5. Théorie statistique de l'estimation

Plan du chapitre 5

Introduction

I. Estimateurs
II. Estimation
III. Intervalle de confiance
IV. Estimateur de l'écart-type

Introduction

Problématique générale :

On renverse la perspective : comment à partir d'information sur les échantillons, obtenir des informations sur la variable aléatoire X qui caractérise P ?

Exemple 1 : loi normale ou quelconque (deux inconnues)

Exemple 2 : loi de Bernouilli (un seul paramètre inconnu)

Introduction

Le problème principal de l'inférence (passage d'un échantillon à l'ensemble de la population) consiste en l'estimation des paramètres de la loi suivie par X.

On utilise des « statistiques » qui sont des variables aléatoires fonctions de X_1, X_2, X_3,…, X_n. La réalisation d'une statistique est un nombre.

I. Estimateurs

A) Définition d'un estimateur

On appelle estimateur d'un paramètre Θ (thêta) inconnu de L(X), une fonction d'un n-échantillon issu de X servant à estimer ce paramètre.

On note cet estimateur $g(x_1, x_2, \dots, x_n)$: c'est une variable aléatoire.

I. Estimateurs

Exemple 1 :

Le QI des étudiants suit une loi normale $\mathcal{N}$(μ,13) avec μ inconnu : c'est le paramètre Θ que l'on doit estimer.

Exemple 2 :

2ème tour des élections de 1988. Chaque individu est caractérisé par une Bernouilli de paramètre p la probabilité de succès de Mitterand.

I. Estimateurs

B) Propriété d'un estimateur

i) Une estimateur de Θ est dit non biaisé si son espérance mathématique est égale au paramètre Θ à estimer.

ii) Un estimateur de Θ est dit convergent su son espérance mathématique tend vers Θ et si sa variance tend vers 0 quand n tend vers l'infini.

I. Estimateurs

$$\lim_{n \to +\infty} E[g(x_1, x_2, \dots, x_n)] = \Theta$$
$$\lim_{n \to +\infty} Var[g(x_1, x_2, \dots, x_n)] = 0$$

iii) Un estimateur est dit efficace (NBVM) quand il est sans biais de variance minimum.

I. Estimateurs

Exemples :

- $\bar{X}$ estimateur de μ
- F_n estimateur de *p*

II. Estimation

A) Définition

On appelle **estimation** d'un paramètre Θ inconnu de $\mathcal{L}$(X), la réalisation de l'estimateur correspondant. C'est un nombre.

II. Estimation

B) Estimation de μ

Exemple : QI des étudiants

C) Estimation de p

Exemple : vote Mitterand en 1988

III. Intervalle d'estimation

A) Définition

Les méthodes d'estimation ponctuelles associent à un échantillon une valeur unique du paramètre inconnu : celle de l'estimation de l'échantillon considéré.

On a autant d'estimations que d'échantillons.

Les méthodes d'estimation par intervalle associent au résultat d'un échantillon un intervalle I_n qui doit contenir la vraie valeur inconnue du paramètre avec une probabilité fixée à l'avance, notée α.

III. Intervalle d'estimation

Définition

On appelle intervalle de confiance à α pour un paramètre Θ un intervalle I_n construit sur le résultat d'un *n* échantillon qui recouvre la vraie valeur du paramètre Θ avec une probabilité α donnée.

III. Intervalle d'estimation

Remarques

- Θ est inconnu mais pas aléatoire
- α est indépendant de Θ : c'est une probabilité que l'on se fixe
- I_n est un intervalle aléatoire puisqu'il dépend de l'échantillon

On fixe souvent $\alpha = 95\ \%$: on dit alors que l'on a une probabilité de 95 % que Θ soit dans I_n.

III. Intervalle d'estimation

B) Estimation par intervalle de l'espérance μ d'une loi normale de variance connue

Soit X une variable aléatoire définie sur un individu d'une population P.

$\mathcal{L}$(X)=$\mathcal{N}$(μ,σ) avec μ inconnu et σ connu

On connait la loi et l'écart-type mais pas l'espérance : μ est le paramètre à estimer.

On va utiliser un *n* échantillon avec remise.

III. Intervalle d'estimation

A priori

Soient X_1, X_2, X_3,…, X_n n variables aléatoires indépendantes telles que :

$\mathcal{L}(X_i)=\mathcal{N}(\mu,\sigma)$, $\forall\ i\ \leq n$

On définit $\bar{X}$ estimateur de μ tel que :

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \text{ ; } \mathcal{L}(\bar{X})=\mathcal{N}(\mu\ ;\ \frac{\sigma}{\sqrt{n}})$$

Avec μ inconnu et $\frac{\sigma}{\sqrt{n}}$ connu

III. Intervalle d'estimation

Soit α une probabilité, on peut trouver un intervalle centré sur μ en utilisant la table 2 du formulaire. La table donne t, le nombre d'écart-type dont on doit s'éloigner de μ pour être dans cet intervalle avec une probabilité α.

$$\alpha = P(\mu - a < \bar{X} < \mu + a)$$
$$= P(\mu - t.\frac{\sigma}{\sqrt{n}} \leq X \leq \mu + t.\frac{\sigma}{\sqrt{n}})$$

III. Intervalle d'estimation

$I_n = \left[\mu - t.\frac{\sigma}{\sqrt{n}}; \mu + t.\frac{\sigma}{\sqrt{n}}\right]$: intervalle théorique contenant la valeur de μ avec une probabilité α.

Exemple : si $\alpha = 0{,}95 = 95\ \%$, alors $t = 1{,}96$ et $I_n = \left[\mu - 1{,}96.\frac{\sigma}{\sqrt{n}}; \mu + 1{,}96.\frac{\sigma}{\sqrt{n}}\right]$.

Seul μ est inconnu dans cet intervalle.

III. Intervalle d'estimation

A posteriori

Soient x_1, x_2, x_3,…, x_n les n valeurs trouvées dans un n échantillon.

On a :

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, une estimation de μ.

$I_n = \left[\bar{x} - t.\frac{\sigma}{\sqrt{n}}; \bar{x} + t.\frac{\sigma}{\sqrt{n}}\right]$: intervalle observé contenant la valeur de μ avec un e probabilité α.

III. Intervalle d'estimation

Exemple 1 : le QI des étudiants

On a $\mathcal{L}$(X)=$\mathcal{N}$(μ,$\frac{13}{\sqrt{100}}$) et on cherche à estimer μ.

1er cas : échantillon de 100 personnes

2ème cas : échantillon de 1000 personnes

III. Intervalle d'estimation

Remarques :

- Plus l'échantillon est grand, plus l'intervalle de confiance est petit.
- Plus α augmente, plus t augmente et plus l'intervalle d'estimation est grand.

III. Intervalle d'estimation

C) Estimation par intervalle de l'espérance μ d'une loi quelconque

- μ est inconnue
- $\mathcal{L}$(X) est inconnue ($\mathcal{L}$(X) = $\mathcal{L}$(μ,σ))
- σ est connu

On tire un n échantillon et on cherche à utiliser le TCL

Il faut donc « n assez grand » soit $n \geq 30$, alors $\mathcal{L}(\bar{X})$ = $\mathcal{L}(\mu, \frac{\sigma}{\sqrt{n}})$ et on se ramène au cas précédent.

III. Intervalle d'estimation

Exemple 2 :

On cherche l'espérance de la charge maximale supportée par un câble fabriqué dans une usine.

X= charge maximale d'un câble

$\mathcal{L}(X) = \mathcal{L}(\mu;0,73)$

- Intervalle à 95 %
- Intervalle à 90 %

III. Intervalle d'estimation

D) Estimation par intervalle de l'espérance μ d'une loi quelconque de variance inconnue

- μ est inconnue
- $\mathcal{L}(X)$ est inconnue ($\mathcal{L}(X) = \mathcal{L}(\mu,\sigma)$)
- σ est inconnu

Il faut estimer μ et σ simultanément, c'est-à-dire trouver le meilleur estimateur (NBVM) et déterminer sa loi de probabilité (voir paragraphe suivant).

III. Intervalle d'estimation

E) Estimation par intervalle d'un pourcentage p

Soit P une population dans laquelle un pourcentage p de personnes présente une caractéristique A. p est inconnue.

X = indicateur de succès

$\mathcal{L}$(X) = $\mathcal{B}$(1,p)

p est le paramètre à estimer

On tire un échantillon de taille n

III. Intervalle d'estimation

A priori

Soient X_1, X_2, X_3,…, X_n un échantillon de variables aléatoires indépendantes issues de X telles $\mathcal{L}$(X) = $\mathcal{B}$(1,p)

Soit $F_n = \frac{\sum_{i=1}^{n} X_i}{n}$ un estimateur de p (proportion théorique d'individus possédant le caractère A)

Peut-on faire une approximation de la loi de F_n par une loi normale ?

III. Intervalle d'estimation

On ne connaît pas *p* et *q* mais on peut ponctuellement les estimer par f_n et $1 - f_n$

On admet que

$\mathcal{L}(F_n) = \mathcal{PB}(n,p) \cong \mathcal{N}(p; \sqrt{\frac{pq}{n}})$ dès que $n.f_n \geq 10$ et $n.(1 - f_n) \geq 10$

On cherche un intervalle centré sur p de probabilité α tel que :

$$P\left[p - t.\sqrt{\frac{pq}{n}}; p + t.\sqrt{\frac{pq}{n}}\right] = \alpha$$

III. Intervalle d'estimation

On a alors :

$I_n = \left[p - t.\sqrt{\frac{pq}{n}}; p + t.\sqrt{\frac{pq}{n}}\right]$ l'intervalle théorique contenant p avec une probabilité α.

A posteriori

On observe un n échantillon x_1, x_2, x_3,…, x_n (les x_i valent 0 ou 1).

On calcule $f_n = \frac{\sum_{i=1}^{n} x_i}{n}$ estimation de p.

III. Intervalle d'estimation

Pour l'estimation par intervalle, on remplace p par la valeur de f_n :

$I_n = \left[f_n - t.\sqrt{\frac{f_n(1-f_n)}{n}} ; f_n + t.\sqrt{\frac{f_n(1-f_n)}{n}} \right]$ intervalle observé contenant la valeur p avec une probabilité α.

On estime avec une probabilité α que le pourcentage de personne ayant le caractère A se trouve dans cet intervalle.

Exemple 2 : les intentions de vote en 1988

III. Intervalle d'estimation

Exercice :

On cherche à estimer la proportion de gauchers dans une population.

On tire un échantillon sans remise de 1000 personnes et on trouve 12 % de gauchers.

IV. Estimateurs de l'écart-type

A) Préliminaires

Soit X une variable aléatoire continue caractérisant un individu d'une population P. Cette variable possède une espérance μ et un écart-type σ.

Chercher un estimateur de σ consiste à trouver un variable aléatoire fonction d'un échantillon issu de X (X_1, X_2, X_3,…, X_n) ayant σ comme espérance mathématique et un écart-type qui tend vers 0 (NBVM)

IV. Estimateurs de l'écart-type

On peut penser que $S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$ est l'estimateur de σ^2 et que $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$ en est l'estimation.

Pour le démontrer il faudrait calculer l'espérance et la variance de cette variable aléatoire.

Deux difficultés :

- Carrés de variables aléatoires
- $\bar{X}$ n'est pas indépendante des X_i.

IV. Estimateurs de l'écart-type

B) Estimateur de σ quand μ est connue

Définition 1 :

Soit X une variable aléatoire de loi normale centrée réduite : $\mathcal{L}(X) = \mathcal{N}(0;1)$.

Soit un n échantillon issu de X : X_1, X_2, X_3,…, X_n avec $\mathcal{L}(X_i) = \mathcal{N}(0;1)$.

Alors la variable aléatoire $Z_n = \sum_{i=1}^{n} {X_i}^2$ suit une loi du Chi-deux à n degrés de liberté (notée ${\chi_n}^2$)

L'espérance mathématique de cette loi est n et sa variance vaut 2n.

IV. Estimateurs de l'écart-type

B) Estimateur de σ quand μ est connue

Définition 1 :

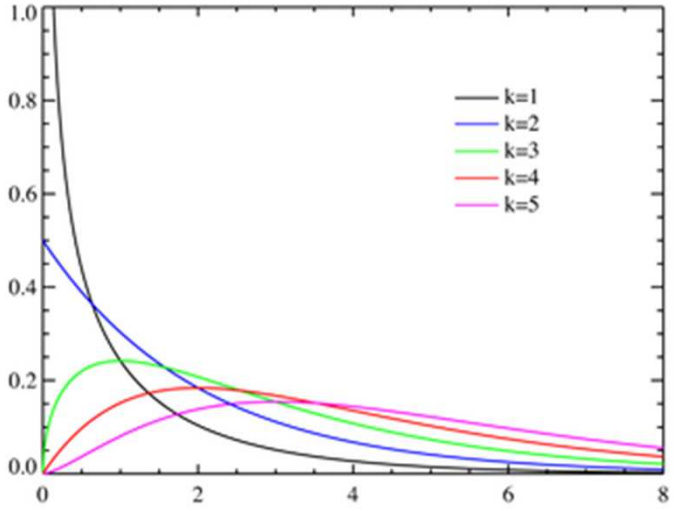
Soit X une variable aléatoire de loi normale centrée réduite : $\mathcal{L}(X) = \mathcal{N}(0;1)$.

Soit un n échantillon issu de X : X_1, X_2, X_3,…, X_n avec $\mathcal{L}(X_i) = \mathcal{N}(0;1)$.

Alors la variable aléatoire $Z_n = \sum_{i=1}^{n} {X_i}^2$ suit une loi du Chi-deux à n degrés de liberté (notée ${\chi_n}^2$)

L'espérance mathématique de cette loi est n et sa variance vaut 2n.

IV. Estimateurs de l'écart-type



IV. Estimateurs de l'écart-type

Remarque :

Quand n tend vers $+\infty$, en vertu du théorème central limite, la loi du $\chi_n{}^2$ tend vers une loi normale $\mathcal{N}(n;\sqrt{2n})$.

Conséquences :

Si dans une population P chaque individu est caractérisé par une variable aléatoire X de loi $\mathcal{N}(\mu,\sigma)$.

On pose $X^* = \frac{X-\mu}{\sigma}$ avec $\mathcal{L}(X^*) = \mathcal{N}(0;1)$.

Soit un n échantillon issu de X : X_1, X_2, X_3,…, X_n avec $\mathcal{L}(X_i) = \mathcal{N}(\mu;\sigma)$

IV. Estimateurs de l'écart-type

Alors la variable aléatoire $Z_n = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2$ suit une loi du chi-deux à n degrés de liberté.

Z_n est composée d'une somme de carrés de lois normales centrées réduites. On a

$$Z_n = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2$$