STA108 – Plans à probabilités inégales

Cours n°3 du 09/10/2020

Sylvie Rousseau

le cnam

Sommaire

1) **Pourquoi échantillonner à probabilités inégales ?**
2) Comment choisir les probabilités d'inclusion ?
3) Comment estimer un total ?
4) Comment estimer une moyenne ?
5) Comment construire des intervalles de confiance pour le total ou la moyenne ?
6) Application dans Sas et dans R

Pourquoi échantillonner à probabilités inégales ?

Pour retenir de préférence les unités les plus porteuses de l'information

⇨ Gain de précision par rapport à un sondage aléatoire simple sans remise où toutes les unités ont même importance

Quand échantillonner à probabilités inégales ?

Lorsque les unités de la population étudiée contribuent inégalement au total d'intérêt

Exemple :

Pour estimer la production d'un secteur que l'on sait assurée par 2 géants du secteur et des centaines de PME, il est légitime de sélectionner d'office les 2 grandes entreprises et échantillonner de manière aléatoire quelques PME

Exemple

- Population de 4 entreprises A, B, C et D de 500, 100, 30 et 20 salariés
- On veut estimer le nombre total de salariés (certes connu : 650) à partir d'un échantillon de taille 2
- Comparons les 2 tirages suivants :
  - un sondage aléatoire simple sans remise
  - un échantillonnage à probabilités inégales

Sondage aléatoire simple

- Il y a $C_4^2 = 6$ échantillons possibles

Échantillon s	Probabilité de tirage p(s)	Estimation du nombre total de salariés
{A, B}	1/6	1 200
{A, C}		1 060
{A, D}		1 040
{B, C}		260
{B, D}		240
{C, D}		100

- En moyenne, on estime parfaitement bien le vrai effectif total de 650 salariés $\frac{1}{6}\times(1200+1060+1040+260+240+100)=650$
- Mais l'estimateur est très dispersé :

  sa variance vaut 207 567 (CV ≅ 70%).

$$\frac{1}{6}\times[(1200-650)^2+\ldots+(100-650)^2]=207567$$

$$CV=\frac{\sqrt{207567}}{650}$$

$$(1-\frac{2}{4})\times\left(\frac{\frac{1}{3}\times[(500-\frac{650}{4})^2+\ldots+(20-\frac{650}{4})^2]}{2}\right)=207567$$

le cnam

Exemple d'un bon choix des probabilités d'inclusion

Avec les probabilités d'inclusion suivantes :

Entreprise k	Effectif salarié X_k	Probabilité d'inclusion π_k
A	500	1
B	100	0,5
C	30	0,25
D	20	0,25
Total	650	2

3 échantillons sont possibles :

Échantillon s	Probabilité de tirage p(s)	Estimation du nombre total de salariés
{A, B}	0,5	700
{A, C}	0,25	620
{A, D}	0,25	580

En moyenne, l'estimateur est aussi sans biais $0{,}5 \times 700 + 0{,}25 \times 620 + 0{,}25 \times 580 = 650$

Et sa variance est beaucoup plus faible : elle vaut 2 700 (CV ≅ 8%)

$$0{,}5 \times (700 - 650)^2 + 0{,}25 \times (620 - 650)^2 + 0{,}25 \times (580 - 650)^2 = 2700$$

le cnam

Exemple d'un mauvais choix des probabilités d'inclusion

Avec les probabilités d'inclusion suivantes :

Entreprise k	Effectif salarié Xk	Probabilité d'inclusion π_k
A	500	0,25
B	100	0,25
C	30	0,5
D	20	1
Total	650	2

3 échantillons sont possibles :

Échantillon s	Probabilité de tirage p(s)	Estimation du nombre total de salariés
{A, D}	0,25	2020
{B, D}	0,25	420
{C, D}	0,5	80

En moyenne, l'estimateur est aussi sans biais $0{,}25\times2020+0{,}25\times420+0{,}5\times80=650$

Mais sa variance est considérable : elle vaut 644 900 (CV $\cong$ 124 %)

$$0{,}25\times(2020-650)^2+0{,}25\times(420-650)^2+0{,}5\times(80-650)^2=644900$$

Principe des plans à probabilités inégales

Aller chercher l'information là où elle se trouve

⇨ Ce qui suppose de disposer avant l'échantillonnage d'information auxiliaire connue sur toute la population et liée positivement avec le caractère d'intérêt

Cette information auxiliaire, connue pour tous les individus de la base de sondage, permet le calcul de probabilités d'inclusion pertinentes, de sorte que les unités les plus contributives au paramètre d'intérêt aient plus de chances d'être sélectionnées

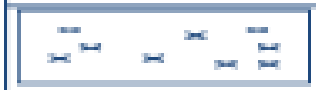
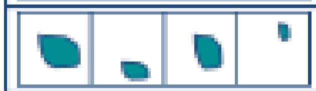
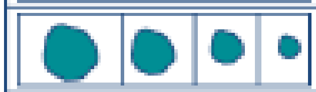
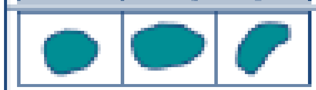
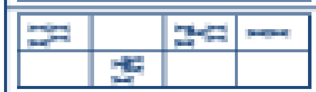
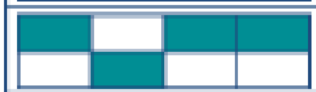
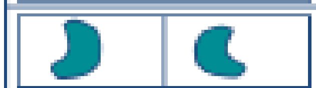
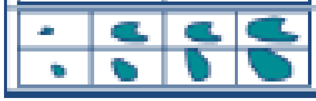
Représentativité ?

Yves Tillé, Théorie des sondages (Dunod, 2001) :

«L'objectif d'un sondage est de fournir un certain nombre d'informations sur une population en n'examinant qu'une partie de celle-ci, appelée échantillon. On dit souvent qu'un échantillon est représentatif d'une population s'il en constitue le modèle réduit. La représentativité est ainsi évoquée en tant qu'argument de validité : un bon échantillon devrait ressembler autant que possible à la population à étudier de sorte que certaines catégories apparaissent en mêmes proportions dans l'échantillon et la population. Pourtant cette théorie, couramment véhiculée par les médias et même par certains ouvrages de méthodologie est erronée. Il est en effet souvent souhaitable de d'effectuer des tirages à probabilités inégales ou de sur représenter certaines fractions de la population. Pour estimer avec précision un paramètre, il faut aller chercher l'information de manière judicieuse plutôt que d'accorder la même importance à chaque unité»

Comparatif des plans de sondage classiques

- Pascal Ardilly, les techniques de sondage, Dunod, 2006

	Plan de sondage	Réalisation du tirage et estimation	précision	coût terrain
	Sondage aléatoire simple	=	=	=
	Sondage stratifié allocation quelconque	-	+	=
	Sondage stratifié alloc proportionnelle	-	+	=
	Sondage stratifié allocation optimale	--	+++	=
	Sondage 'quelconque' à plusieurs degrés	--	-	+
	Sondage en grappes	-	--	++
	Sondages à probabilité inégales	-	Si Yi proportionnel à Xi ++, sinon --	=
	Sondage équilibré	---	+++	=
	Sondage par quota	-	?	++

Sommaire

1) Pourquoi échantillonner à probabilités inégales ?
2) **Comment choisir les probabilités d’inclusion ?**
3) Comment estimer un total ?
4) Comment estimer une moyenne ?
5) Comment construire des intervalles de confiance pour le total ou la moyenne ?
6) Application dans Sas et dans R

Comment choisir les probabilités d’inclusion ?

Cas des plans proportionnels à la taille (ppt)

- Exemple d’une enquête qui s’intéresse au chiffre d’affaires des entreprises d’un secteur donné
  - Si l’on dispose du nombre de salariés de toutes les entreprises de ce secteur
  - Et si on pressent que le chiffre d’affaires est plus ou moins proportionnel au nombre de salariés

Alors il est légitime de calculer les probabilités d’inclusion de toutes les entreprises de manière proportionnelle à leur effectif salarié

Probabilités d'inclusion proportionnelles à la taille

Pour un échantillon de taille fixe n, la probabilité de sélectionner la k[ème] entreprise se calcule ainsi :

$$\forall\, k \in U\,,\quad \pi_k = P(k \in S) = n \frac{X_k}{\sum_{k \in U} X_k}$$

où X_k désigne la variable de taille, ici le nombre de salariés de la k[ème] entreprise de la population U

On vérifie que : $\sum_{k \in U} \pi_k = n$ (plan de taille fixe)

Probabilités d'inclusion proportionnelles à la taille

Attention : l'expression précédente peut conduire à des probabilités supérieures à 1

- Dans ce cas, on sélectionne d'office les unités en question : $\pi_k = 1$ (strate dite « exhaustive »)
- On recalcule les probabilités d'inclusion des autres individus proportionnellement à :
  - la taille de l'échantillon restant à sélectionner
  - et à leur contribution dans le total diminué des valeurs des unités sélectionnées d'office
- Quitte à réitérer la démarche jusqu'à ce que $\pi_k \leq 1, \ \forall\, k \in U$

Exemple

- Population de 4 entreprises A, B, C et D de 500, 100, 30 et 20 salariés
- Sélection de 2 unités proportionnellement à leur effectif salarié
  - A est sélectionnée d'office : $\pi_A = 1$
  - Reste à choisir une entreprise, proportionnellement à son effectif salarié dans la population privée de A

$$\pi_B = 1 \times \frac{100}{150} = \frac{2}{3} \quad ; \quad \pi_C = 1 \times \frac{30}{150} = \frac{1}{5} \quad ; \quad \pi_D = 1 \times \frac{20}{150} = \frac{2}{15}$$

  - On vérifie $\sum_{k \in U} \pi_k = \pi_A + \pi_B + \pi_C + \pi_D = n = 2$

Notations dans la population

- Population finie *U* de *N* objets identifiables (ou individus, unités statistiques) : $U = \{1, 2, ..., k, ..., N\}$
- Variable d'intérêt *Y* de caractéristique individuelle Y_k
- Total : $T_Y = \sum_{k \in U} Y_k$
- Moyenne : $\bar{Y} = \frac{T_Y}{N} = \frac{1}{N} \sum_{k \in U} Y_k$

Notations dans l'échantillon

- Échantillon *s* : sous-ensemble de *U* de taille *n*
- Ensemble des échantillons possibles : $\mathcal{S}$
- Plan de sondage probabiliste : loi de probabilité sur $\mathcal{S}$

$$p(s) \geq 0,\ \forall s \in S,\ \text{ et } \sum_{s \in S} p(s) = 1$$

- Probabilité d'inclusion d'ordre un de *k*

$$\pi_k = P(k \in s) = \sum_{k \in s} p(s) = E(I_k) \qquad I_k = \begin{cases} 1 & k \in S \\ 0 & k \notin S \end{cases}$$

- Probabilité d'inclusion ou double de *k* et *l* (*k* ≠ *l*) :

$$\pi_{kl} = p(k \in s, l \in s) = \sum_{k, l \in s} p(s) = E(I_k I_l)$$

- Covariance entre I_k et I_l pour *k* et *l* :

$$\Delta_{kl} = E(I_{kl}) - E(I_k) E(I_l) = \pi_{kl} - \pi_k \pi_l$$

Sommaire

1) Pourquoi échantillonner à probabilités inégales ?
2) Comment choisir les probabilités d'inclusion ?
3) **Comment estimer un total ?**

   L'estimateur de Horvitz-Thompson

4) Comment estimer une moyenne ?
5) Comment construire des intervalles de confiance pour le total ou la moyenne ?
6) Application dans Sas et dans R

Comment estimer un total ?

- En 1952, **Horvitz et Thompson** ont proposé l'estimateur suivant du total T_y de la variable *Y* :

$$\boxed{\hat{t}_{y\pi} = \sum_{k \in s} \frac{Y_k}{\pi_k}}$$

- On l'appelle aussi **π-estimateur ou estimateur par les valeurs dilatées**

Remarques sur le π-estimateur

- C’est un estimateur linéaire
- Les poids de sondage ne dépendent pas de l’échantillon
- Il permet d'estimer la taille *N* de la population, qu'elle soit connue ou non, en faisant la **somme des poids de sondage** :

$$\hat{N}_{\pi} = \sum_{k \in s} \frac{1}{\pi_k}$$

- Il est valable quel que soit le plan de sondage
- Il généralise les résultats du sondage aléatoire simple sans remise de taille fixe *n*, où $\pi_k = \frac{n}{N}$ pour tout *k* de *U*

le cnam

Espérance du π-estimateur de T_y

- **Si $\pi_k > 0$ pour tout individu *k* de la population *U***, alors l'estimateur d'Horvitz-Thompson du total est sans biais

$$E\left(\hat{T}_{y\pi}\right) = T_y$$

- Si certaines probabilités d'inclusion sont nulles, alors l'estimateur est biaisé
  - Ce biais ne dépend que des unités qui n'ont aucune chance d'être échantillonnées : on parle de problème de couverture (sous-couverture)

Variance du π-estimateur de T_y

Dans le cas général

Si $\pi_k > 0$ pour tout *k* de *U*, alors la variance de l'estimateur d'Horvitz-Thompson du total vaut :

$$\boxed{\mathrm{V}ar\left(\hat{t}_{y\pi}\right) = \sum_{k \in U} \sum_{l \in U} \frac{Y_k}{\pi_k} \frac{Y_l}{\pi_l} \Delta_{kl}}$$

Variance du π-estimateur de T_y

Dans le cas d'un plan de taille fixe

Si $\pi_k > 0$ pour tout k de U et si le plan est de taille fixe, alors Sen, Yates et Grundy ont montré que la variance de l'estimateur d'Horvitz-Thompson du total peut aussi s'écrire :

$$\boxed{Var\left(\hat{t}_{y\pi}\right) = -\frac{1}{2}\sum_{k\in U}\sum_{l\in U}\left(\frac{Y_k}{\pi_k}-\frac{Y_l}{\pi_l}\right)^2 \Delta_{kl}}$$

le cnam

Comment estimer la variance du π-estimateur de T_y ?

Cas général : Si $\pi_{kl} > 0$ pour tous k et l de U, alors la variance de l'estimateur d'Horvitz-Thompson du total peut être estimée sans biais par :

$$\boxed{\hat{V}ar_1\left(\hat{t}_{y\pi}\right)=\sum_{k\in s}\sum_{l\in s}\frac{Y_k}{\pi_k}\frac{Y_l}{\pi_l}\frac{\Delta_{kl}}{\pi_{kl}}}$$

Cas d'un plan de taille fixe : Si $\pi_{kl} > 0$ pour tous k et l de U et si le plan est de taille fixe, alors la variance de l'estimateur d'Horvitz-Thompson du total peut aussi être estimée sans biais par :

$$\boxed{\hat{V}ar_2\left(\hat{t}_{y\pi}\right)=-\frac{1}{2}\sum_{k\in s}\sum_{l\in s}\left(\frac{Y_k}{\pi_k}-\frac{Y_l}{\pi_l}\right)^2\frac{\Delta_{kl}}{\pi_{kl}}}$$

Comparaison de ces 2 estimateurs

- Le 1er estimateur est valable dans le cas général
- Le 2ème estimateur n'est valable **que si le plan est de taille fixe**
- Dans le cas où est le plan est de taille fixe, on dispose généralement de 2 estimateurs concurrents et différents
- Ils sont tous les 2 sans biais dès que tous les π_{kl} sont strictement positifs quels que soient les individus *k* et *l* de la population
- Les 2 estimateurs peuvent prendre des valeurs négatives, mais il existe une condition suffisante pour que le second estimateur soit positif. Cette condition, dite **condition de Sen-Yates-Grundy,** est

$$\forall\, k \neq l \in U \quad \Delta_{kl} \leq 0 \quad \text{soit :} \quad \pi_{kl} - \pi_k \pi_l \leq 0$$

Sommaire

Comment estimer une moyenne ?

Lorsque la taille de la population est connue, on peut estimer la moyenne de *Y* avec l'estimateur de Horvitz-Thompson :

$$\boxed{\hat{\bar{Y}}_{\pi} = \frac{1}{N} \sum_{k \in s} \frac{Y_k}{\pi_k} = \frac{\hat{t}_{y\pi}}{N}}$$

Dans le cas particulier où la variable d'intérêt *Y* est dichotomique et vaut 1 dans *p* % des cas, le π-estimateur de la proportion *p* se calcule pareillement :

$$\hat{p}_{\pi} = \frac{1}{N} \sum_{k \in s} \frac{Y_k}{\pi_k}$$

Propriétés du π-estimateur de la moyenne

- Les propriétés vues pour l'estimateur de Horvitz-Thompson d'un total s'appliquent à une moyenne, en adaptant les formules pour tenir compte de la taille de la population
- Estimateur sans biais : $E\left(\hat{\bar{Y}}_\pi\right)=\bar{Y}$
- De variance :

$$Var\left(\hat{\bar{Y}}_\pi\right)=\frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\frac{Y_k}{\pi_k}\frac{Y_l}{\pi_l}\Delta_{kl}$$

$$Var\left(\hat{\bar{Y}}_\pi\right)=-\frac{1}{2N^2}\sum_{k\in U}\sum_{l\in U}\left(\frac{Y_k}{\pi_k}-\frac{Y_l}{\pi_l}\right)^2\Delta_{kl}$$

Si plan de taille fixe

- Estimée par :

$$\hat{V}ar_1\left(\hat{\bar{Y}}_\pi\right)=\frac{1}{N^2}\sum_{k\in s}\sum_{l\in s}\frac{Y_k}{\pi_k}\frac{Y_l}{\pi_l}\frac{\Delta_{kl}}{\pi_{kl}}$$

$$\hat{V}ar_2\left(\hat{\bar{Y}}_\pi\right)=-\frac{1}{2N^2}\sum_{k\in s}\sum_{l\in s}\left(\frac{Y_k}{\pi_k}-\frac{Y_l}{\pi_l}\right)^2\frac{\Delta_{kl}}{\pi_{kl}}$$

L'estimateur de Hájek de la moyenne

Lorsque la taille de la population est inconnue, on peut estimer la moyenne de *Y* avec l'estimateur de Hájek (1971) :

$$\boxed{\hat{\bar{Y}}_H = \frac{\hat{t}_{y\pi}}{\hat{N}_\pi} = \frac{\sum_{k\in s} \frac{Y_k}{\pi_k}}{\sum_{k\in s} \frac{1}{\pi_k}}}$$

le cnam

Remarques sur l'estimateur de Hájek

- L'estimateur de Hájek est un ratio de 2 π- estimateurs
- Il est biaisé : on montre que le biais est asymptotiquement nul et on le considéra négligeable lorsque la taille de l'échantillon est grande
- Sa précision peut s'avérer supérieure à celle de l'estimateur d'Horvitz-Thompson
- Si $\forall\, k \in U, \quad Y_k = a$ (a constante réelle quelconque), alors :

$$\hat{\bar{Y}}_H = a = \bar{Y}$$

$$\hat{\bar{Y}}_\pi = a\,\frac{\hat{N}_\pi}{N}$$

le cnam

Exemple de Basu

- Basu (1971) relate l'exemple suivant : le propriétaire d'un cirque souhaite estimer le poids total de son troupeau de 50 éléphants
- Il y a 3 années de cela Sambo représentait le poids moyen du troupeau
- En présumant que c'est toujours le cas, il propose de peser à nouveau Sambo et d'en déduire l'estimation suivante du poids du troupeau :

$$50 \times (\text{Poids de Sambo})$$

Exemple de Basu

- Le statisticien du cirque propose quant à lui d'estimer sans biais le poids total en échantillonnant un éléphant à probabilités inégales avec :

$$\begin{cases} \pi_{\text{Sambo}} = {}^{99}\!/_{100} = 0{,}99 \\ \pi_{\text{Autres}} = {}^{1}\!/_{100} \times {}^{1}\!/_{49} = {}^{1}\!/_{4900} \end{cases}$$

- et en calculant :

Poids de Sambo / 0,99	si Sambo est échantillonné
Poids de l'éléphant retenu $\times$ 4900	sinon

- Cela conduit certes à un estimateur sans biais, mais chaque estimation est irréaliste

Sommaire

1) Pourquoi échantillonner à probabilités inégales ?
2) Comment choisir les probabilités d'inclusion ?
3) Comment estimer un total ?
4) Comment estimer une moyenne ?
5) **Comment construire des intervalles de confiance pour le total ou la moyenne ?**
6) Application dans Sas et dans R

Comment construire des intervalles de confiance ?

- On considère que le π-estimateur suit approximativement une loi normale lorsque la taille de l'échantillon est grande

- L'intervalle de confiance au niveau de confiance *1-*α pour le total T_y est donc donné par :

$$IC_{1-\alpha}(t_y)=\left[\hat{t}_{y\pi}-u_{1-\frac{\alpha}{2}}\sqrt{\hat{V}ar(\hat{t}_{y\pi})};\hat{t}_{y\pi}+u_{1-\frac{\alpha}{2}}\sqrt{\hat{V}ar(\hat{t}_{y\pi})}\right]$$

où $u_{1-\alpha/2}$ désigne le fractile d'ordre 1 - α/2 de la loi N(0,1)

- De même, l'intervalle de confiance pour la moyenne est :

$$IC_{1-\alpha}(\bar{Y})=\left[\hat{\bar{Y}}_{\pi}-u_{1-\frac{\alpha}{2}}\sqrt{\hat{V}ar(\hat{\bar{Y}}_{\pi})};\hat{\bar{Y}}_{\pi}+u_{1-\frac{\alpha}{2}}\sqrt{\hat{V}ar(\hat{\bar{Y}}_{\pi})}\right]$$

- En général, on construit des intervalles de confiance à 95 %. Le fractile considéré vaut donc 1,96

Sommaire

1) Pourquoi échantillonner à probabilités inégales ?
2) Comment choisir les probabilités d’inclusion ?
3) Comment estimer un total ?
4) Comment estimer une moyenne ?
5) Comment construire des intervalles de confiance pour le total ou la moyenne ?
6) **Application dans Sas et dans R**

Un exemple

- Exemple (P. Ardilly) : soit une population de 5 communes, Y est le nombre d'habitants et X le nombre de logements
- Estimation du nombre moyen d'habitants (supposé inconnu) par tirage à probabilités proportionnelles au nombre de logements (connu)
- On a les moyens d'échantillonner 2 communes

Communes	Nombre de logements = X	Nombre d'habitants = Y	Probabilité d'inclusion
(1) Antibes	48 812	70 688	0,99
(2) Cagnes	23 227	41 303	0,47
(3) St Laurent du Var	12 383	24 475	0,25
(4) Vence	9 341	15 364	0,19
(5) Villefranche/Mer	4 915	8 123	0,10
Moyenne	19 736	31 991	–

On vérifie bien que la somme des probabilités d'inclusion sur la population vaut 2

Application dans R

```
> n <- 2
> N <- 5
> communes <- c('Antibes', 'Cagnes', 'St laurent du Var', 'Vence', 'Villefranche sur
')
> ID <- c(1,2,3,4,5)
> X <- as.numeric(c(48812,23227,12383,9341,4915))
> Y <- as.numeric(c(70688, 41303, 24475,15364,8123))
> Pi <- as.numeric(c(0.99,0.47,0.25,0.19,0.10))
> YInvPi_PPS <- Y * 1/Pi
> YInvPi_SAS <- Y * N/n
> dataFile <- data.frame(communes,X,Y,Pi,YInvPi_PPS, YInvPi_SAS)
> dataFile
             communes     X     Y   Pi YInvPi_PPS YInvPi_SAS
1             Antibes 48812 70688 0.99   71402.02   176720.0
2              Cagnes 23227 41303 0.47   87878.72   103257.5
3   St laurent du Var 12383 24475 0.25   97900.00    61187.5
4               Vence  9341 15364 0.19   80863.16    38410.0
5 Villefranche sur Mer  4915  8123 0.10   81230.00    20307.5
> sum(dataFile$Pi)
[1] 2
> samples <- combn(ID,2)
> samples
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    1    1    1    1    2    2    2    3    3     4
[2,]    2    3    4    5    3    4    5    4    5     5
> dataFile[samples[,1],]
  communes     X     Y   Pi YInvPi_PPS YInvPi_SAS
1  Antibes 48812 70688 0.99   71402.02   176720.0
2   Cagnes 23227 41303 0.47   87878.72   103257.5
```

Application dans R

```
> 1/N*sum(dataFile[samples[,1],]$YInvPi_PPS)
[1] 31856.15
> estimations <- matrix(0,nrow = 10, ncol = 2)
> for (i in 1:10) {
+   estimations[i,1] <- 1/N * sum(dataFile[samples[,i],]$YInvPi_SAS)
+   estimations[i,2] <- 1/N * sum(dataFile[samples[,i],]$YInvPi_PPS)
+ }
> colnames(estimations) <- c("SAS", "PPS")
> cbind(t(samples), estimations)
               SAS      PPS
 [1,] 1 2 55995.5 31856.15
 [2,] 1 3 47581.5 33860.40
 [3,] 1 4 43026.0 30453.04
 [4,] 1 5 39405.5 30526.40
 [5,] 2 3 32889.0 37155.74
 [6,] 2 4 28333.5 33748.38
 [7,] 2 5 24713.0 33821.74
 [8,] 3 4 19919.5 35752.63
 [9,] 3 5 16299.0 35826.00
[10,] 4 5 11743.5 32418.63
```

Application dans SAS

Tirage d'un échantillon

PROC SURVEYSELECT

DATA = *base de sondage*

METHOD = PPS

SAMPSIZE = *taille de l'échantillon*

OUT = *table échantillon*

OUTSIZE ;

SIZE *variable de taille* ;

STRATA *variable(s) de stratification* ;

ID *liste de variables* ;

RUN;

Application dans SAS

Estimation des paramètres

```
PROC SURVEYMEANS
          DATA = table échantillon
          RATE=taux de sondage
          TOTAL=effectif échantillon
          MISSING
           statistiques ;
       WEIGHT variable de pondération ;
       VAR    liste de variables à estimer ;
       CLASS variable(s) catégorielle(s) à estimer ;
       STRATA variable(s) de stratification ;
RUN;
```

le cnam