Introduction à la théorie des sondages
Challenge "Graines de sondeur"

Camelia Goga

IMB, Université de Bourgogne-Dijon
camelia.goga@u-bourgogne.fr

Dijon, novembre 2015

1. Population, échantillon : définitions
2. Méthodes pour sélectionner un échantillon
   - Sondage de Bernoulli (BE)
   - Sondage systématique (SY)
   - Sondage aléatoire simple sans remise (SAS)
3. Méthodes d'estimation
4. Améliorer l'estimation
   - Sondage stratifié
   - L'estimateur par le ratio et poststratifié
5. La non-réponse
6. Présentation des sujets
7. Quelques compléments théoriques

Sondages versus recensements

- Faire une **enquête par sondage** : interroger seulement une partie d'une population, appelée **échantillon** pour calculer plusieurs quantités d'intérêt (le revenu moyen ou médian, la proportion d'étudiants, ...).
- Quand on interroge **toute** la population, on parle de **recensement**.

sondage	vs	recensement
calcul approché	vs	calcul exact
estimation	vs	vraie valeur

- Faire un recensement coûterait trop cher, prendrait beaucoup de temps. Le dernier recensement de la population française a été réalisé en 1999.

Domaines d'application des sondages

- Souvent, la notion de sondage est assimilée aux sondages d'opinion (sondages politiques).
- Les sondages sont utilisés dans beaucoup d'autres domaines :
  - l'Insee
  - dans l'Education Nationale
  - à Médiamétrie
  - à la Poste, à EDF
  - en agriculture, tourisme, santé...

Petit historique

- Pour la première fois, Laplace propose en 1802 une estimation du nombre de personnes qui habitent en France au lieu de faire un recensement.
- Cette idée d'estimer au lieu de calculer la valeur exacte a été perçue comme peu scientifique.
- Suite à un article de Neyman en 1934 développant les fondements probabilistes de la théorie des sondages, la théorie des sondages a été reconnue.

Etapes d'une enquête par sondage

1. Définir la population cible
2. Définir la base de sondage
3. Décider le plan d'échantillonnage et la taille de l'échantillon
4. Sélectionner l'échantillon
5. Calculer des estimations et les intervalles de confiance associés ;

Population

- La population cible est une collection d'éléments sur lesquels l'information est requise. On note la population cible par $U = \{u_1, u_2, \ldots, u_N\}$ dont la taille N peut être connue ou pas ;
  Un individu est un élément de U qui peut être repéré précisément et sans aucune ambiguïté : l'identifiant k :

$$U = \{1, 2, \ldots, k, \ldots, N\}$$

- Une fois que la population cible a été définie, l'étape suivante est d'établir une **liste** ou base de sondage complète (souvent difficile à réaliser) de tous les individus de cette population.
  Problèmes : sous-couverture, sur-couverture ;
  ex. : l'annuaire téléphonique, la liste SIRENE des entreprises françaises, la liste des établissements scolaires ...

- L'information requise, i.e. les variables d'intérêt, doivent pouvoir être mesurée sur chaque individu.

- Une variable, notée par Y, est **quantitative** ou bien **qualitative** :
  1. **quantitative** : le revenu, la note à une certaine matière, l'audience sur une certaine chaîne de TV, ...
  2. **qualitative** : le sexe, la catégorie socio-professionnelle, ...

Quantités d'intérêt

- **un total** : on veut connaître le nombre total de lettres vertes, le nombre total de touristes dans la région Bourgogne, etc...

$$t_y = y_1 + y_2 + \ldots + y_N = \sum_{k \in U} y_k$$

ex : y_k =le nombre de lettres vertes dans le bureau de poste k et U = la population de bureaux de poste de taille N.

- **une moyenne** : le revenu moyen, les dépenses moyennes des ménages français, ...

$$\overline{y}_U = \frac{1}{N} \sum_{k \in U} y_k$$

ex : y_k = le revenu de l'individu k et U = la population active.

- **une proportion** (cas particulier d'une moyenne) : P=la proportion d'élèves de TS qui veulent faire des études de mathématiques, le taux de chômage, ...

$$P = \frac{1}{N}\sum_{k\in U} y_k$$

où $y_k = 1$ si l'élève k veut faire des maths et $y_k = 0$ sinon et U =la population des éléves de TS.

- des quantités plus compliquées comme **la variance** ou **les quantiles** (ex. la médiane) d'une variables quantitative (revenu, notes,...).
La variance empirique (corrigée) d'une variable y est

$$S^2_{yU} = \frac{1}{N-1}\sum_{k\in U}(y_k - \overline{y}_U)^2$$

Échantillon

Les quantités d'intérêt $t_y, \overline{y}_U, \ldots$ ne peuvent pas être calculées car on ne dispose pas de toute la population. Dans ces conditions, on **estime** (approche) ces quantités à partir d'un **échantillon**.

- Pour obtenir un échantillon, on sélectionne au hasard des numéros dans la base de sondage.
- Il existe beaucoup de méthodes ou *plans de sondage* pour "sélectionner au hasard" un échantillon. On en verra quatre : le sondage de **Bernoulli**, le **sondage systématique**, le **sondage sans remise** et le **sondage stratifié**.
- Pour chaque individu sélectionné, on mesure la valeur de la variable. Si l'individu ne réponde pas, alors on a de la **non-réponse**.

Important : une quantité clé dans le processus d'estimation est π_k=**la probabilité d'un individu d'être dans un échantillon**.

Echantillonnage probabiliste versus non-probabiliste

Il existe aussi des méthodes **non-probabiliste** : les individus sont sélectionnés selon des critères subjectifs (dépendant de l'expérience du praticien) :

- par quotas
- "convenience sampling"
- "judgemental sampling"
- "snowball sampling".

En général, les méthodes non-probabilites sont moins chères que les autres mais les résultats ne peuvent pas être inférés à toute la population.

Exemple (réalisé lors du stage "Trop fort les maths" 2013-2014)

Objectif : estimer le nombre de mots du livre *Les femmes et la science* de G. Chazal.

- Population = ensemble des pages
- taille de la population $N = 90$.
- individu = une page
- variable y_k = nombre de mots sur la page k.

On sélectionne un échantillon de taille $n = 10$ selon les trois méthodes : **sondage systématique**, **sondage sans remise** et **sondage stratifié**.

Dans cette étude, on connaît le nombre total de mots, $t_y = 35170$.

Important : on ne peut pas comparer les différents plans de sondage à partir d'un seul échantillon, il faut tirer beaucoup d'échantillons (ex. 1000).

1. Population, échantillon : définitions
2. Méthodes pour sélectionner un échantillon
   - Sondage de Bernoulli (BE)
   - Sondage systématique (SY)
   - Sondage aléatoire simple sans remise (SAS)
3. Méthodes d'estimation
4. Améliorer l'estimation
   - Sondage stratifié
   - L'estimateur par le ratio et poststratifié
5. La non-réponse
6. Présentation des sujets
7. Quelques compléments théoriques

Sondage sans remise versus avec remise

Deux types de plans de sondage :

1. **sans remise** : une unité est "**enlevée**" de la liste une fois qu'elle a été sélectionnée. Un échantillon ne contient que des unités distinctes.
2. **avec remise** : une unité est "**remise**" dans la liste une fois qu'elle a été sélectionnée. On peut sélectionner une unité plusieurs fois.

En pratique, on utilise le plus souvent les plans de sondage sans remise.

1 Population, échantillon : définitions

2 Méthodes pour sélectionner un échantillon
- Sondage de Bernoulli (BE)
- Sondage systématique (SY)
- Sondage aléatoire simple sans remise (SAS)

3 Méthodes d'estimation

4 Améliorer l'estimation
- Sondage stratifié
- L'estimateur par le ratio et poststratifié

5 La non-réponse

6 Présentation des sujets

7 Quelques compléments théoriques

Sondage de Bernoulli (BE)

On a une pièce parfaite. L'échantillon sera choisi selon le procédé suivant :

- On jette la pièce N fois (une fois pour chaque individu de notre population) ; on note les N résultats ;
- On décide de sélectionner les individus pour lesquels on a obtenu "PILE". Chaque individu a une probabilité $\pi = \frac{1}{2}$ d'être sélectionné dans l'échantillon.

L'échantillon peut être choisi avec une probabilité $\pi \in (0, 1)$ différente de $1/2$ (plus de détails à la fin du cours).

Avantage du plan BE : plan sans remise et indépendance entre les tirages des individus ; très simple à mettre en oeuvre.
Utilisé dans le traitement de la non-réponse (voir chapitre 5).

Inconvénients du plan BE : la taille aléatoire de l'échantillon.

Algorithme pour le plan BE

for $k = 1$ to N **do**
  $u \leftarrow \mathcal{U}_{(0,1)}$
  if $u < \pi$ **then**
    sélectionner l'unité k dans l'échantillon
  else
    passer à l'unité suivante
  end if
end for

Exemple (avec le logiciel R) :

```
x=runif(10)
0.325 0.225 0.897 0.367 0.745 0.813 0.732 0.254 0.178 0.645
as.numeric(x<1/2)
 1 1 0 1 0 0 0 1 1 0
```

1 Population, échantillon : définitions

2 Méthodes pour sélectionner un échantillon
- Sondage de Bernoulli (BE)
- Sondage systématique (SY)
- Sondage aléatoire simple sans remise (SAS)

3 Méthodes d'estimation

4 Améliorer l'estimation
- Sondage stratifié
- L'estimateur par le ratio et poststratifié

5 La non-réponse

6 Présentation des sujets

7 Quelques compléments théoriques

Sondage systématique (SY)

On veut sélectionner un échantillon de taille n.

- On calcule le pas $a = \frac{N}{n}$ et on suppose qu'il est entier.
- On sélectionne au hasard un nombre entier entre 1 et a. Soit k le nombre choisi.
- L'échantillon est composé des individus :

$$s = \{k, k + a, k + 2a, \ldots, k + (n - 1)a\}$$

Illustration sur l'exemple du livre

- Le pas $a = \frac{N}{n} = 9$. On sélectionne au hasard un nombre parmi $1, 2, 3, \cdots, 9$.
- Si le nombre choisi est 2, alors l'échantillon est constitué des pages $2, 11, 20, \cdots, 83$.

1	2	3	$\cdots$	9
10	11	12	$\cdots$	18
19	20	21	$\cdots$	27
$\cdots$	$\cdots$	$\cdots$	$\cdots$	
82	83	84	$\cdots$	90

Remarque : soit
π_k = la probabilité que l'individu k appartient à un échantillon.
Dans le cas du plan SY, on a $\pi_k = \frac{1}{a} = \frac{n}{N}$ pour tous les $k \in U$.

La méthode fractionnaire

Il existe des algorithmes qui permettent de choisir n individus même si la fraction $\frac{N}{n}$ n'est pas un entier.

Exemple d'algorithme (la méthode fractionnaire) : $a = N/n$

Algorithme :

1. On tire u un nombre aléatoire uniforme entre 0 et a; (par exemple $u = a \cdot \varepsilon$ avec ε un nombre aléatoire entre 0 et 1 ;
2. Le départ aléatoire est $1 + [u]$ où $[u]$ est la partie entière de u;
3. L'échantillon est composé des unités ℓ vérifiant $1 + [u + \ell a]$ avec $\ell = 1, \ldots, n - 1$

Exemple

- Soit une population de taille $N = 13$ et nous voulons sélectionner un échantillon de taille $n = 3$ selon un plan systématique ; alors $a = 13/3$.
- Supposons que le nombre aléatoire entre 0 et a est $u = 2,344$.
- Le départ aléatoire est $1 + [u] = 3$. Donc, l'individu 3 est dans l'échantillon.
- Les deux autres individus sélectionnés sont

$$1 + [u + 1 \cdot a] = 7$$

$$1 + [u + 2 \cdot a] = 11$$

- L'échantillon est formé par les individus 3, 7 et 11.

Avantages du plan SY

- très simple à mettre en oeuvre. Il est souvent utilisé dans les enquêtes par téléphone, internet ; Insee utilise ce plan pour le tirage des ménages dans les communes de l'échantillon maître, pour le tirage des logements neufs
- on peut également fixer à l'avance le pas a et non la taille de l'échantillon n; dans ces conditions, si $N = n \cdot a + c$, alors l'échantillon peut être de taille n ou $n + 1$.

Inconvénients du plan SY

Si la population n'est pas bien "mélangée", on risque de tomber sur des individus qui se ressemblent beaucoup. Par conséquent, le sondage n'est pas très efficace dans cette situation.

- Si la liste d'individus de la population est présentée de la façon suivante : H(omme), F(emme), H, F, ... et si on sélectionne l'échantillon avec un pas de 2, il sera composé uniquement d'hommes ou uniquement de femmes.
- Si on veut connaître les dépenses des ménages pour le chauffage et si on échantillonne avec un pas de 4 des appartements dans un immeuble ayant 4 appartements par étage, alors on sous-estime ces dépenses si le départ est un appartement situé au sud.

1 Population, échantillon : définitions

2 Méthodes pour sélectionner un échantillon
- Sondage de Bernoulli (BE)
- Sondage systématique (SY)
- Sondage aléatoire simple sans remise (SAS)

3 Méthodes d'estimation

4 Améliorer l'estimation
- Sondage stratifié
- L'estimateur par le ratio et poststratifié

5 La non-réponse

6 Présentation des sujets

7 Quelques compléments théoriques

Sondage aléatoire simple sans remise (SAS)

De façon intuitive, on peut voir la population comme N billes numérotées de 1 à N placées dans une urne. On effectue n tirages sans remise dans l'urne.

Algorithme (exemple du livre)

- On sélectionne au hasard une page parmi les $N = 90$. On la retire de la liste.
- Une autre page est sélectionnée au hasard parmi les 89 restantes et ensuite retirée de la liste.
- Cette procédure est répétée 10 fois pour obtenir un échantillon de $n = 10$ pages.

Remarque : en pratique, l'échantillon n'est pas choisi en utilisant cet algorithme car il est trop lent surtout pour des populations de grandes tailles ; d'autres algorithmes plus performants sont utilisés.

Exemple d'algorithme plus rapide

Le modéle d'urne est implémenté avec Python et Algobox :
`http://www.irem.univ-mrs.fr/IMG/pdf/tirage_sans_remise-2.pdf`

Sinon, en Python existe la fonction `random.sampling`

Algorithm 2

for $k = 1$ to N **do**
  $u[k] \leftarrow U_{(0;1)}$
end for
Trier la base selon les valeurs décroissantes de u
Sélectionner les n premiers individus

Inconvénient : tri coûteux si N est grand.
Avantage : permet de tirer des échantillons sans recouvrement.

- **Avantages du plan SAS** : c'est une méthode simple et naturelle pour choisir l'échantillon ;

- **Inconvénients du plan SAS** : il peut être coûteux en termes de frais d'enquête ; de plus, l'échantillon peut ne pas être représentatif pour la population étudiée :
  - un échantillon parmi les personnes vivant en France peut contenir des individus situés seulement en Bretagne ou en Provence ;
  - on peut tomber sur un "mauvais" échantillon : échantillon constitué uniquement de pages de début ou de fin de chapitres, donc contenant moins de mots.
  - de point de vue estimation : pour des variables très dispersées (ex. le revenu), l'estimateur a une grande variance (voir chapitre 3).

1 Population, échantillon : définitions

2 Méthodes pour sélectionner un échantillon
- Sondage de Bernoulli (BE)
- Sondage systématique (SY)
- Sondage aléatoire simple sans remise (SAS)

3 Méthodes d'estimation

4 Améliorer l'estimation
- Sondage stratifié
- L'estimateur par le ratio et poststratifié

5 La non-réponse

6 Présentation des sujets

7 Quelques compléments théoriques

Quel est l'aléa ?

- On veut estimer (approcher) le total **inconnu**

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

  à partir de l'échantillon s sélectionné selon un plan de sondage (systématique, aléatoire simple sans remise, ...) ;

- **Important** : les y_k sont considérés **fixés, non-aléatoires** !
- L'aléa est dans le fait d'être échantillonné ou pas !

Indicatrice d'appartenance à l'échantillon

- On considère la variable aléatoire indicatrice d'appartenance à l'échantillon

$$I_k(s) = \begin{cases} 1 & \text{si l'individu k} \in s \\ 0 & \text{sinon} \end{cases}$$

- La variable I_k est une variable de Bernoulli de paramètre π_k = la probabilité que k soit sélectionné ;

$$\begin{aligned} E(I_k) &= 1 \cdot P(I_k = 1) + 0 \cdot P(I_k = 0) \\ &= \pi_k \end{aligned}$$

- **Important** : en général, les variables $I_k, k \in U$ ne sont pas indépendantes sauf pour certains plans comme le plan BE !

Estimateur d'Horvitz-Thompson du total t_y

En théorie des sondages, pour estimer un total, on utilise des sommes pondérées.
L'estimateur d'Horvitz-Thompson (on suppose que $\pi_k > 0, k \in U$) :

- $\hat{t}_y = \sum_{k \in s} w_k y_k = \sum_{k \in s} \frac{1}{\pi_k} y_k = \sum_{k \in U} \frac{1}{\pi_k} y_k I_k$;
- le poids $w_k = \frac{1}{\pi_k} > 1$, et il peut être interprété comme le nombre de personnes de U représentées par l'individu k;
- les w_k ne font appel qu'aux π_k, mais si on dispose d'informations supplémentaires, alors on peut considérer des poids w_k plus compliqués (voir section 4).

Probabilités d'inclusion pour les plans BE, SY et SAS

Considérons les plans de sondages présentés auparavant :

- **le plan BE** : $\pi_k = \pi$ pour tous les $k \in U$;
- **le plan SY** : $\pi_k = \frac{1}{a}$ pour tous les $k \in U$, où a est le pas ;
- **le plan SAS** :

$$\begin{aligned} \pi_k &= \frac{\text{no. d'échantillons sans remise de taille n contenant k}}{\text{no. total d'échantillons sans remise de taille n}} \\ &= \frac{C_{N-1}^{n-1}}{C_N^n} = \frac{n}{N} \quad \text{pour tous les} \quad k \in U \end{aligned}$$

Remarque : pour chacun de ces trois plans, les π_k sont les mêmes pour tout $k \in U$.

Estimations avec les plans BE, SY et SAS

Soit s un échantillon obtenu selon une des méthodes suivantes :

1. **Le plan BE** : $s = \{8, 9, 14, 16, 25, 29, 43, 45, 46, 50, 69, 72\}$ et le nombre de mots $\{419, 464, 250, 445, 81, 435, 447, 440, 423, 484, 175, 414\}$.
   Alors, $\hat{t}_y^{BE} = 40293$.

2. **Le plan SY** : $s = \{2, 11, 20, 29, 38, 47, 56, 65, 74, 83\}$ et le nombre de mots $\{402, 410, 449, 435, 429, 110, 358, 442, 407, 337\}$.
   Alors, $\hat{t}_y^{SY} = 34011$.

3. **Le plan SAS** : $s = \{86, 59, 19, 30, 36, 10, 52, 47, 35, 51\}$ et le nombre de mots $\{418, 449, 407, 448, 40, 485, 449, 110, 331, 405\}$.
   Alors, $\hat{t}_y^{SAS} = 31878$.

Qualité d'un estimateur

Soit θ une quantité inconnue (total, moyenne, ...) et $\hat{\theta}$ son estimateur.
Les indicateurs pour comparer les estimateurs sont (voir aussi le glossaire mis en ligne) :

- **le biais** de $\hat{\theta}$: l'écart entre la moyenne des valeurs de $\hat{\theta}$ prises pour tous les échantillons possibles, $E(\hat{\theta})$, et la vraie valeur θ.

$$\text{Biais}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

  "sans biais" signifie que l'estimateur est bon "en moyenne" et non que l'estimation obtenue à partir d'**un** échantillon est la vraie valeur.

- **la variance** de $\hat{\theta}$: =un indicateur de la dispersion des valeurs de $\hat{\theta}$ autour de sa moyenne, $E(\hat{\theta})$,

$$\text{Var}(\hat{\theta}) = E(\hat{\theta} - E(\hat{\theta}))^2$$

- **l'erreur quadratique moyenne** (EQM) de $\hat{\theta}$: =un indicateur de la dispersion des valeurs de $\hat{\theta}$ autour de sa vraie valeur.

$$EQM(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

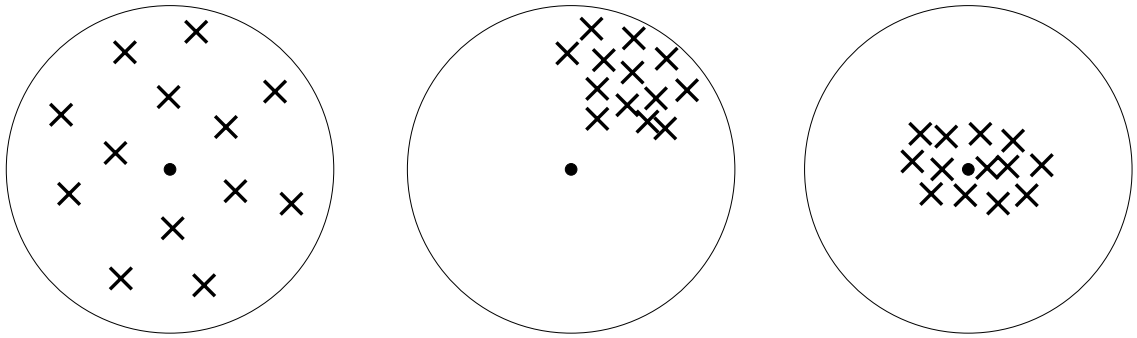


FIGURE : Biais et précision

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).

Dans certaines situations, on préfère un estimateur légèrement biaisé mais avec une EQM plus faible.

Et pour l'estimateur d'Horvitz-Thompson

- $\hat{t}_y$ est sans biais pour t_y :

$$E(\hat{t}_y) = E\left(\sum_{k\in U}\frac{1}{\pi_k}y_k I_k\right) = \sum_{k\in U}\frac{1}{\pi_k}y_k \underbrace{E(I_k)}_{\pi_k} = \sum_{k\in U} y_k$$

- La variance de $\hat{t}_y$ est inconnue (voir chapitre 7) et elle doit être estimée ;
- Pour des variables de type taille (ex.le revenu) et certains plans de sondage (SAS, SYS), l'estimateur d'HT est inefficace car sa variance est grande. Dans ces conditions, il existe des estimateurs biaisés (l'estimateur par le ratio ou poststratifié) mais plus efficaces.

En résumé

Soit $f = \frac{n}{N}$ le taux de sondage.

plan	estimateur	variance	estimateur sans biais de la variance
plan SAS	$\hat{t}_y = N\overline{y}_s$	$N^2 \frac{1-f}{n} S^2_{yU}$	$N^2 \frac{1-f}{n} S^2_{ys}$
plan SY	$\hat{t}_y = a\sum_{k\in s} y_k$	$a\sum_{r=1}^{a}(t_{s_r} - \overline{t})^2$	estimateurs biaisés
plan BE	$\hat{t}_y = \frac{1}{\pi}\sum_{k\in s} y_k$	$\frac{1-\pi}{\pi}\sum_{k\in U} y_k^2$	$\frac{1-\pi}{\pi^2}\sum_{k\in s} y_k^2$

$t_{s_r} = \sum_{k\in s_r} y_k$ et $\overline{t} = \sum_{r=1}^{a} t_{s_r}/a$.

Remarques :

1. l'estimation de la moyenne $\overline{y}_U$ est obtenue en divisant $\hat{t}_y$ par N et les variances par N^2 ; en particulier, avec un plan SAS on obtient $\overline{y}_s = \frac{1}{n}\sum_{k\in s} y_k$;
2. l'estimation d'une proportion P avec un SAS est p=la proportion dans l'échantillon ;

Quelques commentaires sur le plan SAS

- Le facteur $1 - f$ s'appelle *correction en population finie* : très important pour des populations de tailles relativement petites ;
- Pour des populations de grandes tailles, c'est la taille de l'échantillon n qui donne la précision et non le taux de sondage f, c'est à dire l'estimation issue d'un échantillon de taille 1000 dans une population de taille 100.000 aura quasiment la même précision que dans une population de taille 100.000.000
- Le fait que la variable d'intérêt soit peu ou très dispersée a beaucoup d'influence sur la précision.
- Pour des populations de très grandes tailles et si n est très petit par rapport à N, (i.e. $\frac{n}{N} \simeq 0$), alors le sondage sans remise peut être considéré comme un sondage avec remise.

1 Population, échantillon : définitions

2 Méthodes pour sélectionner un échantillon
- Sondage de Bernoulli (BE)
- Sondage systématique (SY)
- Sondage aléatoire simple sans remise (SAS)

3 Méthodes d'estimation

4 Améliorer l'estimation
- Sondage stratifié
- L'estimateur par le ratio et poststratifié

5 La non-réponse

6 Présentation des sujets

7 Quelques compléments théoriques

- Une façon d'améliorer les estimations est de chercher et d'utiliser des informations supplémentaires sur la population étudiée en lien avec l'objectif de l'enquête. Cette information s'appelle **information auxiliaire**.
- Les bases de sondages contiennent toujours ce type d'information (âge, sexe, CSP...) et cette information sera d'autant plus intéressante qu'elle sera reliée fortement à la variable d'étudiée.

Il existe deux façons d'utiliser cette information :

- lors de l'étape de l'échantillonnage en considérant
  - le sondage stratifié
  - proportionnel à la taille
  - équilibré
- lors de l'étape de l'estimation en considérant d'autres estimateurs que l'Horvitz-Thompson :
  - l'estimateur par le ratio
  - l'estimateur poststratifié

1 Population, échantillon : définitions

2 Méthodes pour sélectionner un échantillon
- Sondage de Bernoulli (BE)
- Sondage systématique (SY)
- Sondage aléatoire simple sans remise (SAS)

3 Méthodes d'estimation

4 Améliorer l'estimation
- Sondage stratifié
- L'estimateur par le ratio et poststratifié

5 La non-réponse

6 Présentation des sujets

7 Quelques compléments théoriques

Sondage stratifié

On découpe la population en H sous-populations homogènes appelées "strates" et on sélectionne de façon (aléatoire) **indépendante** un échantillon dans chaque strate.
Grâce à l'indépendance des tirages, on peut utiliser des plans différents à l'intérieur des strates. Le plan stratifié avec SAS à l'intérieur de chaque strate est beaucoup utilisé.

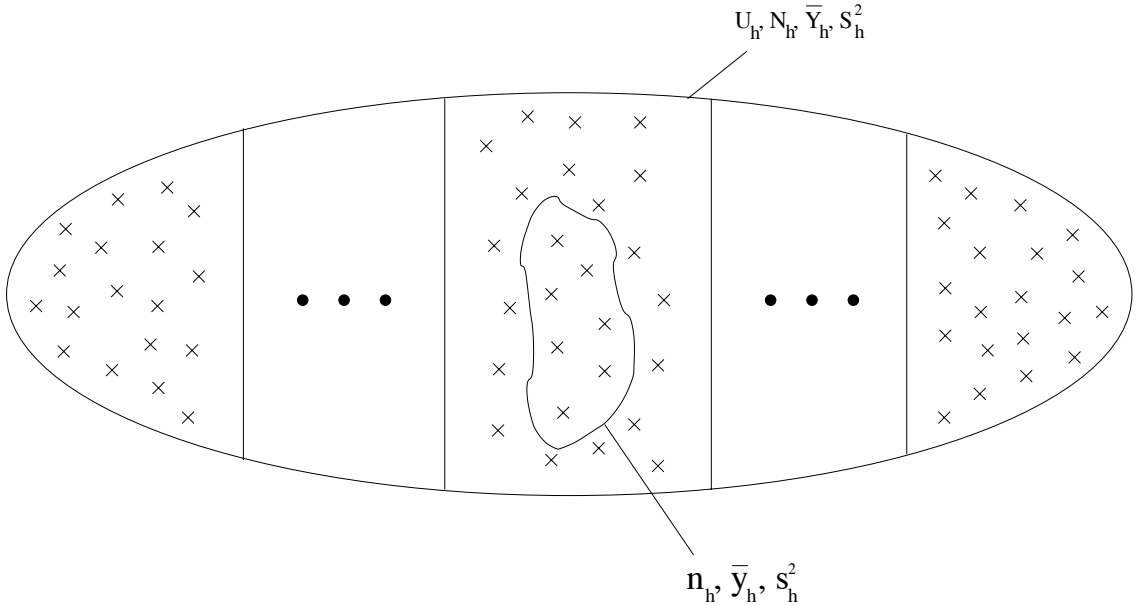


FIGURE : Plan stratifié et SAS dans chaque strate

- Les tailles N_h de strates U_h, $h = 1, \ldots, H$ sont connues et satisfont

$$\sum_{h=1}^{H} N_H = N$$

- Il existe plusieurs méthodes pour choisir la taille n_h de l'échantillon sélectionné à l'intérieur de U_h et tel que :

$$\sum_{h=1}^{H} n_h = n$$

La méthode la plus simple est de choisir n_h de **façon proportionnelle** :

$$\frac{n_h}{N_h} = \frac{n}{N} \quad \text{pour tout} \quad h = 1, \ldots, H.$$

Donc, on a le même taux de sondage dans toutes les strates et égal à celui dans la population totale.

Pourquoi utiliser un sondage stratifié ?

Parmi les plus importantes raisons pour utiliser le plan stratifié, il y a :

- Pour limiter les risques d'obtenir des "mauvais" échantillons comme dans le cas du plan SAS.
- Les strates peuvent correspondre à des souspopulations pour lesquelles des informations sont demandées.
- Dans le cas d'une stratification géographique, l'enquête est plus facile à administrer et son coût plus faible.
- Si les strates sont bien construites (homogènes), le plan stratifié est plus efficace que le plan SAS.

Exemples de populations stratifiés

- Dans une étude sur le nombre moyen de fermes par région, on peut utiliser une stratification géographique.
- Les enquêtes auprès des entreprises françaises utilisent une stratification par activité et taille.
- Les personnes d'âges différents ont une pression du sang différente, alors si on s'intéresse à la pression du sang il faut stratifier par classe d'âge.
- Dans une étude sur la concentration des plantes, il faut stratifier par type de terrain.

Comment estimer le total avec un plan stratifié avec SAS dans chaque strate ?

- Le total t_y est estimé par la somme des H estimateurs pondérés :

$$\begin{aligned}\hat{t} &= \sum_{h=1}^{H} \hat{t}_h = \sum_{\text{unités k dans } s_1} w_{1k} y_{1k} + \ldots + \sum_{\text{unités k dans } s_H} w_{Hk} y_{Hk} \\ &= N_1 \overline{y}_{s_1} + N_2 \overline{y}_{s_2} + \ldots + N_H \overline{y}_{s_H} \quad \text{(pour SAS dans chaque strate)}\end{aligned}$$

- On peut pondérer de façon différente les unités appartenant à des strates différentes.
- Si on choisit les n_h de façon proportionnelle, alors $\frac{n_h}{N_h} = \frac{n}{N}$ pour tous les $h = 1, \ldots, H$, et

$$\hat{t} = N \overline{y}_s$$

donc la même expression que l'estimateur pour un plan SAS. Mais, **très important**, les variances n'ont pas la même expression.

Précision du plan stratifié

- On a grâce à l'indépendance des tirages dans les strates :

$$\text{Var}(\hat{t}) = \sum_{h=1}^{H} \text{Var}(\hat{t}_h)$$

- Si les strates sont homogènes par rapport à la variable y, alors les variances à l'intérieur des strates seront faibles et par conséquent, la variance de $\hat{t}$ avec un plan stratifié sera beaucoup plus faible qu'avec un plan SAS.

Retour à l'exemple du livre

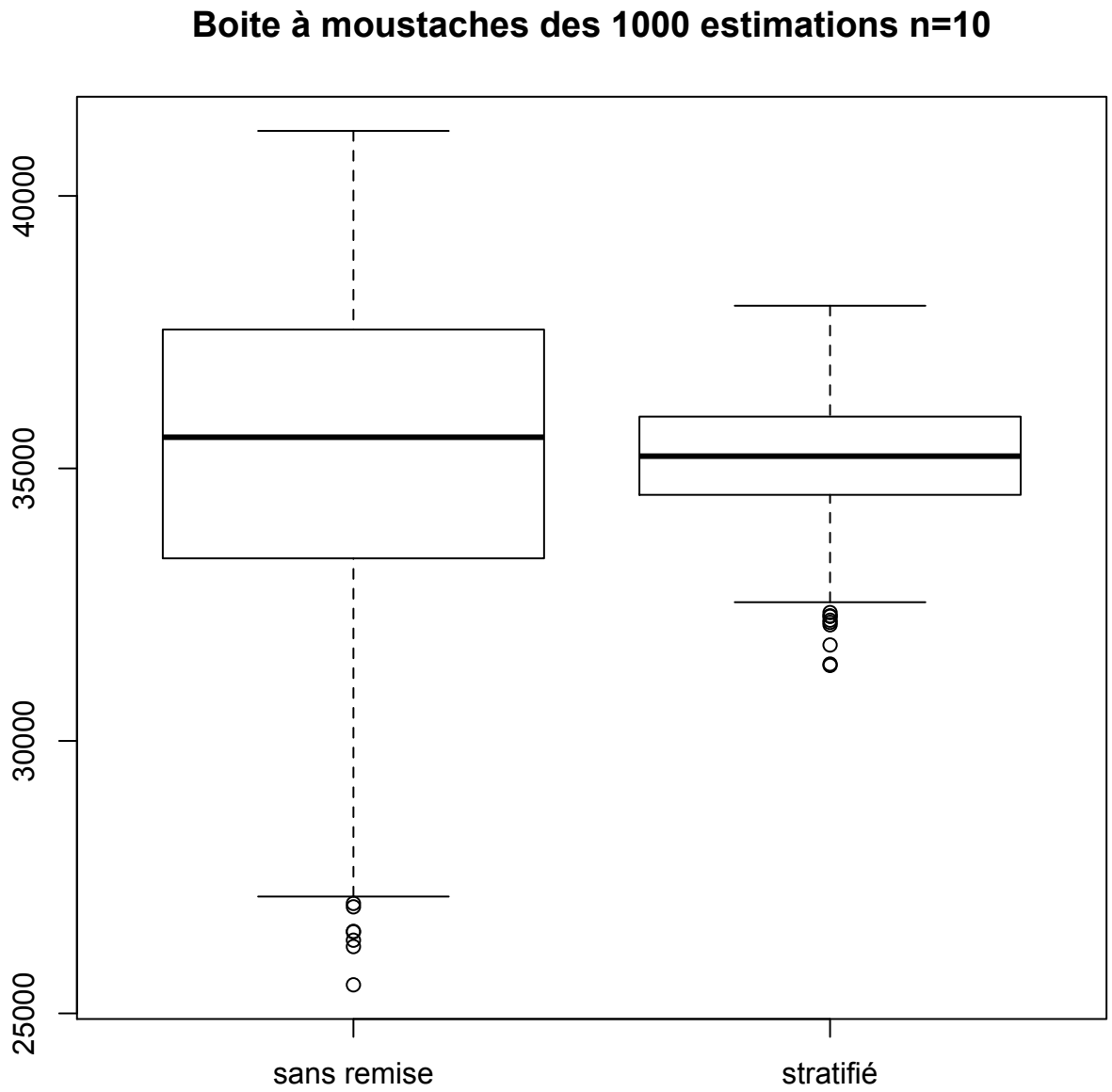
- On peut créer deux strates :
  - **Strate 1** contenant la dernière page de chaque chapitre (de taille $N_1 = 10$) et
  - **Strate 2** contenant toutes les autres pages (de taille $N_2 = 80$).
- L'échantillon sera constitué de 2 pages sélectionnées au hasard dans la strate 1 (donc, parmi 10 pages) et 8 pages sélectionnées au hasard dans la strate 2 (donc, parmi 80 pages).

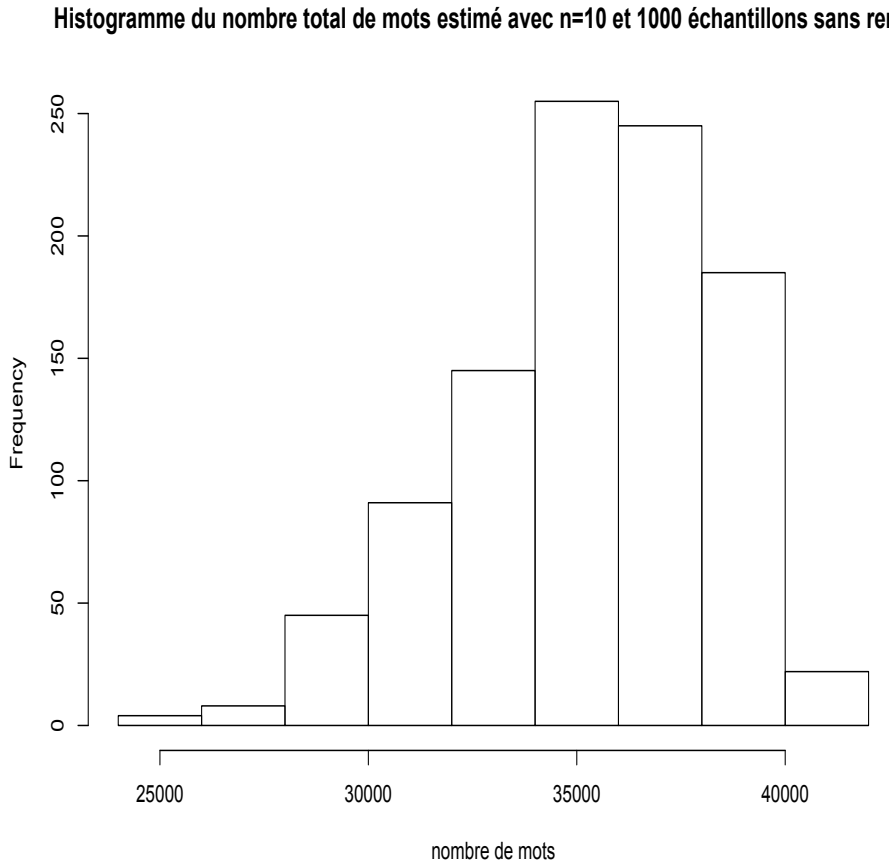
Simulations comme méthode d'expérimentation de statisticiens

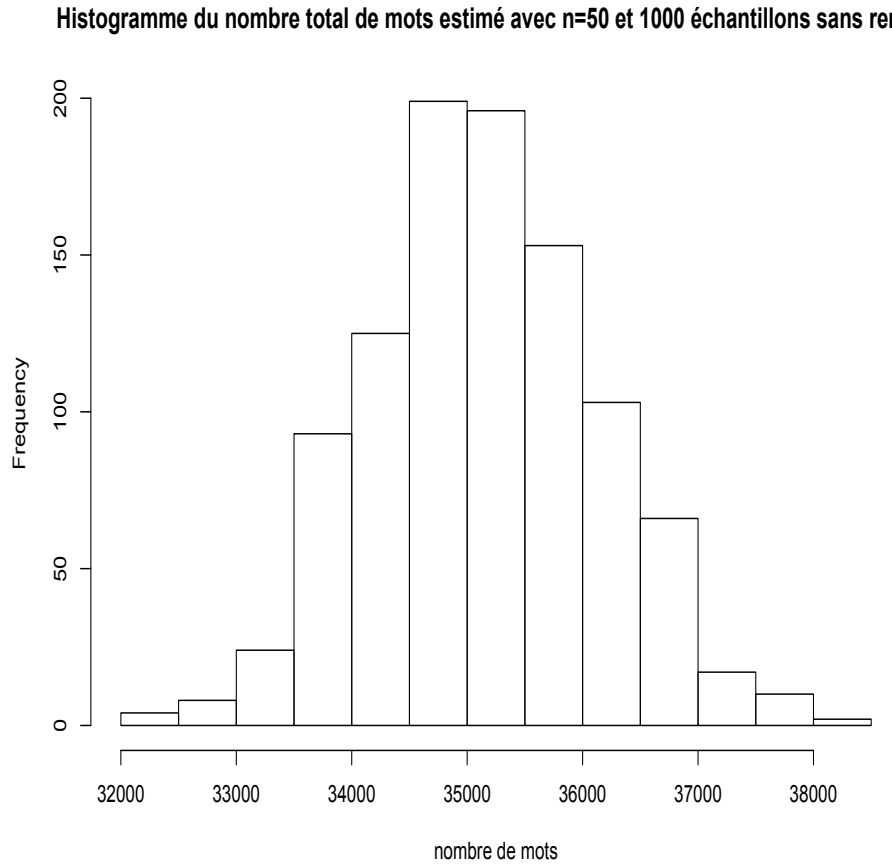
- En statistique, on réalise des **simulations** pour comparer des nouveaux estimateurs, etc et/ou vérifier des propriétés statistiques comme la convergence, le biais ou la normalité asymptotique.
- Le principe des simulations est le suivant :
  - On considère une population pour laquelle on connaît les quantités d'intérêt.
  - On sélectionne à l'aide de l'ordinateur un grand nombre I (=1000, 10000) d'échantillons et pour chaque échantillon sélectionné, on calcule la valeur de l'estimateur (une estimation).
  - On regarde ensuite sur les I réalisations le biais, la normalité...
- **Attention** : en pratique, nous ne pouvons sélectionner qu'un seul échantillon ! Mais les simulations peuvent nous enseigner des inconvénients d'une méthode. On peut essayer ensuite améliorer la méthode avant de la lancer en grandeur nature.

Comparaison des plans SAS et STRAT avec simulations

- On considère l'exemple du livre. Le nombre total des mots est $t_y = 35170$.
- On sélectionne $I = 1000$ échantillons de taille $n = 10$ selon les deux plans : SAS et STRAT.
- Pour chaque échantillon tiré (simulation), on calcule et on garde les estimations $\hat{t}_y$ du nombre total des mots obtenus avec les deux plans ;
- On réalise notre comparaison à partir de ces 1000 estimations.


Boite à moustaches des 1000 estimations n=10
40000
35000
30000
25000
sans remise
stratifié


Histogramme du nombre total de mots estimé avec n=10 et 1000 échantillons sans remise
Frequency
0
50
100
150
200
250
25000
30000
35000
40000
nombre de mots


Histogramme du nombre total de mots estimé avec n=50 et 1000 échantillons sans remise
Frequency
0
50
100
150
200
32000
33000
34000
35000
36000
37000
38000
nombre de mots

1 Population, échantillon : définitions

2 Méthodes pour sélectionner un échantillon
- Sondage de Bernoulli (BE)
- Sondage systématique (SY)
- Sondage aléatoire simple sans remise (SAS)

3 Méthodes d'estimation

4 Améliorer l'estimation
- Sondage stratifié
- L'estimateur par le ratio et poststratifié

5 La non-réponse

6 Présentation des sujets

7 Quelques compléments théoriques

Comment améliorer l'estimation après le tirage de l'échantillon ?

- Le plan stratifié permet d'améliorer l'estimateur d'Horvitz-Thompson. Il existe aussi le plan proportionnel à la taille.
- Ces plans de sondages demandent que l'information auxiliaire soit connue **avant l'enquête** pour chaque individu de la population. Parfois, on n'a que les valeurs agrégées de cette information : le nombre total de filles ou garçons.
  Si on pense que notre information est liée à notre variable d'intérêt, alors comment peut-on l'utiliser ?
  Il s'agit de construire **à postériori** (après l'enquête), un estimateur plus efficace que l'estimateur d'Horvitz-Thompson.
- L'estimateur **par le ratio** est l'exemple le plus simple d'un tel estimateur. Il a été proposé pour la première fois par Laplace (1802).
- Il est possible également de combiner les deux procédés, par exemple le plan stratifié et l'estimateur par le ratio.

Exemple

- Considérons l'exemple d'une enquête agricole américaine de 1992 : on veut estimer la superficie totale d'une région dédiée aux fermes ;
- On ne dispose pas de l'information nécessaire pour stratifier la population de fermes en fonction de leurs tailles, alors il est décidé de faire une enquête par SAS ;
- On dispose par ailleurs de la superficie totale en 1987. La superficie d'une ferme en 1992 est très liée à celle de 1987, donc on aimerait utiliser cette information, mais comment ?
- **Réponse** : on utilise l'estimateur par le ratio.

L'estimateur par le ratio : exemple de Laplace

Objectif : estimer la population de la France vers 1800.

- Laplace sélectionne un échantillon de taille 30 de communes en France et il obtient une estimation de $\hat{t}_{habitants,30com} = 2037615$ habitants ;
- il peut disposer du nombre de naissances dans ces 30 communes et il estime le nombre total de naissances par $\hat{t}_{naissances,30com} = 71866$. Donc, il y a

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

- il connaît également **le nombre total de naissances** en 1802 et il fait le raisonnement que les communes avec beaucoup d'habitants vont avoir plus de naissances, donc il propose d'estimer le nombre total d'habitants en France par

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

Dans l'exemple de Laplace :

- le nombre de naissances est **l'information auxiliaire** qui est très liée à la variable nombre d'habitants ;
- on n'a besoin que du nombre total de naissances, t_x, et du nombre de naissances dans chaque ville sélectionnée : x_k pour $k \in s$.
- L'estimateur proposé par Laplace est beaucoup utilisé en pratique pour estimer des totaux ;
- Il s'appelle **l'estimateur par le ratio** :

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

où $\hat{t}_y = \sum_{k \in s} \frac{y_k}{\pi_k}$ et $\hat{t}_x = \sum_{k \in s} \frac{x_k}{\pi_k}$;

- La variable peut être aussi qualitative (ex. $x_k = 1$ si l'individu k est une fille et zéro sinon).

Propriétés

- L'estimateur par le ratio est une somme pondérée des valeurs de y sur l'échantillon :

$$\hat{t}_{y,ratio} = \sum_{k \in s} \underbrace{\frac{t_x}{\hat{t}_x} \frac{1}{\pi_k}}_{w_k} y_k = \sum_{k \in s} w_k y_k$$

- Ces poids vérifient la propriété suivante :

$$\sum_{k \in s} w_k x_k = \sum_{k \in U} x_k$$

autrement dit, on estime parfaitement le total de la variable auxiliaire quel que soit l'échantillon s :

$$\hat{t}_{x,ratio} = t_x.$$

- En général, cet estimateur est biaisé mais pour des échantillons de grandes tailles, ce biais est très proche de zéro.

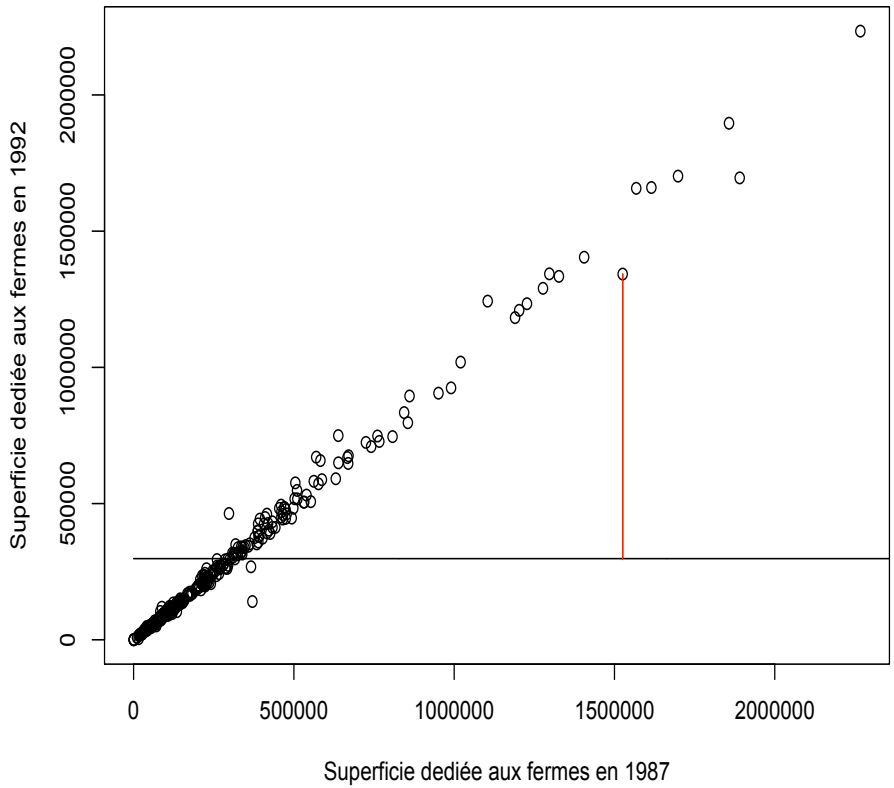
Pourquoi l'estimateur par le ratio est plus efficace que l'estimateur d'Horvitz-Thompson ?

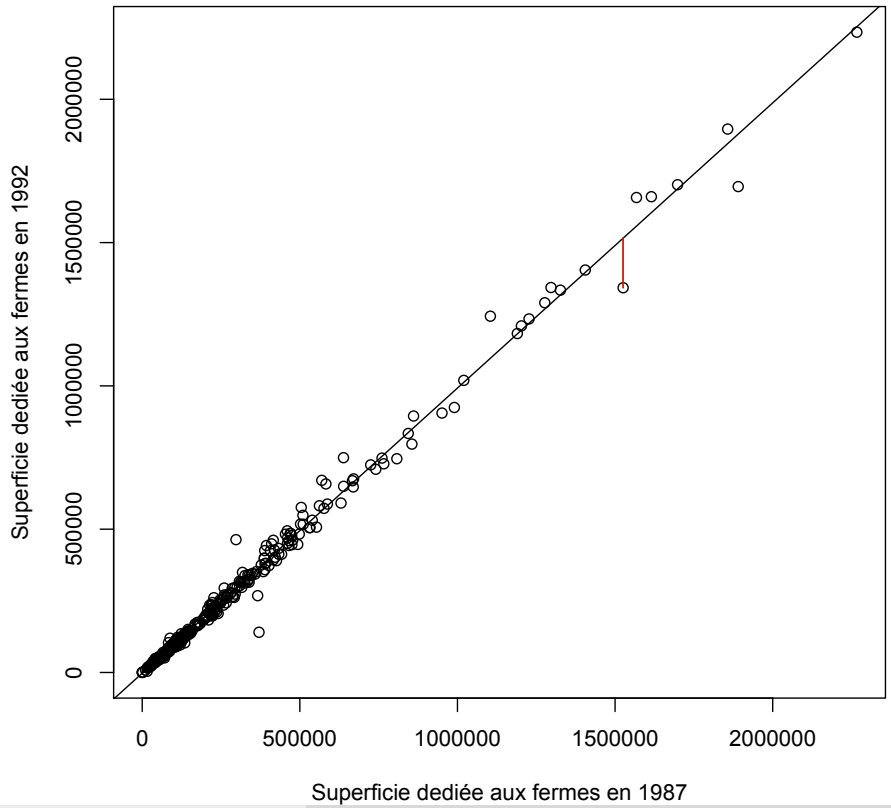
- Si l'information auxiliaire x est presque proportionnelle à la variable d'intérêt alors, pour des échantillons de grandes tailles, $\hat{t}_{y,ratio}$ est plus précis que l'estimateur d'HT qui n'utilise pas d'information auxiliaire (voir aussi l'exemple de la fin du cours).

- Considérons un plan SAS et comparons les estimateurs de variances de $\hat{t}_y$ et $\hat{t}_{y,ratio}$:
  $\hat{Var}(\hat{t}_y) =$ la somme des carrés des écarts de y_k à la moyenne $\bar{y}_s$
  $\hat{Var}(\hat{t}_{y,ratio}) =$ la somme des carrés des écarts de y_k à la droite $y = \hat{R}x$

Retour à l'exemple de fermes américaines

Considérons un extrait d'une enquête américaine et nous souhaitons estimer la superficie totale dédiée aux fermes en 1992. La superficie totale en 1987 est disponible. Ces deux variables sont quasiment proportionnelles.





L'estimateur poststratifié : exemple (1)

Considérons l'exemple suivant :
On veut estimer la somme moyenne dédiée à la nourriture/mois dans les ménages américains. 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.
La consommation est fortement liée à la taille du ménage, donc on aimerait bien utiliser cette information, mais comment ?
On peut faire une **poststratification**. Cela revient à considérer la somme des estimateurs par le ratio à l'intérieur de chaque groupe (type de ménage).

Exemple (2)

Considérons l'exemple suivant :

- on veut estimer le nombre total d'étudiants qui veulent être professeurs après leurs études dans une population de $N = 4000$ étudiants ; donc, la variable d'intérêt est

$$y_k = \begin{cases} 1 \text{ si l'individu } k \text{ veut être prof.} \\ 0 \text{ sinon} \end{cases}$$

- on sait par ailleurs que dans la population totale on a $N_F = 2700$ femmes et $N_H = 1300$ hommes et on pense que la population de femmes (ou d'hommes) est relativement homogène vis-à-vis de la volonté d'être professeur ;
- de nouveau, il n'est pas possible de stratifier (avant l'enquête) la population en deux strates : hommes et femmes. Donc, on prend un échantillon aléatoire simple sans remise de taille $n = 400$;

- dans notre échantillon, il y a $n_F = 240$ femmes dont 84 veulent être des professeurs et $n_H = 160$ hommes dont 40 veulent être des professeurs.
- On utilise cette information et on construit l'estimateur poststratifié :

$$\begin{aligned} \hat{t}_{post} &= N_F \frac{\sum_{s_F} y_k}{n_F} + N_G \frac{\sum_{s_G} y_k}{n_G} \\ &= 2700 \times \frac{84}{240} + 1300 \times \frac{40}{160} = 1270 \end{aligned}$$

Il s'agît d'une somme de deux estimateurs par le ratio par la taille de chaque groupe (règle de trois à l'intérieur de chaque groupe) !

Avantages de la poststratification

- Si les groupes sont homogènes et on a suffisamment d'individus dans chaque groupe, on peut montrer que la poststratification donne une précision similaire à celle obtenue avec un plan stratifié et $n_h = \frac{nN_h}{N}$. Donc, on améliore par rapport à l'estimateur d'HT avec un plan SAS.
- Supposons qu'on veut estimer $P_F = \frac{N_F}{N}$ la proportion de femmes :
  avec HT : $\hat{P}_F = p_F = \frac{n_F}{n} \neq P_F$
  avec post : $\hat{P}_F = \frac{N_F}{Nn_F} \sum_{k \in s_F} \mathbf{1}_{k \in U_F} = P_F$
- La poststratification est utilisée dans le traitement de la non-réponse (voir chapitre suivant).

1 Population, échantillon : définitions

2 Méthodes pour sélectionner un échantillon
- Sondage de Bernoulli (BE)
- Sondage systématique (SY)
- Sondage aléatoire simple sans remise (SAS)

3 Méthodes d'estimation

4 Améliorer l'estimation
- Sondage stratifié
- L'estimateur par le ratio et poststratifié

5 La non-réponse

6 Présentation des sujets

7 Quelques compléments théoriques

Traitement de la non-réponse (NR)

- On a de la **non-réponse** quand le questionnaire ne revient pas (non-réponse totale) ou que l'individu ne répond pas à toutes les questions (non-réponse partielle).
- Ignorer les non-répondants peut conduire à des estimations totalement erronées.
- De nos jours, les taux de non-réponse sont de plus en plus élevées, surtout pour la population de jeunes (16-25 ans).

Traitement de la non-réponse

- **Prévenir** la NR en faisant attention au contenu et à la longueur du questionnaire (les questions "sensibles" augmentent la non-réponse), au moment de l'enquête (éviter les vacances), à choisir une méthode adaptée pour l'enquête (téléphone, face-à-face, papier) ;
- **Réduire** la NR en insistant pour obtenir une réponse ; "follow-ups" et "callbacks".
- Si malgré tous nos efforts, on a la NR, on peut **ré-échantillonner les non-répondants** : on sélectionne un échantillon parmi les non-répondants et on essaie d'obtenir l'information complète pour ce deuxième échantillon ;
- Enfin, **utiliser une méthode adéquate** pour traiter les non-répondants restants.

Difficulté majeure : la probabilité de réponse est inconnue !

ind.	age	sex	years of education	"crime victime"	"violent crime victime"
1	47	H	16	0	0
2	45	F	?	1	1
3	19	H	11	0	0
4	21	F	?	1	1
5	24	H	12	1	1
6	41	F	?	0	0
7	36	H	20	1	?
8	50	H	12	0	0
9	53	F	13	0	?
10	17	H	10	?	?
11	53	F	12	0	0
12	21	F	12	0	0
13	18	F	11	1	?
14	34	H	16	1	0
15	44	H	14	0	0
16	45	H	11	0	0
17	54	F	14	0	0
18	55	F	10	0	0
19	29	F	12	?	0
20	32	F	10	0	0
21	?	?	?	?	?

La répondération comme méthode de traitement de la non-réponse totale

- **Situation** : il manque le questionnaire de l'individu k.
- Soit r=l'ensemble des individus de l'échantillon s qui ont répondu (répondants).
- **Répondération** : on utilise des nouveaux poids

$$w_k = \frac{1}{\phi_k \pi_k}, \quad k \in r$$

où ϕ_k est la probabilité que l'individu k réponde et π_k=la probabilité que l'individu s soit sélectionné.

- De cette façon, les poids des répondants sont augmentés pour qu'ils représentent aussi les non-répondants (en plus de personnes non-échantillonnées).

- Le total $t_y = \sum_{k \in U} y_k$ sera estimé par

$$\hat{t}_r = \sum_{k \in r} w_k y_k = \sum_{k \in r} \frac{y_k}{\phi_k \pi_k}$$

- **Difficulté** : ϕ_k est inconnue et elle doit être estimée ou prédite à l'aide d'un modèle ;
- On classe les répondants en **C classes de répondération** et on suppose qu'à l'intérieur de chaque classe, les individus répondent de façon indépendante et avec la même probabilité. Etant donné l'échantillon s, l'échantillon de répondants à l'intérieur d'une classe peut être vu comme un plan BE de paramètre ϕ_c.
- Les classes de répondération sont constituées à l'aide de l'information auxiliaire (sex, age, ...).

Estimation de la probabilité de réponse

On suppose que l'échantillon s est sélectionné selon un plan SAS dans la population U;

À l'intérieur de la classe c, on estime la probabilité de réponse ϕ_c par :

$$\hat{\phi}_c = \frac{\text{nombre de répondants dans la classe c}}{\text{nombre d'individus dans la classe c}} = \frac{n_{rc}}{n_c}$$

Cas 1 : l'information auxiliaire est connue que sur l'échantillon s

On suppose que par exemple, on peut connaître le sexe (information auxiliaire) pour tous les individus de l'échantillon (répondant ou pas) mais on ne connaît pas les effectifs dans la population.
Dans ce cas, on estime le total t_y par

$$\hat{t}_r = \frac{N}{n} \sum_{c=1}^{C} \frac{n_c}{n_{rc}} \sum_{k \in s_{rc}} y_k = \frac{N}{n} \sum_{c=1}^{C} n_c \overline{y}_{r_c}$$

Si on a que deux classes : Homme/Femme, alors l'estimateur devient

$$\hat{t}_r = \frac{N}{n} \left(\frac{n_F}{n_{rF}} \sum_{k \in s_{rF}} y_k + \frac{n_H}{n_{rH}} \sum_{k \in s_{rH}} y_k \right)$$

Cas 2 : l'information auxiliaire est connue que sur la population U

L'information auxiliaire est connue au niveau de la population mais pas au niveau de l'échantillon ;
Par exemple, on connaît les effectifs d'hommes (N_H) et de femmes (N_F) sur la population, mais pas sur l'échantillon.
Alors, on peut utiliser l'estimateur poststratifié pour estimer le total t_y :

$$\hat{t}_r = \sum_{c=1}^{C} \frac{N_c}{n_{rc}} \sum_{k \in s_{rc}} y_k = \sum_{c=1}^{C} N_c \overline{y}_{r_c}$$

Si on a deux classes, Homme/Femme, alors

$$\hat{t}_{r,post} = \frac{N_F}{n_{rF}} \sum_{k \in s_{rF}} y_k + \frac{N_H}{n_{rH}} \sum_{k \in s_{rH}} y_k$$

L'imputation pour traiter la non-réponse partielle

Situation : l'individu envoie un questionnaire rempli **partiellement** ;
Pour remplacer l'information manquante, on utilise **l'imputation**. On obtient de cette façon, un fichier de données complet ;
En général, les individus sont partagés en classes (construites en utilisant l'information auxiliaire) et on applique, classe par classe, une méthode d'imputation pour remplacer les valeurs manquantes. Voici deux méthodes d'imputation parmi les plus utilisées :

- imputation par la moyenne de la classe : une valeur manquante est remplacée par la moyenne des valeurs de répondants de la même classe ;
- l'imputation hot-deck aléatoire : la valeur manquante est choisie au hasard à l'intérieur de la classe ("un donneur") ;

Exemple

Considérons de nouveau l'exemple suivant :

ind.	age	sex	years of educ.	"c.v."	"v. c. v."
1	47	H	16	0	0
2	45	F	?	1	1
3	19	H	11	0	0
4	21	F	?	1	1
5	24	H	12	1	1
6	41	F	?	0	0
7	36	H	20	1	?
8	50	H	12	0	0
9	53	F	13	0	?
10	17	H	10	?	?
11	53	F	12	0	0
12	21	F	12	0	0
13	18	F	11	1	?
14	34	H	16	1	0
15	44	H	14	0	0
16	45	H	11	0	0
17	54	F	14	0	0
18	55	F	10	0	0
19	29	F	12	?	0
20	32	F	10	0	0

On forme 4 classes en considérant les variables sexe and age :

1. classe 1 : H, $\leq$ 34 ans ;
2. classe 2 : H $\geq$ 35 ans ;
3. classe 3 : F $\leq$ 34 ans ;
4. classe 4 : F $\geq$ 35 ans

Exemple : imputation hot-deck

A l'intérieur de chaque classe, on choisi au hasard un donneur parmi ceux qui ont de réponses complètes. La valeur du donneur est utilisée pour toutes les valeurs manquantes.
Dans la classe 1, on a le donneur 14 ; dans la classe 2 le donneur 16 ; classe 3, le donneur 12 ; classe 4, le donneur 17. La valeur en rouge est imputée par hot-deck.

ind.	age	sex	years of educ.	"c.v."	"v. c. v."	imputation class
3	19	H	11	0	0	1
5	24	H	12	1	1	1
10	17	H	10	1	0	1
14	34	H	16	1	0	1
15	44	H	14	0	0	2
16	45	H	11	0	0	2
1	47	H	16	0	0	2
7	36	H	20	1	0	2
8	50	H	12	0	0	2
4	21	F	12	1	1	3
12	21	F	12	0	0	3
13	18	F	11	1	0	3
19	29	F	12	0	0	3
20	32	F	10	0	0	3
2	45	F	14	1	1	4
6	41	F	14	0	0	4
9	53	F	13	0	0	4
11	53	F	12	0	0	4
17	54	F	14	0	0	4
18	55	F	10	0	0	4

1 Population, échantillon : définitions

2 Méthodes pour sélectionner un échantillon
- Sondage de Bernoulli (BE)
- Sondage systématique (SY)
- Sondage aléatoire simple sans remise (SAS)

3 Méthodes d'estimation

4 Améliorer l'estimation
- Sondage stratifié
- L'estimateur par le ratio et poststratifié

5 La non-réponse

6 Présentation des sujets

7 Quelques compléments théoriques

Sujet 1 : Estimation du niveau d'un lycée en mathématique

Les élèves de Première passent en janvier 2014 une évaluation en mathématiques. Les résultats sont communiqués de façon individuelle en mars, et le lycée reçoit les notes de l'ensemble des élèves début juillet.

En avril, le proviseur souhaite obtenir une estimation de la réussite de son établissement en se basant sur un échantillon de 50 élèves. Il dispose également pour chaque élève de sa note moyenne en mathématiques obtenue lors des épreuves internes du lycée pour le dernier trimestre de l'année 2013-2014.

L'objectif du projet est de proposer des stratégies au proviseur, et de mesurer la pertinence de ces stratégies a posteriori quand tous les résultats sont communiqués au lycée début juillet.

Sujet 2 : Estimation de l'impôt sur le revenu par foyer

A l'été 2010, le conseil régional de Bourgogne souhaite estimer l'impôt sur le revenu moyen payé par foyer fiscal en 2010, et correspondant aux revenus de 2009.

Il a pour cela la possibilité de contacter le centre des impôts de 100 communes, pour obtenir pour chacune d'entre elles l'impôt global acquitté et le nombre total de foyers qui ont déposé une déclaration. Il dispose également pour chaque commune, au travers du site internet de l'Insee, des données communales pour les déclarations de revenu des années 2008 et antérieures, et des données de recensement.

L'objectif du projet est de proposer des stratégies au Conseil Régional, et de mesurer la pertinence de ces stratégies a posteriori quand les statistiques communales exhaustives sont disponibles début 2011.

Sujet 3 : Enquête sur l'usage d'alcool, de tabac ou de drogues

Une enquête est mise en place pour mesurer la proportion et les caractéristiques des lycéens consommant ou ayant consommé de l'alcool, du tabac ou de la drogue. Un échantillon de lycées est pour cela tiré au sort, dont votre propre lycée.

Vous êtes chargés de rendre compte des résultats pour votre établissement. Après saisie des questionnaires, les données sont mises à votre disposition sous la forme d'un tableau où chaque ligne correspond à un questionnaire, et chaque colonne à la réponse à une question.

Vous devez calculer les moyennes ou proportions des réponses aux différentes questions posées sur le thème principal de l'enquête, en proposant des solutions argumentées aux différents problèmes pouvant se poser et notamment les problèmes de réponses manquantes.

Sujet 4 : Algorithmes de sélection d'un échantillon

L'étude statistique d'une population se base généralement sur l'étude d'un échantillon. Le tirage avec remise est une méthode possible de sélection, mais en pratique on se base plutôt sur des méthodes de tirage sans remise.

Parmi les méthodes de sélection les plus utilisées, deux méthodes sont considérées ici : l'échantillonnage systématique, et l'échantillonnage simple sans remise.

L'objectif de ce projet est d'étudier les propriétés de l'un et/ou l'autre de ces deux algorithmes de sélection, et de les comparer à la méthode d'échantillonnage avec remise. Il est également demandé d'implémenter l'un des deux algorithmes, pour une population de taille N quelconque.

Sujet 5 : Réaliser une enquête de votre choix

Dans ce sujet, nous vous laissons libres de concevoir une enquête sur le sujet de votre choix, qu'il vous faudra mener auprès d'un échantillon.

Il vous faudra réfléchir en particulier aux différents aspects suivants, et préciser dans le rapport (en les justifiant) les choix que vous aurez réalisés :

- définition du sujet de l'enquête et de la population visée,
- recherche d'études ou d'informations sur le sujet,
- choix du mode de recueil des données, éventuellement rédaction du questionnaire,
- sélection de l'échantillon,
- réalisation de l'enquête,
- analyse des résultats,
- appréciation de la qualité des résultats.

1 Population, échantillon : définitions

2 Méthodes pour sélectionner un échantillon
- Sondage de Bernoulli (BE)
- Sondage systématique (SY)
- Sondage aléatoire simple sans remise (SAS)

3 Méthodes d'estimation

4 Améliorer l'estimation
- Sondage stratifié
- L'estimateur par le ratio et poststratifié

5 La non-réponse

6 Présentation des sujets

7 Quelques compléments théoriques

Propriétés des variables I_k

1. $Var(I_k) = E(I_k^2) - E^2(I_k) = \pi_k - \pi_k^2$
2. Soient deux individus distincts $k, l \in U$:

$$\begin{aligned} Cov(I_k, I_\ell) &= E(I_k I_\ell) - E(I_k)E(I_\ell) \\ &= \pi_{kl} - \pi_k \pi_l \end{aligned}$$

où π_{kl} est la probabilité que le couple (k, l) soit sélectionné dans un échantillon ;

Compléments sur le plan de BE

- La taille aléatoire de l'échantillon n_s peut s'écrire :

$$n_s = \sum_{k \in U} I_k$$

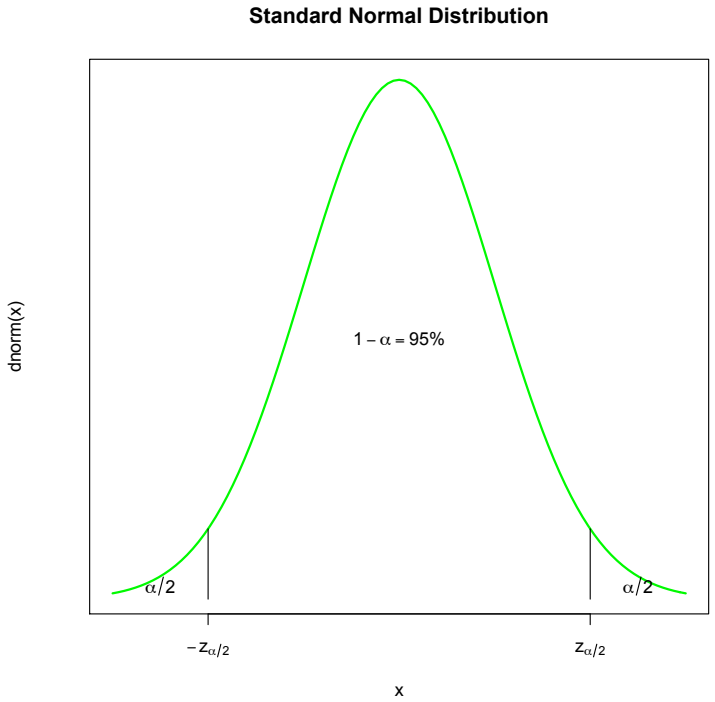
- n_s suit une loi $\mathcal{B}(N, \pi)$ binomiale de paramètre N et π (1/2 si c'est une pièce parfaite).
- On peut choisir π tel que $E(n_s) = n$ où n est la taille fixe souhaitée. On obtient alors

$$N\pi = n \rightarrow \pi = \frac{n}{N}.$$

Intervalle de confiance :1

En plus d'une estimation ponctuelle, on souhaite souvent une estimation par **intervalle de confiance**.

Hypothèse : on suppose que la taille de l'échantillon est suffisamment grande pour qu'on approche la loi de $\frac{\hat{t}_y - t_y}{\sqrt{\text{Var}(\hat{t}_y)}}$ par une loi normale de moyenne nulle et d'écart-type égal à 1



Intervalles de confiance : 2

Sous l'hypothèse de normalité, l'intervalle de confiance $(1-\alpha)\%$ pour $\hat{t}_y$ est donné par

$$\hat{t}_y \in \left[\hat{t}_y - z_{\alpha/2}\sqrt{\hat{\mathrm{V}}(\hat{t}_y)}\,;\, \hat{t}_y + z_{\alpha/2}\sqrt{\hat{\mathrm{V}}(\hat{t}_y)}\right]$$

Si $(1-\alpha)\% = 95\%$, alors $z_{\alpha/2} = 1.96$;
Précision absolue : la demi-longueur de l'intervalle de confiance : $z_{\alpha/2}\sqrt{\hat{\mathrm{V}}(\hat{t}_y)}$.

Précision relative : $\dfrac{z_{\alpha/2}\sqrt{\hat{\mathrm{V}}(\hat{t}_y)}}{\hat{t}_y}$

Comment calculer la taille de l'échantillon

La théorie présentée auparavant suppose que la taille n est connue. En pratique, comment doit-on choisir la taille n ?
Souvent, nous sommes confrontés en pratique aux situations suivantes :

- mesures de plusieurs variables et objectifs ; les objectifs les plus importants doivent être spécifiés.
- une certaine précision est souhaitée et le budget de l'enquête est fixé : un équilibre entre ces deux contraintes doit être trouvé ;

Calcul de la taille de l'échantillon

• La précision souhaitée est souvent exprimée en termes absolus comme suit :

$$P(|\hat{\theta} - \theta| \leq e) = 1 - \alpha$$

où θ peut être une moyenne, un total, une proportion et e=*la marge d'erreur tolérée*.

• Le praticien doit décider quelles valeurs il faut prendre pour α (exemple : $\alpha = 0.05$ and $e = 0.03$).

• On veut n tel que la demi-longueur de l'intervalle de confiance soit au plus égale à e,

$$e \geq z_{\alpha/2}\sqrt{V(\hat{\theta})} \quad (z_{\alpha/2} = 1.96 \quad \text{pour} \quad (1-\alpha)\% = 95\%)$$

Taille d'échantillon pour estimer un total t_y avec SAS

- Soit *e la marge d'erreur tolérée*.
- La variance sous le plan SAS est $V(\hat{t}_\pi) = N^2 \frac{1-f}{n} S_{yU}^2$;
- Il résulte

$$n \geq \frac{N^2 z_{\alpha/2}^2 S_{yU}^2}{e^2 + N z_{\alpha/2}^2 S_{yU}^2}$$

- Si le taux de sondage $f \simeq 0$, alors $n \geq \frac{N^2 z_{\alpha/2}^2 S_{yU}^2}{e^2}$.
- **Difficulté** : on ne connaît pas S_{yU}^2. On l'estime à partir d'une enquête **pilot**.

Taille d'échantillon pour estimer une proportion P avec SAS

- Soit e *la marge d'erreur tolérée*.
- La variance sous le plan SAS est

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

1. Si N est grand et $f \simeq 0$, alors $n \geq \frac{z_{\alpha/2}^2 P(1-P)}{e^2} = n_0$;
2. Si N est grand et f assez grand, alors

$$n \geq \frac{z_{\alpha/2}^2 P(1-P)}{e^2 + \frac{z_{\alpha/2}^2 P(1-P)}{N}} = \frac{n_0}{1 + \frac{n_0}{N}}$$

Difficulté : on ne connaît pas P. Nous avons deux possibilités :

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

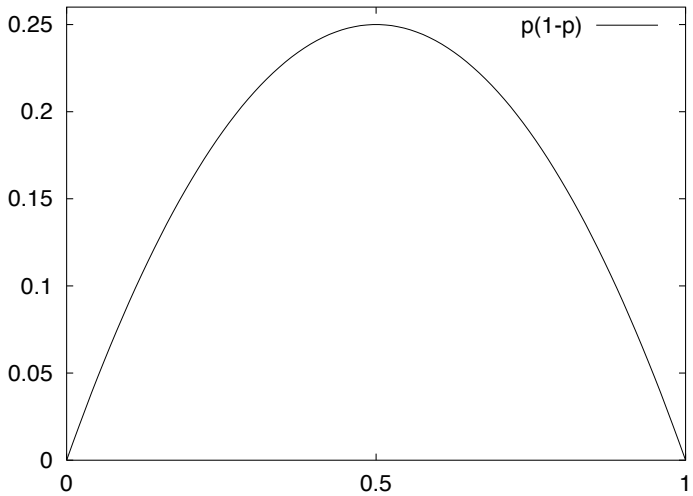


FIGURE : $P \mapsto P(1-P)$

2. on considère une estimation de P issue lors d'une enquête pilot.

Compléments sur l'estimateur par le ratio (1)

- Supposons que nous voulons estimer le nombre de poissons pêchés dans un lac.
- Il y a $N = 4$ endroits pour pêcher autour du lac et on sélectionne selon un plan aléatoire simple sans remise $n = 2$ endroits pour pécher.
- On connaît par ailleurs 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	$t_y = 1400$

- On utilise deux estimateurs : l'estimateur d'Horvitz-Thompson et l'estimateur par le ratio.

Compléments sur l'estimateur par le ratio (2)

L'estimateur par le ratio est $\hat{t}_{yrat} = t_x \times \frac{\sum_s y_i}{\sum_s x_i}$ où s est l'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

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

Compléments sur l'estimateur par le ratio (3)

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.

Compléments sur l'estimateur par le ratio (3) : comparaison avec l'estimateur d'HT

On considère un échantillon SAS de taille n et deux estimateurs :

- sans inf. aux. : l'estimateur d'Horvitz-Thompson (ou par les valeurs dilatées) :

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

avec une variance estimée égale à

$$N^2 \frac{1-f}{n} S^2_{ys} = N^2 \frac{1-f}{n} \frac{1}{n-1} \sum_{k \in s} (y_k - \bar{y}_s)^2;$$

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

$$\hat{t}_{yrat} = t_x \cdot \frac{\hat{t}_y}{\hat{t}_x} = t_x \cdot \frac{\sum_{k \in s} y_k}{\sum_s x_k}$$

avec une variance estimée égale à

$$N^2 \frac{1-f}{n} S^2_{y-\hat{R}x,s} = N^2 \frac{1-f}{n} \frac{1}{n-1} \sum_{k \in s} (y_k - \hat{R}x_k)^2$$