

Modélisation Statistique (MAP-STA1)

M1-Mathématiques Appliquées Cours 1: Introduction et rappels

Christine Keribin

¹Laboratoire de Mathématiques d'Orsay
Université Paris-Sud

2018-2019



Introduction

Que retenir de L3/1A ?

Pour aller plus loin...

Modèle

Estimation

Références

Introduction

Que retenir de L3/1A ?

Pour aller plus loin...

Modèle

Estimation

Introduction

Que retenir de L3/1A ?

Pour aller plus loin...

Modèle

Estimation

Références

- ▶ 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 d'un 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** des propriétés de cette loi pour répondre à une question
- ⇨ résoudre un problème inverse
- ▶ Modéliser
 - ▶ Estimer
 - ▶ Utiliser à des fins explicatives ou prédictives

- 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
 $\mathcal{M} = (\{0, 1\}^n, \mathcal{A}^n, \mathcal{B}_{\pi}^{\otimes n}, \pi \in [0; 1])$

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

Introduction

Que retenir de L3/1A ?

Pour aller plus loin...

Modèle

Estimation

Références

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 ?

Pour aller plus loin...

Les modèles étudiés en L3/1A sont très simples, mais ils ont permis de décrire le cadre de la statistique inférentielle paramétrique.

Pour aller un plus loin, deux types de problématiques

- ▶ **théoriques** : choix de l'estimateur, de la procédure de test, ...
- ▶ **appliquées** : construire des modèles mieux adaptés aux situations observées (et qui induisent des questions d'ordre théoriques, **méthodologiques** et appliquées)
↪ par exemple : la régression

Régression linéaire (simple)

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

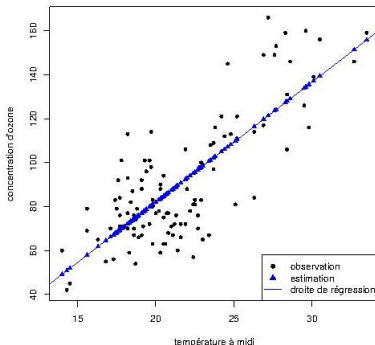


FIGURE – source : Cornillon et Matzner-Løber, 2007

Introduction

Que retenir de L3/1A ?

Pour aller plus loin...

Modèle

Estimation

Références

- ▶ Compléments en théorie de l'estimation
 - ▶ Estimateur du Maximum de Vraisemblance
 - ▶ Recherche d'estimateurs optimaux
 - ▶ Tests de Wald, du rapport de vraisemblance
 - ▶ Région de confiance
- ▶ Applications : modèle linéaire
 - ▶ Régression multiple, anova, ancova
 - ▶ Mise en pratique avec le logiciel R.

Le cours STA201 est un prérequis des cours

- ▶ STA202 (séries chronologiques), dans lequel seront développés les modèles d'observations non indépendantes
- ▶ STA203 (apprentissage statistique), dans lequel seront développés les modèles de régression non linéaires, logistique, ...

Objectifs

Être capable, en utilisant les bases théoriques de la modélisation statistique, de :

- ▶ définir une modélisation adaptée à un jeu de données réelles
- ▶ construire l'estimateur du maximum de vraisemblance, étudier ses propriétés.
- ▶ étudier le risque, l'optimalité et l'efficacité d'un estimateur
- ▶ construire les tests de Wald et du rapport de vraisemblance
- ▶ travailler avec le modèle linéaire
- ▶ estimer un modèle statistique (linéaire) avec un logiciel (R) et interpréter les résultats obtenus ;
- ▶ prendre en compte le risque de toute décision statistique.

$$\text{Note} = (2*EF+CC)/3$$

- ▶ un examen final EF sur table qui pourra comporter des questions théoriques et des questions pratiques d'interprétation de résultats. Le seul document autorisé est une feuille de notes personnelles manuscrites
- ▶ Le note de CC est celle du compte rendu d'un mini projet

Chargés de TD : Baptiste Broto, Timothée Mathieu, Solène Thépaud, Christine Keribin

Lien vers les documents, la bibliographie, ...

[https://www.math.u-psud.fr/~keribin/
EnseignementMA-STA201.htm](https://www.math.u-psud.fr/~keribin/EnseignementMA-STA201.htm)

Introduction

Que retenir de L3/1A ?

Pour aller plus loin...

Modèle

Estimation

Références

Introduction

Modèle

Estimation

Introduction

Que retenir de L3/1A ?

Pour aller plus loin...

Modèle

Estimation

Références

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

- **Modèle** : $\mathcal{M} = (\mathcal{X}^n, \mathcal{A}^n, \mathcal{P})$
 - espace mesurable $(\mathcal{X}^n, \mathcal{A}^n)$
 - muni d'une famille de lois de proba $\mathcal{P} = (\mathbb{P}_\theta^n)_{\theta \in \Theta}$

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 $X_1, \dots, X_n$

Définition (Modèle dominé)

La mesure $\mathbb{P}$ est *dominée* par rapport à une mesure positive σ -finie ξ , ou encore *absolument continue* par rapport à ξ si les ensembles négligeables pour ξ le sont également pour $\mathbb{P}$

- ▶ On peut alors définir *densité* $f = \frac{d\mathbb{P}}{d\xi}$ de la loi de $\mathbb{P} \in \mathcal{P}$
- ▶ si ξ est la mesure de Lebesgue : modèle *continu*
- ▶ si ξ est une mesure de comptage : modèle *discret*

Un modèle statistique est dominé quand la mesure ξ est commune à toutes les probabilités $\mathbb{P} \in \mathcal{P}$.

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

Définition

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

$$\theta \mapsto L(\theta; x_1, \dots, x_n) = f_{\theta}(x_1, \dots, x_n)$$

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

! : Vraisemblance $\neq$ densité

Exemple : loi gaussienne, loi de Bernoulli

Définition

Un modèle à paramétrage dans Θ est *identifiable* si

$$\forall \theta, \theta' \in \Theta, \mathbb{P}_\theta = \mathbb{P}_{\theta'} \Rightarrow \theta = \theta'$$

Introduction

Modèle

Estimation

Introduction

Que retenir de L3/1A ?

Pour aller plus loin...

Modèle

Estimation

Références

Estimer la valeur de θ (inconnu) à partir de la réalisation d'un échantillon

- ▶ Construction d'un estimateur
- ▶ Convergence de l'estimateur
- ▶ Comparaison d'estimateurs
- ▶ Quelle optimalité ?
- ▶ et si la vraie loi n'appartient pas au modèle ?

Estimation ponctuelle, par intervalle de confiance

Un outil : 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

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

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

- ▶ Notation $T_n = \hat{\theta}_n$ ou $T_n = \hat{\theta}$.
- ▶ **Propriétés** (rappel) : biais, variance, risque, consistance
- ▶ **Méthodes de construction d'estimateur** : méthode des moments, méthode du maximum de vraisemblance

Introduction

Que retenir de L3/1A ?

Pour aller plus loin...

Modèle

Estimation

Références

Comment construire un bon résumé de l'échantillon ?

Définition

On dit que T est une **statistique exhaustive** pour $\theta \in \Theta \subset \mathbb{R}^p$ si la loi de $X = (X_1, \dots, X_n)$ conditionnellement à T ne dépend pas de θ

Théorème (de factorisation (Neyman-Fisher))

La statistique $T = t(X_1, \dots, X_n)$ est exhaustive pour θ s'il existe deux applications mesurables positives g et h telles que la densité de l'échantillon puisse se factoriser sous la forme

$$f(x_1, \dots, x_n; \theta) = h(x_1, \dots, x_n)g(t(x); \theta)$$

Exemple : $\sum_i X_i$ est exhaustive pour l'estimation du paramètre de proportion d'une loi de Bernoulli

Introduction

Que retenir de L3/1A ?

Pour aller plus loin...

Modèle

Estimation

Références

- ▶ La notion d'exhaustivité n'implique pas forcément une réduction de dimension : l'échantillon X est exhaustif...
- ▶ Mais c'est une réduction suffisante pour ne pas perdre d'information : **statistique suffisante**
- ▶ Jusqu'à quel point peut-on réduire l'échantillon pour ne pas perdre d'information sur l'estimation du paramètre ?

Définition

On dit que la statistique T_n^ est **exhaustive minimale** si elle est exhaustive, et si pour toute statistique exhaustive T_n , on peut trouver une fonction u telle que $T_n^* = u(T_n)$*

- ▶ Tout estimateur pertinent est fonction d'une statistique exhaustive minimale
- ▶ si $\Theta \subset \mathbb{R}^K$, une statistique exhaustive de dimension K est en règle générale minimale (mais il n'existe pas forcément de stat exhaustive à valeur dans $\mathbb{R}^k$ pour estimer $\theta \in \mathbb{R}^k$.)

Introduction

Que retenir de L3/LA ?

Pour aller plus loin...

Modèle

Estimation

Références

V. Rivoirard et G. Stoltz. *Statistique en action*, Vuibert 2009

J.-F. Delmas, *Introduction au calcul des probabilités et à la statistique*, Les Presses de l'ENSTA 2010

J. Pagès, *Statistique générales pour utilisateurs*, PUR 2005.

P.-A. Cornillon et E. Matzner-Lober, *Régression : Théorie et Applications*, Springer 2007

J.-M. Azaïs et J.-M. Bardet, *Le modèle linéaire par l'exemple*, Dunod 2005

P.-A. Cornillon et autres, *Statistiques avec R*, PUR 2008.