T. D. n° 1
Les méthodes d'échantillonnage Correction

Exercice 1. Correction

1. La distribution de Y est :

$Y=k$	$\mathbb{P}[Y=k]$	f_k	$k \times \mathbb{P}[Y=k]$	$(k-\mu)^2 \times \mathbb{P}[Y=k]$
8	1/4	25%	$8 \times 1/4 = 8/4$	$(8-10)^2 \times 1/4 = 4/4$
10	1/4	25%	$10 \times 1/4 = 10/4$	$(10-10)^2 \times 1/4 = 0$
11	2/4	50%	$11 \times 2/4 = 22/4$	$(11-10)^2 \times 2/4 = 2/4$
Total	1	100%	$\mu = \mathbb{E}[Y] = 10$	$\mathrm{Var}[Y] = 4/4 + 2/4 = 6/4 = 1,5$

D'une part, nous avons d'après les calculs

$$\mathrm{Var}[Y] = \sigma^2 = 1,5$$

et d'autre part, nous avons d'après la définition du cours

$$\sigma_c^2 = \frac{N}{N-1}\sigma^2 = \frac{4}{3} \times \frac{6}{4} = 2.$$

2. Échantillonnage sans remise.

   a) Nous pouvons tirer : $4 \times 3 = 12$ échantillons.

   b) Nous remplissons le tableau ci-dessous en utilisant la définition du cours

$$S_{2,c}^2(obs) = s_{2,c}^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}(Y_i(obs) - \widehat{\mu}_2(obs))^2\right].$$

Échantillon	$Y_1(obs)$	$Y_2(obs)$	$\widehat{\mu}_2(obs)$	$S_{2,c}^2(obs)$
$(1,2)$	11	10	$10,5$	$[(11-10,5)^2 + (10-10,5)^2] = 0,5$
$(1,3)$	11	8	$9,5$	$[(11-9,5)^2 + (8-9,5)^2] = 4,5$
$(1,4)$	11	11	$11,0$	$[(11-11)^2 + (11-11)^2] = 0,0$
$(2,1)$	10	11	$10,5$	$[(10-10,5)^2 + (11-10,5)^2] = 0,5$
$(2,3)$	10	8	$9,0$	$[(10-9)^2 + (8-9)^2] = 2,0$
$(2,4)$	10	11	$10,5$	$[(10-10,5)^2 + (11-10,5)^2] = 0,5$
$(3,1)$	8	11	$9,5$	$[(8-9,5)^2 + (11-9,5)^2] = 4,5$
$(3,2)$	8	10	$9,0$	$[(8-9)^2 + (10-9)^2] = 2,0$
$(3,4)$	8	11	$9,5$	$[(8-9,5)^2 + (11-9,5)^2] = 4,5$
$(4,1)$	11	11	$11,0$	$[(11-11)^2 + (11-11)^2] = 0,0$
$(4,2)$	11	10	$10,5$	$[(11-10,5)^2 + (10-10,5)^2] = 0,5$
$(4,3)$	11	8	$9,5$	$[(11-9,5)^2 + (8-9,5)^2] = 4,5$

Ici la taille de l'échantillon est égale à $n = 2$.

Exemple : Pour l'échantillon $(1,3)$:

$$S^2_{2,c,obs} = \frac{1}{2-1}\left[(11-9,5)^2 + (8-9,5)^2\right] = 4,5.$$

c) La distribution de $\widehat{\mu}_2$ est :

$\widehat{\mu}_2 = x$	$\mathbb{P}[\widehat{\mu}_2 = x]$	$x \times \mathbb{P}[\widehat{\mu}_2 = x]$	$\mathbb{P}[\widehat{\mu}_2 = x] \times (x - \mathbb{E}(\widehat{\mu}_2))^2$
$9,0$	$2/12$	$9,0 \times 2/12 = 18/12$	$2/12 \times (9-10)^2 = 2/12$
$9,5$	$4/12$	$9,5 \times 4/12 = 38/12$	$4/12 \times (9,5-10)^2 = 1/12$
$10,5$	$4/12$	$10,5 \times 4/12 = 42/12$	$4/12 \times (10,5-10)^2 = 1/12$
$11,0$	$2/12$	$11,0 \times 2/12 = 22/12$	$2/12 \times (11-10)^2 = 2/12$
Total	1	$\mathbb{E}[\,\widehat{\mu}_2\,] = 120/12 = 10$	$\mathrm{Var}[\,\widehat{\mu}_2\,] = 6/12 = 1/2 = 0,5$

Nous remarquons que l'estimateur $\widehat{\mu}_2$ est bien sans biais puisque son espérance est égale à 10, qui est la moyenne de la population U. D'autre part, ces calculs nous permettent de vérifier la formule du calcul de la variance de l'estimateur $\widehat{\mu}_2$ dans le cas d'un tirage sans remise. En effet, nous avons :

$$\mathrm{Var}\left[\widehat{\mu}_n\right] = \frac{N-n}{N-1} \times \frac{\sigma^2}{n}.$$

Ici $n = 2$, $N = 4$ et $\sigma^2 = \frac{3}{2}$, par conséquent, nous avons :

$$\mathrm{Var}\left[\widehat{\mu}_2\right] = \frac{4-2}{4-1} \times \frac{3/2}{2} = \frac{1}{2}.$$

d) La distribution de $S^2_{2,c}$ est :

$S^2_{2,c} = x$	$\mathbb{P}[S^2_{2,c} = x]$	$x \times \mathbb{P}[S^2_{2,c} = x]$	$\mathbb{P}[S^2_{2,c} = x] \times (x - \mathbb{E}\left(S^2_{2,c}\right))^2$
$0,0$	$2/12$	$0 \times 2/12 = 0$	$2/12 \times (0-2)^2$
$0,5$	$4/12$	$0,5 \times 4/12 = 2/12$	$4/12 \times (0,5-2)^2$
$2,0$	$2/12$	$2 \times 2/12 = 4/12$	$2/12 \times (2-2)^2$
$4,5$	$4/12$	$4,5 \times 4/12 = 18/12$	$4/12 \times (4,5-2)^2$
Total	1	$\mathbb{E}\left[S^2_{2,c}\right] = 24/12 = 2$	$\mathrm{Var}[\,S^2_{2,c}\,] = 42/12 = 7/2 = 3,5$

e) Voir les réponses c) et d).

f) **Commentaires :** Dans le cas SANS REMISE, aucun échantillon ne fournit la vraie valeur $\mu = 10$ mais en moyenne les échantillons donnent 10. En effet, nous avons

$$\mathbb{E}\left[\widehat{\mu}_2\right] = 10 = \mu.$$

Nous en déduisons de cette égalité que $\widehat{\mu}_2$ est un estimateur sans biais de μ dans le cas d'un tirage sans remise. L'échantillonnage est non biaisé. De plus,

$$\mathbb{E}\left[S^2_{2,c}\right] = 2 = \sigma^2_c$$

et non pas σ^2. Nous en déduisons de cette égalité que $S^2_{n,c}$ est un estimateur sans biais de σ^2_c dans le cas d'un tirage sans remise et ce quelque soit la taille n de l'échantillon.

3. Échantillonnage avec remise.

   a) Nous pouvons tirer : $4 \times 4 = 16$ échantillons.

   b)

Échantillon	$Y_1(obs)$	$Y_2(obs)$	$\widehat{\mu}_2$	$S^2_{2,c}(obs)$
$(1,1)$	11	11	$11,0$	$[(11-11)^2+(11-11)^2]=0,0$
$(1,2)$	11	10	$10,5$	$[(11-10,5)^2+(10-10,5)^2]=0,5$
$(1,3)$	11	8	$9,5$	$[(11-9,5)^2+(8-9,5)^2]=4,5$
$(1,4)$	11	11	$11,0$	$[(11-11)^2+(11-11)^2]=0,0$
$(2,1)$	10	11	$10,5$	$[(10-10,5)^2+(11-10,5)^2]=0,5$
$(2,2)$	10	10	$10,0$	$[(10-10)^2+(10-10)^2]=0,0$
$(2,3)$	10	8	$9,0$	$[(10-9)^2+(8-9)^2]=2,0$
$(2,4)$	10	11	$10,5$	$[(10-10,5)^2+(11-10,5)^2]=0,5$
$(3,1)$	8	11	$9,5$	$[(8-9,5)^2+(11-9,5)^2]=4,5$
$(3,2)$	8	10	$9,0$	$[(8-9)^2+(10-9)^2]=2,0$
$(3,3)$	8	8	$8,0$	$[(8-8)^2+(8-8)^2]=0,0$
$(3,4)$	8	11	$9,5$	$[(8-9,5)^2+(11-9,5)^2]=4,5$
$(4,1)$	11	11	$11,0$	$[(11-11)^2+(11-11)^2]=0,0$
$(4,2)$	11	10	$10,5$	$[(11-10,5)^2+(10-10,5)^2]=0,5$
$(4,3)$	11	8	$9,5$	$[(11-9,5)^2+(8-9,5)^2]=4,5$
$(4,4)$	11	11	$11,0$	$[(11-11)^2+(11-11)^2]=0,0$

en utilisant la formule du cours

$$S^2_{n,c}(obs) = \frac{1}{n-1}\left[\sum_{i=1}^{n}(Y_i(obs)-\widehat{\mu}_n(obs))^2\right].$$

Ici la taille de l'échantillon est égale à $n = 2$.

Exemple : Pour l'échantillon $(1,3)$:

$$S^2_{2,c}(obs) = \frac{1}{2-1}\left[(11-9,5)^2+(8-9,5)^2\right] = 4,5.$$

   c) La distribution de $\widehat{\mu}_2$ est :

$\widehat{\mu}_2 = x$	$\mathbb{P}[\widehat{\mu}_2 = x]$	$x \times \mathbb{P}[\widehat{\mu}_2 = x]$	$\mathbb{P}[\widehat{\mu}_2 = x] \times (x - \mathbb{E}(\widehat{\mu}_2))^2$
$8,0$	$1/16$	$8 \times 1/16 = 8/16$	$1/16 \times (8-10)^2 = 4/16$
$9,0$	$2/16$	$9 \times 2/16 = 18/16$	$2/16 \times (9-10)^2 = 2/16$
$9,5$	$4/16$	$9,5 \times 4/16 = 38/16$	$4/16 \times (9,5-10)^2 = 1/16$
$10,0$	$1/16$	$10 \times 1/16 = 10/16$	$1/16 \times (10-10)^2 = 0$
$10,5$	$4/16$	$10,5 \times 4/16 = 42/16$	$4/16 \times (10,5-10)^2 = 1/16$
$11,0$	$4/16$	$11 \times 4/16 = 44/16$	$4/16 \times (11-10)^2 = 4/16$
Total	1	$\mathbb{E}[\widehat{\mu}_2] = 160/16 = 10$	$\mathrm{Var}[\widehat{\mu}_2] = 12/16 = 0,75$

d) La distribution de $S^2_{2,c}$ est :

$S^2_{2,c} = x$	$\mathbb{P}[S^2_{2,c} = x]$	$x \times \mathbb{P}[S^2_{2,c} = x]$	$\mathbb{P}[S^2_{2,c} = x] \times (x - \mathbb{E}\left(S^2_{2,c}\right))^2$
$0,0$	$6/16$	$0 \times 6/16 = 0$	$6/16 \times (0 - 1,5)^2$
$0,5$	$4/16$	$0,5 \times 4/16 = 2/16$	$4/16 \times (0,5 - 1,5)^2$
$2,0$	$2/16$	$2 \times 2/16 = 4/16$	$2/16 \times (2 - 1,5)^2$
$4,5$	$4/16$	$4,5 \times 4/16 = 18/16$	$4/16 \times (4,5 - 1,5)^2$
Total	1	$\mathbb{E}\left[S^2_{2,c}\right] = 24/16 = 1,5$	$\mathrm{Var}\left[S^2_{2,c}\right] = 54/16 = 3,375$

e) Voir les réponses c) et d).

f) **Commentaires :** Dans le cas AVEC REMISE, un échantillon fournit la vraie valeur $\mu = 10$ et en moyenne les échantillons donnent 10. En effet, nous avons

$$\mathbb{E}\left[\widehat{\mu}_2\right] = 10 = \mu.$$

Nous en déduisons de cette égalité que $\widehat{\mu}_2$ est un estimateur sans biais de μ. L'échantillonnage est non biaisé. De plus, nous remarquons

$$\mathbb{E}[S^2_{2,c}] = 1,5 = \sigma^2$$

et non pas σ^2_c, cette fois-ci. Nous en déduisons de cette égalité que $S^2_{n,c}$ est un estimateur sans biais de σ^2 et ce quelque soit la taille n d'un échantillon dans le cas d'un tirage AVEC REMISE.

D'autre part, nous remarquons que la variance de $\widehat{\mu}_n$ dans le cas d'un tirage AVEC REMISE est plus grande que celle dans le cas d'un tirage SANS REMISE et c'est toujours le cas. Cela veut dire que la distribution des moyennes observées sur les échantillons est plus dispersée AVEC REMISE que SANS REMISE : le plan de sondage est moins précis.

En effet, d'avoir introduit les échantillons $(1,1), ..., (4,4)$ n'améliore pas le plan de sondage car en interrogeant deux fois la même personne nous n'apportons aucune information supplémentaire.

. . . .

Exercice 2. Correction Nous avons un sondage aléatoire simple à probabilités égales sans remise.

1. *Premier cas : l'échantillon est de taille $n = 4$.* Une estimation de la dépense moyenne est égale à :

$$\widehat{\mu}_4 = 12 \text{ euros.}$$

En posant $f = \dfrac{n}{N}$, la précision de cette estimation vaut :

$$\mathrm{Var}\left[\,\widehat{\mu}_4\,\right] = (1 - f)\frac{\sigma^2_c}{n}\cdot$$

Or nous ne connaissons pas σ^2_c qui représente la variance corrigée de la population. Donc il faut introduire un estimateur sans biais de σ^2_c qui est S^2_c.

En effet, nous avons démontré ce résultat dans l'exercice précédent. D'où l'estimateur de la variance de l'estimateur $\widehat{\mu}_4$ est égal à :

$$\widehat{\mathrm{Var}\,[\,\widehat{\mu}_4\,]} = (1-f)\frac{S_c^2}{n}$$

et l'estimation est égale à :

$$\widehat{\mathrm{Var}\,[\,\widehat{\mu}_4\,]} = (1-0,05)\times\frac{4}{3}\times\frac{6^2}{4} = 11,4 \text{ euros}^2.$$

Donc la précision, à 95%, est de :

$$\pm 1,96\sqrt{11,4} = 6,62 \text{ euros.}$$

Donc, nous pouvons écrire :

$$\widehat{\mu}_4 = 12 \pm 6,62 \text{ euros.}$$

Second cas : l'échantillon est de taille $n = 40$. Une estimation de la dépense moyenne est égale à :

$$\widehat{\mu}_{40} = 12 \text{ euros.}$$

En posant $f = \dfrac{n}{N}$, la précision de cette estimation vaut :

$$\mathrm{Var}\,[\,\widehat{\mu}\,] = (1-f)\frac{S_c^2}{n}\cdot$$

Or nous ne connaissons pas S_c^2 qui représente la variance corrigée de la population. Donc il faut introduire un estimateur sans biais de S_c^2 qui est s_c^2 (Nous avons démontré ce résultat dans l'exercice précédent). D'où l'estimateur de la variance de l'estimateur $\widehat{\mu}_{40}$ est égal à :

$$\widehat{\mathrm{Var}\,[\,\widehat{\mu}_{40}\,]} = (1-f)\frac{s_c^2}{n}$$

et l'estimation est égale à :

$$\widehat{\mathrm{Var}\,[\,\widehat{\mu}_{40}\,]} = (1-0,5)\times\frac{40}{39}\times\frac{6^2}{40} = 0,46 \text{ euros}^2.$$

Donc la précision est de :

$$\pm 1,96\sqrt{0,46} = 1,33 \text{ euros.}$$

Donc nous pouvons écrire :

$$\widehat{\mu}_{40} = 12 \pm 1,33 \text{ euros.}$$

2. Nous avons un sondage aléatoire simple à probabilités égales SANS REMISE. Nous estimons π par $\widehat{\pi} = 0,75$. La variance de l'estimateur $\widehat{\pi}$ est égale à

$$\mathrm{Var}\,[\widehat{\pi}] = \frac{N-n}{N-1}\times\frac{\pi(1-\pi)}{n}$$

car l'estimateur $\widehat{\pi}$ suit une loi hypergéométrique. Or cette formule pose un problème puisque nous ne connaissons pas π. Donc nous sommes amenés à construire un estimateur de cette variance. L'estimateur de la variance de l'estimateur $\widehat{\pi}$ est donc égal à :

$$\widehat{\mathrm{Var}[\widehat{\pi}]} = \frac{N-n}{N}\times\frac{\widehat{\pi}(1-\widehat{\pi})}{n-1} = (1-f)\times\frac{\widehat{\pi}(1-\widehat{\pi})}{n-1}\cdot$$

Premier cas : l'échantillon est de taille $n = 4$. Nous avons :

$$\widehat{\mathrm{Var}}[\widehat{\pi}] = (1 - 0,05)\frac{0,1875}{4-1} = 0,059375.$$

La précision est de :

$$\pm 1,96\sqrt{0,059375} \simeq 0,48.$$

Donc nous pouvons écrire :

$$\pi \in [0,27; 1],$$

nous ne pouvons pas dépasser la valeur 1 puisque nous sommes entrain d'estimer une proportion !

Second cas : l'échantillon est de taille $n = 40$. Nous avons :

$$\widehat{\mathrm{Var}}[\widehat{\pi}] = (1 - 0,5)\frac{0,1875}{40-1} = 0,004567308$$

La précision est de :

$$\pm 1,96\sqrt{0,004567308} \simeq 0,13.$$

Donc nous pouvons écrire :

$$\pi \in [0,61; 0,89].$$

3. L'échantillon de taille $n = 4$ introduit des intervalles de confiance très grands, ce qui était prévisible. De plus l'approximation normale dans ce cas n'est peut-être pas légitime. À ce sujet, pour légitimer l'approximation par une loi normale, nous renvoyons au livre « Les techniques de sondage » de Pascal Ardilly, aux éditions Technip, et plus particulièrement aux pages 60 et 61, qui évoque ce problème.

....

Exercice 3. Vérification dans un cas simple

1. Nous avons 20 échantillons possibles, énumérés ci-dessous :

(Y_i, Y_j)	$\overline{y}$	s_c^2	(Y_i, Y_j)	$\overline{y}$	s_c^2
Y_1, Y_2	90	200	Y_2, Y_1	90	200
Y_1, Y_3	100	0	Y_3, Y_1	100	0
Y_1, Y_4	110	200	Y_4, Y_1	110	200
Y_1, Y_5	95	85	Y_5, Y_1	95	85
Y_2, Y_3	90	200	Y_3, Y_2	90	200
Y_2, Y_4	100	800	Y_4, Y_2	100	800
Y_2, Y_5	85	50	Y_5, Y_2	85	50
Y_3, Y_4	110	200	Y_4, Y_3	110	200
Y_3, Y_5	95	52	Y_5, Y_3	95	52
Y_4, Y_5	105	450	Y_5, Y_4	105	450

2. La moyenne dans la population est égale :

$$\mu = \frac{1}{5}(100 + 80 + 100 + 120 + 90) = 98.$$

Chaque échantillon ayant une probabilité de 1/20 d'être tiré, nous avons :

$$\begin{aligned}\mathbb{E}\left[\,\widehat{\mu}_2\,\right] &= \frac{1}{20}(90+100+110+95+90+100+85+110+95+105+90+100+\\ &\qquad 110+95+90+100+85+110+95+105) = 98.\end{aligned}$$

Nous retrouvons le résultat du cours à savoir que :

$$\mathbb{E}\left[\,\widehat{\mu}_2\,\right] = \mu.$$

Donc $\widehat{\mu}_2$ est un estimateur sans biais de la moyenne μ de la population U.

3. La variance de l'estimateur $\widehat{\mu}_2$ peut être calculée, d'une part, directement en utilisant la définition de la variance puisque nous avons la distribution de l'estimateur $\widehat{\mu}_2$:

$$\mathrm{Var}\left[\,\widehat{\mu}_2\,\right] = \frac{1}{20}[(90-98)^2 + (100-98)^2 + \cdots + (105-98)^2] = 66.$$

D'autre part, en utilisant la formule du cours, dans le cadre d'un tirage sans remise, nous avons :

$$\mathrm{Var}\left[\,\widehat{\mu}_2\,\right] = (1-\frac{n}{N})\frac{S_c^2}{n} = 66.$$

4.

$$\mathbb{E}\left[\,s_c^2\,\right] = \frac{1}{20}[200+0+200+\cdots+450] = 220.$$

Nous retrouvons bien :

$$\mathbb{E}\left[\,s_c^2\,\right] = S_c^2.$$

Donc s_c^2 est un estimateur sans biais de la variance corrigée S_c^2 de la population U.

. . . .