

ECHANTILLONNAGE - ESTIMATION PONCTUELLE D'UN PARAMETRE

A. DAOUI

3 décembre 2015

Introduction à la méthode des sondages

Supposons que l'on veuille étudier certaines caractéristiques d'une population de taille N assez grande. Recourir au *recensement* ou une *enquête exhaustive* (individu par individu) demanderait beaucoup de moyens en coût et en temps. Une deuxième méthode dite *enquête par sondage* ou tout simplement *sondage* ne porte que sur une partie de la population étudiée, dite échantillon de taille $n \ll N$, et essayer de trouver une approximation des différentes caractéristiques de la population concernée.

Les enquêtes par sondage n'ont d'intérêt que si l'échantillon est choisi de telle sorte qu'il soit représentatif, autrement dit de façon que les informations collectées puissent être étendues à l'ensemble de la population. Le taux $t = \frac{n}{N}$ est appelé taux de sondage.

Méthodes de sondage ou d'échantillonnage

méthode aléatoire

La méthode la plus utilisée pour prélever un échantillon est la *méthode aléatoire* qui consiste à affecter à chaque individu une probabilité non nulle d'être choisi au sein de l'échantillon et de procéder à un tirage aléatoire. Le plus souvent, cette probabilité est la même pour tous les individus. Les tirages peuvent être effectués de deux façons :

- Avec remise (tirages *indépendants* ou *bernoulliens*) et dans ce cas, le nombre X d'unités-échantillon présentant un caractère déterminé est une variable aléatoire *binomiale*.
- Sans remise (Tirages *exhaustifs*) et de ce cas X est une variable aléatoire *hypergéométrique*.

Méthodes de sondage ou d'échantillonnage : Stratification

Il s'agit d'un procédé utilisé pour améliorer la précision des sondages aléatoires ; il consiste à découper la population étudiée en groupes homogènes, appelés **strates**, et à tirer indépendamment un échantillon aléatoire dans chaque strate pour constituer l'échantillon final. Ces strates seront arrêtées selon des critères ou variables de contrôle fixées d'avance et qui doivent être en corrélation étroite avec les variables étudiées

ESTIMATIONS

Estimation - Estimateur

Commençons par un exemple et supposons qu'un organisme économique, ayant prélevé un échantillon de 10000 ménages, constate que le montant moyen des dépenses consacrées au logement est $\bar{x} = 2000$ Dhs.

Comment ce montant a été calculé ? et comment à partir de cette valeur, estimer la moyenne m des dépenses en logement de toute la population ?

Ce montant est calculé à partir des obsevations ou relevés des 10000 familles : $x_1, x_2, \dots, x_{1000}$ (Dhs)

et $\bar{x} = \frac{1}{1000} \sum_{i=1}^{1000} x_i$; on dit que \bar{x} est une *statistique*.

Mais la moyenne de l'échantillon avant la désignation de celui-ci est la V.A. $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ qui a pour moyenne le paramètre m recherché et pour écart-type $\frac{\sigma}{\sqrt{n}}$ (si le tirage est exhaustif). On dit que la V.A. \bar{X}_n est un estimateur du paramètre m

définitions

Donnons à présent les définitions générales des termes vus dans l'exemple précédent.

- ① Une *statistique* s est une fonction des observations x_1, x_2, \dots, x_n .
Par exemple : $\bar{x}_n = \sum_{i=1}^n \frac{x_i}{n}$, $x^* = \min x_i$, $s_n^2 = \sum_{i=1}^n \frac{(x_i - \bar{x}_n)^2}{n}$
- ② Un *estimateur* d'un paramètre θ est une variable aléatoire T_n à valeurs dans l'ensemble des valeurs possibles de θ .
- ③ Une estimation de θ est une réalisation ou une valeur prise par l'estimateur T_n ; elle peut être donc apparentée à une statistique. Une estimation est donc une valeur déterministe alors qu'un estimateur est une variable aléatoire.

L'estimation qu'on espère être la plus proche de la vraie valeur de θ peut être donnée à l'aide d'une seule valeur vraisemblable (estimation ponctuelle) ou ensemble de valeurs vraisemblables (estimation ensembliste) ou à l'aide d'intervalles dit intervalle de confiance. Notons qu'à chaque intervalle de confiance I correspond un risque d'erreur $\alpha = p(\theta \notin I)$ ou un seuil de confiance $1 - \alpha = p(\theta \in I)$

Méthode des moments (EMM)

Il s'agit de la méthode la plus naturelle, que nous avons déjà utilisée sans évoquer l'appellation. L'idée de base est d'estimer une espérance mathématique par une moyenne empirique, une variance par une variance empirique, etc...

Si le paramètre à estimer est $\theta = E(X)$, alors on peut l'estimer à l'aide de la moyenne empirique ou moyenne d'échantillonnage ; l'estimateur de $\theta = E(X)$ par la méthode des moments (**EMM**) est

$$\tilde{\theta} = \bar{X}_n = \sum_{i=1}^n \frac{X_i}{n}$$

De la même manière, on estime la variance de X par la variance empirique de l'échantillon :

$$S_n^2 = \sum_{i=1}^n \frac{(X_i - \bar{X}_n)^2}{n}$$

Plus généralement, pour un paramètre $\theta \in \mathbb{R}$, si $E(X) = \varphi(\theta)$ avec φ une fonction inversible, alors l'EMM de θ est $\varphi^{-1}(\bar{X}_n)$

Exemples

- ➊ Loi de Bernoulli $\mathcal{B}(p)$
- ➋ Loi exponentielle $\exp(\lambda)$
- ➌ Loi normale $\mathcal{N}(m, \sigma)$

Méthode du maximum de vraisemblance

La fonction de vraisemblance

Définition

On appelle fonction de vraisemblance associée aux observations échantillonnaires x_1, x_2, \dots, x_n , la fonction du paramètre θ :

$$\mathcal{L}(\theta; x_1, x_2, \dots, x_n) = \begin{cases} \prod_{i=1}^n p(X_i = x_i; \theta) & \text{si les } X_i \text{ sont discrètes et indépendantes} \\ \prod_{i=1}^n f_{X_i}(x_i; \theta) & \text{si les } X_i \text{ sont continues et indépendantes} \end{cases}$$

Estimateur du maximum de vraisemblance (EMV)

La méthode consiste à trouver la valeur la plus vraisemblable du paramètre θ pour laquelle les statistiques x_1, x_2, \dots, x_n ont le plus de chance d'être observées sur un tel échantillon. L'**EMV** est donc une solution de l'équation de vraisemblance :

$$\frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta; X_1, X_2, \dots, X_n) = 0$$

vérifiant $\frac{\partial^2}{\partial \theta^2} \ln \mathcal{L}(\theta; X_1, X_2, \dots, X_n) < 0$. Il s'agit donc d'une maximisation de la fonction de vraisemblance et donc des probabilités dépendantes des observations.

Exemples

Exemple introductif

On désire estimer le paramètre p d'une loi binomiale $\mathcal{B}(15, p)$. Pour simplifier on prend ici une seule observation c.à.d. un échantillon de taille $n = 1$. On observe $x_1 = 5$. La fonction de vraisemblance est :

$$\mathcal{L}(p; 5) = p(X_1 = 5; p) = C_{15}^5 p^5 (1 - p)^{10}$$

C'est la probabilité d'avoir observé un 5 quand le paramètre est p . On cherche la valeur de p qui maximise cette probabilité. D'après la table de la loi binomiale, on trouve $p = 0.8$.

Calculons la valeur vraisemblable de p .

Exemples

- ➊ Loi de Bernoulli $\mathcal{B}(p)$
- ➋ Loi exponentielle $\exp(\lambda)$
- ➌ Loi normale $\mathcal{N}(m, \sigma)$

Qualité d'un estimateur

Biais d'un estimateur - estimateur de variance minimum

A priori, n'importe quelle fonction des observations à valeurs dans l'ensemble des valeurs possibles de θ est un estimateur de θ . Mais un estimateur de θ ne sera satisfaisant que si, pour n'importe quelle observation x_1, x_2, \dots, x_n , l'estimation t_n est "proche" de θ . Par exemple les risques $|T_n - \theta|$, $(T_n - \theta)^2$ expriment bien un écart entre T_n et θ . On utilisera des quantités déterministes.

Pour cela, il faut d'abord que si on répète plusieurs fois l'expérience, la moyenne des estimations obtenues soit très proche, et dans l'idéal, égale à θ .

Définition

Un estimateur T_n du paramètre θ est dit sans biais si $E(T_n) = \theta$.
Sinon il sera biaisé de biais

$$B(T_n) = E(T_n) - \theta$$

Un biais est une erreur systématique. Malgré les inconvénients d'une telle erreur, il peut être avantageux d'utiliser un estimateur légèrement biaisé si sa variance $V(T_n) = E[(T_n - E(T_n))^2]$ est petite ; c'est-à-dire s'il présente une faible dispersion.

Definition

Le **risque quadratique** ou **erreur quadratique moyenne** est

$$EQM(T_n) = E[(T_n - \theta)^2]$$

Remarquons que : $EQM(T_n) = V(T_n) + [E(T_n) - \theta]^2$
= Variance de l'estimateur + carré

de son biais

et donc si T_n est sans biais alors $EQM(T_n) = V(T_n)$

On a donc intérêt à ce que l'estimateur soit *sans biais* et de variance minimale (**ESBVM**)

Remarques

- *Un estimateur biaisé peut être plus intéressant si son erreur quadratique moyenne est inférieure à la variance d'un estimateur sans biais.*
- *De deux estimateurs sans biais, le meilleur est celui ayant une variance faible ; on dit que c'est le plus efficace. On étudiera l'efficacité ci-après.*

Enfin, il est logique de s'attendre à ce que plus la taille des données augmente, plus on a d'information sur le phénomène aléatoire observé, donc meilleure sera l'estimation. En théorie, donc une quantité infinie d'observations doit conduire à une estimation sans erreur. On peut traduire ceci par le fait que le risque de l'estimateur T_n doit tendre vers 0.

Définition

On dit que l'estimateur T_n converge en moyenne quadratique vers θ lorsque son erreur quadratique tend vers 0 quand n tend vers l'infini.

$$T_n \xrightarrow{MQ} \theta \Leftrightarrow \lim_{n \rightarrow \infty} E[(T_n - \theta)^2] = 0$$

Remarque

Si T_n est sans biais, il sera convergent en moyenne quadratique si et seulement si sa variance tend vers 0 quand n tend vers l'infini.

Quantité d'information de Fisher, efficacité d'un estimateur

La quantité d'information de Fisher est outil précieux pour évaluer la qualité d'un estimateur et en particulier sa variance ; c.à.d., son efficacité en fournissant un "bon" minorant de sa variance.

Définition

Pour un paramètre réel θ , on appelle quantité d'information de Fisher sur θ fournie par l'échantillon x_1, x_2, \dots, x_n , la quantité :

$$\mathcal{I}_n(\theta) = V \left[\frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta; X_1, X_2, \dots, X_n) \right]$$

Remarque

On peut montrer que $E \left[\frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta; X_1, X_2, \dots, X_n) \right] = 0$; d'où :

$$\mathcal{I}_n(\theta) = E \left[\left(\frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta; X_1, X_2, \dots, X_n) \right)^2 \right]$$

On montre que l'on a également :

Remarque

$$\mathcal{I}_n(\theta) = -E \left[\frac{\partial^2}{\partial \theta^2} \ln \mathcal{L}(\theta; X_1, X_2, \dots, X_n) \right]$$

Cette écriture permet souvent de faciliter les calculs

Proposition

Inégalité de Fréchet-Damois-Cramer-Rao (**FDCR**) :

Pour tout estimateur T_n de θ , on a :

$$V(T_n) \geq \frac{\left[\frac{\partial}{\partial \theta} E(T_n) \right]^2}{I_n(\theta)}$$

Remarque

Dans le cas où T_n est sans biais, ce résultat s'écrit :

$$V(T_n) \geq \frac{1}{I_n(\theta)}$$

La quantité $\frac{1}{I_n(\theta)}$ est appelée borne de **Cramer-Rao** et elle représente donc, dans ce cas, un minorant de la variance de l'estimateur.

Définition

On appelle efficacité d'un estimateur T_n , le rapport :

$$Eff(T_n) = \frac{\left[\frac{\partial}{\partial \theta} E(T_n) \right]^2}{V(T_n) I_n(\theta)}$$

On a $0 \leq Eff(T_n) \leq 1$

T_n sera dit estimateur **efficace** ssi $Eff(T_n) = 1$

T_n sera dit **asymptotiquement efficace** ssi $\lim_{n \rightarrow \infty} Eff(T_n) = 1$

Remarques

- Si T_n est sans biais alors $\text{Eff}(T_n) = \frac{1}{V(T_n)\mathcal{I}_n(\theta)}$ et s'il est en plus efficace alors sa variance est égale à la borne de Cramer-Rao ; c'est donc forcément un ESBVM
- Il est possible qu'il n'existe pas d'estimateur efficace pour θ .
- Si la borne de Cramer-Rao est très grande, il sera impossible d'estimer correctement θ .
- Dans le cas où les variables sont indépendantes et de même loi, on voit facilement que $\mathcal{I}_n(\theta) = n\mathcal{I}_1(\theta)$ (à vérifier dans le cas, par exemple, de V.A. continues)

Exemple

Loi de Bernoulli