STA108 – Plans stratifiés

Cours n°4 du 16/10/2020

Sylvie Rousseau

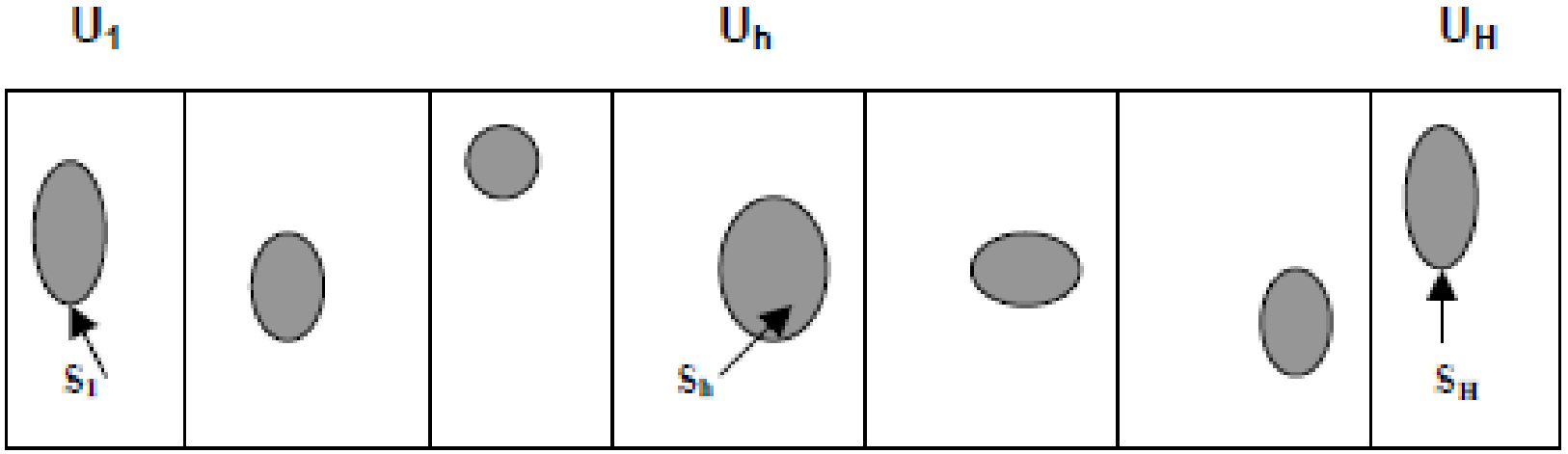
le cnam

Sommaire

1. **Principe et intérêt**
2. Estimation d'un total et d'une moyenne
3. Précision des estimateurs
4. Choix des allocations
5. Considérations pratiques
6. Application dans R et dans SAS

Principe d'un plan stratifié

- Partitionner la population en sous-groupes appelés strates
- Sélectionner un échantillon dans chaque strate, de manière indépendante d'une strate à l'autre



$$S = \bigcup_{h=1}^{H} S_h$$

le cnam

Intérêt d'un plan stratifié

- Objectif : gagner en précision par rapport à un plan simple
  - *Exemple : estimer le revenu moyen en enquêtant des individus de chaque catégorie socio-professionnelle plutôt que de choisir l'échantillon dans la population entière*
- Moyen : disposer d'une information auxiliaire (connue sur toute la population) corrélée positivement à la variable d'intérêt

Exemple : estimer la surface moyenne d'exploitations agricoles

- Supposons N=5 et les vraies valeurs connues y_1=10, y_2=30, y_3=40, y_4=60, y_5=110 ha
- On a $\overline{Y} = 50$ ha et $S_y^2 = 1450$ ha²
- En distinguant U_1={1,2,3} et U_2={4,5} et avec n_1=n_2=1

Echantillon s	*{1,4}*	*{1,5}*	*{2,4}*	*{2,5}*	*{3,4}*	*{3,5}*
Moyenne $\hat{\bar{y}}$	30	50	42	62	48	68

$$E\left(\hat{\bar{Y}}\right) = \bar{Y} = 50 \qquad Var\left(\hat{\bar{Y}}\right) = 156 \simeq 12,5^2$$

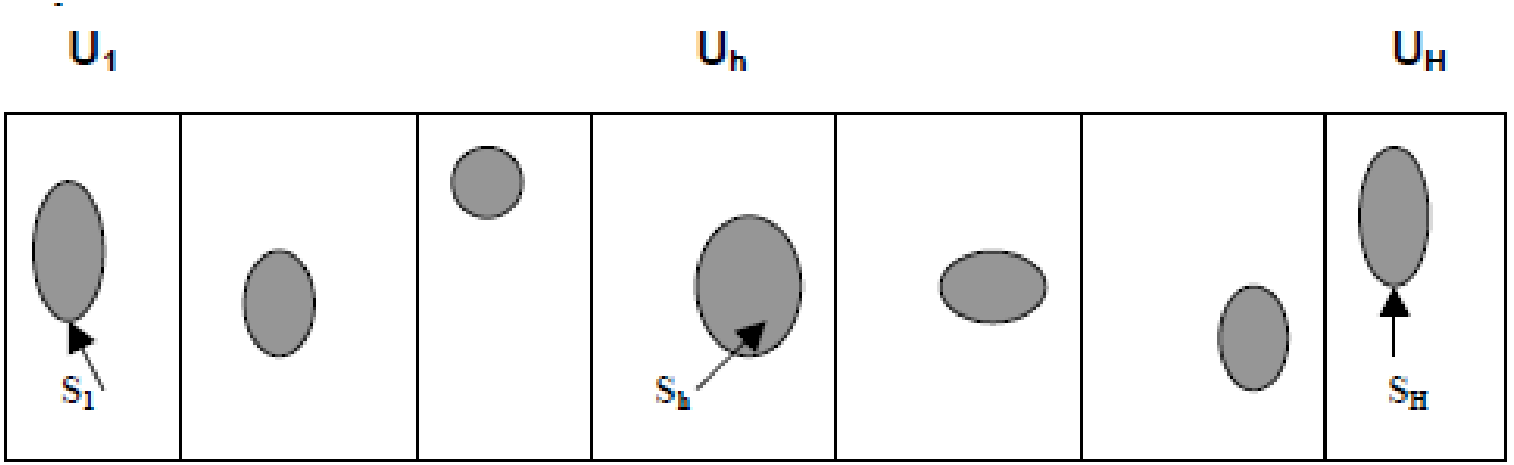
- Avec un plan simple sans remise de taille 2, on a des échantillons « non désirables » d'où une imprécision plus grande

$$Var_{sas,sr}\left(\hat{\bar{Y}}\right) = 435 \simeq 20,9^2 \quad \Rightarrow \quad Deff = \frac{Var\left(\hat{\bar{Y}}\right)}{Var_{sas,sr}\left(\hat{\bar{Y}}\right)} = 0,358$$

Sommaire

1. Principe et intérêt
2. **Estimation d'un total et d'une moyenne**
3. Précision des estimateurs
4. Choix des allocations
5. Considérations pratiques
6. Application dans R et dans SAS

Notations



$$U = \bigcup_{h=1}^{H} U_h \qquad N = \sum_{h=1}^{H} N_h$$

$$S = \bigcup_{h=1}^{H} S_h \qquad n = \sum_{h=1}^{H} n_h$$

Les tirages sont indépendants d'une strate à l'autre.

Dans la suite, on considère que le tirage dans chaque strate obéit à un plan simple sans remise.

Notations sur la population

- Total de la variable Y : $T_y = \sum_{h=1}^{H} T_{yh} = \sum_{h=1}^{H} N_h \bar{Y}_h$

  avec $\bar{Y}_h = \frac{1}{N_h} \sum_{k \in U_h} Y_k$

- Moyenne : $\bar{Y} = \frac{T_Y}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{Y}_h$

- Variance et dispersion :

$$\sigma_y^2 = \frac{1}{N} \sum_{k \in U} \left(Y_k - \bar{Y} \right)^2 = \frac{N-1}{N} S_y^2$$

$$= \sum_{h=1}^{H} \frac{N_h}{N} \sigma_{yh}^2 + \sum_{h=1}^{H} \frac{N_h}{N} \left(\bar{Y}_h - \bar{Y} \right)^2 \quad \text{avec} \quad \sigma_{yh}^2 = \frac{1}{N_h} \sum_{k \in U_h} \left(Y_k - \bar{Y}_h \right)^2$$

$$= \sigma_{y\,\text{intra}}^2 + \sigma_{y\,\text{inter}}^2$$

Estimation d'un total

- Estimateur d'Horvitz-Thompson :

$$\hat{T}_y = N\,\hat{\bar{Y}} = \sum_{h=1}^{H} \hat{T}_{yh} = \sum_{h=1}^{H} N_h\,\hat{\bar{Y}}_h$$

$$= \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in S_h} Y_k$$

$$= \sum_{k \in S} \frac{Y_k}{\pi_k}$$

avec $\pi_k = \frac{n_h}{N_h}$ pour $k \in U_h$

si plan simple sans remise dans U_h

- Sans biais

(car combinaison linéaire d'estimateurs sans biais)

Estimation d'une moyenne

- Estimateur d'Horvitz-Thompson

$$\hat{\bar{Y}} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{\bar{Y}}_h \qquad avec \quad \hat{\bar{Y}}_h = \frac{1}{n_h} \sum_{k \in s} Y_k$$

- Sans biais
- A distinguer de la moyenne empirique

$$\hat{\bar{Y}}_{emp} = \frac{1}{n} \sum_{k \in s} Y_k = \sum_{h=1}^{H} \frac{n_h}{n} \hat{\bar{Y}}_h$$

Estimation d'une proportion

- Cas particulier d'une moyenne
- Estimateur d'Horvitz-Thompson

$$\hat{p} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{p}_h$$

- Sans biais

Exemple

- But : estimer le nombre moyen de patients par médecin et par jour
- Données :

Strate h	N_h	n_h	$\hat{\bar{y}}_h$	$\hat{S}^2_{yh}$
1(débutants)	500	200	10	4
2 (confirmés)	1000	200	15	7
3 (expérimentés)	2500	200	20	10
Ensemble	4000	600		

- Résultat :

$$\frac{500}{4000}10+\frac{1000}{4000}15+\frac{2500}{4000}20=17{,}5 \text{ visites}$$

Sommaire

1. Principe et intérêt
2. Estimation d'un total et d'une moyenne
3. **Précision des estimateurs**
4. Choix des allocations
5. Considérations pratiques
6. Application dans R et dans SAS

Variance de l'estimateur d'une moyenne

- Comme les tirages sont indépendants :

$$Var\left[\hat{\bar{Y}}\right] = Var\left[\sum_{h=1}^{H} \frac{N_h}{N} \hat{\bar{Y}}_h\right] = \sum_{h=1}^{H} Var\left[\frac{N_h}{N} \hat{\bar{Y}}_h\right]$$

- Avec un plan simple sans remise dans chaque strate :

$$Var\left[\hat{\bar{Y}}\right] = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{Yh}^2}{n_h}$$

Variance de l'estimateur d'un total ou d'une proportion

- Pour un total

$$Var\left[\hat{T}_y\right] = Var\left[N\,\hat{\bar{Y}}\right] = N^2\,Var\left[\hat{\bar{Y}}\right] = \sum_{h=1}^{H} N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{S_{Yh}^2}{n_h}$$

- Pour une proportion

$$Var\left[\hat{p}\right] = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(1-\frac{n_h}{N_h}\right)\frac{N_h}{N_h-1}\,\frac{p_h\left(1-p_h\right)}{n_h}$$

Estimation de la variance de l'estimateur d'une moyenne

- Avec un plan simple sans remise dans chaque strate

$$\hat{V}ar\left[\hat{\bar{Y}}\right]=\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(1-\frac{n_h}{N_h}\right)\frac{\hat{S}_{Yh}^2}{n_h}$$

où $$\hat{S}_{Yh}^2=\frac{1}{n_h-1}\sum_{k\in S_h}\left(Y_k-\hat{\bar{Y}}_h\right)^2$$

- Estimateur sans biais

Estimation de la variance de l'estimateur d'un total ou d'une proportion

- Pour un total,

$$\hat{V}ar\left[\hat{T}_y\right] = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{S}_{Yh}^2}{n_h}$$

- Pour une proportion,

$$\hat{V}ar\left[\hat{p}\right] = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{p}_h\left(1-\hat{p}_h\right)}{n_h - 1}$$

Exemple

- Données :

Strate h	N_h	n_h	$\hat{\bar{y}}_h$	$\hat{S}^2_{yh}$
1(débutants)	500	200	10	4
2 (confirmés)	1000	200	15	7
3 (expérimentés)	2500	200	20	10
Ensemble	4000	600		

- Estimation de la variance :

$$\widehat{Var}\left[\hat{\bar{Y}}\right]=\left(\frac{500}{4000}\right)^2\left(1-\frac{200}{500}\right)\frac{4}{200}+\left(\frac{1000}{4000}\right)^2\left(1-\frac{200}{1000}\right)\frac{7}{200}+\left(\frac{2500}{4000}\right)^2\left(1-\frac{200}{2500}\right)\frac{10}{200}$$

$$\widehat{Var}\left[\hat{\bar{Y}}\right]\simeq 0,0199$$

$$IC_{95\%}\left[\bar{Y}\right]=\left[17,5\pm 0,28\right]$$

Sommaire

1. Principe et intérêt
2. Estimation d'un total et d'une moyenne
3. Précision des estimateurs
4. **Choix des allocations**
5. Considérations pratiques
6. Application dans R et dans SAS

Allocations proportionnelles

- Échantillon dit « représentatif »

$$\frac{n_h}{n} = \frac{N_h}{N} \quad \forall h \in \{1, \ldots H\}$$

- Ce qui revient à un taux de sondage constant :

$$\frac{n_h}{N_h} = \frac{n}{N} = f_h = f \quad \forall h \in \{1, \ldots H\}$$

Allocations proportionnelles

- Estimateur de la moyenne : $\hat{\bar{Y}} = \hat{\bar{Y}}_{emp} = \sum_{h=1}^{H} \frac{n_h}{n} \hat{\bar{Y}}_h$
- De variance :

$$Var_{prop}\left[\hat{\bar{Y}}\right] = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{Yh}^2}{n_h} = \left(1 - \frac{n}{N}\right) \frac{1}{n} \sum_{h=1}^{H} \frac{N_h}{N} S_{Yh}^2$$

- Si $\forall$h, N_h est grand, alors

$$Var_{prop}\left[\hat{\bar{Y}}\right] \simeq \left(1 - \frac{n}{N}\right) \frac{1}{n} \sum_{h=1}^{H} \frac{N_h}{N} \sigma_{Yh}^2 = \left(1 - \frac{n}{N}\right) \frac{1}{n} \sigma_{Y\,\mathrm{int}\,ra}^2$$

Allocations proportionnelles

- Avec un plan simple sans remise de même taille dans U

$$Var_{SAS,SR}\left[\hat{\bar{Y}}\right]=\left(1-\frac{n}{N}\right)\frac{1}{n}S_y^2$$

- Gain de précision :

$$Var_{prop}\left[\hat{\bar{Y}}\right]\leq Var_{SAS,SR}\left[\hat{\bar{Y}}\right]$$

Mesuré par l'effet de sondage : $Deff=\frac{Var_{prop}\left(\hat{\bar{Y}}\right)}{Var_{SAS,SR}\left(\hat{\bar{Y}}\right)}\simeq\frac{\sigma^2_{Y\,\text{intra}}}{\sigma^2_Y}$

- **Règle : Constituer des strates homogènes en intra au regard de la variable d'intérêt (ou, ce qui revient au même, hétérogènes en inter)**

Allocations optimales

- But : choix des $(n_h)_{h=1,..,H}$ qui assurent la précision maximale
- Problème d'optimisation sous contrainte :

$$\underset{n_h}{Min}\left(Var\left[\hat{\bar{Y}}\right]\right)=\underset{n_h}{Min}\left[\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(\frac{1}{n_h}-\frac{1}{N_h}\right)S_{Yh}^2\right] \qquad s.c. \qquad n=\sum_{h=1}^{H} n_h$$

- Solution : allocation optimale de Neyman

$$n_h=n\frac{N_h S_{Yh}}{\sum_{l=1}^{H} N_l S_{Yl}}$$

- Précision maximale obtenue :

$$Var_{opt}\left[\hat{\bar{Y}}\right]=\frac{1}{n}\left(\sum_{h=1}^{H}\frac{N_h}{N}S_{Yh}\right)^2-\frac{1}{N}\sum_{h=1}^{H}\frac{N_h}{N}S_{Yh}^2$$

Allocations optimales sous contrainte budgétaire

- Avec un budget total C et un coût unitaire d'enquête C_h dans la strate h

$$\underset{n_h}{Min}\left(Var\left[\hat{\bar{Y}}\right]\right)=\underset{n_h}{Min}\left[\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(\frac{1}{n_h}-\frac{1}{N_h}\right)S_{Yh}^2\right] \quad s.c. \quad C=\sum_{h=1}^{H} n_h C_h$$

- Solution :

$$n_h=C\frac{N_h S_{yh}}{\sqrt{C_h}\sum_{l=1}^{H} N_l S_{yl}\sqrt{C_l}}$$

Exemple

- Calcul des allocations :

Strate h	N_h	n_h	$\hat{\bar{y}}_h$	$\hat{S}^2_{yh}$	$n_{h,prop}$	$n_{h,opt}$
1(débutants)	500	200	10	4	75	52
2 (confirmés)	1000	200	15	7	150	137
3 (expérimentés)	2500	200	20	10	375	411
Ensemble	4000	600			600	600

- Estimation de la variance et de l'effet de sondage :

$$\hat{V}ar_{prop}\left[\hat{\bar{\bar{Y}}}\right] \simeq 0,0120 \qquad \hat{V}ar_{SAS,SR}\left[\hat{\bar{\bar{Y}}}\right] \simeq 0,0297 \qquad \hat{S}^2_y \simeq 20,98$$

$$\hat{V}ar_{opt}\left[\hat{\bar{\bar{Y}}}\right] \simeq 0,0118 \qquad \hat{D}eff_{prop} = \frac{\hat{V}ar_{prop}\left[\hat{\bar{\bar{Y}}}\right]}{\hat{V}ar_{SAS,SR}\left[\hat{\bar{\bar{Y}}}\right]} \simeq 40\,\%$$

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. Principe et intérêt
2. Estimation d'un total et d'une moyenne
3. Précision des estimateurs
4. Choix des allocations
5. **Considérations pratiques**
6. Application dans R et dans SAS

Choix des allocations

- Optimales
  - Si une variable d'intérêt prédomine
  - Si les dispersions S^2_h sont connues ou peuvent être approchées via :
    - Une variable auxiliaire liée à la variable d'intérêt
    - Une enquête passée
    - Une pré-enquête
    - Une approximation du type $\frac{S_{yh}}{\bar{Y}_h} = cste$

d'où $n_h = n\frac{T_{yh}}{T_y} \propto n\frac{T_{xh}}{T_x} \quad avec \; Y \propto X$

- Proportionnelles
  - si les dispersions S_h sont inconnues
  - si les variables d'intérêt sont très variées

Choix des strates

- Constituer des strates homogènes en intra au regard de la variable d'intérêt
- Variable(s) de stratification bien corrélée(s) à la variable d'intérêt
- Nombre maximal de strates mais … il faut avoir suffisamment d'unités dans chaque strate sinon l'estimateur est instable
- Limites des strates
  - Faire coïncider les strates et les domaines d'intérêt
  - Si le critère de stratification est quantitatif, méthode de Dalenius et Hodges

Sommaire

1. Principe et intérêt
2. Estimation d'un total et d'une moyenne
3. Précision des estimateurs
4. Choix des allocations
5. Considérations pratiques
6. **Application dans R et dans SAS**

Application dans SAS

Allocation proportionnelle

PROC SORT DATA= *table base de sondage*;
BY *variables de stratification* ;
PROC SURVEYSELECT
DATA = *base de sondage*
SAMPRATE = *f*
METHOD = SRS
OUT = *table échantillon*
OUTSIZE ;
STRATA *variables de stratification / LIST;*
ID *liste de variables* ;
RUN;

Application dans SAS

Allocation fixe par strate

```
PROC  SURVEYSELECT
                DATA = base de sondage
                SAMPSIZE = n
                METHOD   = SRS
                OUT = table échantillon
                OUTSIZE ;
    STRATA variables de stratification / LIST  ;
    ID liste de variables ;
RUN;
```

Application dans SAS

Allocation différenciée par strate

PROC SURVEYSELECT

DATA = *base de sondage*

$$\begin{cases} \text{SAMPSIZE}=(n_1,n_2,\ldots,n_p) \\ \text{SAMPSIZE}=table\ SAS \\ \text{SAMPRATE}=(f_1,f_2,\ldots,f_p) \\ \text{SAMPRATE}=table\ SAS \end{cases}$$

METHOD = SRS

OUT = *table échantillon*

OUTSIZE ;

STRATA *variables de stratification / LIST* ;

ID *liste de variables* ;

RUN;

Application dans SAS

Structure de la table de données secondaires

Avec SAMPSIZE = *table SAS*

Variable de stratification	_NSIZE_
Modalité 1	*n1*
Modalité 2	*n2*
…	…

Avec SAMPRATE = *table SAS*

Variable de stratification	_RATE_
Modalité 1	*f1*
Modalité 2	*f2*
…	…

Application dans SAS

Estimation des paramètres_

PROC SURVEYMEANS

DATA = *table échantillon*

RATE = f
RATE = *table SAS*
TOTAL = N_h
TOTAL = *table SAS*

MISSING

statistiques ;

STRATA *variable(s) de stratification* ;

WEIGHT *variable de pondération* ;

VAR *liste de variables à estimer* ;

CLASS *liste de variables catégorielles à estimer* ;

BY *variable(s) de stratification* ;

RUN;

Application dans SAS

Structure de la table de données secondaires pour le calcul du facteur $(1-f_h)$

avec TOTAL = *table SAS*

Variable de stratification	_TOTAL_
Modalité 1	*N1*
Modalité 2	*N2*
…	…

Avec RATE = *table SAS*

Variable de stratification	_RATE_
Modalité 1	*f1*
Modalité 2	*f2*
…	…

Application sous R

Fonction StrAlloc du package Prac tools (à installer)

`strAlloc` *Allocate a sample to strata*

Description

Compute the proportional, Neyman, cost-constrained, and variance-constrained allocations in a stratified simple random sample.

Usage

```
strAlloc(n.tot = NULL, Nh = NULL, Sh = NULL, cost = NULL, ch = NULL,
     V0 = NULL, CV0 = NULL, ybarU = NULL, alloc)
```

Arguments

`n.tot`	fixed total sample size
`Nh`	vector of population stratum sizes (N_h) or pop stratum proportions (W_h)
`Sh`	stratum unit standard deviations (S_h), required unless `alloc = "prop"`
`cost`	total variable cost
`ch`	vector of costs per unit in stratum h (c_h)
`V0`	fixed variance target for estimated mean
`CV0`	fixed CV target for estimated mean
`ybarU`	population mean of y ($\bar{y}_U$)
`alloc`	type of allocation; must be one of `"prop"`, `"neyman"`, `"totcost"`, `"totvar"`

Application sous R

Un exemple où le calcul des allocations optimales invite à sélectionner davantage d'individus que la taille de la strate !

- Sélectionner toute la strate
- Reprendre le raisonnement sur la population privée de cette strate exhaustive et une taille d'échantillon diminuée en conséquence