PROJET DE TECHNIQUE D'ECHANTILLONNAGE ET DE SONDAGE

Ce projet est présenté par :
Djibril Diallo (Statistique – Econométrie)
Rama Diallo (Audit et contrôle de gestion)

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 ?

Les probabilités sont inégales, on peut en déduire que nous n'avons pas un sondage aléatoire simple.

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

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

π_2 = P ({1,2}) + P ({2,3}) = 1/ 2 + 1/ 4 = **3/ 4**

π_3 = P ({1,3}) + P ({2,3}) = 1 /4 + 1 / 4 = **1 / 2**

LA SOMME DES PROBABILITES D'INCLUSIONS EST EGALE A 2 CAR IL S 'AGIT D'UN PLAN DE SONDAGE FIXE DE TAILLE 2.

3- Calculons les probabilités d'inclusion d'ordre 2 de :

$\pi_{12} = \Delta_{12} + \pi_1\pi_2$ or $\Delta_{12} =$ 1/ 2 – (3 / 4 * 3 /4) = **- 1/ 16**

PAR SUITE : π_{12} = - 1/ 16 + (3 /4 * 3 /4) = 8 / 16 = 1 / 2

$\Pi_{23} = \Delta_{23} + \pi_2\pi_3$ OR $\Delta_{23} =$ 1 / 4 – (3 / 4 * 1 / 2) = **- 1 /8**

PAR SUITE : π_{23} = -1 /8 + (3 /4 * 1/ 2) = 1 / 4

Δ = MATRICE DE LA VARIANCE COVARIANCE

4- L'estimateur Y_bar si les échantillons sont tirés :

Sachant que nous avons les probabilités d'inclusions d'ordre 1 :

Si {1,2} est tire : **1/ 3 (** $y_1 + y_2$/ 3/ 4**)**

Si {1,3} est tire : **1 / 3 (** y_1/3/4 + y_3 /1/2 **)**

Si {2,3} est tire : **1/3 (**y_2/3/4 + y_3 /1/2**)**

Par suite : L'estimateur **Y_bar = :**

Si { 1,2 } est tire : 4 (y1 + y2)/ 9

Si {1,3 } est tire : (4y1 + 6y3) / 9

Si { 2,3} est tire : (4y2 + 6y3)/ 9

PAR SUITE L'ESTIMATEUR Y_bar= y1 + y2 + y3 / 3

5- Vérifions que l'estimateur est sans biais :

E (Y_bar) – Y_bar = 0

Or

E (Y_bar) = ½ * (4 (y1 + y2)/ 9) + ¼ * (**(**4y1 + 6y3) / 9) + ¼ ((4y2 + 6y3)/ 9)

E (Y_bar) = 3 y1 + 3y2 + 3y3 / 9

E (Y_bar) = y1 + y2 + y3 / 3 = Y_bar

Par suite l'estimateur est sans biais.

6- Ecrire ce que serait 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

PAR SUITE : P= 1/ C^2_3 **= 1 / 3 (EQUIPROBABILITE)**

LE NOMBRE D' ECHANTILLONS POSSIBLE EST DE : C^2_3 **= 3**

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-Quelle taille d'échantillon faut–il sélectionner pour que la longueur totale d'un inter- valle de confiance avec un niveau de confiance 0.95 soit inférieure à 0.02 pour les plans simples avec et sans remise ?

Si l'on suppose que la taille de l'échantillon est suffisamment grande pour que l'approximation selon la loi normale soit acceptable, on a donc un intervalle de confiance à 95% de la forme :

$$\hat{P} \pm 1.96 \sqrt{Var(\hat{p})}.$$

Par suite : on cherche la taille de l'échantillon n telle que :

$2 \times 1{,}96 \sqrt{var(p)} \leq 0{,}02$ ➔

Or var (p) (**sans remise)= (N – n) p (1 – p) / (N – 1) n**

Var (p) (avec remise) = p (1 – p) / n

P = 3 / 10 ; N = 1500

Par application :

Tirage avec remise :

$2 \times 1{,}96 \sqrt{p(1-p)/n} \leq 0{,}02$

$2 \times 196 \sqrt{p(1-p)/n} \leq 2$

$\sqrt{p(1-p)} / n = 2 / 2 \times 196 \quad \Rightarrow \quad \sqrt{p(1-p)} / n = 1 / 196$

$\sqrt{p(1-p)} / n = 196^{-1}$ donc

$(\sqrt{p(1-p)} / n)^2 = (196^{-1})^2$ donc $p(1-p) / n \leq 196^{-2}$

Or $\mathbf{a^{-n} = 1 / a^n}$ donc on obtient : $\mathbf{n \geq 196^2\, p(1-p)}$

n ≥ 8 067 (tirage avec remise)

<u>Tirage sans remise :</u>

Var (p) = (N – n) p (1- p)/ (N – 1) n

(N – n) p (1- p) / (N – 1) n ≤ 196^{-2}

n ≥ 196^2 N p (1 – p) / (N – 1 + 196^2 p(1 – p)

p = 3 / 30 ; N = 1500

n ≥ 1264 (tirage sans remise)

2- <u>Que faire si nous ne connaissons pas la proportion :</u>

Si l'on ne connait pas a priori la proportion de personnes affectées, il faudrait alors remplacer Var [pˆ] par son estimation.

Dans ce cas on obtient :

Var (p) = p (1-p) / (n- p) **: tirage avec remise (AR)**

Var (p) = (N - n) /N * p (1 – p)/ n – 1 : tirage **sans remise (SR)**

Notons que le p est un p avec chapeau comme défini ici Var [pˆ].

Une autre approche consisterait à prendre le cas le plus mauvais (ou pessimiste), c'est à dire la valeur théorique de p telle que Var [pˆ] soit la plus grande possible. Clairement le cas le plus pessimiste correspond au choix p = 0.5.

<u>Cas sur la consommation des 25 automobilistes au 100km</u>

Les hypothèses se traduisent par :

n = 25 ; x_bar = 8,5 ; S (l'ecartype) = 0,8

Nous utiliserons la table de student a n-1 degrés de liberté : n – 1 = 24

L'ecartype de la population est inconnu :

Si $\alpha = 0,05$ alors sur la table de student $t\alpha = 2,064$ pour n= 24

L'intervalle de confiance ayant 95 chance sur 100 de contenir la valeur de la moyenne est de :

$X\_bar - t\alpha * S / \sqrt{n-1} \leq m \leq X\_bar + t\alpha * S / \sqrt{n-1}$

$8,5 - 2,064 * 0,8 / \sqrt{24} \leq m \leq 8,5 + 2,064 * 0,8/\sqrt{24}$

$8,16 \leq m \leq 8,83$

La probabilité que la consommation moyenne soit compris entre [8,16 ; 8,83] est égale a 95 %.

<u>2 – pour une marge d'erreurs de 2 décilitres :</u>

2 dl = 0,2 L

Or $0,2 = t\alpha * S / \sqrt{n - 1}$

$0,2 (\sqrt{n-1}) = t\alpha * S$ donc $0,2^2 (n-1) = (t\alpha S)^2$ par suite :

$n = (t\alpha S)^2 / 0,04 + 1$

Avec $t\alpha = 1, 96$ pour tout $n > 30$ et $\alpha = 0, 05$

$n = (1, 96 * 0,8)^2 / 0, 04 + 1 = 62, 5$

Pour $\alpha = 0,01$, dans la table de student $t\alpha = 2,576$ pout tout $n > 30$

$n = (2,576 * 0,8)^2 / 0, 04 + 1 = 106$

EXERCICE 3 :

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

M = 50 ; m= 5 par suite f = 5/50= 1/10= 0,1

Dans chaque collège, la note est estimée par :

$T_i = N_i * y_i$_bar : 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 de la district est estime par :

$T = M/m * (\sum Ti)$:

T = 50/5 (480 + 160 + 600 + 480 + 528) = 22 480

La note totale estimée est : **22 480**

2 – le nombre d'élèves estimées est de :

$N = M/m * (\sum Ni)$:

N = 50 /5 (40 + 20 + 60 + 40 + 48) = 2 080

Le nombre d'élèves estimée est de : **2 080**

3-pour N = 2000, donnons une estimation de la moyenne et comparons :

Y_bar = 1 /N * T

Y_bar = 1/2000 * 22 480 = 11,24

Par conséquent : la moyenne observée sur N = 50 est de :

: y_bar = 1/50 (10* 12 + 10*8 + 10* 10 + 10* 12 + 10 * 11) = 10,6

Comparons : y_bar n'est un bon estimateur de Y_bar : Y_bar ≠ y_bar

4 – Calculons la variance de l'estimateur total :

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

$M^2 (1- f_i) S^2_1 / m = 50^2 * (1-0,1) * 28620,8\ 4/ 5 = 12\,879\,360$

Maintenant en posant :

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

Par application :

$V1 = 40^2 * (1 - 10/40) * 1,5 / 10 = 180$

$V2 = 20^2 * (1 - 10 / 20) * 1,2 / 10 = 24$ donc dans la même logique,

$V_3 = 480,\ V_4 = 156,\ V_5 = 364,8.$

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

: M/m (v1 + v2 + v3 + v4 +v5)

= 50 /5 (180+24+480+156+364,5) = 10 * 1204,8 = 12 048

L'estimation de la variance de l'estimateur du total est égale :

Var (T)= 12 879 360 + 12 048 = 12 891 408.

On peut en déduire la variance de la moyenne :

Var (y_bar) = $1/N^2$ * 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_bar=y_bar =10,6 ; n=50 et N=2000.

Donc le taux de sondage est égal a :

f= 50/2000 =0,25

L'estimation de la variance de l'estimateur de la moyenne est égal a :

Var Y = (1−f) * S^2/n, ou S^2 est la variance corrigée de l'échantillon.

Dans notre échantillon de taille 50, on a :

variance totale = variance inter + variance intra

Calculons maintenant chaque terme qui compose la variance totale :

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

Variance intra = 1/ 50 * 0,9*10 (1,5 + 1,2 + 1,6 + 1,3 + 2,0) = 1,368

Donc Variance totale = 2,24 + 1,368 = 3,608

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

Et Var (Y_bar) = (1 – 0,25) * 3,68/ 50 = 0, 07.

La précision d'un sondage aléatoire simple a probabilité égale sans remise est supérieure a celle d'un sondage a plusieurs degrés :

Pour un intervalle de confiance de 95%, on a :

Plus ou moins : 1,96√var (Y_bar) donc on a les précisions suivantes 0,52 et 3,25.

EXERCICE 5 :

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

Pour un plan stratifie, la variance est donnée par :

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

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

Pour $z_{0,90}$ = 1,64

U (la moyenne) = 1/N $\sum$ $Nh * yh$ 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 définie tel que U € $[U_1 ; U_2]$

Avec U_1 = U - $Z_{0,90}$ * $\sqrt{VAR}$ (U) = 29, 81 – 1,64 * $\sqrt{}$ 0,055= 29, 43

U_2 = U + $Z_{0,90}$ * $\sqrt{VAR}$ (U)= 29, 81 + 1,64 *$\sqrt{}$ 0,055= 30, 19

Donc U € [29, 43 ; 30, 19]

2. (a) - Pour une allocation proportionnelle :

n_h = n * N_H / N avec N = 1060 ; n = 300

par application : 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

(b) - Pour une allocation optimale :

n_h= n * N_H S_h/ $\sum$ $Nh * Sh$ avec somme $Nh * Sh$ = 500$\sqrt{}$1,5 + 300$\sqrt{}$4+150$\sqrt{}$8+100$\sqrt{}$100+ 10$\sqrt{}$2500

Par application :

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

= 59

Par suite dans la même logique

n1 =59, n2 =57, n3=40, n4 =96, n5=48.

On doit interroger 48 personnes dans la strate 5 alors qu'elle n'en contient que 10. C'est bien entendu impossible, on choisit donc d'interroger les 10 personnes de la strate 5 (n5= 10) et on recalcule les tailles d'échantillons pour les quatre autres strates avec n = 300 − 10 = 290.

Les résultats redeviennent :

n1 = 290 * 500√1, 5 / (500√1, 5 + 300√4+150√8+100√100) = 67, 35

n2 = 290 * 300√4 / (500√1, 5 + 300√4+150√8+100√100) = 70

De même logique n3= 46, 66, n4= 109,98.

Encore une fois, on doit interroger n4 = 110 individus dans la strate 4 qui en contient 100. On les interroge donc toutes (n4 = 100) et on recalcule n1, n4 et n3 avec n = 290 − 100 = 190.

n1 = 190 * 500√1, 5 / (500√1, 5 + 300√4+150√8+100√100) =

n_1 =71, n_2 =70, n_3 =49, n_4 =100, n_5 =10.

3. Pour l'allocation proportionnelle on obtient :

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

Var (u) = $1/ 1060^2$ (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 =

Var (u)= 0, 0819

Pour l'allocation optimale, on obtient :

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

Var (u) = $1/1060^2$ (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=

Var (u) = 0.00974.

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