Statistique (MA101) Cours 1

ENSTA 1ère année

Christine Keribin

christine.keribin@math.u-psud.fr

Laboratoire de Mathématiques
Université Paris-Sud

2017-2018

Sommaire

Statistique

de stare (*établir* en grec) puis status (*état* en latin)

- Ensemble de données observées
  - ↪ Population, échantillon, individus, variables
- Activité qui consiste dans leur recueil, traitement et interprétation
- Discipline mathématique qui fonde l'activité précédente
  - ↪ théorique : statistique mathématique
  - ↪ méthodologique
  - ↪ appliquée : cas pratiques d'étude de jeu de données

Population

- Population = ensemble d'éléments appelés unités statistiques ou individus sur lesquels on observe une ou plusieurs caractéristiques ou variables
  - $\hookrightarrow$ quantitative (valeur numérique associée à une mesure) : discrète ou continue
  - $\hookrightarrow$ qualitative (attribut ou modalité) : nominale ou ordinale
- Si la population est finie de taille N
  - $\hookrightarrow$ étude exhaustive : recensement
  - $\hookrightarrow$ si elle n'est pas possible : sondage avec ou sans remise
- Si la population est infinie, l'échantillonnage avec ou sans remise sont identiques

Un exemple ...

Statistisque inférentielle

Probabilité/statistique : une différence de point de vue

- 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
  - $\hookrightarrow$ définir un modèle
  - $\hookrightarrow$ résoudre un problème inverse
  - $\hookrightarrow$ construire une variable aléatoire fonction de l'échantillon (estimateur) qui a de bonnes propriétés

$\neq$ Statistique descriptive

Statistique inférentielle

- estimation : valeur d'un paramètre d'intérêt, ...
- test : comparaison de deux échantillons, ...
- prédiction pour une nouvelle unité non encore observée
- classification dans un groupe

$\hookrightarrow$ Problématiques :

- choix de l'estimateur, de la procédure de test ?
- fiabilité de l'information obtenue ?

$\hookrightarrow$ Outils mathématiques : variables aléatoires, probabilité et statistique, optimisation

Démarche statistique

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

- Acquérir et préparer les données
- Définir un modèle adapté à la situation observée.
- Estimer les paramètres du modèle grâce aux observations.
- Vérifier l'adéquation de l'estimation aux observations.
- Proposer d'autres modèles et choisir le plus adapté à un objectif donné (interprétation, prédiction)
- Utiliser et décider !

Tous les modèles sont faux, mais certains sont plus utiles que d'autres (G. Box)

De nombreuses applications

Modélisation aléatoire de situations (complexes) pour

- aider à la compréhension
- prendre des décisions

dans tous les domaines : économique, industriel, sciences du vivant, sciences de la nature, etc

- fiabilité de systèmes
- modélisation d'événements extrêmes
- prédiction de consommation électrique
- systèmes de recommandation
- détection d'émotions dans des tweets
- classification automatique d'images
- ...

De la statistique au dataming et à la science des données

Contenu MA101 : Bases de la statistique inférentielle

Estimation, Test, Intervalle de confiance

Objectifs

- Définir un modèle statistique paramétrique
- Construire des estimateurs et en étudier les propriétés (biais, variance, consistance)
- Définir la loi d'une statistique (exacte ou asymptotique)
- Construire un intervalle de confiance d'un paramètre univarié
- Construire un test et savoir en interpréter les résultats

Evaluation

- Note stat : un examen théorique final (seul document autorisé : une feuille de notes personnelles)
- Note finale MA101= (note proba+ note stat)/2

Bibliographie

C. Keribin.
Le poly du cours et le *site pédagogique* : `https://www.math.u-psud.fr/~keribin/EnseignementMA-MA101.htm`

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

J-J. Daudin, S. Robin, C. Vuillet.
Statistique inférentielle. Idées, démarches, exemples.
Presses Universitaires de Rennes, Rennes, 2002.

J. Pagès.
Statistique générales pour utilisateurs.
Presses Universitaires de Rennes, Rennes, 2005.

M. Lejeune.
Statistique La théorie et ses applications.
Springer, 2010.

Sommaire

Modèle statistique

- Modèle : $(\mathcal{X}^n, \mathcal{A}^n, \mathbb{P}_\theta^n, \theta \in \Theta)$.
  $\hookrightarrow$ $\mathcal{X}^n$ espace mesuré par une tribu $\mathcal{A}^n$ et une $(\mathbb{P}_\theta^n)_{\theta\in\Theta}$, famille de lois de probabilité
  Quand il existe $d \in \mathbb{N}^*$ tel que $\Theta \subset \mathbb{R}^d$, le modèle est dit *paramétrique*
- un *n*-échantillon i.i.d. $X = (X_1, \ldots, X_n)$ comporte n variables aléatoires indépendantes (i.) et de même loi (i.d.) : $\mathbb{P}_\theta^n = \mathbb{P}_\theta^{\otimes n}$
  ce qui est supposé dans la suite
- Une observation est une variable aléatoire X à valeur dans $\mathcal{X}^n$ et dont la loi appartient à $(\mathbb{P}^n{}_\theta)_{\theta\in\Theta}$
- Les données sont les réalisations (valeurs) $x_1, \ldots, x_n$ prises par l'échantillon $X_1, \ldots, X_n$

Exemples

- Estimation d'une proportion

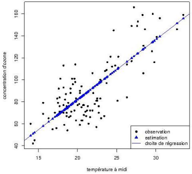
$$X_i \sim_{i.i.d.} \mathcal{B}(1, \theta)$$

- Estimation du rendement d'épis de maïs

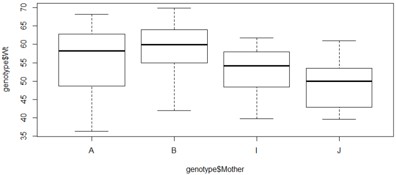
$$X_i \sim_{i.i.d.} \mathcal{N}(\mu, \sigma^2)$$

Exemples

- Régression simple : $Y_i(x_i) \sim_{i.i.d.} \mathcal{N}(\mu + \beta x_i, \sigma^2)$



- Analyse de la variance : $Y_i(g_i) \sim_{i.i.d.} \mathcal{N}(\beta_{g_i}, \sigma^2)$



Statistique

Résumer les n valeurs de l'échantillon par quelques caractéristiques simples

Définition

Une statistique T_n est variable aléatoire, fonction réelle ou vectorielle mesurable de l'échantillon $X = (X_1, \ldots, X_n)$, et ne dépendant pas des caractéristiques de la loi de X

$$T_n = t(X) = t(X_1, \ldots, X_n)$$

Elle est entièrement calculable à partir des données.
Exemple !

Estimateur

Soit θ le paramètre d'une loi $\mathbb{P}_\theta$, $\theta \in \Theta$.

Définition

Un estimateur $\widehat{\theta}$ de θ est une statistique à valeurs dans Θ. Cette définition s'étend au cas d'une grandeur ν calculée à partir de la loi $\mathbb{P}_\theta$: un estimateur $\widehat{\nu}_n$ de $\nu(\theta)$ est une statistique à valeurs dans $\nu(\Theta)$.

Exemple : Soit $X \sim \mathbb{P}_\theta$. L'estimateur empirique de l'espérance $\mu = \mathbb{E}_\theta(X)$ est

$$T_n = \widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

Autres exemples :
$T_n = 0$; $T_n = X_1$; $T_n = \sum_{i=1}^{[n/2]} X_{2i}/[n/2]$.
Comment choisir ?

Biais

Soit $\widehat{\nu}_n$ un estimateur de ν, défini à partir d'un n-échantillon de loi $\mathbb{P}_\theta$:

Définition

Le biais de l'estimateur $\widehat{\nu}_n$ pour estimer ν est défini par

$$B_\theta(\widehat{\nu}_n, \nu) = \mathbb{E}_\theta(\widehat{\nu}_n) - \nu$$

Si $B_\theta(\widehat{\nu}_n, \nu) = 0$, alors $\widehat{\nu}_n$ est dit non biaisé ou sans biais.

- biais = erreur systématique due au fait que $\widehat{\nu}_n$ fluctue en moyenne autour de $\mathbb{E}_\theta(\widehat{\nu}_n)$ au lieu de ν
- Il est souhaitable d'utiliser des estimateurs sans biais
- Attention ! : si T_n est un estimateur sans biais de θ, alors $\nu(T_n)$ n'est pas forcément un estimateur sans biais de $\nu(\theta)$

Variance

Soit $\widehat{\nu}_n$ un estimateur de $\nu(\theta)$, défini à partir d'un n-échantillon de loi $\mathbb{P}_\theta$:

Définition

Le variance de l'estimateur $\widehat{\nu}_n$ de ν est

$$\mathrm{Var}(\widehat{\nu}_n) = \mathbb{E}_\theta[(\widehat{\nu}_n - \mathbb{E}_\theta(\widehat{\nu}_n))^2]$$

- variance = fluctuation aléatoire de $\widehat{\nu}_n$ autour de sa valeur moyenne
- Il est souhaitable d'utiliser des estimateurs de variance la plus faible possible.

Risque d'un estimateur

Définition

Le *risque quadratique* *ou* *erreur quadratique moyenne* *de l'estimateur* $\widehat{\nu}_n$ *pour l'estimation de* ν *est l'espérance de sa perte quadratique :*

$$\nu \mapsto R_\theta(\widehat{\nu}_n, \nu) = \mathbb{E}_\theta[(\widehat{\nu}_n - \nu)^2],$$

Exemple : Le risque quadratique de l'estimateur empirique de l'espérance μ est

$$R_\theta(\bar{X}, \mu) = \mathbb{E}_\theta[(\bar{X} - \mu)^2] = Var_\theta(\bar{X}) = \frac{\sigma^2}{n}.$$

Décomposition du risque quadratique

Avec la fonction de perte quadratique

$$\begin{aligned} R_\theta(\widehat{\nu}_n, \nu) &= \mathbb{E}_\theta[(\widehat{\nu}_n - \nu)^2] \\ &= \mathrm{Var}_\theta(\widehat{\nu}_n) + (B_\theta(\widehat{\nu}_n, \nu))^2 \end{aligned}$$

Définition

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

$$R_\theta(\delta_1, \nu(\theta)) \leq R_\theta(\delta_2, \nu(\theta))$$

cette inégalité est stricte pour au moins une valeur de θ.
Un estimateur est admissible s'il n'existe aucun estimateur le dominant.

- Recherche d'estimateurs Uniformément de Variance Minimale parmi les estimateurs sans Biais

$\hookrightarrow$ A suivre dans le cours de 2A !