UNIVERSITE DE THIES

UFR SCIENCES ECONOMIQUES ET SOCIALES

DEPARTEMENT MANAGEMENT INFORMATISEISE DES ORGANISATIONS

MASTER 1 SCIENCES DE DONNEES ET APPLICATIONS

PROJET TECHNIQUES DE SONDAGE

PRESENTE PAR :

ROKHAYA GUEYE

CHEIKH MBACKE DIOUF

PAPE MOUSSA GUEYE

***Professeur* : Mme DIOP**

ANNEE UNIVERSITAIRE : 2019/2020

EXERCICE 1 :

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

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

1- Est-ce un sondage aléatoire simple ?

Les probabilités étant inégales, nous pouvons en déduire qu'il ne s'agit pas d'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**

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

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

S'agissant d'un plan de sondage fixe de taille 2 alors l'ensemble des probabilités d'inclusions est égal à 2.

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

$\pi 12 = \Delta 12 + \pi 1 \pi 2$ or $\Delta 12 = 1/2 - (3/4 * 3/4) =$ **-1/16**

Ce qui donne par la suite : $\pi 12 = -1/16 + (3/4 * 3/4) = 8/16 =$ **1 / 2**

$\pi 23 = \Delta 23 + \pi 2 \pi 3$ or $\Delta 23 = 1/4 - (3/4 * 1/2) =$ **-1/8**

Ce qui donne par la suite : $\pi 23 = -1/8 + (3/4 * 1/2) =$ **1 / 4**

Δ représentant la matrice de la covariance

4- Calculons le *π*-estimateur $\bar{Y}$. Nous noterons *Y1, Y2* et *Y3* les valeurs respectives de la valeur *Y*.

L'estimateur $\bar{Y}$ si les échantillons sont tirés sachant que nous avons les probabilités d'inclusions d'ordre1 :

a) Si $\{1,2\}$ est tiré : **1/3 (y1 + y2 / 3/4)**

b) Si $\{2,3\}$ est tiré : **1/3 (y2 / 3/4 + y3 / 1/2)**

c) Si $\{1,3\}$ est tiré : **1/3 (y1 / 3/4 + y3 / 1/2)**

Par la suite : **l'estimateur $\bar{Y}$ = : Si {1,2} est tiré : 4 (y1 + y2) / 9**

Si {1,3} est tiré : (4y1 + 6y3) / 9

Si {2,3} est tiré : (4y2 + 6y3) / 9

Ce qui donne par la suite **l'estimateur $\bar{Y}$ = y1 + y2 + y3 / 3**

5- Vérifions que l'estimateur est sans biais :

E ($\bar{Y}$) – $\bar{Y}$ = 0 Or E **($\bar{Y}$)** = 1/2 * (4 (y1 +y2) / 9) + 1/4 * ((4y1 + 6y3) / 9) + 1/4 ((4y2 + 6y3) / 9)
E ($\bar{Y}$) = (3y1 + 3y2 + 3y3) / 9

E ($\bar{Y}$) = (y1 + y2 + y3) / 3 = $\bar{Y}$

Par conséquent, l'estimateur est sans biais.

6- Ecrivons ce que serait les probabilités d'échantillons P et les probabilités d'inclusion π pour un sondage aléatoire simple à probabilités égales sans remise :

Pour un sondage aléatoire simple à probabilités égales sans remise, la loi de probabilité suit une loi telle que : **$P = 1 / C_N^n$**

Par conséquent : **P = 1 / C23 = 1 /3 (Equiprobabilité).**

Le nombre d'échantillons possibles est de : **C23 = 3** alors 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 :

Nous nous intéressons à la proportion d'hommes atteints par une maladie professionnelle dans une entreprise de 1500 travailleurs. Sachant par ailleurs que trois travailleurs sur dix sont ordinairement touchés par cette maladie dans des entreprises du même type, nous nous proposons de sélectionner un échantillon au moyen d'un sondage aléatoire simple.

1- La taille d'échantillon qu'il faut sélectionner pour que la longueur totale d'un intervalle de confiance avec un niveau de confiance 95% soit inférieure à 0.02 pour les plans simples avec et sans remise :

Supposant 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 **P^ ± 1,96 √ var (p^).**

Trouvons la taille de l'échantillon n telle que : **2 * 1,96 √ var (p) ≤ 0,02**

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

P = 3 / 10 ; **N = 1500**

Tirage avec remise : $2 * 1{,}96\sqrt{p(1-p)/n} \leq 0{,}02$

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

$\sqrt{p(1-p)/n} = 2 / 2 * 196$; $\sqrt{p(1-p)/n} = 1 / 196$

$\sqrt{p(1-p)/n} = 196^{-1}$ donc $(\sqrt{p(1-p)/n})^2 = (196\text{-}1)^2$ donc $p(1-p)/n \leq 196^{-2}$ Or $a^{-n} = 1 / a^n$ ainsi on a : $n \geq 196^2\, p(1-p)$; **$n \geq 8\,067$ (tirage avec remise)**

Tirage sans remise : Var (p)= (N – n) p (1 - p) / (N – 1) n

(N – n) p (1 - p) / (N – 1) n $\leq 196^{-2}$

$n \geq 196^2$ N p (1 – p) / (N – 1 + 196^2 p (1 – p)

p = 3 / 30 ; N = 1500 ; **n ≥ 1264 (tirage sans remise)**

2- Que faire si nous ne connaissons pas la proportion :

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

Ainsi, 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 le plus grand possible. Clairement le cas le plus pessimiste correspond au choix **p = 0.5**.

La consommation des 25 automobilistes au 100km :

Les hypothèses se traduisent par : n = 25 ; x_bar = 8,5 ; S (l'écart type) = 0,8

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

L'écart type de la population est inconnu : Si α = 0,05 alors sur la table de student tα= 2,064 pour n= 24.

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

X_bar – tα * S / √n-1 ≤ m ≤ X_bar + tα * S / √n-1

8,5 – 2,064 * 0,8 / √24 ≤ m ≤ 8,5 + 2,064 * 0,8/√24

8,16 ≤ m ≤ 8,83

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

2– Pour une marge d'erreurs de 2 décilitres :

2 dl = 0,2L Or 0,2 = tα * S / √n - 1; 0,2 (√n-1) = tα * S donc $0,2^2$ (n-1) = $(t\alpha S)^2$

Ce qui donne : n = $(t\alpha S)^2$ / 0,04 + 1 avec tα = 1,96 pour tout n > 30 et α = 0,05

n = (1, 96 * 0,8) ² / 0,04 + 1 = 62,5

Pour α = 0,01, dans la table de student tα = 2,576 pour 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 delà f = 5/50 = 1/10 = 0,1

Dans chaque collège, la note est estimée par : **Ti = Ni* Ȳi**

On obtient dans les 05 collèges : T1 = 40 * 12 = 480

T2= 20 * 8 = 160

T3 = 60 * 10 = 600

T4 = 40 * 12 = 480

T5 = 48 * 11 = 528

La note totale du district est estimée par : **T = M/m * (*Ti*) :**

T = 50/5 (480 + 160 + 600 + 480 + 528) = **22480** Ce qui constitue par conséquent la note totale ainsi estimée.

2– Le nombre d'élèves estimées est :

N = M/m ∗ (*Ni*) : N = 50/5 (40 + 20 + 60 + 40 + 48) = 2080 constituant le nombre d'élèves estimé.

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

Ȳ = 1 / N * T

Ȳ = 1/2000 * 22480 = 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

Par comparaison, Ȳ est différent de 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
= **28620,84**

$M^2 (1 - fi) S^2_1 / m = 502 * (1-0,1) * 28620,84 / 5 = 12879360$

En posant : **$Vi = N^2i (1-f_{2,} i) S^2_2 /ni$**

V1 = 40² * (1 – 10 / 40) * 1,5 / 10 = 180

V2 = 202 * (1 – 10 / 20) * 1,2 / 10 = 24

Suivant cette même logique :

V3 = 480, V4 = 156, V5 = 364,8. Donc en multipliant par M/m, on trouve que la quantité cherchée est égale à :

M/m (V1 + V2 + V3 + V4 + V5) = 50 /5 (180 + 24 + 480 + 156 + 364,5) = 10 * 1204,8 = **12048**

L'estimation de la variance de l'estimateur du total donne :

Var (T) = 12879360 + 12048 = **12891408**

On peut en déduire la variance de la moyenne :

Var (y_bar) = 1 / N2 * Var (T) = 1 / 2000 * 12891408 = **3,22**

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

$\bar{Y}$ = y_bar = 10,6 ; n = 50 et N =2000. Ainsi, 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 :

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 : var totale = var inter + var intra

Calcul de chaque terme qui compose la var totale :

Var inter = 1/50 (10 * 122 + 10 * 82 + 10 * 102 + 10 * 122 + 10 * 112) - 10,62 = **2,24**

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

Par conséquent, var totale = 2,24 + 1,368 = **3,608**

La variance corrigée est de : S^2 = 50 / (50 - 1) * 3,608 = **3,68** et var($\bar{Y}$) = (1 – 0,25) * 3,68/50 = **0,07**.

La précision d'un sondage a plusieurs degrés est inférieure à celle d'un sondage aléatoire simple a probabilité égale sans remise.

Pour un intervalle de confiance de 95%, on a plus ou moins : 1,96 √var ($\bar{Y}$), delà on obtient les précisions suivantes **0,52** et **3 75.**

<u>Exercice 4</u> :

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 par conséquent, n = 4/0,05 = **80 enregistrements**

<u>EXERCICE 5</u> :

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

Pour un plan stratifie, la variance est donnéepar :

Var (u) = 1 / N2 $Nh*$ (Nh - nh) / nh * S²h Var (u) = 1 / 1060² (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 z0,90 = 1,64. U (la moyenne) = 1/N $Nh*yh$ 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 € [U1; U2] avec

U1 = U - Z0,90 * √var (U) = 29,81 – 1,64 * √0,055 = **29,43**

U2 = U + Z0,90 * √Var (U) = 29,81 + 1,64 * √0,055 = **30,19**

Donc **U € [29,43 ; 30,19]**

2. (a) - Pour une allocation proportionnelle :

nh = n * NH / 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 :

nh = n * NH Sh / $Nh*Sh$ avec somme $Nh*Sh$ = 500√1,5 + 300√4 + 150√8 + 100√100 + 10√2500

Par application :

n1 = 300 * 500√1,5 / (500√1,5 + 300√4 + 150√8 + 100√100 + 10√2500) = 59

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 10personnes 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**

Suivant cette logique, n3 = **46,66** ; n4 = **109,98**

On doit interroger **n4 = 110** individus dans la strate 4 qui en contient 100.

On les interroge donc tous 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) = **71**

Par conséquent, n2 = **70**, n3 = **49**, n4 = **100**, n5 = **10**

3. Pour l'allocation proportionnelle, on obtient :

Var (u) = 1/N2 $Nh *$ (Nh - nh) / nh *S^2h

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 = Var (u) = 0,0819

Pour l'allocation optimale, on obtient :

Var (u) = 1/N2 $Nh *$ (Nh - nh) / nh * S^2h

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 = **Var (u) = 0.00974**