MAT 3777 – Échantillonnage et sondages – Exercices

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.

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.

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.

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?

Chapitre 2 – Échantillonnage aléatoire simple

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

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 $\frac{\sigma}{\sqrt{n}}$?

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 $p \approx 0.25$.

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.

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%.

   (Ne pas utiliser de fonctions provenant de librairies.)

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?

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.

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

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

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'échantillonnage	2.40	4.12
variance d'échantillonnage	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 ?

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.

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.

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 $t = a_1 y_1 + \cdots + a_n y_n$, 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.

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.]

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 $\overline{y}_{\text{STR}}$?

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.

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 ?

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.

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?

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.

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.

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.

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.

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)).

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.

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.]

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 $\alpha = \frac{1}{\mu}$, 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).

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 superfice de la base des 250 arbres est de $\tau_X = 75$ pieds carrés. Il recueille les données suivantes:

Arbre	**Superfice 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.

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 $\hat{\mu}_{Y;D}$.

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.

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.

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?

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.

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.

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.

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**	**Actuelles**	**Magasin**	**Antérieures**	**Actuelles**
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.

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.

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.

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.]

Chapitre 5 – Conception de questionnaires et collecte automatisée

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:
   i. erreur de couverture;
   ii. erreur de non-réponse;
   iii. erreur de mesure;
   iv. erreur de traitement, et
   v. 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.]

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.

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?

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.

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.]

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

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

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

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?

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 $B = 0.03$?

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. Une 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 (b)?

(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.

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.

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.

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) Décrivre 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 première. 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.

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?

Chapitre 8 – Sujets choisis

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%.

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

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

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

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

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

(b) Montrer que

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

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

(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.

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.

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.

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`).

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.

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`).

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.

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.