Sondage aléatoire simple à probabilités égales

Myriam Maumy[1]

[1]IRMA, Université Louis Pasteur
Strasbourg, France

Master 1ère Année 06-03-2006

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

Qu'appelle-t-on une **probabilité d'inclusion** ?

Réponse

Vous pouvez trouver une définition p: 51 dans le livre de Ardilly.

Remarque

Si l'on reprend 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**. On parle d'un 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**. On parle d'un sondage à probabilités égales sans remise (PESR).

Sommaire

Propriété

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

Remarque

Un même individu peut-être sélectionné plusieurs fois !

Remarque

À chaque tirage, on a toujours *la même population U*.

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** *$\{X_1, \ldots, X_n\}$, où X_i est la valeur observée pour le i-ème individu sélectionné, telles que*

$$\forall i = 1, \ldots, n \quad \mathbb{E}[X_i] = \mu \quad \textit{et} \quad \mathrm{Var}[X_i] = \sigma^2,$$

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

Définition

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

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

Propriété

On montre, par calculs directs, que

$$\mathbb{E}[\widehat{\mu}] = \mu \quad \textit{et} \quad \mathrm{Var}[\widehat{\mu}] = \frac{\sigma^2}{n}.$$

Remarque

L'avant dernière égalité de la dernière propriété implique que $\widehat{\mu}$ est un estimateur sans biais de la moyenne μ de la population U.

Remarque

Dans l'expression de la variance de $\widehat{\mu}$, on remarque que le terme de la variance σ^2 de la population U intervient. Or, dans la plupart des cas, on ne connaît pas la variance σ^2 de la population U. On sera donc amené à construire un estimateur de la variance de $\widehat{\mu}$.

Définition

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

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

où s_c^2 désigne la variance corrigée de l'échantillon $\mathcal{S}$.

Propriété

On montre, par calcul direct, que

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

Remarque

On rappelle que la variance corrigée s_c^2 de l'échantillon $\mathcal{S}$ se définit par

$$s_c^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} (x_i - \widehat{\mu})^2$$

et que s_c^2 est un estimateur sans biais de la variance σ^2 de la population U.

Remarque

De cette dernière propriété, on en déduit que $\dfrac{s_c^2}{n}$ est un estimateur sans biais de la variance de $\widehat{\mu}$.

Définition

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

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

Propriété

On montre, par calculs directs, que

$$\mathbb{E}\left[\widehat{T}\right] = T \quad \textit{et} \quad \mathrm{Var}\left[\widehat{T}\right] = N^2\frac{\sigma^2}{n}.$$

Remarque

L'avant dernière égalité de la dernière propriété implique que $\widehat{T}$ est un estimateur sans biais du total T de la population U.

Remarque

Dans l'expression de la variance de $\widehat{T}$, on remarque que le terme de la variance σ^2 de la population U intervient. Or, dans la plupart des cas, on ne connaît pas la variance σ^2 de la population U. On sera donc amené à construire un estimateur de la variance de $\widehat{T}$.

Définition

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

$$\widehat{\mathrm{Var}\left[\widehat{T}\right]} = N^2 \frac{s_c^2}{n},$$

où s_c^2 désigne la variance corrigée de l'échantillon $\mathcal{S}$.

Propriété

On montre, par calcul direct, que

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

Remarque

De cette dernière propriété, on en déduit que $N^2 \frac{s_c^2}{n}$ est un estimateur sans biais de la variance de $\widehat{T}$.

On rappelle que :

Définition

Un estimateur classique de la variance σ^2 d'une population U se définit par

$$s_c^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} (x_i - \widehat{\mu})^2 .$$

Propriété

On montre, par calculs directs, que

$$\mathbb{E}\left[s_c^2\right] = \sigma^2 \quad \textit{et} \quad \mathrm{Var}\left[s_c^2\right] = \frac{1}{n(n-1)} \left[(n-1)\mu_4 - (n-3)\sigma^4\right] .$$

Remarque

L'avant dernière égalité de la dernière propriété implique que s_c^2 est un estimateur sans biais de la variance σ^2 de la population U.

Remarque

Dans l'expression de la variance de s_c^2, on remarque 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, on ne connaît ni σ^4, ni μ_4. On sera donc amené à construire un estimateur de la variance corrigée s_c^2, 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 $\widehat{\mu}, \widehat{T}$ et s_c^2 sont équivalentes si on renumérote 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 $\mathcal{S}_m$:

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

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.

On montre que, en moyenne, $\widehat{\mu}_m$ est encore égal à μ.

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. On parle alors d'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

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

Propriété

On montre, par calculs (Ardilly, p :196-198), que

$$\mathbb{E}[\widehat{\mu}] = \mu,$$

et

$$\mathrm{Var}[\widehat{\mu}] = \frac{N-n}{N-1}\frac{\sigma^2}{n} = (1-f)\frac{N}{N-1}\frac{\sigma^2}{n} = (1-f)\frac{S_c^2}{n}.$$

Remarque

L'avant dernière égalité de la dernière propriété implique que $\widehat{\mu}$ est un estimateur sans biais de la moyenne μ de la population.

Remarque

Si la taille N de la population U est grande, la variance de $\widehat{\mu}$ vaut :

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

Remarque

Dans l'expression de la variance de $\widehat{\mu}$, on remarque que le terme de la variance corrigée S_c^2 de la population U intervient. On sera donc amené à construire un estimateur de la variance de $\widehat{\mu}$.

Remarque

On rappelle que la variance corrigée s_c^2 de l'échantillon $\mathcal{S}$ se définit par

$$s_c^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \widehat{\mu})^2$$

et que s_c^2 est un estimateur sans biais de la variance corrigée S_c^2 de la population U.

Remarque

De cette dernière propriété, on en déduit que $(1-f)\frac{s_c^2}{n}$ est un estimateur sans biais de la variance de $\widehat{\mu}$.

Définition

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

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

Propriété

On montre, par calculs (Ardilly p :196-198), que

$$\mathbb{E}\left[\widehat{T}\right] = T \quad et \quad \mathrm{Var}\left[\widehat{T}\right] = N^2(1-f)\frac{S_c^2}{n}.$$

Remarque

L'avant dernière égalité de la dernière propriété implique que $\widehat{T}$ est un estimateur sans biais du total T de la population U.

Remarque

Dans l'expression de la variance de $\widehat{T}$, on remarque que le terme de la variance corrigée S_c^2 de la population U intervient. Or, dans la plupart des cas, on ne connaît pas la variance corrigée S_c^2 de la population U. On sera donc amené à construire un estimateur de la variance de $\widehat{T}$.

Définition

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

$$\widehat{\mathrm{Var}\left[\widehat{T}\right]} = N^2(1-f)\frac{s_c^2}{n},$$

où s_c^2 désigne la variance corrigée de l'échantillon $\mathcal{S}$.

Propriété

On montre, par calcul direct, que

$$\mathbb{E}\left[\widehat{\mathrm{Var}\left[\widehat{T}\right]}\right] = N^2(1-f)\frac{S_c^2}{n}.$$

Remarque

De cette dernière propriété, on en déduit que $N^2(1-f)\frac{s_c^2}{n}$ est un estimateur sans biais de la variance de $\widehat{T}$.

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

$$\widehat{\sigma^2} = \frac{N-1}{N} s_c^2 = \frac{N-1}{N} \frac{1}{n-1} \sum_{i \in \mathcal{S}} (x_i - \widehat{\mu})^2 .$$

Propriété

On montre, par calculs (Ardilly et Tillé, p :43-49), que

$$\mathbb{E}\left[\widehat{\sigma^2}\right] = \mathbb{E}\left[\frac{N-1}{N}s_c^2\right] = \sigma^2$$

et

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

Remarque

L'avant dernière égalité de la dernière propriété implique que $\widehat{\sigma^2}$ est un estimateur sans biais de la variance σ^2 de la population U.

Remarque

Dans l'expression de la variance de $\widehat{\sigma^2}$, on remarque 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, on ne connaît ni σ^4, ni μ_4. On sera donc amené à construire un estimateur de la variance de $\widehat{\sigma^2}$, si besoin est.

Sommaire

Remarque

Les deux méthodes conduisent toutes les deux à des estimateurs $\widehat{\mu}$ qui sont, **en moyenne** égaux au paramètre μ de la population.

Remarque

Par contre les variances de $\widehat{\mu}$ ne sont pas égales !

Problème :

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

Pour répondre à cette question, on va utiliser une méthode.

Remarque

En général, les estimateurs que l'on doit 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, on utilise **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 $\widehat{\theta}$.

On rappelle que

Propriété

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

Il s'en suit que

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

Si

$$n > 1,$$

alors

$$N - n < N - 1.$$

Par conséquent

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

Conclusion

La précision de l'estimateur est donc meilleure si l'on utilise 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

1. 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.

2. 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.