Statistiques Fondamentales

Simon Coste

2024-01-04

Table des matières

Organisation

Bienvenue sur la page du cours de Statistiques Fondamentales (STF8) du master Mathématiques Fondamentales et Appliquées de l'Université Paris-Cité.

Les notes de cours sont accessibles à tous. De nombreux auteurs s'y sont succédés ; je suis le dernier en date, mais les versions précédentes ont été travaillées par Clément Levrard, Stéphane Boucheron, Stéphane Gaïffas, Pierre Youssef.

- Les CM ont lieu les jeudi à (8h30 - 10h30), et les vendredi (10h45 - 12h45) **sauf le premier cours qui a lieu lundi 8 janvier à 10h45-12h45**.
- Les TD ont lieu lundi (13h45 - 16h45) et vendredi (13h30 - 15h30), de lundi 8 janvier à vendredi 16 février.
- Il y aura deux contrôles de 2h, le vendredi 26 janvier et lundi 12 février.
- L'examen a lieu le 1er mars de 13h30 à 16h30.
- Il y aura une interro de 5 minutes chaque semaine le jeudi.

Utiliser ce site

Chaque chapitre de ce livre contient une page dédiée au cours théorique, une page dédiée à quelques exemples, et une feuille d'exercices.

La saveur du cours est essentiellement mathématique et nous n'aurons pas de TP d'info ; cependant, je vous recommande vraiment d'essayer d'appliquer tout ça via votre langage de programmation favori, c'est-à-dire ~~Python~~ ~~R~~ ~~SAS~~ ~~C++~~ Julia. J'essaierai autant que possible de fournir des mini-jeux de données avec des petits challenges pour appliquer ce que vous apprenez en cours.

Ces notes sont mises en lignes et totalement accessibles via Quarto. Si vous savez comment utiliser `git`, n'hésitez pas à corriger toutes les erreurs que vous pourriez voir (et Dieu sait qu'elles seront nombreuses) via des pull requests.

1 Introduction

1.1 Un modèle-jouet

1.2 Qu'est-ce qu'un problème statistique ?

Il n'y aurait pas de statistiques s'il n'y avait pas de monde réel, et comme chacun sait, le monde réel est principalement composé de quantités aléatoires.

Un problème statistique tire donc toujours sa source d'un ensemble d'observations, disons n observations notées $x_1, \dots, x_n$; cet ensemble d'observations est appelé un *échantillon*. L'hypothèse de base de tout travail statistique consiste à supposer que cet échantillon suit une certaine loi de probabilité ; l'objectif est de trouver laquelle. Évidemment, on ne va pas partir de rien : il faut bien faire des hypothèse minimales sur cette loi. Ce qu'on appelle un *modèle statistique* est le choix d'une famille de lois de probabilités que l'on suppose pertinentes.

Définition 1.1. Formellement, choisir un modèle statistique revient à choisir trois choses :

- $\mathcal{X}$, l'espace dans lequel vit notre échantillon ;
- $\mathcal{F}$, une tribu sur $\mathcal{X}$, pour donner du sens à ce qui est observable ou non ;
- $(P_\theta)_{\theta \in \Theta}$, une famille de mesures de probabilités sur $\mathcal{X}$ indexée par $\theta \in \Theta$, où Θ est appelé espace des paramètres.

En pratique, dans ce cours, on aura toujours un échantillon $(x_1, \dots, x_n)$ où les x_i vivent dans un même espace, disons $\mathbb{R}^d$ pour simplifier. On devrait donc écrire $\mathcal{X} = \mathbb{R}^{d \times n}$; et l'on fera toujours l'hypothèse que ces observations sont indépendantes les unes des autres, et que ces observations ont la même loi de probabilité. Autrement dit, on se donnera toujours une mesure μ_θ sur $\mathbb{R}^d$ et on supposera que la loi de notre échantillon est $P_\theta = \mu_\theta^{\otimes n}$. Dans ce cadre, les observations x_i sont des *réalisations* de variables aléatoires iid de loi μ_θ.

Définition 1.2. On dit qu'un modèle statistique est identifiable si $\theta \neq \theta'$ entraîne $P_\theta \neq P_{\theta'}$.

Si l'on a bien choisi notre modèle statistique, alors il existe un « vrai » paramètre, disons $\theta_\star$, tel que les observations $x_1, \dots, x_n$ sont des réalisations de loi $\mu_{\theta_\star}$. L'objectif est alors de trouver $\theta_\star$ ou quelque information que ce soit le concernant.

Dans un modèle (identifiable), la statistique inférentielle (classique) permet de faire trois choses:

- Trouver une valeur approchée du vrai paramètre $\theta_\star$ (estimation ponctuelle).
- Donner une zone de Θ dans laquelle le vrai paramètre $\theta_\star$ a des chances de se trouver (intervalle de confiance).
- Répondre à des questions binaires sur $\theta_\star$, par exemple « $\theta_\star$ est-il positif ? ».

1.3 Qu'est-ce qu'un estimateur ?

Définition 1.3. Une *statistique* est une fonction mesurable des observations. Plus formellement, si le modèle statistique fixé est $(\mathcal{X}, \mathcal{F}, P)$, alors une statistique est n'importe quelle fonction mesurable de $(\mathcal{X}, \mathcal{F})$.

Le point important est qu'une statistique ne peut pas prendre θ en argument. Ses valeurs ne doivent dépendre du paramètre θ qu'au travers de P_θ.

Si X est une variable aléatoire et T une statistique, alors $T(X)$ est une variable aléatoire. On peut donc définir des quantités théoriques liées à T: typiquement, si X a pour loi P_θ, on peut définir la valeur moyenne de T sous le modèle P_θ comme

$$\mathbb{E}_\theta[T(X)] = \int_{\mathcal{X}} T(x) P_\theta(dx)$$

ou encore sa variance $\mathbb{E}_\theta[\mathtt{t}(X)^2] - (\mathbb{E}_\theta[\mathtt{t}(X)])^2$, etc. On peut aussi calculer la valeur de cette statistique sur l'échantillon dont on dispose, c'est-à-dire $\mathtt{t}(x_1, \dots, x_n)$. Ce qui ne se voit pas dans la définition, c'est qu'une bonne statistique devrait être facilement calculable ; à la place de *statistique*, on peut penser à *algorithme* : une bonne statistique doit pouvoir être calculée facilement par un algorithme ne prenant en entrée que les échantillons x_i.

Si le but est de deviner la valeur de θ à partir des observations, il est naturel de considérer des statistiques à valeurs dans Θ. C'est précisément la définition d'un estimateur.

Définition 1.4. Dans le modèle $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta \in \Theta})$, un estimateur de θ est une statistique à valeurs dans Θ.

En fait, on n'est pas obligés de vouloir estimer précisément θ. Peut-être qu'on veut estimer quelque chose qui dépend de θ, mais qui n'est pas θ ; disons, une fonction $q(\theta)$. Dans ce cas, un estimateur de $q(\theta)$ sera simplement une statistique à valeurs dans l'espace où vit $q(\theta)$.

1.4 Non-paramétrique

Traditionnellement, on distingue deux grands types de modèles : on dit qu'un modèle est *paramétrique* lorsque le paramètre θ vit dans un espace de dimension finie (autrement dit quand Θ est une partie de $\mathbb{R}^d$), et sinon on dit que le modèle est non-paramétrique. On verra quelques idées d'estimation non-paramétrique dans le dernier chapitre du cours.

2 Estimation de paramètre

La plupart des expériences/modèles statistiques que nous rencontrerons dans ce cours, seront de nature paramétrique, autrement dit indexés par des parties de $\mathbb{R}^d$. Dans de nombreux développements des statistiques, par exemple en estimation de densité, on travaille sur des modèles plus riches qui n'admettent pas de paramétrisation naturelle par une partie d'un espace euclidien de dimension finie. On parle pourtant de paramètre d'une distribution pour désigner ce qui devrait plutôt s'appeler une fonctionnelle. Par exemple, la moyenne, la covariance d'une distribution sur $\mathbb{R}^d$ sont des paramètres de cette distribution. Les quantiles, l'asymétrie, la kurtosis sont d'autres paramètres.

2.1 Précision d'un estimateur

On a fixé un modèle statistique $(\mathcal{X}, \mathcal{F}, (P_\theta))$, et l'on cherche à estimer θ.

Définition 2.1. Le biais de $\hat{\theta}$ est la quantité $\mathbb{E}_\theta[\hat{\theta} - \theta]$. L'estimateur est dit *sans biais* s'il est de biais nul.

Définition 2.2. Le risque quadratique de $\hat{\theta}$ est la quantité $\mathbb{E}_\theta[|\hat{\theta} - \theta|^2]$.

En pratique, on peut vouloir estimer non pas θ lui-même, mais un paramètre $\psi = \psi_\theta$ qui dépend de θ, comme $\cos(\theta)$ ou $|\theta|$ par exemple. Dans ce cas, si $\hat{\psi}$ est un estimateur de ψ alors le biais est défini par $\mathbb{E}_\theta[\hat{\psi} - \psi_\theta]$ et le risque quadratique par $\mathbb{E}_\theta[|\psi^2 - \psi_\theta|^2]$.

Théorème 2.1.

$$\mathbb{E}_\theta[|\hat{\theta} - \theta|^2] = \underbrace{\mathrm{Var}_\theta(\hat{\theta})}_{variance} + \underbrace{\left(\mathbb{E}_\theta[\hat{\theta}] - \theta\right)^2}_{carré\ du\ biais}.$$

La dépendance du risque quadratique vis à vis de la taille de l'échantillon est une question importante en statistique mathématique. Elle concerne la vitesse d'estimation (pour une suite d'expériences donnée, quelles sont les meilleures vitesses envisageables, et comment les obtenir ?).

Pour introduire la notion de consistance d'une suite d'estimateurs, nous aurons besoin des notions de convergence en probabilité et de convergence presque sûre.

Définition 2.3. Une suite de variables aléatoires X_n à valeurs dans $\mathbb{R}^k$ converge en probabilité vers une variable aléatoire X à valeurs dans $\mathbb{R}^k$, vivant sur cet espace probabilisé si et seulement si, pour tout $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0\,.$$

Définition 2.4 (consistance d'un estimateur)**.** Une suite d'estimateurs $(\hat{\theta}_n)$ est consistante pour l'estimation de θ lorsque pour n'importe quel $\theta \in \Theta$, si

$$\forall \varepsilon > 0, \qquad \lim_n P_\theta(|\hat{\theta}_n - \theta| > \varepsilon) = 0 \qquad \text{(convergence en probabilité)}.$$

La suite est fortement consistante si pour n'importe quel $\theta \in \Theta$,

$$\hat{\theta}_n \to \theta \quad P_\theta - \text{p.s.} \qquad \text{(convergence presque sûre)}.$$

2.2 Normalité asymptotique

Définition 2.5 (normalité asymtotique)**.** Soit θ un paramètre à estimer, et $\hat{\theta}_n$ une suite d'estimateurs de θ. On dit que ces estimateurs sont *asymptotiquement gaussiens* (ou *normaux*) si, après les avoir renormalisés convenablement, ils convergent en loi vers une loi gaussienne. Autrement dit, s'il existe une suite a_n de nombres réels tels que

$$a_n(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{\text{loi}} N(0, \Sigma)$$

où Σ est une matrice de covariance qui dépend peut-être de θ — pour éviter les cas dégénérés, on demande à ce que Σ soit non-nulle.

La normalité asymptotique n'est pas intéressante en elle-même, l'idée est de chercher le comportement asymptotique de la statistique recentrée pour pouvoir en déduire des garanties en terme de risque asymptotique ou d'intervalle de confiance. Le théorème central limite indique que le comportement asymptotique normal est relativement fréquent.

2.3 Deux outils sur la normalité asymptotique

Théorème 2.2 (Lemme de Slutsky)**.** *Soit (X_n) une suite de variables aléatoire qui converge en loi vers X et (Y_n) une suite de variables aléatoires qui converge en probabilité (ou en loi) vers une constante c. Alors, le* couple *(X_n, Y_n) converge en loi vers (X, c) ; autrement dit, pour toute fonction continue bornée φ,*

$$\mathbb{E}[\varphi(X_n, Y_n)] \to \mathbb{E}[\varphi(X, c)].$$

Théorème 2.3 (Delta-méthode)**.** *Soit (X_n) une suite de variables aléatoires réelles telle que $\sqrt{n}(X_n - \alpha)$ converge en loi vers $N(0, \sigma^2)$. Pour toute fonction $g : \mathbb{R} \to \mathbb{R}$ dérivable en α (de dérivée non nulle en α), on a*

$$\sqrt{n}(g(X_n) - g(\alpha)) \xrightarrow[n\to\infty]{loi} N(0, g'(\alpha)^2\sigma^2).$$

Plus généralement, si les X_n sont à valeurs dans $\mathbb{R}^d$ et que $\sqrt{n}(X_n - \alpha) \to N(0, \Sigma)$, alors pour toute fonction $g : \mathbb{R}^d \to \mathbb{R}^k$ on a

$$\sqrt{n}(g(X_n) - g(\alpha)) \xrightarrow[n\to\infty]{\text{loi}} N(0, Dg(\alpha)\Sigma Dg(\alpha)^\top)$$

où $Dg(x)$ est la matrice jacobienne de g en x.

3 Intervalles de confiance

a

4 Test d'hypothèse

5 Économétrie

6 Théorie de l'information

7 Estimation de densité

Et après ?

nasuitenasute