Théorie des sondages : cours 1

Camelia Goga

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

Master Besançon

Plan du cours et bibliographie

Plan du cours

- **Chapitre 1** : Généralités
- **Chapitre 2** : Plans simples
- **Chapitre 3** : Sondage stratifié
- **Chapitre 4** : Sondage à deux degrés et en grappes
- **Chapitre 5** : Techniques de linéarisation et de ré-échantillonnage
- **Chapitre 6** : Techniques de redressement

Bibliographie :

- Pascal Ardilly : Les techniques de sondages.
- Yves Tillé : Théorie des sondages.

Problèmes fondamentaux des sondages

Le sondage : bien plus qu'un sondage d'opinion ;

Exemples des domaines qui utilisent les techniques de sondages :

1. la détermination du volume de certaines productions agricoles ;
2. des calculs de grands indices médiatiques : l'indice des prix à la consommation où l'indice du coût à la construction ;
3. en sport : les contrôles antidopage ;
4. le nombre de chômeurs ;
5. ...

Plan d'un sondage

- Population U de taille finie N, connue ou inconnue ;

$$U = \{u_1, \ldots, u_k, \ldots, u_N\} = \{1, \ldots, k \ldots, N\}.$$

Un élément $u_k \in U$ s'appelle individu.
Très important : L'individu $u_k \in U$ est repéré précisément et sans aucune ambiguité : identifiant k.

Exemples : les fermes agricoles (1), les sportifs participants à un concours (3), la population d'un pays avec quelques exceptions (enfants, fonctionnaires) (4)

- Variable d'intérêt : Y qui prend la valeur y_k pour l'individu k ;
  1. quantitative
  2. qualitative

L'objectif d'un sondage : obtenir l'information sur un paramètre Φ qui est une fonction de y_k, $\Phi = \Phi(y_1, \ldots, y_N)$; on ne s'intéresse pas aux valeurs de Y (statistique inférentielle).
Le paramètre Φ est inconnu.

- Si Y est quantitative, alors Φ peut être
  1. $\Phi = \sum_{k \in U} y_k$ le total de Y dans la population U ;
  2. $\Phi = \sum_{k \in U} y_k / N$ la moyenne de Y ;
  3. le quantile (médiane) de Y ;
  4. la variance et l'écart-type ;

Exemples : le revenu total ou moyen, nombre de chômeurs ...

- Si Y est quanlitative, alors Φ peut être essentiellement **des pourcentages** d'individus de la population dont la variable prend telle ou telle modalité.
  Exemple : la proportion d'individus qui ont voté pour monsieur A.

• Échantillon s dans U : une partie d'individus de la population qui sera interrogée $\Longrightarrow$ **une enquête par sondage**. On peut obtenir s selon deux procédés :

1. probabiliste : les individus sont sélectionnés selon un procédé probabiliste $p(s)$ ou chaque individu a une probabilité **donnée connue d'avance** π_k d'appartenir à l'échantillon ;
2. non-probabiliste ou empirique : les sondages par quotas ; coût moins élevé, en France beaucoup utilisés ;

Recensement : on observe tous les éléments de U.

• Chaque individu k de l'échantillon s est interrogé et on note y_k. On obtient

$$\{(k, y_k), \quad k \in s\}.$$

Plusieurs modalités : interview directe, par téléphone, par la poste...

• Les valeurs y_k, avec $k \in s$ sont utilisées pour construire un **estimateur** $\widehat{\Phi}(y_k, k \in s)$ de $\Phi(y_k, k \in U)$.

On veut inférer les résultats de l'échantillon s à la population U.

On regarde la **précision** de $\widehat{\Phi}$;

faire presque aussi bien qu'un recensement mais avec un coût beaucoup plus faible.

Un peu plus de vocabulaire

Définition

Une population cible est une population pour laquelle l'information est requise.

Définition

L'unité d'observation est l'unité sur laquelle on collecte effectivement l'information.

Définition

Les unités d'échantillonnage sont des entités disjointes dont l'union est égale à la population.

Exemple : en Chine, le chef responsable d'un village est le seul autorisé à répondre aux enquêtes et représente l'ensemble de ses administrés : ces derniers constituent l'unité d'observation alors que le chef est l'unité d'échantillonnage. (on supposera que l'unité d'observation et l'unité d'échantillonnage coïncident.)

Définition

La base de sondage donne les moyennes d'identifier les unités d'échantillonnage et de communiquer avec elles.

Base de sondage

Une base parfaite :

1. possibilité de repérer les unités sans ambiguïté : l'identifiant $\Longrightarrow$ liste d'identifiants de bonne qualité ;
   un logement : par la commune, le district, l'immeuble et un rang numérique qu'on lui donne dans l'immeuble.
   un individu : la commune, le no et le nom de la rue, son nom et prénom ;
2. exhaustive ; sinon, on a un défaut de couverture ;
3. sans double compte.
4. contenir de l'information auxiliaire.

Deux types de bases :

1. liste : registres d'état civil, des entreprises, des adresses, annuaire et
2. aréolaire : les unités sont des secteurs géographiques.

Les imperfections d'une base : sous-couverture, sur-couverture, répétition, classification erronée.

Absence d'une base : sondage empirique ou considérer une population intermédiaire ;

Types d'erreurs

Nous avons plusieurs types d'erreurs :

1. **erreurs dues à l'échantillonnage** : conséquence du fait qu'un échantillon a été pris et non toute la population ;
2. **erreurs non dues à l'échantillonnage** :
   - erreurs de couverture entre la base de sondage et la population cible ;
   - erreurs de non-réponse :
     totale : pas de réponse à aucune question,
     partielle : pas de réponse a certaines question mais pas à toutes ;
   - erreurs de mesure : la différence entre la vraie valeur et la valeur inscrite ;
   - erreurs de traitement : le codage et la saisie des données.

Population, échantillon

1. Soit la population
   $U = \{u_1, \ldots, u_k, \ldots, u_N\} = \{1, \ldots, k, \ldots, N\}$ avec N connu ou inconnu avant la mise en œuvre de l'enquête ;
2. Un échantillon s est un sous ensemble de U ;
3. Soit une variable Y et nous sommes intéressés par l'estimation du total de Y,

   $$t_Y = \sum_U y_k$$

   ou la moyenne de Y si N est connu,

   $$\overline{y}_U = \frac{1}{N} \sum_U y_k.$$

Plan de sondage $p(\cdot)$

La notion du plan de sondage est spécifique à la théorie des sondages.

- L'ensemble de toutes les parties non vides de U est $\mathcal{S}$.
  Exemple : Soit $U = \{1, 2, 3\}$ alors
  $\mathcal{S} = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$.
- Soit une variable aléatoire $S : (\Omega, \mathcal{K}, P) \to (\mathcal{S}, \mathcal{B}(\mathcal{S}), p)$ avec

$$P(S(w) = s) = p(s).$$

  En effet, l'échantillon s peut être vu comme la réalisation de S de loi $p(\cdot)$.

Définition Le plan de sondage $p(s)$ est une probabilité sur $\mathcal{S}$.

Propriétés d'un plan de sondage $p(s)$

1. comme toute loi de probabilité, nous avons

$$p(s) \geq 0 \quad \text{et} \quad \sum_{s \in \mathcal{S}} p(s) = 1.$$

2. $p(\cdot)$ détermine les propriétés statistiques de quantités calculées dans l'échantillon (voir chapitre 2).
3. $p(\cdot)$ est un outil mathématique qui n'est pas trop utile dans la sélection de l'échantillon.
4. c'est le sondeur qui décide quel plan de sondage sera utilisé : différence avec la statistique classique.
   Attention : $p(s)$ fixé a priori mais pas forcement connu.

Remarque : on supposera pendant ce cours que $p(\cdot)$ ne dépend pas de la variable d'intérêt ; on dit que le plan est non-informatif.

Définition La taille d'un échantillon n_s est le cardinal de s.

Remarque : n_s peut être le même pour tous les échantillons ou non.

Exemple 1

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 prend

$$p(s_1) = 0,25; p(s_2) = 0,25; p(s_3) = 0,2;$$

$$p(s_4) = 0,1; p(s_5) = 0,1; p(s_6) = 0,1;$$

Exemple 2

Soit une population $U = \{1, 2, 3, 4\}$. On considère les six échantillons de taille 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 prend

$$p(s_1) = 1/3; \quad p(s_2) = 1/6; \quad , p(s_3) = 1/2;$$

$$p(s_4) = p(s_5) = p(s_6) = 0;$$

Les probabilités d'inclusion π_k et π_{kl}

Une propriété d'une population finie U avec des éléments identifiés est que différents individus peuvent avoir différentes probabilités de se trouver dans l'échantillon.

Définition : On appelle variable indicatrice la variable aléatoire $I_k = I_k(S)$ définie de la façon suivante :

$$I_k = \begin{cases} 1 & k \in S \\ 0 & sinon \end{cases} \tag{1}$$

Remarque : les variables I_k ne sont pas forcement indépendantes et identiquement distribuées.

Définition : Pour un plan $p(\cdot)$, on appelle probabilité d'inclusion de premier degré π_k, la probabilité que l'individu k se trouve dans un échantillon :

$$\pi_k = P(k \in S) = P(I_k = 1) = \sum_{s \ni k} p(s)$$

Définition : Pour un plan $p(\cdot)$, on appelle probabilité d'inclusion de deuxième degré π_{kl}, la probabilité que les individus k et l se trouvent dans un échantillon :

$$\begin{aligned} \pi_{kl} &= P(k,l \in S) = P(I_k I_l = 1) = \sum_{s \ni k,l} p(s) \\ \pi_{kk} &= \pi_k \end{aligned}$$

Remarques :

• Un plan de sondage est souvent choisi en respectant des π_k et π_{kl} fixés à l'avance ;

• Les π_k sont connus pour tous $k \in U$ avant même la mise en oeuvre de l'enquête dans le cas d'un sondage direct d'éléments (voir sections ...) ; par contre les π_{kl} sont souvent compliqués, voir impossible à calculer ;

• Les π_k ne sont pas caractéristiques du plan de sondage ;

• Les π_k avec $k \in s$ sont fondamentaux pour le calcul des estimateurs.

Remarque : On supposera dans ce cours que $\pi_k > 0$ pour tout $k \in U$.

Application aux exemples 1 et 2

Exemple 1

Calcul de π_k :

$$\begin{aligned}
\pi_1 &= P(1 \in S) = p(s_1) + p(s_2) + p(s_3) = 0,7 \\
\pi_2 &= P(2 \in S) = p(s_1) + p(s_4) + p(s_5) = 0,45 \\
\pi_3 &= P(3 \in S) = p(s_2) + p(s_4) + p(s_6) = 0,45 \\
\pi_4 &= P(4 \in S) = p(s_3) + p(s_5) + p(s_6) = 0,4
\end{aligned}$$

Calcul de π_{kl} :

$$\begin{aligned}
\pi_{12} &= P(1,2 \in S) = p(s_1) = 0,25 \\
\pi_{13} &= P(1,3 \in S) = p(s_2) = 0,25 \\
\pi_{14} &= P(1,4 \in S) = p(s_3) = 0,2 \\
\pi_{23} &= \pi_{24} = \pi_{34} = 0,1
\end{aligned}$$

Exercice : refaire le calcul pour l'exercice 2.

La notion de statistique et d'estimateur

Définition On appelle statistique une fonction réelle de la variable aléatoire S, $Q(S)$. Pour une réalisation $S = s$, Q prend la valeur $Q(s)$. Nous voulons examiner comment une statistique change en fonction des réalisations s de S.
Exemples : $n_S = \sum_U I_k$; $\sum_S y_k$; $\sum_S y_k / \sum_S z_k \rightarrow$

$$Q(S) = Q((k, y_k, z_k, ...); k \in S).$$

Très important : les variables Y and Z ne sont pas aléatoires ; c'est la variable S qui est l'aléa.
Définition : Un estimateur $\widehat{\Phi}$ d'un paramètre Φ est une statistique (fonction de S),

$$\widehat{\Phi} = \widehat{\Phi}(S)$$

et la quantité $\widehat{\Phi}(s)$ obtenue pour une réalisation s de S est appelée estimation de Φ.

Loi d'un estimateur

Loi d'un estimateur $\widehat{\Phi}$: connaissance des couples $(p(s), \widehat{\Phi}(s))$ pour tous les $s \in \mathcal{S}$.

En pratique : impossible de connaître la vraie loi de $\widehat{\Phi}$ à cause de l'indisponibilité de tous les $\widehat{\Phi}(s)$: si tel était le cas, on n'aurait pas eu besoin de faire un sondage ! !

On peut définir :

1. L'espérance de $\widehat{\Phi}(S)$ est $E(\widehat{\Phi}) = \sum_{s \in \mathcal{S}} p(s)\widehat{\Phi}(s)$;
2. La variance de $\widehat{\Phi}(S)$ est $V(\widehat{\Phi}) = \sum_{s \in \mathcal{S}} p(s)(\widehat{\Phi}(s) - E(\widehat{\Phi}))^2$;
3. La covariance
   $Cov(\widehat{\Phi}_1, \widehat{\Phi}_2) = \sum_{s \in \mathcal{S}} p(s)(\widehat{\Phi}_1(s) - E(\widehat{\Phi}_1))(\widehat{\Phi}_2(s) - E(\widehat{\Phi}_2))$.

La qualité d'un estimateur $\widehat{\Phi}$ est jugé à travers :

- le biais $B(\widehat{\Phi}) = E(\widehat{\Phi}) - \Phi$; on préfère $\widehat{\Phi}$ sans biais ou peu biaisé ;
- la variance $V(\widehat{\Phi})$ (inconnue et estimée à l'aide du même s) ; on choisit l'estimateur qui a une plus petite variance ;
- l'erreur quadratique moyenne $EQM(\widehat{\Phi}) = V(\widehat{\Phi}) + B^2(\widehat{\Phi})$;
- le coefficient de variation $CV(\widehat{\Phi}) = \frac{\sqrt{V(\widehat{\Phi})}}{E(\widehat{\Phi})}$.

Exemple 1 :

- Le vrai revenu moyen est $\Phi = \frac{R_1+R_2+R_3+R_4}{4} = 8000$.
- On considère les échantillons de taille 2 et comme estimateur de Φ la moyenne dans chaque échantillon :

$$\widehat{\Phi}(s_1) = \frac{R_1 + R_2}{2} = 9000 \ldots$$

echantillon, s	$p(s)$	$\widehat{\Phi}$	$p(s) \cdot \widehat{\Phi}$
$\{1,2\}$	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

- L'espérance de $\widehat{\Phi}$ est

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

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

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

- L'erreur quadratique moyenne est

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

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

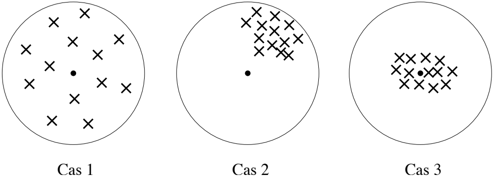

FIGURE: Biais et précision

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).

Intervalles de confiance

Un estimateur peut être sans biais pour un paramètre (la moyenne de ses valeurs sur tous les échantillons possibles) mais nous disposons d'un seul échantillon seulement qui nous fournie une seule estimation pour notre paramètre qui peut être assez éloignée de la vraie valeur (comme vu dans l'exemple prédédent).

On préfère donner une estimation de Φ par intervalles de confiance.

Hypothèse indispensable : $\widehat{\Phi}$ suit une loi normale :

$$IC_\alpha(\widehat{\Phi}) = [\widehat{\Phi} - z_{\alpha/2}\sqrt{V(\widehat{\Phi})}, \widehat{\Phi} + z_{\alpha/2}\sqrt{V(\widehat{\Phi})}]$$

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

Résultat

Soit un plan de sondage $p(\cdot)$. Alors

1. $E(I_k) = \pi_k$;
2. $V(I_k) = \pi_k(1 - \pi_k)$;
3. $Cov(I_k, I_l) = \pi_{kl} - \pi_k \pi_l$.

Résultat

Considérons un plan de sondage $p(\cdot)$ de taille fixe n ($V(n_s) = 0$). Alors,

1. $\sum_U \pi_k = n$;
2. $\sum\sum_{k \neq l} \pi_{kl} = n(n-1)$;
3. $\sum_{l \in U, l \neq k} \pi_{kl} = (n-1)\pi_k$.

Théorie des sondages "versus" statistique inférentielle (1)

Le fait d'avoir d'unités "identifiés" engendre des estimateurs fondamentaux en théorie des sondages et différents de la statistique classique :

Exemple : $N = 3$, $n = 2$ et $s_1 = \{1, 2\}$; $s_2 = \{1, 3\}$ et $s_3 = \{2, 3\}$. On considère $p(s_1) = p(s_2) = p(s_3) = 1/3$ (voir SAS) et on prend

$$t = \begin{cases} t_1 = y_1/2 + y_2/2 & si \quad s_1 \quad \text{tiré} \\ t_2 = y_1/2 + 2y_3/3 & si \quad s_2 \quad \text{tiré} \\ t_3 = y_2/2 + y_3/3 & si \quad s_3 \quad \text{tiré} \end{cases}$$

et la moyenne empirique : $\overline{y}_S = \sum_S y_k/2$. Alors,

- t et $\overline{y}_S$ sont sans biais pour $\overline{y}_U$;
- $V(\overline{y}_S) - V(t) = \frac{y_3(3y_2 - 3y_1 - y_3)}{54} > 0$ pour $y_3(3y_2 - 3y_1 - y_3) > 0$.

Théorie des sondages "versus" statistique inférentielle (2)

Nous avons la possibilité d'améliorer certain estimateurs mais sans pouvoir trouver un unique meilleur estimateur (de variance minimale).

Dans la théorie des sondages :

• le théorème de Rao-Blackwell : pour tout estimateur qui dépend de l'ordre et de la multiplicité des unités dans l'échantillon (pour un tirage avec remise), **on peut trouver un estimateur meilleur qui ne dépend pas de l'ordre ni de la multiplicité**.

Par contre, il n'existe pas une statistique minimale complète et par conséquence, ni d'estimateur de variance uniformément minimale.

Théorie des sondages "versus" statistique inférentielle (3)

- Alors, de façon pratique, grâce au théorème RB, on peut supprimer l'ordre et la multiplicité des unités et considérer que des plans sans remise (sauf certains cas) mais par contre, **nous n'avons pas de méthode pour construire un estimateur**.

- (Godambe, 1955) : Dans la classe des estimateurs sans biais, pour un plan sans remise de taille $n < N$ et $\pi_k > 0$, il n'existe pas d'estimateur optimal de $\overline{y}_U$.

- le théorème de maximum de vraisemblance : il n'existe pas d'estimateur unique de maximum de vraisemblance ;

L'estimateur d'Horvitz-Thompson (HT) du total t_Y

Définition : L'estimateur d'Horvitz-Thompson ou π-estimateur du total t_Y est

$$\hat{t}_\pi = \sum_s \frac{y_k}{\pi_k} = \sum_U \frac{y_k}{\pi_k} I_k. \tag{2}$$

Résultat : (Horvitz-Thompson, 1952)

1. *L'estimateur $\hat{t}_\pi$ est sans biais pour t_Y.*
2. *Supposons que les $\pi_{kl} > 0$ pour tous $k \neq l \in U$. La variance de $\hat{t}_\pi$ est donnée par*

$$V(\hat{t}_\pi) = \sum_U \sum_U \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}, \quad \Delta_{kl} = \pi_{kl} - \pi_k \pi_l. \tag{3}$$

3. *Un estimateur sans biais de la variance est donné par*

$$\widehat{V}(\hat{t}_\pi) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} = \sum_U \sum_U \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} I_k I_l. \tag{4}$$

L'estimateur HT : commentaires

- $\hat{t}_\pi$ est appelé aussi **l'estimateur par les valeurs dilatées** : chaque individu $k \in s$ a un poids dilaté $1/\pi_k > 1$;
- $\hat{t}_\pi$ est le seul estimateur linéaire sans biais dont les poids ne dépendent pas de l'échantillon et de la variable d'intérêt ;
- les doubles sommes de la formule de variance font son calcul difficile ;
- les π_{kl} sont souvent très difficiles à calculer voir impossible pour des plans plus compliqués (à probabilités inégales), alors des formules de variance approchée existent.

Résultat

(Yates-Grundy-Sen, 1953) Si le plan est de taille fixe n, alors

$$V(\hat{t}_\pi) = -\frac{1}{2}\sum_U \sum_U \Delta_{kl}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \quad (5)$$

$$\widehat{V}(\hat{t}_\pi) = -\frac{1}{2}\sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \quad si \quad \pi_{kl} > 0 \quad (6)$$

Chapitre 2 :

- Plans à probabilités égales
  1. Sondage aléatoire simple sans remise (SAS)
  2. Sondage de Bernoulli (BE)
  3. Sondage systématique (SY)
- Sondage stratifié (ST)
- Plan à probabilités inégales
  1. Sondage de Poisson (PO)
  2. Sondage avec remise proportionnel à la taille (PPS)

Sondage aléatoire simple sans remise (SAS) de taille n

Il est très utilisé en pratique.

- Tout échantillon de taille n a la même probabilité d'être sélectionné,

$$p(s) = 1/C_N^n \quad \text{si} \quad s \quad \text{est de taille} \quad n \quad \text{et zéro sinon.}$$

- nombre total d'échantillons : C_N^n;
- $\pi_k = \dfrac{n}{N}$, $k \in U$ et $\pi_{kl} = \dfrac{n(n-1)}{N(N-1)}$, $k \neq l \in U$.
- **Mise en pratique** du SAS : plusieurs façons...

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 : $\dfrac{1}{\mathrm{C}_N^n}$ 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 : un échantillon de taille 2 dans une population de taille 10 ;

• On génère 10 numéros aléatoires uniformes $(0, 1)$:

```
> x=runif(10)
> x
 [1] 0.2887356 0.6560844 0.7098995 0.1535548
 0.6511919 0.2591997 0.1027173 0.
```

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

```
>order(x)
 [1]  7  4  6  1 10  9  8  5  2  3
```

Les individus qui se trouvent sur la 7ème et la 4ème place dans la liste seront sélectionnés (ils ont les deux plus petites nombres aléatoires uniformes).

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

$$5 \quad 8 \quad 10 \quad 12 \quad 15$$

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

• plan SAS, $n = 2$

• Echantillons possibles de taille $n = 2$ et estimations de $\bar{Y}$ par $\bar{y} = \dfrac{y_1 + y_2}{2}$:

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

alors, on peut avoir des "bons échantillons" ou des "mauvais".

Estimation d'un total :

- $\hat{t}_\pi = \frac{N}{n} \sum_s y_k$;
- $V_{SAS}(\hat{t}_\pi) = N^2 \frac{1-f}{n} S^2_{yU}$ avec $S^2_{yU} = \frac{1}{N-1} \sum_U (y_k - \overline{y}_U)^2$ la variance (corrigée) de Y dans la population ;
- $\widehat{V}_{SAS}(\hat{t}_\pi) = N^2 \frac{1-f}{n} S^2_{ys}$ avec $S^2_{ys} = \frac{1}{n-1} \sum_U (y_k - \overline{y}_s)^2$, $\overline{y}_s = \sum_s y_k / n$ la variance de Y dans l'échantillon.

Estimation d'une moyenne $\overline{y}_U$: on divise $\hat{t}_\pi$ par N et par N^2 dans les formules de variance et estimateur de la variance.

Améliorer la qualité : une taille n grande ;
un taux de sondage $f = n/N$ grand ;
une dispersion S^2 faible.

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.

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

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

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

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

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

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

$$
\begin{aligned}
\mathrm{V}(\bar{y_1}) &= 0.9 \times \frac{80}{100} = 0.72 \\
\mathrm{V}(\bar{y_2}) &= 0.9 \times \frac{20}{100} = 0.18
\end{aligned}
$$

- 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

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

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.

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

• L'estimateur HT de P est $\hat{P} = \sum_s y_k/n$ qui est la proportion d'individus ayant A dans l'échantillon s ;
• On a $S^2_{yU} = \frac{N}{N-1}P(1-P)(\simeq P(1-P)$ si N grand) et $V(\hat{P}) = \frac{1-f}{n}S^2_y$;

- 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,

$$e \geq z_{\alpha/2}\sqrt{V(\hat{P})}$$

Il résulte

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

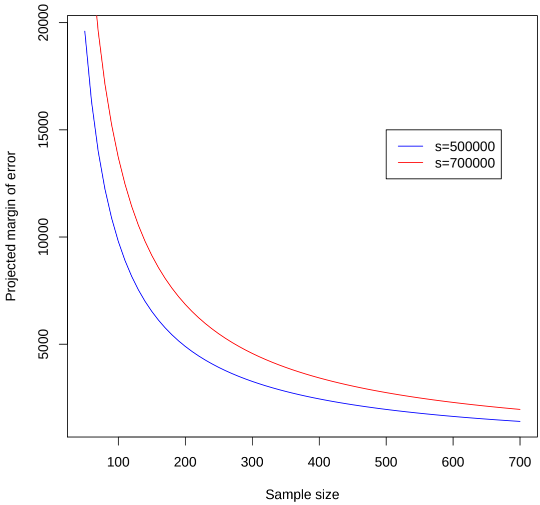

FIGURE: Le graphique de $1.96s/\sqrt{n}$ pour deux valeurs de l'écart-type s

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 $p = 1/2$) ;
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

Sondage de Bernoulli (BE) (1)

C'est le plan pour lequel les variables I_k, $k \in U$, sont indépendantes et de même loi de Bernoulli de paramètre $\pi \in (0, 1)$:

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

- La taille $n_s = \sum_U I_k$ est une variable aléatoire de loi binomiale $\mathcal{B}(N, \pi)$; alors,

$$E(n_s) = N\pi, \quad V(n_s) = N\pi(1 - \pi).$$

- le nombre total d'échantillons est 2^N puisque $s = \emptyset$ ainsi que $s = U$ sont possibles.
- $p(s) = \underbrace{\pi\pi \ldots \pi}_{n_s} \underbrace{(1 - \pi)(1 - \pi) \ldots (1 - \pi)}_{N - n_s} = \pi^{n_s}(1 - \pi)^{N - n_s}$.
- $\pi_k = \pi$ et $\pi_{kl} = \pi^2$.

Sondage de Bernoulli (BE) (2)

Estimation pour un total :

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

Mise en pratique : on génère des variables aléatoires *iid* $\simeq \mathcal{U}_{(0,1)}$, $\varepsilon_1, \ldots, \varepsilon_N$; si $\varepsilon_k < \pi$ alors $k \in s$, sinon on passe à l'unité suivante.
Incovénients : la taille aléatoire et le besoin de parcourir toute la liste pour en avoir s.