TD2–Chapitre 2.Sondages aléatoires simples (SAS) à probabilités égales

Exercice 7. [**Source cours de P. Sarda** 2020**, d'après F. Bertrand**] Soit Y une variable définie sur une population U de taille $N = 4$ individus.

k	1	2	3	4
Y	11	10	8	11

1. Calculer la moyenne $\bar{Y}$, la variance σ_Y^2 et la variance corrigée $\tilde{\sigma}_Y^2$ de Y sur la population.
2. On tire un échantillon de taille fixe $n = 2$ sans remise, à probabilités égales.
   a) Combien d'échantillons peut-on tirer ?
   b) Calculer les valeurs de l'estimateur de la moyenne $\bar{Y}$ et celles de l'estimateur de la variance corrigée $\tilde{\sigma}_Y^2$ pour chaque échantillon.
   c) Donner la loi de probabilité de l'estimateur $\bar{Y}$.
   d) Donner la loi de probabilité de l'estimateur $\tilde{\sigma}_Y^2$.
   e) Calculer $\mathbb{E}(\bar{Y})$, $Var\left(\bar{Y}\right)$, $\mathbb{E}(\tilde{\sigma}_Y^2)$ et $Var\left(\tilde{\sigma}_Y^2\right)$.
   f) Analyser les résultats trouvés.
3. On tire un échantillon de taille $n = 2$ avec remise, en considérant l'ordre, et à probabilités égales. Répondre aux mêmes questions qu'à la partie 2.

Corrigé :

Soit Y une variable définie sur une population U de taille $N = 4$ individus :

k :	1	2	3	4
Y :	11	10	8	11

1. On a : $\bar{Y} = (11 + 10 + 8 + 11)/4 = 10, \sigma_Y^2 = \left(11^2 + 10^2 + 8^2 + 11^2\right)/4 - 10^2 = 1.5$ et $\tilde{\sigma}_Y^2 = 4 \cdot \sigma_Y^2/3 = 2$.
2. — (a) Pour $N = 4$ et $n = 2$, le nombre d'échantillons possibles sans remise est $C_4^2 = 4!/[(4-2)! \cdot 2!] = 4!/[2! \cdot 2!] = 6$
   — (b) Les échantillons en question sont listés dans le tableau ci dessous. Pour $s = \{1, 2\}$, la variable aléatoire $\hat{Y}$ prend pour valeur $\hat{y}_y = (11 + 10) = 2 = 10.5$ et la variable aléatoire $\tilde{S}_Y^2$ prends pour valeur $\tilde{s}_Y^2 = \left((11 - 10.5)^2 + (10 - 10.5)^2\right)/(2 - 1) = 0.5$.

Echantillon	$\hat{Y}$	$\tilde{S}_Y^2$
$\{1, 2\}$	10.5	0.5
$\{1, 3\}$	9.5	4.5
$\{1, 4\}$	11	0
$\{2, 3\}$	9	2
$\{2, 4\}$	10.5	0.5
$\{3, 4\}$	9.5	4.5

   — (c) Le plan de sondage est un sondage aléatoire simple à probabilités égales. Chaque échantillon a ainsi une probabilité égale à $1/6$ d'être tiré. On a : $\mathbb{P}(\hat{Y} = 9) = \mathbb{P}(S = \{2, 3\}) = 1/6, P(\hat{Y} = 9.5) = P(S = \{1, 3\}$ ou $S = \{3, 4\}) = 2 \times 1/6 = 1/3$ etc ... On regroupe la loi de $\hat{Y}$ dans le tableau suivant

$\hat{Y}(= x)$	$\mathbb{P}(\hat{Y} = x)$
9	$1/6$
9.5	$1/3$
10.5	$1/3$
11	$1/6$

   — (d) Par un raisonnement similaire on obtient

$\tilde{S}_Y^2 (= x)$	$\mathbb{P}\left(\tilde{S}_Y^2 = x\right)$
0	$1/6$
0.5	$1/3$
2	$1/6$
4.5	$1/3$

— (e) A l'aide des tableau précédent on en déduit

$$\mathbb{E}[\hat{Y}] = 9 \times 1 = 6 + 9.5 \times 1/3 + 10.5 \times 1/3 + 11 \times 1/6 = 10$$
$$\mathrm{Var}(\hat{Y}) = \left(9^2 \times 1/6 + 9.5^2 \times 1/3 + 10.5^2 \times 1/3 + 11^2 \times 1/6\right) - 10^2 = 0.5$$
$$\mathbb{E}\left(\tilde{S}_Y^2\right) = 0 \times 1/6 + 0.5 \times 1/3 + 2 \times 1/6 + 4.5 \times 1/3 = 2 \text{ et}$$
$$\mathrm{Var}\left(\tilde{S}_Y^2\right) = \left(0^2 \times 1/6 + 0.5^2 \times 1/3 + 2^2 \times 1/6 + 4.5^2 \times 1/3\right) - 2^2 = 3.5$$

— (f) On vérifie bien que $\hat{Y}$ et $\tilde{S}_Y^2$) sont des estimateurs sans biais de $\bar{Y}$ et $\tilde{S}_Y^2$. On voit par ailleurs qu'aucun des échantillons ne fournit une valeur exacte de $\bar{Y}$ et seul l'échantillon $\{2; 3\}$ donne une valeur exacte de $\tilde{S}_Y^2$.

3. Le nombre d'échantillons possibles avec remise et avec ordre est $4 \times 4 = 16$. Les probabilités de tirage étant égales la probabilité de tirer l'un quelconque des échantillons est égale à $1/16$.

Echantillon	$\hat{Y}$	$\tilde{S}_Y^2$
$(1; 1)$	11	0
$(1; 2)$	10,5	0,5
$(1; 3)$	9,5	4,5
$(1; 4)$	11	0
$(2; 1)$	10,5	0,5
$(2; 2)$	10	0
$(2; 3)$	9	2
$(2; 4)$	10,5	0,5
$(3; 1)$	9,5	4,5
$(3; 2)$	9	2
$(3; 3)$	8	0
$(3; 4)$	9,5	4,5
$(4; 1)$	11	0
$(4; 2)$	10,5	0,5
$(4; 3)$	9,5	4,5
$(4; 4)$	11	0

$\hat{Y} (= x)$	$\mathbb{P}(\hat{Y} = x)$
8	$1/16$
9	$2/16$
9,5	$4/16$
10	$1/16$
10,5	$4/16$
11	$4/16$

$\tilde{S}_Y^2 (= x)$	$\mathbb{P}\left(\tilde{S}_Y^2 = x\right)$
0	$6/16$
0,5	$4/16$
2	$2/16$
4,5	$4/16$

À l'aide de calculs identiques à ceux de la question 2, on obtient :

$$\mathbb{E}[\hat{\hat{Y}}] = 10, \quad \mathrm{Var}(\hat{\hat{Y}}) = 0.75, \quad \mathbb{E}\left[\tilde{S}_Y^2\right] = 1.5, \quad \mathrm{Var}\left(\tilde{S}_Y^2\right) = 3.375$$

Dans le cas d'un tirage sans remise, on remarque que $\mathbb{E}[\hat{\hat{Y}}]$ estime sans biais la moyenne de Y sur la population et que $\mathbb{E}\left[\tilde{S}_Y^2\right]$ estime sans biais $\tilde{\sigma}_Y^2$. L'estimateur de la moyenne est moins précis dans le cas d'un tirage avec remise (variance plus grande : $0.75 > 0.5$).

Pour l'estimateur de la variance corrigée de Y sur la population, on remarque que dans le cas d'un tirage avec remise, $\mathbb{E}\left[\tilde{S}_Y^2\right]$ est biaisé (cf. 1.5 au lieu de 2 !) ; que la variance est cependant légèrement plus faible (cf. 3.375 pour les tirages avec remise contre 3.5 pour les tirages sans remise).

Conclusion, il est préférable d'utiliser les tirages sans remise qui assurent globalement des estimateurs de meilleurs qualités.

Exercice 8. (**SAS probas égales - Intervalle de confiance pour la moyenne**) On veut estimer le salaire moyen dans les entreprises d'une ville. Sur 200 entreprises que comprend cette ville, on en tire 10 par sondage aléatoire simple. On note le salaire y_k en milliers d'euros pour l'entreprise k et on obtient sur les 10 individus sondés un total de 18.1 et une somme des carrés des valeurs égale à 36.47.

1. Déterminer l' intervalle de confiance à 95% pour la moyenne des salaires de cette ville.
2. Si l'on avait pris un niveau de confiance de 99%, l'intervalle serait-il plus étendu ou bien moins étendu ?
3. Vérifier votre réponse à la question précédente en calculant l'IC à 99%.

Corrigé :

Nous avons $N = 200, n = 10, \sum_{k\in U} y_k = 18.1$ et $\sum_{k\in U} y_k^2 = 36.47$.

1. La formule générale de l'estimation d'un IC pour estimer un paramètre θ est

$$\hat{IC}_{1-\alpha}(\theta) = \left[\hat{\theta} \pm z_{1-\alpha/2} \cdot \sqrt{\hat{\mathrm{Var}}(\hat{\theta})}\right]$$

avec $z_{1-\alpha/2}$ le quantile de la loi normale centrée réduite. Cela nous donne pour $\theta = \mu = \bar{Y}$, la moyenne sur la population, avec

$$\hat{\mu} = \sum_{k\in U} y_k/N = \hat{t_Y}/N, \quad \hat{\mathrm{Var}}(\hat{\mu}) = \left(1 - \frac{n}{N}\right)\frac{\tilde{s}_Y^2}{n}$$

et $\tilde{s}_Y^2$ est la variance corrigée de Y sur l'échantillon S, i.e. $\tilde{s}_Y^2 = \frac{1}{n-1}\sum_{k\in S}(y_k - \bar{y}_S)^2$ et $\bar{y}_S$ la moyenne sur l'échantillon S.

Comme l'énoncé nous donne la sommes des carrées, il est plus judicieux d'utiliser la formule

$$s_Y^2 = \frac{1}{n}\sum_{k\in S}(y_k - \bar{y}_S)^2 = \frac{1}{n}\sum_{k\in S}\left(y_k^2\right) - (\bar{y}_S)^2 = \frac{1}{n}\sum_{k\in S}\left(y_k^2\right) - \left(\frac{1}{n}\sum_{k\in S} y_k\right)^2$$

Application numérique :

$$\bar{y}_S = 18.1/10 = 1.81; s_Y^2 = 36.47/10 - (1.81)^2 = 0.3709 \text{ et } \tilde{s}_Y^2 = 10 \cdot 0.3709/9 = 0.412111$$

Cela nous amène à $\hat{\mathrm{Var}}(\hat{\mu}) = \left(1 - \frac{10}{200}\right)\frac{0.3709}{10} = 0.03915055$. Pour $\alpha = 0.05$, $z_{1-\alpha/2} \simeq 1.96$

$$\hat{IC}_{95\%}(\hat{\mu}) = [1.81 \pm 1.96 \times \sqrt{0.0352355}] = [1.42; 2.19].$$

Remarque : L'étendue de cet intervalle sans être très grande peut-être améliorée. En effet la taille de l'échantillon est assez petite et pourrait-être augmenté.

2. Si l'on augmente le niveau de confiance, c'est pour s'assurer de prendre plus en compte de valeurs possibles. L'étendue de l'intervalle va donc augmenter et l'estimation plus confiante sera donc plus imprécise.
3. Pour $\alpha = 0.01$, $z_{1-\alpha/2} \simeq 2.56$ et $\hat{IC}_{99\%}(\hat{\mu}) = [0.25; 3.37]$. L'étendue passe de $(3 - 0.61) = 2.39$ à $(3.37 - 0.25) = 3.12$.

Exercice 9. (**SAS probas égales - Proportion - Choix de la taille d'échantillon**) On souhaite estimer la proportion P d'individus ayant été atteints par une maladie professionnelle dans une entreprise de 1500 salariés. On souhaite effectuer un sondage et on se pose la question du choix de la taille d'échantillon afin d'avoir une estimation assez précise avec un intervalle de confiance au niveau 95% ayant une étendue de 0.02.

1. Cas 1 : nous avons une estimation historique de cette proportion dans des entreprises du même type, de 3 salariés sur 10.
2. Cas 2 : que faire lorsque nous n'avons pas d'estimation ponctuelle pour la proportion ou une valeur à laquelle nous référer.

Corrigé :

Dans le cas d'un travail sur une proportion P, nous avons les formules suivantes vues en cours

$$I\hat{C}_\alpha(\hat{P}) = \left[\hat{P} \pm z_{1-\alpha/2} \cdot \sqrt{\hat{\mathrm{Var}}(\hat{P})}\right]$$

avec

$$\hat{\mathrm{Var}}(\hat{P}) = (1-f) \cdot \frac{\hat{P}(1-\hat{P})}{(n-1)}$$

1. Cas 1 : $\hat{P} = 3/10$.
   La formule précédente nous donne pour une demie étendue e donnée où

   $$e = z_{1-\alpha/2} \cdot \sqrt{\hat{\mathrm{Var}}(\hat{P})}$$

   Cela donne

   $$n = \frac{e^2 + z^2_{1-\alpha/2} \cdot \hat{P}(1-\hat{P})}{e^2 + z^2_{1-\alpha/2} \cdot \hat{P}(1-\hat{P})/N}$$

   Application numérique : $N = 1500, \hat{P} = 3/10, e = 0.02, z_{1-\alpha/2} = 1.96, n \geqslant 860.21$. On prendra donc $n = 861$.
2. Cas 2 : dans ce cas on va effectuer une étude de fonction. Voici ci-dessous un petit programme R.

```
# 1 N = 1500
N <- 1500

# 2 Pchap = 3/10
Pchap <- 3/10

# 3 e.bis = 0.02
e.bis <- 0.02

# 4 Calcul de l'expression mathematique
result <- (e.bis^2 + z^2 * Pchap * (1 - Pchap)) /
          (e.bis^2 + (z^2 * Pchap * (1 - Pchap) / N))

# Cas ou l'on n'a pas  d estimation  de pchap ?!
# Etude de fonction

# Creer une sequence de 0.05    0.95 avec un pas de 0.01
Pchap.seq <- seq(0.05, 0.95, 0.01)
```

```
# Calcul des n pour chaque valeur de Pchap
n.seq <- (e.bis^2 + z^2 * Pchap.seq * (1 - Pchap.seq)) /
         (e.bis^2 + (z^2 * Pchap.seq * (1 - Pchap.seq) / N))

# Creation d'une figure
plot(Pchap.seq, n.seq, type = 'l',
     main = 'Sample size choice in function of a given p estimation',
     xlab = 'p estimation', ylab = 'Sample size necessary',
     cex.lab = 1.5, cex.main = 1.5)

# Illustration de l'exemple precedent pour une estimation de p = 0.3
segments(0.3, 860.63, x1 = 0, y1 = 860.63, lwd = 3)
segments(0.3, 0, x1 = 0.3, y1 = 860.63, lwd = 3)
text(0.45, 860, 'First case for p = 0.3, n = 860.63', cex = 1.5)
```

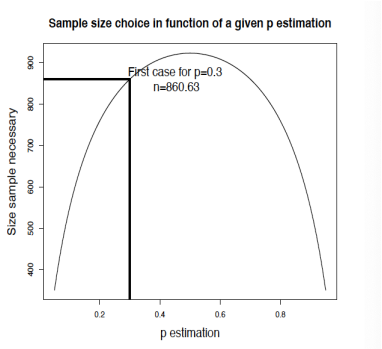

FIGURE 1 – Etude de fonction pour le choix de la taille d'échantillon.

Exercice 10. (**SAS probas égales - Application de l'algorithme sélection-rejet**) On considère une population de taille 8. On souhaite sonder par un plan aléatoire simple sans remise, à probabilités égales, 2 individus statistiques. Voici ci-dessous 8 valeurs aléatoires d'une loi uniforme.

- [1] 0.6062683 0.9376420 0.2643521 0.3800939
- [5] 0.8074834 0.9780757 0.9579337 0.7627319

Déterminer par application de la méthode de sélection - rejet le plan choisi.

Corrigé :

Nous avons ici $N = 8$ et $n = 2$. Rappelons les valeurs aléatoires proposées :

- [1] 0.6062683 0.9376420 0.2643521 0.3800939
- [5] 0.8074834 0.9780757 0.9579337 0.7627319

Rappelons que k correspond à l'indice des individus étudiés et j au nombre de sélections.

1. Etape 1 de "tant que $j < n$ " : $k = 0, j = 0, \quad u = 0.6062683$
   - $(n - j)/(N - k) = (2 - 0)/(8 - 0) = 0.25$
   - u ne vérifie pas l'inégalité $u < (n - j)/(N - k)$
   - On pose donc $k = k + 1$ soit $k = 1$
2. Etape 2 de "tant que $j < n$ " : $k = 1, j = 0, \quad u = 0.9376420$
   - $(n - j)/(N - k) = (2 - 0)/(8 - 1) = 0.286$
   - u ne vérifie pas l'inégalité $u < (n - j)/(N - k)$
   - On pose donc $k = k + 1$ soit $k = 2$

3. Etape 3 de "tant que $j < n$ " : $k = 2, j = 0, \quad u = 0.3800939$
   — $(n - j)/(N - k) = (2 - 1)/(8 - 2) = 0.333$
   — u vérifie l'inégalité $u < (n - j)/(N - k)$.
   — On sélectionne donc l'individu $k + 1 = 3$. On pose $j = j + 1 = 1$ et $k = k + 1 = 3$
4. Etape 4 de "tant que $j < n$ " : $k = 3, j = 1, \quad u = 0.2643521$
   — $(n - j)/(N - k) = (2 - 1)/(8 - 3) = 0.2$
   — u ne vérifie pas l'inégalité $u < (n - j)/(N - k)$.
   — On pose donc $k = k + 1 = 4$.
5. Etape 5 de "tant que $j < n$ " : $k = 4, j = 1, \quad u = 0.8074834$
   — $(n - j)/(N - k) = (2 - 1)/(8 - 4) = 0.25$
   — u ne vérifie pas l'inégalité $u < (n - j)/(N - k)$.
   — On pose donc $k = k + 1 = 5$.
6. Etape 6 de "tant que $j < n$ " : $k = 5, j = 1, \quad u = 0.9780757$
   — $(n - j)/(N - k) = (2 - 1)/(8 - 5) = 0.333$
   — u ne vérifie pas l'inégalité $u < (n - j)/(N - k)$.
   — On pose donc $k = k + 1 = 6$.
7. Etape 7 de "tant que $j < n$ " : $k = 6, j = 1, \quad u = 0.9579337$
   — $(n - j)/(N - k) = (2 - 1)/(8 - 6) = 0.5$
   — u ne vérifie pas l'inégalité $u < (n - j)/(N - k)$.
   — On pose donc $k = k + 1 = 7$.
8. Etape 8 de "tant que $j < n$ " : $k = 7, j = 1, uadu = 0.7627319$
   — $(n - j)/(N - k) = (2 - 1)/(8 - 7) = 1$
   — u vérifie l'inégalité $u < (n - j)/(N - k)$.
   — On sélectionne donc l'individu $k + 1 = 8$. On pose $j = j + 1 = 2$ et $k = k + 1 = 8$.
9. On a $j = 2 \geqslant n = 2$. Fin de la boucle "tant que $j < n$ " : On voit qu'a fortiori, quelque soit la valeur de u, l'étape concernant le dernier élément amènera sa sélection car on a toujours $u < 1$.