Théorie des sondages : cours 2

Camelia Goga

IMB, Université de Bourgogne
e-mail : camelia.goga@u-bourgogne.fr

Master 2 Besançon-2010

Sondage systématique (SY)(1)

Simple à mettre en oeuvre.

• On fixe à l'avance un nombre entier positif **a** appelé **pas d'échantillonnage**.

• On choisit au hasard un élément r appelé **départ aléatoire** parmi les premiers **a** élément de la liste.

• L'échantillon s_r consiste à prendre tous les **a**-ième unités jusqu'à la fin de la liste.

$$N = n \cdot a + c, \quad n = \text{la partie entière de} \quad N/n$$

$$s_r = \{k : k = r + (j-1)a \leq N, j = 1, \ldots, n_s\}$$

où $n_s = n + 1$ si $r \leq c$ et $n_s = n$ si $c \leq r \leq N$.

Exemple : Soit $N = 10$ et on veut $n = 3$. Alors $a = 3$ et $c = 1$ et on choisit r au hasard parmi $1, 2, 3$. Soit $r = 1$ alors l'échantillon est $s_1 = \{1, 4, 7, 10\}$. Si $r = 2$ alors, $s_2 = \{2, 5, 8\}$ et si $r = 3$, $s_3 = \{3, 6, 9\}$.

Sondage systématique (SY)(2)

- Le nombre total d'échantillons est **a** (beaucoup plus petit que dans le cas SAS où BE),

$$\mathcal{S}_{SY} = \{s_1, \ldots, s_{\mathbf{a}}\}.$$

- Le plan est $p(s) = 1/\mathbf{a}$ si $s \in \mathcal{S}_{SY}$ et zéro sinon.
- $\pi_k = 1/\mathbf{a}$ for tous les $k \in U$;
- $\pi_{kl} = 1/a$ si $k \neq l \in s$ et zéro sinon. La condition $\pi_{kl} > 0$ pour tous k, l **n'est pas satisfaite**.
- $U = \cup_{r=1}^{a} s_r$ et $t_y = \sum_U y_k = \sum_{r=1}^{a} t_{s_r}$ avec $t_{s_r} = \sum_{s_r} y_k$.
- $\hat{t}_\pi = \mathbf{a}\mathbf{t_s}$ avec $s \in \mathcal{S}_{SY}$.
- $V_{SY}(\hat{t}_\pi) = \mathbf{a}\sum_{\mathbf{r=1}}^{\mathbf{a}}(\mathbf{t_{s_r}} - \bar{\mathbf{t}})^{\mathbf{2}}$ avec $\bar{t} = \sum_{r=1}^{a} t_{s_r}/\mathbf{a}$.
- Puisque il y a des $\pi_{kl} = 0$, on ne peut pas déduire $\hat{V}_{SY}(\hat{t}_\pi)$.

Sondage systématique (SY) : efficacité et comparaison avec SAS

$V_{SY}(\hat{t}_\pi) = a\sum_{r=1}^{a}(t_{s_r} - \overline{t})^2$ est zéro lorsque les totaux $t_{s_r} = \overline{t}$,

l'efficacité de SY dépend beaucoup de l'ordre de N éléments de U.

Supposons que $N = n \cdot \mathbf{a}$ avec $\mathbf{a}$ entier positif. Alors,
$\hat{t}_\pi = N\sum_{s_r} y_k/n = N\overline{y}_{s_r}$ avec la variance
$V_{SY}(\hat{t}_\pi) = \frac{N^2}{a}\sum_{r=1}^{a}(\overline{y}_{s_r} - \overline{y}_U)^2$. **ANOVA** :

$$\underbrace{\sum_{U}(y_k - \overline{y}_U)^2}_{SST} = \underbrace{\sum_{r=1}^{a}\sum_{s_r}(y_k - \overline{y}_{s_r})^2}_{SSW} + \underbrace{\sum_{r=1}^{a} n(\overline{y}_{s_r} - \overline{y}_U)^2}_{SSB}$$

SST est fixé, alors une diminution de SSW entraîne une majoration de SSB et vice-versa.

$V_{SY}(\hat{t}_\pi) = N{\cdot}SSB \Longrightarrow$ échantillons s_r homogènes implique SY inefficace

Sondage systématique (SY) : efficacité et comparaison avec SAS

Efficacité

$\rho = 1 - \frac{n}{n-1}\frac{SSW}{SST}$ le coefficient de corrélation intra-classes.

$$\rho = 2\frac{\sum_{r=1}^{a}\left[\sum_{k<l}\sum_{s_r}(y_k - \overline{y}_U)(y_l - \overline{y}_U)\right]}{(n-1)(N-1)S_{yU}^2}$$

- ρ interprété comme une mesure de corrélation entre couples d'éléments du même échantillon ;
- $\rho > 0$ si les unités du même échantillon ont des valeurs y_k similaires ;
- $\rho = 1$ si $SSW = 0$ ou homogénéité totale et
- $\rho = -1/(n-1)$ si $SSB = 0$ ou hétérogénéité totale.

Comparaison avec SAS

L'estimateur du total t_Y est le même pour un plan SAS et SY ($N = n \cdot a$) :

- pour un plan SAS : $\hat{t}_\pi = \frac{N}{n} \sum_s y_k$
- pour un plan SY : $\hat{t}_\pi = a \sum_s t_s = \frac{N}{n} \sum_s y_k$

mais les variance ne sont pas les mêmes : on calcule le "design effect" :

$$deff = \frac{V_{SY}(\hat{t}_\pi)}{V_{SAS}(\hat{t}_\pi)} \simeq 1 + (n-1)\rho.$$

- si $\rho \simeq 1$ alors $deff \simeq n$ et SY inefficace par rapport à SAS ;
- si $\rho = 0$ alors $deff \simeq 1$ et SY =SAS ;
- si $\rho < 0$ alors SY plus efficace que SAS.

On retient : SY est plus efficace que SAS si on peut ranger la population pour avoir des y_k hétérogènes dans chaque $s \in \mathcal{S}_{SY}$;

Estimateur de la variance : on a des $\pi_{kl} = 0$ alors nous ne pouvons pas utiliser l'estimateur de la variance de type HT.

Si l'ordre des unités dans la base de sondage est arbitraire, il est coutume d'utiliser l'estimateur de la variance que l'on aurait obtenu si l'échantillon avait été tirait selon un plan SAS,

$$\hat{V} = N^2 \frac{1-f}{n} S^2_{ys_r}.$$

Remarque : le plan SY est utilisé lorsque on ne connaît pas la taille de la population. Dans cette situation, c'est le pas qui est établi à l'avance et par conséquence, la taille de l'échantillon sera aléatoire.

Exemple : Afin de connaître le niveau de scolarité de l'auditoire d'une pièce de théâtre, le metteur en scène décide de tirer un échantillon aléatoire de spectateurs. Il s'installe à la porte du théâtre et interroge chaque dixième spectateur.

Chapitre 3 : Sondage stratifié (ST)

Les tirages présentés dans le chapitre 2 ne prennent pas en compte l'information auxiliaire (une "extra" information sur notre variable d'intérêt).
Et si nous en avons une ?
Comment l'utiliser pour sélectionner notre échantillon ?
De plusieurs façons...dont **la stratification**.
Différents raisons pour utiliser la stratification :

1. éviter les "mauvais échantillons" qui sont possibles avec un plan SAS (sélectionner un échantillons que des femmes ou que des hommes par example) ;
2. des coûts d'enquêtes plus faibles ;
3. avoir une certaine précision pour souspopulations ;
4. une meilleure précision par rapport au plan SI (si la stratification est bien réalisée).

Description du plan stratifié

On découpe la population dans des sous-populations appelées "strates" et on sélectionne de façon indépendante un échantillon (aléatoire simple sans remise) dans chaque strate.

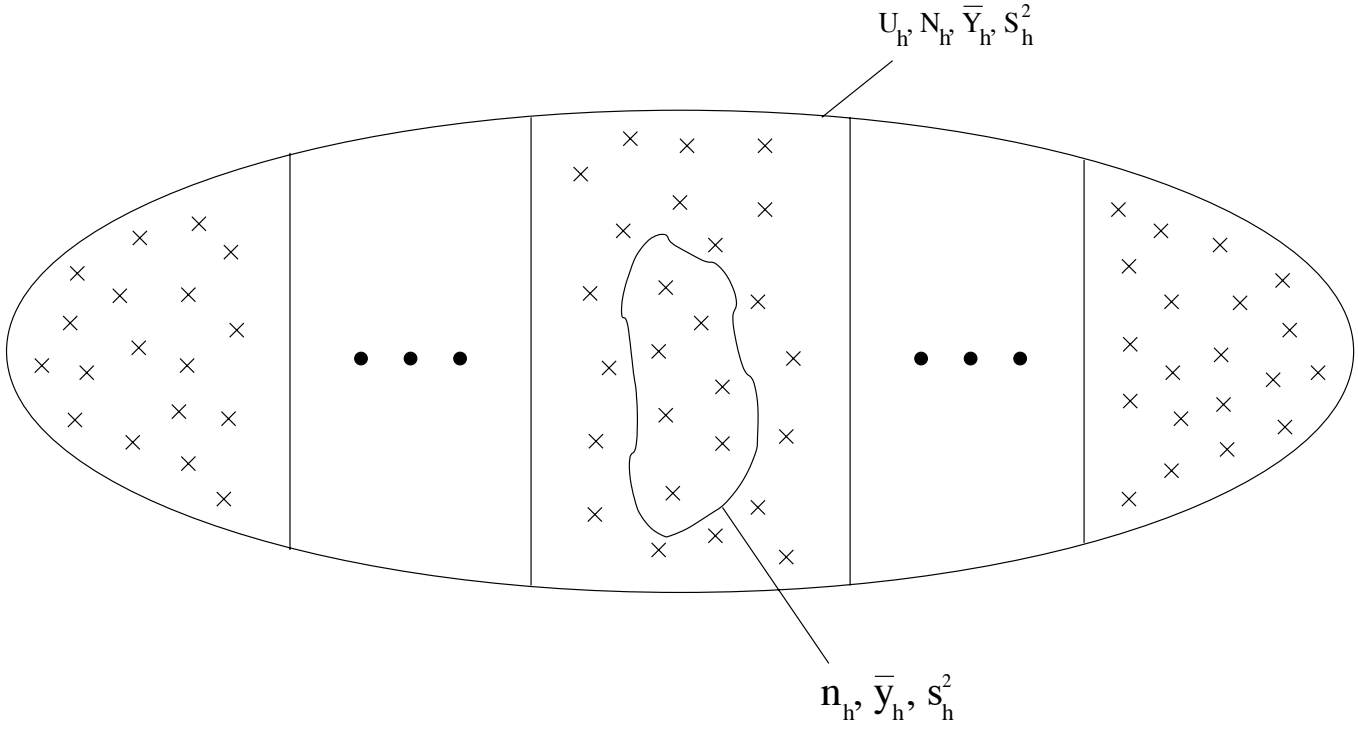


FIGURE: Plan stratifié et SAS dans chaque strate

$$U = \{1, \ldots, k, \ldots, N\} = \bigcup_{h=1}^{H} U_h, \quad U_h \cap U_{h'} = \emptyset \text{ pour } h \neq h'$$

$$|U| = N, \ |U_h| = N_h, \ N = \sum_{h=1}^{H} N_h \quad \text{avec} \quad N_h \quad \text{connu}$$

On tire de façon indépendante $s_h \subset U_h$ de taille n_h selon un plan $p_h(\cdot)$. L'échantillon final est s :

$$s = s_1 \cup s_2 \cup \ldots \cup s_H$$

$$|s_h| = n_h \Rightarrow |s| = \sum_{h=1}^{H} n_h.$$

alors, la probabilité d'obtenir s est :

$$p(s) = p_1(s_1) \ldots p_H(s_H).$$

Exemples de populations stratifiées

1. les personnes de différentes âges ont différentes pressions du sang, alors si on s'intéresse à la pression du sang il faut stratifié par classe d'âge.
2. dans une étude sur la concentration des plantes, il faut stratifié par type de terrain.
3. si on veut estimer le volume des transactions des entreprises américaines avec l'Europe, une stratification en fonction de la taille des entreprises est faite : les grosses entreprises, les moyennes, les petites.

Probabilités d'inclusion

Soient $\pi_k^h = P(k \in s_h)$, $k \in U_h$ et $\pi_{kl}^h = P(k, l \in s_h)$, $k \neq l \in U_h$ les probas d'inclusion par rapport à $p_h(\cdot)$; alors, par rapport à $p(\cdot)$,

$$\pi_k = P(k \in s) = \pi_k^h \quad \text{si} \quad k \in U_h$$

$$\pi_{kl} = P(k, l \in s) = \begin{cases} \pi_k^h \pi_l^{h'} & si \quad k \in U_h, l \in U_{h'}, \quad h \neq h' \\ \pi_{kl}^h & si \qquad k, l \in U_h \end{cases}$$

Estimation d'un total : $t_y = \sum_U y_k = \sum_{h=1}^H t_h = \sum_{h=1}^H N_h \overline{y}_{U_h}$

- $\hat{t}_\pi = \sum_{h=1}^H \hat{t}_{h\pi}$ avec $\hat{t}_{h\pi}$ l'estimateur HT pour la strate h.
- $V(\hat{t}_\pi) = \sum_{h=1}^H V_h(\hat{t}_{h\pi})$ avec $V_h(\hat{t}_{h\pi})$ la variance HT pour $\hat{t}_{h\pi}$ pour tous h.
- $\hat{V}(\hat{t}_\pi) = \sum_{h=1}^H \hat{V}_h(\hat{t}_{h\pi})$ avec $\hat{V}_h(\hat{t}_{h\pi})$ l'estimateur HT pour V_h pour tous h.

Sondage stratifié avec sondage aléatoire simple dans chaque strate

Considérons le cas d'un sondage aléatoire simple de taille n_h dans chaque strate U_h. Alors,

$$\pi_k = \pi_k^h = \frac{n_h}{N_h} \quad \text{si} \quad k \in U_h$$

$$\pi_{kl} = \begin{cases} \frac{n_h}{N_h} \cdot \frac{n_{h'}}{N_{h'}} & si \quad k \in U_h, l \in U_{h'}, \quad h \neq h' \\ \frac{n_h(n_h-1)}{N_h(N_h-1)} & si \qquad k, l \in U_h \end{cases}$$

Estimation d'un total :

- $\hat{t}_\pi = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{s_h} y_k$.
- $V(\hat{t}_\pi) = \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} S_{yU_h}^2$ avec $S_{yU_h}^2$ la variance de $\mathcal{Y}$ dans chaque strate h.
- $\hat{V}(\hat{t}_\pi) = \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} S_{ys_h}^2$.

Précision du plan stratifié

ANOVA décomposition :

$$(N-1)S_{yU}^2 = \sum_{h=1}^{H}(N_h-1)S_{yU_h}^2 + \sum_{h=1}^{H} N_h(\overline{y}_{U_h} - \overline{y}_U)$$

$$SST = SSW + SSB$$

L'estimateur $\hat{t}_\pi$ pour le total $t_y = \sum_U y_k$ a la variance

1. pour un plan SAS : $N^2 \frac{1-f}{n} S_{yU}^2$, alors est liée à SST ;
2. pour un plan STSAS : $\sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} S_{yU_h}^2$, alors elle est liée à SSW.

Il résulte que si les strates sont homogènes, SSW est petit et par conséquence, le plan stratifié STSAS est meilleur que le plan SAS . C'est le contraire du plan SY.

Comment choisir les tailles d'échantillons dans chaque strate ?

On a $n = \sum_{h=1}^{H} n_h$ et comment choisissons-nous chaque n_h ?

- **l'allocation proportionnelle** : on choisi n_h individus dans la strate h de façon que on a le même taux de sondage dans chaque strate que dans la population :

$$\frac{n_h}{N_h} = \frac{n}{N} \quad \text{pour toutes les strates} \quad h$$

Si on veut estimer la moyenne $\overline{y}_U$ avec un plan STSAS et allocation proportionnelle, l'estimateur HT est égal à

$$\hat{\overline{y}}_U = \overline{y}_s = \frac{1}{n} \sum_s y_k \quad \text{échantillon auto-pondéré}$$

Petit exemple pour l'allocation proportionnelle

Une population $U = U_1 \cup U_2$ avec deux strates :

$$N_1 = 40000 \quad N_1/N = 0.8 \quad n_1 = 160 \quad \bar{y}_1 = 12 \quad s_1^2 = 85 \quad s_1 = 9.22$$
$$N_2 = 10000 \quad N_2/N = 0.2 \quad n_2 = 40 \quad \bar{y}_2 = 58 \quad s_2^2 = 930 \quad s_2 = 30.50$$

- $\bar{y}_{\text{st}} = \bar{y} = \dfrac{N_1}{N} \times \bar{y}_1 + \dfrac{N_2}{N} \times \bar{y}_2 = 0.8 \times 12 + 0.2 \times 58 = 21.2$

- $$\begin{aligned} \mathrm{V}(\bar{y}_{\text{st}}) &= \sum_{h=1}^{2} \left(\frac{N_h}{N}\right)^2 V(\bar{y}_{U_h}) = \sum_{h=1}^{2} \left(\frac{N_h}{N}\right)^2 \frac{1-f_h}{n_h} S_{yU_h}^2 \\ &= \frac{1-f}{n}\left(\frac{N_1}{N} S_{yU_1}^2 + \frac{N_2}{N} S_{yU_2}^2\right) \\ &\simeq \left(\frac{n_1}{n}\right)^2 \frac{S_{yU_1}^2}{n_1} + \left(\frac{n_2}{n}\right)^2 \frac{S_{yU_2}^2}{n_2} \end{aligned}$$

  - $\hat{\mathrm{V}}(\bar{y}_{\text{st}}) \simeq 0.64 \times 85/160 + 0.04 \times 930/40 = 1.27$

- **l'allocation optimale de Neyman** : on choisi n_h individus dans la strate h proportionnellement à la dispersion de la variable $\mathcal{Y}$ dans la strate h :

$$n_h = n \frac{N_h S_{yU_h}}{\sum_{h=1}^{H} N_h S_{yU_h}}$$

On va **augmenter** les effectifs échantillonnés dans les strates où

1. la **variabilité** est **grande** et diminuer les effectifs échantillonnés dans les strates homogènes.
2. la **taille relative de la strate** N_h/N est grande.

Remarque : si $n_h > N_h$, alors on fait un recensement dans U_h et on recommence le calcul des taille de n_h dans les strates restantes.

Remarques :

1. l'allocation optimale a été obtenue en minimisant la variance de l'estimateur de Horvitz-Thompson sous une contrainte de coût.
2. cette allocation est idéale et elle ne peut pas être utilisée car elle fait intervenir les dispersions inconnues de $\mathcal{Y}$ dans la strate h, S_{yU_h}.
3. si on dispose d'une variable $\mathcal{X}$ très corrélée avec $\mathcal{Y}$, on peut remplacer cette allocation par **l'allocation $\mathcal{X}$-optimale** :

$$n_h = n\frac{N_h S_{xU_h}}{\sum_{h=1}^{H} N_h S_{xU_h}}$$

Précision du plan stratifié

- **L'allocation proportionnelle** :

$$
\begin{aligned}
V_{STSAS,p}(\hat{t}_\pi) &= N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \sum_h W_h S_{yU_h}^2 \\
W_h &= \frac{N_h}{N}
\end{aligned}
$$

- **L'allocation optimale de Neyman**

$$
\begin{aligned}
V_{STSAS,o}(\hat{t}_\pi) &= \frac{N^2}{n} \left(\sum_h W_h S_{yU_h}\right)^2 - N \sum_h W_h S_{yU_h}^2. \\
W_h &= \frac{N_h}{N}
\end{aligned}
$$

Comparaison entre le plan STSAS avec allocation optimale et proportionnelle

$$V_{STSAS,p}(\hat{t}_\pi) - V_{STSAS,o}(\hat{t}_\pi) = \frac{N^2}{n} Var(S_{yU_h}) \geq 0$$

• On a toujours $V_{STSAS,p}(\hat{t}_\pi) \geq V_{STSAS,o}(\hat{t}_\pi)$ avec égalité si S_{yU_h} est le même pour toutes les strates : dans ce cas, l'allocation proportionnelle est optimale et une allocation proportionnelle suffit.

• Alors, le gain de précision entre les deux types d'allocation dépend de la variance des écarts-types des strates.

• Si les variances $S^2_{yU_h}$ sont très différentes d'une strate à l'autre, alors il y a un gain considérable du plan STSAS avec allocation optimale.

• L'allocation optimale dépend de chaque variable d'intérêt alors que l'allocation proportionnelle dépend que de la taille de chaque strate.

Comparaison entre le plan STSAS avec allocation proportionnelle et le plan SAS

Les probabilités d'inclusion d'ordre un sont égales à

$$\pi_k = \frac{n}{N}$$

pour le plan SAS et STSAS et allocation proportionnelle. Cela montre que les π_k ne définissent pas un plan de sondages.
Alors, l'estimateur HT pour la moyenne est le même pour les deux plans

$$\hat{t}_\pi = N \cdot \overline{y}_s = \frac{N}{n} \sum_s y_k$$

mais des variances différentes :

$$V_{SAS}(\hat{t}_\pi) = \frac{1-f}{N} S_{yU}^2.$$

• **ANOVA** : $SST = SSW + SSB$ ou

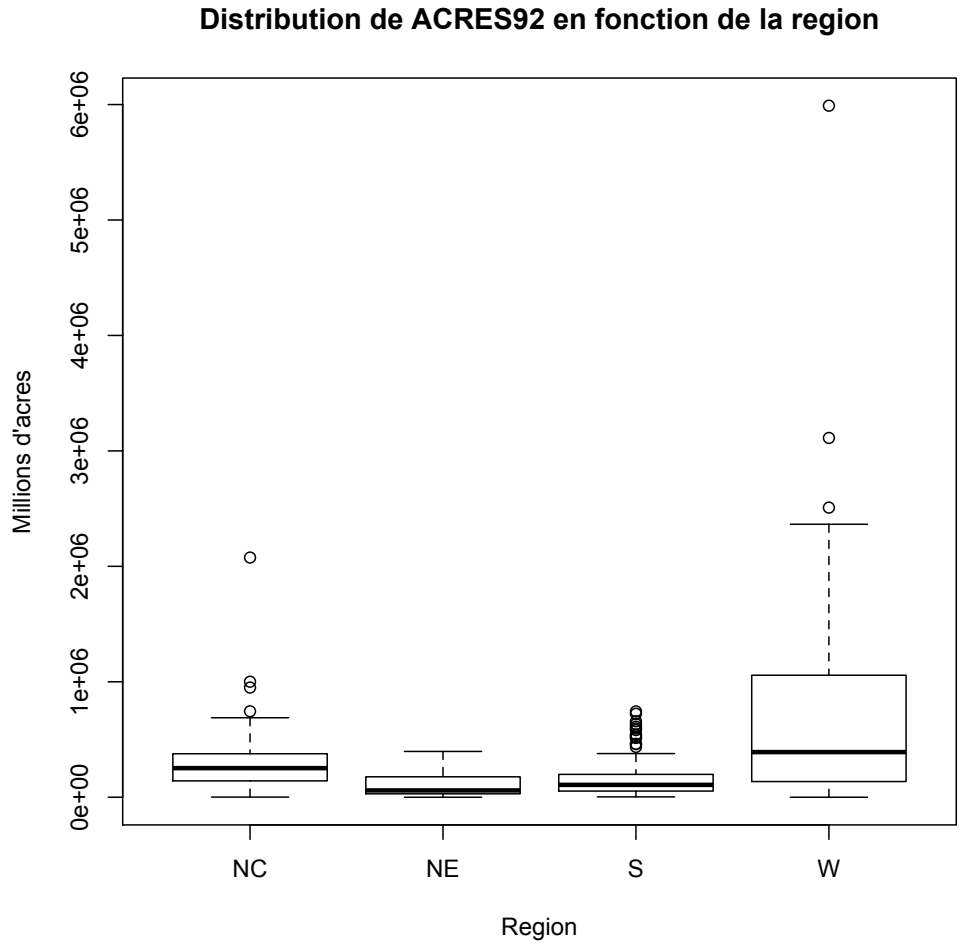
$$(N-1)S_{yU}^2 = \sum_{h=1}^{H}(N_h-1)S_{yU_h}^2 + \sum_{h=1}^{H} N_h(\overline{y}_{U_h} - \overline{y}_U)^2$$

• On peut avoir $V_{STSAS,p}(\hat{t}_\pi) \geq V_{SAS}(\hat{t}_\pi)$ si toutes les moyennes $\overline{y}_{U_h}$ sont égales ou presque (situation plutôt rare).

• Quand les moyennes $\overline{y}_{U_h}$ sont très différentes, alors l'allocation proportionnelle est plus efficace que le plan SAS.

Exemple

On considère la population de fermes (le recensement agricole américain). On prend un échantillon SRS de taille $n = 300$ et on s'interresse à la variable ACRES92



On considère que la population est stratifiée selon la région et on prend un échantillon stratifié avec SRS dans chaque strate :

$$N_1 = 1034, n_1 = nN_1/N \simeq 102$$

$$N_2 = 220, \quad n_2 = nN_2/N \simeq 22$$

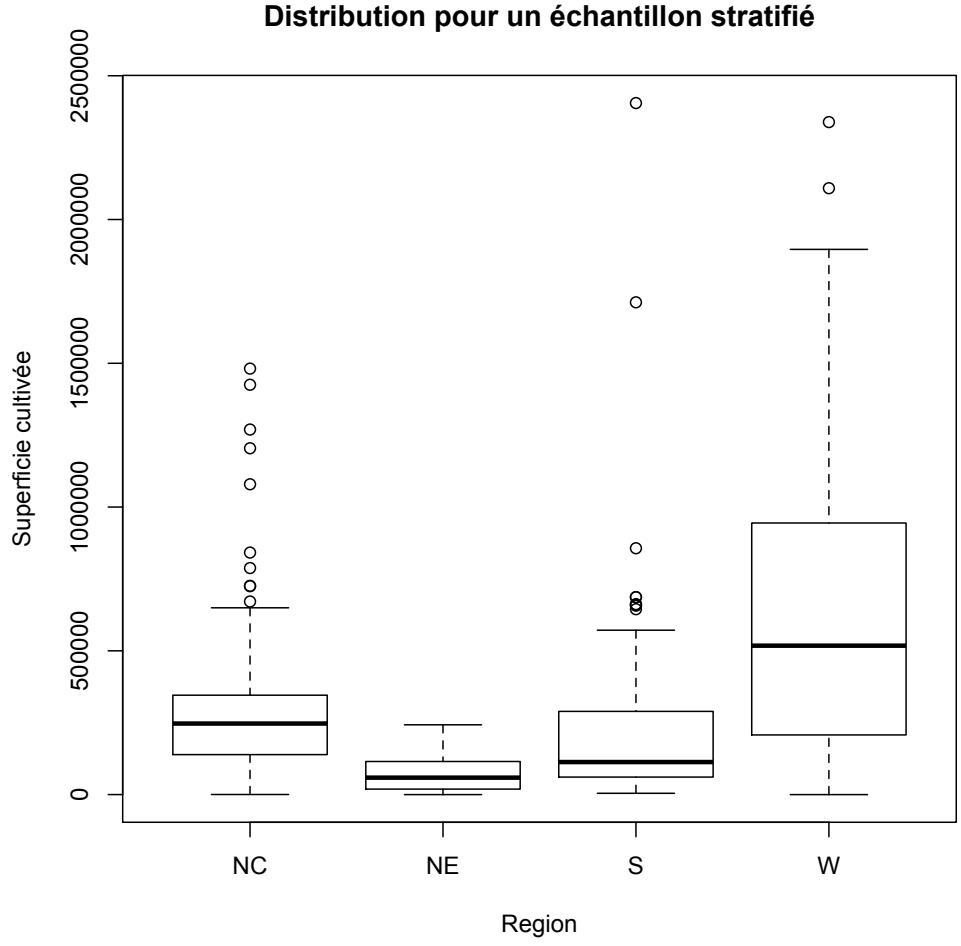
$$N_3 = 1362, \quad n_3 = nN_3/N \simeq 135$$

$$N_4 = 413, \quad n_4 = nN_4/N \simeq 41$$

Région	Taille s	Moyenne	Variance
NC	1034	396672.00	216650390330
NE	220	94563.29	8589632647
S	1362	208838.13	95617819372
W	413	637202.94	444747272381

Exemple

On prend un échantillon STRAT de taille $n = 300$ et on s'interresse à la variable ACRES92



1. estimation du total de la variable ACRES92

$$\begin{aligned}\hat{t}_{strat} &= N_1\overline{y}_{s_1} + \ldots + N_4\overline{y}_{s_4} \\ &= 1034 * 396672 + \ldots + 413 * 637202.94 = 978565119\end{aligned}$$

2. estimation de la variance de l'estimateur $\hat{t}_{strat}$:

$$\begin{aligned}\widehat{\text{Var}}(\hat{t}_{strat}) &= 1034^2 \left(1 - \frac{102}{1034}\right) \frac{216650390330}{102} + \ldots \\ &= 4.914128 \cdot 10^{15}\end{aligned}$$

3. gain du plan stratifié par rapport au plan SRS :

$$\frac{\widehat{\text{Var}}(\hat{t}_{strat})}{\widehat{\text{Var}}(\hat{t}_{srs})} = \frac{4.914128 \cdot 10^{15}}{5.890185 \cdot 10^{15}} \simeq 0.83$$

donc la même précision est obtenue avec un plan STRAT de taille $300 \cdot 0.83 = 249$ que avec un plan SRS de taille 300.

Mise en oeuvre. Avantages et désavantages

- C'est une opération délicate qui demande plus d'information qu'un plan aléatoire simple sans remise.
- Nous avons besoin d'une variable X très corrélée avec la variable d'intérêt Y qui servira comme variable de stratification ; ensuite, on doit connaître X pour chaque individu de la population.
- Pour une stratification efficace, on doit avoir dans la même strate des individus très semblables et très différents des individus des autres strates.
- Difficultés pratiques : comment choisir les bornes des strates et combien de strates il faut en prendre ?
- Les strates ainsi construites conviennent pour une seule variable alors qu'un pratique il y en a plusieurs ; pour palier ça, on peut faire une poststratification.

Chapitre 2 : Sondage à probabilités inégales.

Principe : attribuer aux individus de la base de sondage des π_k données a priori et **inégales**.

Justification : lorsque les unités n'ont pas la même importance, en particulier lorsqu'elle ont des tailles très différentes, il peut être intéressant, voire avantageux, d'attribuer aux différentes unités de chances de sortie inégales, les "grosses" unités ayant plus de chances d'appartenir à l'échantillon.

des π_k inégales se justifie en effet au vu des expressions des formules de variance.

Exemple : on veut estimer le volume d'une production totale Y d'une céréale. Il semble alors préférable que les grosses exploitations agricoles appartiennent à l'échantillon.

Comment définir les grosses exploitations agricoles ?

on utilise **l'information concernant la surface supposée** connue, c'est à dire

$\Longrightarrow$ utilisation d'information auxiliaire au moment de la constitution de l'échantillon ou dans la phase d'échantillonnage : les π_k vont dépendre de cette information auxiliaire.

Sondage de Poisson (PO) (1)

C'est le plan pour lequel les variables I_k, $k \in U$, sont indépendantes de loi Bernoulli de paramètre π_k,

$$P(I_k = 1) = \pi_k, \quad P(I_k = 0) = 1 - \pi_k.$$

- La taille $n_s = \sum_U I_k$ est une variable aléatoire d'espérance et variance

$$E(n_s) = \sum_U \pi_k, \quad V(n_s) = \sum_U \pi_k(1 - \pi_k).$$

- $p(s) = \Pi_{k \in s} \pi_k \Pi_{k \in U - s}(1 - \pi_k)$.
- π_k et π_{kl}.

Sondage de Poisson (PO) (2)

Mise en pratique : on génère des variables aléatoires *iid* $\simeq \mathcal{U}_{(0,1)}$, $\varepsilon_1, \ldots, \varepsilon_N$; si $\varepsilon_k < \pi_k$ alors $k \in s$, sinon on passe à l'unité suivante.

Estimation pour un total :

- $\hat{t}_\pi = \sum_s \frac{y_k}{\pi_k}$;
- $V_{PO}(\hat{t}_\pi) = \sum_U \left(\frac{1}{\pi_k} - 1 \right) y_k^2$;
- $\widehat{V}_{PO}(\hat{t}_\pi) = \sum_s (1 - \pi_k) (y_k/\pi_k)^2$.

Les probabilités π_k optimaux

Objectif : On cherche les π_k qui minimisent la variance $V_{PO}(\hat{t}_\pi)$ sous une contrainte de taille :

$$\min_{\pi_k} V_{PO}(\hat{t}_\pi)$$

$$E(n_s) = \sum_U \pi_k = n$$

Solution : $\pi_k = \dfrac{n y_k}{\sum_U y_k}$ pour tous $k \in U$ et supposant que $y_k < \sum_U y_k / n$.

Incovenient : les y_k ne sont pas connus et par conséquent, les π_k non plus. Alors, si nous disposons d'une variable X pour laquelle on connaît toutes ses valeurs $x_1, \ldots, x_N$ et approximativement proportionnelle avec Y, on peut prendre les $\pi_k \propto x_k$,

$$\pi_k = \frac{n x_k}{\sum_U x_k}$$ probabilités proportionnelles à la taille.

Si $x_k > \sum_U x_k / n$ pour un certain k, alors $\pi_k = 1$.

Plans proportionnels à la taille

La discussion sur le plan PO et les π_k optimaux, nous conduit à la conclusion suivante :

Pour un plan de taille fixe n et les $\pi_k \propto y_k$, l'estimateur HT pour le total est de variance nulle. Dans les mêmes conditions, si on a $\pi_k \propto x_k$, l'estimateur HT a une variance petite.

Deux types des plans proportionnels à la taille :

1. plans de taille fixe sans remise et le π-estimateur (HT) : $\pi_k \propto x_k$ ou des plans πps ;
2. plans avec remise et le *pwr*-estimateur : $p_k \propto x_k$ ou des plans *pps* ;

Plans πps

Dans cette situation, $\pi_k \propto x_k$ avec $x_1, \ldots, x_N$ positives et connues.

L'estimateur HT est $\hat{t}_\pi = \sum_s \frac{y_k}{\pi_k}$ et les rapports $\frac{y_k}{\pi_k}$ sont approximativement constants ce qui implique une petite variance (voir la formule de Yates-Grundy-Sen).

Il faut mettre en oeuvre, un plan de taille fixe n qui possède les propriétés suivantes :

1. la sélection de l'échantillon est relativement simple ;
2. les $\pi_k \propto x_k$;
3. les $\pi_{kl} > 0$ pour tous k, l et facile à calculer ;
4. $\Delta_{kl} < 0$ pour $k \neq l$.

Brewer & Hanif (1983) présentent une soixantaine des plans...
Nous allons faire le plan systématique à probabilités inégales.

Le plan systématique à probabilités inégales

Notons x_k la mesure de taille de l'unité k, connue quel que soit k et $t_x = \sum_U x_k$; les x_k sont une information auxiliaire. Les étapes de ce tirage sont les suivantes :

1. Soit n la taille de l'échantillon qu'on doit tirer. Si une unité a une mesure de taille $\geq t_x/n$ elle est retirée de la population et est mise d'office dans l'échantillon.
2. Posons $p_i = x_i/t_x$ et $\pi_i = np_i$.
3. On forme $V_k = \sum_{i=1}^{k} \pi_i, \forall k \in U$ et $V_0 = 0$.
4. On génère une observation u, d'une v.a. Unif$(0, 1)$.
5. L'échantillon est formé des unités : k_1 telle que $V_{k_1-1} < u \leq V_{k_1}$, k_2 telle que $V_{k_2-1} < u + 1 \leq V_{k_2}, \cdots$, k_n telle que $V_{k_n-1} < u + n - 1 \leq V_{k_n}$.

On voit que la probabilité que l'unité k soit dans l'échantillon est la longueur de l'intervalle $[V_{k-1}, V_k]$, c'est-à-dire la quantité π_k. Dans cette méthode beaucoup de probabilité d'inclusion d'ordre 2 sont nulles. Des approximations de la variance du total ont été proposées.

Plans *pps*

Considérons le modèle de tirage avec remise suivant :
on a une urne avec N billes de trois couleurs : blanc, noire et rouge de proportions p_1, p_2 et p_3. La probabilité de sélectionner une bille blanche est p_1, noire p_2 et rouge p_3. Nous avons des probabilités de sélection différentes. On effectue n tirages avec remise et on regarde le nombre de billes blanches, noires et rouges.
Ce modèle est utilisé lors de sondages à probabilités inégales :

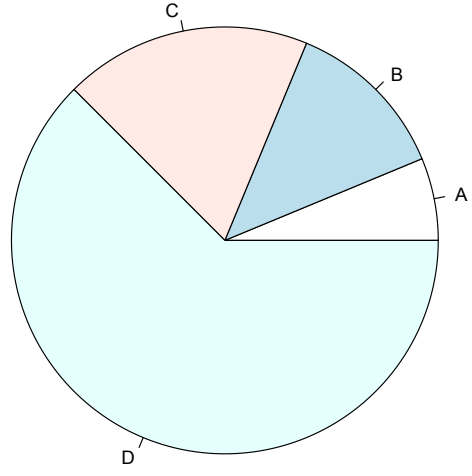
- pour un sondage de ménages, il est logique de considérer que sa probabilité de sélection dépend de sa taille (le nombre de personnes) ;
- pour un sondage d'entreprises, les grosses entreprises ont plus de chances d'être sélectionnées.

Exemple (Lohr, 1999)

- Considérons une ville qui a quatre supermarchés de tailles entre 100 m^2 et 1000 m^2.
- L'objectif est d'estimer le chiffre de vente en prenant un seul supermarché. (c'est qu'un exemple, pour une population de seulement 4 supermarchés on aurait pu faire un recensement).
- On se dit que le chiffre de vente est très lié à la taille du supermarché

supermarché,i	superficie	p_i	chiffre de vente,t_i
A	100	1/16	11
B	200	2/16	20
C	300	3/16	24
D	1000	10/16	245
$Total = t$	1600	1	300

Méthode de tirage : on numérote 16 cartes de 1 à 16 et on tire une carte : si le numéro est 1, on choisi le supermarché A, si le numéro est 2 ou 3, on choisi B, si le numéro est entre 4 et 6, on prend C et pour le reste, D.
Supposons que le supermarché A est tiré. Pour estimer le chiffre de ventes, on multiplie le chiffre de vente de A, 11, par le poids de A : 16. Cela s'explique par le fait que si la superficie de A représente 1/16 de la superficie totale de quatre supermarchés et la superficie et le chiffre de ventes sont fortement liés, alors le chiffre de vente de A va représenter aussi 1/16 du chiffre total.
On obtient $\hat{t} = \frac{t_1}{p_1} = 16 \cdot 11 = 176$



Le plan *pps* : description

- Soient $p_1, \ldots, p_N$, les probabilités de sélection de somme 1,

$$\sum_U p_k = 1$$

- On sélectionne un individu k parmi N avec une probabilité p_k et on le "remet". On répète de façon indépendante m fois le procédé.
- On obtient un échantillon ordonné : $so = (k_1, \ldots, k_m)$ pour m tirages ; un individu peut être sélectionné plusieurs fois.
- Le plan de sondage est $p(so) = p_{k_1} p_{k_2} \ldots p_{k_m}$.
- Le plan aléatoire simple avec remise est un cas particulier : $p_k = \frac{1}{N}$ pour tous $k \in U$.

L'estimateur de Hansen & Hurwitz

Deux types de probabilités :

1. probabilités de sélection : p_k
2. probabilités d'inclusion : π_k

On a $\pi_k = 1 - (1 - p_k)^m$ et pour p_k très petit, $\pi_k \simeq mp_k$.

Pour estimer le total t_Y on prend l'estimateur de Hansen & Hurwitz.

- Soit la variable

$$Z_k = \text{le nombre de fois que l'unité k est dans l'échantillon}$$

- $(Z_1, \ldots, Z_N)$ suit une loi multinomiale de paramètres $(m; p_1, \ldots, p_N)$;
- $Z_k \sim \mathcal{B}(m; p_k)$. Les variables Z_k sont dépendantes entre elles. $\alpha_k = E(Z_k) = mp_k$

Résultat

L'estimateur de Hansen & Hurwitz (1943) est donné par

$$\hat{t}_{HH} = \sum_{j=1}^{m} \frac{y_{k_j}}{\alpha_{k_j}} = \sum_{k \in U} \frac{y_k}{\alpha_k} Z_k = \frac{1}{m} \sum_{i=1}^{m} \frac{y_{k_i}}{p_{k_i}}$$

et il est sans biais pour t_y. Sa variance est donnée par

$$V(\hat{t}_{HH}) = \frac{1}{m} \sum_{U} \left(\frac{y_k}{p_k} - t_y \right)^2 p_k$$

et elle est estimée sans biais par
$\hat{V}(\hat{t}_{HH}) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \left(\frac{y_{k_i}}{p_{k_i}} - \hat{t}_{HH} \right)^2$.

Comment choisir les p_k ?

Si $\frac{y_k}{p_k} = c$ pour tous $k \in U$, alors $V(\hat{t}_{HH}) = 0$. Puisque une telle situation n'est pas possible, nous considérons la situation un peu "moins" idéale qui consiste à prendre $p_k \propto x_k$ avec x_k une variable auxiliaire connue sur toute la population et (approximativement) proportionnelle à y_k. Il résulte

$$p_k = \frac{x_k}{\sum_U x_k}, \quad k = 1 \ldots, N$$

Mise en oeuvre du plan pps : on applique m fois la méthode cumulative :

1. On considère les sommes cumulées $T_1 = p_1$, $T_2 = p_1 + p_2$, ... $T_N = p_1 + \ldots + p_N$ et on génère un nombre aléatoire ε selon $\mathcal{U}_{[0,1]}$. Si $T_{k-1} < \varepsilon \leq T_k$, l'unité k est choisie.
2. On répète m fois l'étape 1.

Calcul de précision sur l'exemple des supermarchés

échantillon	p_i	t_i	$\hat{t}$	$(\hat{t}-t)^2$
A	$1/16$	11	176	15376
B	$2/16$	20	160	19600
C	$3/16$	24	128	29584
D	$10/16$	245	392	8464

• **Le biais** : $E(\hat{t}) = \frac{1}{16} \cdot 176 + \frac{2}{16} \cdot 160 + \frac{3}{16} \cdot 128 + \frac{10}{16} \cdot 392 = 300$, l'estimateur $\hat{t}$ est sans biais.

• **La variance** :
$V(\hat{t}) = \frac{1}{16} \cdot 15376 + \frac{2}{16} \cdot 19600 + \frac{3}{16} \cdot 29584 + \frac{10}{16} \cdot 8464 = 14248$

Comparaison avec le plan à probabilités égales

échantillon	p_i	t_i	$\hat{t}$	$(\hat{t}-t)^2$
A	$1/4$	11	44	65536
B	$1/4$	20	80	48400
C	$1/4$	24	96	41616
D	$1/4$	245	980	462400

• **Le biais** : $E(\hat{t}) = \frac{1}{4} \cdot 44 + \frac{1}{4} \cdot 80 + \frac{1}{4} \cdot 96 + \frac{1}{4} \cdot 245 = 300$, l'estimateur $\hat{t}$ est sans biais.

• **La variance** :
$V(\hat{t}) = \frac{1}{4} \cdot 65536 + \frac{1}{4} \cdot 48400 + \frac{1}{4} \cdot 41616 + \frac{1}{4} \cdot 462400 = 154488.$

Conclusion : les deux estimateurs sons sans biais, mais la variance est nettement plus faible dans le cas d'un tirage proportionnel à la taille car on utilise une information auxiliaire fortement corrélée avec la variable d'intérêt.

Représenter les données avec un échantillon pps

On considère la population REC99HTEGNE et on veut estimer le nombre de logements vacants, variable LOGVAC, avec un échantillon pps de taille $m = 80$.