

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

2022-2023

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

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

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
- ▶ Applications pratiques avec le logiciel R.

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

[//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

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

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.

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

- 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 ?

- ▶ **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**

- **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

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

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

Rappels
Introduction
Construction
Décision et risques
Hyp. composites
Propriétés
NP
Test de Wald
Exemples
p-value

Sommaire

Modèle statistique

Estimation

Vraisemblance

EMV

Tests

Intervalle de confiance

Introduction

Modèle statistique

Estimation

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

Vraisemblance

- Information de Fisher
- Efficacité

EMV

Tests

- Rappels
- Introduction
- Construction
- Décision et risques
- Hyp. composites
- Propriétés
- NP
- Test de Wald
- Exemples
- p-value

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 $X_1, \dots, X_n$

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

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)

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

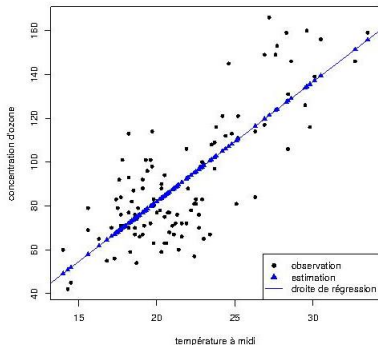


Figure – source des données : Cornillon et Matzner-Løber, 2007

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

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

Un objet mathématique : l'estimateur

Soit $\theta \in \Theta \in \mathbb{R}$ le paramètre d'une loi $\mathbb{P}_\theta$ 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 $\mathbb{P}_\theta$:

- 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

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

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

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

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 $\hookrightarrow$ Recherche d'estimateurs **UVMB**

Performance asymptotique

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

- $\hat{\nu}_n$ 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 6.42 est d'une probabilité forte ou pas sous les indications du constructeur.
 - ▶ accéder à $\mathbb{P}(\bar{X} \geq 6.42)$,
 - ▶ loi de $\bar{X}$

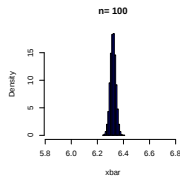
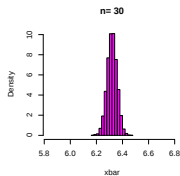
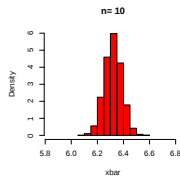
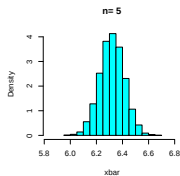
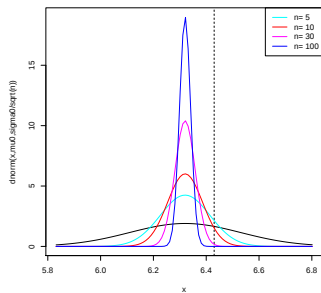
Cas Gaussien

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 :

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

Loi de $\bar{X}$



Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

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

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

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

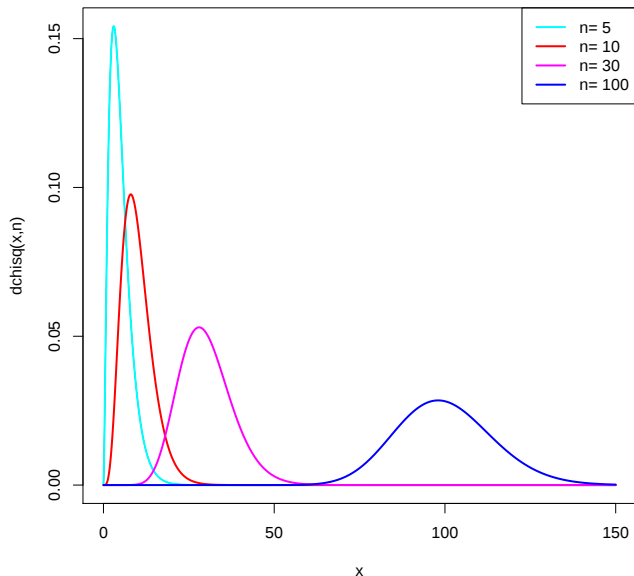
$$\mathbb{E}(K_n) = n; \quad \text{var}(K_n) = 2n$$

Proposition

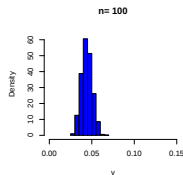
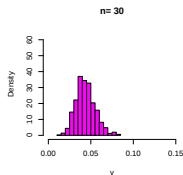
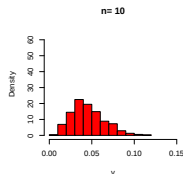
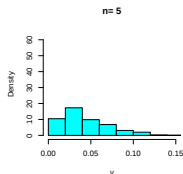
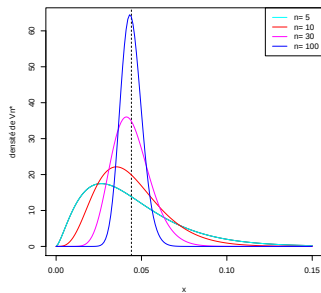
Soit $V_n^* = \frac{1}{n} \sum_i (X_i - \mu)^2$, l'estimateur empirique de la variance d'un échantillon i.i.d. X de loi $\mathcal{N}(\mu, \sigma^2)$, μ connue :

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

Loi du $\chi^2(n)$



Loi de V_n^*



Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

Loi de la variance empirique S_n^2

Proposition

La loi de l'estimateur de la variance empirique S_n^2 d'un n -échantillon i.i.d. gaussien X est telle que

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

*De plus, $\bar{X}$ et S_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

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

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.

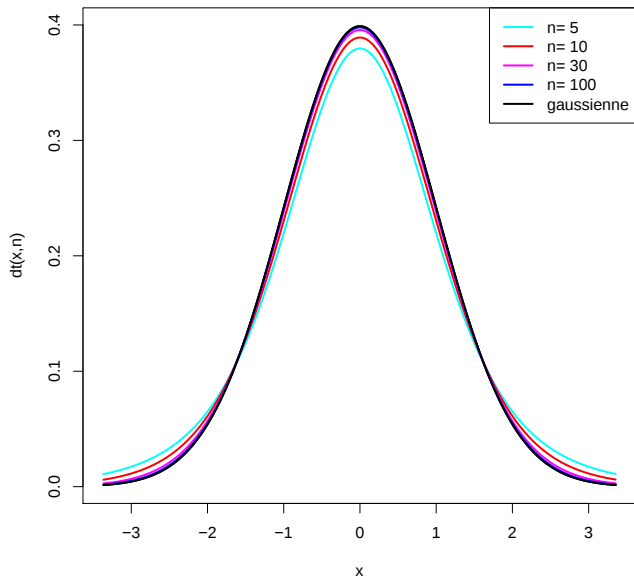
Proposition

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

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

Loi de Fisher

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

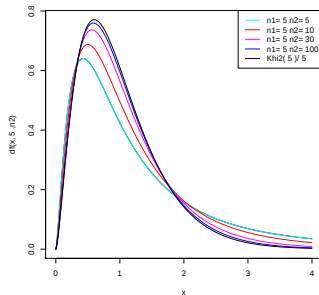
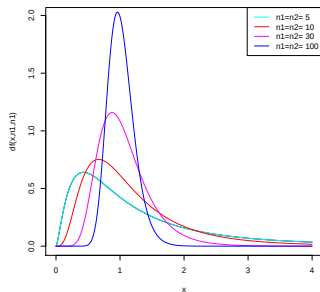
Propriétés

NP

Test de Wald

Exemples

p-value



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

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

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,

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

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

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 $\sqrt{n}$

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

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

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

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

Méthodes de construction d'estimateurs

Méthode des moments

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

- ▶ Les moments de $\mathbb{P}_\theta$ 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

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

Définition

Un modèle paramétrique $(\mathcal{X}, \mathcal{A}, \mathbb{P}_\theta)$, $\theta \in \Theta$ ouvert de $\mathbb{R}^p$, et tel que $\mathbb{P}_\theta$ 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

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

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

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

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

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* :

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

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

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

- 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.

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

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

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

Méthode du maximum de vraisemblance

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

Définition (EMV)

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

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

$\hat{\theta}_n = t(x_1, \dots, x_n)$ 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*

Introduction

Modèle statistique

Estimation

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

Vraisemblance

- Information de Fisher
- Efficacité

EMV

Tests

- Rappels
- Introduction
- Construction
- Décision et risques
- Hyp. composites
- Propriétés
- NP
- Test de Wald
- Exemples
- p-value

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*

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

- *convergent*
- *asymptotiquement normal*

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

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

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

Proposition

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

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

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

Proposition

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

► l'EMV est

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

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

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

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

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

As so far...

- Un estimateur est une variable aléatoire fonction de l'échantillon, permettant d'inférer la valeur d'un paramètre

- biais, variance, risque
- consistance
- loi : exacte ou approchée par l'asymptotique

- VRAI ou FAUX ?

- Un estimateur non biaisé est forcément de risque minimum
- La somme des carrés de n variables aléatoires gaussiennes centrées réduites indépendantes suit une loi du Khi-deux à n ddl
- Soit X un n -échantillon iid gaussien. Alors,

$$T_n = \frac{\bar{X} - \mathbb{E}(X_1)}{\sum (X_i - \bar{X})^2 / (n-1)} \sim \mathcal{T}(n-1);$$

- Tous les estimateurs consistants ont un comportement asymptotiquement normal

Tests : un exemple

Un constructeur automobile annonce une consommation $\mu_0 = 6.32\ell/100\text{ km}$, avec un écart type $\sigma = 0.21\ell/100\text{ km}$, pour des véhicules d'un type donné. Un organisme indépendant suspecte une sous-estimation de cette consommation et indique que la consommation s'élèverait à $\mu_1 = 6.45\ell/100\text{ km}$.

Sur un 30-échantillon $\bar{x} = 6.43\ell/100\text{ km}$. *Qui a raison ?*

(H_0) conso. conforme au constructeur : $X \sim \mathcal{N}(\mu_0, \sigma^2)$
 $\mu = \mu_0 = 6.32$

(H_1) conso. suspectée par l'organisme : $X \sim \mathcal{N}(\mu_1, \sigma^2)$
 $\mu = \mu_1 = 6.45$

- Choisir, à partir d'un n -échantillon, entre les deux hypothèses (H_0) et (H_1) , en assumant le **risque de première espèce** α (5%, 10%,...) de choisir (H_1) alors que (H_0) est vrai.

Un exemple (suite)

- ▶ Statistique $\bar{X}$, moyenne des consommations de $n = 30$ véhicules
- ▶ Loi sous (H_0) ,

$$\bar{X} \sim \mathcal{N}\left(\mu_0, \frac{\sigma^2}{n}\right) \text{ soit } T = \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma} \sim \mathcal{N}(0, 1)$$

- ▶ Choisir a priori un risque α , calibrant la probabilité de rejet de (H_0) à tort ($\alpha = 5\%$ par exemple)

$$\begin{aligned} \alpha &= \mathbb{P}_{H_0}\left(\underbrace{T > q_{1-\alpha}^*}_{\mathcal{R}=]q_{1-\alpha}^*; \infty[, \text{ Région de rejet pour } T}\right) \\ &= \mathbb{P}_{H_0}\left(\underbrace{\bar{X} > \mu_0 + q_{1-\alpha}^* \frac{\sigma}{\sqrt{n}}}_{\mathcal{R}=] \mu_0 + q_{1-\alpha}^* \frac{\sigma}{\sqrt{n}}; \infty[, \text{ Région de rejet pour } \bar{X}}\right) \end{aligned}$$

avec $q_{1-\alpha}^*$ le quantile d'une loi $\mathcal{N}(0, 1)$ d'ordre $1 - \alpha$

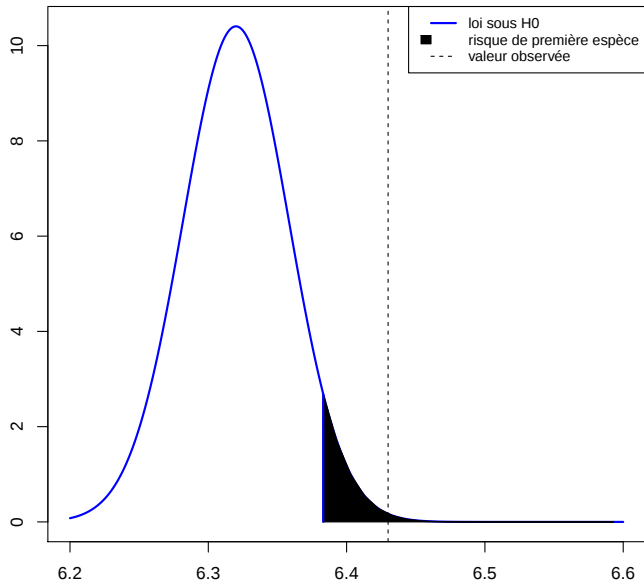
Un exemple (suite)

- Décider :
 - si T est dans la région de rejet, on **rejette** H_0
 - sinon, on **ne rejette pas** [**conserve/accepte**] (H_0) faute de preuves suffisantes
- ici, $\bar{x} = 6.43\ell/100\text{ km}$,

$$t_{obs} = \frac{6.43 - 6.32}{0.21/\sqrt{30}} = 2.86 > 1.64$$

au niveau $\alpha = 5\%$, les données sont significatives pour rejeter (H_0), le constructeur a minimisé la consommation, avec un risque (de première espèce) α .

Représentation graphique



Un exemple (suite)

Un autre cas de figure :

- si la même consommation a été observée sur un échantillon de $n = 9$ véhicules

$$t_{obs} = \frac{6.43 - 6.32}{0.21/\sqrt{9}} = 1.57 < 1.64$$

on ne peut pas rejeter le fait que le constructeur a sous-estimé la consommation, on ne peut rejeter (H_0) qu'on accepte par défaut

- avec quelle erreur ?

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

Un exemple (suite)

Une autre façon de se tromper :

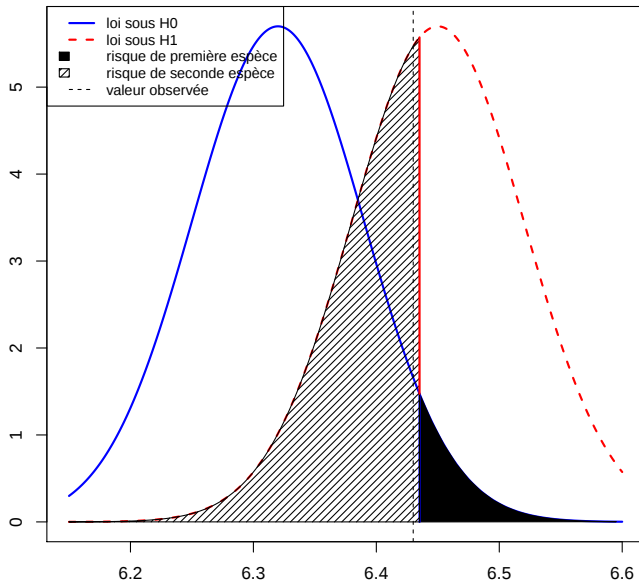
- **erreur de seconde espèce** : ne pas rejeter (H_0) alors que (H_1) est vraie
- Sous (H_1), $\bar{X} \sim \mathcal{N}(\mu_1, \frac{\sigma^2}{n})$ et le risque de seconde espèce est

$$\begin{aligned}\beta &= \mathbb{P}_{H_1} \left(\bar{X} < \mu_0 + q_{1-\alpha}^* \frac{\sigma}{\sqrt{n}} \right) \\ &= F^* \left(\sqrt{n} \frac{\mu_0 - \mu_1}{\sigma} + q_{1-\alpha}^* \right)\end{aligned}$$

- App.Num : $n = 9$, $\beta \simeq 0.41$

la **puissance** $\pi = 1 - \beta$ n'est pas très grande

Représentation graphique



Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

Définition

- Un **test** est une procédure de décision qui permet de trancher, au vu des résultats d'un échantillon, entre deux hypothèses l'**hypothèse nulle** (H_0) et une hypothèse **alternative** (H_1), dont une seule est vraie.
- La **région critique** ou région de **rejet** $\mathcal{R}$ est l'ensemble des valeurs de la variable de décision T qui conduisent à écarter (H_0) au profit de (H_1).
- La région d'**acceptation** du test est $\overline{\mathcal{R}}$.

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

Fonction de test

La **fonction test** est la fonction définissant le rejet du test.
Soit $T(X)$ la statistique de test,

$$\varphi(x) = \mathbb{I}\{T(x) \in \mathcal{R}\}$$

Le test rejette H_0 si et seulement si $\varphi(X) = 1$

Remarques :

- ▶ On note souvent la région de rejet et l'événement $\{T(X) \in \mathcal{R}\}$ de la même façon
- ▶ Quand la région de rejet est sous forme d'un intervalle, on appelle seuil la (les) bornes qui limitent l'intervalle. On peut exprimer le seuil sur la statistique pivotale ou sur une statistique ayant servi à la calculer.

Procédure de test

A l'issue du test, les quatre situations suivantes sont possibles

	Choix (H_0)	Choix (H_1)
(H_0) vraie	$1 - \alpha$ bonne décision	$\alpha = \mathbb{P}_{H_0}(T \in \mathcal{R})$ risque première espèce mauvaise décision
(H_1) vraie	$\beta = \mathbb{P}_{H_1}(T \notin \mathcal{R})$ risque seconde espèce mauvaise décision	$\pi = 1 - \beta$ puissance bonne décision

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

Construire un test

- ▶ Définir le **modèle**
- ▶ Définir les **hypothèses nulle** (H_0) et **alternative** (H_1)
- ▶ Choisir une **statistique de test** $T(X)$, calculer sa **loi sous** (H_0)
- ▶ Définir la **règle de décision** en calibrant la région de rejet $\mathcal{R}$ suivant le risque α
- ▶ Calcul de la puissance π
- ▶ Calcul de la statistique observée et **décision** : rejet ou acceptation de (H_0).

La décision du test, à partir de la valeur observée t de la statistique de test T est :

- ▶ si $t \in \mathcal{R}$, on **rejette** (H_0) au risque α : l'erreur commise est de risque $\alpha = \mathbb{P}_{(H_0)}(T \in \mathcal{R})$.
- ▶ si $t \notin \mathcal{R}$, on conserve (H_0) dans le test de risque α : les données **ne sont pas significatives** pour accepter (H_1).
L'erreur de seconde espèce commise est de risque $\beta = \mathbb{P}_{(H_1)}(T \notin \mathcal{R})$, en général **inconnu**.

La décision est **aléatoire**, elle dépend de l'échantillon

Dissymétrie de la situation de test

- ▶ Le risque n'est contrôlé que pour (H_0)
 - ↪ La véritable décision est celle qui rejette (H_0).
 - ↪ (H_0) et (H_1) ne sont pas interchangeables.
- ▶ Il faut connaître la loi de la statistique de test sous (H_0)
- ▶ Il faut que cette loi soit différente sous (H_1)
- ▶ Entre deux tests de même risque de 1ère espèce α , il faut choisir le plus puissant
 - ▶ La région de rejet de la forme $\{T < q_{0.05}^*\}$ est aussi de risque 5%, mais elle n'a aucune puissance pour détecter le cas $\mu_1 > \mu_0$

Introduction

Modèle statistique

Estimation

Estimateur

Propriétés

Lois

Cas gaussien

Cochran

Approximation gaussienne

Vraisemblance

Information de Fisher

Efficacité

EMV

Tests

Rappels

Introduction

Construction

Décision et risques

Hyp. composites

Propriétés

NP

Test de Wald

Exemples

p-value

Hypothèse alternative composite

La valeur observée n'a pas servi à construire la région de rejet qui a été définie a priori en fonction de la problématique fixée.

- Ainsi, pour tester

$$(H_0) : \theta = \theta_0 \text{ contre } (H_1) : \theta > \theta_0$$

on utilise la même région de rejet que pour le test d'hypothèse simple

$$(H_0) : \theta = \theta_0 \text{ contre } (H_1) : \theta = \theta_1 > \theta_0$$

- mais la **puissance** devient une **fonction** de θ :

$$\theta_1 \in \Theta_1 = \{\theta | \theta > \theta_0\}, \pi(\theta_1) = \mathbb{P}_{\theta_1}(T \in \mathcal{R}) = 1 - \beta(\theta_1).$$

Formes d'hypothèses

► Hypothèses simples

$$(H_0) : \theta = \theta_0 \text{ contre } (H_1) : \theta = \theta_1$$

► Test unilatéral pour une hypothèse nulle composite

$$(H_0) : \theta \leq \theta_0 \text{ contre } (H_1) : \theta > \theta_0$$

► Test unilatéral pour une hypothèse nulle composite

$$(H_0) : \theta \geq \theta_0 \text{ contre } (H_1) : \theta < \theta_0$$

► Test bilaréral pour une hypothèse nulle simple

$$(H_0) : \theta = \theta_0 \text{ contre } (H_1) : \theta \neq \theta_0$$

► De façon générale :

$$(H_0) : \theta \in \Theta_0 \text{ contre } (H_1) : \theta \in \Theta_1 = \Theta \setminus \Theta_0$$

Hypothèse nulle composite, taille, niveau

Quand l'hypothèse nulle est composite,

- le risque de 1ère espèce devient une fonction de θ :

$$\theta \in \Theta_0 = \{\theta | \theta \leq \theta_0\}, \alpha(\theta) = \mathbb{P}_\theta(T \in \mathcal{R})$$

et on définit $\sup_{\alpha \in \Theta_0} \alpha(\theta)$ est la taille du test. Un test est de niveau α si sa taille $\leq \alpha$

- ▶ A **niveau** fixé a priori, on choisit un test de **puissance** maximale
- ▶ Le test est **sans biais** si $1 - \beta(\theta) = \pi(\theta) > \alpha$ pour tout $\theta \in \Theta_1$
- ▶ Le test est **consistant** (ou convergent) si la suite des puissance $\pi_n(\theta)$ tend vers 1 quand n tend vers l'infini