Initiation à la théorie des sondages: cours IREM-Dijon

Camelia Goga

IMB, Université de Bourgogne

Dijon, 12 novembre 2009

Très court historique

- Laplace a présenté à l'Académie des Sciences en 1783 une nouvelle méthode pour estimer (approximer) la population de la France en utilisant les registres de naissances et en prenant un échantillon de communes de la France.
- Les statisticiens de l'époque n'ont pas cru à la validité de cette méthode bien que Laplace l'avait justifié par des théorèmes de base de la théorie des probabilités comme le théorème central limite.

- Au début du 20ème siècle, l'idée de faire un sondage à la place d'un recensement (une énumération exhaustive de la population) réapparaît en Norvège et ensuite, ce procédé a été adopté par (presque) tous les pays.

- Aujourd'hui, il est presque impossible d'obtenir information sans faire un sondage. Pourtant, pour avoir un "bon sondage", plusieurs règles doivent être respectées.

Bibliographie :

1. Deville, J.C., Peut-on croire aux sondages ? *Pour la science*, 2006.
2. Tillé, Y., Théorie des sondages. *Dunod*, 2001.
3. Ardilly, P., Les techniques de sondage, *Editions technip*, 2006.
4. Lohr, S.L. *Sampling : Design and Analysis*, *Duxbury Press*, 1999.

Remarques et vocabulaire (1)

- Une population finie $U = \{u_1, u_2, \ldots, u_N\}$ dont la taille N peut être connue ou pas ;
- On prend un échantillon s dans U, c'est à dire une partie de U;
- On appelle base de sondage une énumération de notre population (l'annuaire, le registre d'entreprises ....)
- Les questionnaires de sondages contiennent plusieurs items ou variables d'intérêt :

  • des questions avec des réponses de type **quantitatif** :
  le salaire (par tranche), somme dédiée aux dépenses pour la nourriture ou loisirs par ménage, superficie attribuée aux fermes ...ou

  • des questions avec des réponses de type **oui-non** :
  avez-vous voté pour le candidat A ?

Remarques et vocabulaire (2)

- L'objectif est de connaître une fonction de ces variables : le salaire moyen de Français, la superficie totale attribuée aux fermes en France ou la proportion de gens avoir voté pour A.
- Pour obtenir une estimation du total d'une variable Y sur toute la population,

$$t_y = \sum_{k \in U} y_k,$$

il ne suffit pas de faire la somme des valeurs sur l'échantillon. Il faut utiliser des poids $w_k > 1$:

$$\hat{t}_y = \sum_{k \in s} w_k y_k \quad \text{estimateur par les valeurs dilatées}$$

- Pour améliorer les résultats, on utilise une variable auxiliaire ; le salaire il y a cinq ans, la superficie totale des fermes il y a deux ans, l'âge, la catégorie socio-professionnelle ...

Quelques outils statistiques

Soit une variable Y de valeurs $y_1, \ldots, y_N$ et $\hat{t}_y$ un estimateur de t_y.

- *Biais*$(\hat{t}_y) = E(\hat{t}_y) - t_y$ avec

$$E(\hat{t}_y) = \sum_{\text{tous les s}} p(s)\hat{t}_y(s)$$

moyenne pondérée des valeurs possibles de $\hat{t}_y$

Un estimateur est sans biais si son biais est nul.

- Variance : mesure de dispersion,

$$V(\hat{t}_y) = \sum_{\text{tous les s}} p(s)(\hat{t}_y(s) - E(\hat{t}_y))^2$$

Variance des $Y_i,\ i = 1, \ldots, N$ $S^2 = \dfrac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2$

Ecart-type (standard deviation) : racine carré de la variance (même unité que la variable).

- **Coefficient de variation** : $cv = \frac{\text{ecart-type}}{\text{estimation}}$ (sans unité de mesure).
- **Erreur quadratique moyenne** :
  $EQM(\hat{t}_y) = E(\hat{t}_y - t_y)^2 = V(\hat{t}_y)^2 + Biais(\hat{t}_y)^2$

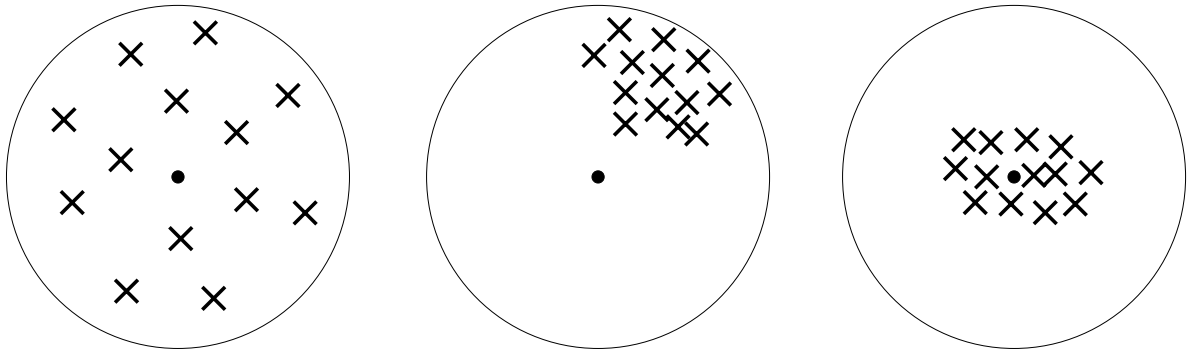


FIGURE : Biais et précision

cas 1= estimateur sans biais (la moyenne des toutes les positions est le centre) ;
cas 2= estimateur précis mais biaisé (les positions sont très proches les unes des autres mais éloignées du centre) ;
cas 3= estimateur "parfait" (les positions sont très proches du centre).

Exemple

Soit une population $U = \{1, 2, 3, 4\}$ et R=le revenu moyen de cette population. On a

$$R_1 = 6000, \quad R_2 = 12000, \quad R_3 = 8000, \quad R_4 = 6000.$$

On veut interroger que deux personnes, alors on a six échantillons de tailles 2 sans remise

$s_1 = \{1, 2\}, s_2 = \{1, 3\}, s_3 = \{1, 4\}, s_4 = \{2, 3\}, s_5 = \{2, 4\}, s_5 = \{3, 4\}.$

On peut prendre par exemple :

$p(s_1) = p(s_2) = 0, 25; p(s_3) = 0, 2; p(s_4) = 0, 1; p(s_5) = 0, 1; p(s_6) = 0, 1;$

echantillon, s	$p(s)$	$\hat{t}(s)$	$p(s) \cdot \hat{t}(s)$
$\{1, 2\}$	0.25	9000	2250
$\{1, 3\}$	0.25	7000	1750
$\{1, 4\}$	0.2	6000	1200
$\{2, 3\}$	0.1	10000	1000
$\{2, 4\}$	0.1	9000	900
$\{3, 4\}$	0.1	7000	700

- La vraie moyenne est $R = \frac{R_1+R_2+R_3+R_4}{4} = 8000$.
- L'espérance de $\hat{t}$ est

$$E(\hat{t}) = 0.25 \cdot 9000 + 0.25 \cdot 7000 + \ldots + 0.1 \cdot 7000 = 7800$$

et le biais est $7800 - 8000 = -200$.

- La variance est

$$\begin{aligned} V(\hat{t}) &= 0.25 \cdot (9000 - 7800)^2 + 0.25 \cdot (7000 - 7800)^2 + \\ &\quad \ldots + 0.1 \cdot (7000 - 7800)^2 = 1860000 \end{aligned}$$

- L'erreur quadratique moyenne est

$$\begin{aligned} EQR(\hat{t}) &= 0.25 \cdot (9000 - 8000)^2 + 0.25 \cdot (7000 - 8000)^2 + \\ &\quad \ldots + 0.1 \cdot (7000 - 8000)^2 = 1900000 = V(\hat{t}) + Biais^2 \end{aligned}$$

Décomposition de la variance ou Anova

groupe 1	groupe 2	...	groupe G
y_{11}	y_{21}	$\cdots$	y_{G1}
y_{12}	y_{22}	$\cdots$	y_{G2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_{1n_1}	y_{2n_2}	$\cdots$	y_{Gn_G}
$\bar{y}_1$	$\bar{y}_2$	$\cdots$	$\bar{y}_G$

variation totale=variation intra-groupes+variation inter-groupes

$$\sum_{i,j}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{G}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2 + \sum_{i=1}^{G} n_i(\bar{y}_i-\bar{y})^2$$

Plan de sondage simple (SASSR)

Principe Pour une population de taille N et une taille d'échantillon fixée à n, tous les échantillons ont la **même probabilité** d'être tirés.

Petit exemple : moyenne des montants des factures de vente d'une société en euros, $N = 5$

- les valeurs sur toute la population :

5	8	10	12	15

- vraie moyenne :

$$\bar{Y} = \frac{5 + 8 + 10 + 12 + 15}{5} = 10$$

- plan SASSR, $n = 2$

Echantillons possibles de taille $n = 2$ et estimations de $\bar{Y}$ par

$$\bar{y} = \frac{y_1 + y_2}{2} :$$

y_1	5	5	5	5	8	8	8	10	10	12
y_2	8	10	12	15	10	12	15	12	15	15
$\bar{y}$	6.5	7.5	8.5	10	9	10	11.5	11	12.5	13.5

Remarques et Vocabulaire

- Echantillon de taille n : **partie** de taille n de la population.

  plan SASSR = plan sans remise ($\simeq$ avec remise donc indépendance si N est grand),

- **Taux de sondage**$= f = \dfrac{n}{N}$ = probabilité pour chaque observation d'être tirée dans l'échantillon (probabilité d'inclusion).

  plan SASSR = plan à probabilités égales (équiprobabilité).

- **Poids de sondage** associé à une observation : inverse de la probabilité d'inclusion de l'observation.

$$\text{plan SASSR} = \frac{N}{n}.$$

- **Estimateur par les valeurs dilatées** :

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{n} \frac{N}{n} y_i = \frac{N}{N} \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Exemple : $\bar{y} = \frac{1}{5} \left(\frac{5}{2} y_1 + \frac{5}{2} y_2 + \frac{5}{2} y_3 \right)$

Mise en oeuvre : 1

Le tirage aléatoire simple sans remise de taille n dans une population de taille N est l'équivalent du tirage sans remise de n *boules noires d'une urne contenant N boules noires*.

Cela permet de calculer la probabilité d'avoir n individus : $\frac{1}{\mathrm{C}_N^n}$ ou :

1. on sélectionne le premier individu avec une probabilité de $\frac{1}{N}$ et on l'enlève de la liste ;
2. on sélectionne le deuxième individu avec une probabilité de $\frac{1}{N-1}$ et on l'enlève de la liste ;
3. ...
4. on sélectionne le n-ième individu avec une probabilité de $\frac{1}{N-n+1}$ et on arrête.

Alors, la probabilité d'avoir un échantillon de taille n est $n! \times \frac{1}{N} \times \frac{1}{N-1} \times \ldots \times \frac{1}{N-n+1} = \frac{1}{\mathrm{C}_N^n}$

Mise en oeuvre : 2

L'algorithme présenté n'est pas utilisé en pratique car il nécessite n lectures du fichier des données et beaucoup des opérations de tri qui peuvent prendre beaucoup de temps si la taille de la population est grande.

Algorithme 2 : on affecte un nombre aléatoire uniforme $(0, 1)$ à chaque individu de la population. On trie ensuite le fichier par ordre croissant (ou décroissant) des nombres aléatoires. On choisit les n premiers (ou derniers) individus du fichier ainsi ordonné. C'est une méthode aisée à mettre en oeuvre mais on doit trier tout le fichier des données (opération longue pour N grand.)

Exemple : pour la population des factures de vente ;
• On génère 10 numéros aléatoires uniformes $(0, 1)$:

```
0.1062524 0.3319940 0.7559061 0.8083069 0.4430856
 0.6938476 0.1101506 0.8585162 0.9977094 0.0980415
```

• On prend les individus qui correspondent aux deux plus petits nombres de la liste :

```
10   1   7   2   5   6   3   4   8   9
```

Les individus qui se trouvent sur la 10ème et la 1ère place dans la liste seront sélectionnés (ils ont les deux plus petits nombres aléatoires uniformes).

Propriétés statistiques

Biais : la moyenne des $\bar{y}$ sur l'ensemble des échantillons est $\bar{Y}$. On dit que $\bar{y}$ est **sans biais**.

Exemple :

$\bar{y}$	6.5	7.5	8.5	10	9	10	11.5	11	12.5	13.5

$$\frac{6.5 + 7.5 + 8.5 + 10 + 9 + 10 + 11.5 + 11 + 12.5 + 13.5}{10} = 10$$

"sans biais" signifie que le résultat est bon "en moyenne" mais pas que le résultat obtenu à partir d'**un** échantillon est exact.

Variance : Variance de $\bar{y}$ pour un plan SASSR :

$$\mathrm{V}(\bar{y}) = (1 - f)\,\frac{S^2}{n}$$

d'autant plus petite que :

- la taille de l'échantillon n est grande,
- la dispersion des données pour la variable considérée est petite (S^2),
- le taux de sondage est grand (f).
- le facteur $1 - f$ s'appelle correction en population finie ; important pour des populations petites.

Remarques

- Pour des populations de grande taille, c'est la taille de l'échantillon n qui donne la précision et non le taux de sondage f.

$$\begin{array}{llll} N_1 = 1000 & n_1 = 10 & f_1 = 0.01 & S_1^2 = 40 \\ N_2 = 1000 & n_2 = 100 & f_1 = 0.1 & S_2^2 = 40 \end{array}$$

$$\begin{aligned} \mathrm{V}(\bar{y_1}) &= 0.99 \times \frac{40}{10} = 3.96 \\ \mathrm{V}(\bar{y_2}) &= 0.9 \times \frac{40}{100} = 0.36 \end{aligned}$$

$$\begin{array}{llll} N_1 = 1000 & n_1 = 100 & f_1 = 0.1 & S_1^2 = 40 \\ N_2 = 100000 & n_2 = 100 & f_1 = 0.001 & S_2^2 = 40 \end{array}$$

$$\begin{aligned} \mathrm{V}(\bar{y_1}) &= 0.9 \times \frac{40}{100} = 0.36 \\ \mathrm{V}(\bar{y_2}) &= 0.999 \times \frac{40}{100} = 0.3996 \end{aligned}$$

- Le fait que la variable d'intérêt soit peu ou très dispersée a beaucoup d'influence sur la précision.

$$\begin{array}{llll} N_1 = 1000 & n_1 = 100 & f_1 = 0.1 & S_1^2 = 80 \\ N_2 = 1000 & n_2 = 100 & f_1 = 0.1 & S_2^2 = 20 \end{array}$$

$$\begin{aligned} V(\bar{y}_1) &= 0.9 \times \frac{80}{100} = 0.72 \\ V(\bar{y}_2) &= 0.9 \times \frac{20}{100} = 0.18 \end{aligned}$$

- Si N est grand ($f \simeq 0$),

$$\mathrm{V}(\bar{y}) = \frac{S^2}{n}$$

et $\sqrt{\mathrm{V}(\bar{y})} = \frac{S}{\sqrt{n}}$ est l'erreur standard (**standard error**) des Y_i.

- Le calcul de la variance V dépend de la valeur de S^2 qui est inconnue. On estime S^2 par

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

et $\mathrm{V}(\bar{y})$ par

$$\hat{\mathrm{V}}(\bar{y}) = (1-f)\,\frac{s^2}{n}$$

Intervalles de confiance : 1

Hypothèse : la loi de $\bar{y}$ est une loi Normale

$$\text{donc} \quad \frac{\bar{y} - \bar{Y}}{\sqrt{V(\bar{y})}}$$

suit une loi normale de moyenne nulle et d'écart-type égal à 1

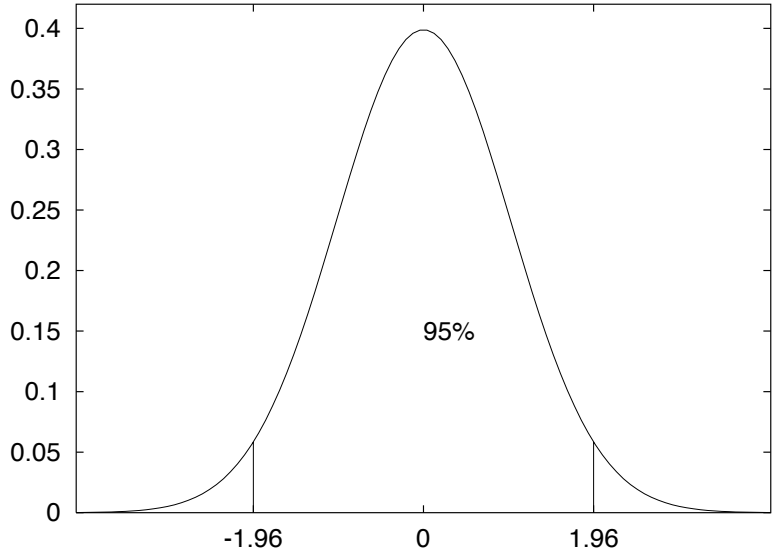


FIGURE : Loi normale

Intervalles de confiance : 2

L'intervalle de confiance pour $\bar{Y}$ à 95% :

$$\bar{Y} \in \left[\bar{y} - 1.96\sqrt{\hat{\mathrm{V}}(\bar{y})}\,;\, \bar{y} + 1.96\sqrt{\hat{\mathrm{V}}(\bar{y})}\right]$$

Petit exemple : $n = 100$, $\bar{y} = 11$, $s^2 = 36$, IC de $\bar{Y}$ à 95%
$= \left[11 - 2 \times \sqrt{36/100}\,;\, 11 + 2 \times \sqrt{36/100}\right] = [9.8\,;\, 12.2]$.

Précision absolue : demi-longueur de l'intervalle de confiance à 95% $\simeq 2\sqrt{\hat{\mathrm{V}}(\bar{y})}$.

Petit exemple : 1.2

Précision relative : $\dfrac{2\sqrt{\hat{\mathrm{V}}(\bar{y})}}{\bar{y}}$

Petit exemple : $1.2/11 = 10.9$ %

Estimation d'une proportion : cas particulier d'une moyenne

Soit une caractéristique A et soit la variable dichotomique Y des valeurs

$$y_k = \begin{cases} 1 & si \quad \text{l'individu k a A} \\ 0 & sinon \end{cases}$$

Objectif : On s'intéresse à la proportion d'individus P dans la population U qui ont caractéristique A.

$$P = \frac{\sum_U y_k}{N}.$$

- Avec un plan SASSR, la proportion P est estimée par

$$\hat{P} = \sum_s y_k / n$$

qui est la proportion d'individus ayant A dans l'échantillon s ;

- On a $S_{yU}^2 = \frac{N}{N-1} P(1-P) (\simeq P(1-P)$ si N grand) et

$$V(\hat{P}) = \frac{1-f}{n} S_{yU}^2.$$

- On a $S_{ys}^2 = \frac{n}{n-1} \hat{P}(1-\hat{P})$ et $\widehat{V}(\hat{P}) = \frac{1-f}{n} S_{ys}^2$;

- L'intervalle de confiance est

$$\widehat{IC}(\hat{P}) = \left[\hat{P} - z_{\alpha/2} \sqrt{\widehat{V}(\hat{P})}, \hat{P} + z_{\alpha/2} \sqrt{\widehat{V}(\hat{P})} \right]$$

Calcul de taille de s pour estimer P avec une précision donnée

Soit e *la marge d'erreur tolérée*. On veut n tel que la démi-longuer de l'intervalle de confiance est au plus égale à e,

$$e \geq z_{\alpha/2}\sqrt{V(\hat{P})} \quad (z_{\alpha/2} = 1.96 \quad \text{pour} \quad \alpha = 95\%)$$

Il résulte

$$n \geq \frac{z^2_{\alpha/2} S^2_{yU}}{e^2 + \frac{z^2_{\alpha/2} S^2_{yU}}{N}} = \frac{z^2_{\alpha/2} \frac{N}{N-1} P(1-P)}{e^2 + z^2_{\alpha/2} \frac{P(1-P)}{N-1}}$$

Difficulté : on ne connaît pas S^2_{yU}. On l'estime par $S^2_{ys} = \frac{n}{n-1}\hat{P}(1-\hat{P})$ et $\hat{P}$ peut être :

1. $\hat{P} = 1/2$ le cas extrême (le maximum de la fonction $p(1-p)$ est atteint pour $p = 1/2$) ;

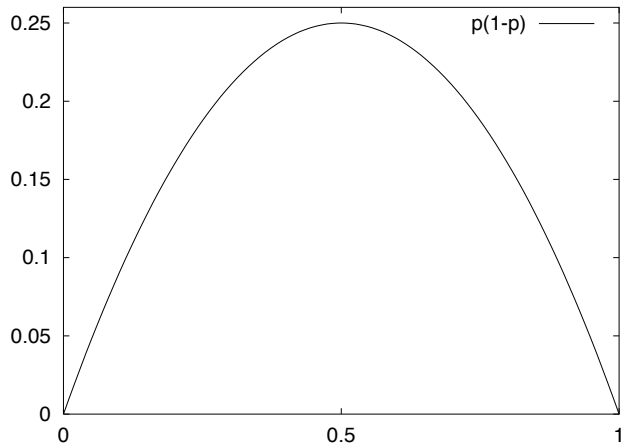


FIGURE : $P \mapsto P(1-P)$

2. une estimation de P issue lors d'une enquête pilote :

$$n \geq \frac{z^2_{\alpha/2} S^2_{ys}}{e^2 + \frac{z^2_{\alpha/2} S^2_{ys}}{N}}$$

Précision absolue de l'estimation d'une proportion en %.

n p	0,05 (0,95)	0,1(0,9)	0,2(0,8)	0,3(0,7)	0,4(0,6)	0,5
100			8	9.2	9.8	10
200		4.3	5.7	6.5	6.9	7.1
300	2.5	3.5	4.6	5.3	5.7	5.8
400	2.2	3	4	4.6	4.9	5
500	2	2.7	3.6	4.1	4.4	5
1000	1.4	1.8	2.5	2.9	3	3.1
2000	1	1.3	1.8	2.1	2.2	2.3
3000	0.8	1.1	1.4	1.6	1.8	1.8
5000	0.6	0.8	1.1	1.3	1.4	1.4
10000	0.4	0.6	0.8	0.9	1	1

Avantages et inconvénients du plan aléatoire simple sans remise

- **Avantage** : très simple à mettre en oeuvre ;
- **Désavantage** : il peut produire des "mauvais échantillons" et par conséquent, on n'a pas un bon échantillon représentatif : on peut avoir que des grandes ou petites valeurs de Y.
- Ce plan est très souvent utilisé dans les comparaisons : un nouvel estimateur est étudié pour ce plan particulier et comparé avec d'autres estimateurs ;
- Ce plan peut être efficace en combinaison avec d'autres méthodes : plans à plusieurs degrés, par régression.

Sondage stratifié (ST)

Le sondage aléatoire simple sans remise ne prend 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 une est **la stratification**.
Différents raisons pour utiliser la stratification :

1. éviter les "mauvais échantillons" qui sont possibles avec un plan SASSR (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 SASSR (si la stratification est bien réalisée).

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 stratifier par classe d'âge.
2. dans une étude sur la concentration des plantes, il faut stratifier 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.

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.

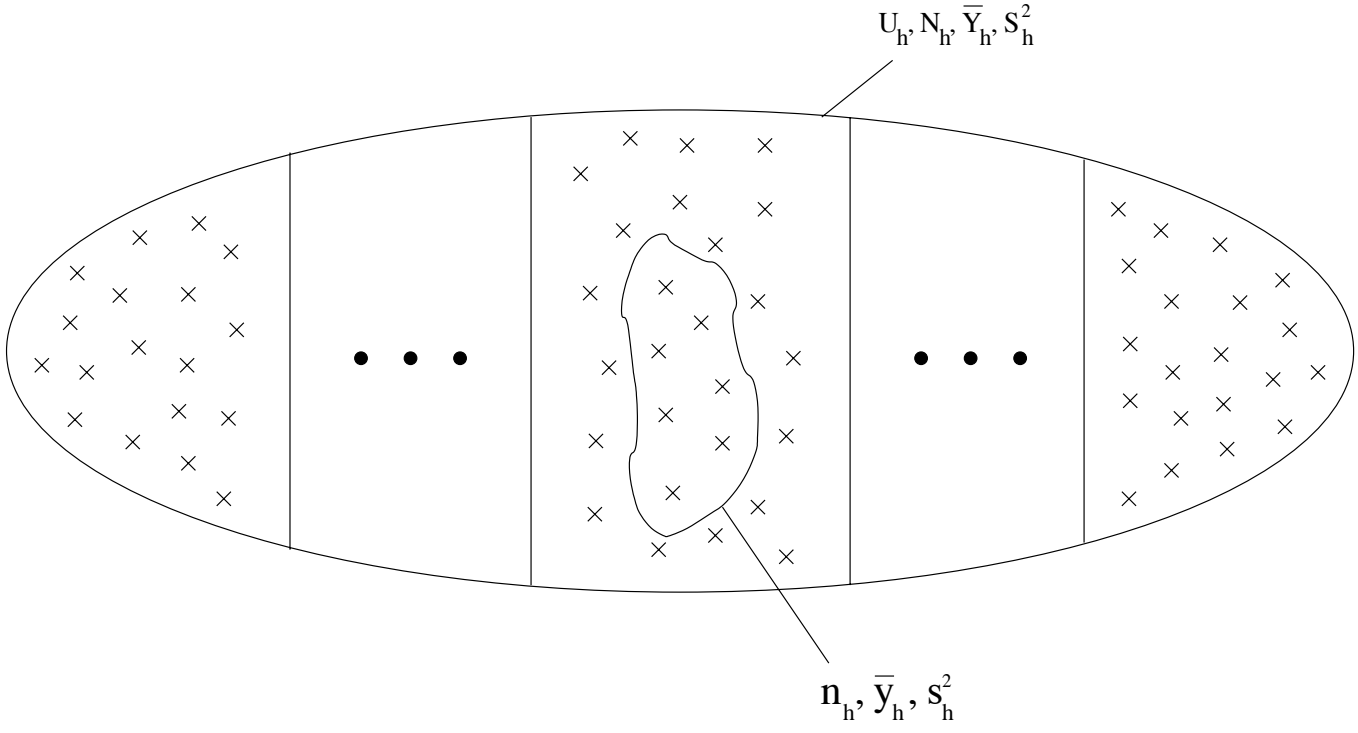


FIGURE : Plan stratifié

Précision du plan stratifié

- La moyenne $\bar{Y}$ est estimée par

$$\bar{y} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_{U_h}$$

- La variance est égale à $V(\bar{y}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 V(\bar{y}_{U_h})$.

En général, le plan stratifié est meilleur que le plan aléatoire simple sans remise.

Décomposition de la variance :

Variation totale = variation intra-strates + variation inter-strates

$$\sum_{k \in U} (y_k - \bar{y}_U)^2 = \sum_{h=1}^{H} N_h (\bar{y}_{U_h} - \bar{y}_U)^2 + \sum_{h=1}^{H} (N_h - 1) S_{yU_h}^2$$

La variance avec un plan stratifié est égale (à un facteur près) à la variance inter-strates qui est plus faible que la variance totale si les strates sont bien réalisées : les individus à l'intérieur de strates se rassemblent et diffèrent de ceux des autres strates.

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 ?

1. **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 :

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

2. **l'allocation optimale** : on choisi n_h individus dans la strate h proportionnel à la dispersion de la variable dans la strate h :

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

Petit exemple pour l'allocation proportionnelle :

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

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

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

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

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

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.

Sondages à probabilités inégales

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.

- Pour les deux exemples, les probabilités de sélection sont proportionnelles à la taille du ménage ou de l'entreprise.
- Plus général, on a besoin de connaître une extra variable ou variable auxiliaire très corrélée avec la variable d'intérêt pour chaque individu de la population.
- Si les probabilités de sélection sont $p_1, \ldots, p_N$ de somme 1 et proportionnelles avec une variable X de valeurs x_k, on obtient

$$p_k = \frac{x_k}{t_x}, \quad t_x \text{ est le total de } X$$

- les probabilités de sélection sont différentes de probabilités d'inclusion dans l'échantillon (une unité peut être sélectionnée plusieurs fois).
- pour estimer le total t_y on utilise

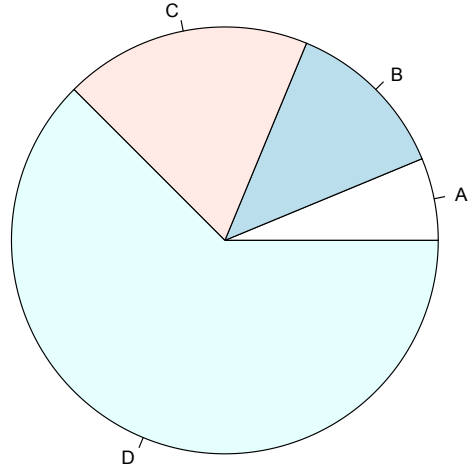
$$\sum_{i=1}^{n} w_i y_i, \quad w_i = 1/(np_i)$$

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 $\hat{t} = \frac{t_1}{p_1} = 16 \cdot 11 = 176$



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.

Sondages par grappes

Le plan aléatoire simple sans remise et le plan stratifié sont des plans “directs” : on interroge les individus lors d'une seule étape. En pratique, ce type de sondage est souvent impossible à cause de plusieurs facteurs :

1. l'absence d'une base de sondage ou le prix pour en fabriquer une est trop élevé ;
2. les éléments de la population sont trop dispersés sur une aréa géographique et le coût d'interroger les individus serait élevé aussi. Dans ce type de situations, les taux de nonréponse sont aussi importants.

Exemple : On veut savoir la proportion d'élèves de terminale de Dijon qui vont faire des études de maths. On sélectionne pour cela un échantillon de lycées et dans les lycées choisis, on interroge tous les élèves de terminale.

Description du plan par grappes

- La population est partagée dans G sous-populations disjointes (différentes de strates) appelées *grappes*. On possède (ou on peut en obtenir facilement) une base de sondage pour cette population des grappes.
- On sélectionne un échantillon de grappes et pour chaque grappe choisi, on interroge tous les individus y appartenant.

Avantages et inconvénients du plan par grappes

- **Avantages** :
  1. on réduit considérablement les coûts et la durée d'une enquête ;
  2. la population d'étude peut être dispersée sur une grande surface géographique ou est présentée (de façon naturelle) sous la forme de grappes :
     dans notre exemple : il est plus facile d'avoir dans un premier temps, une liste des lycées de Dijon et ensuite, pour les lycées sélectionnés, d'obtenir une liste des élèves de terminale que d'avoir une liste d'élèves de terminales pour tous les lycées de Dijon.
- **Désavantages** : le plan par grappes est souvent moins efficace que le sondage aléatoire simple sans remise

  variation totale=variation inter-grappes+variation intra-grappes

  et la variance sous un plan par grappes est liée à la variance **inter-grappes** qui est d'habitude grande.

Strates ou grappes ?

- Sondage stratifié
  1. chaque individu est dans une seule strate ;
  2. on prend un SASSR dans chaque strate ;
  3. la variance de $\bar{y}$ dépend de la variabilité de y à l'intérieur de chaque strate ;
  4. pour une meilleure précision, il faut avoir des strates très homogènes mais de moyennes très différentes ;
- Sondage par grappes
  1. chaque individu est dans une seule grappe ;
  2. on prend un SASSR de grappes ;
  3. la variance de $\bar{y}$ dépend de la variabilité de la moyenne de y entre les grappes ;
  4. pour une meilleure précision, il faut avoir des grappes très hétérogènes mais de la même moyenne ;

Prise en compte de l'information auxiliaire : estimateur par le ratio et poststratifié

Information auxiliaire : une caractéristique de la théorie des sondages.
Ce sont des variables disponibles dans la base de sondage
Ces variables peuvent servir de deux façon pour améliorer la qualité des estimations :

- lors de la conception du plan de sondage : la plan stratifié, à probabilités inégales, par grappes ou à plusieurs degrés ;
- lors de la conception d'un nouvel estimateur.

L'estimateur par le ratio

Laplace (1802) a été le premier a avoir proposé faire un sondage au lieu d'une énumération exhaustive ;

Objectif : estimer la population de la France.

Méthode : l'estimateur par ratio (une "règle de trois") :

- il prend un échantillon de taille 30 de communes de toute la France et il obtient un total de $\hat{t}_{habitants,30com} = 2037615$ habitants ;
- il connaît également le nombre de naissances dans ces 30 communes : $\hat{t}_{naissances,30com} = 71866$ ce qui implique que il y a

$$\text{une naissance pour} \quad \frac{2037615}{71866} = 28.35 \quad \text{personnes}$$

- il connaît le nombre total de naissances par an en 1802 et il fait le raisonnement que les communes avec beaucoup d'habitants vont avoir plus de naissances, il estime le nombre total d'habitants comme

$$t_{habitants} = t_{naissances} \times \frac{\hat{t}_{habitants,30com}}{\hat{t}_{naissances,30com}}$$

Dans l'exemple de Laplace : le nombre total de naissances est l'information auxiliaire qui est bien corrélée avec la variable nombre d'habitants.
L'estimateur proposé par Laplace est beaucoup utilisé pour estimer des totaux : et il s'appelle **l'estimateur par ratio** :

$$\hat{t}_{y,ratio} = t_x \times \frac{\hat{t}_y}{\hat{t}_x} = t_x \times \hat{R}$$

Propriétés : en général, cet estimateur est biaisé mais pour des échantillons de grande taille, ce biais est très proche de zéro. Si l'information auxiliaire est très corrélée avec la variable d'intérêt alors, l'erreur quadratique moyenne de $\hat{t}_{y,ratio}$ est plus petite que si on n'avait pas utilisé d'information auxiliaire.

Exemple (Lohr, 1999)

Supposons que nous voulons estimer le nombre de poissons péchés au long d'un lac. On considère un plan aléatoire simple sans remise de taille $n = 2$ dans une population de taille $N = 4$:

- nous avons $N = 4$ endroits pour pécher ;
- on connaît le nombre x_i de filets qui se trouvent à chaque endroit pour pécher.

Les valeurs pour toute la population se trouvent dans le tableau suivant :

endroit, i	1	2	3	4	total
filet, x_i	4	5	8	5	$t_x = 22$
nombre poissons, y_i	200	300	500	400	$t_y = 1400$

L'estimateur par le ratio est $\hat{t}_{yrat} = t_x \times \frac{\sum_s y_i}{\sum_s x_i}$ où s est un des 6 échantillons possibles de taille 2. Par exemple, si $s = (1, 2)$ $\hat{t}_{yrat} = 22 \times \frac{200+300}{4+5} = 1222$.

échantillon	$\hat{t}_{yrat}$	$(\hat{t}_{yrat} - t_y)^2$
$(1, 2)$	1222	31684
$(1, 3)$	1283	13689
$(1, 4)$	1467	4489
$(2, 3)$	1354	2116
$(2, 4)$	1540	19600
$(3, 4)$	1523	15129

L'espérance de $\hat{t}_{yrat}$ est la moyenne arithmétique des valeurs possibles de $\hat{t}_{yrat}$ pour chaque échantillon, $E(\hat{t}_{yrat}) = 1398.17$ et le biais est

$$1398.17 - 1400 = -1.83$$

L'erreur quadratique moyenne est donnée par

$$EQR(\hat{t}_{yrat}) = E(\hat{t}_{yrat} - t_y)^2 = \frac{1}{6}\sum_s (\hat{t}_{yrat} - t_y)^2 = 14451,2$$

Considérons maintenant l'estimateur par les valeurs dilatées qui ne prend pas en compte l'information auxiliaire : $\hat{t}_y = \frac{N}{n}\sum_s y_i$:

échantillon	$\hat{t}_y$	$(\hat{t}_y - t_y)^2$
$(1,2)$	$2 \cdot (200 + 300) = 1000$	160000
$(1,3)$	1400	0
$(1,4)$	1200	40000
$(2,3)$	1600	40000
$(2,4)$	1400	0
$(3,4)$	1800	160000

Le biais de cet estimateur est 0 mais sa variance est 66 667. Alors, l'estimateur par le ratio est légèrement biaisé mais son EQR est nettement inférieure à celle de l'estimateur par les valeurs dilatées qui est sans biais.

L'estimateur par le ratio plus précis : pourquoi ?

- Nous voulons estimer le total $t_y = \sum_{k \in U} y_k$;
- On considère un échantillon SRSWR de taille n;
  - sans inf. aux. et en utilisant l'estimateur par les valeurs dilatées :

$$\hat{t}_y = \frac{N \sum_s y_k}{n}$$

avec une variance estimée égale à

$$N^2 \frac{1-f}{n} S_{ys}^2 = N^2 \frac{1-f}{n} \frac{1}{N-1} \sum_s (y_k - \bar{y}_s)^2;$$

  - avec l'inf.aux. et l'estimateur par ratio :

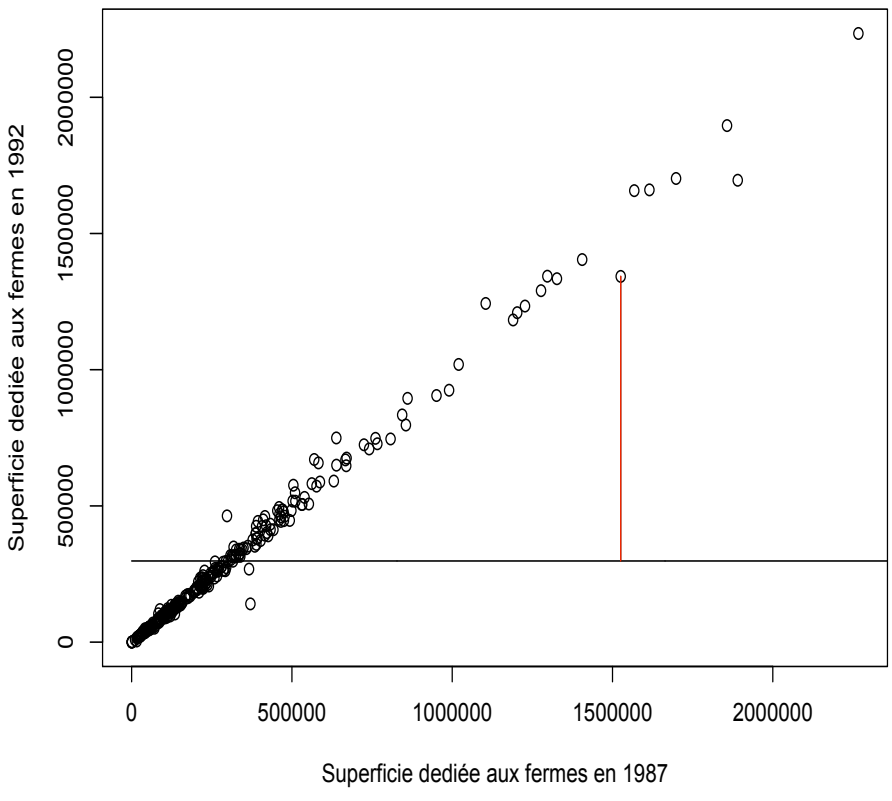
$$\hat{t}_{yrat} = t_x \times \frac{\hat{t}_y}{\hat{t}_x}$$

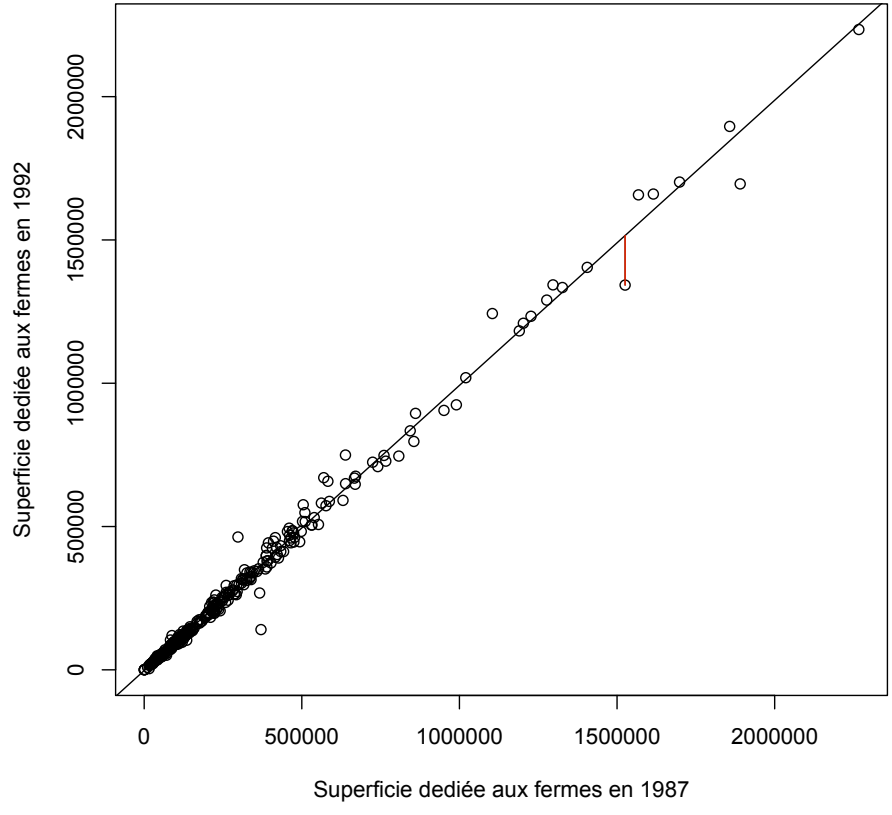
avec une variance estimée égale à

$$N^2 \frac{1-f}{n} S_{y-\hat{R}x,s}^2 = N^2 \frac{1-f}{n} \frac{1}{N-1} \sum_s (y_k - \hat{R}x_k)^2$$

car $\bar{y}_s - \hat{R}\bar{x}_s = 0$.

Si on compare les deux formules, dans la première situation nous avons la somme des carrés des écarts de y_k à la moyenne $\bar{y}_s$ et dans la deuxième situation, nous avons la somme des carrés des écarts de y_k à la droite de régression $y = \hat{R}x$





Poststratification

Considérons l'exemple suivant (Lohr, 1999) :

- on veut estimer le nombre total d'étudiants qui veulent être professeurs après leurs études dans une population de 4000 étudiants ;
- on prend un échantillon aléatoire simple sans remise de taille $n = 400$;
- on sait par ailleurs que dans la population totale on a 2700 femmes et 1300 hommes ;
- dans notre échantillon, il y a 240 femmes dont 84 veulent être des professeurs et 160 hommes dont 40 veulent être des professeurs

On utilise cette information et on obtient l'estimation suivante :

$$\frac{84}{240} \times 2700 + \frac{40}{160} \times 1300 = 1270$$

Il s'agît d'une règle de trois à l'intérieur de chaque groupe !

Propriétés

- La méthode d'estimation utilisée est appelée *poststratification* car utilise l'information auxiliaire (le fait d'appartenir à un certain groupe) après le tirage de l'échantillon.
- Dans la plupart des cas, l'estimateur poststratifié est meilleur que l'estimateur par les valeurs dilatées et il est équivalent au plan stratifié avec l'allocation proportionnelle.

  variation totale=variation inter-groupes+variation intra-groupes

  La poststratification réduit la variance si la variation entre groupes est grande.
- Dans l'exemple donné, l'information auxiliaire est le fait que la population est plus homogène à l'intérieur de chaque groupe que dans la population totale.

Quand utiliser la poststratification ?

1. dans le cas des sondages à plusieurs items : on peut utiliser différentes variables auxiliaires pour différentes variables d'intérêt.
2. dans le traitement de la nonréponse ;
3. l'appartenance d'un individu à un certain groupe n'est pas connue au moment de la mise en place de l'enquête.

Exemple : On veut estimer la somme moyenne dédiée à la nourriture/mois dans les ménages américaines. Nous avons la distribution de la taille des ménages comme suit :

taille du ménage	pourcentage
1	25.75
2	31.17
3	17.50
4	15.58
5+	10.00

Malheureusement, la base de sondage ne contient pas l'information concernant la taille de chaque ménage, par conséquent on ne peut pas faire une stratification.
Mais on peut faire une poststratification.