frac{s^2}{n} \approx \left(\frac{45-9}{45}\right) \frac{16.0461}{9} \approx 1.4263.$$

L'intervalle de confiance de μ à environ 95% est à peu près $19.2889 \pm 2\sqrt{1.4263} \equiv 19.2889 \pm 2.3886$.

ii. $N = 41$, $k = 5$ et $n = 8$. Dans ce cas, la période couverte est $1950 - 1990$ et nous utilisons les 8 premières observations (ou les 8 dernières, etc.). L'estimation fournie par l'échantillonnage systématique est

$$\overline{y}_{\text{SYS}} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{8} \cdot 156.9 = 19.6125,$$

avec

$$\hat{V}\left(\overline{y}_{\text{SYS}}\right) = \left(\frac{N-n}{N}\right) \frac{s^2}{n} \approx \left(\frac{41-8}{41}\right) \frac{17.2613}{8} \approx 1.7367.$$

L'intervalle de confiance de μ à environ 95% est à peu près $19.6125 \pm 2\sqrt{1.7367} \equiv 19.6125 \pm 2.6356$.

iii. $N = 41$, $k = 5$ et $n = 9$ (un EAS). Dans ce cas, la période couverte est $1950 - 1990$ et nous utilisons toutes les observations. L'estimation fournie par l'échantillonnage systématique est

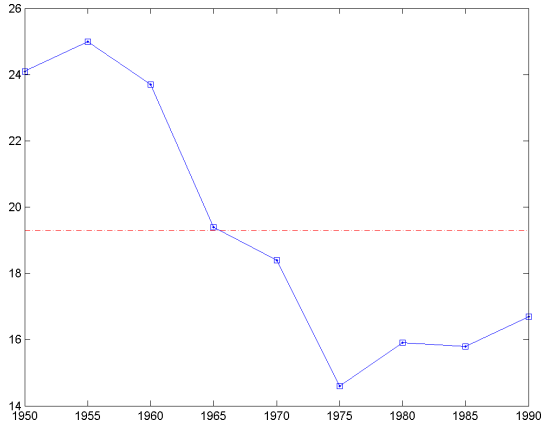
$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{9} \cdot 173.6 = 19.2889,$$

avec

$$\hat{V}\left(\overline{y}\right) = \left(\frac{N-n}{N}\right)\frac{s^2}{n} \approx \left(\frac{41-9}{41}\right)\frac{16.0461}{9} \approx 1.3915.$$

L'intervalle de confiance de μ à environ 95% est à peu près $19.2889 \pm 2\sqrt{1.3915} \equiv 19.2889 \pm 2.3593$.

Quelle que soit l'interprétation choisie, il est peu probable que l'estimation soit un bon prédicteur du taux de natalité en 1995. En effet, un tracé des taux de natalité en fonction de l'année montre une tendance intéressante : avant 1965, tous les taux de natalité échantillonnés sont supérieurs aux estimateurs, après 1965, ils sont tous inférieurs aux estimateurs.



Si ma mémoire est bonne, la société américaine a connu un bouleversement spectaculaire au cours des années 60. Les attitudes par rapport aux grossesses ont changé, et il est fort possible qu'elles aient affecté les taux de natalité postérieurs post-1965. Ainsi, toute prédiction pour 1995 serait probablement plus précise si l'on se limitait aux observations postérieures à 1965. ■

58. Une vérificatrice est confrontée à la longue liste de comptes débiteurs d'une entreprise. Elle doit vérifier les montants figurant sur 10% de ces comptes et estimer la différence moyenne entre les valeurs vérifiées et les valeurs comptables.

(a) Les comptes les plus anciens ont tendance à avoir des valeurs moins élevées. Supposons qu'ils soient classés par ordre chronologique. Lequel des plans SYS ou EAS est préférable dans ce cas? Expliquer.

(b) Supposons maintenant que les comptes sont énumérés de manière aléatoire. Lequel des plans SYS ou EAS est préférable dans ce cas? Expliquer.

(c) Supposons finalement que les comptes sont regroupés par département, puis classés par ordre chronologique au sein des départements dans une longue liste. Là encore, les comptes les plus anciens ont tendance à avoir des valeurs plus faibles.Lequel des plans SYS ou EAS est préférable dans ce cas? Expliquer.

Solution:

(a) Dans ce cas, chaque échantillon systématique comportera certaines des plus petites valeurs ainsi que certaines des plus grandes, ce qui ne serait pas nécessairement le cas avec un échantillonnage aléatoire simple. Cela implique que la variance de l'échantillonnage systématique sera plus faible que celle de l'échantillonnage aléatoire simple, de sorte que l'utilisation de la formule d'échantillonnage aléatoire simple produit une surestimation de la véritable erreur d'échantillonnage. Elle devrait probablement s'en tenir à l'échantillonnage systématique. Sauf si elle veut de mauvaises réponses.

(b) Si les comptes sont disposés de façon aléatoire, l'échantillonnage systématique et l'EAS sont équivalents (à toutes fins pratiques) et l'approximation de la variance à l'aide de la formule pour l'échantillonnage aléatoire simple fonctionne bien pour l'échantillonnage systématique. Elle doit choisir la méthode la plus facile à mettre en œuvre dans son cas particulier, de sorte qu'elle continuera probablement à utiliser l'échantillonnage systématique.

(c) Dans ce cas, les éléments de la population ont des valeurs qui vont suivre un cycle ascendant puis descendant de façon régulière lors de leur énumération. Pour éviter d'échantillonner au rythme de la période du cycle (si tant est qu'il y en ait une), ce qui introduit un biais dans les estimateurs, elle pourrait utiliser l'échantillonnage systématique et choisir un k relativement premier à la période du cycle. Elle peut aussi choisir un échantillon systématique qui touche à la fois les pics et les creux de la tendance cyclique, ce qui se rapproche de la méthode de l'échantillonnage aléatoire simple et permet d'utiliser la formule de variance de l'échantillonnage aléatoire simple comme une approximation raisonnable. Pour éviter le problème de la sous-estimation de la variation, elle pourrait également changer plusieurs fois le point de départ aléatoire ou utiliser un échantillonnage systématique répété. ■

59. Supposons que l'on s'intéresse aux ventes nettes moyennes (en millions de dollars) pour une population de 37 entreprises qui fabriquent du matériel informatique:

```
(1)  42.88   (2)  43.36   (3)  9.08     (4)  40.94   (5)  80.72
(6)  253.20  (7)  103.19  (8)  2869.35  (9)  196.32  (10) 193.34
(11) 18.99   (12) 30.90   (13) 3009.49  (14) 35.52   (15) 21.22
(16) 90.48   (17) 17.33   (18) 7.96     (19) 7.94    (20) 5.21
(21) 6.58    (22) 8.75    (23) 39.98    (24) 17.66   (25) 17.47
(26) 7.30    (27) 4.59    (28) 6.03     (29) 29.93   (30) 21.64
(31) 29.50   (32) 20.52   (33) 8.43     (34) 58.08   (35) 35.52
(36) 21.13   (37) 29.83
```

(a) Supposons qu'un échantillon SYS 1−parmi−7 est prélevé dans cette population afin d'estimer les ventes totales. Si la première entreprise sélectionnée est la troisième de la liste, quel est l'échantillon?

(b) Décrire le plan d'échantillonnage de la partie (a) en termes d'échantillonnage par grappes.

(c) Suite à la réponse de la partie (b), expliquer la difficulté rencontrée lors du calcul de la variance d'échantillonnage du plan décrit en (a).

(d) En supposant que l'échantillon systématique prélevé en (a) puisse être traité comme un EAS, donner un I.C. du total des ventes nettes pour l'année 2000, à environ 95% .

(e) Deux échantillons SYS 1−parmi−7 supplémentaires sont prélevés de la liste. Le premier de ces échantillons est tel que la première entreprise sélectionnée est la septième, tandis que le second est tel que la première entreprise sélectionnée est la deuxième. En utilisant les informations contenues dans ces deux échantillons et celui sélectionné en (a), donner un I.C. du total des ventes nettes pour l'année 2000, à environ 95% en se basant sur l'estimateur $N\overline{y}_G$.

(f) Répèter la partie (e), mais à l'aide de l'estimateur $M\overline{y}_T$.

(g) Est-ce qu'un plan d'échantillonnage systématique répété est une meilleure approche que celle fournie par un plan EAS? Expliquer.

Solution:

(a) Dans un SYS, nous devons avoir $7 = \lfloor \frac{37}{n} \rfloor$, de sorte que nous visons à ce que chaque échantillon soit de taille $n = 5$, même si certains échantillons auront une taille de 6. Ainsi, l'échantillon approprié est

entreprise	3	10	17	24	31
ventes	9.08	193.34	17.33	17.66	29.50

(b) Les grappes sont les colonnes de données, telles qu'elles apparaissent dans la question. Un échantillon systématique de 1−parmi−7 correspond à la sélection d'une des grappes (dans ce cas, la troisième colonne) et de chaque unité dans cette colonne.

(c) Nous avons vu que l'échantillonnage systématique est équivalent à un échantillonnage aléatoire simple de grappes où nous sélectionnons une grappe unique. Lors du calcul de la variance des estimateurs, nous devons calculer s_C^2, qui se trouve être 0 lorsqu'une seule grappe est présente. Ainsi, si nous analysons l'échantillonnage systématique comme un cas particulier de l'échantillonnage en grappes, l'erreur approximative sur la limite d'estimation est toujours nulle, ce qui n'est pas très utile puisque les estimateurs ne sont pas parfaits.

(d) Si on considère l'échantion en (a) en tant que EAS des ventes, nous obtenons

$$M\overline{y} = \frac{37}{5}(9.08 + 193.34 + 17.33 + 17.66 + 29.50) = 1975.134$$

et $s^2 = 6174.3$, de sorte que

$$\hat{V}(M\overline{y}) = M^2\hat{V}(\overline{y}) = M^2\frac{s^2}{n}\left(1 - \frac{n}{M}\right) = \frac{37^2}{5} \cdot 6174.3\left(1 - \frac{5}{37}\right) = 1462074.$$

La marge d'erreur sur l'estimation est alors $B = 2\sqrt{1462074} = 2418.3$, et l'intervalle de confiance correspondant est 1975.134 ± 2418.3... ce n'est pas fameux.

(e) Dans ce cas, on peut résumer les grappes choisies à l'aide du tableau suivant:

grappe i	1	2	3
m_i	6	5	5
$\overline{y}_i$	70.27	53.38	37.37

On peut montrer que l'intervalle de confiance construit à partir de $M\overline{y}_C$ est 2024.29 ± 551.216.

(f) On peut montrer que l'intervalle de confiance construit à partir de $N\overline{y}_T$ est 2042.53 ± 729.226.

(g) Si les données ne sont pas ordonnées en fonction de la variable d'intérêt ou d'une variable auxiliaire à forte corrélation, il ne devrait pas y avoir de différence en général entre l'échantillonnage aléatoire simple et l'échantillonnage systématique. En revanche, l'échantillonnage systématique répété s'apparente à l'échantillonnage en grappes à un seul degré, qui est généralement plus efficace que l'échantillonnage aléatoire simple lorsque les unités ne sont pas homogènes au sein de chaque grappe, ce qui risque de se produire lorsque les données sont ordonnées de façon aléatoire, comme c'est le cas ici. ■

60. On considère une population à "tendance linéaire" de taille N, prenant les valeurs $u_j = j$, $j = 1, \ldots, N$.

(a) Calculer μ et σ^2 pour cette population.

(b) On prélève un EAS de taille n de cette population. Si $N = nM$, montrer que

$$V(\overline{y}) = \frac{(M-1)(N+1)}{12}.$$

(c) Les observations de la population sont énumérées en ordre croissant. On les divises en n strates de taille M, de sorte à ce que $\frac{N_i}{N} = \frac{M}{N} = \frac{1}{n}$ pour $i = 1, \ldots, n$. Expliquer pourquoi $\sigma_i^2 = \frac{(M-1)(M+1)}{12}$ dans chaque strate. De plus, montrer que dans un plan STR où on choisit au hasard une unité par strate, on obtient

$$V(\overline{y}_{\text{STR}}) = \frac{M^2-1}{12n}.$$

(d) Montrer que pour un échantillon SYS $1-$parmi$-M$ prélevé à même cette population, on obtient $\overline{y}_{\text{SYS}(j)} - \mu = j - \frac{M+1}{2}$, $j = 1, \ldots, M$. En conséquence, montrer que

$$V(\overline{y}_{\text{SYS}}) = \frac{M^2-1}{12}.$$

(e) Expliquer pourquoi le plan STR est préférable au plan SYS, qui est à son tour préférable au plan EAS dans cette situation. Quelles implications cela pourrait-il avoir dans la pratique?

Solution:

(a) Nous avons

$$\mu = \frac{1}{N}\sum_{j=1}^{N} u_j = \frac{1}{N}\sum_{j=1}^{N} j = \frac{1}{N}\cdot\frac{N(N+1)}{2} = \frac{N+1}{2}$$

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{N} u_j^2 - \mu^2 = \frac{1}{N}\sum_{j=1}^{N} j^2 - \frac{(N+1)^2}{4} = \frac{1}{N}\cdot\frac{N(N+1)(2N+1)}{6} - \frac{(N+1)^2}{4}$$

$$= (N+1)\left[\frac{2N+1}{6} - \frac{N+1}{4}\right] = \frac{(N+1)(N-1)}{12}$$

(b) Pour un EAS, nous avons

$$V(\overline{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{(N+1)(N-1)}{12n}\left(\frac{N-n}{N-1}\right) = \frac{N+1}{12n}(nM-n) = \frac{(N+1)(M-1)}{12}$$

(c) Dans la $i-$ème strate, les unités sont $y_{(i-1)M+k} = (i-1)M + k$, $k = 1, \ldots, M$. Alors,

$$\overline{y}_i = \frac{1}{M}\sum_{k=1}^{M} y_{(i-1)M+k} = \frac{1}{M}\sum_{k=1}^{M}[(i-1)M+k] = (i-1)M + \frac{M+1}{2}$$

$$\sigma_i^2 = \frac{1}{M}\sum_{k=1}^{M}\left[y_{(i-1)M+k} - \overline{y}_i\right]^2 = \frac{1}{M}\sum_{k=1}^{M}\left[k - \frac{M+1}{2}\right]^2 = \sum_{k=1}^{M}\left[k^2 - (M+1)k + \left(\frac{M+1}{2}\right)^2\right]$$

$$= \frac{(M+1)(2M+1)}{6} - (M+1)\frac{(M+1)}{2} + \left(\frac{M+1}{2}\right)^2 = \frac{(M-1)(M+1)}{12} = \frac{M^2-1}{12}.$$

Si l'on prélève $n_i = 1$ unité de chaque strate, nous obtenons

$$V(\overline{y}_{STR}) = \frac{1}{N^2}\sum_{i=1}^{n} N_i^2\sigma_i^2 = \frac{M^2-1}{12N^2}\sum_{i=1}^{n}\frac{N^2}{n^2} = \frac{M^2-1}{12n}$$

(d) Dans ce cas, il y a M différent échantillon systématique $1-$parmi$-M$:

$$\begin{array}{ccccc} 1 & M+1 & 2M+1 & \cdots & (n-1)M+1 \\ 2 & M+2 & 2M+2 & \cdots & (n-1)M+2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ M & M+M & 2M+M & \cdots & (n-1)M+M \end{array}$$

Si $j = 1, \ldots, n$, nous avons

$$\begin{aligned} \overline{y}_{sy(j)} - \mu &= \frac{1}{n}\sum_{k=0}^{n-1}[kM+j] - \frac{N+1}{2} = \frac{M}{n}\frac{(n-1)n}{2} + \frac{1}{n}nj - \frac{N+1}{2} = \frac{M(n-1)}{2} + j - \frac{N+1}{2} \\ &= \frac{Mn-M}{2} + j - \frac{nM+1}{2} = j + \frac{Mn-M-nM-1}{2} = j - \frac{M+1}{2} \end{aligned}$$

(e) Nous savons déjà que l'échantillonnage systématique est plus efficace que l'échantillonnage aléatoire simple lorsque la population est classée par ordre croissant ou décroissant, en fonction de la variable d'intérêt. Il n'est pas surprenant que le schéma stratifié décrit ci-dessus soit plus efficace que l'échantillon systématique, car le point de départ de l'échantillon systématique est choisi au hasard : si, par accident, le tout premier nombre entier est choisi, le reste des unités échantillonnées par l'échantillon systématique seront les plus petites unités possibles, ce qui créerait une estimation inférieure à la valeur réelle de la moyenne. En revanche, si le point de départ est la plus grande valeur possible qui peut être choisie, l'estimation systématique sera plus grande que la valeur réelle de la moyenne. Avec le schéma d'échantillonnage stratifié décrit ci-dessus, il est possible que certaines des valeurs sélectionnées soient parmi les plus petites valeurs possibles dans leurs strates, et que certaines d'entre elles soient parmi les plus grandes valeurs possibles dans leurs strates, ce qui aura tendance à diminuer la variance de l'estimateur.

Les résultats impliquent non seulement que si la population est classée par ordre croissant (ou décroissant), il pourrait être plus efficace d'utiliser un échantillonnage systématique qu'un échantillonnage aléatoire simple (ce que nous savions déjà), mais aussi que séparer la population en strates successives et choisir un seul élément dans chaque strate pourrait être encore plus efficace que l'échantillonnage systématique. ■