Chapitre 2

Sondage aléatoire simple

2.1 Plans simples sans remise

Un plan est simple sans remise de taille fixe n si et seulement si, pour tout s,

$$p(s) = \begin{cases} \binom{N}{n}^{-1} & \text{si } \#s = n \\ 0 & \text{sinon,} \end{cases}$$

où

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}.$$

On peut en déduire les probabilités d'inclusion

$$\pi_k = \frac{n}{N}, \quad \text{et} \quad \pi_{k\ell} = \frac{n(n-1)}{N(N-1)}.$$

Enfin,

$$\Delta_{k\ell} = \frac{n(N-n)}{N^2} \times \begin{cases} 1 & \text{si } k = \ell \\ \dfrac{-1}{N-1} & \text{si } k \neq \ell. \end{cases}$$

L'estimateur de Horvitz-Thompson du total devient

$$\widehat{Y}_\pi = \frac{N}{n} \sum_{k \in S} y_k.$$

et celui de la moyenne s'écrit

$$\widehat{\overline{Y}}_\pi = \frac{1}{n} \sum_{k \in S} y_k.$$

La variance de $\widehat{Y}_\pi$ vaut

$$\text{var}(\widehat{Y}_\pi) = N^2 \left(1 - \frac{n}{N}\right) \frac{S_y^2}{n},$$

et son estimateur sans biais

$$\widehat{\text{var}}(\widehat{Y}_\pi) = N^2 (1 - \frac{n}{N}) \frac{s_y^2}{n},$$

où

$$s_y^2 = \frac{1}{n-1} \sum_{k \in S} \left(y_k - \widehat{\overline{Y}}_\pi\right)^2.$$

L'estimateur de Horvitz-Thompson de la proportion P_D que représente une sous-population D dans la population totale est

$$p = \frac{n_D}{n},$$

où $n_D = \#(S \cap D)$, et p est la proportion d'individus de D dans S. On vérifie :

$$\text{var}(p) = \left(1 - \frac{n}{N}\right) \frac{P_D(1 - P_D)}{n} \frac{N}{N-1},$$

et on estime sans biais cette variance par

$$\widehat{\text{var}}(p) = \left(1 - \frac{n}{N}\right) \frac{p(1-p)}{n-1}.$$

2.2 Plans simples avec remise

Si m unités sont sélectionnées avec remise et à probabilités égales à chaque tirage dans la population U, alors on note $\tilde{y}_i$ la valeur prise par la variable y sur la i-ème unité sélectionnée dans l'échantillon. On peut sélectionner plusieurs fois la même unité dans l'échantillon. L'estimateur de la moyenne

$$\widehat{\overline{Y}}_{AR} = \frac{1}{m} \sum_{i=1}^{m} \tilde{y}_i,$$

est sans biais, et sa variance vaut

$$\text{var}(\widehat{\overline{Y}}_{AR}) = \frac{\sigma_y^2}{m}.$$

Dans un plan simple avec remise, la dispersion

$$\tilde{s}_y^2 = \frac{1}{m-1} \sum_{i=1}^{m} (\tilde{y}_i - \widehat{\overline{Y}}_{AR})^2,$$

TAB. 2.1 – Plans simples : tableau récapitulatif

Plan de sondage simple	Sans remise	Avec remise
Taille de l'échantillon	n	m
Estimateur de la moyenne	$\widehat{\overline{Y}} = \frac{1}{n} \sum_{k \in S} y_k$	$\widehat{\overline{Y}}_{AR} = \frac{1}{m} \sum_{i=1}^{m} \tilde{y}_i$
Variance de l'estimateur de la moyenne	$\text{var}\left(\widehat{\overline{Y}}\right) = \frac{(N-n)}{nN} S_y^2$	$\text{var}\left(\widehat{\overline{Y}}_{AR}\right) = \frac{\sigma_y^2}{m}$
Espérance de la dispersion dans l'échantillon	$\text{E}\left(s_y^2\right) = S_y^2$	$\text{E}\left(\tilde{s}_y^2\right) = \sigma_y^2$
Estimateur de la variance de l'estimateur de la moyenne	$\widehat{\text{var}}\left(\widehat{\overline{Y}}\right) = \frac{(N-n)}{nN} s_y^2$	$\widehat{\text{var}}\left(\widehat{\overline{Y}}_{AR}\right) = \frac{\tilde{s}_y^2}{m}$

estime σ_y^2 sans biais. Il est cependant possible de montrer que si l'on s'intéresse aux n_S unités de l'échantillon $\widetilde{S}$ des unités distinctes, alors l'estimateur

$$\widehat{\overline{Y}}_{UD} = \frac{1}{n_S} \sum_{k \in \widetilde{S}} y_k,$$

est sans biais de la moyenne et a une variance inférieure à celle de $\widehat{\overline{Y}}_{AR}$.

Exercices

Exercice 2.1 *Surface cultivée*

On veut estimer la surface moyenne cultivée dans les fermes d'un canton rural. Sur $N = 2010$ fermes que comprend ce canton, on en tire 100 par sondage aléatoire simple. On mesure y_k la surface cultivée dans la ferme k en hectares, et on trouve

$$\sum_{k \in S} y_k = 2907 \text{ ha et } \sum_{k \in S} y_k^2 = 154593 \text{ ha}^2.$$

1. Donnez la valeur de l'estimateur sans biais classique de la moyenne

$$\overline{Y} = \frac{1}{N} \sum_{k \in U} y_k.$$

2. Donnez un intervalle de confiance à 95 % pour $\overline{Y}$.

Solution

Dans un plan simple, l'estimateur sans biais de $\overline{Y}$ est

$$\widehat{\overline{Y}} = \frac{1}{n} \sum_{k \in S} y_k = \frac{2907}{100} = 29,07 \text{ ha}.$$

L'estimateur de la dispersion S_y^2 vaut

$$s_y^2 = \frac{n}{n-1} \left(\frac{1}{n} \sum_{k \in S} y_k^2 - \widehat{\overline{Y}}^2 \right) = \frac{100}{99} \left(\frac{154593}{100} - 29,07^2 \right) = 707,945.$$

La taille n étant "suffisamment grande", l'intervalle de confiance à 95% est estimé ainsi en hectares :

$$\left[\widehat{\overline{Y}} \pm 1,96 \sqrt{\frac{N-n}{N} \frac{s_y^2}{n}} \right] = \left[29,07 \pm 1,96 \sqrt{\frac{2010-100}{2010} \times \frac{707,45}{100}} \right] = [23,99; 34,15].$$

Exercice 2.2 *Maladie professionnelle*

On s'intéresse à l'estimation de la proportion d'hommes P atteints par une maladie professionnelle dans une entreprise de 1500 travailleurs. On sait par ailleurs que trois travailleurs sur dix sont ordinairement touchés par cette maladie dans des entreprises du même type. On se propose de sélectionner un échantillon au moyen d'un sondage aléatoire simple.

1. Quelle taille d'échantillon faut-il sélectionner pour que la longueur totale d'un intervalle de confiance avec un niveau de confiance 0,95 soit inférieure à 0,02 pour les plans simples avec remise et sans remise ?
2. Que faire si on ne connaît pas la proportion d'hommes habituellement touchés par la maladie (pour le cas du plan sans remise) ?

Pour éviter les confusions de notation, on mettra l'indice AR aux estimateurs avec remise, et l'indice SR aux estimateurs sans remise.

Solution

1. (a) Plan avec remise.
Si le plan est de taille m, la longueur de l'intervalle de confiance (estimé) à un niveau $(1-\alpha)$ pour une moyenne est donnée par

$$\mathrm{IC}(1-\alpha) = \left[\widehat{\overline{Y}} - z_{1-\alpha/2}\sqrt{\frac{\tilde{s}_y^2}{m}}, \widehat{\overline{Y}} + z_{1-\alpha/2}\sqrt{\frac{\tilde{s}_y^2}{m}}\right],$$

où $z_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ d'une variable aléatoire normale centrée réduite. Si on note $\widehat{P}_{AR}$ l'estimateur de la proportion pour le plan avec remise, on peut écrire

$$\begin{aligned}\mathrm{IC}(1-\alpha) \quad = \quad & \left[\widehat{P}_{AR} - z_{1-\alpha/2}\sqrt{\frac{\widehat{P}_{AR}(1-\widehat{P}_{AR})}{m-1}},\right. \\ & \left. \widehat{P}_{AR} + z_{1-\alpha/2}\sqrt{\frac{\widehat{P}_{AR}(1-\widehat{P}_{AR})}{m-1}}\right].\end{aligned}$$

En effet, dans ce cas,

$$\widehat{\mathrm{var}}(\widehat{P}_{AR}) = \frac{\widehat{P}_{AR}(1-\widehat{P}_{AR})}{(m-1)}.$$

Pour que la longueur totale de l'intervalle de confiance ne dépasse pas 0,02, il faut et il suffit que

$$2z_{1-\alpha/2}\sqrt{\frac{\widehat{P}_{AR}(1-\widehat{P}_{AR})}{m-1}} \leq 0,02.$$

En divisant par deux et en élevant au carré, on obtient

$$z_{1-\alpha/2}^2 \frac{\widehat{P}_{AR}(1-\widehat{P}_{AR})}{m-1} \leq 0,0001,$$

ce qui donne

$$m - 1 \geq z_{1-\alpha/2}^2 \frac{\widehat{P}_{AR}(1-\widehat{P}_{AR})}{0,0001}.$$

Pour un intervalle de confiance à 95%, et avec un estimateur de P de 0,3 issu d'une source extérieure à l'enquête, on a $z_{1-\alpha/2} = 1,96$, et

$$m = 1 + 1,96^2 \times \frac{0,3 \times 0,7}{0,0001} = 8068,36.$$

La taille de l'échantillon (m=8069) est donc plus grande que la taille de la population, ce qui est réalisable (mais pas judicieux) puisque le tirage est avec remise.

(b) Plan sans remise.
Si le plan est de taille n, la longueur de l'intervalle de confiance (estimé) à $1-\alpha$ pour une moyenne est donnée par

$$\mathrm{IC}(1-\alpha) = \left[\widehat{\overline{Y}} - z_{1-\alpha/2}\sqrt{\frac{N-n}{N}\frac{s_y^2}{n}}, \widehat{\overline{Y}} + z_{1-\alpha/2}\sqrt{\frac{N-n}{N}\frac{s_y^2}{n}}\right].$$

Pour une proportion P, on a donc, en notant $\widehat{P}_{SR}$ l'estimateur de la proportion pour le plan sans remise

$$\mathrm{IC}(1-\alpha) = \left[\widehat{P}_{SR} - z_{1-\alpha/2}\sqrt{\frac{N-n}{N}\frac{\widehat{P}_{SR}(1-\widehat{P}_{SR})}{n-1}}, \widehat{P}_{SR} + z_{1-\alpha/2}\sqrt{\frac{N-n}{N}\frac{\widehat{P}_{SR}(1-\widehat{P}_{SR})}{n-1}}\right].$$

Pour que la longueur totale de l'intervalle de confiance ne dépasse pas 0,02, il faut et il suffit que

$$2z_{1-\alpha/2}\sqrt{\frac{N-n}{N}\frac{\widehat{P}_{SR}(1-\widehat{P}_{SR})}{n-1}} \leq 0,02.$$

En divisant par deux et en élevant au carré, on obtient

$$z_{1-\alpha/2}^2\frac{N-n}{N}\frac{\widehat{P}_{SR}(1-\widehat{P}_{SR})}{n-1} \leq 0,0001,$$

ce qui donne

$$(n-1)\times 0,0001 - z_{1-\alpha/2}^2\frac{N-n}{N}\widehat{P}_{SR}(1-\widehat{P}_{SR}) \geq 0,$$

ou encore

$$n\left\{0,0001 + z_{1-\alpha/2}^2\frac{1}{N}\widehat{P}_{SR}(1-\widehat{P}_{SR})\right\} \geq 0,0001 + z_{1-\alpha/2}^2\widehat{P}_{SR}(1-\widehat{P}_{SR}),$$

ou

$$n \geq \frac{0,0001 + z_{1-\alpha/2}^2\widehat{P}_{SR}(1-\widehat{P}_{SR}).}{\left\{0,0001 + z_{1-\alpha/2}^2\frac{1}{N}\widehat{P}_{SR}(1-\widehat{P}_{SR})\right\}},$$

Pour un intervalle de confiance à 95%, et avec un estimateur *a priori* de P de 0,3 issu d'une source extérieure à l'enquête, on a

$$n \geq \frac{0,0001 + 1,96^2\times 0,30\times 0,70.}{\left\{0,0001 + 1,96^2\times\frac{1}{1500}\times 0,30\times 0,70\right\}} = 1264,98.$$

Ici, une taille d'échantillon de 1265 suffit. L'ordre de grandeur obtenu justifie bien l'hypothèse d'une loi normale pour $\widehat{P}_{SR}$. L'impact de la correction de population finie $(1-n/N)$ peut donc être déterminant quand la taille de la population est petite et la précision souhaitée relativement forte.

2. Si la proportion de travailleurs touchés n'est pas estimée *a priori*, on se place dans la situation la plus défavorable, c'est-à-dire celle où la variance est la plus forte : cela conduit à une taille n probablement excessive, mais qui a le mérite de garantir que la longueur de l'intervalle de confiance n'est pas supérieure au seuil fixé de $0,02$. Pour le plan sans remise, cela revient à prendre une proportion de 50%. Dans ce cas, en adaptant les calculs du 1-(b), on trouve $n \geq 1298$. On constate donc qu'une variation importante de la proportion (de 30% à 50%), n'entraîne qu'une variation minime de la taille de l'échantillon (de 1265 à 1298).

Exercice 2.3 *Probabilité d'inclusion et plan avec remise*

Dans un plan aléatoire simple avec remise de taille fixe m dans une population de taille N,

1. Calculez la probabilité qu'un individu k soit au moins une fois dans l'échantillon.
2. Montrez que
$$\Pr(k \in S) = \frac{m}{N} + O\left(\frac{m^2}{N^2}\right),$$
lorsque m/N est petit. Pour rappel, on dit qu'une fonction $f(n)$ de n est d'ordre de grandeur $g(n)$ (noté $f(n) = O(g(n))$) si et seulement si $f(n)/g(n)$ est borné, soit si il existe une quantité M telle que, pour tout $n \in \mathbb{N}$, $|f(n)|/g(n) \leq M$.
3. Que peut-on en conclure ?

Solution

1. On obtient cette probabilité en passant par l'événement complémentaire :
$$\Pr(k \in S) = 1 - \Pr(k \notin S) = 1 - \left(1 - \frac{1}{N}\right)^m.$$
2. Ensuite, on développe
$$\begin{aligned}
\Pr(k \in S) &= 1 - \left(1 - \frac{1}{N}\right)^m \\
&= 1 - \sum_{j=0}^{m} \binom{m}{j} \left(-\frac{1}{N}\right)^{m-j} \\
&= 1 - \left\{ \sum_{j=0}^{m-2} \binom{m}{j} \left(-\frac{1}{N}\right)^{m-j} - \frac{m}{N} + 1 \right\} \\
&= \frac{m}{N} - \sum_{j=0}^{m-2} \binom{m}{j} \left(-\frac{1}{N}\right)^{m-j} \\
&= \frac{m}{N} + O\left(\frac{m^2}{N^2}\right).
\end{aligned}$$
3. On conclut que si le taux de sondage m/N est petit, $(m/N)^2$ est négligeable devant m/N. On retrouve alors la probabilité d'inclusion du tirage sans remise, car les deux modes de tirage deviennent indiscernables.

Exercice 2.4 *Taille d'échantillon*

Quelle taille d'échantillon retenir, si on choisit un sondage aléatoire simple, pour connaître à deux points de pourcentage près (au plus) et avec 95 chances sur 100, la proportion des parisiens qui portent des lunettes ?

Solution

Il y a 2 positions raisonnables à adopter d'emblée :

- La taille de la ville de Paris est très grande : le taux de sondage est donc négligeable.
- N'ayant manifestement aucune information *a priori* sur la proportion recherchée, on se place dans la situation qui conduit à une taille d'échantillon maximale (position maximaliste "de précaution"), soit $P = 50$ %. Si la réalité est différente (ce qui est presque certain), on a *in fine* une incertitude inférieure à celle qui était fixée au départ (2 points de pourcentage).

On fixe n de manière à ce que :

$$1,96 \times \sqrt{\frac{P(1-P)}{n}} = 0,02, \text{ avec } P = 0,5,$$

soit $n = 2\ 401$ personnes.

Exercice 2.5 *Nombre d'ecclésiastiques*

On veut estimer le nombre d'ecclésiastiques dans la population française. Pour cela, on choisit d'échantillonner par sondage aléatoire simple n individus. Si la véritable proportion (inconnue) d'ecclésiastiques parmi la population est de 0,1 %, combien faut-il tirer de personnes pour obtenir un coefficient de variation CV de 5 % ?

Solution

Par définition :

$$\text{CV} = \frac{\sigma(Np)}{NP} = \frac{\sigma(p)}{P},$$

où P est la vraie proportion à estimer (ici 0,1 %) et p son estimateur sans biais, c'est-à-dire la proportion d'ecclésiastiques dans l'échantillon tiré. Un CV de 5 % correspond à une précision que l'on peut raisonnablement qualifier de "moyenne". Or ,

$$\text{var}(p) \approx \frac{P(1-P)}{n} \quad (f \textit{ a priori} \text{ négligeable devant } 1).$$

Donc,

$$\text{CV} = \sqrt{\frac{(1-P)}{nP}} \approx \frac{1}{\sqrt{nP}} = 0,05,$$

ce qui donne

$$n = \frac{1}{0,001} \times \frac{1}{0,05^2} = 400\ 000.$$

Cette énorme taille, impossible à obtenir en pratique, découle mécaniquement de la rareté de la sous-population étudiée.

Exercice 2.6 *Taille pour des proportions*

Dans une population de 4 000 personnes, on s'intéresse à 2 proportions :

P_1 = proportion des individus possédant un lave-vaisselle,
P_2 = proportion des individus possédant un ordinateur portable.

D'après des renseignements "sûrs", on sait qu'*a priori* :

$$45\ \% \leq P_1 \leq 65\ \%, \quad \text{et} \quad 5\ \% \leq P_2 \leq 10\ \%.$$

Quelle doit être la taille de l'échantillon n dans le cadre d'un sondage aléatoire simple si on veut connaître *à la fois* P_1 à $\pm$ 2 % près et P_2 à $\pm$ 1 % près, avec un niveau de confiance de 95 % ?

Solution

On estime sans biais $P_i, (i = 1, 2)$ par la proportion p_i calculée dans l'échantillon :

$$\text{var}(p_i) = \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{N}{N-1} P_i(1 - P_i).$$

On veut que :

$$1,96 \times \sqrt{\text{var}(p_1)} \leq 0,02, \quad \text{et que} \quad 1,96 \times \sqrt{\text{var}(p_2)} \leq 0,01.$$

Or ,

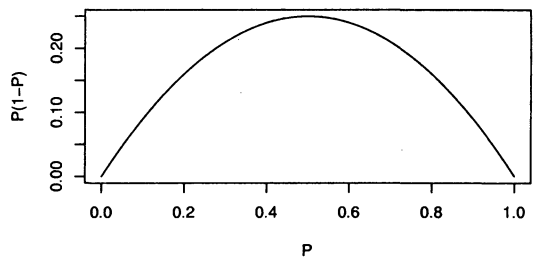
$$\max_{45\,\% \leq P_1 \leq 65\,\%} P_1(1 - P_1) = 0,5(1 - 0,5) = 0,25,$$

et

$$\max_{5\,\% \leq P_2 \leq 10\,\%} P_2(1 - P_2) = 0,1(1 - 0,1) = 0,09.$$

La valeur maximale de $P_i(1 - P_i)$ va conduire au n maximum (à titre de garantie pour

FIG. 2.1 – Variance selon la proportion : exercice 2.6



atteindre au moins la précision souhaitée).
Il faut *conjointement* :

$$\begin{cases} \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{N}{N-1} \times 0,25 \leq \left(\frac{0,02}{1,96}\right)^2 \\ \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{N}{N-1} \times 0,09 \leq \left(\frac{0,01}{1,96}\right)^2, \end{cases}$$

ce qui implique que

$$\begin{cases} n \geq 1\,500,62 \\ n \geq 1\,854,74. \end{cases}$$

La condition sur la précision de p_2 étant la plus exigeante, on conclut en choisissant : $n = 1\,855$.

Exercice 2.7 *Estimation de la dispersion*

Montrez que

$$\sigma_y^2 = \frac{1}{N} \sum_{k \in U} (y_k - \overline{Y})^2 = \frac{1}{2N^2} \sum_{k \in U} \sum_{\substack{\ell \in U \\ \ell \neq k}} (y_k - y_\ell)^2 . \tag{2.1}$$

Utilisez cette égalité pour trouver (facilement) un estimateur sans biais de la dispersion S_y^2 dans le cas du sondage aléatoire simple où $S_y^2 = N\sigma_y^2/(N-1)$.