MAT 3777 – Échantillonnage et sondages – Solutions

Chapitre 1 – Introduction

1. Pour chacune des situations suivantes, discuter des mérites relatifs de l'utilisation d'entretiens personnels, d'entretiens téléphoniques, et de questionnaires postaux comme méthodes de collecte de données.

   (a) Une responsable de la télévision veut donner un estimé de la proportion de téléspectateurs dans le pays qui regardent sa chaîne à une certaine heure.

   (b) Le service de santé d'une municipalité souhaite évaluer la proportion de chiens qui ont été vaccinés contre la rage au cours de l'année écoulée.

   (c) Un commissaire municipal souhaite connaître l'avis des propriétaires sur une proposition de changement de zonage.

   (d) Le rédacteur en chef d'un journal souhaite sonder l'attitude du public à l'égard du type de couverture médiatique proposé par son journal.

   Solution: En général, les entretiens personnels sont coûteux, mais ils fournissent plus d'informations, tandis que les questionnaires postaux sont moins chers, mais ils ont des taux de non-réponse élevés. Les entretiens téléphoniques se situent souvent entre ces deux cas.

   (a) **Entretiens personnels:** (Avantage) le taux de réponse est plus élevé lorsqu'on est confronté à une personne "vivante." (Inconvénient) la question étant simple, nous pouvons obtenir d'aussi bons résultats qu'avec des entretiens téléphoniques, mais à un coût plus élevé, puisque la base de sondage contient des téléspectateurs de tout le pays (frais interurbains).

   Interviews téléphoniques: (Avantage) Comme la question est simple, le temps nécessaire pour réaliser l'interview est faible. (Inconvénient) Beaucoup de gens n'aiment pas être dérangés chez eux pour des futilités.

   Questionnaires postaux: (Avantage) Faible coût. (Inconvénient) Pour le sujet en question, la non-réponse peut biaiser davantage les observations puisque les personnes qui regardent ce réseau particulier peuvent être plus enclines à répondre que celles qui ne le sont pas.

   Verdict: les entretiens téléphoniques semblent plus appropriés ici, car les taux de réponse sont maximisés en raison de la courte durée des entretiens.

   (b) **Entretiens personnels:** (Avantage) si on utilise une base de sondage de zones, cela permet de limiter les coûts. De plus, j'imagine que les questions de zonage sont plus susceptibles d'intéresser les habitants de la ville/du quartier, donc un entretien personnel permettra de mieux enregistrer les nuances qui peuvent être exprimées. (Inconvénient) Les enquêteurs peuvent avoir à faire face à des personnes en colère. Et toujours la question du coût de la collecte d'informations : elle pourrait probablement être réalisée à un coût moindre.

   Entretiens téléphoniques et questionnaires postaux: (Avantage) Puisque le changement de zonage affecte les propriétaires (dans un certain sens), ils sont plus susceptibles de répondre.

   Verdict: J'opterais probablement pour l'entretien personnel dans ce cas, car il serait plus facile d'obtenir toute la gamme des nuances possibles dans les sentiments des propriétaires.

(c) **Entretiens personnels et téléphoniques:** (Inconvénient) Les répondants propriétaires de chiens qui n'ont pas fait vacciner leurs chiens pourraient se sentir coupables et donner une fausse réponse en personne ou au téléphone. La sélection du cadre pourrait également poser problème; je ne suis pas familier avec les questions relatives aux animaux, mais comment construire un cadre approprié de propriétaires d'animaux ? Je peux facilement m'imaginer qu'il y ait des problèmes de couverture.

Questionnaires postaux: (Avantage) Pourrait être plus rentable si la base de sondage est difficile à construire. (Inconvénient) Non-réponse de la part des personnes qui n'ont pas donner la piqûre à leur chien ?

Verdict: si nous pouvons obtenir une base de sondage appropriée, je pense que le questionnaire envoyé par la poste est plus judicieux ici.

(d) **Entretiens personnels:** (Avantages) La question est complexe et un enquêteur en personne peut noter les indices non verbaux et les réactions. Un exemplaire du journal peut être apporté pour le montrer à la personne interrogée. (Inconvénient) Si un exemplaire est apporté pour être lu, le choix particulier de l'exemplaire peut biaiser la réponse de la personne.

Entretiens téléphoniques: (Avantage) Pour un journal régional, il est moins probable qu'il y ait des frais d'appel longue distance, donc cette méthode est probablement moins chère que les entretiens personnels. (Inconvénient) La question étant plus complexe, l'interview peut avoir tendance à être longue et le répondant peut rapidement devenir un non-répondant.

Questionnaires postaux: (Avantage) Les personnes interrogées peuvent prendre plus de temps pour répondre aux questions complexes, et lire le journal plus souvent pour se faire une opinion. (Inconvénient) Les personnes qui ne lisent pas le journal en question peuvent prendre le questionnaire envoyé par la poste pour une offre d'abonnement et ne pas ouvrir l'enveloppe (?). De plus, si elles prennent plus de temps à répondre, c'est peut-être pour composer un long tôme, et nous pourrions avoir du mal à traiter leurs réponses.

Verdict: un pile ou face entre les entretiens personnels et le questionnaire envoyé par courrier.

Vos réponses peuvent différer, bien sûr, tant qu'elles sont justifiées.

2. Le ministère de la chasse et de la pêche du Québec est préoccupé par l'orientation de ses futurs programmes de chasse. Afin de prévoir un plus grand potentiel pour la chasse future, le département cherche à déterminer la proportion de chasseurs recherchant deux types de gibier. Un échantillon de taille $n = 1250$, prélevé parmis les $N = 95,675$ chasseurs titulaires d'un permis, a été obtenu. Expliquer pourquoi le ministère a préféré une enquête par sondage à un recensement.

   Solution: Une enquête par sondage est moins dispendieuse et moins longue qu'un recensement, et sa qualité est généralement suffisante. Elle peut même être plus précise, car un échantillon plus petit permet de consacrer des fonds à une meilleure conception du questionnaire et à la formation du personnel de terrain. ■

3. Dans laquelle des situations suivantes pouvez-vous généraliser raisonnablement de l'échantillon à la population ?

   (a) On se sert des étudiant.e.s de ce cours afin d'obtenir une estimation du pourcentage des étudiant.e.s de l'Université d'Ottawa qui étudient au moins deux heures par jour.

   (b) On utilise le revenu annuel moyen des ambassadeurs auprès de l'ONU pour obtenir une estimation du revenu moyen par habitant à l'échelle mondiale.

   (c) En 2018, un maison de sondage réputée a échantillonné 500 résidents canadiens âgés de 18 à 29 ans afin de donner un estimé du pourcentage de tous les résidents canadiens âgés de 18 à 29 ans qui étaient favorables à une réduction des dépenses militaires.

   Solution:

   (a) Possiblement. La taille de la classe est relativement élevée, et la qualité des étudiant.e.s dans ce cours ne devrait pas être bien différente de celle dans n'importe lequel des cours du premier cycle à l'UdO, et donc leurs habitudes d'étude devraient être sensiblement les mêmes, en moyenne. Ceci étant dit, je ne serais pas surpris de découvrir que les habitudes d'étude varient d'un programme d'études à l'autre, ou d'un cycle à l'autre.

   (b) Pas dans ce cas. Bien que n'étant pas au courant du processus de sélection, je suis à peu près certain que les ambassadeurs sont généralement choisis en fonction de leurs relations politiques et qu'ils sont donc, en règle générale, (beaucoup, beaucoup) plus riches que les citoyen.ne.s moyen.ne.s qu'ils représentent. Mais même si ce n'était pas le cas, puisque la population de certains pays est beaucoup plus imposante que celle d'autres pays, les différents ambassadeurs devraient avoir des "poids" différents lors du calcul du salaire moyen, ce qui n'est pas le cas ici.

   (c) Dans ce cas, oui. L'institut de sondage étant "réputé", on peut supposer que l'échantillon aléatoire a été tiré au sort (en utilisant la méthode d'échantillonnage jugée appropriée) et que les résultats sont aussi exempts de biais que possible. De plus, l'échantillon semble suffisamment grand (mais ce n'est pas une garantie!). ■

   Vos réponses peuvent différer, bien sûr, tant qu'elles sont justifiées.

4. On cherche à évaluer la distance moyenne quotidienne parcourue par les voitures Ontariennes, ainsi que la consommation d'essence quotidienne. Discuter de diverses approches à utiliser. Quels sont les enjeux et difficultés?

Solution: les enjeux et difficultés pourrait inclure:

- **Population cible:** toutes les voitures Ontariennes – est-ce une population bien définie? Les autobus sont-elles de la partie? Les camions de type "pick-up"? etc.
- **Population à l'étude:** le MTO gère les enregistrements des voitures de la province – à quelle fréquence cette liste est-elle mise à jour?
- **Base de sondage:** la liste contient-elle les adresses des propriétaires? Ou un autre moyen de les contacter? Si on loue un véhicule, est-ce la propriétaire ou l'utilisatrice qui se retrouve dans cette liste? Peut-on avoir accès a cette liste? Peut-on concevoir d'autres bases de sondage?
- **Population répondante:** quel proportion des propriétaires identifiés répondraient s'ils.elles étaient sélectionné.e.s.?
- **Variables réponses:** comment mesure/calcule-t-on la distance moyenne quotidienne parcourue ou la comsommation quotidienne de carburant? Une auto-déclaration? Un GPS/compteur électronique? Y a-t-il des enjeux envers la protection de la vie privée? Que faire si on remarque qu'un conducteur dépasse la limite de vitesse ou est impliqué dans une collision?
- **Attributs:** est-ce suffisant de ne mesurer que les deux variables réponses? Devrait-on aussi mesurer des réponses auxiliaires? De l'information au sujet des conducteurs? Des passagers? Sur quelle période?
- **Erreur de couverture:** est-ce que toutes les voitures sont représentées dans la base de sondage? S'il y a des voitures qui ne sont pas représentées, leur absence est elle systématique (rurale, urbaine, etc.)?
- **Erreur de non-réponse:** les conducteurs qui font de la vitesse ou qui ont des comportements illicites sont-ils moins disposés à participer? Ceux qui ne conduisent pas (ou presque jamais) sont-ils moins portés à répondre?
- **Erreur d'échantillonnage:** les individus de la population répondante sélectionnés par le sondage peuvent ne pas être représentatifs de la population répondante, côté habitudes de conduite (trop de conducteurs de banlieues, par exemple)
- **Erreur de mesure et de traitement:** peux-t-on faire confiance aux auto-déclarations? Les GPS/compteurs électroniques sont-ils assez simple à installer? Les données électroniques peuvent-elles être corrompues?
- **Variabilité d'échantillonnage:** différents échantillons de voitures de la base de sondage et de la population répondante pourraient produire des résultats différents, côté habitudes de conduite
- **Variabilité des mesures:** pour certains ou tous les membres de l'échantillon réalisé, on pourrait obtenir des réponses différentes en fonction de la période de sondage ■

Vos réponses peuvent différer, bien sûr, tant qu'elles sont justifiées.

5. Considérons une population de taille $N = 5$ contenant les valeurs $\{0, 1, 2, 3, 4\}$. Supposons que nous choisissions un échantillon aléatoire simple de taille $n = 3$. Soit μ la moyenne de la population et σ^2 sa variance.

   (a) Quelle est la fonction de distribution de probabilité de la moyenne de l'échantillon $\overline{y}$?

   (b) Démontrer que $E(\overline{y}) = \mu$.

   (c) Démontrer que $V(\overline{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$.

   Solution:

   (a) Il y a $\binom{5}{3} = 10$ façons de sélectionner des échantillons aléatoires simples de taille 3 dans la population. Elles sont énumérées ci-dessous, avec leur moyenne respective :

			Moyenne
0	1	2	1
0	1	3	4/3
0	1	4	5/3
0	2	3	5/3
0	2	4	2
0	3	4	7/3
1	2	3	2
1	2	4	7/3
1	3	4	8/3
2	3	4	3

   La fonction de distrbution de probabilité de la moyenne de l'échantillon est ainsi

$$f(\overline{y}) = \begin{cases} \frac{1}{10} & \text{si } \overline{y} = 1, \frac{4}{3}, \frac{8}{3}, 3 \\ \frac{1}{5} & \text{si } \overline{y} = \frac{5}{3}, 2, \frac{7}{3} \end{cases}$$

   (b) Nous savons que $\mu = \frac{1}{5}(0 + 1 + 2 + 3 + 4 + 5) = 2$. Puisque

$$E(\overline{y}) = \sum \overline{y} f(\overline{y}) = \left(1 + \frac{4}{3} + \frac{8}{3} + 3\right)\frac{1}{10} + \left(\frac{5}{3} + 2 + \frac{7}{3}\right)\frac{1}{5} = 2,$$

   on en déduit que $E(\overline{y}) = \mu$.

   (c) Nous savons que $\sigma^2 = \frac{1}{5}(0^2 + 1^2 + 2^2 + 3^2 + 4^2) - 2^2 = 2$. Puisque

$$V(\overline{y}) = \sum \overline{y}^2 f(\overline{y}) - E(\overline{y})^2 = \left(1 + \frac{16}{9} + \frac{64}{9} + 9\right)\frac{1}{10} + \left(\frac{25}{9} + 4 + \frac{49}{9}\right)\frac{1}{5} - 2^2 = \frac{1}{3}$$

   et

$$\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{2}{3}\left(\frac{5-3}{5-1}\right) = \frac{1}{3},$$

   on en déduit que $V(\overline{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$. ■

6. (a) Produire une population de taille $N = 100$ avec une variable y provenant d'une distribution de Poisson avec paramètre $\lambda = 9250$.

   (b) En utilisant des échantillons de taille $n = 10$ sans remise, sélectionner 1500 échantillons aléatoires simples de façon répétée dans la population obtenue en (a).

   (c) Calculer la moyenne de y pour chacun des 1500 échantillons.

   (d) Afficher les moyennes d'échantillon, et produire une distribution d'échantillonnage empirique de la moyenne $\overline{y}$ pour des échantillons de taille $n = 10$.

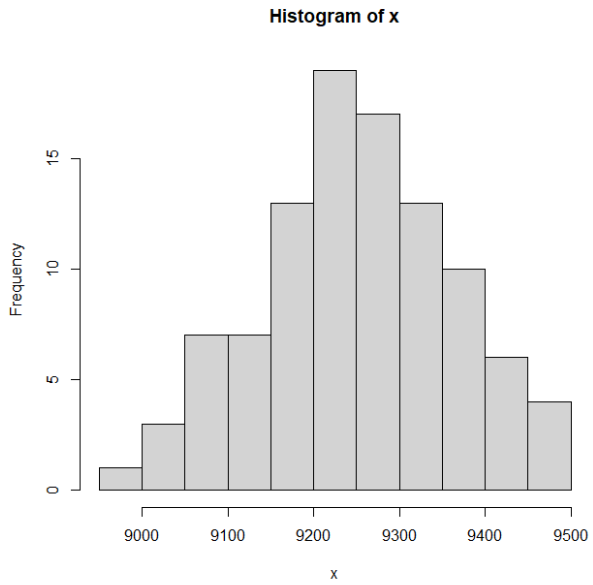
   (e) Décrire la forme de la distribution d'échantillonnage empirique. Semble-t-elle normale? Pourquoi ou pourquoi pas?

   (f) Calculer l'écart-type des moyennes d'échantillons produites. Comment se compare-t-il à la valeur théorique $\frac{\sigma}{\sqrt{n}}$?

Solution:

(a) On pourrait se servir du code R suivant, par exemple:

```
> N=100; lambda=9250
> x <- rpois(N,lambda)
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   8975    9171    9252    9253    9321    9499
> hist(x)
```



Vos valeurs seront différentes, selon le "seed" utilisé.

(b) On pourrait utiliser le code R suivant:

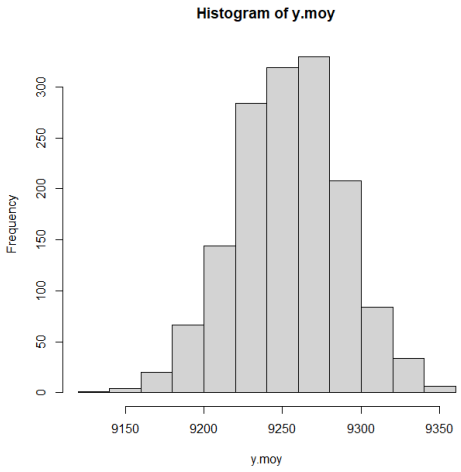
```
> n=10
> M=1500
> y=list()
> for(j in 1:M){
    y[[j]]=sample(x,n,replace=FALSE)
> }
```

(c) On pourrait utiliser le code R suivant:

```
> y.moy=c()
> for(j in 1:M){
    y.moy[j]=mean(y[[j]])
> }
```

(d) On pourrait utiliser le code R suivant:

```
> summary(y.moy)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9134    9231    9255    9254    9277    9357
> hist(y.moy)
```

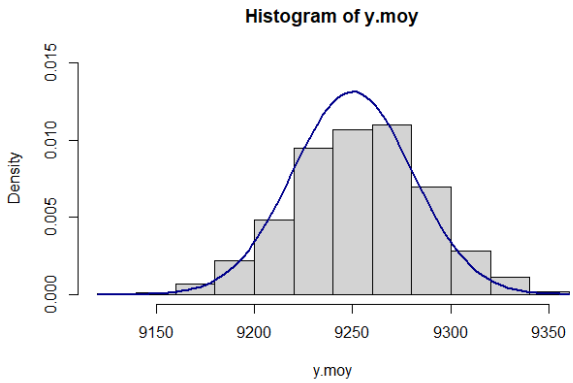


(e) Cet histogramme est à peu près symétrique autour de de 9250, et il semble certainement suivre une loi plus ou moins normale.

(f) Selon le TLC, on s'attend à ce que $\overline{y}$ suive une loi normale dont la moyenne serait $E(\overline{y}) = 9250$ et dont l'écart-type serait $\sqrt{V(\overline{y})} = \frac{\sigma}{\sqrt{10}} = \frac{\sqrt{9250}}{\sqrt{10}} = 30.4$. En pratique, ces valeurs sont

```
> mean(y.moy)
[1] 9254.093
> sd(y.moy)
[1] 34.0155
```

Vous conviendrez que cela se rapproche des valeurs théoriques. ■



7. Une étude sociologique menée dans un village s'intéresse à la proportion de ménages dont au moins un membre est âgé de plus de 65 ans. Le village compte 631 ménages selon l'annuaire municipal le plus récent. Un échantillon aléatoire simple de $n = 75$ ménages a été sélectionné dans l'annuaire. Au terme du travail de terrain, sur les 75 ménages échantillonnés, il n'y en avait que 13 qui contenaient au moins un membre âgé de plus de 65 ans.

   (a) Donner un estimé de la véritable proportion p de ménages dont au moins un membre est âgé de plus de 65 ans au village.

   (b) Quelle est la marge d'erreur sur l'estimation?

   (c) Construire un intervalle de confiance de p à environ 95%.

   (d) Quelle taille d'échantillon faut-il utiliser afin d'estimer p avec une marge d'erreur sur l'estimation de 0.07? Supposer que la proportion réelle $p \approx 0.25$.

   Solution:

   (a) La proportion réelle p est approchée par $\hat{p} = \frac{13}{75} \approx 0.1733$.

   (b) La marge d'erreur sur l'estimation est $2\sqrt{\hat{V}(\hat{p})}$, où

$$\hat{V}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1}\left(\frac{N-n}{N}\right) = \frac{\frac{13}{75}\cdot\frac{62}{75}}{75-1}\left(\frac{631-75}{631}\right) \approx 0.001706185,$$

   d'où la marge d'erreur sur l'estimation de p est $2\sqrt{0.001706185} \approx 0.0826$.

   (c) En supposant que la moyenne de l'échantillon suive une loi normale approximative,

$$\hat{p} \pm 2\sqrt{\hat{V}(\hat{p})} \equiv 0.1733 \pm 0.0826$$

   forment les extrémités de l'I.C. à 95% de p, d'où l'intervalle en question est $[0.0907, 0.2559]$.

   (d) Puisque $p = 0.25$, la taille d'échantillon requise afin de donner un estimé de p avec une marge d'erreur de $B = 0.07$ est

$$n = \frac{Np(1-p)}{(N-1)\frac{B^2}{4} + p(1-p)} = \frac{631(0.25)(0.75)}{630\frac{(0.07)^2}{4} + 0.25(0.75)} = 123.3375.$$

   Ainsi, il suffit de choisir $n \geq 124$. ■

8. 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) Quelle est la population cible? Que sont les unités de la population?

(b) Quelle est la variable réponse? Quel est l'attribut de la population d'intérêt?

(c) Supposons que nous décidons de procéder à une estimation de la moyenne des ventes pour toutes les entreprises en sélectionnant un échantillon aléatoire simple de taille $n = 8$, en utilisant les observations 3, 4, 12, 15, 21, 22, 25, 30. Quelle valeur obtient-on pour la moyenne de votre échantillon ?

(d) En supposant que les ventes nettes des 37 entreprises ont été mesurées sans erreur, trois autres types d'erreur d'enquête peuvent être présents : l'erreur de couverture, l'erreur de non-réponse et l'erreur d'échantillonnage. Indiquer si chacun des trois autres types d'erreur est présent lors de l'estimation de la moyenne et expliquer pourquoi.

Solution:

(a) La population cible est constituée de 37 entreprises qui fabriquent du matériel informatique. Les unités de la population cible sont donc les entreprises qui fabriquent du matériel informatique et qui se retrouvent dans la liste des 37.

(b) La variable réponse d'intérêt est le chiffre d'affaires net de chaque entreprise, que nous désignerons par u_j, $j = 1, \ldots, N = 37$. L'attribut de population d'intérêt est la moyenne des u_j, c'est-à-dire

$$\mu = \frac{1}{37} \sum_{j=1}^{37} u_j.$$

(c) L'échantillon correspondant est prsénté dans le tableau suivant:

i	3	4	12	15	21	22	25	30
y_i	9.08	40.94	30.90	21.22	6.58	8.75	17.47	21.64

La moyenne d'échantillon est ainsi

$$\overline{y} = \frac{1}{8} \sum_{i=1}^{8} y_i = \frac{1}{8} \left(9.08 + 40.94 + 30.90 + 21.22 + 6.58 + 8.75 + 17.47 + 21.64\right) \approx 19.58.$$

(d) Puisque chaque unité de la population cible a été identifiée, la population étudiée et la population cible sont les mêmes. Par conséquent, il ne peut y avoir d'erreur de couverture. De plus, puisque nous avons la valeur de la variable de réponse pour chaque unité de la population cible, il ne peut y avoir d'erreur de non-réponse. La seule erreur est l'erreur d'échantillonnage, puisque la moyenne de l'échantillon n'est pas nécessairement égale à la moyenne de la population (et qu'elle dépend de l'échantillon choisi). ■

9. Utiliser les observations de la question précédente.

   (a) Écrire et exécuter un programme unique qui:

      i. calcule la valeur moyenne des ventes pour la population de 37 entreprises;

      ii. prélève un échantillon aléatoire simple de ces entreprises, de taille $n = 8$, et,

      iii. calcule la valeur moyenne des ventes pour cet échantillon.

   (b) Répéter la partie (a) pour trois autres échantillons. En considérant les valeurs des ventes pour les 37 entreprises de la population, expliquer pourquoi les moyennes de l'échantillon prennent des valeurs inférieures à 130, entre 360 et 500, ou entre 735 et 850.

   (c) Écrire et exécuter un autre programme qui:

      i. prélève un unique échantillon aléatoire de $n = 8$ entreprises, et

      ii. utilise les observations de l'échantillon afin de déterminer un estimé des ventes moyennes pour l'ensemble des 37 entreprises, tout en donnant une approximation de la marge d'erreur sur l'estimation de la moyenne, et un intervalle de confiance de la moyenne à environ 95%.

Solution:

(a) On pourrait se servir du code R suivant, par exemple:

```
> x <- c(42.88,43.36,9.08,40.94,80.72,253.20,103.19,2869.35,196.32,193.34,
       18.99,30.90,3009.49,35.52,21.22,90.48,17.33,7.96,7.94,5.21,
       6.58,8.75,39.98,17.66,17.47,7.30,4.59,6.03,29.93,21.64,
       29.50,20.52,8.43,58.08,35.52,21.13,29.83)
> n=8
> set.seed(0) # replicabilite
> (x.ech <- sample(x,n,replace=FALSE))
[1] 35.52 40.94 42.88 58.08 39.98 18.99 29.83  7.96
> mean(x.ech)
[1] 34.2725
```

(b) On répète le code trois fois supplémentaires et on obtient

```
> (x.ech <- sample(x,n,replace=FALSE))
[1]   8.43   6.58  21.13 193.34 103.19 196.32  21.22  35.52
> mean(x.ech)
[1] 73.21625
> (x.ech <- sample(x,n,replace=FALSE))
[1] 29.83 17.47 58.08 21.13 21.22 42.88  5.21  9.08
> mean(x.ech)
[1] 25.6125
> (x.ech <- sample(x,n,replace=FALSE))
[1] 253.20 193.34   5.21   6.03  35.52   7.30  30.90  17.47
> mean(x.ech)
[1] 68.62125
```

Les deux plus grandes valeurs de la population sont 3009.49 et 2869.35. Un échantillon de taille 8 peut ne contenir aucune de ces valeurs, exactement l'une d'entre elles, ou les deux. Dans le cas où l'échantillon n'en contient aucune, la moyenne la plus élevée que nous pouvons obtenir est par l'échantillon

$$43.36, 58.08, 80.72, 90.48, 103.19, 193.34, 196.32, 253.20, \quad \text{moy} = 127.3362.$$

Dans le cas où l'échantillon contient les deux plus grands nombres, la moyenne la plus élevée que nous puissions obtenir est par l'échantillon

$$80.72, 90.48, 103.19, 193.34, 196.32, 253.20, 2869.35, 3009.49, \quad \text{moy} = 849.5112;$$

la moyenne la plus basse que l'on peut obtenir dans les mêmes conditions est celle de l'échantillon

$$4.59, 5.21, 6.03, 6.58, 7.30, 7.94, 2869.35, 3009.49, \quad \text{moy} = 739.5612, 288.$$

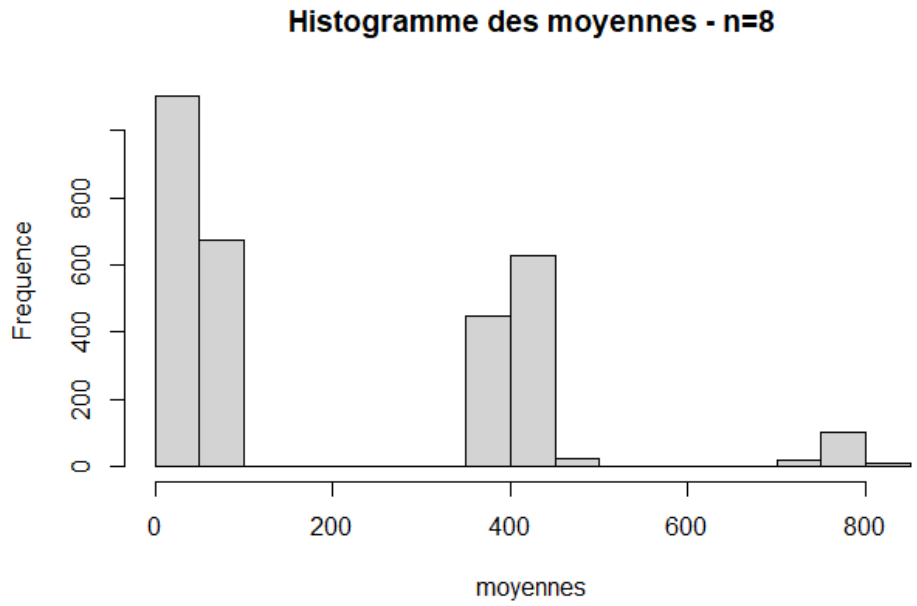
Enfin, si l'échantillon contient exactement l'un des deux plus grands nombres, la moyenne la plus élevée que nous pouvons obtenir est par l'échantillon

$$58.08, 80.72, 90.48, 103.19, 193.34, 196.32, 253.2, 3009.49, \quad \text{moy} = 498.1025;$$

la moyenne la plus basse que nous pouvons obtenir dans la même condition est à travers l'échantillon

$$4.59, 5.21, 6.03, 6.58, 7.30, 7.94, 7.96, 2869.35, \quad \text{moy} = 364.37.$$

En bref, les moyennes de l'échantillon tombent bien dans les tranches données.
Puisque la moyenne de la population est en fait 201.09, la moyenne de l'échantillon ne sera jamais à moins de 73.76 unités de la moyenne de la population, même si elle est un estimateur sans biais.



Ensuite, il y a le problème de la variance de l'échantillon : tout échantillon contenant au moins l'une des deux plus grandes valeurs aura une très grande variance, ce qui rendra les intervalles de confiance très larges. En somme, un échantillon aléatoire simple de taille $n = 8$ ne semble pas être un très bon plan d'échantillonnage dans ce cas.

(c) La marge d'erreur sur l'estimation pour un échantillon aléatoire simple est approximée par

$$2\sqrt{\hat{V}(\overline{y})} = 2\sqrt{\frac{s^2}{n}\left(1 - \frac{n}{N}\right)},$$

où s^2 est la variance de l'échantillon, $n = 8$ et $N = 37$. Il faut donc calculer la variance de l'échantillon.

```
> (x.ech <- sample(x,n,replace=FALSE))
[1]  30.90  21.13 196.32   5.21  90.48   9.08  39.98  17.66
> (x.moy=mean(x.ech))
[1] 51.345
> s.2=var(x.ech)
> (B=2*sqrt(s.2/n*(1-n/N)))
[1] 43.69876
> c(x.moy-B,x.moy+B)
[1]  7.646238 95.043762
```

Vos réponses peuvent différer, bien sûr, tant qu'elles sont justifiées. ■

10. On cherche à donner un estimé du nombre de touffes de mauvaises herbes d'un certain type dans un champ.

    (a) Quelle est la population et que sont les unités d'échantillonnage?

    (b) Comment pourrait-on construire une base de sondage pour cette tâche?

    (c) Comment pourrait-on sélectionner un échantillon aléatoire simple?

    (d) Si une unité d'échantillonnage est une superficie (1 m^2, par exemple), la taille choisie pour une unité d'échantillonnage a-t-elle une incidence sur la fiabilité des résultats?

    (e) Quelles considérations entreraient dans le choix de la taille des unités d'échantillonnage?

    Solution:

    (a) La population cible est constituée de tous les touffes de mauvaises herbes d'un certain type dans le champ, en supposant que cette notion de "touffe" soit bien définie : je suppose que nous parlons de zones "contiguës" où l'on trouve les mauvaise herbes. Ma suggestion pour les unités d'échantillonnage serait d'utiliser des parcelles de terrain carrées égales et disjointes. Cela crée deux problèmes : il est peu probable que l'union de toutes ces parcelles carrées couvre le champ et uniquement le champ (ce qui pose des problèmes de couverture), et je ne sais pas quoi faire d'un groupe qui chevaucherait deux ou plusieurs parcelles (ce qui poserait également des problèmes de mesure). J'imagine qu'une certaine attention peut être apportée à la disposition des carrés afin de minimiser le nombre de chevauchements ; de même, si les carrés sont suffisamment petits, leur union sera à peu près égale au champ dans son intégralité. En conséquence, je maintiens ma proposition initiale. La variable réponse serait alors le nombre de touffes de mauvaises herbes par unité d'échantillonnage, dénotée par u_j pour la $j$ième unité de ce type. L'attribut de population qui nous intéresse dans ce cas est

    $$\tau = \sum_{j=1}^{N} u_j.$$

    (b) Si des fonds sont disponibles, une photographie aérienne ou par satellite du champ pourrait alors être utilisée pour produire une grille numérotée de parcelles carrées. Sinon, une carte topographique du champ pourrait être utilisée dans le même but.

    (c) Une fois qu'une taille particulière de parcelle carrée a été sélectionnée (et donc que le nombre N d'unités dans la population étudiée a été fixé) et que la taille de l'échantillon n a été choisie afin d'estimer τ avec une limite d'erreur prescrite, un logiciel (ou une table de nombres aléatoires) peut être utilisé pour sélectionner un EAS de n entiers parmi les N premiers entiers. Chaque parcelle carrée correspondant à un entier choisi i sera ensuite examinée afin de produire sa réponse y_i, $i = 1, \ldots, n$. Si la résolution de la photographie est suffisamment élevée, on pourrait imaginer de l'utiliser pour compter les touffes de mauvaises herbes dans chaque unité d'échantillonnage. Sinon, nous devrons envoyer un étudiant diplômé pour faire le compte en personne (pourquoi pas...).

    (d) Tout d'abord, notez que le fait de changer la superficie des unités d'échantillonnage modifiera automatiquement les constantes N et n. Pour répondre à la question, si les unités d'échantillonnage sont trop petites, il y aura de nombreuses unités de ce type où aucune touffes de mauvaises herbes n'est située. Ainsi, modifier la superficie des unités d'échantillonnage modifie également la variance de la population σ^2. Nous courons également le risque de sélectionner un échantillon d'unités dans lesquelles peu ou pas de touffes de mauvaises herbes ne se retrouvent. Ce problème particulier peut affecter

directement l'estimation ponctuelle de τ. Peut-il aussi affecter également sa variance ? Rappelez-vous que la variance de τ est estimée par

$$\hat{V}(\tau) = N^2 \frac{s^2}{n} \left(1 - \frac{n}{N}\right).$$

La question à se poser est la suivante : le fait d'avoir une zone plus petite (ou plus grande) pour les unités d'échantillonnage, donc des valeurs plus grandes (ou plus petites de n, N) et des valeurs plus petites (ou plus grandes) de s^2 réduit-il ou augmente-t-il la valeur de $\hat{V}(\tau)$? Si $\frac{n}{N}$ demeure plus ou moins constant quelle que soit la surface de l'unité d'échantillonnage, il en va de même pour $\frac{N}{n}$. Le dernier terme d'intérêt serait alors Ns^2. Puisque nous avons émis l'hypothèse que s^2 diminue lorsque N augmente, et *vice-versa*, il se pourrait très bien que Ns^2 reste plus ou moins constant. Dans ce cas (et selon une succession d'hypothèses pas tout à fait probables), la surface de l'unité d'échantillonnage n'affecterait pas la précision de l'estimation.

(e) La marge d'erreur sur l'estimation, certainement. Consulter la réponse en partie (d) pour plus de renseignements. ■

11. Une population de $N = 5$ unités prend les valeurs $u_1 = 3$, $u_2 = 1$, $u_3 = 0$, $u_4 = 1$, $u_5 = 5$.

(a) Calculer la moyenne, μ, et la variance, σ^2, de cette population.

(b) Supposons qu'un échantillon aléatoire simple de taille 3 soit prélevé dans cette population. Si y_1, y_2, et y_3 représentent la première, la deuxième, et la troisième unité sélectionnées dans l'échantillon, respectivement, montrer que $P(y_3 = u_j) = \frac{1}{N}$.

(c) Énumérer tous les échantillons possibles de taille 3 qui peuvent être prélevés dans cette population.

(d) Pour chaque échantillon obtenu en (c), calculer sa moyenne $\overline{y}$.

(e) Attribuer une probabilité de sélection à chaque échantillon énuméré en (c) si un échantillonnage aléatoire simple est utilisé pour sélectionner l'un des échantillons.

(f) À l'aide des valeurs de $\overline{y}$ calculées en (d) et des probabilités spécifiées en (e), vérifier que

$$E(\overline{y}) = \sum_{\text{all } \overline{y}} \overline{y} p(\overline{y}) = \mu \quad \text{et} \quad V(\overline{y}) = \sum_{\text{all } \overline{y}} \overline{y}^2 p(\overline{y}) - \left[E(\overline{y})\right]^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).$$

(g) Quelle est la médiane, M, de la population des cinq unités?

(h) Déterminer la médiane, $\tilde{y}$, de chaque échantillon obtenu en (c). Utiliser ces valeurs et les probabilités spécifiées en (e) afin de déterminer $E(\tilde{y})$ et $V(\tilde{y})$.

(i) Comparer $\overline{y}$ et $\tilde{y}$ en tant qu'estimateurs de leurs paramètres de population respectifs, en faisant référence au biais d'échantillonnage et à la variabilité d'échantillonnage.

Solution:

(a) Selon les définitions,

$$\mu = \frac{1}{5}\sum_{i=1}^{5} u_i = \frac{1}{5}(3+1+0+1+5) = 2$$

$$\sigma^2 = \frac{1}{5}\sum_{i=1}^{5}(u_i-\mu)^2 = \frac{1}{5}\left((3-2)^2+(1-2)^2+(0-2)^2+(1-2)^2+(5-2)^2\right) = \frac{16}{5}$$

(b) Soient $A : y_1 \neq u_j$, $B : y_2 \neq u_j$ et $C : y_3 = u_j$. Alors, selon le théorème de la probabilité conditionnelle et la règle de la multiplication,

$$P(y_3 = u_j) = P(C) = \frac{P(A,B,C)}{P(A,B|C)} = \frac{P(A)P(B|A)P(C|A,B)}{P(A,B|C)}.$$

Par construction, nous avons $P(A,B|C) = 1$, d'où

$$P(C) = \frac{P(A)P(B|A)P(C|A,B)}{P(A,B|C)} = \frac{P(A)P(B|A)P(C|A,B)}{1} = P(A)P(B|A)P(C|A,B).$$

De plus, on note que $P(A) = \frac{N-1}{N}$, $P(B|A) = \frac{N-2}{N-1}$, et $P(C|A,B) = \frac{1}{N-2}$, de sorte que

$$P(C) = P(A)P(B|A)P(C|A,B) = \frac{N-1}{N}\cdot\frac{N-2}{N-1}\cdot\frac{1}{N-2} = \frac{1}{N},$$

ce qui complète la démonstration.

(c) Il existe $\binom{5}{3} = 10$ échantillons de taille $n = 3$:

(y_1, y_2, y_3)	**Échantillon**
(u_1, u_2, u_3)	$(3, 1, 0)$
(u_1, u_2, u_4)	$(3, 1, 1)$
(u_1, u_2, u_5)	$(3, 1, 5)$
(u_1, u_3, u_4)	$(3, 0, 1)$
(u_1, u_3, u_5)	$(3, 0, 5)$
(u_1, u_4, u_5)	$(3, 1, 5)$
(u_2, u_3, u_4)	$(1, 0, 1)$
(u_2, u_3, u_5)	$(1, 0, 5)$
(u_2, u_4, u_5)	$(1, 1, 5)$
(u_3, u_4, u_5)	$(0, 1, 5)$

(d) Les moyennes d'échantillon correspondantes sont

(y_1, y_2, y_3)	**Échantillon**	**Moyenne d'échantillon** $\overline{y}$
(u_1, u_2, u_3)	$(3, 1, 0)$	$4/3$
(u_1, u_2, u_4)	$(3, 1, 1)$	$5/3$
(u_1, u_2, u_5)	$(3, 1, 5)$	3
(u_1, u_3, u_4)	$(3, 0, 1)$	$4/3$
(u_1, u_3, u_5)	$(3, 0, 5)$	$8/3$
(u_1, u_4, u_5)	$(3, 1, 5)$	3
(u_2, u_3, u_4)	$(1, 0, 1)$	$2/3$
(u_2, u_3, u_5)	$(1, 0, 5)$	2
(u_2, u_4, u_5)	$(1, 1, 5)$	$7/3$
(u_3, u_4, u_5)	$(0, 1, 5)$	2

(e) Les probabilités de sélection d'un échantillon donné par échantillonnage aléatoire simple sont les suivantes

(y_1, y_2, y_3)	**Échantillon**	**Moyenne d'échantillon** $\overline{y}$	**Probabilité de sélection** $p(\overline{y})$
(u_1, u_2, u_3)	$(3, 1, 0)$	$4/3$	0.1
(u_1, u_2, u_4)	$(3, 1, 1)$	$5/3$	0.1
(u_1, u_2, u_5)	$(3, 1, 5)$	3	0.1
(u_1, u_3, u_4)	$(3, 0, 1)$	$4/3$	0.1
(u_1, u_3, u_5)	$(3, 0, 5)$	$8/3$	0.1
(u_1, u_4, u_5)	$(3, 1, 5)$	3	0.1
(u_2, u_3, u_4)	$(1, 0, 1)$	$2/3$	0.1
(u_2, u_3, u_5)	$(1, 0, 5)$	2	0.1
(u_2, u_4, u_5)	$(1, 1, 5)$	$7/3$	0.1
(u_3, u_4, u_5)	$(0, 1, 5)$	2	0.1

Un tableau un peu plus simple peut être construit si notre objectif est de trouver les probabilités de sélectionner un échantillon avec une moyenne d'échantillon particulière (bien que ce tableau ne corresponde pas tout à fait à ce qui est demandé, il facilitera légèrement les calculs dans la partie (f)).

Moyenne d'échantillon $\overline{y}$	**Probabilité de sélection $p(\overline{y})$**
2/3	0.1
4/3	0.2
5/3	0.1
2	0.2
7/3	0.1
8/3	0.1
3	0.2

(f) Selon les définitions,

$$E(\overline{y}) = \left(\frac{2}{3}+\frac{5}{3}+\frac{7}{3}+\frac{8}{3}\right)(0.1)+\left(\frac{4}{3}+2+3\right)(0.2) = \frac{22}{3}(0.1)+\frac{19}{3}(0.2) = 2$$
$$V(\overline{y}) = \left[\left(\frac{2}{3}\right)^2+\left(\frac{5}{3}\right)^2+\left(\frac{7}{3}\right)^2+\left(\frac{8}{3}\right)^2\right](0.1)+\left[\left(\frac{4}{3}\right)^2+2^2+3^2\right](0.2)-2^2$$
$$= \frac{142}{9}(0.1)+\frac{133}{9}(0.2)-4 = \frac{8}{15},$$

qui est en effet équivaut à $\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{16/5}{3}\left(\frac{5-3}{5-1}\right) = \frac{16}{15}\cdot\frac{2}{4} = \frac{8}{15}$.

(g) Ordonnons la population selon

$$w_1 = u_3 = 0 < w_2 = w_3 = u_2 = u_4 = 1 < w_4 = u_1 = 3 < w_5 = u_5 = 5.$$

Comme il y a $N = 5$ unités dans la population, la médiane est tout simplement $M = w_3 = u_2 = u_4 = 1$.

(h) Les médianes d'échantillon sont

(y_1, y_2, y_3)	**Échantillon**	**Médiane d'échantillon $\tilde{y}$**
(u_1, u_2, u_3)	$(3, 1, 0)$	1
(u_1, u_2, u_4)	$(3, 1, 1)$	1
(u_1, u_2, u_5)	$(3, 1, 5)$	3
(u_1, u_3, u_4)	$(3, 0, 1)$	1
(u_1, u_3, u_5)	$(3, 0, 5)$	3
(u_1, u_4, u_5)	$(3, 1, 5)$	3
(u_2, u_3, u_4)	$(1, 0, 1)$	1
(u_2, u_3, u_5)	$(1, 0, 5)$	1
(u_2, u_4, u_5)	$(1, 1, 5)$	1
(u_3, u_4, u_5)	$(0, 1, 5)$	1

Les probabilités correspondantes sont ainsi

Médiane d'échantillon $\tilde{y}$	**Probabilité de sélection $p(\tilde{y})$**
1	0.7
3	0.3

Selon les définitions, nous obtenons

$$E(\tilde{y}) = (1)(0.7)+(3)(0.3) = 1.6$$
$$V(\tilde{y}) = (1)^2(0.7)+(3)^2(0.3)-1.6^2 = 0.84$$

(i) D'après les calculs précédents, nous obtenons les biais suivants :

$$\text{Biais}(\overline{y}) = E(\overline{y} - \mu) = E(\overline{y}) - E(\mu) = \mu - \mu = 0$$
$$\text{Biais}(\tilde{y}) = E(\tilde{y} - M) = E(\tilde{y}) - E(M) = 1.6 - 1 = 0.6$$

En d'autres termes, $\overline{y}$ est un estimateur sans biais de μ, alors que $\tilde{y}$ est un estimateur biaisé de M. D'autre part, nous avons les variances suivantes :

$$V(\overline{y}) = 0.53$$
$$V(\tilde{y}) = 0.84$$

Nous obtenons ainsi un estimateur sans biais avec une variance assez faible ($\overline{y}$) et un estimateur biaisé où le biais relatif est assez grand et avec une variance plus élevée ($\tilde{y}$). Lequel est le "meilleur" estimateur du paramètre qu'il tente d'estimer ? Cela dépend des préférences de la personne qui fait l'expérience, bien sûr. Mais nous pouvons utiliser une mesure de la variation totale pour répondre à la question.
La variation totale d'un estimateur $\hat{\theta}$ par rapport à la valeur actuelle du paramètre θ est donnée par

$$\text{EQM}(\hat{\theta}) = V(\hat{\theta}) + \Big(\text{Biais}(\hat{\theta})\Big)^2.$$

Ici, nous obtenons

$$\text{EQM}(\overline{y}) = V(\overline{y}) + (\text{Biais}(\overline{y}))^2 = 0.53 + 0^2 = 0.53$$
$$\text{EQM}(\tilde{y}) = V(\tilde{y}) + (\text{Biais}(\tilde{y}))^2 = 0.84 + (0.6)^2 = 1.2$$

Il semblerait donc que $\overline{y}$ est un 'meilleur' estimateur de μ que $\tilde{y}$ n'est un estimateur de M (dans cet example, du moins). ■

12. La variance d'une population de N unités prenant les valeurs u_j, $j = 1, \ldots, N$ est donnée par

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(u_j - \mu)^2, \quad \text{où} \quad \mu = \frac{1}{N}\sum_{j=1}^{N} u_j.$$

Démontrer que

$$\sigma^2 = \frac{1}{N}\left[\sum_{j=1}^{N} u_j^2 - \frac{1}{N}\left(\sum_{j=1}^{N} u_j\right)^2\right] = \frac{1}{N}\sum_{j=1}^{N} u_j^2 - \mu^2.$$

Démonstration: Selon la définition,

$$\begin{aligned}
\sigma^2 &= \frac{1}{N}\sum_{j=1}^{N}(u_j - \mu)^2 = \frac{1}{N}\sum_{j=1}^{N}(u_j^2 - 2u_j\mu + \mu^2) = \frac{1}{N}\sum_{j=1}^{N} u_j^2 - \frac{2\mu}{N}\sum_{j=1}^{N} u_j + \mu^2 \\
&= \frac{1}{N}\sum_{j=1}^{N} u_j^2 - 2\mu^2 + \mu^2 = \frac{1}{N}\sum_{j=1}^{N} u_j^2 - \mu^2 = \frac{1}{N}\sum_{j=1}^{N} u_j^2 - \left(\frac{1}{N}\sum_{j=1}^{N} u_j\right)^2 \\
&= \frac{1}{N}\sum_{j=1}^{N} u_j^2 - \frac{1}{N^2}\left(\sum_{j=1}^{N} u_j\right)^2 = \frac{1}{N}\left[\sum_{j=1}^{N} u_j^2 - \frac{1}{N}\left(\sum_{j=1}^{N} u_j\right)^2\right],
\end{aligned}$$

ce qui complète la "démonstration" (que nous avions déjà vue en classe). ■

13. Les gestionnaires de ressources d'une forêt giboyeuse (riche en gibier) s'inquiètent de la taille des populations de cerfs et de lapins en hiver. Pour donner un estimé de la taille de la population, ils proposent d'utiliser le nombre moyen d'excréments de lapins et de cerfs par parcelle de 30 m^2. La forêt est divisée en 10 000 telles parcelles à l'aide d'une photo aérienne. Un échantillon aléatoire simple de 250 parcelles a été prélevé et le nombre d'excréments de lapins et de cerfs a été observé dans chaque parcelle. Les résultats de ce sondage sont résumés dans le tableau ci-dessous.

	cerfs	**lapins**
moyenne d'échantillon	2.40	4.12
variance d'échantillon	0.61	0.93

(a) Donner des estimations du nombre moyen d'excréments par parcelle pour les cerfs et les lapins, et donner un estimé de la marge d'erreur sur l'estimation pour chacun d'entre eux.

(b) Combien de parcelles supplémentaires faudrait-il échantillonner afin de donner un estimé du nombre moyen d'excréments de cerfs par parcelle avec une marge d'erreur de 0.05 ?

Solution:

(a) En posant $N = 10,000$ et $n = 250$, on obtient les résultats suivants.

Cerfs: le nombre moyen d'excréments par parcelle de 30 m^2 est $\overline{y}_{\text{cerfs}} = 2.40$, et la marge approximative sur l'erreur d'estimation est donnée par

$$B = 2\sqrt{\hat{V}\left(\overline{y}_{\text{cerfs}}\right)} = 2\sqrt{\frac{s^2_{\text{cerfs}}}{n}\left(1 - \frac{n}{N}\right)} = 2\sqrt{\frac{0.61}{250}\left(1 - \frac{250}{10,000}\right)} \approx 0.0975,$$

d'où $2.40 \pm 0.0975 \equiv [2.302, 2.498]$ forme les extrémités de l'intervalle de confiance à environ 95%.

Lapins: le nombre moyen d'excréments par parcelle de 30 m^2 est $\overline{y}_{\text{lapin}} = 4.12$, et la marge approximative sur l'erreur d'estimation est donnée par

$$B = 2\sqrt{\hat{V}\left(\overline{y}_{\text{lapin}}\right)} = 2\sqrt{\frac{s^2_{\text{lapin}}}{n}\left(1 - \frac{n}{N}\right)} = 2\sqrt{\frac{0.93}{250}\left(1 - \frac{250}{10000}\right)} \approx 0.1204,$$

d'où $4.12 \pm 0.1204 \equiv [4.000, 4.240]$ forme les extrémités de l'intervalle de confiance à environ 95%.

(b) La taille n d'un EAS provenant d'une population de taille N et de variance σ^2 requise afin d'atteindre une marge sur l'erreur d'estimation B de la moyenne est

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2}, \quad \text{où } D = \frac{B^2}{4}.$$

Lorsque nous ne connaissons pas la variance de la population (comme c'est le cas ici), nous pouvons l'estimer en utilisant la variance de l'échantillon. En utilisant cette formule, et en approximant σ^2_{cerfs} par s^2_{cerfs}, on obtient

$$n_{\text{cerfs}} \geq \frac{10,000 s^2_{\text{cerfs}}}{(10000-1)\frac{(0.05)^2}{4} + s^2_{\text{cerfs}}} = \frac{10,000(0.61)}{9999\frac{0.0025}{4} + 0.61} = 889.23.$$

Comme nous avions déjà sélectionné 250 parcelles lors de l'enquête pilote, nous devons donc sélectionner au moins $890 - 250 = 640$ observations supplémentaires afin d'atteindre la marge requise. ■

14. Une vérificatrice choisit au hasard 20 comptes clients parmi les 573 comptes d'une certaine entreprise. La vérificatrice répertorie le montant de chaque compte (en dollars) et vérifie si les documents sous-jacents sont conformes aux procédures énoncées. Les données sont les suivantes:

Client	**Montant**	**Conforme?**	**Client**	**Montant**	**Conforme?**
1	278	O	11	188	N
2	192	O	12	212	N
3	310	O	13	92	O
4	94	N	14	56	O
5	86	O	15	142	O
6	335	O	16	37	O
7	310	N	17	186	N
8	290	O	18	221	O
9	221	O	19	219	N
10	168	O	20	305	O

(a) Donner un estimé du total des comptes à recevoir pour les 573 comptes de l'entreprise et donner une approximation de la limite de l'erreur d'estimation. Le montant moyen des créances de l'entreprise dépasse-t-il 250$? Expliquer.

(b) Quelle taille d'échantillon est nécessaire afin de donner un estimé du montant total des comptes à recevoir avec une marge d'erreur sur l'estimation de $10,000?

(c) Donner un estimé de la proportion des comptes de l'entreprise qui n'est pas conforme aux procédures énoncées. Donnez une approximation de la marge d'erreur sur l'estimation. La proportion réelles des comptes qui se conforment aux procédures énoncées dépasse-t-elle 80%? Expliquer.

(d) Pour une marge d'erreur sur l'estimation de 0.12, déterminer la taille de l'échantillon nécessaire pour donner un estimé de la proportion de comptes qui ne sont pas conformes aux procédures énoncées dans les deux cas suivants:

i. on utilise un estimé de p donné par les 20 comptes échantillonnés, ou

ii. aucun estimé de p n'est disponible.

Solution: Pour les comptes clients échantillonnés $i = 1, \ldots, 20$, y_i désigne le montant dû; la conformité aux procédures est désignée par la variable

$$w_i = \begin{cases} 1, & \text{si les documents sous-jacents sont conformes aux procédures énoncées} \\ 0, & \text{s'ils ne le sont pas} \end{cases}$$

(a) On peut établir que $\overline{y} = 197.1$ et $s_{\mathrm{y}} \approx 90.86$. L'estimation du total des créances pour les 573 comptes de l'entreprise, en se référant à l'EAS est donc

$$\tau = N\overline{y} = 573(197.1) = 112938.3$$

et la marge d'erreur sur l'estimation du total est environ

$$2\sqrt{\hat{V}(\tau)} = 2Ns\sqrt{\frac{1}{n} - \frac{1}{N}} = 2(573)(90.86)\sqrt{\frac{1}{20} - \frac{1}{573}} = 22873.24.$$

Ainsi, $112938.3 \pm 22873.24 \equiv [90065.06, 135811.5]$ est un I.C. à environ 95% pour le total des comptes à recevoir selon cet EAS.

Par ailleurs, la moyenne des comptes débiteurs, μ, est estimée sans biais par $\overline{y} = 197.1$. La marge d'erreur sur l'estimation correspondante peut facilement être obtenue à partir de la borne sur l'erreur d'estimation pour le total : il suffit de diviser cette dernière par $N = 573$, ce qui donne une borne d'environ 39.92.

Cela signifie que $197.1 \pm 39.92 \equiv [157.2, 237.0]$ forme un intervalle de confiance à environ 95% pour la moyenne μ. Puisque 250 se situe au-dessus de cet intervalle, il est très peu probable que $\mu > 250$.

(b) La taille n d'un EAS provenant d'une population de taille N et de variance σ^2 requise afin d'atteindre une marge sur l'erreur d'estimation B du total est

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2} \approx \frac{Ns^2}{(N-1)D + s^2}, \quad \text{où } D = \frac{B^2}{4N^2}.$$

En utilisant cette formule, on obtient

$$n_{\text{y}} \geq \frac{573(90.86)^2}{572\frac{(10{,}000)^2}{4(573)^2} + (90.86)^2} = 91.3.$$

(c) On calcule également que $\overline{w}_{=}0.7$ et $s_{\text{w}} \approx 0.47$. L'estimation de la proportion p de comptes débiteurs dont les documents sous-jacents ne sont pas conformes aux procédures énoncées est donc la suivante $\hat{p} = 1 - 0.7 = 0.3$ et la marge d'erreur sur l'estimation pour cette proportion est approximée par

$$2\sqrt{\hat{V}(\hat{p})} = 2\sqrt{\frac{\hat{p}\hat{q}}{n-1}\left(1 - \frac{n}{N}\right)} = 2\sqrt{\frac{(0.3)(0.7)}{19}\left(1 - \frac{20}{573}\right)} \approx 0.206,$$

d'où $0.3 \pm 0.206 \equiv [0.094, 0.506]$ forme un I.C. à environ 95% pour la proportion de comptes non conformes.

La marge d'erreur sur l'estimation pour les proportions est invariante sous les permutations de (p, q); $0.7 \pm 0.206 \equiv [0.494, 0.906]$ forme un I.C. à environ 95% pour la proportion q de comptes clients conformes. Puisque 0.8 se retrouve à l'intérieur de l'intervalle de confiance, nous ne pouvons pas dire avec la certitude requise si q dépasse 0.8 ou non.

(d) La taille n d'un EAS provenant d'une population de taille N et de variance σ^2 requise afin d'atteindre une marge sur l'erreur d'estimation B de la proportion p est

$$n = \frac{Npq}{(N-1)D + pq}, \quad \text{où } D = \frac{B^2}{4}.$$

i. Si l'on approxime la proportion sur la force des 20 comptes clients choisis, on utilise $p = 0.3$ et $q = 0.7$, d'où

$$n \geq \frac{573(0.3)(0.7)}{572\frac{(0.12)^2}{4} + (0.3)(0.7)} = 53.0275.$$

ii. Sans valeur de p, on obtient la taille de l'échantillon en prenant $p = q = 0.5$. Alors,

$$n \geq \frac{573(0.5)(0.5)}{572\frac{(0.12)^2}{4} + (0.5)(0.5)} = 62.03$$

Bien sûr, on peut faire d'une pierre, deux coup en utilisant $n \geq 92$ et en obtenant à la fois une marge d'erreur sur l'estimation de $10,000$ pour le total des comptes débiteurs, de 0.12 pour la proportion réelle des comptes qui ne respectent pas les procédures énoncées, au niveau de signification standard de l'échantillonnage. ■

15. Considérer les données suivantes extraites d'un article de presse de 1992 : 56% des femmes et 45% des hommes ont déclaré que le gouvernement américain devrait faire de la lutte contre la criminalité et la violence une priorité absolue. Les résultats proviennent d'un échantillon national de $n_1 = 611$ femmes et $n_2 = 609$ hommes. La marge d'erreur sur l'estimation est de 0.03 pour l'échantillon combiné, et de 0.06 dans chacun des sous-populations.

    (a) Déterminer un I.C. (à environ 95%) de la proportion des femmes qui pensent que la lutte contre la criminalité et la violence devrait être une priorité absolue.

    (b) Déterminer un I.C. (à environ 95%) de la proportion des hommes qui pensent que la lutte contre la criminalité et la violence devrait être une priorité absolue.

    (c) Déterminer un I.C. (à environ 95%) de la différence entre la proportion des femmes et la proportion des hommes qui pensent que la lutte contre la criminalité et la violence devrait être une priorité absolue.

    (d) Y a-t-il une différence statistiquement significative entre les opinions des femmes et des hommes sur la question de savoir si la lutte contre la criminalité et la violence devrait être une priorité absolue. Expliquer.

    Solution:

    (a) Soit p_{f} la proportion qui nous intéresse. Nous avons $n_{\mathrm{f}} = 611$, $\frac{n_{\mathrm{f}}}{N_{\mathrm{f}}} \approx 0$ et $\hat{p}_{\mathrm{f}} = 0.56$. La marge d'erreur sur l'estimation de p_{f} est alors approximée par

$$2\sqrt{\hat{V}\left(\hat{p}_{\mathrm{f}}\right)} \approx 2\sqrt{\frac{\hat{p}_{\mathrm{f}}\left(1-\hat{p}_{\mathrm{f}}\right)}{n_{\mathrm{f}}-1}} = 2\sqrt{\frac{(0.56)(0.44)}{610}} \approx 0.0402;$$

    un I.C. approximatif à 95% pour p_{w} est donné par $\hat{p}_{\mathrm{f}} \pm 2\sqrt{\hat{V}\left(\hat{p}_{\mathrm{f}}\right)} \equiv 0.56 \pm 0.04 \equiv [0.52, 0.60]$.

    (b) Soit p_{h} la proportion qui nous intéresse. Nous avons $n_{\mathrm{h}} = 609$, $\frac{n_{\mathrm{h}}}{N_{\mathrm{h}}} \approx 0$ et $\hat{p}_{\mathrm{h}} = 0.45$. La marge d'erreur sur l'estimation de p_{h} est alors approximée par

$$2\sqrt{\hat{V}\left(\hat{p}_{\mathrm{h}}\right)} \approx 2\sqrt{\frac{\hat{p}_{\mathrm{h}}\left(1-\hat{p}_{\mathrm{h}}\right)}{n_{\mathrm{h}}-1}} = 2\sqrt{\frac{(0.45)(0.55)}{608}} \approx 0.0403;$$

    un I.C. approximatif à 95% pour p_{h} est donné par $\hat{p}_{\mathrm{h}} \pm 2\sqrt{\hat{V}\left(\hat{p}_{\mathrm{h}}\right)} \equiv 0.45 \pm 0.04 \equiv [0.41, 0.49]$.

    (c) Soit $p_{\mathrm{d}} = p_{\mathrm{f}} - p_{\mathrm{h}}$ la difference recherchée, approximée par $\hat{p}_{\mathrm{d}} = \hat{p}_{\mathrm{f}} - \hat{p}_{\mathrm{h}} = 0.56 - 0.45 = 0.11$. Afin de trouver une marge d'erreur approximative sur l'estimation, notons tout d'abord que

$$V\left(\hat{p}_{\mathrm{d}}\right) = V\left(\hat{p}_{\mathrm{f}} - \hat{p}_{\mathrm{h}}\right) = V(\hat{p}_{\mathrm{f}}) + V\left(\hat{p}_{\mathrm{h}}\right) + 2\mathrm{Cov}(\hat{p}_{\mathrm{f}}, \hat{p}_{\mathrm{h}}).$$

    Si $\hat{p}_{\mathrm{w}}$ et $\hat{p}_{\mathrm{m}}$ sont indeépendents, cette variance devient

$$V\left(\hat{p}_{\mathrm{d}}\right) = V(\hat{p}_{\mathrm{f}}) + V\left(\hat{p}_{\mathrm{h}}\right).$$

    Les deux termes sont approximés par $\hat{V}(\hat{p}_{\mathrm{f}})$ et $\hat{V}(\hat{p}_{\mathrm{h}})$ (voir un peu plus haut) de sorte qu'une marge d'erreur approximative sur l'estimation est donnée par

$$2\sqrt{\hat{V}\left(\hat{p}_{\mathrm{d}}\right)} = 2\sqrt{\frac{\hat{p}_{\mathrm{f}}\left(1-\hat{p}_{\mathrm{f}}\right)}{n_{\mathrm{f}}-1} + \frac{\hat{p}_{\mathrm{h}}\left(1-\hat{p}_{\mathrm{h}}\right)}{n_{\mathrm{h}}-1}} \approx 2\sqrt{0.00081} \approx 0.05695636;$$

cela correspond aux informations de l'énoncé du problème. Par ailleurs, la marge d'erreur sur l'estimation pour l'échantillon combiné est en effet d'environ 0.03, puisque

$$2\sqrt{\frac{\left(\frac{0.56(611)+0.45(609)}{1220}\right)\left(1-\frac{0.56(611)+0.45(600)}{1220}\right)}{1219}} \approx 0.02864017.$$

Un intervalle de confiance à environ 95% pour la différence entre les deux proportions est donné par

$$(\hat{p}_{\mathrm{f}} - \hat{p}_{\mathrm{h}}) \pm 2\sqrt{\hat{V}\left(\hat{p}_{\mathrm{f}} - \hat{p}_{\mathrm{h}}\right)} \approx 0.11 \pm 0.057 \equiv= [0.053, 0.167].$$

(d) Les données semblent confirmer qu'il y a une différence réelle: si nous répétions cette enquête 100 fois, la différence réelle se situerait dans l'intervalle de confiance correspondant environ 95 fois. Il y a donc 19 chances sur 20 pour que la différence réelle se situe à l'intérieur de l'IC 95% de la partie (c). Puisque l'intervalle entier se trouve dans l'axe réel positif, cela confirme fortement la différence. ■

16. Un échantillon aléatoire $y_1, \ldots, y_n$ est prélevé d'une population de taille N, dont la moyenne est μ et la variance σ^2. Considérons la combinaison linéaire $t = a_1y_1 + \cdots + a_ny_n$, où $a_1, \ldots, a_n$ sont des constantes.

(a) Montrer que pour que t soit un estimateur sans biais de μ, on doit avoir $a_1 + \cdots + a_n = 1$.

(b) Montrer que

$$V(t) = \frac{N\sigma^2}{N-1}\sum_{i=1}^{n} a_i^2 - \frac{\sigma^2}{N-1}\left(\sum_{i=1}^{n} a_i\right)^2.$$

(c) Supposons que t soit un estimateur sans biais de μ. Trouver les valeurs de a_i qui minimisent la variance de t, sujettes à la restriction donnée en (a).

(d) Discuter de l'implication du résultat obtenu en (c) par rapport à l'utilisation de la moyenne empirique provenant d'un EAS en tant qu'estimateur de la moyenne μ.

Solution:

(a) Suppose that t is an unbiased estimator of μ. Then

$$0 = \mu - E(t)\mu - E\left(\sum_{i=1}^{n} a_iy_i\right) = \mu - \sum_{i=1}^{n} a_iE(y_i) = \mu - \sum_{i=1}^{n} a_i\mu = \mu - \mu\sum_{i=1}^{n} a_i = \mu\left(1 - \sum_{i=1}^{n} a_i\right).$$

As a result, either $\mu = 0$ or $a_1 + \cdots + a_n = 1$, or both. But I am providing an example where t is an unbiased estimator of μ and $a_1 + \cdots + a_n \neq 1$. Let $N = 3$, $n = 2$, and $u_1 = -1$, $u_2 = 0$, $u_3 = 1$. Clearly,

$$\mu = E(u) = \sum_{i=1}^{3} u_ip(u_i) = \frac{1}{3}(-1 + 0 + 1) = 0.$$

Now, let $t = y_1 + y_2$. We have

$$E(t) = E(y_1) + E(y_2) = \mu + \mu = 2\mu = 0 = \mu,$$

so that t is an unbiased estimator of μ. However, $a_1 = 1$ and $a_2 = 1$, which yields $a_1 + a_2 = 2 \neq 1$.

(b) Using the formula for the variance of a linear combination of random variables, we have

$$\begin{aligned} V(t) &= \sum_{i=1}^{n} a_i^2V(y_i) + \sum_{j\neq k} a_ja_k\mathrm{Cov}(y_j, y_k) = \sum_{i=1}^{n} a_i^2\sigma^2 - \sum_{j\neq k} a_ja_k\frac{\sigma^2}{N-1} \\ &= \sigma^2\sum_{i=1}^{n} a_i^2 - \frac{\sigma^2}{N-1}\sum_{j\neq k} a_ja_k = \sigma^2\sum_{i=1}^{n} a_i^2 - \frac{\sigma^2}{N-1}\left[\left(\sum_{i=1}^{n} a_i\right)^2 - \sum_{i=1}^{n} a_i^2\right] \\ &= \sigma^2\sum_{i=1}^{n} a_i^2 - \frac{\sigma^2}{N-1}\left(\sum_{i=1}^{n} a_i\right)^2 + \frac{\sigma^2}{N-1}\sum_{i=1}^{n} a_i^2 = \frac{N\sigma^2}{N-1}\sum_{i=1}^{n} a_i^2 - \frac{\sigma^2}{N-1}\left(\sum_{i=1}^{n} a_i\right)^2. \end{aligned}$$

(c) We would like to minimize $f(a_1, \ldots, a_n) = V(t)$ subject to $g(a_1, \ldots, a_n) = a_1 + \cdots + a_n - 1 = 0$. Since

$$\nabla f(a_1, \ldots, a_n) = \frac{2\sigma^2}{N-1}\left(Na_1 - \sum_{i=1}^{n} a_i, \ldots, Na_n - \sum_{i=1}^{n} a_i\right) \quad \text{and} \quad \nabla g(a_1, \ldots, a_n) = (1, \ldots, 1),$$

we want to find solutions to the following system of equations:

$$\frac{2\sigma^2}{N-1}\left(Na_j - \sum_{i=1}^{n} a_i\right) = \lambda, \quad j = 1, \ldots, n$$
$$\sum_{i=1}^{n} a_i = 1$$

This yields

$$a_j = \frac{1}{N}\left(\lambda\frac{N-1}{2\sigma^2} + 1\right), \quad j = 1, \ldots, n.$$

Finally, note that

$$1 = \sum_{j=1}^{n} a_j = \sum_{j=1}^{n} \frac{1}{N}\left(\lambda\frac{N-1}{2\sigma^2} + 1\right) = \frac{n}{N}\left(\lambda\frac{N-1}{2\sigma^2} + 1\right).$$

As a result, $a_j^* = \frac{1}{n}$ $(j = 1, \ldots, n)$ is the sole critical point of f subject to $g = 0$. That it is a minimizer is easy to see since there is at least one point on $g = 0$ for which the variance is greater than $f(a_1^*, \ldots, a_L^*)$:

$$f(1, 0, \ldots, 0) = \sigma^2 > f(a_1^*, \ldots, a_L^*) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).$$

(d) Amongst all unbiased estimators of the population mean built from linear combinations of the (simply random) sampled units' variates' values, the sample mean has the smallest variance. It is thus the most accurate of all such estimators. And that's good news, because it sure seems to me that's the only one I've ever seen. Do we use any other estimators in practice? ■

17. On cherche à donner un estimé de la distance quotidienne moyenne parcourue durant la saison hivernale 2012 en Ontario par certains types de véhicules. La consommation de carburant quotidienne est aussi d'intérêt, tout comme la proportion des véhicules qui ne sont pas utilisés. Un EAS est prélevé à même la flotte Ontarienne (de taille $N = 7,868,359$); les données relatives aux répondants sont recueillies dans le fichier `Autos_EAS.xlsx`. Discuter des enjeux pouvant venir influencer la qualité des données. Donner un sommaire numérique et visuel des données de l'échantillon réalisé. Donner un intervalle de confiance pour chaque moyenne de population recherchée, à environ 95%, avec coefficient de variation correspondant. [La majorité des notes seront attribuées pour la discussion et la présentation des résultats.]

Solution: Nous ne présentons que les calculs. ■

Chapitre 3 – Échantillonnage aléatoire stratifié

18. Les valeurs de la variable de réponse d'une population sont: $u_1 = 2$, $u_2 = 3$, $u_3 = 4$, $u_4 = 5$, $u_5 = 7$, $u_6 = 9$.

    (a) Déterminer la moyenne μ et la variance σ^2 de la population.

    (b) Calculer la moyenne et la variance de $\overline{y}$ pour un échantillon aléatoire simple de taille 4 de cette population.

    (c) Supposons que la population soit divisée en deux strates: la strate 1 contient $u_1 = 2$, $u_2 = 3$, $u_3 = 4$, tandis que la strate 2 contient $u_4 = 5$, $u_5 = 7$, $u_6 = 9$. Déterminer la moyenne et la variance empirique dans chacune des deux strates.

    (d) Énumérer tous les échantillons de taille 4 qui peuvent être sélectionnés en choisissant des échantillons aléatoires simples de 2 unités dans chacune des strates. Pour chaque échantillon global de taille 4, donner la probabilité qu'il soit sélectionné.

    (e) Pour chaque échantillon obtenu en (d), calculer la moyenne stratifiée $\overline{y}_{\text{STR}}$ de l'échantillon.

    (f) Vérifier que $E(\overline{y}_{\text{STR}}) = \mu$ et

$$V(\overline{y}_{\text{STR}}) = \frac{1}{N^2} \sum_{i=1}^{L} N_i^2 \frac{\sigma_i^2}{n_i} \left(\frac{N_i - n_i}{N_i - 1} \right).$$

    (g) Pour la population considérée, comparer $\overline{y}_{\text{EAS}}$ et $\overline{y}_{\text{STR}}$ comme estimateurs de μ en termes de biais d'échantillonnage et de variabilité. Pourquoi la variance de $\overline{y}_{\text{EAS}}$ est-elle plus élevée que celle de $\overline{y}_{\text{STR}}$?

Solution:

(a) C'est simple: nous avons un population de taille $N = 6$ et

$$\mu = (2 + 3 + 4 + 5 + 7 + 9)/6 = 30/6 = 5$$
$$\sigma^2 = (2^2 + 3^2 + 4^2 + 5^2 + 7^2 + 9^2)/6 - 5^2 = 17/3 = 5.6667.$$

(b) Pour un EAS de taille $n = 4$,

$$E(\overline{y}_{\text{EAS}}) = \mu = 5 \text{ et } V(\overline{y}_{\text{EAS}}) = \frac{\sigma^2}{n} \left(\frac{N - n}{N - 1} \right) = \frac{17/3}{4} \left(\frac{6 - 4}{6 - 1} \right) = 17/30 = 0.56667.$$

(c) Puisque $N_1 = N_2 = 3$, nous obtenons

$$\mu_1 = (2 + 3 + 4)/3 = 3, \quad \sigma_1^2 = (2^2 + 3^2 + 4^2)/3 - 3^2 = 2/3$$
$$\mu_2 = (5 + 7 + 9)/3 = 7, \quad \sigma_2^2 = (5^2 + 7^2 + 9^2)/3 - 7^2 = 8/3.$$

(d) Il y a $\binom{3}{2} \cdot \binom{3}{2} = 9$ différents échantillons de taille $(2, 2)$:

Échantillon	$\overline{y}_{\text{STR}}$
$(2, 3, 5, 7)$	4.25
$(2, 4, 5, 7)$	4.50
$(3, 4, 5, 7)$	4.75
$(2, 3, 5, 9)$	4.75
$(2, 4, 5, 9)$	5.00
$(3, 4, 5, 9)$	5.25
$(2, 3, 7, 9)$	5.25
$(2, 4, 7, 9)$	5.50
$(3, 4, 7, 9)$	5.75

(e) Voir réponse précédente.

(f) On calcule les quantités recherchées directement:

$$E(\overline{y}_{\text{STR}}) = (4.25 + 4.5 + 4.75 + 4.75 + 5 + 5.25 + 5.25 + 5.5 + 5.75)/9 = 5$$
$$V(\overline{y}_{\text{STR}}) = (4.25^2 + 4.5^2 + 4.75^2 + 4.75^2 + 5^2 + 5.25^2 + 5.25^2 + 5.5^2 + 5.75^2)/9 - 5^2 = 5/24 = 0.2083;$$

cette dernière quantité s'obtient aussi directement à l'aide de la formule:

$$V(\overline{y}_{\text{STR}}) = \frac{1}{N^2}\sum_{i=1}^{2} N_i^2 \frac{\sigma_i^2}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right) = \frac{1}{6^2}\left[3^2 \cdot \frac{2/3}{2}\left(\frac{3-2}{3-1}\right) + 3^2 \cdot \frac{8/3}{2}\left(\frac{3-2}{3-1}\right)\right] = 5/24.$$

(g) Ni l'un, ni l'autre des estimateurs n'admet de biais d'échantillonnage, mais l'estimateur de la moyenne de l'échantillon stratifié est beaucoup plus précis selon la mesure de la variation totale. Cela est dû au fait que l'ensemble des moyennes d'échantillons STR possibles est un sous-ensemble propre de l'ensemble des moyennes d'échantillons possibles selon un EAS. Par conséquent, $V(\overline{y}) = V(\overline{y}_{\text{st}}) + K^2$ pour un certain K. ■

19. Pour une population divisée en M strates distinctes, le coût total d'obtention d'un échantillon STR de taille n (contenant n_i unités dans la $i$ème strate, $i = 1, \ldots, M$) est donné par

$$C = c_0 + \sum_{i=1}^{M} c_i n_i^{3/4}.$$

Si l'on souhaite utiliser $\overline{y}_{\text{STR}}$ afin de donner un estimé de la moyenne de la population μ, déterminer les poids d'échantillonnage qui minimiseront $V(\overline{y}_{\text{STR}})$ en respectant la contrainte de coût total ci-dessus.

Solution: Nous utilisons la méthode des multiplicateurs de Lagrange afin de minimiser la fonction

$$f(n_1, \ldots, n_M) = V(\overline{y}_{\text{st}}) = \frac{1}{N^2} \sum_{i=1}^{M} N_i^2 \frac{\sigma_i^2}{n_i} \left(\frac{N_i - n_i}{N_i - 1} \right),$$

soumise à la contrainte

$$g(n_1, \ldots, n_M) = c_0 + \sum_{i=1}^{M} c_i n_i^{3/4} - C = 0.$$

En différentiant, on obtient

$$\nabla f(n_1, \ldots, n_M) = -\frac{1}{N^2} \left(\frac{N_1^3 \sigma_1^2}{n_1^2 (N_1 - 1)}, \cdots, \frac{N_M^3 \sigma_M^2}{n_M^2 (N_M - 1)} \right)$$
$$\nabla g(n_1, \ldots, n_M) = \frac{3}{4} \left(\frac{c_1}{n_1^{1/4}}, \cdots, \frac{c_M}{n_M^{1/4}} \right)$$

On résoud $\nabla f = \lambda \nabla g$ en fonction de $(n_1, \ldots, n_M)$ et l'on obtient:

$$(n_1, \ldots, n_M) = -\left(\frac{4}{3N^2 \lambda} \right)^{4/7} \left(\left(\frac{N_1^3 \sigma_1^2}{c_1 (N_1 - 1)} \right)^{4/7}, \cdots, \left(\frac{N_M^3 \sigma_M^2}{c_M (N_M - 1)} \right)^{4/7} \right).$$

Puisque $n = n_1 + \cdots + n_M$, le schéma général d'allocation optimale est donné par

$$w_i = \frac{n_i}{n} = \frac{\left(\frac{N_i^3 \sigma_i^2}{c_i (N_i - 1)} \right)^{4/7}}{\sum_{k=1}^{M} \left(\frac{N_k^3 \sigma_k^2}{c_k (N_k - 1)} \right)^{4/7}}, \quad i = 1, \ldots, M.$$

Si nous supposons en outre que $N_i - 1 \approx N_i$, $i = 1, \ldots, M$, les poids d'échantillonnage sont approximativement les suivants

$$w_i \approx \frac{\left(\frac{N_i^2 \sigma_i^2}{c_i} \right)^{4/7}}{\sum_{k=1}^{M} \left(\frac{N_k^2 \sigma_k^2}{c_k} \right)^{4/7}}, \quad i = 1, \ldots, M,$$

ce qui n'est pas bien différent du cas où nous ne faisons pas l'hypothèse simplificatrice. ■

20. Une chercheuse souhaite donner un estimé du revenu moyen des employés d'une grande entreprise de Montréal. Les employés sont répertoriés par ancienneté (en général, le salaire augmente avec l'ancienneté). Discuter des mérites relatifs de l'EAS et de l'échantillonnage STR dans ce cas. Laquelle de ces approches devrait-on préconiser? À quoi ressemblerait le plan d'échantillonnage ?

Solution: En général, un STR présente plusieurs avantages par rapport à un EAS :

(a) un STR offre une plus grande précision (variance plus faible) qu'un EAS de même taille;

(b) en raison de cette plus grande précision, la taille requise par un STR afin d'obtenir la même précision qu'un EAS est en général plus petite – un STR est ainsi moins dispendieux, en géeral;

(c) un STR peut fournir une protection contre les échantillons "non représentatifs".

Le principal inconvénient du STR est qu'il peut nécessiter un effort administratif plus important que le EAS.

Cependant, dans ce cas, nous disposons déjà d'une base de sondage stratifiée : les dossiers répertoriant les employés par ancienneté. Comme les salaires augmentent avec l'ancienneté (en général), il semble adéquat de stratifier les employés en fonction de leur ancienneté. Mais pour tirer parti du STR, nous devons disposer de strates dans lesquelles la variance est relativement faible. Nous devons donc supposer que les employés de différents secteurs (par exemple, ceux de la recherche et les employés de bureau réguliers) qui ont la même ancienneté ont des salaires comparables. Ensuite, nous devons sélectionner les strates : étant donné que les employés les plus récents ont "plus de marge" pour évoluer (promotions, augmentation de salaire, etc.) et que les employés plus âgés pourraient atteindre un "plafond salarial", on pourrait s'y prendre avec les strates suivantes:

- moins d'un an d'ancienneté;
- entre 1 et 3 ans d'ancienneté;
- entre 3 et 7 ans d'ancienneté;
- entre 7 et 12 ans d'ancienneté;
- plus de 12 ans d'ancienneté,

avec une allocation proportionnelle, puisque le coût et les variances sont égaux entre les strates. On peut aussi utiliser des strates et des allocations différentes (en justifiant les choix, bien entendu). ■

21. Dans l'utilisation de l'estimateur STR $\overline{y}_{\text{STR}}$ en tant qu'estimateur de $\overline{Y}$, il peut s'avérer avantageux de trouver une répartition et une taille d'échantillon qui minimise la variance $V(\overline{y}_{\text{STR}})$, pour un coût fixe C. Autrement dit, le coût C autorisé pour l'enquête est fixe, et nous cherchons la meilleure répartition des ressources qui permet de maximiser l'information au sujet de $\overline{Y}$. la répartition optimale dans ce cas demeure toujours

$$n_i = n\left(\frac{N_i\sigma_i/\sqrt{c_i}}{\sum N_j\sigma_j/\sqrt{c_j}}\right).$$

(a) Montrer que le meilleur choix pour n est

$$n = \frac{(C-c_0)\sum N_k\sigma_k/\sqrt{c_k}}{\sum N_k\sigma_k\sqrt{c_k}},$$

où c_0 représente les frais généraux fixes du sondage. (Noter que $C = c_0 + \sum c_k n_k$.)

(b) Si $V(\overline{y}_{\text{STR}}) = V$ est fixe, montrez que le choix approprié de n est

$$n = \frac{\left(\sum \frac{N_k\sigma_k/\sqrt{c_k}}{N}\right)\left(\sum \frac{N_k\sigma_k\sqrt{c_k}}{N}\right)}{V + \sum \frac{N_k\sigma_k^2}{N^2}}.$$

(c) Une entreprise souhaite obtenir des renseignements sur l'efficacité d'une machine commerciale qu'elle produit. Elle demande aux répondants d'évaluer l'équipement sur une échelle numérique. Le coût par entretien et les variances approximatives des évaluations et du nombre d'éléments pour trois strates sont donnés par (c_1 = \$9, $\sigma_1^2 = 2.25$, $N_1 = 112$), (c_2 = \$25, $\sigma_2^2 = 3.24$, $N_2 = 68$) et (c_3 = \$36, $\sigma_3^2 = 3.24$, $N_3 = 39$). L'entreprise veut donner un estimé de la note moyenne tout en respectant la condition $V(\overline{y}_{\text{STR}}) = 0.1$. Déterminer la taille d'échantillon n qui permet d'atteindre cette borne, et trouver la répartition appropriée.

Solution:

(a) On le constate par la simple manipulation suivante :

$$C - c_0 = \sum n_k c_k = \sum \frac{n(N_k\sigma_k/\sqrt{c_k})}{M} c_k = \frac{n}{M}\sum N_k\sigma_k\sqrt{c_k},$$

où

$$M = \sum \frac{N_k\sigma_k}{\sqrt{c_k}},$$

d'où

$$n = \frac{M(C-c_0)}{\sum n_k\sigma_k\sqrt{c_k}} = \frac{(C-c_0)\sum N_k\sigma_k/\sqrt{c_k}}{\sum N_k\sigma_k\sqrt{c_k}}.$$

(b) Soit a_i tel que $n_i = a_i n$. Puisque

$$V(\overline{y}_{\text{st}}) = \frac{1}{N^2}\sum N_k^2\left(\frac{N_k - a_k n}{N_k}\right)\left(\frac{\sigma_k^2}{a_k n}\right) = V,$$

on obtient (après quelques simplifications)

$$V = \frac{1}{N^2}\sum\left(N_k - \frac{n}{M}(N_k\sigma_k/\sqrt{c_k})\right)\left(\frac{M}{n}\sigma_k\sqrt{c_k}\right)$$

où M est comme à la partie (a). Ainsi ,

$$VN^2 = \frac{1}{n}\sum N_k M\sigma_k\sqrt{c_k} - \sum N_k\sigma_k^2$$

de sorte que

$$\begin{aligned} n &= \frac{\sum N_k M\sigma_k\sqrt{c_k}}{VN^2+\sum N_k\sigma_k^2} = \frac{M\sum N_k\sigma_k\sqrt{c_k}}{VN^2+\sum N_k\sigma_k^2} = \frac{\left(\sum \frac{N_k\sigma_k}{\sqrt{c_k}}\right)\left(\sum N_k\sigma_k\sqrt{c_k}\right)}{VN^2+\sum N_k\sigma_k^2} \\ &= \frac{1/N^2}{1/N^2}\cdot\frac{\left(\sum \frac{N_k\sigma_k}{\sqrt{c_k}}\right)\left(\sum N_k\sigma_k\sqrt{c_k}\right)}{VN^2+\sum N_k\sigma_k^2} = \frac{\left(\sum \frac{N_k\sigma_k/\sqrt{c_k}}{N}\right)\left(\sum \frac{N_k\sigma_k\sqrt{c_k}}{N}\right)}{V+\sum \frac{N_k\sigma_k^2}{N^2}}. \end{aligned}$$

(c) En utilisant les données disponibles, on obtient

$$\begin{aligned} n &= \frac{\left(\frac{N_1\sigma_1}{N\sqrt{c_1}}+\frac{N_2\sigma_2}{N\sqrt{c_2}}+\frac{N_3\sigma_3}{N\sqrt{c_3}}\right)\left(\frac{N_1\sigma_1\sqrt{c_1}}{N}+\frac{N_2\sigma_2\sqrt{c_2}}{N}+\frac{N_3\sigma_3\sqrt{c_3}}{N}\right)}{V+\left(\frac{N_1\sigma_1^2}{N^2}+\frac{N_2\sigma_2^2}{N^2}+\frac{N_3\sigma_3^2}{N^2}\right)} \\ &= \frac{\left(\frac{112\sigma_1}{219\sqrt{9}}+\frac{68\sigma_2}{219\sqrt{25}}+\frac{39\sigma_3}{219\sqrt{36}}\right)\left(\frac{112\sigma_1\sqrt{9}}{219}+\frac{68\sigma_2\sqrt{25}}{219}+\frac{39\sigma_3\sqrt{36}}{219}\right)}{0.1+\left(\frac{112\sigma_1^2}{219^2}+\frac{68\sigma_2^2}{219^2}+\frac{39\sigma_3^2}{219^2}\right)} = 26.266. \end{aligned}$$

En calculant $n_k = \frac{n}{M}N_k\sigma_k$, où $M = 92.18$, on obtient l'allocation $(16, 7, 3)$. ■

22. Pour donner un estimé du nombre total, τ, de sièges du parti social-démocrate dans tous les conseils municipaux d'un pays, la population a été stratifiée en quatre strates en utilisant le nombre total de sièges dans chaque conseil. On retrouve des renseignements sur ces strates dans le tableau suivant.

# sièges	N_i	$\sum_k Y_{k,i}$ (pop)	$\sum_k Y_{k,i}^2$ (pop)	$\sum_k y_{k,i}$ (éch)	$\sum_k y_{k,i}^2$ (éch)
$31-40$	44	756	13784	89	1647
$41-50$	168	3383	72223	441	9735
$51-70$	56	1545	44529	250	8294
71+	16	617	24137	102	5294

(a) Distribuer un échantillon total de taille $n = 40$ dans les 4 strates en utilisant la répartition proportionnelle.

(b) Donner un estimé du total des sièges socio-démocratiques à l'aide d'un échantillon STR de taille $n = 40$ selon cette répartition. Construire un intervalle de confiance pour le total à environ 95%.

(c) Donner un estimé du nombre total de sièges socio-démocratiques si un EAS avait été utilisé à la place afin de sélectionner un échantillon de taille $n = 40$. Construire un intervalle de confiance pour le total à environ 95%.

(d) Laquelle des deux méthodes utilisées en (b) et (c) est la plus efficace? Pourquoi?

Solution:

(a) Sous l'allocation proportionnelle, $n_1 = n\frac{N_i}{N}$. Ici, $n = 40$ et $N = 284$; avec les données, la répartition appropriée est $(n_1, n_2, n_3, n_4) = (6, 24, 8, 2)$.

(b) Dans ce cas, le total τ est approché à l'aide de

$$\hat{\tau} = \sum_{i=1}^{4} N_i \overline{y}_{\mathrm{st,i}} = \sum_{i=1}^{4} \frac{N_i}{n_i} \sum_k y_{k,i} = 6305,67$$

et la variance de τ par

$$\hat{V}(\tau) = \sum_{i=1}^{4} N_i^2 \left(\frac{N_i - n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right),$$

où l'écart type de l'échantillon dans chaque strate est

$$s_i^2 = \frac{1}{n_i - 1}\left(\sum_k y_{i,k}^2 - \frac{1}{n_i}\left(\sum_k y_{k,i}\right)^2\right).$$

En utilisant les données fournies dans le tableau, on obtient

$$(s_1^2, s_2^2, s_3^2, s_4^2) = (65.37, 70.94, 68.70, 92.00),$$

de sorte que $\hat{V}(\tau) = 123141.4..$ Ainsi,

$$\hat{\tau} \pm 2\sqrt{\hat{V}(\overline{y}_{\mathrm{st}})} = 6305.67 \pm 701.83.$$

est un I.C. à environ 95

(c) Dans ce cas, le total τ est estimé par

$$N\overline{y} = \frac{284}{40}\sum_{i=1}^{40} y_i = \frac{284}{40}\left(\sum_{i=1}^{4}\sum_{k} y_{k,i}\right) = \frac{284 \cdot 882}{40} = 6262.2$$

et la variance de τ est approchée par

$$\hat{V}(\tau) = N^2\left(\frac{N-n}{N}\right)\left(\frac{s^2}{n}\right),$$

où $s^2 = \frac{1}{39}(24970 - 40 \cdot 22.05^2) = 141.59$. Ainsi,

$$\hat{V}(\tau) = 284^2\left(\frac{244}{284}\right)\left(\frac{141.59}{40}\right) = 245290.52$$

et on obtient un I.C. à environ 95% pour τ à l'aide de

$$N\overline{y}_{\text{st}} \pm 2\sqrt{\hat{V}(\overline{y}_{\text{st}})} = 6262.2 \pm 990.54$$

.

(d) L'estimateur calculé avec un échantillonnage stratifié est plus efficace que l'estimateur calculé avec un échantillonnage aléatoire simple car sa variance est la plus petite des deux; ce n'est pas surprenant puisque les strates sont bien différentes les unes des autres. ■

23. Les salariés d'une grande entreprise sont stratifiés en deux classes: les cadres et les employé.e.s de bureau, la première de taille $N_1 = 121$ et la seconde de taille $N_2 = 589$. On cherche à évaluer l'attitude à l'égard de la politique de congé de maladie en prélevant des échantillons aléatoires indépendants de taille $N_1 = n_2 = 35$ dans chacune des classes. On sépare de plus les réponses selon le genre des répondants. Dans le tableau des résultats, a =nombre d'individus qui aiment la politique; b =nombre d'individus qui n'aiment pas la politique, et c =nombre d'individus qui n'ont pas d'opinion.

	Cadres $N_1 = 121$	**Bureau** $N_2 = 589$	**Total** $N = 710$
Hommes	$a = 3$ $b = 15$ $c = 3$	$a = 10$ $b = 2$ $c = 1$	34
Femmes	$a = 6$ $b = 6$ $c = 2$	$a = 15$ $b = 7$ $c = 0$	36
Total	$n_1 = 35$	$n_2 = 35$	$n = 70$

Donner un estimé et une variance approximative de cet estimé pour les paramètres suivants:

(a) Proportion des cadres en faveur de cette politique.

(b) Proportion des employé.e.s en faveur de cette politique.

(c) Nombre total d'employées qui ne supportent pas la politique.

(d) Différence entre la proportion de cadres masculins et la proportion de cadres féminins en faveur de la politique.

(e) Différence entre la proportion des cadres en faveur de la politique et les cadres qui ne supportent pas la politique.

Solution:

(a) Nous observons que $\hat{p} = \frac{3+6}{35} \approx 0.26$, et que

$$\hat{V}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n_1 - 1}\left(\frac{N_1 - n_1}{N_1}\right) = \frac{0.26(0.74)}{34}\left(\frac{86}{121}\right) = 0.004.$$

(b) Nous observons que $\hat{p}_{\text{st}} = \frac{1}{N}[N_1 p_1 + N_2 p_2] = \frac{1}{710}[121(0.26) + 589(0.71)] = 0.637$, et que

$$\begin{aligned}\hat{V}(\hat{p}_{\text{st}}) &= \frac{1}{N_2}\sum_{i=1}^{2} N_i^2 \left(\frac{N_i - n_i}{N_i}\right)\frac{p_i(1-p_i)}{n_i - 1} \\ &= \frac{1}{(710)^2}\left[(121)^2\left(\frac{86}{121}\right)\frac{0.26(0.74)}{34} + (589)^2\left(\frac{554}{589}\right)\frac{0.71(0.29)}{34}\right] = 0.004.\end{aligned}$$

(c) Nous définissons une nouvelle variables sur les cadres selon

$$z_{1,i} = \begin{cases} 1 & \text{le/la } i\text{ème cadre n'aime pas la politique} \\ 0 & \text{autrement} \end{cases}$$

et sur les employé.e.s de bureau selon

$$z_{2,i} = \begin{cases} 1 & \text{le/la } i\text{ème employé.e de bureau n'aime pas la politique} \\ 0 & \text{autrement} \end{cases}$$

Nous obtenons alors:

Cadre	Bureau
$N_1 = 121$	$N_2 = 589$
$n_1 = 35$	$n_2 = 35$
$\sum z_{1,k} = 21$	$\sum y_{2,k} = 9$
$\sum z_{1,k}^2 = 21$	$\sum y_{2,k}^2 = 9$

Avec ces données, nous obtenons $n = n_1 + n_2 = 70$, $s_1^2 = \frac{34}{35}(21/35(1 - 21/35)) = 0.233$, $s_2^2 = \frac{34}{35}(9/35(1 - 9/35)) = 0.186$, $A_1 = \frac{N_1}{N} = \frac{121}{710}$ et $A_2 = \frac{N_2}{N} = \frac{589}{710}$, d'où

$$\hat{\tau} = N\overline{z} = N\sum_i A_i \overline{y}_i = 710\left[\frac{121}{710}(21/35) + \frac{580}{710}(9/35)\right] = 221.7,$$

et la variance approximative est

$$\begin{aligned}\hat{V}(\hat{\tau}) = N^2\hat{V}(\overline{z}) &= \left[N_1^2\frac{s_1^2}{n_1}\left(1 - \frac{n_1}{N_1}\right) + N_2^2\frac{s_2^2}{n_2}\left(1 - \frac{n_2}{N_2}\right)\right] \\ &= \left[(121)^2\frac{0.233}{35}\left(1 - \frac{35}{121}\right) + (589)^2\frac{0.186}{35}\left(1 - \frac{35}{589}\right)\right] = 1803.358.\end{aligned}$$

(d) Selon un communiqué de presse de l'Organisation internationale du travail émis en 1997, les femmes représenteraient 42% des cadres canadiens en 1997 (ils ne donnent pas l'intervalle de confiance, donc nous ne devrions probablement pas leur faire confiance à ce point, mais en tout cas). Nous allons donc supposer que $0.42(121) \approx 51$ des cadres de l'entreprise sont des femmes et donc que $121 - 51 = 70$ des cadres sont des hommes (vos chiffres pourraient être différents), d'où

$$\hat{p}_{\mathrm{F}} - \hat{p}_{\mathrm{H}} = \frac{6}{14} - \frac{3}{21} = \frac{2}{7} = 0.286$$

et

$$V(\hat{p}_{\mathrm{H}}) = \frac{(3/21)(1 - 3/21)}{21}\left(\frac{70 - 21}{69}\right) = 0.0041,\ V(\hat{p}_{\mathrm{F}}) = \frac{(6/14)(1 - 6/14)}{14}\left(\frac{51 - 14}{50}\right) = 0.0127,$$

d'où

$$V(\hat{p}_{\mathrm{F}} - \hat{p}_{\mathrm{H}}) = V(\hat{p}_{\mathrm{F}}) + V(\hat{p}_{\mathrm{H}}) = 0.0129 + 0.0041 = 0.0171,$$

puisque les proportions sont indépendantes.

(e) Ces quantités sont corrélées (négativement), mais nous supposons ici qu'elles ne le sont pas. La différence entre les proportions est ainsi

$$\hat{p}_1 - \hat{p}_2 = \frac{15 + 6}{35} - \frac{6 + 3}{35} = 0.343,$$

et

$$V(\hat{p}_1 - \hat{p}_2) = V(\hat{p_1}) + V(\hat{p_2}) = \left(\frac{121 - 35}{120(35)}\right)((21/35)(1 - 21/35) + (9/35)(1 - 9/35)) = 0.0088,$$

ce qui complète l'exercice. ■

24. (a) Prélever un échantillon aléatoire de 20 tailles d'hommes à partir d'une distribution binomiale (la taille correspond au nombre de succès de n expériences de Bernouilli indépendentes avec chance de succès p) avec paramètres $n = 142$ et $p = 0.5$, et un échantillon aléatoire distinct de 20 tailles de femmes à partir d'une distribution binomiale avec paramètres $n = 130$ et $p = 0.5$. À partir de ces données, donner un estimé de la taille moyenne des adultes et calculez la marge d'erreur sur l'estimation.

    (b) Prélever un EAS de 40 tailles d'adultes à partir d'une distribution binomiale avec paramètres $n = 135$ et $p = 0.5$. À partir de ces données, donner un estimé de la taille moyenne de tous les adultes et donner une marge d'erreur sur l'estimation.

    (c) Comparer les résultats de (a) et (b). Discuter des cas où la stratification semble produire des gains en précision des estimations.

Solution: On suppose que $N_{\mathrm{H}} \approx N_{\mathrm{F}} \approx 0.5N$ et que

$$\frac{n_{\mathrm{H}}}{N_{\mathrm{H}}} \approx \frac{n_{\mathrm{F}}}{N_{\mathrm{F}}} \approx \frac{n}{N} \approx 0.$$

(a) On se sert du programme suivant afin d'obtenir la moyenne et la variance de chacuns des échantillons:

```
> set.seed(0) # replicabilite
> n = 20
> x.h = rbinom(n,142,0.5) # ech. hommes
> x.f = rbinom(n,130,0.5) # ech. femmes
> (x.h.moy = mean(x.h))   # moy. ech. hommes
[1] 69.85
> (x.h.s2 = var(x.h))     # var. ech. hommes
[1] 29.18684
> (x.f.moy = mean(x.f))   # moy. ech. femmess
[1] 66.3
> (x.f.s2 = var(x.f))     # var. ech. femmes
[1] 26.95789
> (x.st = 0.5*x.h.moy + 0.5*x.f.moy) # moy. str.
[1] 68.07
> (V=(0.5)^2/n*(x.h.s2+x.f.s2)) # var. str.
[1] 0.7018092
> (B=2*sqrt(V))
[1] 1.675481
```

(b) On s'y prend de la même manière:

```
> set.seed(5) # replicabilite
> n = 40
> x = rbinom(n,135,0.5)
> (x.moy = mean(x))
[1] 68.375
> (x.s2 = var(x))
[1] 39.88141
> (V=x.s2/n)
[1] 0.9970353
> (B=2*sqrt(V))
[1] 1.997033
```

(c) Dans ces exemples, la marge d'erreur pour l'échantillon STR est légèrement inférieure à celle de l'échantillon EAS. Est-ce un accident? Le code suivant répète la procédure à 400 reprises pour voir ce qui en découle.

```
> set.seed(6) # replicabilite
> M = 400
> str.meilleure = c()
> for(j in 1:M){
    # STR
    n = 20
    x.h = rbinom(n,142,0.5)
    x.f = rbinom(n,130,0.5)
    x.h.moy = mean(x.h)
    x.f.moy = mean(x.f)
    x.h.s2 = var(x.h)
    x.f.s2 = var(x.f)
    V=(0.5)^2/n*(x.h.s2+x.f.s2)
    B=2*sqrt(V)

    # EAS
    n = 40
    x = rbinom(n,135,0.5)
    x.moy = mean(x)
    x.s2 = var(x)
    V=x.s2/n
    B1=2*sqrt(V)

    # comparaison entre STR et SRS
    str.meilleure[j] = B<B1
> }
> summary(str.meilleure)
   Mode   FALSE    TRUE
logical     210     190
```

L'approche STR semble plus ou moins équivalente à l'EAS avec les paramètres du problème.

En général, lorsque la distribution est multimodale, ce qui n'est pas le cas en (a), la stratification centrée sur les différents modes produit de meilleures estimations, car la variabilité est réduite. Ou encore, si la distribution de la population a une "grande" variance et est une "somme" de distributions distinctes avec de "petites" variances, la stratification est préférable. ■

25. Une école souhaite donner un estimé du score moyen de ses élèves de sixième année à un examen de compréhension de l'écrit. Les élèves de l'école sont regroupés en trois filières: les élèves plus rapides étant regroupés dans la filière I et les élèves plus lents dans la filière III. L'école décide de stratifier par rapport aux filières. La sixième année compte 50 élèves dans la voie I, 90 dans la voie II et 60 dans la voie III. Un échantillon stratifié de 50 élèves est réparti proportionnellement dans les filières (on obtient $n_{\mathrm{I}} = 14$, $n_{\mathrm{II}} = 20$ et $n_{\mathrm{III}} = 16$, respectivement). Les résultats de l'échantillon sont présentés ci-dessous:

Filière i	$\overline{y}_i$	s_i^2
I	79.71	105.14
II	64.75	158.20
III	37.44	186.13

(a) En considérant l'enquête ci-dessus comme une étude pilote, trouver la taille de l'échantillon nécessaire pour donner un estimé du score moyen avec une marge d'erreur sur l'estimation $B = 4$. Utiliser la répartition proportionnelle.

(b) Répéter la partie (a) en utilisant la répartition de Neyman. Comparer les résultats.

Solution: Nous avons $N_I = 50$, $N_{II} = 90$, $N_{III} = 60$ et $N = 200$. On utilise $\sigma_i^2 \approx s_i^2$.

(a) Dans un scénario de répartition proportionnelle, la taille de l'échantillon est

$$n = \frac{N_I\sigma_I^2 + N_{II}\sigma_{II}^2 + N_{III}\sigma_{III}^2}{NB^2/4 + \frac{1}{200}(N_I\sigma_I^2 + N_{II}\sigma_{II}^2 + N_{III}\sigma_{III}^2)} = \frac{50\sigma_I^2 + 90\sigma_{II}^2 + 60\sigma_{III}^2}{200(4^2/4) + \frac{1}{200}(50\sigma_I^2 + 90\sigma_{II}^2 + 60\sigma_{III}^2)}$$
$$= \frac{50(105.14) + 90(158.20) + 60(186.13)}{200(4) + \frac{1}{200}(50(105.14) + 90(158.20) + 60(186.13))} = 32.16443 \approx 33.$$

La répartition proportionnelle serait alors

$$(n_I, n_{II}, n_{III}) = \frac{n}{N}(N_I, N_{II}, N_{III}) = (8.04, 14.47, 9.65) \approx (8, 14, 10);$$

mais cela ne nous donne que 32 unités, alors qu'il n'en faut au moins 33. Le meilleur candidat est donc $(8, 15, 10)$.

(b) Dans un scénario de répartition de Neyman, la taille de l'échantillon est

$$n = \frac{(N_I\sigma_I + N_{II}\sigma_{II} + N_{III}\sigma_{III})^2}{N^2B^2/4 + (N_I\sigma_I^2 + N_{II}\sigma_{II}^2 + N_{III}\sigma_{III}^2)} = \frac{(50\sigma_I + 90\sigma_{II} + 60\sigma_{III})^2}{200^2(4^2/4) + (50\sigma_I^2 + 90\sigma_{II}^2 + 60\sigma_{III}^2)}$$
$$= \frac{(50\sqrt{105.14} + 90\sqrt{158.20} + 60\sqrt{186.13})^2}{200^2(4) + (50(105.14) + 90(158.20) + 60(186.13))} = 31.82409 \approx 32.$$

Pour ceux que cela intéresserait, la répartition de Neyman serait alors

$$(n_I, n_{II}, n_{III}) = \frac{n}{N_I\sigma_I + N_{II}\sigma_{II} + N_{III}\sigma_{III}}(N_I\sigma_I, N_{II}\sigma_{II}, N_{III}\sigma_{III}) \approx (6.62, 14.62, 10.57).$$

Si on arrondi, on obtient $(7, 15, 11)$, ce qui donne 33 unités. À toutes fins pratiques, il n'y a pas vraiment de différence entre la répartition de Neyman et la répartition proportionnelle étant données les variances de strates et la marge d'erreur recherchée, si ce n'est qu'avec l'allocation de Neyman, on donne un peu plus d'importance à la strate qui a une variance plus élevée (compatible avec la définition de cette répartition). ■

26. Une entreprise souhaite obtenir des renseignements sur l'efficacité d'une imprimante commerciale. Un certain nombre de chefs de division seront interrogés par téléphone et il leur sera demandé d'évaluer l'équipement sur une échelle numérique. Les divisions sont situées en Amérique du Nord, en Europe et en Asie. Par conséquent, un échantillonnage stratifié est utilisé. Les coûts sont plus élevés pour les entretiens avec les chefs de division situés en dehors de l'Amérique du Nord. Le tableau suivant indique les coûts par entretien, les variances approximatives des évaluations et la taille des strates.

Strate	N_i	σ_i^2	c_i
Amérique du Nord	127	2.31	\$9
Europe	58	3.33	\$25
Asie	79	3.21	\$36

(a) L'entreprise souhaite donner un estimé de la cote moyenne en préservant $V(\overline{y}_{\text{STR}}) = 0.1$. Obtenir la taille d'échantillon stratifié n qui permet d'atteindre cette marge et trouvez la répartition appropriée.

(b) Un budget de 800\$ est disponible, duquel 125\$ doivent être réservés pour les frais généraux fixes. Déterminer la taille de l'échantillon et la taille optimale des échantillons dans chaque strate.

(c) Répéter (a) et (b) en utilisant un logiciel.

Solution:

(a) Selon la répartition optimale, nous avons

$$n = \frac{\left(\sum_{i=1}^{3} N_i\sigma_i\sqrt{c_i}\right)\left(\sum_{i=1}^{3} \frac{N_i\sigma_i}{\sqrt{c_i}}\right)}{N^2V(\overline{y}_{\text{st}}) + \sum_{i=1}^{3} N_i\sigma_i^2}$$

$$= \frac{\left(127\sqrt{2.31}\sqrt{9} + 58\sqrt{3.33}\sqrt{25} + 79\sqrt{3.21}\sqrt{36}\right)\left(\frac{127\sqrt{2.31}}{\sqrt{9}} + \frac{58\sqrt{3.33}}{\sqrt{25}} + \frac{79\sqrt{3.21}}{\sqrt{36}}\right)}{(264)^2(0.1) + 127(2.31) + 58(3.33) + 79(3.21)} = 27.7;$$

nous devons ainsi choisir (au moins) $n = 28$ chefs de division.
Les poids d'échantillonnage sont donnés par

$$w_i = \frac{n_i}{n} = \frac{\frac{N_i\sigma_i}{\sqrt{c_i}}}{\sum_{i=1}^{3} \frac{N_i\sigma_i}{\sqrt{c_i}}} = \frac{1}{142.548} \cdot \frac{N_i\sigma_i}{\sqrt{c_i}}, \quad i = 1, 2, 3.$$

Ainsi,

$$w_1 = \frac{1}{109.1} \cdot \frac{127\sqrt{2.31}}{\sqrt{9}} = 0.590, \; w_2 = \frac{1}{109.1} \cdot \frac{58\sqrt{3.33}}{\sqrt{25}} = 0.194, \; w_3 = \frac{1}{109.1} \cdot \frac{79\sqrt{3.21}}{\sqrt{36}} = 0.216.$$

En gros, nous devrions échantillonner $0.589n = 16.49$ chefs de division en Amérique du Nord, $0.1940n = 5.43$ chefs de division en Europe et $0.216n = 5.98$ chefs de division en Asie (cela ne donne que 27... on en rajoute un en Europe, mettons, puisque la variance dans cette strate est plus élevée): $(n_1, n_2, n_3) = (16, 6, 6)$. On pourrait aussi utiliser $(17, 5, 6)$.

(b) Nous avons $C = 800$ and $c_0 = 125$. La taille totale de l'échantillon minimisant $V(\overline{y}_{\text{st}})$ est donnée par

$$n = (C - c_0)\frac{\sum_{i=1}^{3} \frac{N_i\sigma_i}{\sqrt{c_i}}}{\sum_{i=1}^{3} N_i\sigma_i\sqrt{c_i}} = \frac{(800 - 125)\left(\frac{127\sqrt{2.31}}{\sqrt{9}} + \frac{58\sqrt{3.33}}{\sqrt{25}} + \frac{79\sqrt{3.21}}{\sqrt{36}}\right)}{127\sqrt{2.31}\sqrt{9} + 58\sqrt{3.33}\sqrt{25} + 79\sqrt{3.21}\sqrt{36}} = 37.62;$$

de sorte que nous choisissons 38 chefs de division dans ce schéma. Les poids d'échantillonnage sont les mêmes que ceux en (a), d'où $(n_1, n_2, n_3) \approx (22, 8, 8)$.

(c) Pour la troisième question, on laisse tomber. ■

27. Un gouvernement municipal souhaite agrandir les installations d'une garderie pour enfants à besoins spéciaux. Cette extension augmentera le coût d'inscription d'un enfant dans la garderie. Un sondage sera mené afin de donner un estimé de la proportion de familles ayant des enfants à mobilité réduite qui utiliseront les nouvelles installations. Les familles sont divisées entre celles qui utilisent les installations existantes et celles qui ne les utilisent pas. Certaines familles vivent dans la municipalité, d'autres dans les banlieues et les zones rurales environnantes. On utilise donc un plan d'échantillonnage STR avec les strates suivantes: (1) utilisateurs actuels provenant de la municipalité, (2) utilisateurs actuels provenant des régions environnantes, (3) non-utilisateurs actuels provenant de la municipalité, et (4) non-utilisateurs actuels provenant des régions environnantes. Le coût d'obtention d'une observation pour un utilisateur actuel est de \$4; il est de \$8 pour un non-utilisateur actuel. Selon les dossiers de la municipalité, les populations sont $N_1 = 97$, $N_2 = 43$, $N_3 = 45$ et $N_4 = 68$.

(a) Déterminer la taille de l'échantillon et la répartition requise afin de donner un estimé de la proportion de la population avec une marge d'erreur sur l'estimation de $B = 0.05$.

(b) Supposons que l'enquête soit menée et qu'elle donne les proportions suivantes: $\hat{p}_1 = 0.87$, $\hat{p}_2 = 0.93$, $\hat{p}_3 = 0.60$ et $\hat{p}_4 = 0.53$. Estimez la proportion dans la population et placer une borne sur l'erreur d'estimation. La limite souhaitée en (a) a-t-elle été atteinte?

(c) Supposons qu'un budget de 475\$ soit disponible, mais que 75\$ doivent être réservés pour les frais généraux fixes. Déterminer la taille de l'échantillon STR et la taille optimale de l'échantillon dans chaque strate en utilisant les informations de l'énoncé du problème comme valeurs plausibles pour les proportions des strates (et non celles de la partie (b)).

Solution: On résume la situation comme suit :

Strate i	N_i	c_i	$\hat{p}_{\mathrm{i}}$
1	97	4	0.87
2	43	4	0.93
3	45	8	0.60
4	68	8	0.53

(a) Nous ne connaissons pas les proportions exactes p_{i}, alors nous utilisons $p_{\mathrm{i}} = 0.5$ afin de déterminer les poids d'échantillonnage pour la répartition optimale:

$$w_i = \frac{n_i}{n} \approx \frac{\dfrac{N_i(0.5)}{\sqrt{c_i}}}{\displaystyle\sum_{i=1}^{4} \frac{N_i(0.5)}{\sqrt{c_i}}} = 0.0091 \cdot \frac{N_i}{\sqrt{c_i}}, \quad i = 1, 2, 3, 4.$$

Nous obtenons alors

$$w_1 = 0.009 \cdot \frac{97}{2} \approx 0.44,\ w_2 = 0.009 \cdot \frac{43}{2} \approx 0.20,\ w_3 = 0.009 \cdot \frac{45}{\sqrt{8}} \approx 0.14.\ w_4 = 0.009 \cdot \frac{68}{\sqrt{8}} \approx 0.22,$$

d'où

$$n = \frac{\left(\displaystyle\sum_{i=1}^{4} N_i(0.5)\sqrt{c_i}\right)\left(\displaystyle\sum_{i=1}^{4} \frac{N_i(0.5)}{\sqrt{c_i}}\right)}{N^2 D + \displaystyle\sum_{i=1}^{4} N_i(0.25)}, \quad \text{avec} \quad D = \frac{B^2}{4} = 0.000625.$$

Ainsi, $n = 159.622 \approx 160$. La répartition $n_i = w_i n$, $i = 1, 2, 3, 4$ devient:

$$(n_1, n_2, n_3, n_4) = (70.577, 31.287, 23.152, 34.895) \approx (71, 31, 23, 35).$$

Le coût associé au sondage est donc $4(71 + 31) + 8(23 + 35) = 872\$$ (sans compter les frais généraux).

(b) L'estimateur de la proportion prend la valeur

$$\hat{p}_{\mathrm{st}} = \frac{1}{N} \sum_{i=1}^{4} N_{\mathrm{i}} \hat{p}_{\mathrm{i}} = \frac{1}{253} \left(97(0.87) + 43(0.93) + 45(0.60) + 68(0.53)\right) = 0.741.$$

La marge d'erreur sur l'estimation est ainsi

$$B = 2\sqrt{\hat{V}\left(\hat{p}_{\mathrm{st}}\right)} = \frac{2}{N} \sqrt{\sum_{i=1}^{4} N_{\mathrm{i}}^2 \frac{\hat{p}_{\mathrm{i}} \hat{q}_{\mathrm{i}}}{n_{\mathrm{i}} - 1} \left(1 - \frac{n_{\mathrm{i}}}{N_{\mathrm{i}}}\right)} = 0.045;$$

la marge souhaitée a été atteinte.

(c) Puisque nous ne connaissons pas les proportions exactes p_{i}, nous utilisons $p_{\mathrm{i}} = 0.5$ afin de déterminer la taille de l'échantillon :

$$n = (C - c_0) \frac{\displaystyle\sum_{i=1}^{4} \frac{N_i(0.5)}{\sqrt{c_i}}}{\displaystyle\sum_{i=1}^{4} N_i(0.5)\sqrt{c_i}} = (C - c_0) \frac{\displaystyle\sum_{i=1}^{4} \frac{N_i}{\sqrt{c_i}}}{\displaystyle\sum_{i=1}^{4} N_i \sqrt{c_i}} = 400 \cdot \frac{109.951}{599.612} \approx 73.$$

Les poids d'échantillonnage de la répartition optimale sont toujours valides: on utilise ainsi $n_i = w_i(73)$, $i = 1, 2, 3, 4$:

$$(n_1, n_2, n_3, n_4) \approx (32, 14, 11, 16),$$

et le coût total du sondage est alors $75 + 4(32 + 14) + 8(11 + 16) = 475$. ■

28. Un forestier souhaite donner un estimé du nombre total d'acres agricoles plantés d'arbres dans sa province. Comme la superficie des arbres varie considérablement en fonction de la taille de l'exploitation en question, il décide de procéder à une stratification en fonction de la taille des exploitations. Les 263 fermes de la province sont placées dans l'une des quatre catégories en fonction de leur taille. Un échantillon aléatoire stratifié de 40 exploitations, sélectionné en utilisant la répartition proportionnelle, donne les résultats indiqués dans le tableau ci-dessous.

Strate	N_i	n_i	$\overline{y}_\text{i}$	s_i
<200 acres	96	14	63.36	32.74
200 à <400 acres	82	12	183.0	95.2
400 à <600 acres	55	9	340.6	129.6
600+ acres	30	5	472.0	269.0

(a) Donner un estimé de la superficie totale (en acres) d'arbres dans les exploitations de la province, et donner une marge d'erreur sur l'estimation.

(b) Supposons que l'on souhaite obtenir une marge d'erreur sur l'estimation de 5000 acres. En considérant ce qui précède comme une enquête préliminaire, trouver la taille de l'échantillon nécessaire pour atteindre cette borne si on utilise la répartition de Neyman.

Solution:

(a) L'estimateur du total prend la valeur

$$\tau_\text{st.} = N\overline{y}_\text{st.} = \sum_{i=1}^{4} N_\text{i}\overline{y}_\text{i} = 96(63.36) + 82(183.0) + 55(340.6) + 30(472.0) = 53981.56,$$

et la marge d'erreur sur l'estimation approche

$$\begin{aligned} B &= 2\sqrt{\hat{V}\left(\tau_\text{st.}\right)} = 2\sqrt{\sum_{i=1}^{4} N_i^2 \cdot \frac{s_\text{i}^2}{n_\text{i}}\left(1-\frac{n_\text{i}}{N_\text{i}}\right)} \\ &= 2\sqrt{96^2 \cdot \frac{32.74^2}{14}\left(1-\frac{14}{96}\right) + 82^2 \cdot \frac{95.2^2}{12}\left(1-\frac{12}{82}\right) + 55^2 \cdot \frac{129.6^2}{9}\left(1-\frac{9}{55}\right) + 30^2 \cdot \frac{269^2}{5}\left(1-\frac{5}{30}\right)} \\ &= 9058.391. \end{aligned}$$

(b) Avec la répartion de Neyman, la taille d'échantillon minimale devrait être

$$n = \frac{\left(\sum_{i=1}^{4} N_i\sigma_i\right)^2}{\frac{B^2}{4} + \sum_{i=1}^{4} N_i\sigma_i^2} \approx \frac{\left(\sum_{i=1}^{4} N_i s_i\right)^2}{\frac{B^2}{4} + \sum_{i=1}^{4} N_i s_i^2} = 67.08 \approx 68$$

La répartition de Neyman donne

$$n_i = w_i n \approx \left(\frac{N_i s_i}{\sum N_j s_j}\right) \cdot 68 \approx (8, 20, 19, 21),$$

ce qui donne bien une taille de $n = 68$. ■

29. On cherche à donner un estimé de la distance quotidienne moyenne parcourue durant la saison hivernale 2012 en Ontario par certains types de véhicules. La consommation de carburant quotidienne est aussi d'intérêt, tout comme la proportion des véhicules qui ne sont pas utilisés. Un échantillon STR est prélevé à même la flotte Ontarienne (de taille $N = 7,868,359$) contenant des information au sujet du type de véhicule, de l'âge du véhicule, et de la région; les données relatives aux répondants sont recueillies dans le fichier `Autos_STR.xlsx`. Discuter des enjeux pouvant venir influencer la qualité des données. Donner un sommaire numérique et visuel des données de l'échantillon réalisé. Donner un intervalle de confiance pour chaque moyenne de population recherchée, à environ 95%, avec coefficient de variation correspondant. Faire de même dans chaque strate (et chaque combinaison). [La majorité des notes seront attribuées pour la discussion et la présentation des résultats.]

Solution: On ne présente que les calculs. ■

Chapitre 4 – Estimation par le quotient, par la régression, et par la différence

30. La caractéristique de la population à laquelle on s'intéresse dans une enquête est $\alpha = 1/\mu$, où μ est la moyenne de la population. Dans un EAS de taille $n = 105$, on obtient $\overline{y} = 5.25$ et $s = 0.37$. Dans ce qui suit, nous considérons $\hat{\alpha} = \overline{y}^{-1}$ comme estimateur de α.

    (a) Utiliser un développement en série de Taylor (de deuxième ordre) de $\hat{\alpha}$ autour de $\overline{y} = \mu$ afin d'obtenir une expression approximative du biais de $\hat{\alpha}$ en tant qu'estimateur de α.

    (b) Utiliser un développement en série de Taylor (du premier ordre) de $\hat{\alpha}$ autour de $\overline{y} = \mu$ afin d'obtenir une expression approximative du biais de $\hat{\alpha}$ en tant qu'estimateur de α.

    (c) En supposant que la distribution de $\hat{\alpha}$ suit approximativement une loi normale pour des valeurs de n suffisamment élevées, utiliser le résultat de (b) afin d'obtenir un I.C. de α à environ 95%. [Ignorer le biais de $\hat{\alpha}$ et le facteur de correction de la population finie, en supposant dans ce dernier cas que N est très grand.]

    (d) Trouver un I.C. de α à environ 95% en trouvant d'abord un intervalle analogue pour μ, puis en inversant les bornes. Comparer avec le résultat obtenu en (c).

    Solution: Soient un EAS $\{y_i\}$ de taille n provenant d'une population $\{u_j\}$ de taille N, et l'estimateur

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Dans le contexte EAS, nous savons que $E(\overline{y}) = \mu$ et $V(\overline{y}) = \frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)$, où σ^2 represente la variance de la population.

(a) La série de Taylor de deuxième ordre de $f(x) = \frac{1}{x}$ autour de $x = \mu$ est

$$\frac{1}{x} \approx \frac{1}{\mu} - \frac{1}{\mu^2}(x-\mu) + \frac{1}{\mu^3}(x-\mu)^2.$$

Ainsi, $\overline{y}^{-1} \approx \frac{3}{\mu} - \frac{3\overline{y}}{\mu^2} + \frac{\overline{y}^2}{\mu^3}$ et

$$\begin{aligned}\text{Biais}\,(\hat{\alpha}) = E(\hat{\alpha}-\alpha) = E(\hat{\alpha}) - \alpha = E\left(\overline{y}^{-1}\right) - \frac{1}{\mu} &\approx E\left(\frac{3}{\mu} - \frac{3\overline{y}}{\mu^2} + \frac{\overline{y}^2}{\mu^3}\right) - \frac{1}{\mu} \\ &= \frac{3}{\mu} - \frac{3}{\mu^2}E(\overline{y}) + \frac{1}{\mu^3}E\left(\overline{y}^2\right) - \frac{1}{\mu} \\ &= \frac{3}{\mu} - \frac{3}{\mu^2}\mu + \frac{1}{\mu^3}\left[V(\overline{y}) + (E(\overline{y}))^2\right] - \frac{1}{\mu} = \frac{1}{\mu^3}\left[V(\overline{y}) + \mu^2\right] - \frac{1}{\mu} = \frac{1}{\mu^3}V(\overline{y})\end{aligned}$$

L'approximation de deuxième ordre du Biais($\hat{\alpha}$) dans un EAS est alors $\frac{1}{\mu^3}\cdot\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$.

(b) L'approximation de premier ordre de $f(x) = \frac{1}{x}$ autour de $x = \mu$ est

$$\frac{1}{x} \approx \frac{1}{\mu} - \frac{1}{\mu^2}(x-\mu).$$

Ainsi, $\overline{y}^{-1} \approx \frac{2}{\mu} - \frac{\overline{y}}{\mu^2}$ et

$$\begin{aligned}\text{Biais}\,(\hat{\alpha}) = E(\hat{\alpha}-\alpha) = E(\hat{\alpha}) - \alpha = E\left(\overline{y}^{-1}\right) - \frac{1}{\mu} &\approx E\left(\frac{2}{\mu} - \frac{\overline{y}}{\mu^2}\right) - \frac{1}{\mu} \\ &= \frac{2}{\mu} - \frac{1}{\mu^2}E(\overline{y}) - \frac{1}{\mu} = \frac{2}{\mu} - \frac{1}{\mu} - \frac{1}{\mu} = 0\end{aligned}$$

L'approximation de premier ordre du Biais($\hat{\alpha}$) dans un EAS est donc nulle.

(c) Si $\hat{\alpha}$ suit approximativement une loie normale lorsque la taille n est élevée, et si Biais$(\hat{\alpha}) \approx 0$, alors $\hat{\alpha} \pm 2\sqrt{\hat{V}(\hat{\alpha})}$ représente un I.C. de α à environ 95%. On se sert de l'expansion de premier ordre de la partie (b) et l'on obtient

$$V(\hat{\alpha}) = V\left(\overline{y}^{-1}\right) \approx V\left(\frac{2}{\mu} - \frac{\overline{y}}{\mu^2}\right) = V\left(\frac{\overline{y}}{\mu^2}\right) = \frac{1}{\mu^4} V(\overline{y}) = \frac{1}{\mu^4} \cdot \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).$$

On constate alors, en ignorant le facteur de correction en population finie, que

$$\hat{V}(\hat{\alpha}) \approx \overline{y}^{-4} \cdot \frac{s^2}{n}\left(1 - \frac{n}{N}\right) \approx \frac{\overline{y}^{-4} s^2}{n}.$$

Conséquemment,

$$\hat{\alpha} \pm 2\sqrt{\hat{V}(\hat{\alpha})} \equiv \overline{y}^{-1} \pm 2 \cdot \frac{\overline{y}^{-2} s}{\sqrt{n}} \equiv \frac{1}{5.25} \pm 2 \cdot \frac{0.37}{(5.25)^2\sqrt{105}} \equiv (0.1878561, 0.1930963)$$

est l'I.C. recherché.

(d) Dans un EAS, si l'on suppose que le FCPF est ≈ 0, on obtient un I.C. de μ à environ 95% à l'aide de

$$\overline{y} \pm 2\sqrt{\hat{V}\left(\overline{y}\right)} \approx \overline{y} \pm 2 \cdot \frac{s}{\sqrt{n}} \equiv 5.25 \pm 2 \cdot \frac{0.37}{\sqrt{105}} \equiv (5.177783, 5.322217).$$

En inversant les bornes, on obtient un autre I.C. de α à environ 95%:

$$\left(\frac{1}{5.322217}, \frac{1}{5.177783}\right) = (0.1878916, 0.1931329).$$

■

31. Notre ami forestier souhaite maintenant donner un estimé du volume total des arbres d'une vente de bois ($N = 250$). Il prélève un EAS (de taille $n = 12$) de ces arbres et enregistre le volume de chaque arbre dans l'échantillon. En outre, il mesure la superficie de la base de chaque arbre marqué pour la vente. Il utilise ensuite un estimateur par le quotient pour le volume total. Soit X la superficie de la base et Y le volume en pieds cubes d'un arbre. Le total de la superficie de la base des 250 arbres est de $\tau_X = 75$ pieds carrés. Il recueille les données suivantes:

Arbre	**Superficie de la base**	**Volume**	**Arbre**	**Superficie de la base**	**Volume**
1	0.3	6	7	0.6	12
2	0.5	9	8	0.5	9
3	0.4	7	9	0.8	20
4	0.9	19	10	0.4	9
5	0.7	15	11	0.8	18
6	0.2	5	12	0.6	13

(a) Obtenir les moyennes et les écarts types de l'échantillon pour la superficie de la base et pour le volume, ainsi qu'une estimé de la corrélation entre les deux variables.

(b) En utilisant les résultats de (a), donner un estimé du volume total des arbres marqués pour la vente en utilisant l'estimation par le quotient, et une marge d'erreur sur l'estimation.

Solution: Soit X la superficie de la base, et Y le volume. On utilise le code `R` suivant:

```
> x=1/10*c(3,5,4,9,7,2,6,5,8,4,8,6)
> y=c(6,9,7,19,15,5,12,9,20,9,18,13)

> mean(x)
> sd(x)

> mean(y)
> sd(y)

> cor(x,y)
```

(a) Ainsi, $\mu_X = 0.55833$, $s_X = 0.21515$, $\overline{y} = 11.83333$, $s_Y = 5.18448$, et $\hat{\rho} = \frac{s_{XY}}{s_X s_Y} = 0.97123$.

(b) L'estimateur par le quotient du total du volume est

$$\hat{\tau}_Y = r\tau_X = \frac{\overline{y}}{\mu_X}\tau_X = \frac{11.83333}{0.55833}(75) = (21.19)(75) = 1589.561281,$$

tandis que la marge d'erreur sur l'estimation est

$$\begin{aligned} B &= 2\sqrt{\hat{V}\left(\hat{\tau}_Y\right)} = 2\sqrt{\hat{V}\left(r\tau_X\right)} = 2\tau_X\sqrt{\hat{V}\left(r\right)} \approx 2\tau_X\sqrt{\frac{s_W^2}{n\mu_X^2}\left(1-\frac{n}{N}\right)} = 2N\sqrt{\frac{s_W^2}{n}\left(1-\frac{n}{N}\right)} \\ &= 2N\sqrt{\frac{s_Y^2 + r^2 s_X^2 - 2r\hat{\rho}s_X s_Y}{n}\left(1-\frac{n}{N}\right)} \\ &= 2(250)\sqrt{\frac{(5.2)^2 + (21.2)^2(0.2)^2 - 2(21.2)(0.97)(0.2)(5.2)}{12}\left(1-\frac{12}{250}\right)} \\ &\approx 186.321 \end{aligned}$$

■

32. On souhaite donner un estimé de la moyenne μ_Y d'une population donnée. Un EAS contient les observations y_i et l'information auxiliaire x_i, $i = 1, \ldots, n$, (la moyenne μ_X de la population est connue). Discuter des mérites relatifs de l'utilisation de:

(a) La moyenne de l'échantillon $\overline{y}$.

(b) L'estimateur par le quotient $\hat{\mu}_{Y;R}$.

(c) L'estimateur par la régression $\hat{\mu}_{Y;L}$.

(d) L'estimateur par la différence $\hat{\mu}_{Y;D}$.

Solution:

(a) La moyenne de l'échantillon $\overline{y}$ représente l'estimateur EAS.

Bénéfices: Facile à calculer. Pas besoin d'informations auxiliaires x. S'il y a des informations auxiliaires x mais que la corrélation avec y est faible, l'estimateur EAS fournit un estimateur plus efficace.

Inconvénients: S'il y a une forte corrélation entre x et y, l'estimateur EAS perd en précision en n'utilisant pas les informations auxiliaires.

(b) L'estimateur par le quotient $\hat{\mu}_{Y;R} = \frac{\overline{y}}{\mu_X}\mu_X$ présume l'existence d'une relation linéaire forte entre x et y, et que la droite de régression passe par l'origine.

Bénéfices: Si les hypothèses se réalisent, l'estimateur du ratio est plus efficace que l'estimateur EAS.

Inconvénients: Si ces hypothèses ne sont pas vérifiées, l'estimateur du ratio peut être inefficace.

(c) L'estimateur par la régression $\hat{\mu}_{Y;L} = a + b\mu_X$ présume l'existence d'une rlation linéaire forte entre x et y, mais la droite de régression ne passe pas nécessairement pas l'origine.

Bénéfices: L'estimateur par la régression est plus efficace que l'estimateur par le quotient, sauf si $b = \frac{\overline{y}}{\mu_X}$, auquel cas ils sont équivalents.

Inconvénients: Si $a \approx 0$, l'estimateur par le quotient devrait être utilisé car il est plus facile à mettre en œuvre. L'estimateur par la régression peut présenter un biais important si la corrélation entre x et y est faible.

(d) L'estimateur par la différence est $\hat{\mu}_{Y;D} = \mu_X + d = \mu_X + (\overline{y} - \mu_X)$.

Bénéfices: Plus facile à calculer que l'estimateur par la régression.

Inconvénients: Peut être moins efficace que l'estimateur par la régression lorsque la corrélation entre x et y n'est pas forte. ■

33. Une société souhaite donner un estimé du revenu total des ventes d'un produit durant une période de trois mois. Pour chacun des $N = 123$ bureaux de district, le total des revenus est disponible durant la période de trois mois correspondante de l'année précédente: $\tau_X = 128,200$. Un EAS de 13 bureaux de district est prélevé parmi les 123 bureaux de la société. Les données résultantes sont présentées dans le tableau ci-dessous.

Bureau i	1	2	3	4	5	6	7	8	9	10	11	12	13
Précédent x_i	550	720	1500	1020	620	980	928	1200	1350	1750	670	729	1530
Actuel y_i	610	780	1600	1030	600	1050	977	1440	1570	2210	980	865	2020

(a) Tracer un graphique de dispersion de y_i en fonction de x_i et appliquer un modèle linéaire simple. Quel estimateur le modèle suggère-t-il? Expliquer.

(b) Utiliser un estimateur par le quotient afin de donner un estimé de la moyenne des revenus actuels μ_Y (par bureau) et donner une marge d'erreur sur l'estimation.

(c) Utiliser un estimateur par le quotient afin de donner un estimé du total τ_Y des revenus actuels (société) et donner une marge d'erreur sur l'estimation.

Solution:

(a) Les sommes et les données qui seront nécessaires pour calculer les paramètres et les coefficients sont les suivantes :

$$\sum_{i=1}^{13} x_i = 13547 \qquad \sum_{i=1}^{13} y_i = 15732 \qquad \sum_{i=1}^{13} x_i y_i = 18748141$$
$$\sum_{i=1}^{13} x_i^2 = 15963525 \qquad \sum_{i=1}^{13} y_i^2 = 22230054.$$

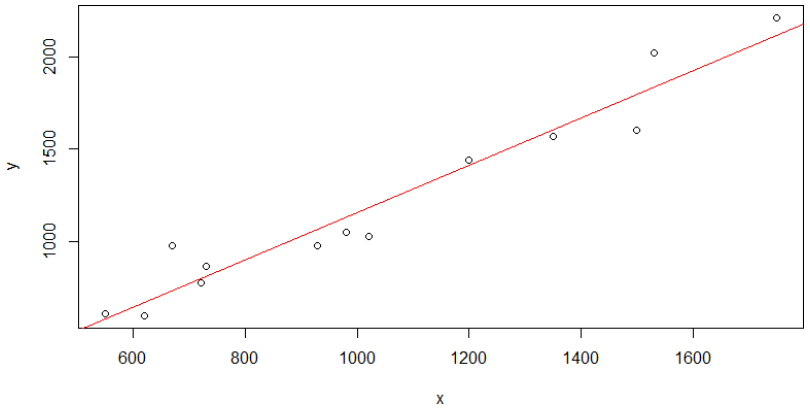
Les moyennes sont alors $\overline{x} = 1042.077$ et $\overline{y} = 1210.154$. Selon les formules, les paramètres de régression de la droite sont les suivants :

$$b = \frac{S_{xy}}{S_{xx}} = \frac{\sum x_i y_i - 13\overline{xy}}{\sum x_i^2 - 13\overline{x}^2} = \frac{18748141 - 13(1042.077)(1210.154)}{15963525 - 13(1042.0769)^2} = 1.2749$$
$$a = \overline{y} - b\overline{x} = 1210.154 - 1.2749(1042.0769) = -118.390,$$

et la droite de régression est tout simplement $\hat{y} = -118.390 + 1.275x$. Le coefficient de corrélation entre y et x est de

$$\rho = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = 0.9697.$$

Le nuage de points et la droite des moindres carrés sont présentés ci-dessous.



L'utilisation de l'estimateur par le quotient est acceptable lorsque $1 \geq \rho \gg 0$ et $a = 0$. Ici, $1 \geq \rho = 0.9697 \gg 0$. À première vue, il semblerait que $a = -118.390 \neq 0$. On peut tester

$$\begin{cases} H_0 : a = 0 \\ H_a : a \neq 0 \end{cases}$$

à l'aide d'un test t bilatéral. D'une part, note that

$$s\{a\} = \sqrt{\text{EQM}}\sqrt{\frac{1}{13} + \frac{\overline{x}}{S_{xx}}} = 107.30798.$$

Selon la théorie, la statistique $t^* = \frac{a}{s\{a\}}$ suit une loi $t(n-2)$ lorsque H_0 est valide. Pour $\alpha = 0.05$, la valeur critique $t(1-\alpha/2; 11) = t(0.975; 11) = 2.200985$ est plus élevée que $|t^*| = \frac{118.390}{107.30798} = 1.10$. Conséquemment, il n'y a pas assez d'évidence pour rejeter H_0 et on considère que l'estimateur par le quotient est acceptable.

(b) L'estimateur de μ_Y par le quotient est

$$\hat{\mu}_{Y;R} = \frac{\overline{y}}{\overline{x}}\mu_X = \frac{1210.154}{1042.0769} \cdot \frac{128200}{123} = 1210.386.$$

L'approximation de la variance de l'estimateur par le quotient est :

$$\begin{aligned} \hat{V}\left(\hat{\mu}_{Y;R}\right) &= \frac{N-n}{Nn}s_r^2 = \frac{N-n}{Nn}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \frac{\overline{y}}{\overline{x}}x_i\right)^2\right] \\ &= \frac{123-13}{123(13)} \cdot \frac{1}{12}214317.9834 = 1228.6313. \end{aligned}$$

On peut alors construire un intervalle de confiance pour μ_Y à environ 95%:

$$1210.386 \pm 2\sqrt{1228.6313} = 1210.386 \pm 70.1.$$

(c) L'estimateur de τ_Y par le quotient est

$$\hat{\tau}_{Y;R} = \frac{\overline{y}}{\overline{x}}\tau_X = \frac{1210.154}{1042.0769} \cdot 128200 \approx 148877.44.$$

L'approximation de la variance de l'estimateur par le quotient est :

$$\hat{V}\left(\hat{\tau}_{Y;R}\right) = N^2\hat{V}\left(\hat{\mu}_{Y;R}\right) = 123^2 \cdot 1228.6313 \approx 18587962.9377.$$

On peut alors construire un intervalle de confiance pour τ_Y à environ 95%:

$$148877.44 \pm 2\sqrt{18587962.9377} = 148877.44 \pm 8622.7520.$$

■

34. Une gestionnaire de ressources forestières souhaite donner un estimé du nombre de sapins morts dans une zone de 400 acres. À l'aide d'une photo aérienne, elle divise la zone en 200 parcelles de 2 acres. Soit x le compte des sapins morts sur la photo et y le compte réel au sol pour un EAS de $n = 10$ parcelles. Le nombre total de sapins morts obtenu à partir du compte photographique est $X = 4300$. Les données résultantes sont présentées dans le tableau ci-dessous.

Parcelle i	1	2	3	4	5	6	7	8	9	10
Compte photo x_i	12	30	24	24	18	30	12	6	36	42
Compte réel y_i	18	42	24	36	24	36	14	10	48	54

(a) Tracer un graphique de dispersion de y_i en fonction de x_i et appliquer un modèle linéaire simple. Quel estimateur le modèle suggère-t-il? Expliquer.

(b) Utiliser un estimateur par le quotient afin de donner un estimé du nombre total τ_Y de sapins mort dans la zone de 400 acres et donner une marge d'erreur sur l'estimation.

(c) Utiliser un estimateur par la régression afin de donner un estimé du nombre total τ_Y de sapins mort dans la zone de 400 acres et donner une marge d'erreur sur l'estimation.

(d) Utiliser un estimateur par la différence afin de donner un estimé du nombre total τ_Y de sapins mort dans la zone de 400 acres et donner une marge d'erreur sur l'estimation.

(e) Quel estimateur est préférable pour ce problème? Expliquer.

Solution:

(a) On utilise le code suivant:

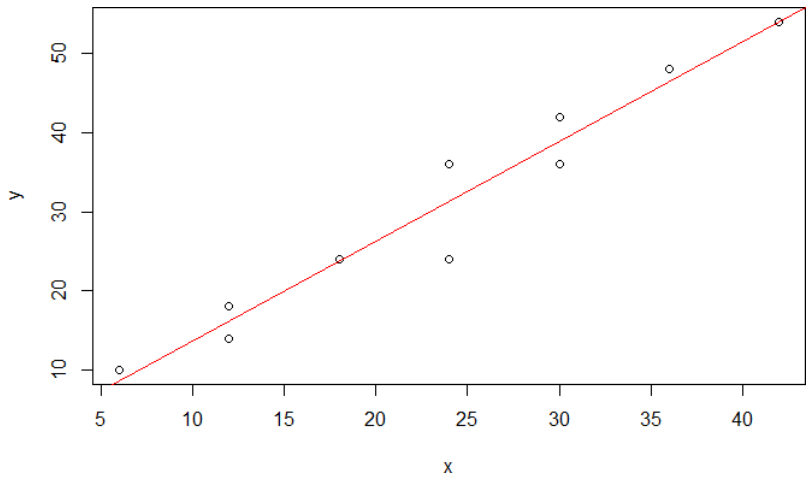
```
> x=c(12,30,24,24,18,30,12,6,36,42)
> y=c(18,42,24,36,24,36,14,10,48,54)
> plot(x,y)
> abline(lm(y ~ x), col = "red")
> summary(lm(y~x))

Residuals:
    Min      1Q  Median      3Q     Max
-7.3556 -1.6884  0.7568  1.7006  4.6444

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.1307     2.7286   0.414    0.689
x             1.2594     0.1057  11.911 2.27e-06 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.639 on 8 degrees of freedom
Multiple R-squared:  0.9466,Adjusted R-squared:   0.94
F-statistic: 141.9 on 1 and 8 DF,  p-value: 2.269e-06
```

Le nuage de points et la droite de régression se retrouvent plus bas:



La droite de régression est $\hat{y} = 1.1307 + 1.2594x$ et la correlation de x et y dans l'échantillon est

```
> cor(x,y)
[1] 0.9729
```

L'estimateur par le quotient est approprié lorsque $1 \geq \rho \gg 0$ et $a = 0$.
Ici, nous avons $1 \geq \rho = 0.9729 \gg 0$ et $a = 1.1307$ et l'erreur-type de l'ordonnée à l'origine est

$$s\{a\} = 2.7286;$$

la statistique observée $t^* = \frac{a}{s\{a\}}$ suit une distribution $t(n-2) = t(8)$ si la valeur réelle de a est nulle. Si $\alpha = 0.05$, la valeur critique $t(1-\alpha/2; 8) = t(0.975; 8) = 2.306$ est plus élevée que la valeur observée $|t^*| = 1.1307/2.7286 = 0.4144$. Nous n'avons donc pas assez d'évidence afin de rejeter l'hypothèse $a = 0$. Conséquemment, l'estimateur par le quotient est préférable.

(b) Nous calculons aisément que $\overline{x} = 23.4$ et $\overline{y} = 30.6$. L'estimateur par le quotient de τ_Y est

$$\hat{\tau}_{Y;R} = \frac{\overline{y}}{\overline{x}}\tau_X = \frac{30.6}{23.4}4300 = 5623.077,$$

et sa variance approximative est

$$\hat{V}\left(\hat{\tau}_{Y;R}\right) = \hat{V}\left(N\hat{\mu}_{Y;R}\right) = N^2\left(\frac{N-n}{Nn}\right)s_r^2 = N^2\left(\frac{N-n}{Nn}\right)\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \frac{\overline{y}}{\overline{x}}x_i\right)^2\right]$$
$$= (200)^2\frac{200-10}{200(10)}\cdot\frac{1}{9}(12.08) \approx 45890.$$

On forme alors un intervalle de confiance à environ 95% de τ_Y à l'aide de $5623.077 \pm 2\sqrt{45890} \approx 5623.01 \pm 428.44$.

(c) L'estimateur par le régression est à propos lorsqu'il y a une forte corrélation linéaire entre x et y: puisque le coefficient de corrélation est $\rho = 0.9727$, c'est une hypothèse adéquate.

L'estimateur du total τ_Y par la régression est

$$\hat{\tau}_{Y;L} = N(a + b\mu_X) = Na + b\tau_X = 200(1.1307) + 1.2594(4300) = 5641.56,$$

et sa variance approximative est

$$\hat{V}\left(\hat{\tau}_{Y;L}\right) = \hat{V}\left(N\hat{\mu}_{Y;L}\right) = N^2\left(\frac{N-n}{Nn}\right)\text{EQM} = N^2\left(\frac{N-n}{Nn}\right)\left[\frac{1}{n-2}\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2 - b\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2\right]$$
$$= (200)^2\frac{200-10}{200(10)}\cdot\frac{1}{8}\left(13.24\right) \approx 50312.$$

On forme alors un intervalle de confiance à environ 95% de τ_Y à l'aide de $5641.56 \pm 2\sqrt{50312} \approx 5641.56 \pm 448.61$.

(d) Pas de solution.

(e) Pas de solution.

35. Un contrôle traditionnel exprime les ventes au détail comme étant l'inventaire d'ouverture plus les achats du magasin, duquel on retranche l'inventaire de fermeture, sur une période de 6 semaines afin de rapporter les ventes totales. De telles données, provenant de plusieurs magasins et recueillies pour une variété de marques concurrentes, permettent de donner un estimé dus parts de marché. Mais les méthodes de vérification des ventes de la fin de semaine et des achats en magasin offrent des méthodes plus rapides pour donner un estimé dus parts de marché. La première élimine les achats en magasin, car les achats sont minimes le fin de semaine, mais utilise une période plus courte et est sujette à des irrégularités dues aux promotions de fin de semaine. La seconde utilise uniquement l'information sur les achats pour calculer la part de marché et n'implique aucune vérification des stocks. Pour une certaine marque de bière, les données sur les parts de marché calculées par les trois méthodes [traditionnelle (T), fin de semaine (F), achats (A)] sont présentées dans le tableau ci-dessous [les observations ont été effectuées à six périodes différentes au cours de l'année].

Traditionnelle (T)	**Fin de semaine** (F)	**Achats** (A)
15	16	12
18	17	14
16	17	20
14	16	11
13	12	8
16	18	15

(a) Présenter un estimé du quotient de la part de marché moyenne calculée avec la méthode F par celle calculée avec la méthode T, et donner une marge d'erreur sur l'estimation.

(b) Présenter un estimé du quotient de la part de marché moyenne calculée avec la méthode A par celle calculée avec la méthode T, et donner une marge d'erreur sur l'estimation.

(c) Quelle méthode se compare le plus favorablement à la méthode traditionnelle?

(d) Y a-t-il des obstacles qui se manifestent dans les divers diagrammes de dispersion?

Solution:

(a) Les sommes et les données qui seront nécessaires pour calculer les paramètres et les coefficients sont les suivantes :

$$\sum_{i=1}^{6} t_i = 92 \qquad \sum_{i=1}^{6} f_i = 96 \qquad \sum_{i=1}^{6} t_i f_i = 1486$$
$$\sum_{i=1}^{6} t_i^2 = 1426 \qquad \sum_{i=1}^{6} f_i^2 = 1558$$

Les moyennes sont $\overline{t} = 15.3333$, $\overline{f} = 16$. Les paramètres de régression sont ainsi :

$$b_F = \frac{1486 - 6(15.3333)(16)}{1426 - 6(15.3333)^2} = 0.9130$$
$$a_F = 16 - 0.9130(15.3333) = 2$$

et la droite de régression est $\hat{f} = 2 + 0.9130t$. De plus, $s\{a_F\} = 5.9764$ et $\rho_{T,F} = 0.7623$. Alors $r_F = \frac{\overline{f}}{\overline{t}} = 1.043$ et

$$\hat{V}(r_F) = \left(\frac{N-n}{Nn}\right) \frac{1}{\overline{T}^2} s_{r_F}^2.$$

Puisque $\overline{T}$ et N sont inconnus, nous utilisons $\bar{t} \approx \overline{T}$ et $\frac{N-n}{N} \approx 1$. Alors

$$\hat{V}(r_F) = \frac{1}{n\overline{T}^2} s^2_{r_F} = \frac{1}{5(15.3333)^2} 1.895 = 0.00134,$$

et on obtient un I.C. de R_F à environ 95% à l'aide de $1.043 \pm 2\sqrt{0.00134} = 1.043 \pm 0.0733$.

(b) Les sommes et les données qui seront nécessaires pour calculer les paramètres et les coefficients sont les suivantes :

$$\sum_{i=1}^{6} t_i = 92 \qquad \sum_{i=1}^{6} a_i = 80 \qquad \sum_{i=1}^{6} t_i a_i = 1250$$
$$\sum_{i=1}^{6} t_i^2 = 1426 \qquad \sum_{i=1}^{6} a_i^2 = 1150$$

Les moyennes sont $\bar{t} = 15.3333$, $\bar{a} = 13.3333$. Les paramètres de régression sont ainsi :

$$b_A = \frac{1250 - 6(15.3333)(13.3333)}{1426 - 6(15.3333)^2} = 1.5217$$
$$a_A = 13.3333 - 1.5217(15.3333) = -10$$

et la droite de régression est $\hat{a} = -10 + 1.5217t$. De plus, $s\{a_A\} = 13.6135$ et $\rho_{T,A} = 0.6528$. Alors $r_A = \frac{\bar{a}}{\bar{t}} = 0.8696$ et

$$\hat{V}(r_A) = \left(\frac{N-n}{Nn}\right) \frac{1}{\overline{T}^2} s^2_{r_A}.$$

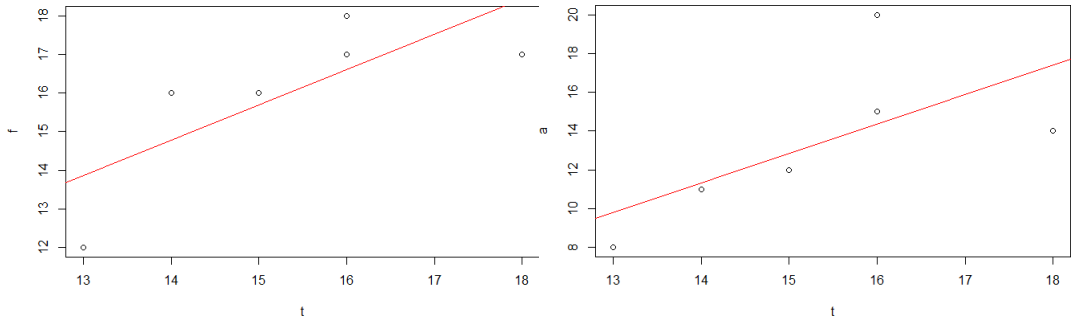
Puisque $\overline{T}$ et N sont inconnus, nous utilisons $\bar{t} \approx \overline{T}$ et $\frac{N-n}{N} \approx 1$. Alors

$$\hat{V}(r_A) = \frac{1}{n\overline{T}^2} s^2_{r_A} = \frac{1}{5(15.3333)^2} 10.868 = 0.0077,$$

et on obtient un I.C. de R_A à environ 95% à l'aide de $0.868 \pm 2\sqrt{0.0077} = 0.868 \pm 0.1755$.

(c) La méthode F est favorable à la méthode A par le quotient puisque $\hat{V}(r_F) < \hat{V}(r_A)$, et r_F est plus près de 1 que r_A ne l'est.

(d) Les méthodes par le quotient utilisées en (a) et (b) nécessitent $a_A, a_P \approx 0$ et $1 \geq \rho_{T,F}, \rho_{T,P} \gg 0$. On peut montrer à l'aide de tests t qu'il n'est pas impossible que $a_A, a_F \approx 0$ et les coefficients de corrélation $1 \geq \rho_{T,F}, \rho_{T,A} \gg 0$ sont relativement élevés comme on peut le voir dans les nuages. ■



36. Une population est composée de $N = 5$ unités dont les valeurs de X et Y sont les suivantes:

$(X_1, Y_1) = (3, 2), \quad (X_2, Y_2) = (5, 3), \quad (X_3, Y_3) = (3, 3), \quad (X_4, Y_4) = (4, 2), \quad (X_5, Y_5) = (6, 5).$

(a) Déterminer le quotient R dans cette population.

(b) Pour chaque échantillon possible de taille $n = 3$, déterminer le quotient r. Calculer ensuite le biais d'échantillonnage de r, à savoir $E[r - R]$.

(c) Nous avons développé, en classe, l'approximation théorique de l'erreur systématique:

$$E[r - R] \approx \frac{1}{n\mu_X^2} \left(\frac{N-n}{N-1} \right) (R\sigma_X^2 - \rho\sigma_X\sigma_Y).$$

Calculer la valeur de l'approximation théorique de l'erreur systématique pour cette population, et comparer avec la valeur réelle.

(d) Calculer les deux estimations de la moyenne de la population, $\overline{y}_{\text{EAS}}$ et $\hat{\mu}_{Y;R}$, pour chaque échantillon. À partir de ces résultats, calculer $V(\overline{y}_{\text{EAS}})$ et $E\left[(\hat{\mu}_{Y;R} - \mu_Y)^2\right]$. Discutez des avantages et des inconvénients de l'utilisation respective de y_{EAS} et de $\hat{\mu}_{Y;R}$ en tant qu'estimateurs de μ_Y.

Solution:

(a) Le quotient est

$$R = \frac{\mu_Y}{\mu_X} = \frac{\sum Y_i}{\sum X_i} = \frac{2+3+3+2+5}{3+5+3+4+6} = \frac{15}{21} = \frac{5}{7}.$$

(b) Il y a $\binom{5}{3} = 10$ échantillons de taille $n = 3$:

$(x_1, y_1), (x_2, y_2), (x_3, y_3)$	**Échantillon**	r
$(X_1, Y_1), (X_2, Y_2), (X_3, Y_3)$	$(3,2), (5,3), (3,3)$	8/11
$(X_1, Y_1), (X_2, Y_2), (X_4, Y_4)$	$(3,2), (5,3), (4,2)$	7/12
$(X_1, Y_1), (X_2, Y_2), (X_5, Y_5)$	$(3,2), (5,3), (6,5)$	5/7
$(X_1, Y_1), (X_3, Y_3), (X_4, Y_4)$	$(3,2), (3,3), (4,2)$	7/10
$(X_1, Y_1), (X_3, Y_3), (X_5, Y_5)$	$(3,2), (3,3), (6,5)$	5/6
$(X_1, Y_1), (X_4, Y_4), (X_5, Y_5)$	$(3,2), (4,2), (6,5)$	9/13
$(X_2, Y_2), (X_3, Y_3), (X_4, Y_4)$	$(5,3), (3,3), (4,2)$	2/3
$(X_2, Y_2), (X_3, Y_3), (X_5, Y_5)$	$(5,3), (3,3), (6,5)$	11/14
$(X_2, Y_2), (X_4, Y_4), (X_5, Y_5)$	$(5,3), (4,2), (6,5)$	2/3
$(X_3, Y_3), (X_4, Y_4), (X_5, Y_5)$	$(3,3), (4,2), (6,5)$	10/13

Le biais d'échantillonnage est ainsi

$$E(r - R) = E(r) - R = \frac{1}{10}\left(\frac{8}{11} + \frac{7}{12} + \frac{5}{7} + \frac{7}{10} + \frac{5}{6} + \frac{9}{13} + \frac{2}{3} + \frac{11}{14} + \frac{2}{3} + \frac{10}{13}\right) - \frac{5}{7} = -0.0004045954.$$

(c) Les valeurs importantes sont :

$$\mu_X = \frac{1}{5}(3+5+3+4+6) = 4.2$$
$$\sigma_X^2 = \frac{1}{5}\left(3^2+5^2+3^2+4^2+6^2\right) - (4.2)^2 = 1.36$$
$$\mu_Y = \frac{1}{5}(2+3+3+2+5) = 3$$
$$\sigma_Y^2 = \frac{1}{5}\left(2^2+3^2+3^2+2^2+5^2\right) - 3^2 = 1.2$$
$$\text{Cov}(X,Y) = E(XY) - \mu_X\mu_Y = \frac{1}{5}((3\cdot 2)+(5\cdot 3)+(3\cdot 3)+(4\cdot 2)+(6\cdot 5)) - (4.2)(3) = 1.$$

L'approximation théorique du biais is

$$E(r-R) \approx \frac{1}{n\mu_X^2}\left(\frac{N-n}{N-1}\right)\left(R\sigma_X^2 - \mathrm{Cov}(X,Y)\right)$$
$$= \frac{1}{3(4.2)^2}\left(\frac{5-3}{5-1}\right)\left(\frac{5}{7}(1.36)-1\right) = -0.0002699492.$$

(d) Nous avons la table suivante :

$(x_1,y_1),(x_2,y_2),(x_3,y_3)$	**Échantillon**	r	$\overline{y}$	$\hat{\mu}_{Y;R} = r\mu_X$
$(X_1,Y_1),(X_2,Y_2),(X_3,Y_3)$	$(3,2),(5,3),(3,3)$	8/11	8/3	3.054545
$(X_1,Y_1),(X_2,Y_2),(X_4,Y_4)$	$(3,2),(5,3),(4,2)$	7/12	7/3	2.45
$(X_1,Y_1),(X_2,Y_2),(X_5,Y_5)$	$(3,2),(5,3),(6,5)$	5/7	10/3	3
$(X_1,Y_1),(X_3,Y_3),(X_4,Y_4)$	$(3,2),(3,3),(4,2)$	7/10	7/3	2.94
$(X_1,Y_1),(X_3,Y_3),(X_5,Y_5)$	$(3,2),(3,3),(6,5)$	5/6	10/3	3.5
$(X_1,Y_1),(X_4,Y_4),(X_5,Y_5)$	$(3,2),(4,2),(6,5)$	9/13	3	2.907692
$(X_2,Y_2),(X_3,Y_3),(X_4,Y_4)$	$(5,3),(3,3),(4,2)$	2/3	8/3	2.8
$(X_2,Y_2),(X_3,Y_3),(X_5,Y_5)$	$(5,3),(3,3),(6,5)$	11/14	11/3	3.3
$(X_2,Y_2),(X_4,Y_4),(X_5,Y_5)$	$(5,3),(4,2),(6,5)$	2/3	10/3	2.8
$(X_3,Y_3),(X_4,Y_4),(X_5,Y_5)$	$(3,3),(4,2),(6,5)$	10/13	10/3	3.230769

Puisque $\overline{y}$ est un estimateur sans biais de μ_Y, nous avons ainsi

$$\mathrm{EQM}(\overline{y}) = V(\overline{y}) = \frac{1}{10}\sum_{\overline{y}} \overline{y}^2 - \mu_Y^2 = \frac{1}{10}((8/3)^2 + \cdots + (10/3)^2) - 3^2 = 9.2 - 9 = 0.2$$

$$\mathrm{EQM}(\hat{\mu}_{Y;R}) = \frac{1}{10}\left[(3.0... - 3)^2 + \cdots + (3.2... - 3)^2\right] + (E(\hat{\mu}_{Y;R} - \mu_Y))^2 = 0.079 + (-0.017)^2 = 0.079.$$

Ceci suggère que l'estimateur par le quotient, quoique biaisé, est plus précis et sans doute préférable dans ce cas. En général, $\hat{\mu}_Y$ est un bon choix d'estimateur pour μ_Y si :

i. la relation entre Y et X est linéaire et passe par l'origine, et si

ii. si la variance de Y le long de la droite de régression est proportionnelle à la valeur prise par X.

Dans ce cas, cependant, il n'y a pas vraiment assez de points dans la population pour déterminer si les hypothèses sont réelement valides. ■

37. Les données relatives à la taille de la famille x_i et aux dépenses alimentaires y_i au cours de la semaine d'enquête sont enregistrées pour chaque famille d'un échantillon de 33 familles provenant d'une grande population de familles.

(a) Exprimer les dépenses alimentaires (pour cette semaine) par personne dans la population sous forme de quotient de populations.

(b) En utilisant les données de l'échantillon, nous obtenons

$$\sum_{i=1}^{33} x_i = 123, \quad \sum_{i=1}^{33} x_i^2 = 533, \quad \sum_{i=1}^{33} y_i = 2721.30, \quad \sum_{i=1}^{33} y_i^2 = 254196, \quad \sum_{i=1}^{33} x_i y_i = 10786.5$$

donner un estimé et un I.C. (à environ 95%) des dépenses alimentaires par capita dans la population.

Solution:

(a) Le quotient recherché est $R = \frac{\sum_{j=1}^{N} y_j}{\sum_{j=1}^{N} x_j}$.

(b) Avec les données du problème, l'estimé ponctuel pour le quotient est $r = \frac{\sum_{i=1}^{33} y_i}{\sum_{i=1}^{33} x_i} = \frac{2721.3}{123} = 22.12$. Si on néglige le facteur de correction en population finie, la variance de l'estimateur est environ

$$\hat{V}(r) = \frac{s_r^2}{33\overline{x}^2} = \frac{\sum_{i=1}^{33}(y_i - rx_i)^2}{32 \cdot 33 \cdot \overline{x}^2} = \frac{1}{123^2 \cdot 32/33}\left[\sum_{i=1}^{33} y_i^2 - 2r\sum_{i=1}^{33} x_i y_i + r^2 \sum_{i=1}^{33} x_i^2\right]$$
$$= \frac{254196 - 2(22.12)10786.5 + (22.12)^2 533}{123^2 \cdot 32/33} = 2.58.$$

La marge d'erreur sur l'estimateur r est ainsi $B_r = 2\sqrt{\hat{V}(r)} = 3.21$, d'où l'intervalle de confiance recherché est 22.12 ± 3.21. ■

38. Donner un estimé du volume total (en pieds cube) des arbres marqués pour la vente (cf. donnés de la question 31) en utilisant l'estimation par la régression et l'estimation par EAS, et placer une limite sur l'erreur d'estimation dans les deux cas.

Solution: On commence par l'EAS, et on termine avec l'estimateur de régression.

(a) Puisque $\overline{y} = \frac{142}{12} = 11.83$, l'estimateur du total dans le contexte EAS est

$$\tau = N\overline{y} = 250(11.83) = 2958.33.$$

La variance de l'estimateur est environ

$$\begin{aligned}\hat{V}(\hat{\tau}) &= N^2 \frac{s^2}{n}\left(1 - \frac{n}{N}\right) = 250^2 \frac{s^2}{12}\left(1 - \frac{12}{250}\right) = \frac{14875}{3}s^2 \\ &= \frac{14875}{3} \cdot \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right] = \frac{14875}{3} \cdot \frac{1}{11}\left[1976 - 12(11.83)^2\right] = 133700.6\end{aligned}$$

et la marge d'erreur sur l'estimation est $2\sqrt{\hat{V}(\tau)} = 731.302$.

(b) Nous obtenons :

i. $\sum_{i=1}^{12} X_i = 6.7000$

ii. $\sum_{i=1}^{12} Y_i = 142.0000$

iii. $\sum_{i=1}^{12} X_i Y_i = 91.2000$

iv. $\sum_{i=1}^{12} X_i^2 = 4.2500$

v. $\sum_{i=1}^{12} Y_i^2 = 1976.0000$

vi. $n = 12$.

Les moyennes sont ainsi $\overline{X} = 0.5583$ et $\overline{Y} = 11.8333$, et conséquemment, les coefficients de la régression sont

$$b_1 = \frac{91.2000 - 12(0.5583)(11.8333)}{4.2500 - 12(0.5583)^2} = 23.4043 \quad \text{et} \quad b_0 = 11.8333 - 23.4043(0.5583) = -1.2340$$

et la droite de régression est $\hat{Y} = -1.2340 + 23.4043X$. La corrélation entre X et Y est forte ($\rho = .9712$). L'estimateur ponctuel du total est donc

$$\hat{\tau}_{Y;L} = N\hat{\mu}_{Y;L} = 250(-1.2340 + 23.4043 \cdot \frac{75}{250}) = 1446.822,$$

et la variance d'échantillonnage est

$$\hat{V}(\hat{\tau}_{Y;L}) = N^2 \cdot \frac{s_Y^2(1-\rho)}{n}\left(\frac{N-n}{N-1}\right) = 250^2 \cdot \frac{26.96(1-0.9712)}{12}\left(\frac{250-12}{250-1}\right) = 3865.349,$$

et la marge d'erreur sur l'estimation est $2\sqrt{\hat{V}(\hat{\tau}_{Y;L})} = 2\sqrt{3865.349} = 124.3439$, ce qui est de loin meilleur. ■

39. Une agence de publicité s'inquiète de l'effet que peut avoir une nouvelle campagne promotionnelle régionale sur les ventes totales en dollars d'un produit particulier. Un EAS de 20 magasins a été constitué à partir de la population de 452 magasins dans lesquels le produit est vendu. Les données trimestrielles sur les ventes ont été obtenues pour la période de trois mois en cours et la période de trois mois précédant la nouvelle campagne et sont présentées dans le tableau ci-dessous. On sait également que les ventes totales pour l'ensemble des 452 magasins au cours de la période de trois mois précédant la nouvelle campagne étaient de 216,256.

Magasin	**Antérieures** x	**Actuelles** y	**Magasin**	**Antérieures** x	**Actuelles** y
1	208	239	11	599	626
2	400	428	12	510	538
3	440	472	13	828	888
4	259	276	14	473	510
5	351	363	15	924	998
6	880	942	16	110	171
7	273	294	17	829	889
8	487	514	18	257	265
9	183	195	19	388	419
10	863	897	20	244	257

(a) Tracer un diagramme de dispersion des valeurs des ventes actuelles par rapport aux valeurs des ventes antérieures. Quelle méthode d'estimation semble plus appropriée? Expliquer.

(b) Déterminer un I.C. pour les ventes totales actuelles à environ 95% en utilisant l'estimation par le quotient.

(c) Répéter l'étape (b) en utilisant l'estimation par la régression.

(d) Comparer les marges d'erreur sur l'estimation pour les intervalles de confiance obtenus aux étapes (b) et (c). Laquelle est la plus élevée? Est-ce conforme aux attentes? Expliquer.

(e) Répéter l'étape (b) en utilisant l'estimation par la différence.

(f) L'estimation par la différence est-elle une approche raisonnable pour donner un estimé du total des ventes en cours? Expliquer.

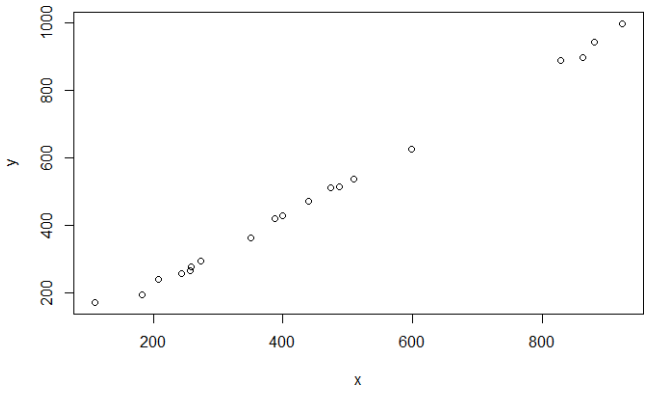
(g) Comparer les marges d'erreur sur l'estimation pour les intervalles de confiance obtenus aux étapes (c) et (f). Laquelle est la plus élevée? Est-ce conforme aux attentes? Expliquer.

(h) Combien de magasins faudrait-il échantillonner afin de donner un estimé du total des ventes actuelles en préservant une marge d'erreur sur l'estimation de 2500$ si l'on utilise l'estimation par le quotient?

(i) Répéter l'étape (h) en utilisant l'estimation par la régression et l'estimation par la différence.

Solution:

(a) Voici le nuage de points:



La droite de régression est $y = 10.11 + 1.05x$; la relation entre les deux variables est clairement fortement linéaire (et la droite de régression croise l'axe des Y très près de l'origine, à 10.1052), mais la variance ne semble pas être proportionnelle à X, et l'estimateur de régression est probablement plus efficace que l'estimateur du quotient.

(b) On obtient 231611.63 ± 3073.83 selon la méthode du quotient.

(c) On obtient 231581.43 ± 2950.85 selon la méthode de la régression.

(d) L'estimateur du quotient est effectivement moins "serré" que l'estimateur de régression, mais la différence n'est pas si importante que cela.

(e) La méthode de la différence est sans doute adéquate, puisque la relation entre la variable auxiliaire X et la variable réponse Y est fortement linéaire, et puisque la pente de la droite de régression s'approche de 1.

(f) On obtient 231510.78 ± 3849.01 selon la méthode de la différence.

(g) L'estimateur de régression est plus "serré" que celui de la différence, ce qui n'est pas surprenant puisque c'est toujours le cas sauf si $\rho\frac{\sigma_Y}{\sigma_X} = 1$.

(h) Selon la méthode du quotient, nous devons avoir

$$n = \frac{N\sigma_W^2}{(N-1)D + \sigma_W^2} \approx \frac{452(241.940)}{\frac{451(2500)^2}{4(452)^2} + 241.940} = 29.63;$$

on devrait donc prélever au moins 30 unités.

(i) Selon la méthode de l'estimateur de régression, nous avons

$$V(\hat{\mu}_{Y;L}) \approx \frac{\text{EQM}}{n}\left(\frac{N-n}{N-1}\right).$$

On résoud $B = 2N\sqrt{\frac{\text{EQM}}{n}\left(\frac{N-n}{N-1}\right)}$ pour n, ce qui donne $n \approx \frac{N\cdot\text{EQM}}{(N-1)D+\text{EQM}}$, où $D = \frac{B^2}{4N^2}$. Avec la marge requise, cela nous donne

$$n \approx \frac{452(222.968)}{\frac{451(2500)^2}{4(452)^2} + 222.968} = 27.44,$$

d'où il nous faudrait un échantillon d'au moins 28 unités. Finalement, pour l'estimateur de la différence, nous avons

$$V(\hat{\mu}_{Y;D}) \approx \frac{\sigma_D^2}{n}\left(\frac{N-n}{N-1}\right).$$

On résoud, $B = 2N\sqrt{\frac{\sigma_D^2}{n}\left(\frac{N-n}{N-1}\right)}$ pour n, ce qui donne $n \approx \frac{N\cdot\sigma_D^2}{(N-1)D+\sigma_D^2}$, où $D = \frac{B^2}{4N^2}$. Avec la marge requise, cela nous donne

$$n \approx \frac{452(379.355)}{\frac{451(2500)^2}{4(452)^2} + 379.355} = 44.78,$$

d'où il nous faudrait prélever au moins 45 unités. ■

40. On examine la relation entre la consommation de carburant au ralenti ("idling") Y et la capacité du moteur X en prélevant un EAS de taille $n = 15$ à même une population de $N = 227,133$ automobiles ayant en moyenne une cylindrée de $\mu_X = 2.5$L:

Véhicule	C.C. au ralenti	Cylindrée	Véhicule	C.C. au ralenti	Cylindrée
1	0.18	1.2	9	0.45	2.5
2	0.21	1.2	10	0.52	3.4
3	0.17	1.2	11	0.61	3.4
4	0.31	1.8	12	0.44	3.4
5	0.34	1.8	13	0.62	4.2
6	0.29	1.8	14	0.65	4.2
7	0.42	2.5	15	0.59	4.2
8	0.39	2.5			

(a) Donner un estimé de la consommation moyenne de carburant au ralenti pour la population de 227,133 automobiles à l'aide d'une estimation par le quotient, et déterminer la marge d'erreur sur l'estimation.

(b) Répéter l'étape (a) en utilisant l'estimation par la régression.

(c) Répéter l'étape (a) en utilisant l'estimation par la différence.

(d) Expliquer les résultats obtenus aux étapes (a), (b), et (c) en utilisant un diagramme de dispersion et une ligne de meilleur ajustement.

Solution:

```
> y = c(0.18, 0.21, 0.17, 0.31, 0.34, 0.29,
      0.42, 0.39, 0.45, 0.52, 0.61, 0.44,
      0.62, 0.65, 0.59)
> x = c(1.2, 1.2, 1.2, 1.8, 1.8, 1.8, 2.5, 2.5, 2.5, 3.4, 3.4, 3.4, 4.2, 4.2, 4.2)
> plot(x,y)
> n = 15
> N = 227133
> mux = 2.5
> taux = mux*N
> sumx = sum(x)
> sumy = sum(y)
> sumxy = sum(x*y)
> sumx2 = sum(x^2)
> sumy2 = sum(y^2)
> sy2 = 1/(n-1)*sum((y-mean(y))^2)
> sx2 = 1/(n-1)*sum((x-mean(x))^2)
> sxy = 1/(n-1)*sum((y-mean(y))*(x-mean(x)))
> r = sumy/sumx
> swr2 = 1/(n-1)*(sumy2-2*r*sumxy+r*r*sumx2)
> Vr = 1/mux^2*swr2/n*(1-n/N)
> muyr = r*mux
> Br = 2*sqrt(swr2/n*(1-n/N))
> rho = sxy/sqrt(sy2*sx2)
> b = rho*sqrt(sy2)/sqrt(sx2)
> a = mean(y)-b*mean(x)
```

```
    > swl2 = (n-1)/(n-2)*sy2*(1-rho^2)
    > muyl = mean(y)+b*sum(mux-mean(x))
    > Bl = 2*sqrt(swl2/n*(1-n/N))
    > d = y-x
    > dm = mean(d)
    > muyd = mux+dm
    > sd2 = 1/(n-1)*sum((d-dm)^2)
    > Bd = 2*sqrt(sd2/n*(1-n/N))
    > c(muyr,  Br, muyr-Br,muyr+Br)
[1] 0.3937659 0.0241291 0.3696368 0.4178950
    > c(muyl,  Bl, muyl-Bl,muyl+Bl)
[1] 0.39581956 0.02285444 0.37296512 0.41867400
    > c(muyd,  Bd, muyd-Bd,muyd+Bd)
[1]  0.2926667  0.4956889 -0.2030222  0.7883555
```

■

41. Le modèle théorique utile $Y_i = \beta X_i + D_i$, où D_i représente l'écart par rapport à la droite, peut être utilisé afin de comparer divers estimateurs de quotients. Pour une valeur donnée de $X = x$, supposons que les valeurs de Y soient éparpillées autour de la droite, de sorte que l'espérance et la variance des écarts soient

$$E[D \mid X = x] = 0 \quad \text{et} \quad V[D \mid X = x] = \sigma^2 x^{2a}.$$

Considérons un estimateur général de β ayant la forme $b = \sum_{i=1}^{n} c_i y_i$, où c_i peut dépendre de x_i.

(a) Trouver une condition sur les coefficients afin de garantir que b est un estimateur non biaisé de β, étant donné les x observés.

(b) Déterminer une expression pour la variance de b en fonction de a, conditionnellement aux $x > 0$ observés.

(c) Pour une valeur donnée de a, trouver l'estimateur non biaisé de la classe ci-dessus avec une variance conditionnelle minimale.

(d) Si $a = 0$, quel estimateur présente la plus petite variance conditionnelle? Et si $a = 0.5$? Et pour $a = 1$?

(e) Discuter des conséquences de cette analyse pour l'estimation de μ_Y par le quotient et par la régression.

Solution:

(a) On doit avoir $E(b) = \beta$. Puisque $E(D_i|X = x_i) = 0$,

$$E(y_i|X = x_i) = E(\beta x_i|X = x_i) + E(D_i|X = x_i) = \beta x_i + 0 = \beta x_i \quad \text{for all } i = 1, \ldots, n.$$

Mais

$$\begin{aligned} E(b) &= E\left(c_1 y_1 + \cdots + c_n y_n\right) = c_1 E(y_1|X = x_1) + \cdots + c_n E(y_n|X = x_n) = c_1 \beta x_1 + \cdots + c_n \beta x_n \\ &= \beta(c_1 x_1 + \cdots + c_n x_n) \end{aligned}$$

Pour que $E(b) = \beta$, on doit avoir $c_1 x_1 + \cdots + c_n x_n = 1$, à moins, bien sûr, que $\beta = 0$, auquel cas b est nécessairement un estimateur sans biais de β.

(b) Puisque $V(D_i|X = x_i) = \sigma^2 x_i^{2a}$,

$$V(y_i|X = x_i) = V(\beta x_i + D_i|X = x_i) = V(D_i|X = x_i) = \sigma^2 x_i^{2a} \quad \text{pour tout } i = 1, \ldots, n.$$

Si les x_i sont indépendants, les y_i le sont également et

$$\begin{aligned} V(b) &= V\left(c_1 y_1 + \cdots + c_n y_n\right) = c_1^2 V(y_1|X = x_1) + \cdots + c_n^2 V(y_n|X = x_n) \\ &= c_1^2 \sigma^2 x_1^{2a} + \cdots + c_n^2 \sigma^2 x_n^{2a} = \sigma^2 \left(c_1^2 x_1^{2a} + \cdots + c_n^2 x_n^{2a}\right) \end{aligned}$$

(c) Supposons que $\beta \neq 0$. On cherche à minimiser $V(b) = f(x) = \sigma^2 \left(c_1^2 x_1^{2a} + \cdots + c_n^2 x_n^{2a}\right)$, sujet à la contrainte $g(x) = c_1 x_1 + \cdots + c_n x_n - 1 = 0$. Selon la méthode des multiplicateurs de Lagrange, on cherche les points x tels que

$$\begin{aligned} \nabla f(x) &= \lambda \nabla g(x) \\ g(x) &= 0 \end{aligned}$$

c'est-à-dire que

$$2a\sigma^2 c_i^2 x_i^{2a-1} = \lambda c_i, \quad i = 1, \ldots, n$$
$$c_1 x_1 + \cdots + c_n x_n = 1$$

On résoud pour c_i: la première ligne nous donne soit $c_i = 0$ ou soit $c_i = \frac{\lambda}{2a\sigma^2 x_i^{2a-1}}$ pour tout $i = 1, \ldots, n$. Ainsi, $c_i x_i = 0$ ou $c_i x_i = \frac{\lambda}{2a\sigma^2 x_i^{2a-2}}$ pour tout $i = 1, \ldots, n$.
Mais $c_i \neq 0$ pour au moins un i, sinon $g(x) = -1 \neq 0$. Soit $C = \{i : c_i \neq 0\} \neq \varnothing$; nous devons avoir

$$1 = \sum_{i \in C} \frac{\lambda}{2a\sigma^2 x_i^{2a-2}} = \frac{\lambda}{2a\sigma^2} \sum_{i \in C} x_i^{2-2a},$$

d'où $\lambda = \frac{2a\sigma^2}{\sum_{i \in C} x_i^{2-2a}}$. Ainsi,

$$c_i = \frac{\lambda}{2a\sigma^2 x_i^{2a-1}} = \left(x_i^{2a-1} \sum_{j \in C} x_j^{2-2a} \right)^{-1}, \quad \text{pour } i \in C, \quad \text{et} \quad c_i = 0, \quad \text{pour } i \notin C.$$

Si $c_i = 0$, on aurait pu tout aussi bien ne pas choisir le $i-$ème point dans le modèle. Conséquemment, on suppose que $c_i \neq 0$ pour tout i, de sorte que

$$c_i = \frac{x_i^{1-2a}}{x_1^{2-2a} + \cdots + x_n^{2-2a}}, \quad i = 1, \ldots, n.$$

(d) Lorsque $a = 0$, nous avons $c_i = \frac{x_i}{x_1^2 + \cdots + x_n^2}$, $i = 1, \ldots, n$, d'où l'estimateur sans biais ayant la plus faible variance est

$$b = c_1 y_1 + \cdots + c_n y_n = \frac{1}{\sum x_i^2} (x_1 y_1 + \cdots + x_n y_n) = \frac{\sum x_i y_i}{\sum x_i^2};$$

c'est l'estimateur des moindres carrés de la pente de la droite de régression à travers l'origine.
Lorsque $a = 0.5$, nous avons $c_i = \frac{1}{x_1 + \cdots + x_n}$, $i = 1, \ldots, n$, d'où l'estimateur sans biais ayant la plus faible variance est

$$b = c_1 y_1 + \cdots + c_n y_n = \frac{1}{\sum x_i} (y_1 + \cdots + y_n) = \frac{\sum y_i}{\sum x_i};$$

c'est l'estimateur du quotient.
Lorsque $a = 1$, nous avons $c_i = \frac{1}{x_i}$, $i = 1, \ldots, n$, , d'où l'estimateur sans biais ayant la plus faible variance est

$$b = c_1 y_1 + \cdots + c_n y_n = \frac{y_1}{x_1} + \cdots + \frac{y_n}{x_n} = \sum \frac{y_i}{x_i}.$$

(e) Si nous pouvons montrer que le modèle de la ligne (non-horizontale) passant par l'origine est approprié et que l'hypothèse de variance constante se vérifie, l'estimateur sans biais le plus efficace de μ_Y est basé sur l'estimateur de régression. Si, par contre, il est démontré que la variance est proportionnelle au niveau des x, l'estimateur sans biais le plus efficace de μ_Y est basé sur l'estimateur du quotient. ■

42. On cherche à donner un estimé de la distance quotidienne moyenne parcourue durant la saison hivernale 2012 en Ontario par certains types de véhicules. Un échantillon est prélevé à même la flotte Ontarienne (de taille $N = 7,868,359$) contenant des information au sujet du type et de l'âge du véhicule, et de la consommation quotidienne de carburant; les données relatives aux répondants sont recueillies dans le fichier `Autos_RLD.xlsx`. Donner un intervalle de confiance pour la distance quotidienne moyenne, à environ 95%, avec coefficient de variation correspondant, en utilisant l'estimation par le quotient, l'estimation par la régression, et l'estimation par la différence. Faire de même dans chaque strate (et chaque combinaison). [La majorité des notes seront attribuées pour la discussion et la présentation des résultats.]

Solution: ■

43. La lettre d'information suivante a été envoyée à tous les membres du personnel et du corps enseignant de l'Université de XXXXXX.

> **CENTRE DE RECHERCHE SUR LES LENTILLES DE CONTACT**
> **ÉCOLE D'OPTOMÉTRIE, UNIVERSITÉ DE XXXXXX**
> **ARRÊT DU PORT DE LENTILLES DE CONTACT**
>
> On estime qu'en Amérique du Nord, entre 10% et 30% des gens portant des lentilles de contact ont cessé de le faire.
>
> Grâce à ce questionnaire, nous espérons identifier le pourcentage de porteurs de lentilles de contact de la ville de XXXXXX qui ont abandonné le port de lentilles, les raisons de cet abandon, et le stade auquel il s'est produit après la première adaptation. Les renseignements recueillis nous aideront à améliorer notre compréhension de l'abandon prématuré du port de lentilles de contact, ce qui sera au bénéfice des porteurs actuels et éventuels de lentilles de contact. Une confidentialité totale est assurée dans le cadre de ce projet.
>
> Si vous avez cessé de porter des lentilles de contact ou si vous en portez présentement, il est essentiel pour le succès de cette enquête que vous remplissiez ce questionnaire avec soin et attention. Cela ne devrait prendre qu'environ 10 minutes de votre temps. Veuillez retourner le questionnaire à xxxxxx xxxxx, Centre de recherche sur les lentilles de contact de l'École d'optométrie (XXXX XXX XXX). Pour de plus amples informations, veuillez téléphoner au xxx-xxxx x xxxx.
>
> Si vous n'avez jamais porté de lentilles de contact ou si vous ne souhaitez pas participer, veuillez renvoyer le questionnaire sans réponse à l'adresse ci-dessus.
>
> En appréciation de votre temps et de votre attention, je vous remercie.
>
> Ce questionnaire a été approuvé par le Bureau de la recherche sur les humains et les animaux de l'Université XXXXXX.

Puisque la chercheuse a demandé aux gens qui n'ont jamais porté de lentilles de contact de ne pas remplir le questionnaire, on peut supposer qu'elle a défini sa population à l'étude comme étant l'ensemble du personnel et du corps enseignant de l'Université de XXXXXX qui ont déjà porté ou portent actuellement des lentilles de contact.

(a) Que sont
   i. la population cible;
   ii. la population répondante;
   iii. l'échantillon visé, et
   iv. l'échantillon réalisé.

(b) Énumérer des variables-réponse et des attributs de population d'intérêt dans cette étude.

(c) En ce qui concerne les attributs de la population d'intérêt sélectionnés à l'étape (b), expliquer comment chacune des catégories suivantes d'erreurs non dues à l'échantillonnage est susceptible de se produire dans cette enquête: erreur de couverture, erreur de non-réponse, erreur de mesure, erreur de traitement, et erreur d'échantillonnage.

(d) La chercheuse a-t-elle tenté de minimiser les problèmes liés à la non-réponse? Expliquer.

(e) Élaborer un questionnaire pour cette étude. [La majorité des notes seront attribuées à la présentation et à la logique qui sous-tend le type et l'ordre des questions.]

Solution:

(a) i. La population cible est constituée de tous les habitants de la ville de Waterloo (et de Kitchener?) qui ont porté ou portent actuellement des lentilles de contact.

ii. La population des répondants comprend tous les membres de la population étudiée qui répondraient s'ils étaient sélectionnés pour participer à l'enquête, c'est-à-dire tous les employés et les professeurs de l'Université de Waterloo qui ont porté ou portent actuellement des lentilles de contact et qui répondraient s'ils étaient sélectionnés pour participer à l'enquête.

iii. L'échantillon visé est le sous-ensemble des membres de la population étudiée dont l'enquêteur a l'intention de mesurer les caractéristiques, c'est-à-dire, dans le cas présent, tous les membres du personnel et de la faculté de l'Université de Waterloo qui ont porté ou portent actuellement des lentilles de contact et qui résident dans la ville de Waterloo (et peut-être également Kitchener).

iv. L'échantillon réalisé est le sous-ensemble des membres de la population étudiée dont l'enquêteur mesure effectivement les caractéristiques, ce qui, dans le cas présent, est le sous-ensemble de tous les employés et professeurs de l'Université de Waterloo qui ont porté ou portent actuellement des lentilles de contact (qu'ils résident à Waterloo/Kitchener ou non) et qui ont rempli le questionnaire et l'ont renvoyé. Il pourrait y avoir un problème concernant les membres du personnel et du corps professoral figurant sur les listes d'employés de l'Université même s'ils/elles ont pris leur retraite ou sont décédés, ou des omissions dues à des embauches effectuées après la production de la liste d'employés ; dans ce cas, il suffit de redéfinir la population étudiée comme étant l'ensemble du personnel et du corps professoral actuels de l'Université de Waterloo qui ont porté ou portent actuellement des lentilles de contact et qui se retrouvent dans la population de référence du chercheur.

(b) L'une des variables réponses est désignée par la variable caractéristique $u_j = 1$ si la $j-$ième unité de la population cible a porté des verres de contact dans le passé mais a cessé de le faire à un moment donné, et $u_j = 0$ si la $j$ème unité de la population cible porte toujours des verres de contact. La proportion p est la proportion des habitants de la ville de Waterloo qui ont porté des verres de contact dans le passé et qui ont maintenant cessé de le faire.

(c) i. L'erreur de couverture est due à des différences dans l'étude et la population cible : le personnel et les professeurs de l'Université de Waterloo qui ont porté ou portent actuellement des verres de contact peuvent ne pas résider dans la ville de Waterloo elle-même. Dans le même ordre d'idées, certains nouveaux employés de l'Université de Waterloo qui ont porté ou portent actuellement des verres de contact peuvent résider dans la ville sans pour autant se retrouver dans la population cible. J'ai également fait une remarque sur le personnel et le corps enseignant à la retraite ci-dessus. Enfin, je ne sais pas s'il est logique que l'abandon du port de verres de contact par le personnel et le corps enseignant soit considéré comme représentatif de la population générale des porteurs de verres de contact. Je me risquerais à penser que ces personnes lisent plus que la moyenne et que cela peut influencer leur décision de garder leurs verres plus longtemps qu'elles ne le feraient autrement (mais je ne suis pas certain de cela).

ii. L'erreur de non-réponse est due aux différences entre les populations étudiées et les populations interrogées. Je ne sais pas ce qu'il en est du personnel, mais si le corps professoral de l'Université de Waterloo ressemble un tant soit peu à celui de l'Université d'Ottawa, ce serait un petit miracle si la chercheuse obtenait un

taux de réponse supérieur à 10 % (je serais même surpris que de nombreux non-répondants retournent le questionnaire comme demandé). Soyons francs, même si le sujet du sondage peut être très intéressant pour l'enquêteur, j'ai du mal à imaginer que le taux de réponse soit très élevé, à moins qu'il y ait une sorte de récompense monétaire. Puisqu'aucune telle récompense n'est mentionnée... Je pense que les répondants seront principalement des personnes qui ont arrêté de porter des verres de contact, ce qui conduirait à une grande valeur de biais de non-réponse.

iii. Une erreur de mesure survient lorsque la valeur réelle de la variable de réponse n'est pas évaluée correctement. À moins que la question ne ressemble à celle du référendum sur l'indépendance du Québec de 1980, il est difficile d'imaginer que quelqu'un ne réponde pas correctement dans cette situation. "Êtes-vous un ancien porteur de verres de contact qui a cessé de les porter à un moment donné dans le passé ?" Il est peu probable qu'il s'agisse d'un facteur important dans cette étude.

iv. Son échantillon est constitué de l'ensemble de la population étudiée, puisqu'elle envoie un questionnaire à chaque unité de sa population étudiée. Ainsi, l'ensemble de la population interrogée est étudiée et il n'y a pas d'erreur d'échantillonnage.

(d) Étant donné que les personnes qui souhaitent ne pas répondre sont invitées à renvoyer le questionnaire de toute façon, il est logique que certains non-répondants potentiels décident de le remplir avant de le renvoyer, puisque cela ne devrait prendre que 10 minutes. Bien sûr, je continue de penser que la plupart des gens ne retourneront même pas le questionnaire, mais c'est mieux que si elle n'avait rien fait au sujet des non-réponses.

(e) Allez-y! ■

44. Dans un sondage mené près de 2227 canadien.ne.s, 214 des personnes interrogées reconnaissent avoir falsifié leur déclaration d'impôt sur le revenu. Pensez-vous que cette fraction est proche de la proportion réelle de personnes ayant commis cette infraction? Pourquoi? Discuter des difficultés à obtenir des renseignements précis sur une question de ce type.

Solution: Il est peu probable que la proportion soit proche de la réalité, car il s'agit d'un sujet très personnel et sensible. L'écart par rapport à la proportion réelle dépend de la méthode de collecte : en personne, la présence sociale de l'enquêteur peut faire craindre aux répondants d'être jugés ou dénoncés aux autorités. Au téléphone, les mêmes craintes existent, mais à un degré moindre, bien qu'un entretien téléphonique donne aux gens la possibilité de quitter l'entretien à mi-chemin s'ils n'aiment pas la direction que prennent les questions. Quoi qu'il en soit, si l'enquête a été menée sans contrôle à double insu (ce n'est peut-être pas la bonne expression pour décrire ce que je veux dire : donner aux gens deux questions, l'une sensible, l'autre non, leur demander de tirer à pile ou face et de répondre honnêtement à l'une ou l'autre des deux questions si elle tombe sur pile, et de répondre à n'importe quelle question comme ils le veulent si elle tombe sur face), le taux de non-réponse était probablement assez élevé. ■

45. Il y a une trentaine d'années, les lecteurs et lectrices du magazine *Popular Science* ont été invités à contacter un numéro de téléphone afin de donner leur réponse à la question suivante:

> Les États-Unis doivent-ils construire davantage de centrales à combustibles fossiles ou des nouveaux générateurs nucléaires dits "sûrs" afin répondre à une crise énergétique qui pourrait survenir dans les 10 prochaines années?

Sur le total des appels, 83% ont choisi l'option nucléaire. Est-ce que le sondage a bien été mené? Qu'en est-il de la formulation de la question? Les résultats semblent-ils constituer une bonne estimation de l'état d'esprit qui règnait au pays à l'époque?

Solution: Le plan d'échantillonnage de ce sondage est sans aucun doute de nature non probabiliste ; en fait, il s'agit d'un exemple d'échantillonnage volontaire. Dans ce type d'échantillonnage, les sujets controversés ont tendance à générer des réponses uniquement de la part de ceux qui ont des opinions très fortes sur le sujet du sondage, dans un sens comme dans l'autre, ce qui peut donner lieu à un biais de couverture important, car la majorité silencieuse est peu susceptible de se porter volontaire et d'exprimer son opinion.

Quant à la formulation de la question, avant même de lire les résultats du sondage, j'avais la nette impression que la question favoriserait l'option de la centrale nucléaire : je pense que les mots "sûrs" et "crise" (et "nouveau", compte tenu du public) devraient sans doute être supprimés.

J'ai également l'impression qu'il y a un peu de "double jeu" là-dedans : J'étais un peu jeune à l'époque, mais je crois me souvenir qu'une bonne partie du public américain ne pensait pas qu'il y aurait une crise énergétique dans les années 1990. Ne sont-ils pas en train de forcer les personnes interrogées à convenir qu'il va y avoir une crise énergétique ?

En outre, les centrales à combustibles fossiles et les centrales nucléaires sont-elles les seuls types de centrales électriques que l'on peut construire ? Qu'en est-il des centrales solaires ou des centrales hydroélectriques ? Il me semble que les options présentées dans cette question fermée ne sont pas collectivement exhaustives.

Pourquoi ne pas essayer quelque chose comme ce qui suit ?

(a) Pensez-vous que les États-Unis produisent autant d'énergie qu'ils en ont besoin ?
____ Oui ____ Non

(b) Pensez-vous que cela restera le cas au cours des 10 prochaines années ?
____ Oui ____ Non

(c) À la lumière de vos réponses aux questions (a) et (b), quel plan d'action proposeriez-vous ?
____ Construire davantage de centrales à combustibles fossiles
____ Construire davantage de centrales hydroélectriques
____ Construire de nouveaux générateurs nucléaires
____ Construire d'autres types de centrales électriques (éoliennes, solaires, etc.)
____ Maintenir le statu quo
____ Autre (veuillez préciser)

C'est plus difficile qu'il n'y paraît. Il faudrait y penser un peu plus.

Enfin, en raison de l'énorme biais de couverture, l'inférence devient inutile, comme c'est généralement le cas avec un échantillonnage non probabiliste. Mais, je doute fortement que les lecteurs de *Popular Science* soient représentatifs du public américain sur ce sujet particulier. En 1990, la catastrophe de Three Mile Island devait être suffisamment présente dans l'esprit de nos voisins pour que la victoire convaincante de l'option nucléaire ait très peu de chances de se produire lors d'un recensement de la population américaine. En fait, je doute des bonnes intentions des sondeurs... ■

46. On cherche à évaluer la distance moyenne quotidienne parcourue par les voitures Ontariennes en 2012, ainsi que la consommation d'essence quotidienne, le nombre de voyages quotidiens, le nombre de passagers, la proportion de l'utilisation au ralenti (idling), la vitesse, etc. Discuter du mode de collecte des données, de la base de sondage, des erreurs d'échantillonnage (et contre-mesures), et des problèmes éventuels. Élaborer un plan d'échantillonnage et un questionnaire permettant de répondre à ces questions et d'éviter les embuscades.

Solution: Vous aurez tous et toutes, bien sûr, vos propres réponses. ■

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:

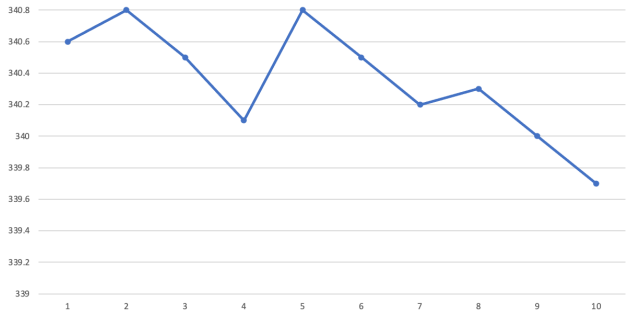
(a) L'intervalle de confiance de μ à environ 95% est

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

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. ■

52. On cherche à donner un estimé de la distance quotidienne moyenne parcourue durant la saison hivernale 2012 en Ontario par certains types de véhicules. La consommation de carburant quotidienne est aussi d'intérêt, tout comme la proportion des véhicules qui ne sont pas utilisés. Comment pourrait-on utiliser l'échantillonnage par grappe pour y arriver? Expliquer les hypothèses et suppositions, et donner des intervalles de confiance à environ 95% pour les quantités d'intérets (utiliser l'ensemble de données `Autos_STR.xlsx`).

53. Ue chercheuse travaillant dans une zone urbaine souhaite estimer la valeur moyenne d'une variable fortement corrélée à l'ethnicité. Elle pense qu'elle devrait utiliser un EPG, en utilisant les pâtés de maisons en tant que grappes et les adultes vivant dans les pâtés de maisons en tant qu'unités. Expliquer si un EPG est un plan d'échantillonnage approprié ou non dans chacune des situations suivantes.

    (a) La plupart des adultes de certains pâtés de maisons appartiennent à l'ethnie la plus répandue, tandis que la plupart des adultes d'autres pâtés de maisons appartiennent à d'autres ethnies.

    (b) La proportion d'adultes n'appartenant pas à l'ethnie la plus répandue est à peu près la même dans tous les pâtés de maisons et ne se rapproche ni de 0, ni de 1.

    (c) La proportion d'adultes n'appartenant pas à l'ethnie la plus répandue diffère d'un pâté de maisons à l'autre de la manière à laquelle on pourrait s'y attendre si on assignait les adultes dans les grappes de manière aléatoire.

54. Un fabricant de scies à ruban souhaite estimer le coût moyen des réparations mensuelles des scies qu'il a vendues. Il ne peut pas obtenir un coût de réparation pour chaque scie, mais il a accès au coût de réparation de toutes les scies et au nombre de scies possédées par chaque client. Il décide donc d'utiliser un EPG, chaque client constituant une grappe. Le fabricant sélectionne une EAS de taille $m = 20$ parmi les $M = 96$ clients qu'il dessert. Les données pour le mois dernier sont les suivantes.

Client	1	2	3	4	5	6	7	8	9	10
# Scies	3	7	11	9	2	12	14	3	5	9
Coût	50	110	230	140	60	280	240	45	60	230
Client	11	12	13	14	15	16	17	18	19	20
# Scies	8	6	3	2	1	4	12	6	5	8
Coût	140	130	70	50	10	60	280	150	110	120

(a) Donner un estimé du coût moyen de réparation par scie au cours du mois dernier, ainsi que la marge d'erreur sur l'estimation.

(b) Donner un estimé du montant total dépensé par les $M = 96$ clients pour la réparation des scies à ruban, ainsi que la marge d'erreur sur l'estimation.

(c) Après avoir vérifié ses registres de vente, le fabricant découvre qu'il a vendu un total de $N = 710$ scies à ruban à ses $M = 96$ clients. À l'aide de ces informations supplémentaires, donner un estimé du montant total dépensé par ses clients pour la réparation des scies, ainsi que la marge d'erreur sur l'estmation correspondante.

(d) Le même fabricant souhaite estimer le coût moyen de réparation mensuelle par scie pour le mois à venir. Combien de grappes doit-il sélectionner dans son échantillon s'il veut que la marge d'erreur sur l'estimation soit inférieure à 3?

55. Un type de carte de circuit comporte 12 micro-puces par carte. Lors de l'inspection de contrôle de la qualité de dix de ces cartes, le nombre de micropuces défectueuses sur chacune des dix cartes était le suivant : 2, 0, 1, 3, 2, 0, 0, 1, 3, 4.

    (a) Donner un estimé de la proportion de micro-puces défectueuses dans la population dont l'échantillon a été tiré, ainsi que la marge d'erreur sur l'estimation.

    (b) Si l'échantillon de dix cartes utilisé provient d'un lot de 250 cartes de ce type, donner un estimé du nombre total de micro-puces défectueuses dans le lot, ainsi que la marge d'erreur sur l'estimation.

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%.

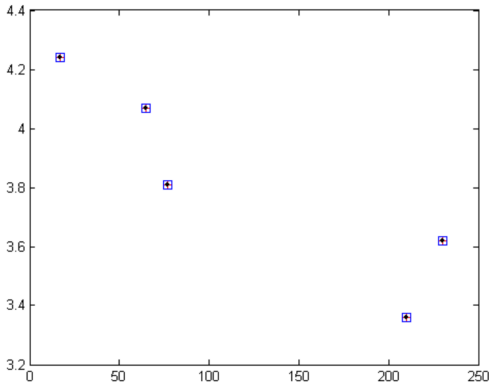
(c) Nous avons introduit l'estimateur $M\overline{y}_T$ afin de déterminer le total dans la population. En divisant cet estimateur par N, on obtient un estimateur pour la moyenne dans la population. En utilisant l'estimateur $\frac{M}{N}\overline{y}_T$, déterminer un I.C. pour le score moyen de tous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

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

■

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

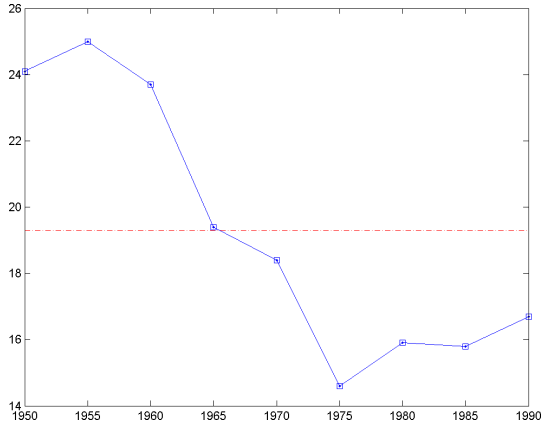
$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{9} \cdot 173.6 = 19.2889,$$

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. ■

61. On considère une population de $N = 10$ cartes de circuits imprimés, ayant chacune un nombre différent de composantes, comme indiqué dans le tableau ci-dessous. Le nombre de composantes défectueux sur chaque carte est également indiqué.

Carte	1	2	3	4	5	6	7	8	9	10
# Composantes	10	12	22	8	16	24	9	10	8	31
# Défauts	1	1	3	1	2	3	1	1	0	3

Prélever un échantillon PPT (avec remise) de taille $n = 3$ et déterminer un intervalle de confiance pour le nombre total de défauts sur la collection de cartes de circuits imprimés, à environ 95%.

Solution: Le tableau des étendues cumulées est

Carte i	**# composantes** x_i	**Étendue associée**	π_i	**# défauts** u_i
1	10	001-010	10/150	1
2	12	011-022	12/150	1
3	22	023-044	22/150	3
4	8	045-052	8/150	1
5	16	053-068	16/150	2
6	24	069-092	24/150	3
7	9	093-101	9/150	1
8	10	102-111	10/150	1
9	8	112-119	8/150	0
10	31	120-150	31/150	3

On prélève un échantillon PPT (avec remise) de $n = 3$ unités, où le nombre de composantes d'une carte représente sa taille. On choisit $n = 3$ entiers au hasard entre 1 et 150: on obtient 148, 90, et 93 (vos entiers seront sans doute différents, bien sûr). Les unités associées sont la 10ième, la 6ième, et la 7ième:

$$\{y_1 = u_{10} = 3, y_2 = u_6 = 3, y_3 = u_7 = 1\}.$$

Les poids des 3 unités dans l'échantillon sont

$$w_1 = \frac{u_{10}}{\pi_{10}} = \frac{3}{31/150}, \quad w_2 = \frac{u_6}{\pi_6} = \frac{3}{24/150}, \quad w_3 = \frac{u_7}{\pi_7} = \frac{1}{9/150}.$$

L'estimateur

$$\hat{\tau}_{\text{ppt}} = \frac{1}{3}\sum_{i=1}^{3} w_i = \frac{1}{3}\left[\frac{3}{31/150} + \frac{3}{24/150} + \frac{1}{9/150}\right] = 16.64,$$

sa variance d'échantillonnage est

$$\hat{\text{V}}(\hat{\tau}_{\text{ppt}}) = \frac{1}{3(3-1)}\left[\sum_{i=1}^{3} w_i^2 - 3\hat{\tau}_{\text{ppt}}^2\right] = \frac{1}{6}\left[\left(\frac{3}{31/150}\right)^2 + \left(\frac{3}{24/150}\right)^2 + \left(\frac{1}{9/150}\right)^2 - 3(16.64)^2\right] = 1.49.$$

L'intervalle de confiance recherché est ainsi $16.64 \pm 2\sqrt{1.49} \equiv (14.20, 19.08)$. ■

62. Montrer que $\hat{V}(\hat{\tau}_{\text{ppt}})$ est un estimateur non-biasé de $V(\hat{\tau}_{\text{ppt}})$.

Solution: Il suffit de montrer que

$$E\left[\hat{V}(\hat{\tau}_{\text{ppt}})\right] = V(\hat{\tau}_{\text{ppt}}).$$

Notons au départ que

$$\begin{aligned}
V(w_i) &= \sum_{j=1}^{N}(w_j-\tau)^2 P(w_i=w_j) = \sum_{j=1}^{N}\left(\frac{u_j}{\pi_j}-\tau\right)^2\pi_j \\
&= \sum_{j=1}^{N}\left(\frac{u_j^2}{\pi_j^2}-2\frac{u_j}{\pi_j}\tau+\tau^2\right)\pi_j = \sum_{j=1}^{N}\left(\frac{u_j^2}{\pi_j}-2u_j\tau+\tau^2\pi_j\right) \\
&= \sum_{j=1}^{N}\frac{u_j^2}{\pi_j}-2\tau\sum_{j=1}^{N}u_j+\tau^2\sum_{j=1}^{N}\pi_j = \sum_{j=1}^{N}\frac{u_j^2}{\pi_j}-2\tau^2+\tau^2 = \sum_{j=1}^{N}\frac{u_j^2}{\pi_j}-\tau^2.
\end{aligned}$$

De plus, $V(w_i) = E(w_i^2) - E^2(w_i) = E(w_i^2) - \tau^2$, d'où

$$E(w_i^2) = V(w_i) + \tau^2 = \sum_{j=1}^{N}\frac{u_j^2}{\pi_j} - \tau^2 + \tau^2 = \sum_{j=1}^{N}\frac{u_j^2}{\pi_j}.$$

De la même manière, $V(\hat{\tau}_{\text{ppt}}) = E(\hat{\tau}_{\text{ppt}}^2) - E^2(\hat{\tau}_{\text{ppt}}) = E(\hat{\tau}_{\text{ppt}}^2) - \tau^2$, d'où $E(\hat{\tau}_{\text{ppt}}^2) = V(\hat{\tau}_{\text{ppt}}) + \tau^2$.

Ainsi,

$$\begin{aligned}
E\left[\hat{V}(\hat{\tau}_{\text{ppt}})\right] &= E\left[\frac{1}{n(n-1)}\left(\sum_{i=1}^{n}w_i^2 - n\hat{\tau}_{\text{ppt}}^2\right)\right] = \frac{1}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}E(w_i^2) - E(\hat{\tau}_{\text{ppt}}^2)\right] \\
&= \frac{1}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{N}\frac{u_j^2}{\pi_j}\right] - (V(\hat{\tau}_{\text{ppt}}) + \tau^2)\right] = \frac{1}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{N}\frac{u_j^2}{\pi_j}\right] - \tau^2 - V(\hat{\tau}_{\text{ppt}})\right] \\
&= \frac{1}{n-1}\left[\underbrace{\sum_{j=1}^{N}\frac{u_j^2}{\pi_j} - \tau^2}_{nV(\hat{\tau})} - V(\hat{\tau}_{\text{ppt}})\right] = \frac{1}{n-1}[nV(\hat{\tau}_{\text{ppt}}) - V(\hat{\tau}_{\text{ppt}})] = V(\hat{\tau}_{\text{ppt}}),
\end{aligned}$$

ce qui termine la démonstration. ■

63. Soit $\mathcal{Y} = \{y_1, \ldots, y_n\}$ un échantillon PPT prélevé **sans remise** à partir de la population $\mathcal{U} = \{u_1, \ldots, u_N\}$. Soient $\pi_j = P(y_i = u_j \in \mathcal{Y})$ et $\pi_{j,\ell} = P(y_i = u_j, y_k = u_\ell \in \mathcal{Y})$. Pour chaque unité $u_j \in \mathcal{U}$, posons $t_j = 1$ si $u_j \in \mathcal{Y}$ et $t_j = 0$ si $u_j \notin \mathcal{Y}$. L'expression

$$\hat{\tau}_{\mathrm{HT}} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \sum_{j=1}^{N} t_j \frac{u_j}{\pi_j}$$

représente l'estimateur de Horvitz-Thompson du total de la population $\tau = u_1 + \cdots + u_N$.

(a) Montrer que $\hat{\tau}_{\mathrm{HT}}$ est un estimateur non-biasé de τ.

(b) Montrer que

$$V(\hat{\tau}_{\mathrm{HT}}) = \sum_{j=1}^{N} \frac{1-\pi_j}{\pi_j} u_j^2 + \sum_{j=1}^{N} \left\{ \sum_{\ell=1}^{j-1} \frac{\pi_{j,\ell} - \pi_j \pi_\ell}{\pi_j \pi_\ell} u_j u_\ell + \sum_{\ell=j+1}^{N} \frac{\pi_{j,\ell} - \pi_j \pi_\ell}{\pi_j \pi_\ell} u_j u_\ell \right\}.$$

Indice: les t_j ne sont pas indépendants. Quelle forme prennent $V(t_j)$ et $\mathrm{Cov}(t_j, t_\ell)$?

(c) Montrer que

$$\hat{V}(\hat{\tau}_{\mathrm{HT}}) = \sum_{i=1}^{n} \frac{1-\pi_i}{\pi_i^2} y_i^2 + \sum_{i=1}^{n} \left\{ \sum_{k=1}^{i-1} \frac{\pi_{i,k} - \pi_i \pi_k}{\pi_{i,k}} \cdot \frac{y_i}{\pi_i} \cdot \frac{y_k}{\pi_k} + \sum_{k=i+1}^{n} \frac{\pi_{i,k} - \pi_i \pi_k}{\pi_{i,k}} \cdot \frac{y_i}{\pi_i} \cdot \frac{y_k}{\pi_k} \right\}$$

est un estimateur non-biasé de $V(\hat{\tau}_{\mathrm{HT}})$. Indice: ré-écrire les sommes à l'aide des t_j.

Solution:

(a) Puisque nous avons

$$E(t_j) = 0 \cdot P(t_j = 0) + 1 \cdot P(t_j = 1) = 0 \cdot (1 - \pi_j) + 1 \cdot \pi_j = \pi_j,$$

on voit que

$$E(\hat{\tau}_{\mathrm{HT}}) = E\left(\sum_{j=1}^{N} t_j \frac{u_j}{\pi_j} \right) = \sum_{j=1}^{N} E\left(t_j \frac{u_j}{\pi_j} \right) = \sum_{j=1}^{N} \frac{u_j}{\pi_j} E\left(t_j\right) = \sum_{j=1}^{N} \frac{u_j}{\pi_j} \pi_j = \sum_{j=1}^{N} u_j = \tau.$$

(b) La variable aléatoire t_j est binaire, d'où $V(t_j) = \pi_j(1 - \pi_j)$. De plus,

$$\mathrm{Cov}(t_j, t_\ell) = E(t_j t_\ell) - E(t_j)E(t_\ell) = \pi_{j,\ell} - \pi_j \pi_\ell.$$

Puisque les $\{t_j\}$ ne sont pas indépendents, nous avons

$$\begin{aligned}
V(\hat{\tau}_{\mathrm{HT}}) &= V\left(\sum_{j=1}^{N} t_j \frac{u_j}{\pi_j} \right) = \sum_{j=1}^{N} V\left(t_j \frac{u_j}{\pi_j} \right) + \sum_{\ell \neq j=1}^{N} \mathrm{Cov}\left(t_j \frac{u_j}{\pi_j}, t_\ell \frac{u_\ell}{\pi_\ell} \right) \\
&= \sum_{j=1}^{N} \left(\frac{u_j}{\pi_j} \right)^2 V\left(t_j\right) + \sum_{\ell \neq j=1}^{N} \frac{u_j}{\pi_j} \frac{u_\ell}{\pi_\ell} \mathrm{Cov}\left(t_j, t_\ell\right) \\
&= \sum_{j=1}^{N} \left(\frac{u_j}{\pi_j} \right)^2 \pi_j(1 - \pi_j) + \sum_{\ell \neq j=1}^{N} \frac{u_j}{\pi_j} \frac{u_\ell}{\pi_\ell} (\pi_{j,\ell} - \pi_j \pi_\ell) \\
&= \sum_{j=1}^{N} \frac{1-\pi_j}{\pi_j} u_j^2 + \sum_{\ell \neq j=1}^{N} \frac{\pi_{j,\ell} - \pi_j \pi_\ell}{\pi_j \pi_\ell} u_j u_\ell.
\end{aligned}$$

(c) Soit $u_j \in \mathcal{U}$. Si $u_j \in \mathcal{Y}$, u_j est la $i-$ème unité de l'échantillon $y_i = u_j$. Par conséquent, pour toute function g,

$$g(u_j)t_j = \begin{cases} g(y_i) & \text{si } u_j \text{ est une unité de } \mathcal{Y} \text{ (la } i-\text{ème, disons)} \\ 0 & \text{autrement} \end{cases}$$

Ainsi,

$$\begin{aligned} \hat{V}(\hat{\tau}_{\text{HT}}) &= \sum_{i=1}^{n} \frac{1-\pi_i}{\pi_i^2} y_i^2 + \sum_{k \neq i=1}^{n} \frac{\pi_{i,k} - \pi_i \pi_k}{\pi_{i,k}} \cdot \frac{y_i}{\pi_i} \cdot \frac{y_k}{\pi_k} \\ &= \sum_{j=1}^{N} \frac{1-\pi_j}{\pi_j^2} u_j^2 t_j + \sum_{\ell \neq j=1}^{N} \frac{\pi_{j,\ell} - \pi_j \pi_\ell}{\pi_{j,\ell}} \cdot \frac{u_j t_j}{\pi_j} \cdot \frac{u_\ell t_\ell}{\pi_\ell}, \end{aligned}$$

et

$$\begin{aligned} E\left(\hat{V}(\hat{\tau}_{\text{HT}})\right) &= \sum_{j=1}^{N} \frac{1-\pi_j}{\pi_j^2} u_j^2 E(t_j) + \sum_{\ell \neq j=1}^{N} \frac{\pi_{j,\ell} - \pi_j \pi_\ell}{\pi_{j,\ell}} \cdot \frac{u_j}{\pi_j} \cdot \frac{u_\ell}{\pi_\ell} E(t_j t_\ell) \\ &= \sum_{j=1}^{N} \frac{1-\pi_j}{\pi_j^2} u_j^2 \pi_j + \sum_{\ell \neq j=1}^{N} \frac{\pi_{j,\ell} - \pi_j \pi_\ell}{\pi_{j,\ell}} \cdot \frac{u_j}{\pi_j} \cdot \frac{u_\ell}{\pi_\ell} \pi_{j,\ell} \\ &= \sum_{j=1}^{N} \frac{1-\pi_j}{\pi_j} u_j^2 + \sum_{\ell \neq j=1}^{N} \frac{\pi_{j,\ell} - \pi_j \pi_\ell}{\pi_j \pi_\ell} u_j u_\ell = V(\hat{\tau}_{\text{HT}}), \end{aligned}$$

ce qui termine la démonstration. ■

64. Donner un intervalle de confiance de l'espérance de vie moyenne des pays de la planète en 2011, à environ 95%, à l'aide d'un échantillon PPT, où la taille est donnée par le logarithme du produit national brut. Justifier la réponse.

Solution: ■

65. Donner un intervalle de confiance de l'espérance de vie moyenne et du logarithme du produit national brut moyen des pays de la planète en 2011, à environ 95%, à l'aide d'un EAS2D (premier degré: continents; second degré: pays). Justifier la réponse.

Solution: ■

66. On cherche à donner un estimé de la distance quotidienne moyenne parcourue durant la saison hivernale 2012 en Ontario par certains types de véhicules. La consommation de carburant quotidienne est aussi d'intérêt, tout comme la proportion des véhicules qui ne sont pas utilisés. Comment pourrait-on utiliser l'échantillonnage à plusieurs degrés pour y arriver? Expliquer les hypothèses et suppositions, et donner des intervalles de confiance à environ 95% pour les quantités d'intérets (utiliser l'ensemble de données `Autos_STR.xlsx`).

Solution: ■

67. Donner des intervalle de confiance de l'espérance de vie moyenne des pays de la planète en 2011, à environ 95%, à l'aide d'un EAS2P (caractéristique principale: espérance de vie; caractéristique auxiliaire: logarithme du produit national brut). Justifier la réponse.

Solution: ■

68. On cherche à donner un estimé de la consommation de carburant quotidienne (caractéristique principale) moyenne parcourue durant la saison hivernale 2012 en Ontario par certains types de véhicules, à l'aide de la distance quotidienne parcourue (caractéristique auxiliaire). Comment pourrait-on utiliser l'échantillonnage à plusieurs phases pour y arriver? Expliquer les hypothèses et suppositions, et donner des intervalles de confiance à environ 95% pour les quantités d'intérets (utiliser l'ensemble de données `Autos_STR.xlsx`).

Solution: ■

69. Une biologiste cherche à estimer la taille d'une population de carouges à epaulettes dans une région. Elle en capture 500 et elle les marque avant de les remettre en liberté. Un mois plus tard, elle en recapture 50, et elle découvre qu'elle avait déjà capturé 10 de ces oiseaux. Donner un intervalle de confiance de la taille de la population de carouges à epaulettes, à environ 95%, dans la région en question.

Solution: On calcule l'estimé ponctuel à l'aide de $\hat{N} = \frac{500 \cdot 50}{10} = 2500$. De plus, $\hat{p} = \frac{X}{n_2} = \frac{10}{50} = 0.2$, d'où

$$2\sqrt{\hat{\mathrm{V}}(\hat{N})} = 2\sqrt{\frac{500^2 \cdot (1-0.2)}{50 \cdot (0.2)^3}} = 1414.214,$$

d'où

$$\mathrm{IC}(N; 0.95): \quad 2500 \pm 1414.214 \equiv (1085.786, 3914.214).$$

La taille de l'intervalle de confiance relativement élevée est-elle surprenante, étant donnés les paramètres de la question? ■

70. un EAS de $n = 800$ étudiant.e.s de plus de 18 ans est prélevé des universités de la région ($N \gg 800$). On demande à chaque répondant.e de tirer une carte au hasard (avec remise) d'un paquet typique de 52 cartes. Si la carte choisie est un valet, une dame, ou un roi, ils ou elles doivent répondre honnêtement à la question "As-tu eu des rapports sexuels consensuels avant l'âge de 18 ans?"; sinon, la question à laquelle ils ou elles doivent répondre devient "Es-tu né en janvier?". On sait que la date de naissance de 8.5% des étudiant.e.s tombe en janvier. Des 800 réponses obtenues, 175 sont des "Oui". Donner un intervalle de confiance de la proportion des étudiant.e.s ayant eu des rapports sexuels consensuels avant l'âge de 18 ans, à environ 95%, en ne tenant pas compte du FPCF.

Solution: On a $\hat{\phi} = 175/800$, $\theta = 0.085$, $\rho = 12/52 = 3/13$. Il suffit alors de calculer

$$\hat{p}_{\text{ra}} = \frac{175/800 - 0.085(1 - 3/13)}{3/13} = 0.665$$
$$\hat{\text{V}}(\hat{p}_{\text{ra}}) \approx \frac{1}{(3/13)^2} \cdot \frac{175/800(1 - 275/800)}{800 - 1} = 0.0040,$$

d'où $\text{IC}_{\text{ra}}(p; 0.95) = 0.665 \pm 2\sqrt{0.0040} \equiv (0.538, 0.792)$. ■