Statistiques Fondamentales

Simon Coste

2024-01-18

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, et contiendra dans un futur proche une page 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

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 ;
- $\mathcal{F}$, 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. 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 $(\mathcal{X}, \mathcal{F}, P)$, alors une statistique est n'importe quelle fonction mesurable de $(\mathcal{X}, \mathcal{F})$.

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.

2 Estimation de paramètre

On fixe un modèle statistique $(\mathcal{X}, \mathcal{F}, (P_\theta))$, et l'on cherche à estimer le paramètre θ ou un autre paramètre qui dépend de θ. Dans ce chapitre, on explique comment juger la qualité d'un estimateur, et l'on donne une technique générale pour construire de bons estimateurs dans des situations assez naturelles.

2.1 Précision d'un estimateur

Définition 2.1 (Biais , risque quadratique)**.**

- 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.
- 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 $\mathbb{E}_\theta[|\hat{\psi}^2 - \psi_\theta|^2]$.

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 ?).

Théorème 2.1 (Décomposition biais-variance)**.**

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

2.2 Consistances

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 2.2. 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 2.3 (consistance d'un estimateur)**.** Une suite d'estimateurs $(\hat{\theta}_n)$ est consistante pour l'estimation de θ lorsque, pour tout $\theta \in \Theta$,

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

La suite est fortement consistante si, pour tout $\theta \in \Theta$,

$$\hat{\theta}_n \to \theta \quad P_\theta - \text{p.s.}$$

2.3 Normalité asymptotique

Lorsqu'un estimateur est consistant, on peut se demander à quoi ressemblent ses fluctuations autour de sa valeur limite. Le théorème central limite indique que le comportement asymptotique gaussiens est relativement fréquent, et beaucoup d'estimateurs sont des sommes de réalisations de variables indépendantes.

Définition 2.4 (normalité asymptotique)**.** 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.

2.4 Deux outils sur la normalité asymptotique

La normalité asymptotique n'est pas intéressante en elle-même ; l'idée est plutôt 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. Nous utiliserons cela de nombreuses fois dans la suite ; la normalité asymptotique sera par exemple la clé de la construction de nombreux intervalles de confiance. Aussi, prouver que des estimateurs sont asymptotiquement normaux est une tache importante, qui est grandement simplifiée par les deux outils suivants.

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,*

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

Démonstration. Fixons une fonction test φ continue à support compact, donc bornée par un certain M. Il faut montrer que $\mathbb{E}[\varphi(X_n, Y_n) - \varphi(X, c)]$ tend vers zéro. L'intégrande est égal à la somme de $A = \varphi(X_n, Y_n) - \varphi(X_n, c)$ et de $B = \varphi(X_n, c) - \varphi(X, c)$.

Comme X_n tend en loi vers X et que $t \to \varphi(t, c)$ est continue bornée, l'espérance de B tend vers zéro. Il faut donc montrer que l'espérance de A tend vers zéro. On fixe un $\varepsilon > 0$.

- Par le théorème de Heine, φ est uniformément continue : il existe $\delta > 0$ tel que $|(x, y) - (x', y')| < \delta$ entraîne que $|\varphi(x, y) - \varphi(x', y')| < \varepsilon/2$.
- On introduit l'événement $\{|Y_n - c| \leqslant \delta\}$. Par le point précédent, sur cet événement on a $|A| < \varepsilon/2$. Hors de cet événement, on peut toujours borner $|A|$ par $2M$. On a donc

$$|\mathbb{E}A| \leqslant \mathbb{P}(|Y_n - c| \leqslant \delta)\varepsilon/2 + \mathbb{P}(|Y_n - c| > \delta)2M.$$

- Comme Y_n converge en probabilité vers c, lorsque n est assez grand on a $\mathbb{P}(|Y_n - c| > \delta) < \varepsilon/4M$.

- En regroupant tout ce qui a été dit, on obtient bien $|\mathbb{E}A| \leqslant \varepsilon$ dès que n est assez grand, ce qui montre bien que $\mathbb{E}A \to 0$.

□

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 application $g : \mathbb{R}^d \to \mathbb{R}^k$, la suite $\sqrt{n}(g(X_n) - g(\alpha))$ converge en loi vers

$$N(0, Dg(\alpha)\Sigma Dg(\alpha)^\top)$$

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

Démonstration. À écrire.

□

2.5 La méthode des moments

Il existe plusieurs techniques générales pour *construire* des estimateurs. La méthode des moments est la plus naturelle, et donne beaucoup des estimateurs avec de bonnes propriétés.

Dans un modèle statistique, supposons qu'on dispose d'une statistiques intégrable T (pas forcément réelle), dont la moyenne n'est pas le paramètre θ lui-même, mais plutôt une *fonction* de θ :

$$\mathbb{E}_\theta[T(X)] = \varphi(\theta).$$

C'est cette fonction φ qu'on appelle *moment*. Typiquement,

- la moyenne d'une loi $\mathcal{E}(\theta)$ n'est pas θ mais $1/\theta$.
- la moyenne d'une loi log-normale de paramètres $(0, \sigma^2)$ est $e^{\sigma^2/2}$.

Prenons la moyenne empirique associée à cet estimateur, $\bar{T}_n$. Par la loi des grands nombres,

$$\bar{T}_n = \frac{1}{n}\sum_{i=1}^{n} T(X_i) \to \varphi(\theta) \qquad P_\theta - ps,$$

ce qui permet d'estimer $\varphi(\theta)$. Peut-on alors estimer θ ? Si la fonction φ est inversible et si $\bar{T}_n$ appartient presque sûrement à l'ensemble de définition de φ^{-1}, alors $\varphi^{-1}(\bar{T}_n)$ est bien définie. Pour qu'en plus cette quantité converge presque sûrement vers θ, il faut s'assurer que φ^{-1} est continue. C'est par exemple le cas lorsque l'ensemble des paramètres Θ est un ouvert, et que φ est un difféomorphisme sur son image — une situation si fréquente qu'elle mérite son propre théorème, et si agréable qu'elle garantit que l'estimateur associé est asymptotiquement normal.

Théorème 2.4 (Estimation par moments)**.** *Sous l'hypothèse mentionnée ci-dessus (la fonction φ est un difféomorphisme), l'estimateur*

$$\hat{\theta}_n = \varphi^{-1}(\bar{T}_n)$$

est presque sûrement bien défini pour tout n suffisamment grand ; il est également consistant pour l'estimation de θ. En outre, si T est de carré intégrable, cet estimateur est asymptotiquement normal, au sens où $\sqrt{n}(\hat{\theta}_n - \theta)$ converge en loi vers une gaussienne centrée de matrice de covariance

$$(D\varphi(\theta))^{-1}\mathrm{Cov}_\theta(T)(D\varphi(\theta)^\top)^{-1}.$$

Démonstration. La première partie a essentiellement été démontrée un peu plus haut. Pour la seconde, il faut d'abord remarquer que si T est de carré intégrable, alors $\sqrt{n}(\bar{T}_n - \varphi(\theta))$ converge vers une loi $N(0, \mathrm{Cov}_\theta(T))$ par le TCL. Une simple application de la delta-méthode (Théorème 2.3) donne alors le résultat, puisque la matrice jacobienne de φ^{-1} en $\varphi(\theta)$ n'est autre que l'inverse de la matrice jacobienne de φ en θ.

□

2.6 Deux estimateurs importants

Deux estimateurs sont omniprésents en statistique : la moyenne empirique et la variance empirique. Ils sont pertinents dans n'importe quel modèle où les observations sont des réalisations de variables iid possédant une moyenne μ et une variance σ^2.

La moyenne empirique est définie par

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Il est évident que $\mathbb{E}[\bar{X}_n] = \mathbb{E}[X] = \mu$. Cet estimateur est donc toujours sans biais, et son risque quadratique est égal à sa variance, c'est-à-dire $\frac{\sigma^2}{n}$.

L'estimateur de la variance empirique est défini comme

$$\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Théorème 2.5. *L'estimateur $\hat{\sigma}_n^2$ est sans biais.*

Démonstration. À écrire.

□

3 Exercices

3.1 Questions

1. Montrer que la convergence en loi vers une constante implique la convergence en probabilité.
2. Montrer que, si un modèle statistique n'est pas identifiable, alors il ne peut exister aucun estimateur convergent.
3. Trouver un couple de variables aléatoires (X_n, Y_n) tel que X_n converge en loi et Y_n converge en loi, mais le couple ne converge pas en loi.
4. On observe un échantillon de lois de Poisson de paramètre λ, que l'on estime par la moyenne empirique. Calculer le risque quadratique de cet estimateur.
5. Quelle est la loi d'une somme de lois de Bernoulli indépendantes ? L'écart-type ?

3.2 Exercices

Exercice 3.1 (Variance empirique)**.** On se donne $Y_1, \dots, Y_n$, i.i.d de moyenne μ et variance σ^2.

1. On suppose μ connu. Donner un estimateur non biaisé de σ^2.
2. On suppose μ inconnu. Calculer l'espérance de $\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$. En déduire un estimateur non biaisé de σ^2.

Exercice 3.2 (Estimation de masse)**.** Au cours de la seconde guerre mondiale, l'armée alliée notait les numéros de série $X_1, \dots, X_n$ de tous les tanks nazis capturés ou détruits, afin d'obtenir un estimateur du nombre total N de tanks produits.

1. Proposer un modèle pour le tirage de $X_1, \dots, X_n$.
2. Calculer l'espérance de $\bar{X}_n$. En déduire un estimateur non biaisé de N. Indication: la loi de n tirages sans remise est échangeable.
3. Étudier la loi de $X_{(n)}$ et en déduire un estimateur non biaisé de N.
4. Proposer deux intervalles de confiance de niveau $1-\alpha$ de la forme $[aS, bS]$ avec $a, b \in \mathbb{R}$ et S une statistique. On pourra utiliser le fait que l'inégalité de Hoeffding s'applique également aux tirages sans remise.

Selon Ruggles et Broodie (1947, JASA), la méthode statistique a fourni comme estimation une production moyenne de 246 tanks/mois entre juin 1940 et septembre 1942. Des méthodes d'espionnage traditionnelles donnaient une estimation de 1400 tanks/mois. Les chiffres officiels du ministère nazi des Armements ont montré après la guerre que la production moyenne était de 245 tanks/mois.

Exercice 3.3 (Lois uniformes (1))**.** On considère $(X_1, \dots, X_n)$ un échantillon de loi uniforme sur $]\theta, \theta+1[$.

1. Donner la densité de la loi de la variable $R_n = X_{(n)} - X_{(1)}$, où $X_{(1)} = \min(X_1, \dots, X_n)$ et $X_{(n)} = \max(X_1, \dots, X_n)$.
2. Étudier les différents modes de convergence de R_n quand $n \to \infty$.
3. Étudier le comportement en loi de $n(1 - R_n)$ quand $n \to \infty$.

Exercice 3.4 (Lois uniformes (2))**.** Soit $X_1, \dots, X_n$ un échantillon de loi $\mathcal{U}([0, \theta])$, avec $\theta > 0$. On veut estimer θ.

1. Déterminer un estimateur de θ à partir de $\bar{X}_n$.
2. On considère l'estimateur $X_{(n)} = \max_{1 \leq i \leq n} X_i$. Déterminer les propriétés asymotptiques de ces estimateurs.
3. Comparer les performances des deux estimateurs.

Exercice 3.5 (Lois Gamma)**.** La loi Gamma $\Gamma(\alpha, \beta)$ de paramètres $\alpha, \beta > 0$ a pour densité

$$x \mapsto \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha - 1} \exp(-\beta x), \quad x > 0.$$

On se donne un échantillon $(X_1, \dots, X_n)$ de loi $\Gamma(\alpha, \beta)$ et on chercche à estimer les paramètres.

1. On suppose le paramètre β connu. Proposer un estimateur de α par la méthode des moments.
2. On suppose à présent que les deux paramètres α, β sont inconnus. Proposer un estimateur de (α, β) par la méthode des moments.

Exercice 3.6 (Lois de Gumbel)**.** La loi de Gumbel (centrale) de paramètre β a pour fonction de répartition $F(x) = e^{-e^{-x/\beta}}$. On observe un échantillon de lois de Gumbel et l'on cherche à estimer β.

1. Calculer la densité des lois de Gumbel, ainsi que leur moyenne et variance [indice : 0.57721...]
2. En déduire un estimateur convergent dont on calculera le risque quadratique et les propriétés asymptotiques.

Exercice 3.7 (Lois de Yule-Simon)**.** Une variable aléatoire X suit la loi de Yule-Simon de paramètre $\rho > 0$ lorsque $\mathbb{P}(X = n) = \rho B(n, 1 + \rho)$, où $n \geqslant 1$ et B est la fonction beta.

1. Montrer que si $\rho > 1$, alors $\mathbb{E}[X] = \rho/(\rho - 1)$.
2. Trouver un estimateur de ρ et donner ses propriétés.

4 Intervalles de confiance

4.1 Principe

Dans un modèle statistique, l'estimation du paramètre d'intérêt θ par intervalles de confiance consiste à spécifier un intervalle calculable à partir des données, et qui contient θ avec grande probabilité : en d'autres termes, une *région de confiance* pour θ.

Pour simplifier, on supposera d'abord que θ est un paramètre réel.

Définition 4.1 (intervalle de confiance)**.** Un intervalle de confiance de niveau $1-\alpha$ est un intervalle $I = [A, B]$ dont les bornes A, B sont des statistiques, et tel que pour tout θ,

$$P_\theta(\theta \in I) \geqslant 1 - \alpha.$$

Un intervalle de confiance de niveau asymptotique $1-\alpha$ est une *suite* d'intervalles $I_n = [A_n, B_n]$ dont les bornes A_n, B_n sont des statistiques, et tels que pour tout n,

$$P_\theta(\theta \in I_n) \geqslant 1 - \alpha.$$

Le terme « niveau » désigne $1-\alpha$; la vocation de ce nombre est d'être proche de 1, typiquement 99%. Le nombre α est parfois appelé « erreur », « marge d'erreur » ou encore « niveau de risque » ; la vocation de ce nombre est d'être proche de zéro, typiquement 1%.

Il n'y a rien d'autre à savoir sur les intervalles de confiance ; tout l'art de la chose consiste à savoir les construire. Commençons par des exemples essentiels à plusieurs titres : le cas d'un échantillon gaussien, et le cas de lois de Bernoulli.

4.2 Exemples gaussiens

On dispose de variables aléatoires $X_1, \dots, X_n$ de loi $N(\mu, \sigma^2)$. On va donner des intervalles de confiance pour l'estimation des paramètres μ et σ dans plusieurs cas de figure.

4.2.1 Estimation de μ

Lorsque σ est connue.

Nous avons déjà vu que la moyenne empirique $\bar{X}_n$ est un estimateur sans biais de μ. Or, nous savons aussi la loi *exacte* de $\bar{X}_n$, qui est $N(\mu, \sigma^2/n)$. Autrement dit,

$$\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \sim N(0, 1). \tag{4.1}$$

Dans cette équation, on a trouvé une variable aléatoire dont la loi ne dépend plus de μ. Il est donc possible de déterminer un intervalle dans lequel elle fluctue à l'aide des quantiles de la loi normale, qui sont étudiés dans Section 5.1. Si l'on se donne une marge d'erreur $\alpha = 1\%$, alors

$$\mathbb{P}((\sqrt{n}/\sigma)|\bar{X}_n - \mu| > z_{0.99}) = 1\%$$

où $z_{0.99} \approx 2.32$. Or,

$$\frac{\sqrt{n}}{\sigma}|\bar{X}_n - \mu| > z_{0.99} \tag{4.2}$$

est équivalent à

$$\bar{X}_n - \frac{z_{0.99}\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{X}_n + \frac{z_{0.99}\sigma}{\sqrt{n}}. \tag{4.3}$$

Le passage de Équation 4.2 à Équation 4.3 est souvent appelé *pivot* et sert à passer d'un intervalle de fluctuation à un intervalle de confiance.

Nous avons donc les deux bornes de notre intervalle de confiance :

$$A = \bar{X}_n - \frac{z_{0.99}\sigma}{\sqrt{n}}$$

$$B = \bar{X}_n + \frac{z_{0.99}\sigma}{\sqrt{n}}.$$

Ces deux quantités sont bien des statistiques, car σ est connu. De plus, nous venons de montrer que $P_\mu(\mu \in [A, B]) = 99\%$. Ici, le choix de la marge d'erreur $\alpha = 1\%$ ne jouait aucun rôle particulier ; ainsi, un intervalle de confiance de niveau $1 - \alpha$ pour l'estimation de μ est donné par

$$\left[\bar{X}_n - \frac{z_{1-\alpha}\sigma}{\sqrt{n}} \quad ; \quad \bar{X}_n + \frac{z_{1-\alpha}\sigma}{\sqrt{n}}\right]. \tag{4.4}$$

Lorsque σ est inconnue.

Lorsque σ n'est pas connu, les bornes A, B ci-dessus ne sont pas des statistiques, car elles dépendent de σ. Heureusement, on peut estimer σ sans biais via l'estimateur

$$\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Que se passe-t-il si, dans la statistique Équation 4.1, on remplace σ par son estimation $\hat{\sigma}_n^2$? On obtient la statistique dite *de Student*,

$$T_n = \frac{\sqrt{n}}{\sqrt{\hat{\sigma}_n^2}}(\bar{X}_n - \mu). \tag{4.5}$$

Sa loi n'est plus une loi gaussienne, mais une loi de Student à $n-1$ paramètres de liberté $\mathcal{T}(n-1)$: le calcul de la densité est fait en détails dans Section 5.2.3 - Section 5.2.4. Les quantiles des lois de Student ont été calculés avec précision. On notera $t_{k,\alpha}$ le quantile symétrique de niveau α de $\mathcal{T}(k)$. Alors,

$$P_{\mu,\sigma^2}(|T_n| > t_{n-1,\alpha}) \leqslant \alpha.$$

Par le même raisonnement que tout à l'heure, l'inégalité

$$\left|\frac{\sqrt{n}}{\hat{\sigma}_n^2}(\bar{X}_n - \mu)\right| > t_{n-1,\alpha}$$

est équivalente à

$$\bar{X}_n - \frac{t_{n-1,\alpha}\hat{\sigma}_n^2}{\sqrt{n}} \leqslant \mu \leqslant \bar{X}_n + \frac{t_{n-1,\alpha}\hat{\sigma}_n^2}{\sqrt{n}}.$$

et les deux côtés de ces inégalités sont des statistiques; en les notant A, B, on a bien trouvé un intervalle de confiance de niveau α, c'est-à-dire tel que $P_{\mu,\sigma^2}(\mu \in [A, B]) = \alpha$. Cet intervalle de confiance est d'une grande importance en pratique et mérite son propre théorème. Il est dû à William Gosset.

Théorème 4.1 (Intervalle de Student)**.** *Un intervalle de confiance de niveau* $1-\alpha$ *pour l'estimation de* μ *lorsque* σ *n'est pas connue est donné par*

$$\left[\bar{X}_n - \frac{t_{n-1,1-\alpha}\hat{\sigma}_n}{\sqrt{n}} \quad ; \quad \bar{X}_n + \frac{t_{n-1,1-\alpha}\hat{\sigma}_n}{\sqrt{n}}\right].$$

4.2.2 Estimation de σ

Supposons maintenant qu'on désire estimer la variance σ^2.

Lorsque μ est connue.

En supposant que μ est connue, l'estimateur des moments le plus naturel pour estimer σ^2 est évidemment

$$\tilde{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

Comme les $(X_i - \mu)/\sigma$ sont des variables aléatoires gaussiennes centrées réduites, l'estimateur $\tilde{\sigma}_n^2 \times (n/\sigma^2)$ est une somme de n gaussiennes standard indépendantes. La loi de cette statistique est connue : c'est une loi du chi-deux à n paramètres de liberté comme démontré dans Section 5.2.2. Cette loi n'est pas symétrique, puisqu'elle est supportée sur $[0, \infty[$. On note souvent $k_{n,\alpha}^-$ et $k_{n,\alpha}^+$ les nombres les plus éloignées possible (ils exisent) tels que $\mathbb{P}(k_{n,\alpha}^- < \chi^2(n) < k_{n,\alpha}^+) = 1 - \alpha$. Ainsi,

$$P_{\sigma^2}(k_{n,\alpha}^- < \frac{n\tilde{\sigma}_n^2}{\sigma^2} < k_{n,\alpha}^+) = \alpha.$$

En pivotant comme dans les exemples précédents, on obtient que l'intervalle

$$\left[\frac{n\tilde{\sigma}_n^2}{k_{n,\alpha}^+} \quad ; \quad \frac{n\tilde{\sigma}_n^2}{k_{n,\alpha}^-}\right]$$

est un intervalle de confiance de niveau α pour σ^2.

Lorsque μ est inconnue.

Cette fois, on utilise l'estimateur déjà évoqué plus tôt, à savoir

$$\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

La loi de $(n-1)\hat{\sigma}_n^2/\sigma^2$ est encore une loi du chi-deux, mais à $n-1$ paramètres de liberté. Ainsi, le même raisonnement que ci-dessus donne l'intervalle de confiance de niveau α suivant :

$$\left[\frac{(n-1)\hat{\sigma}_n^2}{k_{n-1,\alpha}^+} \quad ; \quad \frac{(n-1)\hat{\sigma}_n^2}{k_{n-1,\alpha}^-}\right].$$

4.3 Exemples asymptotiques

4.3.1 Estimation du paramètre p dans un modèle de Bernoulli.

Soient $X_1, \dots, X_n$ des variables indépendantes de loi $\mathcal{B}(p)$, dont on cherche à estimer le paramètre $p \in]0,1[$. Un estimateur naturel est donné par la moyenne empirique, $\hat{p}_n = (X_1 + \cdots + X_n)/n$. Cet estimateur est non biaisé et son risque quadratique est égal à $p(1-p)/n$. De plus, la loi de $\hat{p}_n$ est connue : $n\hat{p}_n \sim \mathrm{Bin}(n,p)$. Par conséquent, si l'on connaît les quantiles de $\mathcal{B}in(n,p) - p$, on pourra construire des intervalles de confiance de niveau $1-\alpha$. Ces quantiles peuvent être calculés par des méthodes numériques, mais il existe des façons plus simples de faire.

Inégalité BT. L'inégalité de Bienaymé-Tchebychev dit que

$$P_p(|\hat{p}_n - p| > t) \leqslant \frac{p(1-p)}{nt^2}. \tag{4.6}$$

Si l'on choisit

$$t = \sqrt{\frac{p(1-p)}{n\alpha}},$$

cette probabilité est plus petite que α. En pivotant, on en déduit que l'intervalle $[\hat{p_n} \pm \sqrt{p(1-p)/n\alpha}]$ contient p avec une probabilité supérieure à $1-\alpha$. Mais les bornes de cet intervalle ne sont pas des statistiques, car elles dépendent de p ! Fort heureusement, on sait que p est entre 0 et 1, ce qui entraîne que $p(1-p)$ est plus petit que $1/4$, donc l'intervalle ci-dessus est contenu dans l'intervalle plus grand

$$\left[\hat{p}_n \pm \frac{1}{2\sqrt{n\alpha}}\right].$$

Ce dernier est bien un intervalle de confiance de niveau $1-\alpha$ pour l'estimation de p.

TCL. On a mentionné que les quantiles des lois binomiales pourraient être calculés ; or, ils peuvent également être approchés grâce au théorème central-limite. Celui-ci dit que

$$\frac{\sqrt{n}(\hat{p}_n - p)}{\sqrt{p(1-p)}} \to N(0,1). \tag{4.7}$$

Si z_α est le quantile symétrique d'ordre α de $N(0,1)$, alors on en déduit que

$$\mathbb{P}\left(\left|\frac{\sqrt{n}(\hat{p}_n - p)}{\sqrt{p(1-p)}}\right| > z_\alpha\right) \to \alpha.$$

En pivotant, on voit alors que l'intervalle

$$\left[\hat{p}_n \pm z_\alpha \sqrt{p(1-p)/n}\right]$$

contient p avec une probabilité *qui tend lorsque* $n \to \infty$ *vers* $1-\alpha$. Là encore, cet intervalle n'est pas un intervalle de confiance. On pourrait utiliser deux techniques.

1. Comme tout à l'heure, l'intervalle ci-dessus est contenu dans l'intervalle plus grand $[\hat{p}_n \pm z_\alpha/2\sqrt{n}]$ qui est un intervalle de confiance *asymptotique* de niveau $1-\alpha$.

2. Il y a plus fin. Nous savons par la loi des grands nombres que $\hat{p}_n \to p$ en probabilité. Ainsi, $\sqrt{\hat{p}_n(1-\hat{p}_n)} \to \sqrt{p(1-p)}$ en probabilité. Le lemme de Slutsky nous assure alors que dans Équation 4.8, on peut remplacer le dénominateur par $\sqrt{\hat{p}_n(1-\hat{p}_n)}$ pour obtenir

$$\frac{\sqrt{n}(\hat{p}_n - p)}{\sqrt{\hat{p}_n(1-\hat{p}_n)}} \to N(0,1). \tag{4.8}$$

Le reste du raisonnement est identique, et l'on obtient l'intervalle de confiance asymptotique de niveau $1-\alpha$ suivant :

$$\left[\hat{p}_n \pm z_\alpha \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right]$$

Hoeffding. L'inégalité de Bienaymé-Tchebychev n'est pas très fine. Il existe de nombreuses autres inégalités de concentration : l'inégalité de Hoeffding (Théorème 5.4) concerne les variables bornées, comme ici où les X_i sont dans $[0,1]$. Cette inégalité dit que

$$\mathbb{P}(|\hat{p}_n - p| > t) \leqslant 2e^{-2nt^2}.$$

Le choix

$$t = \sqrt{\frac{1}{2n}\ln\left(\frac{1}{\alpha}\right)}$$

donne une probabilité inférieure à α, et fournit donc l'intervalle de confiance **non-asymptotique** de niveau $1-\alpha$ suivant :

$$\left[\bar{X}_n \pm \frac{\ln(1/\alpha)}{\sqrt{2n}}\right].$$

4.3.2 Estimation de moyenne dans un modèle non-gaussien.

Les deux techniques ci-dessus n'ont rien de spécifique au cas de variables de Bernoulli. En fait, elles s'appliquent à tout modèle statistique iid dont on cherche à estimer la moyenne μ, pourvu que la variance existe.

La première méthode utilisant Bienaymé-Tchebychev nécessite de borner la variance. Cela peut se faire dans certains cas, mais pas dans tous.

La seconde méthode s'applique systématiquement en utilisant l'estimateur de la variance empirique $\hat{\sigma}_n^2$. En effet, la convergence

$$\frac{\sqrt{n}}{\hat{\sigma}_n}(\bar{X}_n - \mu) \to N(0,1)$$

est toujours vraie d'après le théorème de Slutsky.

Théorème 4.2. *Soient $X_1, \ldots, X_n$ des variables iid possédant une variance. L'intervalle*

$$\left[\bar{X}_n \pm \frac{z_\alpha \hat{\sigma}_n}{\sqrt{n}}\right]$$

est un intervalle de confiance asymptotique de niveau α pour l'estimation de la moyenne des X_i.

5 Outils pour les IC

5.1 Quantiles

Si X est une variable aléatoire sur $\mathbb{R}$, un quantile d'ordre $\beta \in]0,1[$, noté q_β, est un nombre tel que $\mathbb{P}(X \leqslant q_\beta) = \beta$. Lorsque X est continue, un tel nombre existe forcément, car la fonction de répartition $F(x) = \mathbb{P}(X \leqslant x)$ est une surjection continue. Les quantiles symétriques z_β sont, eux, définis par $\mathbb{P}(|X| \leqslant z_\beta) = \beta$.

Si la loi de X est de surcroît symétrique, les quantiles symétriques s'expriment facilement en fonction des quantiles classiques. En effet, $\mathbb{P}(|X| \leqslant z)$ est égal à $\mathbb{P}(X \leqslant z) - \mathbb{P}(X \leqslant -z)$. Or, si la loi de X est symétrique, alors $\mathbb{P}(X \leqslant -z) = 1 - \mathbb{P}(X \leqslant z)$, et donc

$$\mathbb{P}(|X| \leqslant z) = 2\mathbb{P}(X \leqslant z) - 1.$$

Il suffit alors de choisir pour z le quantile $q_{\frac{1+\beta}{2}}$ pour obtenir $\mathbb{P}(|X| \leqslant z) = \beta$. Lorsque β est de la forme $1-\alpha$ avec α petit (comme les niveaux des intervalles de confiance), on trouve alors $z_{1-\alpha} = q_{1-\alpha/2}$.

Les quantiles s'obtiennent en inversant la fonction de répartition : lorsque celle-ci est une bijection sur $]0,1[$, alors $q_\beta = F^{-1}(\beta)$. En règle générale, il n'y a pas de forme fermée. Par exemple, pour une loi gaussienne standard,

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du$$

qui elle-même n'a pas d'écriture plus simple. Fort heureusement, les outils de calcul numérique permettent d'effectuer ces calculs avec une grande précision. La table suivante donne les quantiles symétriques de la gaussienne.

β	90%	95%	98%	99%	99.9%	99.99999%
z_β	1.64	1.96	2.32	2.57	3.2	5.32

Voir aussi la règle 1-2-3. Il existe de nombreuses tables de quantiles pour les lois usuelles.

Théorème 5.1 (Queues de distribution de la gaussienne)**.** *Si x est plus grand que 1,*

$$\left(\frac{1}{x} - \frac{1}{x^3}\right) \frac{e^{-x^2/2}}{\sqrt{2\pi}} \leqslant \mathbb{P}(X > x) \leqslant \frac{1}{x} \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

En particulier, si x est grand, $\mathbb{P}(X \geqslant x) \sim e^{-x^2/2}/x\sqrt{2\pi}$ avec une erreur d'ordre $O(e^{-x^2/2}/x^3)$.

À titre d'exemple, pour $x = 2.32$ cette approximation donne 98.83%, ce qui est remarquablement proche de 98%. Pour $x = 2.57$ on trouve 99.42%.

Démonstration. À écrire.

□

5.2 Calculs de lois

5.2.1 Lois Gamma

Une variable aléatoire suit une loi Gamma de paramètres $\lambda > 0, \alpha > 0$ lorsque sa densité est donnée par

$$\gamma_{r,\alpha}(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-\lambda x} x^{\alpha-1} \mathbf{1}_{x>0}.$$

Les lois Gamma rassemblent les lois exponentielles ($\Gamma(\lambda, 1) = \mathcal{E}(\lambda)$) et les lois du chi-deux qu'on verra ci-dessous ($\Gamma(1/2, n/2) = \chi_2(n)$). La transformée de Fourier $\varphi_{\lambda,\alpha}$ d'une loi $\Gamma(\lambda, \alpha)$ se calcule facilement par un changement de variables :

$$\varphi_{\lambda,\alpha}(t) = \left(1 - \frac{it}{\lambda}\right)^{-\alpha}. \tag{5.1}$$

Cette identité montre également que si $X_1, \dots, X_n$ sont des variables indépendantes de loi $\Gamma(\lambda, \alpha_i)$, alors leur somme est une variable de loi $\Gamma(\lambda, \alpha_1 + \dots + \alpha_n)$.

5.2.2 Loi du chi-deux

Soit X une loi gaussienne standard. Calculons la densité de X^2 ; pour toute fonction-test φ, $\mathbb{E}[\varphi(X^2)]$ est donné par

$$\frac{1}{\sqrt{2\pi}} \int e^{-x^2/2} \varphi(x^2) dx.$$

Cette intégrale est symétrique, donc on peut ajouter un facteur 2 et intégrer sur $[0, \infty[$. En posant $u = x^2$, on obtient alors la valeur

$$\frac{2}{\sqrt{2\pi}} \int_0^\infty e^{-u/2} \varphi(u) \frac{1}{2\sqrt{u}} du.$$

On reconnaît la densité d'une loi Gamma de paramètres $(1/2, 1/2)$. Cette loi s'appelle *loi du chi-deux* et on la note $\chi_2(1)$. Sa tranformée de Fourier est donnée par

$$\mathbb{E}[e^{itX^2}] = \frac{1}{\sqrt{1 - 2it}}.$$

Soient maintenant $X_1, \dots, X_n$ des variables de loi $N(0, 1)$ indépendantes. Chaque X_i^2 est une $\chi_2(1)$; leur somme a pour loi la convolée n fois de $\chi_2(1)$. Calculons sa tranformée de Fourier :

$$\mathbb{E}[e^{it(X_1^2 + \dots + X_n^2)}] = \mathbb{E}[e^{itX_1^2}]^n \tag{5.2}$$

$$= (1 - 2it)^{-\frac{n}{2}}. \tag{5.3}$$

On reconnaît la transformée de Fourier d'une loi $\Gamma(n/2, 1/2)$; cette loi s'appelle *loi du chi-deux à n paramètres de liberté* et elle est notée $\chi_2(n)$. Sa densité est donnée par

$$\frac{1}{2^{n/2}\Gamma(n/2)} e^{-x/2} x^{n/2-1} \mathbf{1}_{x>0}. \tag{5.4}$$

5.2.3 Loi de Student

Soit X une variable de loi $N(0,1)$ et Y_n une variable de loi $\chi_2(n)$ indépendante de X. On va calculer la loi de $T_n = X/\sqrt{Y_n/n}$. Soit φ une fonction test ; l'espérance $\mathbb{E}[\varphi(T_n)]$ est égale à

$$\frac{1}{Z_n\sqrt{2\pi}} \int_0^\infty \int_{-\infty}^\infty \varphi\left(\frac{x}{\sqrt{y/n}}\right) e^{-\frac{x^2}{2}} e^{-\frac{y}{2}} y^{\frac{n}{2}-1} dxdy$$

où $Z_n = 2^{n/2}\Gamma(n/2)$. Dans l'intégrale en x, on effectue le changement de variable $u = x/\sqrt{y/n}$ afin d'obtenir

$$\frac{1}{Z_n\sqrt{2\pi}} \int_0^\infty \int_{-\infty}^\infty \varphi(u) e^{-\frac{yu^2}{2n}} e^{-\frac{y}{2}} y^{\frac{n}{2}-1} \sqrt{\frac{y}{n}} dxdy.$$

La densité de T_n est donc donnée par

$$t_n(u) = \frac{1}{Z_n\sqrt{2\pi n}} \int_0^\infty e^{-\frac{yu^2}{2n}-\frac{y}{2}} y^{\frac{n+1}{2}-1} dy.$$

Le changement de variables $z = y(1+u^2/n)/2$ nous ramène à

$$t_n(u) = \frac{1}{Z_n\sqrt{2\pi n}} \left(\frac{2}{1+\frac{u^2}{n}}\right)^{\frac{n+1}{2}} \int_0^\infty e^{-z} z^{\frac{n+1}{2}-1} dz.$$

On reconnaît $\Gamma((n+1)/2)$ à droite. La densité $t_n(x)$ est donc

$$t_n(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(\frac{1}{1+\frac{x^2}{n}}\right)^{\frac{n+1}{2}}.$$

Cette loi s'appelle *loi de Student* de paramètre n ; on dit parfois *à n degrés de liberté*. Elle est notée $\mathcal{T}(n)$. La loi de Student de paramètre $n = 1$ est tout simplement une loi de Cauchy.

5.2.4 Loi de la statistique de Student

Soient $X_1, \ldots, X_n$ des variables gaussiennes $N(\mu, \sigma^2)$ indépendantes, et soit $T_n = (\bar{X}_n - \mu)/\sqrt{\hat{\sigma}_n^2}$, où

$$\hat{\sigma}_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n-1}.$$

Théorème 5.2.

$$T_n \sim \mathcal{T}(n-1).$$

Démonstration. On va montrer 1° que $\bar{X}_n$ et $\sqrt{\hat{\sigma}_n^2/\sigma^2}$ sont indépendantes, et 2° que $\sqrt{\hat{\sigma}_n^2/\sigma^2}$ a bien la même loi que $\sqrt{Y_{n-1}/(n-1)}$ où Y_{n-1} est une $\chi_2(n-1)$. Dans la suite, on supposera que $\mu = 0$ et $\sigma = 1$, ce qui n'enlève rien en généralité.

Premier point. Le vecteur $X = (X_1, \dots, X_n)$ est gaussien. Posons $Z = (X_1 - \bar{X}_n, \dots, X_n - \bar{X}_n)$. Le couple $(\bar{X}_n, Z_n)$ est linéaire en X, donc ce couple est aussi un vecteur gaussien. Or, la covariance de ses deux éléments est nulle. Par exemple, $\mathrm{Cov}(\bar{X}_n, Z_1)$ est égale à $\mathrm{Cov}(\bar{X}_n, X_1) - \mathrm{Var}(\bar{X}_n)$, ce qui par linéarité donne $1/n - 1/n = 0$. Ainsi, $\bar{X}_n$ et Z sont deux variables conjointement gaussiennes et décorrélées : elles sont donc indépendantes. Comme $\hat{\sigma}_n$ est une fonction de Z, elle est aussi indépendante de $\bar{X}_n$.

Second point. Z est la projection orthogonale de X sur le sous-espace vectoriel $\mathcal{V} = \{x \in \mathbb{R}^n : x_1 + \dots + x_n = 0\}$. Soit $(f_i)_{i=2,\dots,n}$ une base orthonormale de $\mathcal{V}$, de sorte que $Z = \sum_{i=2}^n \langle f_i, X\rangle f_i$. Par l'identité de Parseval,

$$|Z|^2 = \sum_{i=2}^n |\langle f_i, X\rangle|^2.$$

Or, les $n-1$ variables aléatoires $G_i = \langle f_i, X\rangle$ sont des gaussiennes standard iid. En effet, on vérifie facilement que $\mathrm{Cov}(G_i, G_j) = \langle f_i, f_j\rangle = \delta_{i,j}$. On en déduit donc que $|Z|^2$ suit une loi $\chi_2(n-1)$.

□

La seconde partie de la démonstration est un cas particulier du théorème de Cochran, que nous verrons dans le chapitre sur la régression linéaire.

5.3 Inégalités de concentration

Les outils de base pour construire des intervalles de confiance dans des circonstances générales (non gaussiennes) sont les inégalités de concentration. Une inégalité de concentration pour une variable aléatoire intégrable X consiste à borner $\mathbb{P}(|X - \mathbb{E}[X]| > x)$ par quelque chose de petit quand x est grand : on cherche à contrôler la probabilité pour que les réalisations de la variable aléatoire X soient éloignées de leur valeur moyenne $\mathbb{E}[X]$ de plus de x.

5.4 Inégalité de Bienaymé-Tchebychev

Théorème 5.3. *Soit X une variable aléatoire de carré intégrable. Alors,*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geqslant x) \leqslant \frac{\mathrm{Var}(X)}{x^2}.$$

Démonstration. Élever au carré les deux membres de l'inégalité dans $\mathbb{P}$, puis appliquer l'inégalité de Markov à la variable aléatoire positive $|X - \mathbb{E}X|^2$ dont l'espérance est $\mathrm{Var}(X)$.

□

5.5 Inégalité de Hoeffding

Théorème 5.4 (Inégalité de Hoeffding)**.** *Soient $X_1, \dots, X_n$ des variables aléatoires indépendantes, pas forcément de même loi. On suppose que chaque X_i est à valeurs dans un intervalle borné $[a_i, b_i]$ et on pose $S_n = X_1 + \dots + X_n$. Pour tout $t > 0$,*

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geqslant t) \leqslant e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}} \tag{5.5}$$

et

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geqslant t) \leqslant 2e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}. \tag{5.6}$$

La démonstration se fonde sur le lemme suivant.

Lemme 5.1 (lemme de Hoeffding)**.** *Soit X une variable aléatoire à valeurs dans $[a, b]$. Pour tout t,*

$$\mathbb{E}[e^{t(X - \mathbb{E}[X])}] \leqslant e^{\frac{t^2 (b-a)^2}{8}}. \tag{5.7}$$

Démonstration. Soit X une variable aléatoire, que par simplicité on supposera centrée et à valeurs dans l'intervalle $[a, b]$ (a est forcément négatif). En écrivant

$$x = a \times \frac{b - x}{b - a} + b \times \left(1 - \frac{b - x}{b - a}\right)$$

et en utilisant la convexité de la fonction $x \mapsto e^{tx}$, on obtient $e^{tX} \leqslant (b - X)e^{ta}/(b - a) + (1 - (b - x)/(b - a))e^{bt}$), puis en prenant l'espérance et le fait que X est centrée et en simplifiant,

$$\mathbb{E}[e^{tX}] \leqslant \frac{be^{ta} - ae^{tb}}{b - a}.$$

Notons $f(t)$ le terme à droite ; pour montrer Équation 5.7, il suffit de montrer que $\ln f(t) \leqslant t^2(b - a)^2/8$. La formule de Taylor dit que

$$\ln f(t) = \ln f(0) + t(\ln f)'(0) + \frac{t^2}{2}(\ln f)''(\xi)$$

pour un certain ξ. Or, $\ln f(0) = \ln 1 = 0$, $(\ln f)'(0) = f'(0)/f(0) = 0$, et il suffit donc de montrer que $(\ln f)''(t)$ est toujours plus petit que $(b - a)^2/4$ pour conclure. Un simple calcul montre que $\ln f(t) = \ln(b/(b - a)) + ta + \ln(1 - ae^{t(b-a)}/b)$, et donc

$$(\ln f)''(t) = \frac{(a/b)(b - a)e^{t(b-a)}}{(1 - ae^{t(b-a)}/b)^2}.$$

L'inégalité $uv/(u - v)^2 \leqslant 1/4$ appliquée à $u = a/b$ et $v = e^{t(b-a)}$ permet alors de conclure.

□

Preuve de l'inégalité de Hoeffding. En remplaçant X_k par $X_k - \mathbb{E}[X_k]$, on peut supposer que tous les X_i sont centrés et étudier seulement $\mathbb{P}(S_n > t)$. Écrivons $\mathbb{P}(S_n > t) = \mathbb{P}(e^{\lambda S_n} > e^{\lambda t})$, où λ est un nombre positif que l'on choisira plus tard. L'inégalité de Markov borne cette probabilité par $\mathbb{E}[e^{\lambda S_n}]e^{-\lambda t}$. Comme les X_i sont indépendantes, $\mathbb{E}[e^{tS_n}]$ est le produit des $e^{\varphi_k(\lambda)}$ où $\varphi_k(t) = \ln \mathbb{E}[e^{itX_k}]$. En appliquant le lemme de Hoeffding à chaque φ_k, on borne $\mathbb{P}(S_n > t)$ par

$$\exp\left(\sum_{i=1}^{n} \frac{(b_i - a_i)^2\lambda^2}{8} - t\lambda\right).$$

Le minimum en λ du terme dans l'exponentielle est atteint au point $4t/\sum(a_i - b_i)^2$ et la valeur du minimum est le terme dans l'exponentielle de Équation 5.5. On déduit Équation 5.6 par une simple borne de l'union.

La démonstration de l'inégalité de Hoeffding ne dépend pas directement du fait que X est bornée, mais plutôt de Équation 5.7. Toutes les variables aléatoires qui vérifient une inégalité de type $\mathbb{E}[e^{tX}] \leqslant e^{ct^2}$ pour une constante c peuvent donc avoir leur propre inégalité de Hoeffding.

6 Exercices

6.1 Questions

1. Soit X_n une variable aléatoire de loi de Student de paramètre n. Montrer que X_n converge en loi vers $N(0,1)$.
2. Soit $X_n \sim \chi_2(n)$. La suite (X_n) est-elle asymptotiquement normale ?
3. Donner un intervalle de confiance de la forme $[A, +\infty[$ pour la moyenne d'un échantillon gaussien.
4. Même question pour la variance dans un modèle gaussien centré.
5. Dans l'estimation de la moyenne μ d'un modèle gaussien où la variance σ^2 est connue, montrer que l'intervalle de confiance obtenu (Équation 4.4) est le plus grand possible de niveau $1-\alpha$.
6. Démontrer le théorème Théorème 5.1 sur l'asymptotique des queues de distribution de la loi gaussienne.
7. Montrer la borne suivante sur les quantiles de loi gaussienne standard: $q_\beta < \sqrt{\ln \frac{1}{\beta\sqrt{2\pi}}}$ (pour tout $1/2 < \beta < 1$).

8. Comparer les queues de distribution des lois $N(0,1), \chi_2(n)$ et $\mathcal{T}(n)$.
9. Expliquer à votre grand-mère la différence entre un intervalle de fluctuation et un intervalle de confiance.
10. L'intervalle de confiance de niveau $1-\alpha$ pour la moyenne d'un modèle $N(\mu,1)$ avec n observations est $I_n = [\bar{X}_n \pm z_\alpha/\sqrt{n}]$. Supposons qu'on obtienne une nouvelle observation indépendante des autres, disons Z. La probabilité $\mathbb{P}(Z \in I_n)$ est-elle plus grande ou plus petite que $1-\alpha$?
11. Comparer la longueur des intervalles de confiance obtenus par les différentes méthodes de la section Section 4.3.1.

6.2 Exercices

Exercice 6.1 (Lois de Poisson)**.** On suppose que l'on observe $X_1, \dots, X_n$ i.i.d de loi $\mathcal{P}(\theta)$.

1. Étudier $\bar{X}_n$.
2. Montrer que $\sqrt{\bar{X}_n} \xrightarrow[n\to\infty]{\mathbb{P}} \sqrt{\theta}$.
3. Donner deux intervalles de confiance au niveau 98% pour $\sqrt{\theta}$, et les comparer.

Exercice 6.2 (Lois uniformes)**.** Soit $X_1, \dots, X_n$ des variables aléatoires iid de loi $\mathcal{U}[0,\theta]$. Donner un intervalle de confiance non asymptotique pour θ en utilisant l'estimateur $\hat{\theta}_n = \max_{i=1,\dots,n} X_i$.

Exercice 6.3 (Lois exponentielles décalées)**.** Soit $X_1, \dots, X_n$ des variables aléatoires iid de densité $e^{\theta-x}\mathbf{1}_{x>\theta}$, où $\theta > 0$.

1. Calculer $\mathbb{E}_\theta[X_1]$ et en déduire un estimateur de θ que l'on notera $\widehat{\theta}_n$. Étudier ses propriétés (risque quadratique, convergence) et l'utiliser pour construire un premier intervalle de confiance $I_1(\alpha)$ non-asymptotique pour θ de niveau $1-\alpha$.
2. Construire un intervalle de confiance asymptotique $I_2(\alpha)$ pour θ à partir de $\widehat{\theta}_n$.
3. Montrer que l'estimateur $\theta_n^\star := \min_{1\leq i\leq n} X_i$ est meilleur que $\widehat{\theta}_n$ au sens du risque quadratique, puis l'utiliser pour construire un intervalle de confiance $I_3(\alpha)$ de niveau $1-\alpha$.
4. Comparer les longueurs de tous ces différents intervalles de confiance.

Exercice 6.4 (Lois exponentielles)**.** Soit $X_1, \dots, X_n$ des variables aléatoires iid exponentielles de paramètre $\lambda > 0$.

1. Quelle est la loi de $S_n = X_1 + \cdots + X_n$?
2. Construire un intervalle de confiance de niveau $1-\alpha$ pour λ.

Exercice 6.5 (Inégalité d'Azuma)**.** Montrer que l'inégalité de Hoeffding reste valable lorsque les X_i ne sont plus supposés indépendants, mais que la suite $S_k = X_1 + \cdots + X_k$ est une martingale. Indice : $\mathbb{E}[e^{\lambda S_{n+1}}] = \mathbb{E}[e^{\lambda S_n}\mathbb{E}[e^{\lambda X_{n+1}}|S_n]]$.

Ce raffinement s'appelle *inégalite de Hoeffding-Azuma*. C'est celui que nous avons utilisé dans l'exercice (**ex-tanks?**), lorsque les $X_1, \dots, X_n$ sont des tirages *sans remise* dans une urne à N éléments.

7 Test d'hypothèses

Si l'on essaie d'estimer le rendement μ d'un actif financier, on cherche implicitement à savoir si l'on va investir ou pas. Cette décision dépendra de notre estimation : pour faire simple, on peut considérer que si nous estimons que le rendement est positif ($\hat{\mu} > 0$), alors il faut investir. Sinon, on n'investira pas.

Les tests d'hypothèses visent à formaliser cela. Faire une *hypothèse* dans un modèle statistique $(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta \in \Theta_0})$, c'est supposer que θ appartient à une certaine région de $H_0 \subset \Theta$. Les *tests* visent à construire des procédures pour tester une hypothèse nulle, que l'on notera H_0, contre une hypothèse alternative, notée H_1.

Dans le cadre ci-dessus, on peut se placer dans un modèle où les rendements sont $\mathcal{N}(\mu, \sigma^2)$. On veut tester l'hypothèse nulle $H_0 : \ \mu \in]-\infty, 0]$ contre l'hypothèse alternative $H_1 : \ \mu \in]0, +\infty[$.

Définition 7.1. Un test est un événement qui, s'il survient, nous incite à rejeter l'hypothèse nulle. Cet événement sera noté rejeter et son complémentaire sera noté accepter.

- L'erreur de première espèce est la probabilité de rejeter l'hypothèse nulle à tort : $\alpha = \sup_{\theta \in H_0} P_\theta(\text{rejeter})$. Le **niveau d'un test** est $1 - \alpha$. C'est la probabilité d'accepter l'hypothèse nulle à raison.
- L'erreur de seconde espèce est la probabilité de ne pas rejeter l'hypothèse nulle, à tort : $\beta = \sup_{\theta \in H_1} P_\theta(\text{accepter})$. La **puissance d'un test** est $1 - \beta$. C'est la probabilité de « détecter » l'hypothèse alternative à raison.
- L'affinité d'un test est la somme des erreurs de première et seconde espèce.
-

Par « événement », on veut bien dire « un élément de $\mathcal{F}$ », c'est-à-dire qui n'est déterminé que par les observations et pas par θ. Formellement on écrit souvent qu'un test est une statistique, disons T, à valeurs dans $\{0, 1\}$. L'événement $\{T = 0\}$ est rejeter, l'événement $\{T = 1\}$ est accepter.

Un des grands objectifs de la statistique mathématique est de construire des familles de tests qui, pour un niveau de confiance $1 - \alpha$ fixé, ont la plus grande puissance possible ; autrement dit, **trouver un événement hautement improbable sous l'hypothèse nulle, et hautement probable sous l'hypothèse alternative**.

Comme on verra dans les exemples, le rôle des deux hypothèses n'est pas interchangeable. Maximiser le niveau et la puissance ne reviennent pas au même. Le choix des hypothèses H_0 et H_1 n'est pas anodin : l'hypothèse H_0 est une hypothèse que l'on cherche implicitement à réfuter.

1. Si $\theta \in H_0$ quel qu'il soit, les probablités pour qu'un certain événement rejeter sont infimes – disons, 1%.
2. Si cet événement arrive, par contraposée, on est amenés à rejeter l'hypothèse selon laquelle θ est dans H_0.

C'est pour cela que les tests sont une forme de logique statistique. Le raisonnement de base une contraposée : en logique, $A \Rightarrow B$ est équivalent à $\neg B \Rightarrow \neg A$. En statistiques, on pourrait écrire $\theta \in H_0 \Rightarrow \mathsf{accepter}$ (avec grande probabilité), donc $\mathsf{rejeter} \Rightarrow \theta \notin H_0$ (probablement).

7.1 Exemples de tests gaussiens

On se place dans un modèle où $X_1, \dots, X_n$ sont des gaussiennes $N(\mu, \sigma^2)$. Nous avons déjà vu plusieurs fois que $\bar{X}_n \sim N(\mu, \sigma^2)$.

7.1.1 Construction du test

On cherche à réfuter l'hypothèse selon laquelle ces variables aléatoires sont centrées ; autrement dit, on posera $H_0 = \{0\}$. Sous cette hypothèse, nos variables aléatoires sont donc des variables $N(0, \sigma^2)$.

Supposons dans un premier temps que σ^2 est connue. Sous H_0, on a donc

$$\frac{\sqrt{n}\bar{X}_n}{\sigma} \sim N(0,1)$$

et par conséquent, $P_0(|\bar{X}_n| < z_{1-\alpha}\sigma/\sqrt{n}) = 1 - \alpha$. Autrement dit, sous l'hypothèse $\mu = 0$, on devrait observer l'événement

$$\bar{X}_n \in \left[\pm \frac{z_{1-\alpha}\sigma}{\sqrt{n}}\right]$$

avec probabilité élevée $1 - \alpha$. Si cet événement n'est pas observé, il est alors très douteux que μ soit effectivement égal à zéro ! On pose donc

$$\mathsf{rejeter}_\alpha = \{|\bar{X}_n| > z_{1-\alpha}\sigma/\sqrt{n}\}.$$

Le niveau de ce test est bien $1 - \alpha$: nous l'avons construit pour cela.

Supposons maintenant que σ n'est pas connue. En l'estimant via $\hat{\sigma}_n$, nous savons que (toujours sous l'hypothèse selon laquelle $\mu = 0$)

$$\frac{\sqrt{n}\bar{X}_n}{\hat{\sigma}_n} \sim \mathcal{T}(n-1).$$

On reproduit alors le raisonnement ci-dessus : comme $\mathbb{P}(|\bar{X}_n| < t_{n-1,1-\alpha}\sqrt{\sigma}_n/\sqrt{n}) = \alpha$ où $t_{n-1,1-\alpha}$ est le quantile symétrique de $\mathcal{T}(n-1)$, on voit que l'événement

$$\mathsf{rejeter}_\alpha = \{|\bar{X}_n| > t_{n-1,1-\alpha}\hat{\sigma}_n/\sqrt{n}\}$$

est bien un test de niveau $1 - \alpha$.

7.1.2 Calcul de la puissance et hypothèse alternative

Nous n'avons pas encore eu besoin de spécifier une hypothèse alternative, mais nous allons en avoir besoin pour calculer la puissance du test. Pour commencer, on va supposer que, si μ n'est pas nulle, alors elle ne peut être égale qu'à 1. Autrement dit, $H_1 = \{1\}$. Ce genre d'hypothèse alternative ne peut évidemment avoir de pertinence qu'en fonction du problème réel sous-jacent !

Sous l'hypothèse alternative, donc, nous savons que $\bar{X}_n \sim N(1, \sigma^2)$. La puissance du test est définie par $1 - \beta$ où $\beta = P_1(\mathsf{accepter}_\alpha)$ c'est-à-dire

$$\begin{aligned}
\beta &= P_1(|\bar{X}_n| \leqslant z_{1-\alpha}\sigma/\sqrt{n}) && (7.1)\\
&= P_1\left(-\frac{z_{1-\alpha}\sigma}{\sqrt{n}} \leqslant \bar{X}_n \leqslant \frac{z_{1-\alpha}\sigma}{\sqrt{n}}\right) && (7.2)\\
&= P_1\left(-\frac{z_{1-\alpha}\sigma}{\sqrt{n}} - 1 \leqslant \bar{X}_n - 1 \leqslant \frac{z_{1-\alpha}\sigma}{\sqrt{n}} - 1\right) && (7.3)\\
&= \Phi(-\sqrt{n}/\sigma + z_{1-\alpha}) - \Phi(-\sqrt{n}/\sigma + z_{1-\alpha}). && (7.4)
\end{aligned}$$

où $\Phi(x) = \mathbb{P}(N(0,1) \leqslant x)$. Cette expression ne peut pas plus se simplifier, mais on peut quand même la borner par $F(-\sqrt{n}/\sigma + z_{1-\alpha})$. Lorsque x est grand, nous avons vu (Théorème 5.1) que $F(x) < e^{-x^2}/|x|\sqrt{2\pi}$. Ainsi, l'erreur de première espèce est bornée par $O(e^{-n/\sigma^2/2}/\sqrt{n})$. Cela tend extrêmement vite vers 0 ; en fait, dès que n est plus grand que 10 et $\sigma = 1$, cette erreur est inférieure à 0.1%, donc dans ce cas le test aura une puissance supérieure à 99.9%.

Que se serait-il passé si notre hypothèse alternative n'avait pas été $\mu = 1$ mais $\mu = m$ pour n'importe quel $m \neq 0$? Dans ce cas, on aurait eu $H_1 = \mathbb{R} \setminus \{0\}$. L'erreur de première espèce aurait alors été $\beta = \sup_{m\neq 0} \beta_m$ où

$$\beta_m = P_m(\mathsf{accepter}_\alpha).$$

On revoyant les calculs ci-dessus, on voit que

$$\beta_m = \Phi(-m\sqrt{n}/\sigma + z_{1-\alpha}) - \Phi(-m\sqrt{n}/\sigma + z_{1-\alpha}).$$

En particulier,

$$\lim_{m\to 0} \beta_m = \Phi(-z_{1-\alpha}) - \Phi(-z_{1-\alpha}) = 1 - \alpha$$

par continuité de Φ et par définition de $z_{1-\alpha}$. Ainsi, $1 - \beta = \alpha$: pour cette seconde hypothèse alternative, la puissance de notre test… est extrêmement faible.

Cela vient du fait que notre hypothèse alternative contient des situations quasiment indiscernables de notre hypothèse nulle. Par exemple, il est quasiment impossible de distinguer $\mu = 0$ de $\mu = 10^{-100}$ par exemple. Cet exemple illustre la dissymétrie entre H_0 et H_1.

7.2 La notion de p-valeur

La construction d'un test dépend du niveau de risque α. Si le niveau de risque acceptable est de plus en petit, alors l'événement $\mathsf{rejeter}_\alpha$ devrait être de moins en moins probable. D'ailleurs, $\mathsf{rejeter}_0 = \emptyset$ et $\mathsf{accepter}_0 = \Omega$: si l'on ne tolère aucun niveau de risque de première espèce, c'est qu'on ne veut pas rejeter l'hypothèse nulle.

Très souvent, si $\alpha < \beta$, on a même

$$\mathsf{rejeter}_\alpha \subset \mathsf{rejeter}_\beta.$$

Définition 7.2. La p-valeur d'une famille croissante de tests est le plus petit niveau de risque qui nous amène à rejeter l'hypothèse nulle compte tenu des observations. Formellement,

$$p = \inf\{\alpha > 0 : \ \mathsf{rejeter}_\alpha\}.$$

C'est donc une statistique.

8 Théorie des tests simples

8.1 La distance en variation totale

Lorsqu'on cherche à tester une hypothèse de type loi $= P$ contre une hypothèse de type loi $= Q$ (c'est-à-dire, deux hypothèses simples), on en revient à chercher un événément très improbable sous la loi P, et très probable sous la loi Q. On peut se demander en toute généralité quels sont les événements A pour lesquels ces probabilités diffèrent le plus, c'est-à-dire les événements qui maximisent $P(A) - Q(A)$. Cela mène directement à la définition de la *variation totale*.

Définition 8.1 (distance en variation totale)**.** Soient P, Q deux mesures de probabilité sur un même espace $(\mathcal{X}, \mathcal{F})$. Leur distance en variation totale est

$$\mathrm{d_{TV}}(P, Q) = \sup_{A \in \mathcal{F}} P(A) - Q(A).$$

La distance en variation totale est un objet important en probabilités, qui possède de nombreuses propriétés. Parmi elles,

1. C'est une distance sur l'espace des mesures de probabilité.
2. Elle génère une topologie plus fine que celle de la convergence en loi ; autrement dit, si $\mathrm{d_{TV}}(P_n, Q) \to 0$ alors P_n converge en loi vers Q mais l'inverse n'est pas vrai.

Proposition 8.1. *Soit ν une mesure telle que P et Q sont absolument continues par rapport à ν, de densités respectives p et q par rapport à ν. Alors,*

$$\mathrm{d_{TV}}(P, Q) = \int_{\mathcal{X}} (p(x) - q(x))_+ d\nu = \frac{1}{2} \int_{\mathcal{X}} |p(x) - q(x)| d\nu. \tag{8.1}$$

De plus, notons E l'ensemble mesurable $\{p(x) > q(x)\}$. Alors,

$$\mathrm{d_{TV}}(P, Q) = P(E) - Q(E). \tag{8.2}$$

Démonstration. Pour tout événement $A \in \mathcal{F}$, la différence $P(A) - Q(A)$ est égale à $\int_A p(x) - q(x) d\nu$, qui peut elle-même s'écrire sous la forme

$$\int_{A \cap E} p(x) - q(x) d\nu + \int_{A \cap \bar{E}} p(x) - q(x) d\nu.$$

Le second terme est négatif, puisque si $x \notin E$ alors $p(x) \leqslant q(x)$. Ainsi, $P(A) - Q(A)$ est plus petit que le premier terme, lequel est à son tour plus petit que $\int_E (p - q) d\nu = P(E) - Q(E)$. Cela montre directement Équation 8.2. Au passage, il est évident que

$$\int_E (p(x) - q(x)) d\nu = \int_{\mathcal{X}} (p(x) - q(x))_+ d\nu,$$

ce qui montre la première égalité de Équation 8.1. La seconde égalité résulte de la première, puisque comme p et q sont des densités de probabilité, on a forcément $\int (p-q)_+ = \int (p-q)_-$.

□

8.2 Meilleur test possible au sens de l'affinité

L'affinité d'un test est la somme de ses erreurs de première et seconde espèce : c'est la probabilité de « se tromper » en général, quelle que soit l'hypothèse.

Théorème 8.1. *Soit $\mathfrak{T}$ l'ensemble des tests possibles de l'hypothèse H_0 : $P = P_0$ contre l'hypothèse alternative H_1 : $P = P_1$. Alors, le test possédant la meilleure affinité possible parmi tous les tests possibles vérifie*

$$\inf_{T \in \mathfrak{T}} \{\alpha_T + \beta_T\} = 1 - \mathrm{d_{TV}}(P_0, P_1).$$

En particulier, le test optimal pour l'affinité est donné par la région de rejet

$$\mathsf{rejeter}_\star = \{p(x) > q(x)\}.$$

Démonstration. Soit T n'importe quel test. Son affinité $P_1(\{T = 0\}) + P_0(\{T = 1\})$. En passant au complémentaire dans le second terme, on obtient

$$1 - (P_0(\{T = 0\}) - P_1(\{T = 0\})).$$

Cette quantité est forcément plus petite que $1 - \mathrm{d_{TV}}(P_0, P_1)$ par la définition même de la variation totale. De plus, cette borne est atteinte en choisissant le test T donné dans l'énoncé, d'où l'égalité.

□

8.3 Bornes sur la variation totale

Le théorème précédent semble donner au problème de la construction de tests une réponse définitive : il donne le test optimal au sens de l'affinité. C'est pourtant trompeur, car la construction de ce test nécessite le calcul de la distance en variation totale, laquelle peut être notoirement difficile. En pratique, on cherche plutôt à *borner* cette distance par d'autres quantités plus faciles à calculer. Parmi ces quantités, la *divergence de Kullback-Leibler* joue un rôle extrêmement important, notamment pour son lien avec le maximum de vraisemblance que nous verrons plus tard.

Définition 8.2. Soient P et Q deux mesures, P étant absolument continue par rapport à Q. Alors,

$$\mathrm{d_{KL}}(P \mid Q) = \int \ln\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) dP.$$

Cette quantité n'est pas une distance, et c'est pour cela qu'on l'appelle *divergence* et qu'on la note avec une barre plutôt qu'une virgule : elle n'est pas symétrique en général. Cependant, elle est toujours positive, et n'est nulle que si $P = Q$.

Théorème 8.2 (Borne de Bretagnole-Huber-Pinsker)**.**

$$\mathrm{d_{TV}}(P,Q) \leqslant \sqrt{1-e^{-\mathrm{d_{KL}}(P|Q)}}. \tag{8.3}$$

Il est très facile de vérifier que $\sqrt{1-e^{-x}} \leqslant \sqrt{x}$ lorsque $x > 0$. Ainsi, Équation 8.3 entraîne la borne plus simple $\mathrm{d_{TV}} \leqslant \sqrt{\mathrm{d_{KL}}}$. La borne *classique* de Pinsker dit qu'en fait, on a légèrement mieux, puisqu'en toute généralité $\mathrm{d_{TV}} \leqslant \sqrt{\mathrm{d_{KL}}/2}$.

Démonstration. À écrire.

□

9 Modèle linéaire

9.1 Ajustement affine en une dimension.

On dispose de variables x_i, dites *explicatives*, et de variables y_i, dites *à expliquer*. On suppose qu'il existe une relation, peut-être imparfaite, de la forme

$$y_i \approx \alpha + \beta x_i$$

où α, β sont deux nombres réels. Pour trouver les meilleurs α, β possibles, on calcule la distance entre le nuage de points (x_i, y_i) et la droite d'équation $y = \alpha + \beta x$. Cette distance au carré est donnée par

$$\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2.$$

On cherchera donc à minimiser la fonction de deux variables

$$(\alpha, \beta) \mapsto \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2.$$

Appelons cette fonction L. C'est manifestement une fonction quadratique qui tend vers $+\infty$ lorsque $(\alpha, \beta) \to \infty$, par conséquent cette fonction possède un unique minimiseur $(\hat{\alpha}, \hat{\beta})$, et ce minimiseur est le seul point en lequel les dérivées partielles s'annulent (*conditions de premier ordre*) : $\partial_\alpha L(\hat{\alpha}, \hat{\beta}) = 0$ et $\partial_\beta L(\hat{\alpha}, \hat{\beta}) = 0$. Or,

$$\partial_\alpha L(\alpha, \beta) = \sum_{i=1}^{n} (\alpha + \beta x_i - y_i)$$

$$\partial_\beta L(\alpha, \beta) = \sum_{i=1}^{n} x_i (\alpha + \beta x_i - y_i).$$

Les conditions de premier ordre deviennent donc $n\alpha + \beta(x_1 + ... + x_n) - (y_1 + \cdots + y_n) = 0$ soit encore $\alpha + \beta\bar{x} - \bar{y} = 0$, et d'autre part $\alpha(x_1 + \cdots + x_n) + \beta(x_1^2 + \cdots + x_n^2) - (x_1 y_1 + \cdots + x_n y_n) = 0$, soit $\alpha\bar{x} + \beta\overline{xx} - \overline{xy} = 0$, où $\overline{xx}$ est la moyenne des carrés des x_i et $\overline{xy}$ la moyenne des $x_i y_i$. En résolvant ces équations, on trouve d'abord α puis β :

$$\beta = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{xx} - \bar{x}\bar{x}}, \qquad \alpha = \bar{y} - \hat{\beta}\bar{x}.$$

Le coefficient β n'est rien d'autre que la covariance empirique des x_i et des y_i, normalisé par la variance empirique des x_i.

L'inégalité de Cauchy-Schwartz dit que $|\overline{xy} - \bar{x}\bar{y}| \leqslant \tilde{\sigma}_x \tilde{\sigma}_y$, où l'on a noté

$$\tilde{\sigma}_x^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)^2$$

l'estimateur naïf de la variance[1]. L'inégalité n'est une égalité que si x et y sont effectivement colinéaires, c'est-à-dire si $y_i = \hat{\alpha} + x_i\hat{\beta}$ pour tous les i. La qualité de l'ajustement affine est donc bien mesurée par la quantité

$$r^2 = \frac{\overline{xy} - \bar{x}\bar{x}}{\tilde{\sigma}_x\tilde{\sigma}_y}.$$

9.2 Cadre général

Dans le cadre général, les variables explicatives ne sont pas de dimension 1 mais d. On notera $x = (x_1, \dots, x_d)$ un élément de $\mathbb{R}^d$; les variables explicatives seront alors $x_1, \dots, x_n$. On cherchera des nombres θ_i tels que y_i est aussi proche que possible de

$$\theta_1 x_{i,1} + \cdots + \theta_d x_{i,d} = \langle \theta, x \rangle = x_i\theta^\top$$

On pose X la matrice $n \times d$ dont la i-ème ligne est x_i. On cherche à trouver le θ qui minimise $|Y - X\theta|^2$, autrement dit

$$\hat{\theta} = \arg\min_\theta |Y - X\theta|^2.$$

Théorème 9.1. *Si $d \leqslant n$ et si X est de rang d, alors*

$$\hat{\theta} = (X^\top X)^{-1} X^\top Y. \tag{9.1}$$

Démonstration. La projection orthogonale sur le sous-espace vectoriel engendré par les colonnes d'une matrice X est la matrice $X(X^\top X)^{-1}X^\top$. Ainsi, la projection de Y sur ce sous-espace est $X(X^\top X)^{-1}X^\top Y$, et c'est aussi (par définition de l'argmin) $X\hat{\theta}$. Comme X est injective en vertu du théorème du rang, on en déduit le résultat.

□

Le vecteur $\hat{\varepsilon} = Y - X\hat{\theta}$ est appelé *vecteur des résidus*. S'il est nul ou très petit, cela veut dire que les Y sont presque parfaitement des fonctions linéaires des X.

9.3 Modèle gaussien

À ce stade, nous n'avons fait aucune hypothèse statistique ni probabiliste sur le modèle : les x_i, y_i étaient donnés tels quels. Dans le *modèle linéaire gaussien* avec variables explicatives $x_1, \dots, x_n$ exogènes consiste à supposer que $Y = X\theta + \varepsilon$, où $\varepsilon = N(0, \sigma^2 I_n)$. Formellement, le modèle est indexé par θ et σ^2, et donné par

$$P_{\theta,\sigma^2} = N(X\theta, \sigma^2 I_d).$$

Dans ce modèle, la loi de l'estimateur Équation 9.1 est connue.

[1] Celui qui est biaisé, contrairement à $\hat{\sigma}_n^2$.

Théorème 9.2. *Sous le modèle linéaire gaussien P_{θ,σ^2},*

$$\widehat{\theta} \sim N(\theta, \sigma^2(X^\top X)^{-1}),$$

$$\widehat{\varepsilon} \sim N(0, \sigma^2(I_n - H)),$$

et ces deux variables aléatoires sont indépendantes. En particulier, $|\widehat{\varepsilon}|^2/\sigma^2 \sim \chi_2(d)$.

Démonstration. Ce n'est rien de plus que le théorème de Cochran appliqué à notre problème.

□

10 Outils gaussiens

10.1 Théorème de Cochran

Théorème 10.1 (Théorème de Cochran)**.** *Soit $X \sim N(0, I_d)$ et soient $E_1, \dots, E_k$ des sous-espaces orthogonaux de $\mathbb{R}^n$ tels que $\mathbb{R}^n = \oplus_{j=1}^k E_j$. On note $\pi_j(X)$ la projection orthogonale de X sur E_j. Alors, la famille $(\pi_j(X))_{j=1,\dots,k}$ est une famille de vecteurs gaussiens indépendants. De plus,*

$$|\pi_j(X)|^2 \sim \chi_2(\dim E_j).$$

Démonstration. À écrire.

□

10.2 Loi de Fisher

Soient X, Y deux variables aléatoires indépendantes, de lois respectives $\chi_2(p)$ et $\chi_2(q)$.

Théorème 10.2. *La loi du rapport $(X/p)/(Y/q)$ s'appelle loi de Fisher de paramètres p, q. Sa densité est donnée par*

$$f_{p,q}(x) = \frac{\mathbf{1}_{x>0}}{Z_{p,q}} \frac{\left(\frac{px}{px+q}\right)^{\frac{p}{2}} \left(1 - \frac{px}{px+q}\right)^{\frac{q}{2}}}{x} \tag{10.1}$$

où la constante $Z_{p,q}$ est $B(p/2, q/2)$, c'est-à-dire

$$Z_{p,q} = \int_0^1 u^{\frac{p}{2}-1}(1-u)^{\frac{q}{2}-1} du.$$

Le calcul est facile, puisque les lois du χ_2 ont une densité connue donnée par Équation 5.4. Soit φ une fonction test et soit $F = (X/p)/(Y/q)$. Alors, $\mathbb{E}[\varphi(F)]$ vaut

$$\frac{1}{C_p C_q} \int_0^\infty \int_0^\infty \varphi\left(\frac{uq}{vp}\right) e^{-\frac{u}{2}-\frac{v}{2}} u^{\frac{p}{2}-1} v^{\frac{q}{2}-1} du dv$$

avec $C_n = 2^{n/2}\Gamma(n/2)$. Dans l'intégrale en v, on pose $x = uq/vp$, de sorte que l'intégrale ci-dessus devient

$$\frac{(p/q)^{\frac{p}{2}}}{C_p C_q} \int_0^\infty \varphi(x) x^{\frac{p}{2}-1} \int_0^\infty e^{-\frac{vpx}{2q}-\frac{v}{2}} v^{\frac{p}{2}-1} v^{\frac{q}{2}} dv dx.$$

On reconnaît dans l'intégrale en v une fonction Gamma ; son intégrale est égale à

$$\frac{\Gamma(p/2+q/2)}{\left(\frac{px+q}{2q}\right)^{\frac{p+q}{2}}}.$$

L'espérance $\mathbb{E}[\varphi(F)]$ vaut donc

$$\frac{(p/q)^{p/2}\Gamma\left(\frac{p+q}{2}\right)}{C_pC_q(2q)^{\frac{p+q}{2}}}\int_0^\infty \varphi(x)\frac{x^{\frac{p}{2}-1}}{(px+q)^{\frac{p+q}{2}}}dx.$$

En simplifiant, on trouve exactement la densité donnée par Équation 10.1.

11 Théorie de l'information

Cours 8-9-10

12 Estimation de densité

Cours 11-12

Et après ?

nasuitenasute