Sondage aléatoire simple à probabilités égales

Myriam Maumy-Bertrand[1]

[1]IRMA, Université de Strasbourg
Strasbourg, France

Master 1ère Année 26-09-2013

Ce chapitre s'appuie essentiellement sur trois livres :

- « éléments de statistiques »,
  de Jean-Jacques Droesbeke,
  Université de Bruxelles, 2001.
- « Les techniques de sondage »
  de Pascal Ardilly,
  éditions Technip, 2006.
- « Exercices corrigés de méthodes de sondage »
  de Pascal Ardilly et de Yves Tillé,
  éditions Ellipses, 2003.

Sommaire

Définition

Un sondage aléatoire est simple (SAS) si tous les échantillons de taille n fixée a priori, prélevés au sein d'une population U d'effectif N, sont réalisables avec la même probabilité.

Remarque

Dans ce cas, les individus de la population U ont tous la même probabilité d'être choisis pour faire partie de l'échantillon $\mathcal{S}$: leur **probabilité d'inclusion** est une constante.

Remarque

Rappelez ce qu'est une **probabilité d'inclusion** !

Réponse

Vous pouvez trouver une définition p: 51 dans le livre de Ardilly ou alors en allant regarder sur le lien internet suivant : « images.math.cnrs.fr/pdf2006/Lejeune.pdf » ou encore en consultant le cours intitulé « Notations ».

Remarque

Si nous reprenons le choix d'*une seule observation*, chaque individu de la population U a une probabilité égale à $1/N$ d'être prélevé dans la population U afin de constituer l'échantillon $\mathcal{S}$.

Il y a deux méthodes pour sélectionner des individus pour constituer un échantillon $\mathcal{S}$.

La première méthode

Elle consiste à replacer chaque valeur observée dans la population U avant le tirage suivant et cela n fois de suite.
$\Rightarrow$ Prélèvement **avec remise**. Ce type de sondage est dit sondage à probabilités égales avec remise (PEAR).

La deuxième méthode

Elle consiste à ne pas remettre l'individu dans la population U à chaque tirage.
$\Rightarrow$ Prélèvement **sans remise**. Ce type de sondage est dit sondage à probabilités égales sans remise (PESR).

Sommaire

Propriété

Dans ce cas, il y a N^n échantillons $\mathcal{S}$ possibles.

Remarques

- Un même individu peut-être sélectionné plusieurs fois !
- À chaque tirage, la population U est toujours *la même*.

Chaque valeur observée est prise *indépendamment* des autres.

Propriété

L'échantillon $\mathcal{S}$ est alors considéré comme une suite de variables aléatoires **indépendantes et équidistribuées** *$\{Y_1, \ldots, Y_n\}$, où Y_i est la valeur observée pour le i-ème individu sélectionné, telles que*

$$\forall i = 1, \ldots, n \quad \mathbb{E}[Y_i] = \overline{Y} = \mu_Y \quad \textit{et} \quad \mathrm{Var}[Y_i] = \sigma^2_Y,$$

où μ_Y est la moyenne de la population U et σ^2_Y la variance de la population U.

Définition

Un estimateur classique de la moyenne μ d'une population U se définit par :

$$\widehat{\mu}_n = \frac{1}{n} \sum_{i \in \mathcal{S}} Y_i.$$

Propriété

Un calcul direct montre que :

$$\mathbb{E}\left(\widehat{\mu}_n\right) = \mu_Y \quad \textit{et} \quad \mathrm{Var}\left(\widehat{\mu}_n\right) = \frac{\sigma_Y^2}{n}.$$

Remarques

- L'avant dernière égalité de la dernière propriété implique que $\widehat{\mu}_n$ est un estimateur sans biais de la moyenne μ_Y de la population U.
- Dans l'expression de la variance de $\widehat{\mu}_n$, nous remarquons que le terme de la variance σ_Y^2 de la population U intervient. Or, dans la plupart des cas, nous ne connaissons pas la variance σ_Y^2 de la population U. Nous serons donc amené à construire un estimateur de la variance de $\widehat{\mu}_n$.

Définition

Un estimateur de la variance de $\widehat{\mu}_n$ se définit par :

$$\widehat{\mathrm{Var}\left[\widehat{\mu}_n\right]} = \frac{S_{n,c}^2}{n},$$

où $S_{n,c}^2$ désigne la variance corrigée de l'échantillon $\mathcal{S}$.

Propriété

Un calcul direct montre que :

$$\mathbb{E}\left[\widehat{\mathrm{Var}\left[\widehat{\mu}_n\right]}\right] = \frac{\sigma_Y^2}{n}.$$

Remarques

- Rappelons que la variance corrigée $S^2_{n,c}$ de l'échantillon $\mathcal{S}$ se définit par :

$$S^2_{n,c} = \frac{1}{n-1} \sum_{i \in \mathcal{S}} (Y_i - \widehat{\mu}_n)^2$$

  et que $S^2_{n,c}$ est un estimateur sans biais de la variance σ^2_Y de la population U.
- De cette dernière propriété, nous en déduisons que $S^2_{n,c}/n$ est un estimateur sans biais de la variance de $\widehat{\mu}_n$.

Définition

Un estimateur classique du total T_Y d'une population U se définit par :

$$\widehat{T}_n = N\widehat{\mu}_n = \frac{N}{n}\sum_{i\in\mathcal{S}} Y_i.$$

Propriété

Un calcul direct montre que

$$\mathbb{E}\left(\widehat{T}_n\right) = T_Y \quad \textit{et} \quad \mathrm{Var}\left(\widehat{T}_n\right) = N^2\frac{\sigma_Y^2}{n}.$$

Remarques

- L'avant dernière égalité de la dernière propriété implique que $\widehat{T}_n$ est un estimateur sans biais du total T_Y de la population U.
- Dans l'expression de la variance de $\widehat{T}_n$, nous remarquons que le terme de la variance σ_Y^2 de la population U intervient. Or, dans la plupart des cas, nous ne connaissons pas la variance σ_Y^2 de la population U. Nous serons donc amené à construire un estimateur de la variance de $\widehat{T}_n$.

Définition

Un estimateur de la variance de $\widehat{T}_n$ se définit par :

$$\widehat{\mathrm{Var}\left(\widehat{T}_n\right)} = N^2 \frac{S^2_{n,c}}{n},$$

où $S^2_{n,c}$ désigne la variance corrigée de l'échantillon $\mathcal{S}$.

Propriété

Un calcul direct montre que :

$$\mathbb{E}\left[\widehat{\mathrm{Var}\left(\widehat{T}_n\right)}\right] = N^2 \frac{\sigma^2_Y}{n}.$$

Remarque

De cette dernière propriété, nous en déduisons que $N^2 \frac{S^2_{n,c}}{n}$ est un estimateur sans biais de la variance de $\widehat{T}_n$.

Définition

Un estimateur classique de la variance σ_Y^2 d'une population U se définit par : $S_{n,c}^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} (Y_i - \widehat{\mu}_n)^2$.

Propriété

Des calculs montrent que :

$$\mathbb{E}\left(S_{n,c}^2\right) = \sigma_Y^2$$

et

$$\mathrm{Var}\left(S_{n,c}^2\right) = \frac{1}{n(n-1)}\left((n-1)\mu_{Y,4} - (n-3)\sigma_Y^4\right).$$

Remarques

- L'avant dernière égalité de la dernière propriété implique que $S^2_{n,c}$ est un estimateur sans biais de la variance σ^2_Y de la population U.
- Dans l'expression de la variance de $S^2_{n,c}$, nous remarquons que le terme σ^4, qui est le carré de la variance de la population U, intervient ainsi que le moment d'ordre 4, $\mu_{Y,4}$. Or, dans la plupart des cas, nous ne connaissons ni σ^4_Y, ni $\mu_{Y,4}$. Nous serons donc amené à construire un estimateur de la variance de $S^2_{n,c}$, si besoin est.

Le **prélèvement avec remise** est susceptible de fournir plusieurs fois un individu de la population. Deux situations se présentent.

Les n tirages fournissent n individus distincts.

Dans ce cas, $\mathcal{S}$ correspond à un sous-ensemble de U de taille n.

Les définitions de $\widehat{\mu}_n, \widehat{T}_n$ et S_c^2 sont équivalentes si nous renumérotons les individus de la population U de telle sorte que

$$\mathcal{S} = \{1, \ldots, n\}.$$

Les n tirages fournissent m individus, où $m < n$.

Dans ce cas, deux comportements sont à envisager.

- Le premier consiste à prendre en compte les observations autant de fois qu'elles ont été recueillies.
- Le second consiste de prendre la moyenne des m valeurs distinctes observées dont l'ensemble est désigné par $\mathcal{S}_m$:

$$\widehat{\mu}_m = \sum_{k \in \mathcal{S}_m} Y_k.$$

Il est clair que dans ce cas, la taille de n de l'échantillon n'est plus une constante mais devient elle-même une v.a., fonction du processus de prélèvement.

Nous montrons que, en moyenne, $\widehat{\mu}_m$ est encore égal à μ_Y.

Sommaire

Définition

Un sondage aléatoire simple est **sans remise** *si l'observation prélevée au i-ème tirage n'est pas replacée dans la population avant les prélèvements suivants. Ce type de sondage est appelé un sondage à probabilités égales sans remise (PESR)*

Remarque

Un individu est choisi au plus une fois, chaque tirage fait décroître la population U d'une unité.
$\Rightarrow$ Les observations ne sont plus des variables aléatoires indépendantes les unes des autres.

Définition

Un estimateur classique de la moyenne μ d'une population U se définit par :

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Propriété

Des calculs (Ardilly, p :259-261) montrent que :

$$\mathbb{E}\left(\widehat{\mu}_n\right) = \mu_Y,$$

et

$$\mathrm{Var}\left(\widehat{\mu}_n\right) = \frac{N-n}{N-1}\frac{\sigma_Y^2}{n} = (1-f)\frac{N}{N-1}\frac{\sigma_Y^2}{n} = (1-f)\frac{\sigma_{Y,c}^2}{n}.$$

Remarques

- L'avant dernière égalité de la dernière propriété implique que $\widehat{\mu}_n$ est un estimateur sans biais de la moyenne μ_Y de la population.
- Si la taille N de la population U est grande, la variance de $\widehat{\mu}_n$ vaut :

$$\mathrm{Var}\left(\widehat{\mu}_n\right) \approx (1-f)\frac{\sigma_Y^2}{n}.$$

- Dans l'expression de la variance de $\widehat{\mu}_n$, nous remarquons que le terme de la variance corrigée $\sigma_{Y,c}^2$ de la population U intervient. Or, dans la plupart des cas, nous ne connaissons pas la variance corrigée $S_{n,c}^2$ de la population U. Nous serons donc amené à construire un estimateur de la variance de $\widehat{\mu}_n$.

Remarques

- Rappelons que la variance corrigée $S^2_{n,c}$ de l'échantillon $\mathcal{S}$ se définit par :

$$S^2_{n,c} = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \widehat{\mu}_n)^2$$

  et que $S^2_{n,c}$ est un estimateur sans biais de la variance corrigée $\sigma^2_{Y,c}$ de la population U.
- De cette dernière propriété, nous en déduisons que $(1-f)\frac{S^2_{n,c}}{n}$ est un estimateur sans biais de la variance de $\widehat{\mu}_n$.

Définition

Un estimateur classique du total T d'une population U se définit par :

$$\widehat{T}_n = N\widehat{\mu}_n = \frac{N}{n}\sum_{i=1}^{n} Y_i.$$

Propriété

Des calculs (Ardilly p :196-198) montrent que :

$$\mathbb{E}\left(\widehat{T}_n\right) = T_Y \quad et \quad \mathrm{Var}\left(\widehat{T}_n\right) = N^2(1-f)\frac{\sigma^2_{Y,c}}{n}.$$

Remarques

- L'avant dernière égalité de la dernière propriété implique que $\widehat{T}_n$ est un estimateur sans biais du total T_Y de la population U.
- Dans l'expression de la variance de $\widehat{T}_n$, nous remarquons que le terme de la variance corrigée σ_c^2 de la population U intervient. Or, dans la plupart des cas, nous ne connaissons pas la variance corrigée σ_c^2 de la population U. Nous serons donc amené à construire un estimateur de la variance de $\widehat{T}_n$.

Définition

Un estimateur de la variance de $\widehat{T}_n$ se définit par :

$$\widehat{\mathrm{Var}\left(\widehat{T}_n\right)} = N^2(1-f)\frac{S^2_{n,c}}{n},$$

où $S^2_{n,c}$ désigne la variance corrigée de l'échantillon $\mathcal{S}$.

Propriété

Un calcul direct montre que :

$$\mathbb{E}\left(\widehat{\mathrm{Var}\left(\widehat{T}_n\right)}\right) = N^2(1-f)\frac{\sigma^2_{Y,c}}{n}.$$

Remarque

De cette dernière propriété, nous en déduisons que

$N^2(1-f)\dfrac{S^2_{n,c}}{n}$ est un estimateur sans biais de la variance de $\widehat{T}_n$.

Définition

Un estimateur de la variance σ^2 d'une population U, dans le cas d'un sondage aléatoire simple à probabilités égales sans remise, se définit par :

$$\widehat{\sigma}_n^2 = \frac{N-1}{N} S_{n,c}^2 = \frac{N-1}{N} \frac{1}{n-1} \sum_{i \in \mathcal{S}} (Y_i - \widehat{\mu}_n)^2 .$$

Propriété

Des calculs (Ardilly et Tillé, p :43-49) montrent que :

$$\mathbb{E}\left(\widehat{\sigma}_n^2\right) = \mathbb{E}\left(\frac{N-1}{N} S_{n,c}^2\right) = \sigma^2$$

et

$$\begin{aligned}\mathrm{Var}\left[\widehat{\sigma}_n^2\right] &= \frac{(N-n)}{n(n-1)N(N-2)(N-3)} \\ &\quad \times \Big\{ \mu_4 (N-1)\left[N(n-1)-(n+1)\right] \\ &\quad -\sigma^4\left[N^2(n-3)+6N-3(n+1)\right] \Big\}.\end{aligned}$$

Remarques

- L'avant dernière égalité de la dernière propriété implique que $\widehat{\sigma}_n^2$ est un estimateur sans biais de la variance σ^2 de la population U.
- Dans l'expression de la variance de $\widehat{\sigma}_n^2$, nous remarquons que le terme σ^4, qui est le carré de la variance de la population U, intervient ainsi que le moment d'ordre 4, μ_4. Or, dans la plupart des cas, nous ne connaissons ni σ^4, ni μ_4. Nous serons donc amené à construire un estimateur de la variance de $\widehat{\sigma}_n^2$, si besoin est.

Sommaire

Remarques

- Les deux méthodes conduisent toutes les deux à des estimateurs $\widehat{\mu}_n$ qui sont, **en moyenne** égaux au paramètre μ_Y de la population.
- Par contre les variances de $\widehat{\mu}_n$ ne sont pas égales !

Problème

Qui est le meilleur estimateur de la moyenne μ_Y de la population parmi ces deux estimateurs ?

Pour répondre à cette question, nous allons utiliser une méthode.

Remarque

En général, les estimateurs que nous devons comparer sont en moyenne égaux au paramètre à estimer. Ils ne différent que par leur variance. (La variance est un paramètre de précision de l'estimateur.)

Proposition

Pour comparer deux estimateurs ou deux méthodes qui produisent des estimateurs différents, nous utilisons **l'effet de sondage**.

Définition

L'effet de sondage *est défini par :*

$$D(\widehat{\theta^*}|\widehat{\theta}) = \frac{\text{Var}\left[\widehat{\theta^*}\right]}{\text{Var}\left[\widehat{\theta}\right]}.$$

Remarque

Si $D(\widehat{\theta^*}|\widehat{\theta}) < 1$, alors $\widehat{\theta^*}$ sera plus précis que $\widehat{\theta}$.

Rappelons

Propriété

$$\mathrm{Var}\left(\widehat{\mu}_{PEAR}\right) = \frac{\sigma^2}{n} \quad et \quad \mathrm{Var}\left(\widehat{\mu}_{PESR}\right) = \frac{\sigma^2}{n}\frac{N-n}{N-1}.$$

Il s'en suit que :

$$D\left(\mathrm{PESR}|\mathrm{PEAR}\right) = \frac{N-n}{N-1}.$$

Si $n > 1$, alors nous avons $N - n < N - 1$. Par conséquent, nous obtenons :

$$D(\mathrm{PESR}|\mathrm{PEAR}) < 1.$$

Conclusion

La précision de l'estimateur est donc meilleure si nous utilisons un échantillon aléatoire simple PESR qu'un échantillon aléatoire simple PEAR.

Remarque

Ce dernier résultat est intuitif car il y a une perte d'information dès que certains individus sont observés plus d'une fois, ce qui est impossible lors d'un tirage sans remise.

Remarques

- Si la taille de la population est grande, l'effet de sondage est tel que

$$D(\text{PESR}|\text{PEAR}) = \frac{N-n}{N-1} \approx \frac{N-n}{N} = 1-f,$$

où f est le taux de sondage. L'amélioration de la précision est d'autant meilleure que f est grand.

- La différence entre les deux procédures faiblit quand la taille de l'échantillon est petite par rapport à celle de la population, i.e. quand f est faible ! Dans ce cas l'effet de sondage est proche de 1, les deux méthodes fournissent des estimateurs de précision analogue.