Chapitre 4 : Estimation
MA 361 : Probabilités

Pierre-Alain TOUPANCE
pierre-alain.toupance@esisar.grenoble-inp.fr

Grenoble INP - ESISAR

3$^{\text{ième}}$ année

6 novembre 2016

Introduction

Le but de l'estimation consiste à partir d'un échantillon de prévoir des informations sur la population totale.
On effectue deux types d'estimations :

- estimation ponctuelle
- estimation par intervalle de confiance

Soit Ω dont on considère un caractère :
On prend un échantillon de n individus et l'on obtient les données $x_1, x_2, ..., x_n$
On pose :

$$\widehat{m} = \frac{x_1 + x_2 + ... + x_n}{n}$$

$$\widehat{s}^2 = \frac{(x_1 - \widehat{m})^2 + ...(x_n - \widehat{m})^2}{n}$$

Soit $X_1, X_2, ..., X_n$ les variables aléatoires correspondant à ces données de moyenne m et de variance s^2, on pose :

$$\overline{X}_n = \frac{X_1 + ... + X_n}{n}$$

$$S_n^2 = \frac{(X_1 - \overline{X_n})^2 + ... + (X_n - \overline{X_n})^2}{n}$$

Estimateur

Soient g_n des fonctions de n variables réelles et soit Y_n la VAR $Y_n = g_n(X_1, ..., X_n)$

1. On dit que (g_n) est un estimateur de θ lorsque :

Estimateur

Soient g_n des fonctions de n variables réelles et soit Y_n la VAR $Y_n = g_n(X_1, ..., X_n)$

1. On dit que (g_n) est un estimateur de θ lorsque :

$$\lim_{n \to +\infty} E(Y_n) = \theta$$

   On dit aussi que Y_n est un estimateur de θ.
   La valeur de Y_n, calculée à partir de l'échantillon observé, càd $g_n(x_1, ...x_n)$ est appelée estimation de θ

Estimateur

Soient g_n des fonctions de n variables réelles et soit Y_n la VAR $Y_n = g_n(X_1, ..., X_n)$

1. On dit que (g_n) est un estimateur de θ lorsque :

$$\lim_{n \to +\infty} E(Y_n) = \theta$$

   On dit aussi que Y_n est un estimateur de θ.
   La valeur de Y_n, calculée à partir de l'échantillon observé, càd $g_n(x_1, ...x_n)$ est appelée estimation de θ

2. On dit que l'estimateur de g est sans biais lorsque

$$\forall n \in \mathbb{N}^*, \ E(Y_n) = \theta.$$

Estimateur

Soient g_n des fonctions de n variables réelles et soit Y_n la VAR $Y_n = g_n(X_1, ..., X_n)$

1. On dit que (g_n) est un estimateur de θ lorsque :

$$\lim_{n \to +\infty} E(Y_n) = \theta$$

   On dit aussi que Y_n est un estimateur de θ.
   La valeur de Y_n, calculée à partir de l'échantillon observé, càd $g_n(x_1, ...x_n)$ est appelée estimation de θ

2. On dit que l'estimateur de g est sans biais lorsque

$$\forall n \in \mathbb{N}^*, \ E(Y_n) = \theta.$$

3. On dit que l'estimateur g est convergent lorsque $\lim V(Y_n) = 0$

Propriété : estimation ponctuelle

Estimation ponctuelle

On a :

- $E[\overline{X}_n] = m$,
  $\widehat{m}$ est donc une estimation ponctuelle sans biais de $E[X]$
- $E(S_n^2) = \dfrac{n-1}{n} s^2$,
  ainsi $\dfrac{n}{n-1}\widehat{s}^2$ est une estimation sans biais de $V(X)$

Démonstration :

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

Démonstration :

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

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

Démonstration :

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

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

$$S_n^2 = \frac{1}{n}\left(\sum_{i=1}^{n}(X_i - m)^2 + \sum_{i=1}^{n}(\overline{X}_n - m)^2 - 2\sum_{i=1}^{n}(X_i - m)(\overline{X}_n - m)\right)$$

Démonstration :

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

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

$$S_n^2 = \frac{1}{n}\Big(\sum_{i=1}^{n}(X_i - m)^2 + \sum_{i=1}^{n}(\overline{X}_n - m)^2 - 2\sum_{i=1}^{n}(X_i - m)(\overline{X}_n - m)\Big)$$

$$S_n^2 = \frac{1}{n}\Big(\sum_{i=1}^{n}(X_i - m)^2 + n(\overline{X}_n - m)^2 - 2(\overline{X}_n - m)\sum_{i=1}^{n}(X_i - m)$$

$$\sum_{i=1}^{n}(X_i - m) =$$

$$\sum_{i=1}^{n}(X_i - m) = \sum_{i=1}^{n} X_i - nm$$

$$\sum_{i=1}^{n}(X_i - m) = \sum_{i=1}^{n} X_i - nm = n(\overline{X}_n - m)$$

$$\sum_{i=1}^{n}(X_i - m) = \sum_{i=1}^{n} X_i - nm = n(\overline{X}_n - m)$$

Ainsi

$$\sum_{i=1}^{n}(X_i - m) = \sum_{i=1}^{n} X_i - nm = n(\overline{X}_n - m)$$

Ainsi

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

On a :

$$E[\frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2] =$$

et

$$E[(\overline{X}_n - m)^2] =$$

On a :

$$E[\frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2] = \frac{1}{n}\sum_{i=1}^{n}V(X_i) = s^2$$

et

$$E[(\overline{X}_n - m)^2] =$$

On a :

$$E[\frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2] = \frac{1}{n}\sum_{i=1}^{n}V(X_i) = s^2$$

et

$$E[(\overline{X}_n - m)^2] = V(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n}V(X_i) = \frac{1}{n}s^2$$

Par conséquent :

$$E[S_n^2] = s^2 - \frac{s^2}{n}$$

D'où

$$E[S_n^2] = \frac{n-1}{n} s^2$$

On choisit comme estimation de θ la valeur qui maximise la probabilité de provoquer l'apparition de l'échantillon effectivement observé.

Maximum de vraissemblance

Soit $x_1,\ x_2, ..., x_n$ un échantillon, on note :

$$p = P(x_1, \theta)...P(x_n, \theta) = L(x_1, ..., x_n,\ theta)$$

Dans le cas continue,

$$p = f(x_1, \theta)...f(x_n, \theta)dx_1...dx_n \text{ où } f \text{ est la densité des VAR}$$

On cherche alors à résoudre :

$$L(x_1, ..., x_n, \widehat{\theta}) = \max_{\theta} L(x_1, ..., x_n, \theta)$$

En général, on cherche à rendre maximal $\ln(L)$

Propriété

L'estimateur du maximum de vraisemblance du paramètre θ est la solution de l'équation :

$$\frac{\partial Ln(X,\theta)}{\partial\theta} = 0$$

$$\frac{\partial^2 Ln(X,\theta^2)}{\partial\theta} < 0$$

Exemple

Soit une population, avec $X \rightsquigarrow \mathcal{N}(m, \sigma)$.
On cherche à estimer m et σ

$$L = \Pi_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i - m)^2/(2\sigma^2)}$$

Exemple

Soit une population, avec $X \rightsquigarrow \mathcal{N}(m, \sigma)$.
On cherche à estimer m et σ

$$L = \Pi_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i - m)^2/(2\sigma^2)}$$

$$L = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\sum (x_i - m)^2/(2\sigma^2)}$$

$$\ln L = -n \ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2$$

$$\ln L = -n \ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m)^2$$

$$\frac{\partial L}{\partial m} = 0 \Leftrightarrow \frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(x_i - m) = 0$$

$$\ln L = -n\ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2$$

$$\frac{\partial L}{\partial m} = 0 \Leftrightarrow \frac{1}{2\sigma^2}\sum_{i=1}^{n}2(x_i - m) = 0$$

$$\frac{\partial L}{\partial \sigma} = 0 \Leftrightarrow -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - m)^2 = 0$$

intervalle de confiance

Il est préférable de compléter l'estimation ponctuelle par une fourchette, c'est à dire, nous cherchons a et b tel que :
$P(a \leq \theta \leq b) = 1 - \alpha$ où θ est la valeur que l'on souhaite estimée (moyenne ou variance)
$1 - \alpha$ est appelé **niveau de confiance** de l'intervalle, on dit aussi que $[a; b]$ est un intervalle de confiance de θ **au risque** α
On cherchera a et b tel que $P(\theta \leq a) = P(\theta \geq b) = \dfrac{\alpha}{2}$

intervalle de confiance de la moyenne

1er cas : σ est connu Si les variables aléatoires $X_1, X_2, ..., X_n$ suivent des lois normales de paramètre (m, σ) ou si $n > 30$ alors $\overline{X}_n \rightsquigarrow \mathcal{N}(m, \frac{\sigma}{\sqrt{n}})$ ou $\overline{X}_n$ tend vers $Y \rightsquigarrow \mathcal{N}(m, \frac{\sigma}{\sqrt{n}})$.

intervalle de confiance de la moyenne

1er cas : σ est connu On pose

$$U = \frac{\overline{X}_n - m}{\sigma/\sqrt{n}} \rightsquigarrow \mathcal{N}(0,1)$$

Ainsi on cherche a tel que : $P(\overline{X}_n \leq a) = \alpha/2$
Or

$$P(\overline{X}_n \leq a) = \frac{\alpha}{2} \Leftrightarrow P(U \leq \frac{a-m}{\sigma/\sqrt{n}}) = \frac{\alpha}{2}$$

intervalle de confiance de la moyenne

1er cas : σ est connu On obtient ainsi $t_{\alpha/2}$ sur la table de la loi normale centrée réduite tel que :

$$\frac{a-m}{\sigma/\sqrt{n}} = t_{\alpha/2}$$

On a donc :

$$a = t_{\alpha/2}\frac{\sigma}{\sqrt{n}} + m$$

1er cas : σ est connu De la même façon on obtient, on cherche b tel que $P(\overline{X}_n \leq b) = 1 - \frac{\alpha}{2}$
Or

$$P(\overline{X}_n \leq b) = 1 - \frac{\alpha}{2} \Leftrightarrow P(U \leq \frac{b-m}{\sigma/\sqrt{n}}) = 1 - \frac{\alpha}{2}$$

1er cas : σ est connu On obtient ainsi $t_{\alpha/2}$ sur la table de la loi normale centrée réduite tel que :

$$\frac{b-m}{\sigma/\sqrt{n}} = t_{1-\alpha/2}$$

On a donc :

$$b = t_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} + m$$

On a $t_{\alpha/2} = -t_{1-\alpha/2}$

1er cas : σ est connu

L'intervalle de confiance de m au risque α est :

$$\boxed{I_\alpha = [\widehat{m} - \frac{\sigma}{\sqrt{n}} t_{1-\alpha/2} ; \widehat{m} + \frac{\sigma}{\sqrt{n}} t_{1-\alpha/2}]}$$

2ième cas : σ est inconnu

Dans ce cas, on pose $T = \dfrac{\overline{X}_n - m}{\sqrt{S_n^2/(n-1)}}$

On a aussi

$$T = \frac{\overline{X}_n - m}{\sqrt{S'^{\,2}_n/n}} = \frac{(\overline{X}_n - m)/(\sigma/\sqrt{n})}{\sqrt{\dfrac{(n-1)(S'^{\,2}_n/\sigma^2)}{n-1}}}$$

T suit une loi du Student à n-1 degré de liberté, on utilise la table numérique du Student pour obtenir a et b tel que $P(a \leq \overline{X}_n \leq b) = 1 - \alpha$

2ième cas : σ est inconnu

On a donc

$$P(\frac{a-m}{\sqrt{s'^2/n}} \leq \frac{\overline{X}_n - m}{\sqrt{S'^{\,2}_n/n}} \leq \frac{b-m}{\sqrt{s'^2/n}} = 1-\alpha$$

L'intervalle de confiance de m au risque α est :

$$\boxed{I_\alpha = \left[\widehat{m} - \frac{\widehat{s'}}{\sqrt{n}} t_{1-\alpha/2}; \widehat{m} + \frac{\widehat{s'}}{\sqrt{n}} t_{1-\alpha/2}\right]}$$

2ième cas : σ est inconnu : Exemple On réceptionne des pièces et on sait que la longueur des pièces à une distribution normale.
On prélève un échantillon de 20 pièces qui a une moyenne de 10cm et un écart type de 2.
Déterminons un intervalle de confiance de la moyenne de la longueur des pièces au risque de 5

2ième cas : σ est inconnu :

Exemple

Soit $X_1, X_2, ...X_{20}$ les variables aléatoires égales à la longueur des 20 pièces prélevées.

On pose

$$\overline{X}_{20} = \frac{\sum X_i}{20}$$

$$T = \frac{\overline{X}_{20} - m}{\sqrt{{S'_{20}}^2/20}}$$

T suit une loi du Student à 19 ddl.

On cherche a et b tel que

$$P(a \leq \overline{X}_{20} \leq b) = 0,95$$

on obtient

$$P(\frac{a-m}{\sqrt{{s'_{20}}^2/20}} \leq T \leq \frac{b-m}{\sqrt{{s'_{20}}^2/20}}) = 0,95$$

utilisant la table de la loi du Student, on obtient :

$$\frac{a-m}{\sqrt{{s'_{20}}^2/20}} = -2,093$$

$$\frac{b-m}{\sqrt{{s'_{20}}^2/20}} = 2,093$$

Ainsi l'intervalle de confiance est :

$$I = [10 - 2,093 \times \sqrt{\frac{4}{19}}; 10 + 2,093 \times \sqrt{\frac{4}{19}}]$$

$$I = [9,0397; 10,9603]$$

intervalle de confiance d'une proportion

Soit K_n la variable aléatoire qui compte le nombre d'individu ayant la propriété P dans un échantillon de taille n non exhaustif.
On a $K_n \rightsquigarrow \mathcal{B}(n, f)$

Si l'approximation par une loi normale est justifiée $F_n = \frac{K_n}{n}$ est proche d'une loi normale $F_n \rightsquigarrow \mathcal{N}(f, \sqrt{\frac{f(1-f)}{n}})$

On pose $U = \frac{F - f}{\sqrt{\frac{f(1-f)}{n}}} \rightsquigarrow \mathcal{N}(0, 1)$.

On obtient ainsi l'intervalle de confiance au risque α

$$\boxed{I_\alpha = \left[\widehat{f} - \sqrt{\frac{\widehat{f}(1-\widehat{f})}{n}}t_{1-\alpha/2} ; \widehat{f} + \sqrt{\frac{\widehat{f}(1-\widehat{f})}{n}}t_{1-\alpha/2}\right]}$$

où $\widehat{f}$ est la fréquence de l'échantillon

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

On pose :

$$Y = (n-1)\frac{S_n'^2}{\sigma^2}$$

Y suit une loi du Khi deux à (n-1) degrés de liberté.

On détermine a et b dans la table du Khi deux tel que :

$$P(a \leq (n-1)\frac{S_n'^2}{\sigma^2} \leq b) = 1 - \alpha$$

On obtient ainsi l'intervalle de confiance de la variance :

$$\boxed{I_\alpha = \left[\frac{(n-1)\widehat{s'}^2}{b}; \frac{(n-1)\widehat{s'}^2}{a}\right]}$$

Exemple : On a un échantillon de 16 chiffres d'affaires tel que $s'_{16} = 72,53$.
Déterminer l'intervalle de confiance de la variance de niveau 0,95.