TD1–Chapitre 1. Introduction à la théorie des sondages

Exercice 3. (**Que faire de plusieurs échantillons?**) Soit X une variable aléatoire suivant une loi $\mathcal{B}(p)$. On effectue un premier échantillon de taille n_1, $\left(X_i^1\right)_{1\leqslant i\leqslant n_1}$. Puis, plus tard on a l'opportunité de réaliser un second échantillon de taille n_2, $\left(X_i^2\right)_{1\leqslant i\leqslant n_2}$, que l'on supposera réalisé indépendamment du premier. On a donc deux estimateurs de p

$$\hat{P}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i^1 \text{ et } \hat{P}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_i^2$$

Considérons la variable $\tilde{P} = a\hat{P}_1 + b\hat{P}_2$ avec a et b des réels positifs.

1. Calculons l'espérance et la variance de $\tilde{P}$: $\mathbb{E}[\tilde{P}] = (a+b)p$.
2. Y a-t il un choix de a et b qui permette de créer un estimateur de p de meilleure qualité ? On entend par là, un estimateur P qui soit sans biais et de variance la plus faible possible.
3. Conclure

Corrigé :

1. On a $\tilde{P} = a\hat{P}_1 + b\hat{P}_2$ où a et b sont des réels positifs.

   - En utilisant l'expression de $\hat{P}_1$ et celle de $\hat{P}_2$, on a

   $$\mathbb{E}[\hat{P}_1] = \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbb{E}[X_1^i] = p, \quad \mathbb{E}[\hat{P}_2] = \frac{1}{n_2}\sum_{i=1}^{n_2}\mathbb{E}[X_2^i] = p.$$
   $$\mathbb{E}[\tilde{P}] = a\mathbb{E}[\hat{P}_1] + b\mathbb{E}[\hat{P}_2] = ap + bp = (a+b)p.$$

   - On a

   $$\text{Var}(\tilde{P}) = a^2\text{Var}(\hat{P}_1) + b^2\text{Var}(\hat{P}_2).$$

   Puisque $\hat{P}_1$ et $\hat{P}_2$ sont indépendants. La variance de $\hat{P}_1$ et $\hat{P}_2$ est :

   $$\text{Var}(\hat{P}_1) = \frac{p(1-p)}{n_1}, \quad \text{Var}(\hat{P}_2) = \frac{p(1-p)}{n_2},$$
   $$\text{Var}(\tilde{P}) = a^2\frac{p(1-p)}{n_1} + b^2\frac{p(1-p)}{n_2} = p(1-p)\left(\frac{a^2}{n_1} + \frac{b^2}{n_2}\right).$$

2. Choix de a et b pour un estimateur de meilleure qualité. Pour répondre à cette question, on procède comme suit:

   - L'estimateur $\tilde{P}$ soit sans biais (condition de non-biais) : D'après la question 1, on a $\mathbb{E}[\tilde{P}] = (a+b)p$. Pour que $\tilde{P}$ soit sans biais, il faut que $a+b=1$.
   - La variance de $\tilde{P}$ est plus faible ou minimale (Minimisation de la variance): On souhaite minimiser la variance de $\tilde{P}$ sous la contrainte $a+b=1$. On

   $$\text{Var}(\tilde{P}) = p(1-p)\left(a^2\frac{1}{n_1} + (1-a)^2\frac{1}{n_2}\right).$$

   Sous la contrainte $a+b=1$, on remplace b par $1-a$ et on minimise par rapport à a. La fonction à minimiser devient :

   $$\text{Var}(\tilde{P}) = p(1-p)\left(a^2\frac{1}{n_1} + (1-a)^2\frac{1}{n_2}\right).$$

En dérivant cette expression par rapport à a, on obtient :

$$\frac{d}{da}\left(a^2\frac{1}{n_1}+(1-a)^2\frac{1}{n_2}\right)=2a\frac{1}{n_1}-2(1-a)\frac{1}{n_2}$$
$$2a\frac{1}{n_1}=2(1-a)\frac{1}{n_2},$$
$$a\frac{1}{n_1}=(1-a)\frac{1}{n_2}.$$

En réarrangeant, on trouve : $a\left(\frac{1}{n_1}+\frac{1}{n_2}\right)=\frac{1}{n_2}$,
d'où : $a=\frac{n_2}{n_1+n_2}$. Ainsi, $b=1-a=\frac{n_1}{n_1+n_2}$.

Cela donne un estimateur $\tilde{P}$ qui est sans biais et qui minimise la variance. Il s'agit de l'estimateur de p le plus efficace parmi les combinaisons linéaires de $\hat{P}_1$ et $\hat{P}_2$.

3. S'y prendre à deux fois pour échantillonner **la même population** ne semble rien apporter de plus à la qualité de l'estimation par rapport à effectuer un prélèvement de la taille totale dès le départ. Inversement, si pour des raisons de coût ou d'organisation un échantillon de taille suffisante ne peut être effectué en une seule fois, plusieurs prélèvements peuvent être envisagés, en s'assurant bien sur que des covariables éventuelles ne modifient pas les valeurs relevées entre temps.

Exercice 4. (**Estimateur de la variance**) Soit m l'espérance de X que l'on supposera connue. On considère $\tilde{S}^2$ l'estimateur de la variance empirique corrigée et soit T l'estimateur suivant:

$$T=\frac{1}{n}\sum_{i=1}^{n}(X_i-m)^2\,.$$

Déterminer lequel des deux estimateurs est le plus efficace pour estimer la variance de X. (Source Saporta p.286) NB : on utilisera le fait que dans le cas général (c'est-à-dire pas forcément dans le cas gaussien), la variance de l'estimateur de la variance non-corrigée est la suivante

$$\mathrm{Var}\left(S_n^2\right)=\frac{n-1}{n^3}\cdot\left[(n-1)\mu_4-(n-3)\sigma^4\right]$$

Corrigé :

Montrons tout d'abord que $T=\frac{1}{n}\sum_{i=1}^{n}(X_i-m)^2$ est sans biais pour estimer la variance d'une loi $\mathcal{N}(m,\sigma)$ dont l'espérance m est connu.

$$\mathbb{E}[T]=\frac{1}{n}\cdot\sum_i\mathbb{E}\left[(X_i-m)^2\right]=\frac{1}{n}\cdot\sum_i\mathbb{E}\left[(X_i-\mathbb{E}\left[X_i\right])^2\right]$$
$$=\frac{1}{n}\cdot\sum_i\mathrm{Var}\left(X_i\right)=\frac{1}{n}\cdot n\sigma^2.$$

On a donc $\mathbb{E}[T]=\sigma^2$, T est donc bien sans biais pour estimer la variance de X. Calculons maintenant la variance de T. On a

$$\mathrm{Var}(T)=\frac{1}{n^2}\cdot\mathrm{Var}\left(\sum_i(X_i-m)^2\right)$$

Comme les X_i sont indépendants les variables $(X_i-m)^2$ le sont aussi. On en déduit

$$\mathrm{Var}\left(\sum_i(X_i-m)^2\right)=\sum_i\mathrm{Var}\left[(X_i-m)^2\right]=n\cdot\mathrm{Var}\left[(X-m)^2\right].$$

On a donc

$$\begin{aligned}\mathrm{Var}(T) &= \frac{1}{n}\cdot \mathrm{Var}\left[(X-m)^2\right] \\ &= \frac{1}{n}\left\{\mathbb{E}\left[(X-m)^4\right] - \mathbb{E}\left[(X-m)^2\right]^2\right\} \\ &= \frac{1}{n}\cdot\left(\mu_4 - \sigma^4\right).\end{aligned}$$

On sait de plus (cf. énoncé) que

$$\mathrm{Var}\left(S_n^2\right) = \frac{n-1}{n^3}\cdot\left[(n-1)\mu_4 - (n-3)\sigma^4\right].$$

De là

$$\mathrm{Var}\left(\tilde{S_n^2}\right) = \frac{1}{n}\left[\mu_4 - \frac{n-3}{n-1}\cdot\sigma^4\right].$$

On a clairement $\frac{n-3}{n-1} < 1$ d'où

$$\mathrm{Var}\left(\tilde{S_n^2}\right) > \mathrm{Var}(T)$$

On en déduit que si la moyenne de X est connue, T est un meilleur estimateur que $\tilde{S_n^2}$ pour la variance.

Exercice 5. (**Sondage à taille fixe**) Considérons une population de taille $N = 5$ et des échantillons de taille fixe $n = 3$. On supposera ici que tous les individus ont même probabilité d'être sélectionnés.

1. Déterminez le nombre d'échantillons différents possibles.
2. Calculez les probabilités d'inclusion d'ordre 1 et vérifiez la propriété

$$\sum_{k\in U}\pi_k = n$$

3. Calculez les probabilités d'inclusion d'ordre 2 et vérifiez la propriété

$$\sum_{k,\ell\in U, k\neq\ell}\pi_k = n\cdot(n-1)$$

4. Montrez que dans ce cadre, l'estimateur de Horvitz Thomson pour la moyenne est identique à la moyenne empirique usuelle

---

Corrigé :

1. Nombre d'échantillons possibles : c'est le nombre de façon de choisir $n = 3$ individus dans une population de $N = 5$, soit C_N^n. Cela donne ici

$$C_5^3 = \frac{5!}{3!(5-3)!} = \frac{5\times 4\times 3}{3\times 2} = 10$$

Si L'ensemble de la population peut être représenté comme $U = 1, 2, 3, 4, 5$. Nous cherchons tous les sous-ensembles de taille 3 dans cet ensemble. Alors, Les 10 échantillons possibles sont donc les suivants (chaque triplet représente un échantillon de taille 3) :

$$123;\quad 124;\quad 125;\quad 134;\quad 135;\quad 145;\quad 234;\quad 235;\quad 245;\quad 345 \qquad (1)$$

2. Probabilités d'inclusion d'ordre 1 : prenons par exemple l'individu 1. Il intervient dans 6 échantillons sur les 10 possibles. Cela donne $\pi_1 = 6/10 = 3/5$. Et l'on peut effectuer le même calcul pour tous les individus. De manière générale, il existe C_{N-1}^{n-1} échantillons comprenant un individu k donné, ainsi
$$\pi_k = \frac{C_{N-1}^{n-1}}{C_N^n} = n/N.$$
On voit alors que $\sum_{k\in U} \pi_k = n$.

3. Probabilités d'inclusion d'ordre 2 : de la même manière que pour celles d'ordre 1, prenons un exemple avec $k = 1$ et $\ell = 2$. On a alors 3 échantillons qui possèdent à la fois 1 et 2 sur les 10 échantillons, cela nous amène donc $\pi_{k\ell} = 3/10$. Cela est valable pour tous les 10 couples d'individus. On a alors
$$\sum_{k\neq\ell\in U} \pi_{k\ell} = 3.$$
En effet, I
   - Il y a 10 couples distincts dans la population.
   - Pour chaque couple, la probabilité qu'il soit inclus dans un échantillon est
   - La somme des probabilités d'inclusion d'ordre 2 pour tous les couples distincts est donc $\sum_{k\neq\ell\in U} \pi_{k\ell} = 10 \times \frac{3}{10} = 3$.

   Il y a de manière générale C_{N-2}^{n-2} échantillons comprenant un couple donné d'individus sur les C_N^n au total. Cela donne
$$\pi_{k\ell} = \frac{C_{N-2}^{n-2}}{C_N^n} = \frac{n(n-1)}{N(N-1)} = \frac{3}{10}.$$
Et il y a C_N^2 couple d'individus différents au total, en ne contant dans ce raisonnement qu'une seule fois les couples (k,ℓ) et (ℓ,k) ainsi
$$\sum_{k<\ell\in U} \pi_{k\ell} = C_N^2 \cdot \frac{n(n-1)}{N(N-1)} = \frac{n(n-1)}{2} = 3.$$
Si l'on distingue les couples (k,ℓ) et (ℓ,k), il suffit de multiplier par 2 la somme précédente. Cela nous ramène bien à la formule du cours :
$$\sum_{k\neq\ell\in U} \pi_{k\ell} = 2 \times 3 = n(n-1)$$

4. L'estimation de la moyenne par de Horvitz-Thomson pour est
$$\hat{\bar{y}}_\pi = \frac{1}{N} \sum_{k\in s} y_k/\pi_k$$
avec s le sondage, ou échantillon considéré. Comme ici tous les individus ont même probabilité d'inclusion $\pi_k = n/N$, cela donne $\hat{\bar{y}}_\pi = \frac{1}{n}\sum_{k\in S} y_k = \bar{Y}$ qui est bien la moyenne empirique usuelle.

Exercice 6. (**Estimateur du total**) Soit Y une variable définie sur une population U de taille $N = 4$ individus.

k	1	2	3	4
Y	11	10	8	11

On tire un échantillon de taille fixe $n = 2$ sans remise, en supposant l'équiprobabilité des individus.

1. Combien d'échantillons peut-on tirer?
2. Calculez les probabilités d'inclusion d'ordre 1 et 2.
3. On suppose que l'on a tiré/observé l'échantillon composé des individus 1 et 2.
   (a) Quelle est la valeur de l'estimateur de Horvitz-Thomson pour le total sur cet échantillon ?
   (b) Déterminez une estimation de la variance de cet estimateur à partir de l'échantillon tiré.
   (c) Est-il possible d'établir un intervalle de confiance à partir de l'échantillon tiré ?

Corrigé :

1. Nombre d'échantillons possibles (sans remise) : C_N^n. ela donne ici
$$C_4^2 = \frac{4!}{2!(4-2)!} = \frac{4 \times 3 \times 2}{2 \times 2} = 6$$
qui sont les suivants $12;\quad 13;\quad 14;\quad 23;\quad 24;\quad 34$ pour $U = \{1; 2; 3; 4\}$.
2. Probabilités d'inclusion d'ordre 1 : il existe C_{4-1}^{2-1} échantillons comprenant un individu k donné, et
$$\pi_k = \frac{C_{N-1}^{n-1}}{C_N^n} = \frac{n}{N} = \frac{2}{4} = 0.5$$
3. Probabilités d'inclusion d'ordre 2 : prenons un exemple avec $k = 1$ et $\ell = 2$. On a 1 échantillons qui possèdent à la fois 1 et 2 sur les 6 échantillons, cela donc $\pi_{k\ell} = \frac{1}{6}$ et cela est valable pour tous les 6 couples d'individus.
4. L'estimateur du total Horvitz-Thomson sur l'échantillon 1 et 2 nous donne
$$\hat{t_\pi} = \frac{11}{0.5} + \frac{10}{0.5} = 42$$
que l'on compare à la vraie valeur égale à 40 .
5. Estimation de la variance sur l'échantillon : l'application de la formule donnée en cours
$$\hat{\text{Var}}\left(\hat{t_\pi}\right) = -\frac{1}{2} \sum_{k,\ell \in s, k \neq \ell} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell}\right)^2 \cdot \frac{\Delta_{k\ell}}{\pi_{k\ell}}$$
nous donne dans le cas d'un échantillon s de taille 2
$$\hat{\text{Var}}\left(\hat{t_\pi}\right) = -\frac{1}{2}\left[\left(\frac{y_1}{\pi_1} - \frac{y_2}{\pi_2}\right)^2 \cdot \frac{\Delta_{12}}{\pi_{12}} + \left(\frac{y_2}{\pi_2} - \frac{y_1}{\pi_1}\right)^2 \cdot \frac{\Delta_{21}}{\pi_{21}}\right]$$
Avec $\pi_1 = \pi_2 = 1/2, \pi_{12} = \pi_{21} = 1/6$ et $\Delta_{k\ell} = \pi_{k\ell} - \pi_k\pi_\ell = \Delta_{\ell k}$ nous avons $\Delta_{21} = \Delta_{12} = -1/12$ et $\hat{\text{Var}}\left(\hat{t}_\pi\right) = 2$.
6. Pour un niveau de confiance à 95%, le quantile de la loi normale correspondant est égale à 1.96 , d'où l'intervalle de confiance
$$IC = [42 \pm 1.96 \cdot \sqrt{2}] = [39.23; 44.77].$$
(cf. formule du cours $IC_{1-\alpha}(\theta) = \left[\hat{\theta} - u_{1-\alpha/2}\sqrt{\text{Var}(\hat{\theta})}; \hat{\theta} + u_{1-\alpha/2}\sqrt{\text{Var}(\hat{\theta})}\right]$)