Statistiques Fondamentales

Simon Coste

2025-01-08

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.

Le cours reprendra l'année prochaine, en janvier 2025.

- **Calcul de la note finale**. Trois quantités entrent en jeu : la note de contrôle continu $c \in [0, 10]$, la note du partiel $p \in [0, 20]$, et la note que vous obtiendrez à l'examen, $e \in [0, 20]$. La note finale $f \in [0, 20]$ est donnée par la formule

$$f = 20^{\theta} \left(\frac{\max\{p, e\} + e}{2} \right)^{1-\theta}$$

  où $\theta = 0.3 \times c/10$. C'est une moyenne géométrique qui ne pénalise personne, et qui ne peut que tirer votre note vers le haut si vous avez une note de contrôle continu c strictement positive.

Utiliser ce site

Chaque chapitre de ce livre contient une page dédiée au cours théorique, et contiendra dans un futur proche une page d'exercices.

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

Les outils des statistiques furent créés pour analyser des phénomènes quantitatifs dans lesquels la présence de *bruit* ou de *hasard* rendait l'analyse classique moins opérante. Il peut typiquement s'agir de problèmes dans lesquels de nombreuses données indépendantes ont été générées par le même phénomène. Dans cette section, nous allons développer un exemple pour bien comprendre les questions qui se posent et la façon de les résoudre, puis nous poserons quelques bases qui nous permettront d'utiliser le langage des mathématiques.

1.1 Un exemple pour fixer les idées

Une grande enseigne de distribution possède $n = 100$ magasins identiques, qui génèrent chaque année un chiffre d'affaire annuel (CA, en millions d'euros). Ce chiffre oscille autour d'une valeur de référence μ. Cette valeur n'est pas observée ; ce qui est observé, ce sont tous les chiffres d'affaires des n magasins, qui fluctuent tous autour de la vraie valeur μ. Ces fluctuations proviennent de nombreuses sources : erreurs comptables, perturbations des ventes dues aux fournisseurs ou aux prix, etc. Ce qu'on observe, c'est donc des chiffres $x_1, \dots, x_n$ qui ne sont pas tous égaux ; comment avoir une idée de la véritable valeur de μ ?

Estimation. Évidemment, la moyenne empirique

$$\bar{x}_n = \frac{x_1 + \cdots + x_n}{n}$$

vient naturellement à l'esprit. En faisant le calcul, on trouve $\bar{x}_n \approx 21,6$. Cette valeur est une *estimation* du CA moyen μ. Ce chiffre peut être utilisé par l'enseigne, par exemple pour jauger la rentabilité d'un possible plan d'ouverture de nouveaux magasins.

Précision. On pourrait se demander à quel point cette estimation est précise ou, disons, essayer de quantifier l'erreur possible qu'on fait si l'on dit que μ est égal à 21,6 millions d'euros. Cela nécessite de faire quelques hypothèses sur le hasard qui génère les fluctuations des x_i autour de μ. Ces fluctuations observées au cours de l'année proviennent de l'agrégation de toutes les fluctuations quotidiennes, lesquelles sont à peu près indépendantes, et pour cette raison on peut supposer (pour commencer) que ces fluctuations sont gaussiennes et ont à peu près la même variance, disons $\sigma^2 = 1$. Comme on a supposé que les x_i sont des réalisations d'une loi gaussienne $N(\mu, 1)$, alors on sait que $\bar{x}_n$ est la réalisation d'une loi $N(\mu, 1/n)$, ou encore que $\bar{x}_n - \mu$ est la réalisation d'une gaussienne centrée de variance $1/n$. Les lois gaussiennes sont bien connues ; par exemple, avec probabilité supérieure à 99%, une gaussienne $N(0, \sigma^2)$ est comprise entre les valeurs $-2,96\sigma$ et $2,96\sigma$. Autrement dit, il y a 99% de chances pour que le nombre $|\bar{x} - \mu|$, qui représente l'erreur d'estimation, soit plus petite que $2,96/\sqrt{n} = 2,96/10 \approx 0,3$.

Ce dernier raisonnement peut être vu d'une autre façon. Dire que $\bar{x}_n$ et μ ne diffèrent pas de plus de $0,3$, c'est équivalent à dire que μ appartient à l'intervalle $[\bar{x} - 0,3, \bar{x} + 0,3]$. En d'autres termes, avec une probabilité supérieure à 99%, le vrai CA μ de chaque magasin se situe entre $21,3$ et $21,9$. Cela laisse tout de même une chance de 1% que le paramètre μ ne soit pas dans cette région.

Tests. Il existe encore un autre point de vue sur ce problème. Par exemple, le conseil d'administration de la firme veut s'assurer que le dirigeant a bien tenu sa promesse selon laquelle le CA de chaque magasin était supérieur à 21 millions d'euros. La valeur exacte de μ n'est pas le plus important : ce qui nous intéresse maintenant, c'est plutôt d'être sûrs que μ n'est pas inférieur au seuil de 21. Le dirigeant, fin statisticien, effectue alors un raisonnement par l'absurde *en probabilité*. Supposons que le CA μ soit effectivement égal à 21 (ou même, inférieur). Alors, par les mêmes calculs que ci-dessus, cela voudrait dire qu'avec 99% de chances, $\bar{x}_n$ et 21 ne devraient pas différer de plus de $0,3$; autrement dit, que $\bar{x}_n$ devrait se situer entre $20,7$ et $21,3$. Ce n'est pas le cas, puisque $\bar{x}_n = 21,6$. Si μ est réellement plus petit que 21, alors ce qu'on a observé est extrêmement peu probable. Par contraposée probabiliste, il est raisonnable de rejeter l'hypothèse selon laquelle μ est inférieur à 21.

Les trois points de vue donnés ci-dessus sont en quelque sorte les piliers de l'analyse statistique. L'estimation consiste à deviner une valeur cachée dans du bruit ; les intervalles de confiance consistent à donner une région dans laquelle se trouve cette valeur ; les tests d'hypothèse permettent de raisonner de façon logique sur cette valeur.

L'objectif du cours de statistiques de quantifier l'incertitude liée au hasard dans chacun de ces objectifs. Comme dans les exemples donnés ci-dessus, c'est un ensemble de méthodes scientifiques qui s'appuient sur la théorie des probabilités ; dans ce cours, on fera des *hypothèses* sur le hasard qui est en jeu, et on en tirera des conséquences *probables* sur le modèle sous-jacent. En théorie des probabilités, le jeu est plutôt inverse : partant d'un modèle probabiliste fixé, on essaie de déterminer quel sera le comportement des réalisations de ce modèle. Il semble difficile de faire l'un sans l'autre.

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 ;
- $\mathscr{F}$, 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. On écrira fréquemment $\mathbb{E}_\theta$ ou Var_θ pour désigner des espérances, variances, etc., calculées avec la loi P_θ.

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 p_θ sur $\mathbb{R}^d$ et on supposera que la loi de notre échantillon est $P_\theta = p_\theta^{\otimes n}$. Dans ce cadre, les observations x_i sont des *réalisations* de variables aléatoires X_i iid de loi p_θ.

Il faut prendre garde à distinguer les variables aléatoires X_i, qui sont des objets théoriques, de leurs réalisations x_i, qui, elles, sont bel et bien observées.

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 $p_{\theta_\star}$. 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 $(\mathcal{X}, \mathscr{F}, P)$, alors une statistique est n'importe quelle fonction mesurable de $(\mathcal{X}, \mathscr{F})$.

1) Le premier 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.

2) Le second point important est que, 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 $\mathbb{E}_\theta[T(X)^2] - (\mathbb{E}_\theta[T(X)])^2$, etc. On peut aussi calculer la valeur de cette statistique sur l'échantillon dont on dispose, c'est-à-dire $T(x_1, \dots, x_n)$. Par exemple, la moyenne empirique d'un n-échantillon réel est la fonction $T : (a_1, \dots, a_n) \to n^{-1}(a_1 + \cdots + a_n)$. Si les x_i sont des réalisations des variables aléatoires X_i, alors $T(x_1, \dots, x_n)$ est une réalisation de la variable aléatoire $T(X_1, \dots, X_n)$.

3) 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 $\varphi(\theta)$. Dans ce cas, un estimateur de $\varphi(\theta)$ sera simplement une statistique à valeurs dans l'espace où vit $\varphi(\theta)$.

1.4 Points de vue

Inférence paramétrique. La plupart des expériences/modèles statistiques que nous rencontrerons dans ce cours, seront de nature dite *paramétrique*, autrement dit indexés par des parties de $\mathbb{R}^d$. Le mot "paramètre" est en lui-même trompeur : on parle souvent 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.

Statistique non paramétrique. Tous les modèles ne sont pas paramétriques au sens ci-dessus : 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. C'est ce qu'on appelle l' *estimation non-paramétrique*. Nous y reviendrons au dernier chapitre.

Statistique bayésienne. En statistique paramétrique, les paramètres θ déterminent le hasard qui génère les observations x_i. La statistique bayésienne consiste à renverser le point de vue, et à rendre le paramètre θ lui-même aléatoire ; sa loi, appelée *prior*, mesure "le degré de connaissance a priori" qu'on en a. La règle de Bayes explique comment cette loi est modifiée par les observations. C'est un point de vue qui ne sera pas abordé dans ce cours.