

Projet Techniques de Sondage  
Master 1 de Sciences de Données et Applications

**EXERCICE 1**

Soit la population  $\{1, 2, 3\}$  et le plan de probabilité suivante :

$$P(\{1,2\}) = 1/2, P(\{1,3\}) = 1/4, P(\{2,3\}) = 1/4$$

1- Est-ce un sondage aléatoire simple ?

Non, il ne s'agit pas d'un sondage aléatoire simple.

En effet, on a

$$P(\{1,2\}) = 1/2, P(\{1,3\}) = 1/4, P(\{2,3\}) = 1/4$$

2- Calculons les probabilités d'inclusions d'ordre 1 :

$$\pi_1 = P(\{1,2\}) + P(\{1,3\}) = 1/2 + 1/4 = 3/4$$

$$\pi_2 = P(\{1,2\}) + P(\{2,3\}) = 1/2 + 1/4 = 3/4$$

$$\pi_3 = P(\{1,3\}) + P(\{2,3\}) = 1/4 + 1/4 = 1/2$$

3- Calculons les probabilités d'inclusions d'ordre 2 :

$$\pi_{kl} = \sum_{s \in kl} p(s)$$

$$\pi_{12} = 1/16 + (3/4 * 3/4) = 8/16 = 1/2$$

$$\pi_{23} = 1/8 + (3/4 * 1/2) = 1/4$$

$$\pi_{13} = 1/8 + (3/4 * 1/2) = 1/4$$

4- L'estimateur  $\bar{Y}$  si les échantillons sont tirés :

On suppose que le  $\pi$ -estimateur est celui de **M. Thomson**

$$\text{Si } \{1,2\} \text{ est tiré : } 1/3 (y_1 + y_2/3/4) = 8/9$$

$$\text{Si } \{1,3\} \text{ est tiré : } 1/3 (y_1/3/4 + y_3/1/2) = 10/9$$

$$\text{Si } \{2,3\} \text{ est tiré : } 1/3 (y_2/3/4 + y_3/1/2) = 10/9$$

5- Vérifions que l'estimateur est sans biais :

$$E(\bar{Y}) - Y = 0$$

$$\text{Or } E(\bar{Y}) = 1/2 * (4(y_1 + y_2)/9) + 1/4 * ((4y_1 + 6y_3)/9) + 1/4 * ((4y_2 + 6y_3)/9)$$

$$E(\bar{Y}) = 3y_1 + 3y_2 + 3y_3/9$$

$$E(\bar{Y}) = y_1 + y_2 + y_3/3 = Y$$

*L'estimateur  $\bar{Y}$  est sans biais*

6- Ecrire ce que seraient les probabilités d'échantillons P et les probabilités d'inclusion d'un sondage aléatoire simple sans remise :

Pour un sondage aléatoire simple sans remise a probabilité égale.

La loi de probabilité suit une loi telle que :

$$P = 1 / C_n^N$$

$$P = 1 / C_{23}^3 = 1 / 3 \text{ (EQUIPROBABILITE)}$$

LE NOMBRE D' ECHANTILLONS POSSIBLE EST DE :  $C_{23}^3 = 3$

$$\text{Donc } P(\{1,2\}) = P(\{1,3\}) = P(\{2,3\}) = 1 / 3$$

Ces probabilités d'inclusions sont définies par la relation  $n/N = 2 / 3$ .

### **EXERCICE 2 :**

On s'intéresse à la proportion d'hommes atteints par une maladie professionnelle dans une entreprise de 1500 travailleurs. On sait par ailleurs que trois travailleurs sur dix sont ordinairement touchés par cette maladie dans des entreprises du même type. On se propose de sélectionner un échantillon au moyen d'un sondage aléatoire simple.

1-La taille de l'échantillon pour les plans simples avec ou sans remise:  
paramètres d'intérêt :

$y_i$  représente la présence ou non de la maladie  
estimation du paramètre

$$\text{Variance de l'estimateur : } \text{Var}(\bar{p}) = \begin{cases} P(1-P) / n & \text{avec remise} \\ (N-n) / (n-1) * P(1-P) / n & \text{sans remise} \end{cases}$$

***P : proportion d'hommes atteints***

***Puisque  $y_i^2 = y_i$***

***Variance de la population :***

***Intervalle de confiance à 95% :***

$$\bar{p} \pm 1.96 * \sqrt{\text{Var}(\bar{p})}$$

***Taille de l'échantillon n telle que la longueur totale de l'intervalle de confiance***

$$(2 * 1.96 * \sqrt{\text{Var}(\bar{p})}) < 0.02 \quad \text{donc } \text{Var}(\bar{p}) < 0.00051$$



$$\left\{ \begin{array}{l} P(1-P)/n < 0.00051 \text{ avec remise} \\ (N-n)/(n-1) * P(1-P)/n < 0.00051 \text{ sans remise} \end{array} \right.$$

$$\left\{ \begin{array}{l} n > 411.76 \text{ avec remise} \\ n > 323.29 \text{ sans remise} \end{array} \right.$$

2-Si nous ne connaissons pas la proportion d'hommes habituellement touchés par la maladie, il faudrait pour choisir l'estimateur  $\bar{p}$  qui maximise  $Var(\bar{p})$ .

Pour le cas du plan sans remise

$$f(\bar{p}) = Var(\bar{p}) = (N-n) / (n-1) * P(1-P) / n \quad \text{On a } f' = 1-2\bar{p}$$

$\bar{p}$	1/2
$Var(\bar{p})$	
$Var(\bar{p})$	0 

P optimale est

$$\bar{p} = 1/2$$

### EXERCICE 3 :

1- Donnons une estimation totale des notes dans le district :

$$f=m/ M \text{ avec } M = 50 ; m = 5$$

$$AN : f = 5/50 = 1/10 = 0,1$$

Dans chaque collège, la note est estimée par :

$$T_i = N_i * \bar{y}_i$$

On obtient dans les 05 collèges :

$$T_1 = 40 * 12 = 480$$

$$T_2 = 20 * 8 = 160$$

$$T_3 = 60 * 10 = 600$$

$$T_4 = 40 * 12 = 480$$

$$T_5 = 48 * 11 = 528$$

La note totale dans le district est estimée par :

$$T = 1/f * \sum_i T_i$$

$$AN: T = 50/5 (40 * 12 + 20 * 8 + 60 * 10 + 40 * 12 + 48 * 11) = 22\ 480$$

La note totale estimée est :  $T=22\ 480$

2 – Estimation du nombre d'élèves:

$$N = 1/f * \sum_i N_i$$

$$AN : N = 50 / 5 (40 + 20 + 60 + 40 + 48) = 2\ 080$$

Le nombre d'élèves estimée est de :  $N=2\ 080$

3-Sachant que  $N = 2000$ , on a :

$$Y^- = 1 / N * T$$

$$Y^- = 1/2000 * 22\ 480 = 11,24$$

La moyenne observée est donnée par:

$$Y^- = 1/M * \sum_i y_{i\_bar}$$

$$Y^- = 1/50 (10*12 + 10*8 + 10*10 + 10*12 + 10*11) = 10,6$$

Etant donné que  $Y^-$  et  $Y^-$  ne sont pas les mêmes alors  $Y^-$  est biaisé.

4 – Calcul de  $V(T)$  :

$$S^2_1 = 1/4 [(480-449,6)^2 + (160-449,6)^2 + (600-449,6)^2 + (480-449,6)^2 + (528-449,6)^2] = 28\ 620,8$$

$$M^2(1-f) S^2_1 / m = 502 * (1-0,1) * 28620,8 / 5 = 12879360$$

Maintenant en posant :

$$V_i = N^2 i (1-f_{2,i}) S^2_2 / n_i$$

$$AN : V_1 = 402 * (1 - 10/40) * 1,5 / 10 = 180$$

$$V_2 = 202 * (1 - 10 / 20) * 1,2 / 10 = 24$$

Par Analogie,

$$V_3 = 480,$$

$$V_4 = 156,$$

$$V_5 = 364,8.$$

Ainsi en multipliant par  $M/m$ , on obtient que la quantité cherchée est égale :

$$M/m (v_1 + v_2 + v_3 + v_4 + v_5) = 50 / 5 (180+24+480+156+364,5) = 10 * 1204,8 = 12\ 048$$

$$V(T) = 12\ 879\ 360 + 12\ 048 = 12\ 891\ 408$$

$$V(T) = 12\ 891\ 408$$

On peut en déduire la variance de la moyenne :

$$\text{Var}(Y^-) = 1/N^2 * \text{Var}(T) = 1/2000 * 12\ 891\ 408 = 3,22$$

5 - Comparaison avec un sondage aléatoire simple a probabilité est égale sur les mêmes données :

$$Y^- = 10,6 ; n=50 \text{ et } N=2000.$$

Donc le taux de sondage est égal à :

$$f = 50/2000 = 0,25$$

L'estimation de la variance de l'estimateur de la moyenne est égale à :

$$\text{Var } Y = (1-f) * S^2 / n,$$

Avec  $S^2$  la variance corrigée de l'échantillon.

Dans notre échantillon de taille 50, on a :

$$V(T) = \text{variance inter} + \text{variance intra}$$

Calculons maintenant chaque terme qui compose la variance totale :

$$\text{Variance inter} = 1/50 (10*12^2 + 10*8^2 + 10*10^2 + 10*12^2 + 10*11^2) - 10,6^2 = 2,24$$

$$\text{Variance intra} = 1/50 * 0,9 * 10 (1,5 + 1,2 + 1,6 + 1,3 + 2,0) = 1,368$$

$$\text{Donc } V(T) = 2,24 + 1,368 = 3,608$$

$$\text{La variance corrigée est de : } S^2 = 50 / (50 - 1) * 3,608 = 3,68$$

$$\text{Et } \text{Var}(Y^-) = (1 - 0,25) * 3,68 / 50 = 0,07.$$

La précision d'un sondage aléatoire simple a probabilité égale sans remise est supérieure à celle d'un sondage à plusieurs degrés : Pour un intervalle de confiance de 95%, on a plus ou moins :

$$1. 96 * \sqrt{\text{Var}(\bar{p})}$$

Donc on a les précisions suivantes 0,52 et 3,25.

#### **Exercice 4 :**

1-Le nombre maximum d'erreur est de :  $e = n * p$

Pour  $n = 200$  :  $e = 200 * 0,05 = 10$  erreurs

Pour  $n = 400$  :  $e = 400 * 0,05 = 20$  erreurs

Pour  $n = 600$  :  $e = 600 * 0,05 = 30$  erreurs

Pour  $n = 1000$  :  $e = 1000 * 0,05 = 50$  erreurs

2- Le nombre d'enregistrements en tolérant 4 erreurs au plus :

$$4 = n * 0,05$$

Donc  $n = 4/0,05 = 80$  enregistrements

#### **EXERCICE 5 :**

1- Un intervalle de confiance de niveau 0.90 est donné par :

Pour un plan stratifié, la variance est donnée par :

$$\text{Var}(u) = 1/N^2 \sum N_h * (N_h - n_h) / n_h * S^2_h$$

$$\text{Var}(u) = 1/ 1060^2 (500*1,5*500-130/130 + 300*4*300-80/80 + 150*8*150-60/60 + 100*100*100-25/25 + 10*2500*10-5/5) = 0,055$$

$$\text{Pour } z(0, 90) = 1,64$$

$$U (\text{la moyenne}) = 1/N \sum N_h * y_h$$

$$\text{avec } N = 130+80+60+25+5 = 300$$

$$U = 1/ 300(130*5+12*80+30*60+150*25+600*5) = 29, 81$$

L'intervalle de confiance est défini tel que  $U \in [U1 ; U2]$

$$\begin{aligned} \text{Avec } U1 &= U - Z(0, 90) * \sqrt{\text{VAR}(U)} \\ &= 29, 81 - 1,64 * \sqrt{0,055} = 29, 43 \\ U2 &= U + Z(0, 90) * \sqrt{\text{VAR}(U)} \\ &= 29, 81 + 1,64 * \sqrt{0,055} = 30, 19 \\ \text{Donc } U &\in [29, 43 ; 30, 19] \end{aligned}$$

2. (a) - Pour une allocation proportionnelle :

$$n_h = n * N_h / N$$

$$\text{avec } N = 1060 ; n = 300$$

**AN :**

$$\left\{ \begin{array}{l} n1 = 300 * 500/1060 = 142 \\ n2 = 300 * 300/ 1060 = 85 \\ n3 = 300 * 150/ 1060 = 42 \\ n4 = 300 * 100/1060= 28 \\ n5 = 300 * 10/1060 = 3 \end{array} \right.$$

(b) - Pour une allocation optimale :

$$nh = n * Nh Sh / \sum Nh * Sh$$

$$\text{avec somme } Nh * Sh = 500\sqrt{1,5} + 300\sqrt{4} + 150\sqrt{8} + 100\sqrt{100} + 10\sqrt{2500}$$

$$AN : n1 = 300 * 500\sqrt{1,5} / (500\sqrt{1,5} + 300\sqrt{4} + 150\sqrt{8} + 100\sqrt{100} + 10\sqrt{2500}) = 59$$

**Par analogie :**

$$n2 = 57, n3 = 40, n4 = 96, n5 = 48.$$

$$(n5 = 10)$$

$$\text{avec } n = 300 - 10 = 290$$

Les résultats redeviennent :

$$n1 = 290 * 500\sqrt{1,5} / (500\sqrt{1,5} + 300\sqrt{4} + 150\sqrt{8} + 100\sqrt{100}) = 67,35$$

$$n2 = 290 * 300\sqrt{4} / (500\sqrt{1,5} + 300\sqrt{4} + 150\sqrt{8} + 100\sqrt{100}) = 70$$

**Par analogie**

$$n3 = 46.66, n4 = 109,98 \approx 110$$

$$n4 = 110 \text{ dans la strate 4 qui en contient 100.}$$

**On les interroge donc toutes (n4 = 100) et on recalcule**

$$n1, n4 \text{ et } n3$$

$$\text{avec } n = 290 - 100 = 190$$

$$n1 = 190 * 500\sqrt{1,5} / (500\sqrt{1,5} + 300\sqrt{4} + 150\sqrt{8} + 100\sqrt{100})$$

$$n1 = 71$$

**Par analogie**

$$n2 = 70, n3 = 49, n4 = 100, n5 = 10.$$

**3. Pour l'allocation proportionnelle on obtient :**

$$\text{Var}(u) = 1/N^2 * \sum h * (Nh - nh) / nh * S^2h$$

$$\text{Var}(u) = 1/10602 (500 * 1,5 * (500 - 142) / 142 + 300 * 4 * (300 - 85) / 85 + 150 * 8 * (150 - 42) / 42 + 100 * 100 * (100 - 28) / 28 + 10 * 2500 * (10 - 3) / 3)$$

$$\text{Var}(u) = 0,0819$$

Pour l'allocation optimale, on obtient :

$$\text{Var}(u) = 1/N^2 * \sum h * (Nh - nh) / nh * S^2h$$

$$\text{Var}(u) = 1/10602 (500 * 1,5 * (500 - 71) / 71 + 300 * 4 * (300 - 70) / 70 + 150 * 8 * (150 - 49) / 49 + 100 * 100 * (100 - 100) / 100 + 10 * 2500 * (10 - 10) / 10)$$

$$\text{Var}(u) = 0.00974.$$