

# Cours / TP de sondages approfondis

2<sup>ème</sup> séance : inférence sur des totaux pour des plans complexes

Xavier d'Haultfœuille

## Introduction

- un estimateur n'a de valeur qu'avec sa distribution.
- En sondages, la variance d'une statistique dépend de deux éléments :
  - la statistique considérée ;
  - le plan de sondage.
- Il y a donc une difficulté supplémentaire par rapport à la statistique classique.
- Pour l'instant, on va se “contenter” d'examiner le cas des totaux.

Plan :

- Normalité asymptotique
- Approximation et estimation de la variance
- Quelques algorithmes de tirage

# 1 Normalité asymptotique

- En sondages, on suppose toujours que les estimateurs sont normaux. Cette approximation est-elle justifiée ?

- Pour étudier cela, on définit l'asymptotique en sondages par une suite de populations  $(U_\nu)_{\nu \in \mathbb{N}}$  et des plans de sondage  $(p_\nu(\cdot))_{\nu \in \mathbb{N}}$ . On supposera (au moins) que

$$\begin{aligned} N_\nu &\xrightarrow{\nu \rightarrow +\infty} +\infty \\ \sum_{i=1}^{N_\nu} \pi_{i\nu} &\xrightarrow{\nu \rightarrow +\infty} +\infty \end{aligned}$$

(dans la suite on omet le paramètre  $\nu$ ). La question est alors, pour un estimateur  $\hat{t}_y$  :

$$\frac{\hat{t}_y - t_y}{\sqrt{V(\hat{t}_y)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) ?$$

## 1. Normalité asymptotique

Considérons le  $\pi$ -estimateur :

$$\hat{t}_{y\pi} = \sum_{k \in U} \frac{y_k}{\pi_k} I_k$$

Pourquoi ne peut-on pas appliquer les théorèmes centraux limites standards ?

La normalité asymptotique du  $\pi$ -estimateur a été mise en évidence dans quelques plans de sondage :

- le sondage aléatoire simple (Madow, 1948) ;
- le sondage réjectif (Hájek, 1964, Barbé, 2003) ;
- le sondage successif (Rosen, 1972) ;
- Quelques sondages à plusieurs degrés (Sen, 1988).

## 1. Normalité asymptotique

Le sondage réjectif.

Soit  $(\alpha_1, \dots, \alpha_N) \in [0, 1]^N$  tels que  $\sum_{i=1}^N \alpha_i = 1$ . Le sondage réjectif consiste à tirer un échantillon avec remise avec les probabilités  $\alpha_i$ . Si un individu est tiré deux fois, on retire l'échantillon. On note  $r(\cdot)$  un plan de sondage réjectif,  $R$  l'échantillon correspondant.

**Lemme 1** *Le sondage réjectif est un sondage poissonnien conditionnel à la taille, avec  $\alpha_i \propto \tilde{\pi}_i / (1 - \tilde{\pi}_i)$  (où les  $\tilde{\pi}_i$  sont les probabilités de tirage du sondage poissonnien inconditionnel).*

**Preuve :** exercice.

**Proposition 2** *pour tout jeu de probabilité  $(\pi_1, \dots, \pi_N)$  tels que  $\sum_{k \in S} \pi_k = n$ , il existe  $(\alpha_1, \dots, \alpha_N)$  (ou de façon équivalent des  $\tilde{\pi}_i$ ) tels que  $P_\alpha(k \in R) = \pi_k$ .*

**Preuve :** voir par exemple Chen, Dempster et Liu (1994).

## 1. Normalité asymptotique

Le sondage réjectif existe donc toujours. Le  $\pi$ -estimateur d'un tel sondage est asymptotiquement normal.

**Proposition 3** *On suppose que  $\sum_{i=1}^N \pi_i(1 - \pi_i) \rightarrow +\infty$  (+ conditions techniques). Alors*

$$\frac{\hat{t}_y - t_y}{\sqrt{V(\hat{t}_y)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) ?$$

**Intuition :** d'après le TCL de Lindeberg, pour le plan de Poisson non conditionnel associé aux probabilités  $(\tilde{\pi}_k)$ , on a

$$\left( \begin{array}{c} \frac{\sum_{k \in U} y_k [(I_k / \tilde{\pi}_k) - 1]}{\sqrt{V(\hat{t}_y)}} \\ \frac{\sum_{k \in U} I_k - n}{\sqrt{\sum_{k \in U} \tilde{\pi}_k (1 - \tilde{\pi}_k)}} \end{array} \right) \xrightarrow{\mathcal{L}} \mathcal{N} \left( 0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

Si ces deux variables  $(T_1, T_2)$  étaient *exactement* normales, on aurait

$$T_1 | T_2 = 0 \sim \mathcal{N}(0, 1 - \rho^2).$$

Le sondage réjectif est un plan de Poisson conditionnel à  $T_2 = 0$ . Il est donc raisonnable de penser que l'approximation est valable.

## 1. Normalité asymptotique

Pour généraliser ce résultat on va s'intéresser à l'entropie d'un plan de sondage.

**Définition 4** On appelle entropie du plan  $p(\cdot)$  la quantité  $-\sum_s p(s)\ln(p(s))$ .

**Proposition 5** le sondage réjectif est le plan à entropie maximale de taille fixe et de probabilités d'inclusion d'ordre 1 fixée.

**Preuve :** exercice.

Intuition : les plans de grandes entropies vont se comporter comme les plans réjectifs et le  $\pi$ -estimateur sera donc asymptotiquement normal. On note, pour un plan  $p(\cdot)$ ,

$$D(p||r) = \sum_s p(s) \ln \left( \frac{p(s)}{r(s)} \right) (\geq 0)$$

la divergence entre  $p(\cdot)$  et  $r(\cdot)$ . Un plan proche de l'entropie maximale (donc de  $r(\cdot)$ ) aura une divergence faible.

## 1. Normalité asymptotique

Formalisation (Berger, 1998).

**Proposition 6** *Si  $D(p||r) \rightarrow 0$ , alors*

$$T(S) = \frac{\hat{t}_y - t_y}{\sqrt{V_r(\hat{t}_y)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

où  $V_r(\hat{t}_y)$  est la variance correspondant au sondage réjectif.

**Preuve :** il s'agit de montrer que pour tout  $u$ ,

$$|P_p(T(S) \leq u) - \Phi(u)| \rightarrow 0.$$

où  $P_p$  est la probabilité sous le plan de sondage  $p(\cdot)$ . Or d'après précédemment

$$|P_r(T(S) \leq u) - \Phi(u)| \rightarrow 0.$$

On va donc montrer

$$|P_p(T(S) \leq u) - P_r(T(S) \leq u)| \rightarrow 0.$$

Le résultat suivra de l'inégalité triangulaire. On a

$$\begin{aligned} |P_p(T(S) \leq u) - P_r(T(S) \leq u)| &= \left| \sum_{s/T(s) \leq u} (p(s) - r(s)) \right| \\ &\leq \sum_{s/T(s) \leq u} |p(s) - r(s)| \\ &\leq \sum_s |p(s) - r(s)| \end{aligned}$$

Or (lemme) :

$$\sum_s |p(s) - r(s)| \leq \sqrt{2D(p||r)}.$$

Donc si  $D(p||r) \rightarrow 0$ , on a  $|P_p(T(S) \leq u) - P_r(T(S) \leq u)| \rightarrow 0$ .

## 1. Normalité asymptotique

Berger (1998, 2005) montre que cette condition est satisfaite pour :

- le plan de Rao-Sampford (1965) où l'on tire le 1er individu au hasard avec le jeu de probabilité  $(\pi_1/n, \dots, \pi_N/n)$ , puis les autres avec remise avec une probabilité proportionnelle à  $\pi_k/(1 - \pi_k)$ .
- le plan successif (condition, en gros :  $n^2/N \rightarrow 0$ ). Dans ce plan, on tire à chaque étape (avec remise) un individu  $k$  avec une probabilité  $\alpha_k$ . Si l'individu a déjà été tiré, on en retire un autre.
- Le plan de Chao (1982), où l'on met l'échantillon à jour à chaque étape. Condition :  $n/N \rightarrow 0$ .

Par ailleurs, on peut se rapprocher du plan réjectif à l'aide d'un tri aléatoire.

**Proposition 7** *En moyenne, un tri aléatoire du fichier avant le tirage diminue la divergence d'un tirage :*

$$E_p \left[ \ln \left( \frac{p(S)}{r(S)} \right) \right] \leq E_\nu \left\{ E_p \left[ \ln \left( \frac{p(S|\nu)}{r(S)} \right) \mid \nu \right] \right\}$$

où  $\nu$  est une variable aléatoire uniforme sur l'ensemble des permutations de  $\{1, \dots, N\}$ .

**Preuve :** exercice.

## 2. Calcul de variance

Pour un tirage à probabilités inégales  $\pi_k > 0$ , on a

$$V(\widehat{t}_{y\pi}) = \sum_{k,l} \frac{y_k y_l}{\pi_k \pi_l} \Delta_{kl}$$

Inconvénient de cette formule :

- double somme, très lourde!
- les  $\pi_{kl}$  peuvent être impossibles ou difficiles à calculer.

Nécessité d'une formule « universelle » de variance, s'exprimant comme une somme simple.

Pour cela on va s'appuyer encore sur l'entropie.

## 2. Calcul de variance

Partons du plan réjectif...Définissons :

$$\sigma^2(\pi) = \frac{N}{N-1} \left[ \sum_{k \in U} \frac{y_k^2(1-\pi_k)}{\pi_k} - d(\pi)G(\pi)^2 \right]$$

avec  $d(\pi) = \sum_{k \in U} \pi_k(1-\pi_k)$  et

$$G(\pi) = \frac{1}{d(\pi)} \sum_{k \in U} y_k(1-\pi_k).$$

**Proposition 8** (Hájek, 1964) *Pour un plan de sondage réjectif, on a, lorsque  $d(\pi) \rightarrow +\infty$ ,*

$$\left| \frac{V_r(\hat{t}_{y\pi})}{\sigma^2(\pi)} - 1 \right| \rightarrow 0$$

**Intuition** : le sondage réjectif de probabilité  $(\pi_k)$  est un sondage poissonnien conditionnel de probabilités  $\tilde{\pi}_j$ . Supposons en 1ère approximation que  $\tilde{\pi}_j = \pi_j$ . On a à peu près

$$\begin{pmatrix} \sum_{k \in U} (y_k/\pi_k)(I_k - \pi_k) \\ \sum_{k \in U} (I_k - \pi_k) \end{pmatrix} \sim \mathcal{N}(0, \Sigma)$$

avec

$$\begin{aligned}\Sigma_{11} &= V(\widehat{t}_{y\pi}) = \sum_{k \in U} \frac{y_k^2}{\pi_k} (1 - \pi_k) \\ \Sigma_{22} &= V(\widehat{n}) = \sum_{k \in U} \pi_k (1 - \pi_k) = d(\pi) \\ \Sigma_{12} &= Cov(\widehat{t}_{y\pi}, \widehat{n}) = \sum_{k \in U} \frac{y_k}{\pi_k} \pi_k (1 - \pi_k) = \sum_{k \in U} y_k (1 - \pi_k) = d(\pi)G(\pi)\end{aligned}$$

Donc

$$\begin{aligned}V(\widehat{t}_y | \widehat{n} = n) &\simeq \Sigma_{11} - \frac{\Sigma_{12}^2}{\Sigma_{22}} \\ &\simeq \sum_{k \in U} \frac{y_k^2}{\pi_k} (1 - \pi_k) - d(\pi)G(\pi)^2\end{aligned}$$

Le facteur correctif  $N/N - 1$  assure que la formule est exacte pour un SAS.

N.B. : on pourrait trouver une meilleure approximation en tenant compte du fait que les  $\widetilde{\pi}_k$  sont différents des  $\pi_k$ . cf. Tillé (2001) et Deville et Tillé (2004).

## 2. Calcul de variance

Cette formule reste valable lorsqu'on considère des plans à entropie maximale.

**Proposition 9** (*Berger, 1998*) si  $D(p||r) \rightarrow 0$ , alors

$$\left| \frac{V_p(\hat{t}_{y\pi})}{\sigma^2(\pi)} - 1 \right| \rightarrow 0$$

**Preuve :** on part de

$$\begin{aligned} |V_p(\hat{t}_{y\pi}) - V_r(\hat{t}_{y\pi})| &= \left| \sum_s p(s)(\hat{t}_{y\pi}(s) - t_y)^2 - \sum_s r(s)(\hat{t}_{y\pi}(s) - t_y)^2 \right| \\ &= \sum_s |p(s) - r(s)| (\hat{t}_{y\pi}(s) - t_y)^2 \end{aligned}$$

et on montre, comme précédemment,

$$\sum_s |p(s) - r(s)| (\hat{t}_{y\pi}(s) - t_y)^2 \leq \beta \sqrt{D(p||r)} \sum_s r(s) (\hat{t}_{y\pi}(s) - t_y)^2$$

## 2. Calcul de variance

On peut également approcher  $\sigma^2(\pi)$  de manière convergente par

$$\hat{\sigma}^2(\pi) = \frac{n\hat{d}(\pi)}{(n-1)d(\pi)} \left[ \sum_{k \in S} \left( \frac{y_k}{\pi_k} \right)^2 (1 - \pi_k) - \hat{d}(\pi)\hat{G}(\pi)^2 \right]$$

Cet estimateur est exactement égal à l'estimateur de variance standard dans le cas du SAS. Il est également « bon » lorsque la divergence de  $p(\cdot)$  est faible.

**Proposition 10** (*Berger, 1998*) si  $D(p||r) \rightarrow 0$ , alors

$$\left| \frac{\hat{V}_p(\hat{t}_{y\pi})}{\hat{\sigma}^2(\pi)} - 1 \right| \rightarrow 0$$

## 2. Calcul de variance

Comment fait-on en pratique, où les enquêtes présentent plusieurs degrés de tirage ?

Problème : le fait d'introduire des degrés de tirage fait augmenter sensiblement la divergence du plan de sondage. L'approximation n'est pas valide au global.

Méthode retenue dans le logiciel de calcul de variance POULPE de l'Insee :

- On utilise de façon récursive la formule de décomposition de la variance pour calculer chaque degré. Supposons que l'échantillon  $S = S_0$  corresponde au tirage de  $S_1$ , puis de  $S_2$  parmi les individus de  $S_1$ , etc. On utilise :

$$V(\hat{t}_y|S_n) = E [V(\hat{t}_y|S_{n+1})|S_n] + V [E(\hat{t}_y|S_{n+1})|S_n]$$

On estime chaque terme comme dans le plan à deux degrés.

- A chaque degré de tirage, on suppose que le plan est à entropie maximale.

## 2 Algorithmes de tirage

Algorithme de tirage du plan réjectif.

1ère étape : calcul des  $\alpha_k$  correspondant aux  $\pi_k$ . Notons  $\pi_k(n)$  les probabilités d'inclusion d'un plan réjectif de taille  $n$  correspondant aux  $(\alpha_k)$ . On a

$$\begin{aligned}\pi_k(n) &\propto \sum_{s \ni k, \#s=n} \prod_{i \in s} \alpha_i \\ &\propto \alpha_k \sum_{\substack{s \subset \mathcal{P}(U) \setminus \{k\} \\ \#s=n-1}} \prod_{i \in s} \alpha_i \\ &\propto \alpha_k (1 - \pi_k(n-1))\end{aligned}$$

(le coefficient de proportionnalité est tel que  $\sum_{k \in U} \pi_k = n$ ). De plus

$$\pi_k(1) = \alpha_k.$$

Donc on peut calculer les  $\pi_k(n)$  par récurrence à partir des  $\alpha_k$ . On connaît donc la fonction  $\phi_n$  telle que  $\phi_n(\alpha) = \pi(n)$  (où  $\alpha = (\alpha_1, \dots, \alpha_N)$ ). On peut alors (cf. Tillé, 2001) inverser numériquement  $\phi_n$ .

### 3. Algorithmes de tirage

Une fois les  $(\alpha_k)$  obtenus, considérons :

$$\pi_k(n, V) = P(k \in R | R \subset V) = \frac{\sum_{s \ni k, s \subset V} r(s)}{\sum_{\substack{s \subset V \\ \#s=n}} r(s)} = \frac{\sum_{s \ni k, s \subset V} \prod_{i \in s} \alpha_i}{\sum_{\substack{s \subset V \\ \#s=n}} \prod_{i \in s} \alpha_i}$$

où  $V$  est un sous-ensemble de  $U$ . Connaissant les  $\alpha_k$ , on peut calculer comme précédemment les  $\pi_k(n, V)$  :

$$\pi_k(n, V) = \frac{\alpha_k(1 - \pi_k(n-1, V))}{\sum_{l \in V} \alpha_l(1 - \pi_l(n-1, V))}$$

L'algorithme de Chen, Dempster et Liu (1994) est alors :

- 1ère étape : on sélectionne une unité avec la probabilité  $\pi_k(n, U)/n$ . Soit  $S_1$  l'unité échantillonnée ;
- A la  $i$ -ème étape, on tire une unité à probabilité inégales  $\pi_k(n-i+1, U \setminus \{S_{i-1}\})/(n-i+1)$ .

On note  $S_i$  l'ensemble des unités échantillonnées jusqu'à cette étape.

**Preuve** : cf. Chen, Dempster et Liu (1994).

Remarque : ce n'est pas très simple...

### 3. Algorithmes de tirage

L'algorithme de Chao (1982). Supposons les  $\pi_k$  proportionnels à une variable  $x_k$ . On définit  $U_i = \{1, \dots, i\}$  et pour  $k \leq i$ ,

$$\pi(i; k) = \begin{cases} \frac{nx_k}{\sum_{j=1}^i x_j} & \text{si } i > n; \\ 1 & \text{sinon.} \end{cases}$$

Notons que  $\pi(N; k) = \pi_k$ . Algorithme :

1) Initialisation : on pose  $s_n = U_n$ .

2) A partir d'un échantillon  $s_i$  de taille  $n$  sur  $U_i$  vérifiant les probabilités  $(\pi(i; k))_{k \leq i}$ , on définit  $s_{i+1}$  comme suit :

- on sélectionne  $i + 1$  avec une probabilité  $\pi(i + 1; i + 1)$ ;

- s'il n'est pas tiré,  $s_{i+1} = s_i$

- s'il est tiré  $s_{i+1} = s_i \cup \{i + 1\} \setminus \{j\}$  où  $j$  est sélectionné à partir de  $s_i$  avec la probabilité  $R(j; k)$  :

$$R(i; j) = \frac{1}{\pi(i + 1; i + 1)} \left( 1 - \frac{\pi(i + 1; j)}{\pi(i; j)} \right)$$

Echantillon final :  $s = s_N$ .

### 3. Algorithmes de tirage

L'algorithme tire bien chaque individu avec la probabilité  $\pi_k$ .

On montre par récurrence qu'à chaque étape, la probabilité de tirage vaut  $\pi(i; k)$  :

- Vrai au rang 1 ;
- Si la proposition est vraie au rang  $i$ , alors

$$\begin{aligned} Pr(k \in S_{i+1}) &= \pi(i; k) \times ((1 - \pi(i+1; i+1))R_{ik}) \\ &= \pi(i+1; k) \end{aligned}$$

Donc au rang  $N$ , la probabilité de tirage vaut  $\pi(N; k) = \pi_k$ .

### 3. Algorithmes de tirage

Une méthode générale de tirage à probabilités inégales : la méthode de scission.

Méthode générale de tirage à probabilités inégales. A partir de probabilités  $(\pi_1, \dots, \pi_N)$ , on tire :

- avec une probabilité  $\lambda$  un jeu  $(\pi_1^{(1)}, \dots, \pi_N^{(1)})$  ;
- avec une probabilité  $1 - \lambda$  un jeu  $(\pi_1^{(2)}, \dots, \pi_N^{(2)})$ .

On suppose par ailleurs :

$$\pi_k = \lambda \pi_k^{(1)} + (1 - \lambda) \pi_k^{(2)}$$

$$\sum_{k \in U} \pi_k^{(1)} = \sum_{k \in U} \pi_k^{(2)} = n$$

S'il existe  $k$  tel que  $\pi_k^{(j)} \in \{0, 1\}$ , on est ramené à un problème plus simple. On itère le procédé jusqu'à obtention de l'échantillon final.

**Exemple** : cf. exercice.