

Cours accéléré Statistiques

Partie 1: Bases de la statistique inférentielle

Master 2 Mathématiques et Applications

Christine Keribin

¹Laboratoire de Mathématiques d'Orsay
Université Paris-Saclay

2024-2025

université
PARIS-SACLAY

FACULTÉ
DES SCIENCES
D'ORSAY



Cours accéléré
Statistiques
Partie 1: Bases de
la statistique
inférentielle

Christine Keribin

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Introduction

NP

Test de Wald

Exemples

p-value

Intervalle de
confiance

Introduction

Construction

Région de confiance 1/116

Contenu

- ▶ Concepts fondamentaux en statistique : estimateur, intervalle de confiance, test ; estimateur du maximum de vraisemblance
- ▶ Modèle linéaire et applications : régression multiple, anova
- ▶ Principes de l'analyse bayésienne

https:

[//www.imo.universite-paris-saclay.fr/~keribin/EnseignementRaN.htm](https://www.imo.universite-paris-saclay.fr/~keribin/EnseignementRaN.htm)

Cours accéléré
Statistiques

Partie 1: Bases de
la statistique
inférentielle

Christine Keribin

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Introduction

NP

Test de Wald

Exemples

p-value

Intervalle de
confiance

Introduction

Construction

Région de confiance 2/116

Bibliographie



V. Rivoirard et G. Stoltz.

Statistique en action.

Vuibert, Paris, 2009.



M. Lejeune.

Statistique : la théorie et ses applications

Springer, 2010.



J. Pagès.

Statistique générales pour utilisateurs.

Presses Universitaires de Rennes, Rennes, 2005.



P.-A. Cornillon et autres.

Statistiques avec R.

Presses Universitaires de Rennes, Rennes, 2008.



C. Keribin.

Modélisation statistique.

poly.

Cours accéléré

Statistiques

Partie 1: Bases de

la statistique

inférentielle

Christine Keribin

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Introduction

NP

Test de Wald

Exemples

p-value

Intervalle de
confiance

Introduction

Construction

Région de confiance 3/116

Introduction

- ▶ On dispose d'un échantillon de 6 relevés du temps de trajet (en min) domicile/bureau d'un employé

$x = (15, 17, 15, 18, 16, 15)$

$$\bar{x} = \frac{1}{6} \sum_i x_i = 16$$

- ▶ Quelle est la durée moyenne d'un trajet sur l'année ?
- ▶ Peut-on affirmer avec peu de risque que la durée moyenne d'un trajet est supérieure à 15 min ?
- ▶ La durée d'un trajet a-t-elle augmenté d'une année sur l'autre ?
- ▶ Y a-t-il des facteurs qui influencent la durée de trajet ?
- ▶ Quelle sera la durée de son trajet demain ?

Introduction

- ▶ **Population** : ensemble d'objets "équivalents" (**individus, unités statistiques**), sur lesquels on observe des caractéristiques (**variables** qualitatives ou quantitatives)
 - ▶ finie : recensement
 - ▶ infinie : sondage
- ▶ Étude de la **variabilité**
- ▶ Statistique **exploratoire** / statistique **inférentielle**

Introduction

Modèle statistique

Estimation

Estimateur
Propriétés
Lois
Cas gaussien
Cochran
Approximation gaussienne

Vraisemblance

Information de Fisher
Efficacité

EMV

Tests

Introduction
NP
Test de Wald
Exemples
p-value

Intervalle de confiance

Introduction
Construction
Région de confiance

- ▶ **Probabilité** : étudier les propriétés d'une loi connue
- ▶ **Statistique** : à partir d'un ensemble d'**observations** d'une loi inconnue, **inférer/apprendre** des propriétés de cette loi pour répondre à une question

↔ résoudre un problème inverse

- ▶ Modéliser
- ▶ Estimer
- ▶ Utiliser

Introduction

Modèle statistique

Estimation

Estimateur
Propriétés
Lois
Cas gaussien
Cochran
Approximation gaussienne

Vraisemblance

Information de Fisher
Efficacité

EMV

Tests

Introduction
NP
Test de Wald
Exemples
p-value

Intervalle de confiance

Introduction
Construction
Région de confiance

Quelques objectifs de la statistique inférentielle

- ▶ **estimation** : valeur d'un paramètre d'intérêt, **intervalle de confiance**
- ▶ **test** : comparaison à une situation de référence, de deux échantillons, ...
- ▶ **prédiction** pour une nouvelle unité non encore observée
- ▶ **classification** dans un groupe

Problématiques :

- ▶ construction, comparaison, choix des procédures
- ▶ fiabilité (**risque**) de l'information obtenue ?

Introduction

Modèle statistique

Estimation

Estimateur
Propriétés
Lois
Cas gaussien
Cochran
Approximation gaussienne

Vraisemblance

Information de Fisher
Efficacité

EMV

Tests

Introduction
NP
Test de Wald
Exemples
p-value

Intervalle de confiance

Introduction
Construction
Région de confiance

Sommaire

Modèle statistique

Estimation

Vraisemblance

EMV

Tests

Intervalle de confiance

Cours accéléré
Statistiques
Partie 1: Bases de
la statistique
inférentielle

Christine Keribin

Introduction

Modèle statistique

Estimation

Estimateur
Propriétés
Lois
Cas gaussien
Cochran
Approximation gaussienne

Vraisemblance

Information de Fisher
Efficacité

EMV

Tests

Introduction
NP
Test de Wald
Exemples
p-value

Intervalle de
confiance

Introduction
Construction
Région de confiance

Modélisation statistique

Modéliser l'expérience, c'est proposer une loi théorique pour l'échantillon $X = (X_1, \dots, X_n)$.

- ▶ **Modèle** : espace probabilisé $\mathcal{M} = (\mathcal{X}^n, \mathcal{A}^n, (\mathbb{P}_\theta^n)_{\theta \in \Theta})$
 - ▶ \mathcal{X}^n espace des réalisations, \mathcal{A}^n tribu des événements
 - ▶ $(\mathbb{P}_\theta^n)_{\theta \in \Theta}$ famille de lois de probabilité

Quand il existe $d \in \mathbb{N}^*$ tel que $\Theta \subset \mathbb{R}^d$, le modèle est dit *paramétrique*

- ▶ Une **observation** est une variable aléatoire X à valeur dans \mathcal{X}^n et dont la loi appartient à $(\mathbb{P}_\theta^n)_{\theta \in \Theta}$
- ▶ Un **n -échantillon i.i.d.** est une observation $X = (X_1, \dots, X_n)$ de n variables aléatoires de même loi η_θ et indépendantes. Alors, $\mathbb{P}_\theta^n = (\eta_\theta)^{\otimes n}$
- ▶ Les **données** sont les réalisations (valeurs) x_1, \dots, x_n prises par l'échantillon X_1, \dots, X_n

Exemples de modèles (simples)

- ▶ Etude de la moyenne (espérance) d'un temps de trajet :
 $\mathcal{M} = (\mathbb{R}^n, \mathcal{A}^n, \mathbb{P}_\theta^{\otimes n}, \theta \in \Theta)$

$$X_i \sim_{i.i.d.} \mathbb{P}(\mu, \sigma^2)$$

- ▶ Comparaison du rendement de maïs sous deux conditions de culture

$$X \sim \mathcal{N}^{\otimes n}(\mu_x, \sigma_x^2) \text{ et } Y \sim \mathcal{N}^{\otimes m}(\mu_y, \sigma_y^2) \text{ indépendants}$$

- ▶ Estimation d'une proportion par sondage
 $\mathcal{M} = (\{0, 1\}^n, \mathcal{A}^n, \mathcal{B}_\pi^{\otimes n}, \pi \in [0; 1])$

$$X_i \sim_{i.i.d.} \mathcal{B}(1, \pi)$$

Régression linéaire (simple)

$$Y_i|X_i=x_i = \alpha + \beta x_i + \varepsilon_i \text{ où } \varepsilon_i|X_i=x_i \sim i.i.d. \mathcal{L}(0, \sigma^2 I_n).$$

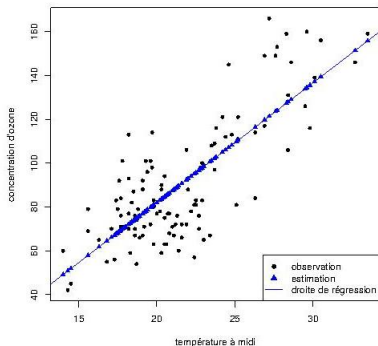


Figure – source des données : Cornillon et Matzner-Løber, 2007

Les étapes de la démarche statistique

A partir des données d'un n -échantillon, déduire -ou inférer- certaines propriétés du modèle inconnu

1. Acquérir et **préparer** les données
2. Définir un **modèle** adapté à la situation observée.
3. **Estimer** les paramètres du modèle grâce aux observations.
4. Vérifier l'**adéquation** de l'estimation aux observations.
5. **Utiliser** le modèle à des fins de décision ou de prédiction.

Sommaire

Modèle statistique

Estimation

Vraisemblance

EMV

Tests

Intervalle de confiance

Cours accéléré
Statistiques
Partie 1: Bases de
la statistique
inférentielle

Christine Keribin

Introduction

Modèle statistique

Estimation

- Estimateur
- Propriétés
- Lois
- Cas gaussien
- Cochran
- Approximation gaussienne

Vraisemblance

- Information de Fisher
- Efficacité

EMV

Tests

- Introduction
- NP
- Test de Wald
- Exemples
- p-value

Intervalle de
confiance

- Introduction
- Construction
- Région de confiance

Un objet mathématique : l'estimateur

Soit $\theta \in \Theta \in \mathbb{R}$ le paramètre d'une loi \mathbb{P}_θ et soit $X = (X_1, \dots, X_n)$ un n -échantillon issu de cette loi

Définition

Une *statistique* est une *variable aléatoire* T_n , fonction mesurable de l'échantillon et calculable à partir de l'échantillon

$$T_n = t(X_1, \dots, X_n).$$

Un *estimateur* est une statistique utilisée pour estimer un paramètre ou une quantité d'intérêt $\nu(\theta)$

- ▶ Notation $T_n = \hat{\nu}_n$ ou $T_n = \hat{\nu}$.
- ▶ Exemples : estimateur empirique de l'espérance, de la variance

Performance en moyenne

Soit $\hat{\nu}_n$ un estimateur de $\nu(\theta)$, fonction du paramètre d'une loi \mathbb{P}_θ :

- ▶ On appelle **biais** de $\hat{\nu}_n$ pour $\nu(\theta)$ la valeur

$$b_\theta(\hat{\nu}_n) = \mathbb{E}_\theta(\hat{\nu}_n) - \nu(\theta)$$

Si $b_\theta(\hat{\nu}_n) = 0$ pour tout $\theta \in \Theta$, T_n est **sans biais** pour $\nu(\theta)$

- ▶ On appelle **variance** de $\hat{\nu}_n$ la valeur

$$\text{Var}_\theta(\hat{\nu}_n) = \mathbb{E}_\theta\left(\left(\hat{\nu}_n - \mathbb{E}_\theta(\hat{\nu}_n)\right)^2\right)$$

- ▶ On appelle **risque quadratique** de $\hat{\nu}_n$ la valeur

$$R_\theta(\hat{\nu}_n) = \mathbb{E}_\theta\left(\left(\hat{\nu}_n - \nu(\theta)\right)^2\right)$$

Décomposition du risque quadratique

$$R_{\theta}(\hat{\nu}_n) = \text{Var}_{\theta}(\hat{\nu}_n) + (b_{\theta}(\hat{\nu}_n))^2$$

Définition

Un estimateur δ_1 de $\nu(\theta)$ **domine** l'estimateur δ_2 si, pour tout $\theta \in \Theta$,

$$R_{\theta}(\delta_1) \leq R_{\theta}(\delta_2),$$

cette inégalité étant stricte pour au moins une valeur de θ .

Un estimateur est **admissible** s'il n'existe aucun estimateur le dominant.

- Soit $\theta_0 \in \Theta$. L'estimateur constant $\hat{\nu}_n = \theta_0$ est-il admissible ?

Il n'existe en général pas d'estimateur dominant tous les autres \leftrightarrow Recherche d'estimateurs **UVMB**

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Introduction

NP

Test de Wald

Exemples

p-value

Intervalle de
confiance

Introduction

Construction

Région de confiance

Performance asymptotique

Soit $\hat{\nu}_n$ un estimateur de $\nu(\theta)$, défini à partir d'une observation de loi \mathbb{P}_θ :

- ▶ $\hat{\nu}_n$ est **asymptotiquement sans biais** pour $\nu(\theta)$:

$$\forall \theta \in \Theta, \lim_{n \rightarrow \infty} b_\theta(\hat{\nu}_n) = 0$$

- ▶ $\hat{\nu}_n$ est **consistant** : $\hat{\nu}_n$ tend en probabilité vers $\nu(\theta)$ quand $n \rightarrow \infty$:

$$\forall \theta \in \Theta, \forall \epsilon, \lim_{n \rightarrow \infty} \mathbb{P}_\theta(|\hat{\nu}_n - \nu(\theta)| > \epsilon) = 0$$

- ▶ $\hat{\nu}$ est **fortement consistant** ssi

$$\forall \theta \in \Theta, \mathbb{P}_\theta(\lim_{n \rightarrow \infty} \hat{\nu}_n = \nu(\theta)) = 1$$

- ▶ $\hat{\nu}$ est **consistant en moyenne quadratique** :

$$\forall \theta \in \Theta, \lim_{n \rightarrow \infty} R_\theta(\hat{\nu}_n) = 0$$

Aller plus loin : s'intéresser à la loi d'un estimateur

- ▶ Un constructeur automobile indique une consommation de $c_0 = 6.32\ell/100km$ pour les véhicules d'un type donné, dans des conditions précises de roulage, avec un écart-type de $\sigma_0 = 0.21\ell/100km$
- ▶ Un organisme indépendant prend 30 véhicules au hasard, et les soumet aux conditions de roulage nominales. Il observe $\bar{x} = 6.43\ell/100km > c_0$, $\hat{\sigma} = 0.25\ell/100km$.
 - ▶ Est-ce dû à la variabilité naturelle de l'expérience ?
 - ▶ Ou le constructeur a-t-il sous-estimé la consommation de ses véhicules ?
- ▶ Déterminer si le fait d'observer une moyenne plus grande que 6.43 est d'une probabilité forte ou pas sous les indications du constructeur.
 - ▶ accéder à $\mathbb{P}(\bar{X} \geq 6.43)$,
 - ▶ loi de \bar{X}

Cas Gaussien

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Introduction

NP

Test de Wald

Exemples

p-value

Intervalle de confiance

Introduction

Construction

Région de confiance

Proposition

L'estimateur empirique \bar{X} de l'espérance d'une loi gaussienne $\mathcal{N}(\mu, \sigma^2)$, calculé à partir d'un échantillon i.i.d. de cette loi, est gaussien :

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Introduction

NP

Test de Wald

Exemples

p-value

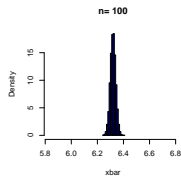
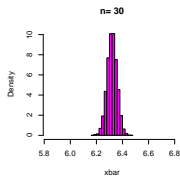
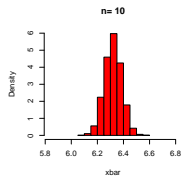
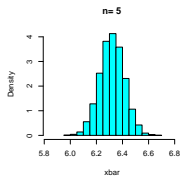
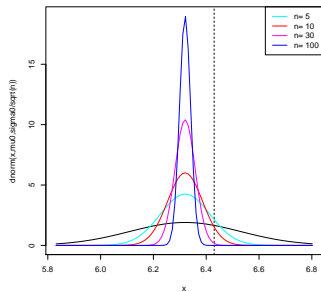
Intervalle de
confiance

Introduction

Construction

Région de confiance

Loi de \bar{X}



Introduction

Modèle statistique

Estimation

Estimateur
Propriétés
Lois

Cas gaussien

Cochran
Approximation gaussienne

Vraisemblance

Information de Fisher
Efficacité

EMV

Tests

Introduction
NP
Test de Wald
Exemples
p-value

Intervalle de confiance

Introduction
Construction
Région de confiance

Applications

- ▶ Si on considère que la mesure de pollution d'un véhicule de marque donnée suit $X_1 \sim \mathcal{N}(c_0 = 6.32, \sigma_0^2 = 0.21^2)$, alors

$$\begin{aligned}\mathbb{P}(\bar{X} \geq 6.43) &= \mathbb{P}\left(\frac{\bar{X} - c_0}{\sigma_0/\sqrt{n}} \geq \frac{6.43 - c_0}{\sigma_0/\sqrt{n}}\right) \\ &= 1 - F_{\mathcal{N}}\left(\frac{6.43 - c_0}{\sigma_0/\sqrt{n}}\right) \simeq 0.002\end{aligned}$$

- ▶ On peut aussi se demander quelle valeur q de consommation moyenne sur 30 véhicules est dépassée avec une probabilité donnée (par ex, $\alpha = 5\%$)

$$\mathbb{P}(\bar{X} \geq q) = 1 - F_{\mathcal{N}}\left(\frac{q - c_0}{\sigma_0/\sqrt{n}}\right) = 0.05$$

Soit

$$q = c_0 + \underbrace{F_{\mathcal{N}}^{-1}(0.95)}_{\text{quantile d'ordre 95\% de } \mathcal{N}(0,1)} \frac{\sigma_0}{\sqrt{n}} = 6.38$$

Loi de l'estimateur de la variance à espérance connue

Proposition

Soit $V_n^* = \frac{1}{n} \sum_i (X_i - \mu)^2$, l'estimateur empirique de la variance d'un échantillon i.i.d. X de loi $\mathcal{N}(\mu, \sigma^2)$, μ connue :

$$n \frac{V_n^*}{\sigma^2} \sim \chi^2(n)$$

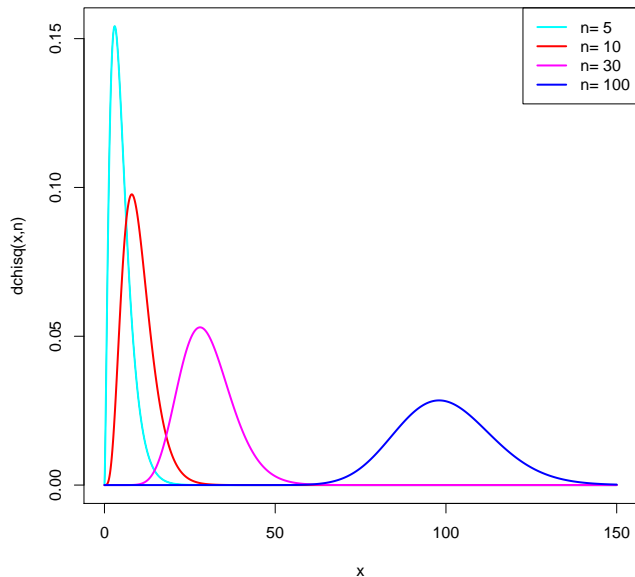
Définition (loi du Khi-deux)

Soit Z un vecteur gaussien **centré réduit** de dimension n . La loi de la somme du carré de ses composantes est la loi du **Khi-deux** (centré) à n degrés de liberté

$$K_n = \sum_i Z_i^2 \sim \chi^2(n); \quad \psi_{K_n}(t) = \mathbb{E}(e^{tK_n}) = \frac{1}{(1 - 2t)^{n/2}}$$

$$\mathbb{E}(K_n) = n; \quad \text{var}(K_n) = 2n$$

Loi du $\chi^2(n)$



Christine Keribin

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Introduction

NP

Test de Wald

Exemples

p-value

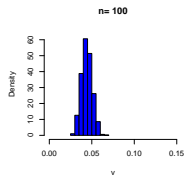
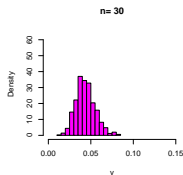
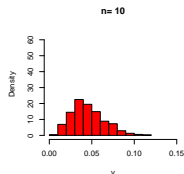
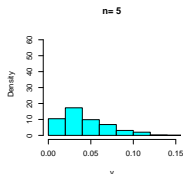
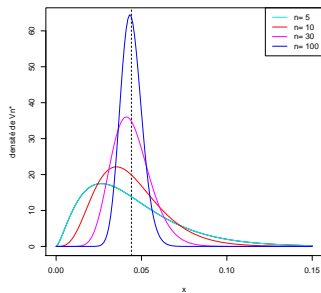
Intervalle de
confiance

Introduction

Construction

Région de confiance

Loi de V_n^*



Introduction

Modèle statistique

Estimation

Estimateur
Propriétés
Lois

Cas gaussien

Cochran
Approximation gaussienne

Vraisemblance

Information de Fisher
Efficacité

EMV

Tests

Introduction
NP
Test de Wald
Exemples
p-value

Intervalle de confiance

Introduction
Construction
Région de confiance

Loi de la variance empirique V_n

Proposition

La loi de l'estimateur de la variance empirique

$V_n^2 = \sum_i (X_i - \bar{X})^2 / n$ d'un n -échantillon i.i.d. gaussien X
est telle que

$$n \frac{V_n^2}{\sigma^2} \sim \chi^2(n-1)$$

De plus, \bar{X} et V_n^2 sont **indépendants**.

Application :

$$\mathbb{P}(\hat{\sigma}_n > 0.25) = 1 - F_{\chi^2}((n-1)(0.25/\sigma_0)^2, n-1) \simeq 0.067$$

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Introduction

NP

Test de Wald

Exemples

p-value

Intervalle de
confiance

Introduction

Construction

Région de confiance

Théorème (Cochran)

Si $Y \sim \mathcal{N}_n(\mu, \sigma^2 I_n)$, et si $E_1 \oplus \dots \oplus E_r = \mathbb{R}^n$ est une décomposition de \mathbb{R}^n en r sous-espaces orthogonaux, alors les **projections orthogonales** $\Pi_1(Y), \dots, \Pi_r(Y)$ sur ces sous-espaces sont des **vecteurs gaussiens indépendants** tels que, pour tout $j = 1, \dots, r$

$$\|\Pi_j(Y)\|^2 \sim \sigma^2 \chi^2(d_j = \text{Dim}(E_j), \mu_j = \|\Pi_j(\mu)/\sigma\|^2).$$

Preuve Th. de Cochran

- ▶ $Z = Y/\sigma$
- ▶ Pour tt j , soit $(e_{j1}, \dots, e_{jd_j})$ une base orthonormée de E_j

$$\Pi_j Z = \sum_{k=1}^{d_j} \langle e_{jk}, Z \rangle e_{jk}$$

- ▶ Soit $U = (e_{11} \dots e_{rd_r})'$ matrice de passage : $UU' = Id_n$.
On a : $UZ \sim \mathcal{N}_n(U\mu/\sigma, Id_n)$.

Les variables $e'_{jk}Z$ sont indépendantes quand j et k varient.

Donc $\Pi_1(Z), \dots, \Pi_r(Z)$ sont indépendantes

- ▶ Pour un sous-espace E_j , et pour $k = 1, \dots, k_j$

$$e'_{jk}Z \sim \mathcal{N}(e'_{jk}\mu/\sigma, e'_{jk}e_{jk} = 1)$$

- ▶ d'où , avec $\mu_j = \|\Pi_j\mu/\sigma\|^2 = \sum_{k=1}^{d_j} (e'_{jk}\mu/\sigma)^2$

$$\|\Pi_j(Z)\|^2 = \sum_{k=1}^{d_j} \|e'_{jk}Z\|^2 \sim \chi^2(d_j, \mu_j),$$

Loi de Student (W. Gosset)

Définition (Loi de Student)

Soit deux variables Z et K **indépendantes** telles que $Z \sim \mathcal{N}(0, 1)$ et $K \sim \chi^2(p)$. Alors, la v.a.

$$T = \frac{Z}{\sqrt{\frac{K}{p}}} \sim \mathcal{T}(p)$$

suit une loi appelée loi de **Student** à p degrés de liberté.

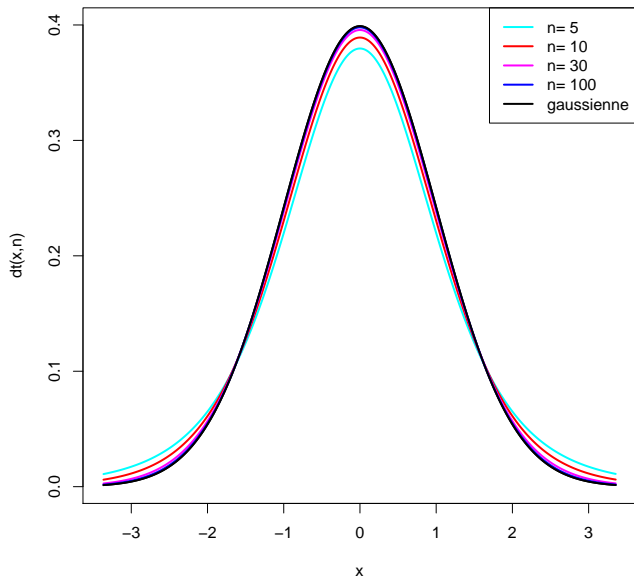
Proposition

Si X_1, \dots, X_n est un n -échantillon gaussien de loi $\mathcal{N}(\mu, \sigma^2)$,

$$\sqrt{n} \frac{\bar{X} - \mu}{\hat{\sigma}_n} \sim \mathcal{T}(n-1) \quad \text{avec} \quad \hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Application : $\mathbb{P}(\bar{X} > 6,43) = 1 - F_{\mathcal{T}}\left(\frac{6.43 - c_0}{0.25/\sqrt{n}}, n-1\right) \simeq 0.018$

Loi de Student [W. Gosset]



Définition (Loi de Fisher)

Soit deux variables K_1 et K_2 **indépendantes** telles que $K_1 \sim \chi^2(n_1)$ et $K_2 \sim \chi^2(n_2)$. Alors, la v.a.

$$F = \frac{K_1/n_1}{K_2/n_2} \sim \mathcal{F}(n_1, n_2)$$

suit une loi appelée loi de **Fisher** à (n_1, n_2) degrés de liberté.

Proposition

$\mathbb{E}(F)$ existe pour $n_2 \geq 2$ et vaut $\mathbb{E}(F) = \frac{n_2}{n_2-2}$. $\text{var}(F)$ existe pour $n_2 \geq 5$ et vaut $\text{var}(F) = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}$

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Introduction

NP

Test de Wald

Exemples

p-value

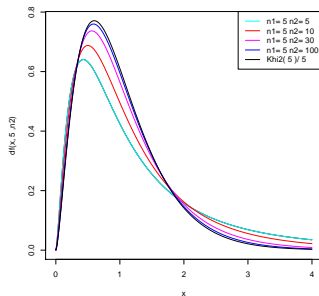
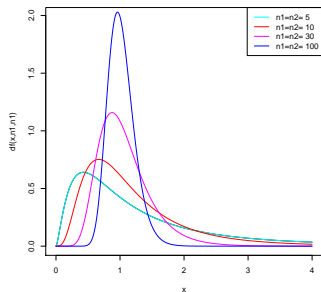
Intervalle de
confiance

Introduction

Construction

Région de confiance

Loi de Fisher



Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Introduction

NP

Test de Wald

Exemples

p-value

Intervalle de
confiance

Introduction

Construction

Région de confiance

Proposition (Loi du rapport des estimateurs de variance)

Soient deux échantillons gaussiens indépendants de taille n_1 et n_2 , de **même variance** σ^2 , et soient $\hat{\sigma}_1^2$ et $\hat{\sigma}_2^2$ les estimateurs **non biaisés** de la variance σ^2 dans chacun des deux échantillons. Alors, la v.a.

$$\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \sim \mathcal{F}(n_1 - 1, n_2 - 1)$$

Approximation gaussienne

Si la loi mère n'est pas gaussienne, et si la loi de \bar{X} est difficile à identifier, on peut utiliser des **approximations gaussiennes** pour des échantillons **suffisamment grands**.

Théorème (de limite centrale)

Soit $\{X_n\}$ une suite de variables aléatoires i.i.d. admettant une espérance μ et une variance σ^2 finie. Alors, la suite des variables $\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}$ converge en loi vers la v.a. $\mathcal{N}(0, 1)$ quand $n \rightarrow \infty$

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Introduction

NP

Test de Wald

Exemples

p-value

Intervalle de
confiance

Introduction

Construction

Région de confiance

Application

La loi de l'estimateur empirique de l'espérance d'une loi **quelconque** de moment d'ordre 2 fini peut être approximée par une loi gaussienne

$$\bar{X} \stackrel{appr}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Exemple

► si $X \sim \mathcal{B}(\pi)$, si $n\pi > 5$ et $n(1 - \pi) > 5$

$$\bar{X} \stackrel{appr}{\sim} \mathcal{N}\left(\pi, \frac{\pi(1 - \pi)}{n}\right), \text{ ou } n\bar{X} \stackrel{appr}{\sim} \mathcal{N}(n\pi, n\pi(1 - \pi))$$

Proposition

si $X_n \xrightarrow{\mathcal{L}} X$ et $Y_n \xrightarrow{\mathcal{L}} c$ où c est une constante, alors

$$(X_n, Y_n) \xrightarrow{\mathcal{L}} (X, c)$$

En particulier :

$$X_n + Y_n \xrightarrow{\mathcal{L}} X + c, X_n Y_n \xrightarrow{\mathcal{L}} cX; X_n/Y_n \xrightarrow{\mathcal{L}} X/c$$

Application : Si X_1, \dots, X_n est un n -échantillon de loi d'espérance μ et de variance σ^2 finie,

$$\sqrt{n} \frac{\bar{X} - \mu}{\hat{\sigma}_n} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \quad \text{avec} \quad \hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$\sqrt{n} \frac{\bar{X} - \mu}{\sqrt{V_n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \quad \text{avec} \quad V_n = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Introduction

NP

Test de Wald

Exemples

p-value

Intervalle de
confiance

Introduction

Construction

Région de confiance

Définition

Si un estimateur $\hat{\nu}_n$ de $\nu \in \mathbb{R}^p$ de variance $\text{var}(\hat{\nu}_n) = V_n \rightarrow 0$ a un comportement asymptotique tel que

$$V_n^{-1/2}(\hat{\nu}_n - \nu) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, Id_p)$$

on dit que l'estimateur est **asymptotiquement normal**. Si $nV_n \rightarrow V_0$ où $V_0 > 0$ est finie, on dit que la **vitesse** de l'estimateur est en \sqrt{n}

\Leftrightarrow Un estimateur est d'autant meilleur que sa vitesse de convergence est rapide et sa loi limite concentrée autour de 0.

Proposition

Si h est une fonction différentiable de $\nu \in \mathbb{R}^p$ et $\widehat{\nu}_n$ un estimateur asymptotiquement normal, alors $h(\widehat{\nu}_n)$ est un estimateur asymptotiquement normal de $h(\nu)$

$$(D_\nu V_n(\nu) D'_\nu)^{-1/2} (h(\widehat{\nu}_n) - h(\nu)) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, Id_p)$$

avec $D_\nu = \left(\partial h(\nu) / \partial \nu_1 \quad \dots \quad \partial h(\nu) / \partial \nu_p \right)$

Sommaire

Modèle statistique

Estimation

Vraisemblance

EMV

Tests

Intervalle de confiance

Cours accéléré
Statistiques
Partie 1: Bases de
la statistique
inférentielle

Christine Keribin

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Introduction

NP

Test de Wald

Exemples

p-value

Intervalle de
confiance

Introduction

Construction

Région de confiance

Méthodes de construction d'estimateurs

Méthode des moments

Soit $X_1, \dots, X_n \sim \mathbb{P}_\theta$.

- ▶ Les moments de \mathbb{P}_θ dépendent de θ . Moment d'ordre k

$$\mathbb{E}(X_1^k) = m_k(\theta)$$

- ▶ On définit le **moment empirique** d'ordre k
 $\hat{m}_k = \frac{1}{n} \sum_{i=1}^n X_i^k$ fortement consistant si $\mathbb{E}(|X^k|)$ existe.
- ▶ On exprime $\nu(\theta) = g(m_1, \dots, m_k)$ en fonction des moments
- ▶ On remplace chaque moment par son estimateur empirique (méthode *plug-in*) : $\widehat{\nu(\theta)} = g(\hat{m}_1, \dots, \hat{m}_k)$

Méthode du maximum de vraisemblance

Modèle statistique $\mathcal{M} = (\mathcal{X}^n, \mathcal{A}^n, \mathcal{P} = (\mathbb{P}_\theta^n))$ dominé

Modèle paramétrique **dominé** : les lois $(\mathbb{P}_\theta)_{\theta \in \Theta}$ admettent une **densité** $p_\theta(x)$ par rapport à une mesure commune ξ (mesure de Lebesgue, mesure de comptage)

$$\forall A \in \mathcal{A}, \quad \mathbb{P}_\theta(A) = \int_A p_\theta(x) d\xi(x)$$

Définition

Dans un modèle paramétrique dominé, on appelle **vraisemblance** d'une réalisation (x_1, \dots, x_n) du n -échantillon, la fonction de θ :

$$\theta \mapsto L(\theta; x_1, \dots, x_n) = p_\theta(x_1, \dots, x_n)$$

Pour un échantillon *i.i.d.* : $L(\theta; x_1, \dots, x_n) = \prod_{i=1}^n p_\theta(x_i)$

! : Vraisemblance \neq densité

Exemple : loi gaussienne, loi de Bernoulli

Définition

Un modèle paramétrique $(\mathcal{X}, \mathcal{A}, \mathbb{P}_\theta)$, $\theta \in \Theta$ ouvert de \mathbb{R}^p , et tel que \mathbb{P}_θ admette une densité $f(\cdot; \theta)$ par rapport à une mesure dominante ν , est **régulier** si

- ▶ Le support des lois $f(\cdot; \theta)$ est indépendant de $\theta \in \Theta$
- ▶ $\theta \mapsto \log f(x; \theta)$ est deux fois continûment différentiable sur Θ , pour tout $x \in \mathcal{X}$
- ▶ Pour tout $A \in \mathcal{A}$, l'intégrale $\int_A f(x; \theta) d\nu(x)$ est au moins deux fois dérivable sous le signe d'intégration et on peut permuter intégration et dérivation

Ex : Bernoulli, gaussien ; Contre-ex : $\mathcal{U}[0, \theta]$

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Introduction

NP

Test de Wald

Exemples

p-value

Intervalle de
confiance

Introduction

Construction

Région de confiance

Soit $\ell_\theta = \log f_\theta$

Définition

Dans un modèle paramétrique dominé, si pour tout $x \in \mathcal{X}$ la vraisemblance est différentiable, le vecteur gradient de la log-vraisemblance est le **vecteur aléatoire** appelé **score** (de Fisher) et défini par

$$\dot{\ell}_\theta(X) = \begin{pmatrix} \frac{\partial}{\partial \theta_1} \ell_\theta(X) \\ \vdots \\ \frac{\partial}{\partial \theta_p} \ell_\theta(X) \end{pmatrix}$$

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Introduction

NP

Test de Wald

Exemples

p-value

Intervalle de
confiance

Introduction

Construction

Région de confiance

Vecteur du score : propriétés

- ▶ le score est **additif** : pour deux variables aléatoires indépendantes X et Y de même loi,

$$\dot{\ell}_{\theta}(X, Y) = \dot{\ell}_{\theta}(X) + \dot{\ell}_{\theta}(Y)$$

- ▶ Dans un modèle régulier, le score est un vecteur aléatoire centré

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Introduction

NP

Test de Wald

Exemples

p-value

Intervalle de confiance

Introduction

Construction

Région de confiance

Définition

Dans un modèle paramétrique régulier, on appelle *Information de Fisher* au point $\theta \in \Theta \subset \mathbb{R}^p$ la *matrice de variance du score* :

$$I_n(\theta) = \mathbf{E}_\theta[\dot{\ell}_\theta(X)[\dot{\ell}_\theta(X)]'] \underset{\text{modèle régulier}}{=} \text{var}(\dot{\ell}_\theta(X))$$

où la notation prime ' indique la transposée. C'est une matrice de taille $p \times p$, **symétrique**, **semi-définie positive**.

- ▶ Si $X = (X_1, \dots, X_n)$ est un n -échantillon iid d'information $I_n(\theta)$, alors (**additivité**)

$$I_n(\theta) = nI_1(\theta)$$

- ▶ modèle régulier. Soit $\ddot{\ell}_\theta(X)$ la matrice des dérivées secondes en θ de la log-vraisemblance, alors :

$$I_n(\theta) = \mathbb{E}_\theta[\dot{\ell}_\theta(X)[\dot{\ell}_\theta(X)]'] = -\mathbb{E}_\theta[\ddot{\ell}_\theta(X)]$$

Exemple : modèle de Bernoulli, Gaussien

Théorème (FDCR)

Soit h une fonction différentiable de Θ , un ouvert de \mathbb{R}^k .
Dans un modèle est *régulier* d'information de Fisher finie et
inversible, pour tout estimateur T_n de $h(\theta)$, *sans biais* et de
carré intégrable et on a

$$\text{var}(T_n) \geq [\dot{h}(\theta)]' I_n(\theta)^{-1} \dot{h}(\theta)$$

- ▶ La limite inférieure de la variance des estimateurs sans biais s'appelle **borne de Cramér-Rao**
- ▶ L'information de Fisher est la **précision** maximale avec laquelle la fonction $h(\theta)$ peut être estimée.

Définition

Un estimateur *sans biais* T_n est *efficace* pour $h(\theta)$ s'il atteint la borne de Cramér-Rao, ie pour tout $\theta \in \Theta$,

$$\text{var}(T_n) = [\dot{h}(\theta)]' I_n(\theta)^{-1} \dot{h}(\theta)$$

et il est donc UVMB, optimal parmi les estimateurs sans biais.

Efficacité mais...

Théorème (admis)

La borne de Cramér-Rao n'est atteinte que si

- (a) la loi des observations est d'une *famille exponentielle* : modèle dominé dont la densité de la loi mère peut s'écrire sous la forme

$$f(x; \theta) = \exp \left(a(x)\alpha(\theta) + \beta(\theta) + c(x) \right)$$

pour tout $x \in \mathbb{R}$

- (b) et pour l'estimation d'une *fonction particulière* de θ définie à une transformation affine près

$$h(\theta) = \mathbb{E}_{\theta}(a(X)).$$

Rem : il peut ne pas exister d'estimateurs efficaces - Il peut y avoir des estimateurs optimaux (UVMB) non efficaces

Sommaire

Modèle statistique

Estimation

Vraisemblance

EMV

Tests

Intervalle de confiance

Cours accéléré
Statistiques
Partie 1: Bases de
la statistique
inférentielle

Christine Keribin

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Introduction

NP

Test de Wald

Exemples

p-value

Intervalle de
confiance

Introduction

Construction

Région de confiance

Méthode du maximum de vraisemblance

La valeur θ_1 de θ est plus vraisemblable que la valeur θ_2 , si $L(\theta_1; \mathbf{x}) > L(\theta_2; \mathbf{x})$

Définition (EMV)

On appelle *estimation du maximum de vraisemblance*, une valeur $\hat{\theta}_n$ maximisant la vraisemblance

$$\hat{\theta}_n \in \text{Arg max}_{\theta \in \Theta} L(\theta; \mathbf{x}).$$

$\hat{\theta}_n = t(x_1, \dots, x_n)$ est une fonction des données, ce qui induit la statistique $t(X_1, \dots, X_n)$ que l'on note (abusivement) avec la même notation :

$\hat{\theta}_n = t(X_1, \dots, X_n)$ est appelé *Estimateur du Maximum de Vraisemblance*

Calcul de l'EMV

Remarque : quand l'échantillon est i.i.d. on utilise plutôt la log-vraisemblance

$$\ell_n(\theta; \mathbf{x}) = \log L(\theta; \mathbf{x}) = \sum_{i=1}^n \log p_\theta(x_i)$$

Méthode :

- Chercher $\hat{\theta}_n$ annulant les équations de vraisemblance (ou équations du **score**)

$$U_n(\hat{\theta}_n) := \dot{\ell}_n(\hat{\theta}_n; \mathbf{x}) = \left(\frac{\partial}{\partial \theta_k} \ell_n(\hat{\theta}_n; \mathbf{x}) \right)_{k=1, \dots, \dim(\theta)} = 0,$$

- Vérifier que $\hat{\theta}_n$ est bien un maximum : $H_n(\theta) = \ddot{\ell}_n(\theta; \mathbf{x})$ est définie négative autour de $\hat{\theta}_n$.

Exemple : loi gaussienne, loi de Bernoulli

- ▶ si la vraisemblance n'est pas strictement concave pour tout θ , il peut exister des optima locaux
- ▶ l'EMV n'est pas forcément unique
- ▶ l'EMV peut ne pas exister
- ▶ pb de dérivabilité, par ex : $\mathcal{U}(0, \theta)$
- ▶ utilisation d'un schéma numérique si le calcul analytique n'est pas possible

mais

- ▶ Pour n'importe quelle application g de Θ , si $\hat{\theta}$ est l'EMV de θ , alors $g(\hat{\theta})$ est l'EMV de $g(\theta)$.
- ▶ de bonnes propriétés asymptotiques

Proposition

Sous des conditions de *régularité* du modèle, l'EMV $\hat{\theta}_n$ du paramètre θ calculé sur un n -échantillon *i.i.d.* est

- ▶ *asymptotiquement sans biais*

$$\lim_{n \rightarrow \infty} \mathbf{E}_{\theta}(\hat{\theta}_n) = \theta$$

- ▶ *convergent*
- ▶ *asymptotiquement normal*

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I_1(\theta)^{-1})$$

où $I_1(\theta) = \mathbf{E}_{\theta}[\dot{\ell}_n(X_1)(\dot{\ell}_n)']$ est l'*information de Fisher*

Proposition

Des propriétés intéressantes pour les "grands" échantillons
(par ex $n \geq 30$)

- ▶ L'EMV $\hat{\theta}_n$ est *asymptotiquement efficace* et donc UVMB :
 - ↪ très bonne qualité d'estimation
- ▶ $\hat{\theta}_n$ tend à devenir gaussien :
 - ↪ *loi approchée* à distance finie

$$\hat{\theta}_n \overset{\text{appr}}{\sim} \mathcal{N}(0, [nI(\theta)]^{-1} = [I_n(\theta)]^{-1})$$

Proposition

Si X_1, \dots, X_n est un n -échantillon gaussien de loi $\mathcal{N}(\mu, \sigma^2)$,

- ▶ l'EMV est

$$\hat{\mu}_n = \bar{X}, \quad V_n = \frac{(n-1)\widehat{\sigma}_n^2}{n}$$

où $\widehat{\sigma}_n^2 = \sum_i (X_i - \bar{X})^2 / (n-1)$ est l'estimateur sans biais de la variance.

- ▶ \bar{X} et $\widehat{\sigma}_n^2$ sont indépendants et vérifient

$$\sqrt{n}(\bar{X} - \mu) \sim \mathcal{N}(0, \sigma), \quad \frac{(n-1)\widehat{\sigma}_n^2}{\sigma^2} \sim \chi_{n-1}^2$$

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Introduction

NP

Test de Wald

Exemples

p-value

Intervalle de
confiance

Introduction

Construction

Région de confiance