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 $n = 10$.

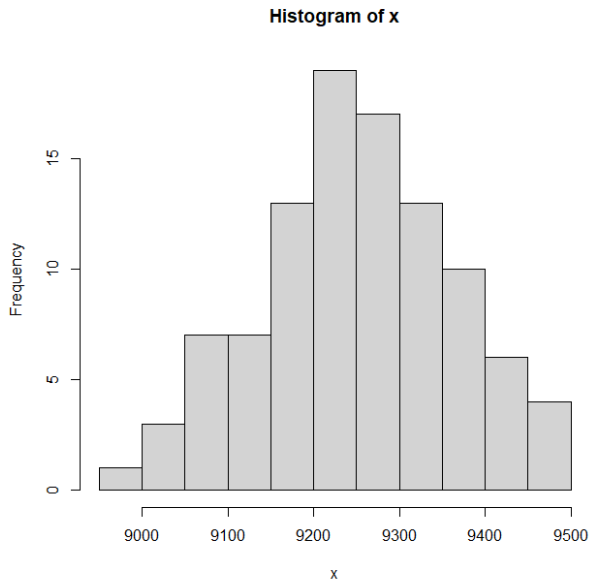
   (e) Décrire la forme de la distribution d'échantillonnage empirique. Semble-t-elle normale? Pourquoi ou pourquoi pas?

   (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:

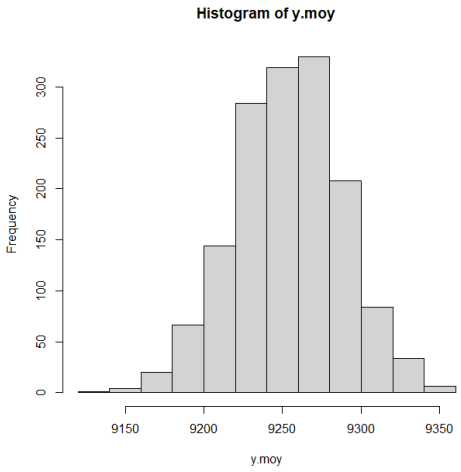
```
> n=10
> M=1500
> y=list()
> for(j in 1:M){
    y[[j]]=sample(x,n,replace=FALSE)
> }
```

(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)
```

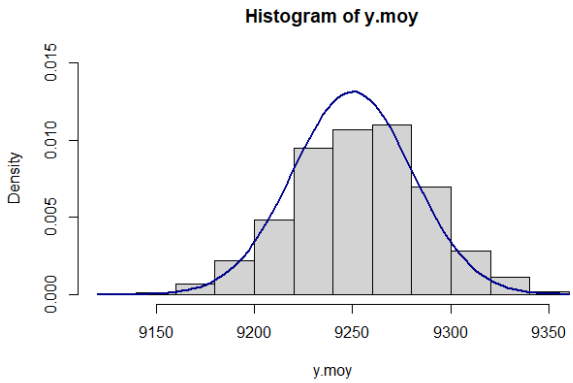


(e) Cet histogramme est à peu près symétrique autour de de 9250, et il semble certainement suivre une loi plus ou moins normale.

(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

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

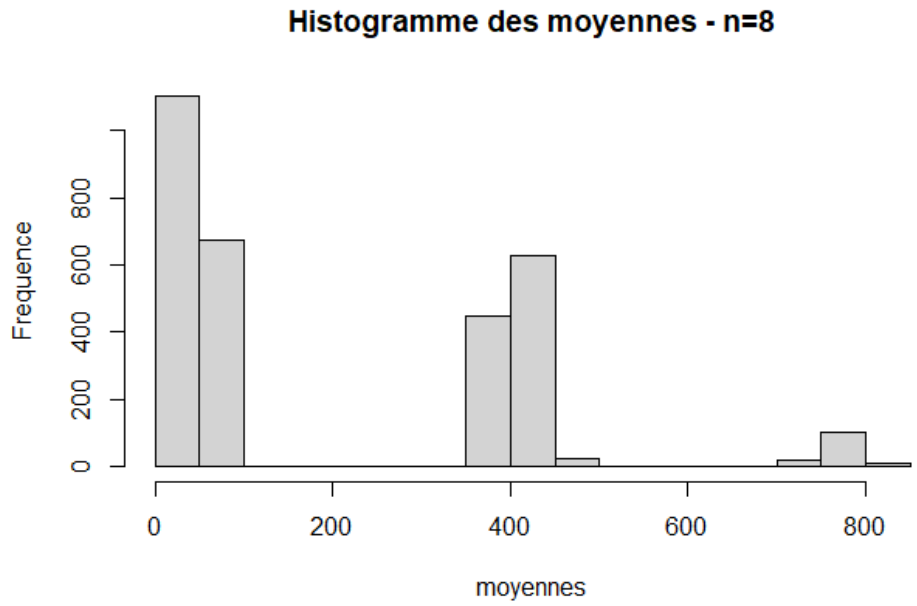
Enfin, si l'échantillon contient exactement l'un des deux plus grands nombres, la moyenne la plus élevée que nous pouvons obtenir est par l'échantillon

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

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. ■