UNIVERSITE DE THIES
UFR SES/SET
MASTER EN SCIENCES DE DONNEES ET APPLICATION
PROJET DE TECHNIQUE DE SONDAGE

PROFESSEUR Mme Diop

Presente par
ABDOULAYE FALL
AWA WADE
NDEYE COUMBA CISSE

Exercice 1 : Probabilité d'inclusion

Soit la population {1,2,3} et le plan de sondage suivant :

P {1,2} =1/2 ; P {1,3} =1/4 ; P {2,3} =1/4.

1) Ici on n'a pas un sondage aléatoire simple, car tous les échantillons n'ont pas la même probabilité d'^être sélectionné.

2) Calculons la probabilité d'inclusion d'ordre 1 π1, π2 et π3

π1 = ∑P(S)= P ({1,2}) + P ({1,3}) = 3/4

π2 = ∑P(S)= P ({1,2}) + P ({2,3}) = 3/4

π3 = ∑P(S)= P ({1,3}) + P ({2,3}) = ½

3) Calculons les probabilités d'inclusion d'ordre 2 π12 et π23

π12=∑P(S)= X12-

π23= ∑P(S)= X

4) Le π-estimateur de ȳ

t^ȳ.π =1/N ∑yk / πk=1/3 (y1/ π1+ y2/ π2+ y3/ π3)

On a : y1=1.5 ; y2=2 ; y3=2.5

t^ȳ.π =1/3[1.5/ (3/4) +2 / (3/4) + 2.5/ (1/2)]

t^ȳ.π=1/3(2+2.6+5)

t^ȳ.π=0,88

On note y1,y2,y3 les valeurs respectives de la variable y

a) Si l'échantillon {1,2} est tiré ?

b) Si l'échantillon {1,3} est tiré ?

c) Si l'échantillon {2,3} est tiré ?

5) Vérifions que le π-estimateur est un estimateur sans biais

Théorème : Si πk > 0 pour tout k ∈ U, alors t^ȳ.π estime tȳ sans biais

Or on a : π1=3/4

π2=3/4 => π2 > 0 ∀ k ∈ U.

π3=1/2 donc π-estimateur est un estimateur sans biais.

6) Dans le cas d'un sondage aléatoire simple à probabilité égales sans remise ,

nous aurons : P({1,2})= P({1,3})= P({2,3})=1/3

Et pour les probabilités d'inclusion on aura :

π1 = π2 = π3 = 2/3

EXERCICE 2

. Le paramètre d'intérêt est donné par

P= 1/N ∑yk, k € U

où les y_k sont des indicatrices codant la présence ou non de la maladie. On estimera ce paramètre par p® =1/n ∑Yk k € U

et la variance de cet estimateur est donnée par

Var(ˆp) = β^2 y/n ,avec remise,

Var(ˆp)= N-n/N *s^2y/n , sans remise

mais puisque $y_k^2 = y_k$, la variance et la variance corrigée sur la population sont égales à

= 1/N ∑yk, k € U -- (1/N ∑yk, k € U)= P-P*P= P(1-P) ; $S_y^2 = \frac{N}{N-1}p(1-p)$.

Ainsi on a donc

Var(ˆp) = p(1-p) n , avec remise, Var(ˆp) = N-n /N-1. p(1-p)/ n , sans remise.

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

pˆ± 1.96 *racine*Var(pˆ).

Ainsi on cherche donc la taille de l'échantillon $_n$ telle que

2 ◊ 1.96 ◊ $\sqrt{\text{Var}(\hat{p})}$ Æ 0.02 ≈Δ Var($\hat{p}$) Æ 196≠^{2}

$\frac{p(1\neq p)}{n}$ ^{2}

$\frac{N\neq n}{N\neq 1}\frac{p(1\neq p)}{n}$ Æ 196≠^{2}

≈Δ]Y[Æ 196≠ , , avec remise avec remise

≈Δ $_n$n ØØ 196196$^{22p}{}_{Np}$(1(1≠≠^{p})$_p$)/{N ≠ 1 + 196$^2{}_p$(1 ≠ p)} sans remise avec remise.

En prenant $_p$ = 3/10 et $_N$ = 1500 on trouve alors que

Y]8067, avec remise *n* >

1264, sans remise.

Notons qu'avec remise la taille d'échantillon requise est supérieure à la taille de la population :-(

Exercice 3

1. Au 1^{er} degré, on a

$$M = 50 \text{ colleges}, \quad m = 5 \text{ colleges}, \quad f_1 = 0{,}1.$$

Au $2^{eme`}$ degré, on a

Observation	N_i	n_i	$\bar{y}_i$	s_{2i}	$\widehat{T}_i$
1	40	10	12	1,5	480
2	20	10	8	1.2	160
3	60	10	10	1.6	600
4	40	10	12	1.3	480
5	48	10	11	2.0	528
Total	208	50			2248

Dans chaque collège, on estime la note totale T_i par

$$\widehat{T}_i = N_i \bar{y}_i.$$

Ce qui donne avec les valeurs numériques T_{b1} =

$$40 \cdot 12 = 480, \; T_{b2} = \ldots$$

On estime la note totale dans le district par

$$\begin{aligned} \widehat{T} &= \frac{M}{m} \sum_{i=1}^{m} \widehat{T}_i \\ &= \frac{50}{5} \cdot 2\,248 \\ &= 22\,480. \end{aligned}$$

2. Le nombre estimé d''élèves est égal a

$$\begin{aligned} \widehat{N} &= \frac{M}{m} \sum_{i=1}^{m} N_i \\ &= \frac{50}{5} \cdot (40 + 20 + \ldots + 48) \\ &= 2\,080. \end{aligned}$$

$$1$$

3. Si l'on sait que $N = 2000$ alors

$$\begin{aligned} \widehat{\overline{Y}} &= \frac{1}{N} \cdot \widehat{T} \\ &= \frac{1}{2\,000} \cdot 22\,480 \\ &= 11,24. \end{aligned}$$

Par conséquent la moyenne observée sur l''échantillon de taille 50 est égale a

$$\overline{y} = \frac{1}{50} \cdot (10 \cdot 12 + \ldots + 10 \cdot 11) = 10,6$$

En général $\overline{y}$ n'est pas un bon estimateur de $\overline{Y}$. Il n'y a que dans le cas particulier ou` les taux de sondage $f_i = n_i/N_i$ sont constants et si toutes les unités primaires ont la même taille que $Y = y$.

4. Calculons la variance de l'estimateur du total. On ne peut pas la calculer. Donc on calculera une estimation de cette variance. Elle est égale a

$$\widehat{\mathrm{Var}}\left(\widehat{T}\right) = M^2 \left(1 - f_1\right) \frac{s_1^2}{m} + \frac{M}{m} \sum_i N_i^2 \left(1 - f_{2,i}\right) \frac{s_{2,i}^2}{n_i}$$

Ou

$$s_1^2 = \frac{1}{m-1} \sum_{i=1}^{m} \left(\widehat{T}_i - \frac{\widehat{T}}{M}\right)^2$$

qui est la variance observée entre les unités primaires et

$$s_{2,i}^2 = \frac{1}{n_i - 1} \sum_j \left(y_{i,j} - \overline{y}_i\right)^2$$

qui est la variance observée dans les unités secondaires. $s_{2,i}^2$ est donnée dans le tableau. Il ne reste plus qu'a calculer s_1^2.

$$s_1^2 = \frac{1}{4} \left[(480 - 449,6)^2 + \ldots + (528 - 449,6)^2 +\right] = 28\,620,8$$

qui peut se calculer également grâce `a la formule d´développée suivante

$$\frac{1}{m-1} \sum_{i=1}^{m} \widehat{T}_i^{\,2} - \frac{m}{m-1} \widehat{T}^2$$

.

Maintenant, on peut calculer le premier terme de l'estimation de la variance de l'estimateur du total qui vaut alors

$$M^2 \left(1 - f_1\right) \frac{s_1^2}{m} = 50^2 \cdot 0,9 \cdot \frac{28620,8}{5} = 12\,879\,360$$

.

Maintenant en posant

$$V_i = N_i^2 \left(1 - f_{2,i}\right) \frac{s_{2,i}^2}{n_i}$$

2 ,

on peut calculer le second terme de l'estimation de la variance de l'estimateur du total qui vaudra alors la somme des quantités suivantes pondérée par la

quantite M/m,

$$\begin{aligned} V_1 &= 40^2 \cdot \left(1 - \frac{10}{40}\right) \cdot \frac{1,5}{10} = 180 \\ V_2 &= 24, \\ V_3 &= 480, \\ V_4 &= 156 \\ V_5 &= 364,8. \end{aligned}$$

,

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

$$\frac{M}{m}\sum_{i=1}^{5} V_i = \frac{50}{5} \cdot 1204,8 = 12\,048$$

.

Finalement, l'estimation de la variance de l'estimateur du total est égale a

$$\widehat{\mathrm{Var}}\left(\widehat{T}\right) = 12\,879\,360 + 12\,048 = 12\,891\,408$$

.

On en d´déduit la variance de la moyenne qui est égale `a

$$\begin{aligned} \widehat{\mathrm{Var}}\left(\overline{Y}\right) &= \frac{1}{N^2}\widehat{\mathrm{Var}}\left(\overline{T}\right) \\ &= \frac{1}{(2\,000)^2} \cdot 12\,891\,408 \\ &= 3,22. \end{aligned}$$

On obtient ainsi la précision qui est égale a

$$\pm 1{,}96\sqrt{3{,}22} = 3{,}5,$$

et qui va nous permettre de calculer un intervalle de confiance de la moyenne, `a 95% qui vaut

$$11{,}2 \pm 3{,}5$$

autrement dit la précision est très mauvaise. Cela est duˆ `a la grande dispersion des totaux.

5. Comparaison avec un sondage aléatoire simple `a probabilités égales sur les mêmes données.

On prend

$$\widehat{\overline{Y}} = \overline{y} = 10,6, \quad n = 50 \quad \text{et} \quad N = 2000.$$

Donc le taux de sondage est égal `a

$$f = \frac{50}{2\,000} = 0,25$$

.

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

$$\widehat{\mathrm{Var}}\left(\widehat{\overline{Y}}\right) = (1-f)\frac{s^2}{n}$$

,

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 $= \frac{1}{50} \cdot \left(10 \cdot 12^2 + \ldots + 10 \cdot 11^2\right) - 10,6^2 = 2,24$
- variance intra = $\frac{1}{50} \cdot \left(10 \cdot \frac{9}{10} \cdot 1,5 + \ldots + 10 \cdot \frac{9}{10} \cdot 2,0\right)$

$$= 1,368.$$

Donc la variance totale est égale a

$$2,24 + 1,368 = 3,608.$$

et par conséquent la variance corrigée est égale a

$$s^2 = \frac{50}{49} \cdot 3,608 = 3,68.$$

On en déduit que

$$\widehat{\mathrm{Var}}\left(\widehat{\overline{Y}}\right) = (1 - 0,25) \cdot \frac{3,68}{50} = 0,07.$$

La précision est égale a

$$\pm 1,96\sqrt{0,07} \simeq \pm 0,52.$$

D'ou`

$$\overline{Y} = 10,6 \pm 0,52.$$

Conclusion : La précision d'un sondage aléatoire simple `a probabilités égales sans remise est supérieure `a celle d'un sondage `a plusieurs degrés, surtout que les classes sont peu homogènes générant une grande variance au 1er degré.

Exercice 4

1) le nombre maximum d'erreurs qu'on peut acceptation dans cet échantillon sans remettre en cause le niveau d'acceptation

on a Nombre d'erreurs=n*P

***POUR n=200 on a :**

Nombre d'erreurs= 0,05*200=10 erreurs

Même question avec n=400, n=600 et n=1000

***POUR n=400 on a :**

Nombre d'erreurs =0,05*400= 20 erreurs

***POUR n=600 on a :**

Nombre d'erreurs =0,05*600= 30 erreurs

***POUR n=1000 on a :**

Nombre d'erreurs = 0,05*1000= 50 erreurs

2) le nombre d'enregistrement supplémentaire qu'on doit effectuer pour que l'hypothèse soit acceptée

0 ,05*n= (7+4) => n=11/0,05=180

Donc on doit faire 180 enregistrements supplémentaires pour que l'hypothèse d'un niveau d'acceptation de 5% puisse être raisonnement raisonnablement retenue.

EXERCICE 5

1. Un intervalle de confiance de niveau 0.90 est donné par

$$IC_{0.90} = \left[\hat{\mu} - z_{0.95}\sqrt{\mathbf{V}(\hat{\mu})}, \hat{\mu} + z_{0.95}\sqrt{\mathbf{V}(\hat{\mu})}\right],$$

avec $z_{0.95} \simeq 1.64$. On calcule $\mathbf{V}(\hat{\mu})$ grâce à (B.3) et on obtient $\mathbf{V}(\hat{\mu})$ =

$$0.055.$$

On calcule $\hat{\mu}$ = 29.81 et on déduit

$$IC_{0.90} = [29.43;30.19].$$

2. (a) Pour une allocation proportionnelle $n_h = n\frac{N_h}{N}$, donc

$n_1 = 141.51$, $n_2 = 84.91$, $n_3 = 42.45$, $n_4 = 28.30$, $n_5 = 2.83$,

en arrondissant

$n_1 = 142$, $n_2 = 85$, $n_3 = 42$, $n_4 = 28$, $n_5 = 3$.

(b) (plus difficile) Pour une allocation optimale

$$n_h = n \times \frac{{}_H N^h S_h}{\mathrm{P}_{h=1} N_h S_h},$$

ce qui donne $n_1 = 58.57,\quad n_2 = 57.39,\quad n_3 = 40.58,\quad n_4 = 95.64,$

$n_5 = 47.82,$

en arrondissant

$$n_1 = 59,\quad n_2 = 57,\quad n_3 = 40,\quad n_4 = 96{,}4\quad n_5 = 48.$$

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 ($n_5 = 10$) et on recalcule les tailles d'échantillons pour les quatre autres strates avec $n = 300 - 10 = 290$. On a par exemple pour n_1

$$n_1 = 290\frac{500\sqrt{1.5}}{500\sqrt{1.5} + 300\sqrt{4} + 150\sqrt{8} + 100\sqrt{100}} = 67.35,$$

de même $n_2 = 65.99$, $n_3 = 46.66$, $n_4 = 109.98$.

Encore une fois, on doit interroger $n_4 = 110$ individus dans la strate 4 qui en contient 100. On les interroge donc toutes ($n_4 = 100$) et on recalcule n_1,n_2 et n_3 avec $n = 290 - 100 = 190$. On obtient après arrondi

$$n_1 = 71,\quad n_2 = 70,\quad n_3 = 49.$$

Pour résumer

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

3. Pour l'allocation proportionnelle on obtient grâce à (B.3)

$$\mathbf{V}(\hat{\mu}) = 0.0819.$$

Pour l'allocation optimale, on obtient :

$$\mathbf{V}(\hat{\mu}) = 0.00974.$$