Quelques élément de statistique
27 avril 2017
Pierre Rousselin

La dernière version de ce document est disponible sur le site :

https://www.math.univ-paris13.fr/~rousselin/

N'hésitez pas à me faire part de vos commentaires et à me signaler les coquilles, erreurs et imprécisions à l'adresse indiquée sur ce site.

1 Modèle statistique

Définition 1. Un modèle statistique est un espace mesurable $(\Xi, \mathcal{E})$ appelé ensemble des issues, muni d'une famille de probabilités $(P_\theta)_{\theta\in\Theta}$ indexée par l'ensemble des paramètres Θ. Une *observation* est une variable aléatoire X à valeurs dans $(\Xi, \mathcal{E})$ dont la loi est un élément de la famille $(P_\theta)_{\theta\in\Theta}$.

Remarque 2.
— Dans la suite de ce cours, on verra qu'il y a en général une distance canonique sur l'ensemble des issues Ξ et la tribu $\mathcal{E}$ sera de façon automatique la tribu borélienne pour la topologie associée à cette distance.
— On fera toujours l'hypothèse que le modèle est *identifiable*, c'est-à-dire que l'application $\theta \mapsto P_\theta$ est injective.
— Enfin, on supposera que le modèle est *paramétrique*, c'est-à-dire qu'il existe un entier d tel que Θ est inclus dans $\mathbb{R}^d$.

Exemple 3. Avant un référendum, on fait un sondage sur une population. On suppose qu'on peut assimiler le choix du panel de sondés à n tirages uniformes avec remise dans la population.

En notant 1 la réponse « oui » et 0 la réponse « non », on peut donc poser $\Xi = \{0\,;1\}^n$, la tribu $\mathcal{E}$ étant bien sûr la tribu discrète 2^Ξ. Soit θ la proportion des individus dans la population qui vont voter « oui » au référendum. L'observation $X = (X_1, X_2, \dots, X_n)$ est le n-uplet des intentions de vote des n sondés [1]. Les variables aléatoires $X_1, X_2, \dots, X_n$ sont indépendantes de même loi de Bernoulli $\mathcal{B}(1, \theta)$. L'ensemble des paramètres est $\Theta = [0\,;1]$ et la famille de lois de probabilité est

$$\left(\mathcal{B}(1,\theta)^{\otimes n}\right)_{\theta\in[0;1]}.$$

Définition 4. Lorsque l'observation X a la forme d'un n-uplet $X = (X_1, \dots, X_n)$, où les variables $X_1, X_2, \dots, X_n$ sont i.i.d. de même loi ν_θ, on dit que X a la forme d'un n-échantillon. [2]

L'objet des parties suivantes est, à partir d'*une* réalisation d'*une* observation, de donner des informations (mêmes partielles et soumises à des aléas) sur le paramètre θ permettant éventuellement d'aider à un *prise une décision*.

Avec ce formalisme, l'espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ sur lequel est définie l'observation X n'a pas d'importance. Pour se souvenir de la dépendance en le paramètre θ, on notera $\mathbb{P}_\theta$ la mesure de probabilité sur Ω, lorsque la loi de l'observation X est P_θ.

L'exemple précédent, bien qu'important, est l'un des plus simples possible. Cependant, il n'est pas systématique que l'observation ait la structure d'un n-échantillon même dans le cas où elle a la forme $(X_1, \dots, X_n)$. Il est possible que ses composantes ne soient pas indépendantes ou n'aient pas la même loi, comme c'est le cas par exemple dans le modèle linéaire gaussien.

1. Dans ce modèle de tirage avec remise, un même individu peut être, en théorie, sondé plusieurs fois. Si la taille de la population est assez grande par rapport à la taille n de l'échantillon, on peut assimiler un tirage sans remise à un tirage avec remises.

2. Bien sûr, dans ce cas, le modèle est du type $(E^n, \mathcal{A}^n)$, où E est l'espace d'arrivée de X_1 et la famille de probabilités est $P_\theta = \nu_\theta^{\otimes n}$.

Le paragraphe suivant est à prendre avec *beaucoup de précautions* car il est, à la maigre connaissance de l'auteur, assez personnel. Il s'agit de donner une définition alternative du modèle statistique pour pouvoir faire tendre la taille de l'échantillon vers l'infini de manière rigoureuse. Si vous n'êtes pas particulèrement obsessionnel, vous devriez passer directement à la deuxième partie.

Modèle canonique Nous construisons ici un espace de probabilités explicite sur lequel vivront les observations de toutes tailles. On procède comme dans la théorie des chaînes de Markov à temps discret, lorsque l'on construit la chaîne de Markov canonique.

Nous partons d'un espace mesurable $(E, \mathcal{A})$ et supposons donnée pour tout paramètre θ dans l'ensemble Θ une suite de probabilités $\left(P_\theta^{(n)}\right)_{n\geq 1}$ *cohérente*, c'est-à-dire vérifiant,

$$\forall n \geq 1,\ \forall A \in \mathcal{A}^n, \quad P_\theta^{(n+1)}\left(A \times E\right) = P_\theta^{(n)}\left(A\right).$$

Le théorème d'extension de Kolmogorov nous garantit alors qu'il existe une unique probabilité $\mathbb{P}_\theta$ sur l'espace Ω des suites indexées par $\{1, 2, \dots\}$ à valeurs dans E, muni de la tribu cylindrique [3] $\mathcal{F}$ telle que

$$\forall n \geq 1,\ \forall A \in \mathcal{A}^n, \quad \mathbb{P}_\theta\left(A \times E \times E \times \cdots\right) = P_\theta^{(n)}\left(A\right).$$

Ici, l'ensemble $A \times E \times E \times \cdots$ désigne l'ensemble des suites $(u_n)_{n\geq 1}$ d'éléments de E qui sont telles que le n-uplet $(u_1, u_2, \dots, u_n)$ appartienne à A.

Dans ce formalisme, on considère ensuite, pour tout entier $n \geq 1$, les applications $X_n : \Omega \to E$ et $X^{(n)} : \Omega \to E^n$ définies par,

$$\forall \omega = (\omega_1, \omega_2, \dots) \in \Omega, \quad X_n(\omega) = \omega_n \quad \text{et} \quad X^{(n)}(\omega) = (\omega_1, \omega_2, \dots, \omega_n).$$

Il découle de la construction précédente que la loi de $X^{(n)}$ est $P_\theta^{(n)}$. Nous appelons alors *observations* ces variables aléatoires et appelons le modèle $(\Omega, \mathcal{F}, \mathbb{P}_\theta, \theta \in \Theta)$ le *modèle canonique* associé à la famille de probabilités $\left(P_\theta^{(n)};\ n \geq 1,\ \theta \in \Theta\right)$.

2 Estimateurs

2.1 Définitions

Soit $(\Xi, \mathcal{E}, P_\theta, \theta \in \Theta)$ un modèle statistique et X une observation. Étant donnés un entier naturel k et une certaine fonction mesurable $g : \Theta \to \mathbb{R}^k$, on cherche à estimer la quantité $g(\theta)$, uniquement à l'aide de l'observation X.

Définition 5. Un *estimateur* de $g(\theta)$ est une variable aléatoire $\widehat{g}$ à valeurs dans $\mathbb{R}^k$, mesurable par rapport à X, c'est-à-dire de la forme $\widehat{g} = h(X)$, où h est une fonction mesurable de Ξ dans $\mathbb{R}^k$.

Cette définition est sémantiquement pauvre. Elle ne fait même pas intervenir la quantité à estimer et par exemple, les fonctions constantes sont toujours des estimateurs. Donnons un meilleur exemple.

Exemple 6. On se place dans le cas d'un n-échantillon $(X_1, \dots, X_n)$ où la loi commune est la loi de Poisson, de paramètre $\theta > 0$ inconnu. On cherche à estimer le paramètre θ (donc ici la fonction g est tout simplement l'identité). Un estimateur possible est

$$\widehat{\theta}_n = h_n(X_1, \dots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i =: \overline{X}_n.$$

3. c'est-à-dire la plus petite tribu rendant toutes les applications X_n définies ci-après mesurables

Pour l'instant nous n'avons aucune raison de préférer la moyenne empirique à, par exemple, l'estimateur constant égal à 0 ou la variable aléatoire $X_1^2 - 3X_2$. Notre travail va donc consister à donner une liste de critères et de propriétés souhaitées, permettant de choisir certains estimateurs plutôt que d'autres.

Définition 7. Dans le cas où l'observation à la forme $X = (X_1, \dots, X_n)$, on dit que l'estimateur [4] $\widehat{g}_n = h_n(X_1, X_2, \dots, X_n)$ est *consistant* (resp. *fortement consistant*) lorsque, pour tout paramètre $\theta \in \Theta$, la suite de variables aléatoires $\widehat{g}_n$ converge en $\mathbb{P}_\theta$-probabilité (resp. $\mathbb{P}_\theta$-p.s.) vers $g(\theta)$.

Exemple 8. Dans l'exemple précédent, d'après la loi forte des grands nombres, on sait que $\overline{X}_n$ converge $\mathbb{P}_\theta$-p.s. vers l'espérance de la loi de Poisson, qui vaut θ. L'estimateur donné est donc fortement consistant.

Le critère suivant n'est pas asymptotique.

Définition 9. Dans le cas où, pour tout $\theta \in \Theta$, l'espérance $\mathbb{E}_\theta\left[\|\widehat{g}\|\right]$ est finie, on appelle *biais* de l'estimateur $\widehat{g}$ la quantité
$$b(\theta) \coloneqq \mathbb{E}\left[\widehat{g}\right] - g(\theta).$$
L'estimateur $\widehat{g}$ est dit *sans biais* si, pour tout $\theta \in \Theta$, $b(\theta) = 0$.

Comme signalé dans l'ouvrage [RS12], la classe des estimateurs sans biais peut être vide.

Exemple 10. Pour l'exemple précédent, on a bien sûr, pour tout $n \geq 1$ et tout $\theta > 0$,
$$\mathbb{E}\left[\overline{X}_n\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i\right] = \theta,$$
donc la moyenne empirique est un estimateur sans biais du paramètre θ.

2.2 Méthode des moments

Dans cette sous-partie, on se limite au cas où l'observation à la forme d'un n-échantillon $X = (X_1, \dots, X_n)$ de loi commune P_θ sur l'espace mesuré $(E, \mathcal{A})$. L'exemple précédent est bien sûr généralisable. Si l'on suppose que E est inclus dans un espace $\mathbb{R}^m$, et que toutes les lois P_θ ont une espérance, la moyenne empirique $\overline{X}_n$ fournit toujours un estimateur sans biais et fortement consistant de la quantité $g(\theta) = \mathbb{E}_\theta\left[X_1\right]$. Plus généralement, si $\phi : E \to \mathbb{R}$ est une fonction mesurable telle que pour toute valeur du paramètre θ, $\mathbb{E}_\theta\left[|\phi(X_1)|\right] < \infty$, on peut estimer la quantité $g(\theta) = \mathbb{E}_\theta\left[\phi(X_1)\right]$ par la moyenne empirique
$$\widehat{g}_n = \frac{1}{n}\sum_{i=1}^{n}\phi(X_i).$$
Cet estimateur est sans biais et est fortement consistant d'après la loi forte des grands nombres.

Supposons maintenant que E est inclus dans $\mathbb{R}$ et que les lois P_θ ont toutes un moment d'ordre 2, c'est-à-dire,
$$\forall \theta \in \Theta, \quad \mathbb{E}_\theta\left[X_1^2\right] < \infty.$$
Toujours d'après la loi des grands nombres (appliquée deux fois), un estimateur fortement consistant de $\sigma^2 \coloneqq \mathbb{V}\mathrm{ar}_\theta\left[X_1\right]$ est donné par la *variance empirique*
$$\widehat{s}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i)^2 - \left(\overline{X}_n\right)^2. \tag{1}$$

4. On fait ici un petit abus de langage, on devrait dire *la suite d'estimateurs*.

Exercice 2.1. 1. Justifier la formule

$$\widehat{s}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2.$$

2. Montrer que le biais de cet estimateur vaut

$$b\left(\theta\right) = -\frac{1}{n}\sigma^2.$$

3. Montrer que la variance empirique « débiaisée »

$$\widehat{\sigma}_n^2 \coloneqq \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

est un estimateur sans biais et fortement consistant de la variance σ^2.

2.3 Maximum de vraisemblance

La méthode des moments fournit facilement des estimateurs dans le cadre asymptotique. Cependant, s'il est tout à fait naturel pour le mathématicien de faire tendre la taille de l'échantillon vers l'infini, cela l'est beaucoup moins pour le biologiste qui fait des expériences sur des animaux ou le laboratoire qui teste un nouveau médicament. Contrairement à une idée répandue, la statistique ne concerne pas seulement les grands échantillons. On cherche dans ce paragraphe, même lorsque la taille de l'échantillon est petite, à déterminer une valeur *vraisemblable* du paramètre.

On suppose ici que toutes les lois P_θ sont absolument continues par rapport à une mesure de référence λ sur $(\Xi, \mathcal{E})$. Le plus souvent, λ sera la mesure de comptage ou la mesure de Lebesgue. Notons, pour tout paramètre θ dans Θ, p_θ la densité de P_θ par rapport à λ.

Définition 11. La *vraisemblance* de θ par rapport à l'observation X est la densité p_θ appliquée à l'observation X, c'est-à-dire :

$$V_X\left(\theta\right) \coloneqq p_\theta\left(X\right).$$

L'estimateur du maximum de vraisemblance, lorsqu'il existe, est l'unique valeur $\widehat{\theta}$ de vraisemblance maximale :

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} V_X\left(\theta\right).$$

L'idée est que, faute de mieux, on cherche une valeur du paramètre qui rende notre observation « la plus probable possible ».

Exemple 12. On se place dans le cas d'un n-échantillon dont la loi commune est la loi uniforme sur $[0\,;\theta]$, où $\theta \in [0\,;\infty[$. Pour toute valeur de θ, la loi de l'échantillon a pour densité

$$p_\theta\left(x_1, x_2, \ldots, x_n\right) = \frac{1}{\theta^n}\prod_{i=1}^{n}\mathbf{1}_{[0;\theta]}\left(x_i\right).$$

On cherche donc la valeur de θ qui, à $X_1, \ldots, X_n$ fixés, rend maximal la vraisemblance

$$V_X\left(\theta\right) = \frac{1}{\theta^n}\prod_{i=1}^{n}\mathbf{1}_{\{X_i\leq\theta\}}$$

Ici, le produit est nul ou vaut 1, lorsque $\theta \geq \max_{1\leq i\leq n} X_i$, et dans ce cas la vraisemblance est d'autant plus grande que θ est petit, donc l'estimateur du maximum de vraisemblance est

$$\widehat{\theta}_n = \max_{1\leq i\leq n} X_i.$$

Exemple 13. Considérons maintenant le cas d'un n-échantillon dont la loi commune est la loi de Bernoulli, de paramètre inconnu $\theta \in \,]0\,;1[$. On commence par calculer la densité par rapport à la mesure de comptage (sur $\{0;1\}^n$). Pour tout vecteur $(x_1,\ldots,x_n) \in \{0;1\}^n$,

$$p_\theta(x_1,x_2,\ldots,x_n) = \mathbb{P}_\theta(X_1 = x_1,\ldots,X_n = x_n) = \theta^{\#\{i|x_i=1\}}(1-\theta)^{\#\{i|x_i=0\}}.$$

Comme souvent lorsqu'on travaille avec des n-échantillons, il est plus pratique de chercher à maximiser la *log-vraisemblance*

$$L_X(\theta) \coloneqq \log V_X(\theta) = \left(\sum_{i=1}^n X_i\right)\log\theta + \left(n - \sum_{i=1}^n X_i\right)\log(1-\theta).$$

Une rapide étude de fonction montre que l'estimateur du maximum de vraisemblance n'est autre que la moyenne empirique : $\widehat{\theta}_n = \overline{X}_n$.

Exercice 2.2. Encore des n-échantillons !

1. Écrire les détails manquants dans l'exemple précédent.
2. Dans le cas où la loi commune est la loi normale $\mathcal{N}(m,\sigma^2)$, *où σ est supposée connue*, montrer que l'estimateur du maximum de vraisemblance de m est la moyenne empirique.
3. Dans le cas où la loi commune est la loi normale $\mathcal{N}(m,\sigma^2)$, *où m est supposée connue*, montrer que l'estimateur du maximum de vraisemblance de σ^2 est la variance empirique.
4. Dans le cas de la loi normale $\mathcal{N}(m,\sigma^2)$ où aucun des paramètres n'est connu, montrer que l'estimateur du maximum de vraisemblance de (m,σ^2) est le couple formé de la moyenne empirique et de la variance empirique. On pourra étudier les points critiques de la log-vraisemblance.
5. Dans le cas où la loi commune est la loi géométrique de paramètre $\theta \in \,]0\,;1[$, montrer que l'estimateur du maximum de vraisemblance est l'inverse de la moyenne empirique.

Exercice 2.3. Retour sur l'exemple 12
On se place dans le cas d'un n-échantillon dont la loi commune est la loi uniforme sur $[0\,;\theta]$, où $\theta \in \,]0\,;\infty[$ est le paramètre du modèle. On rappelle que l'estimateur du maximum de vraisemblance est

$$\widehat{\theta}_n = \max_{1\le i\le n} X_i.$$

1. Soit $\varepsilon > 0$ assez petit. Calculer
$$\mathbb{P}_\theta\left(\theta - \widehat{\theta}_n > \varepsilon\right).$$
En déduire que l'estimateur $\widehat{\theta}_n$ est consistant.
2. Montrer qu'il est même fortement consistant. *On pourra utiliser le lemme de Borel-Cantelli.*
3. On s'intéresse maintenant à la vitesse de convergence (bref, on voudrait un analogue du théorème central limite). Montrer que pour tout $x > 0$,
$$\lim_{n\to\infty} \mathbb{P}_\theta\left(n\left(\theta - \widehat{\theta}_n\right) > x\right) = e^{-\frac{x}{\theta}}.$$
En déduire que la suite $n\left(\theta - \widehat{\theta}_n\right)$ converge en loi et en préciser la loi limite.

3 Intervalles de confiance

3.1 Définition

Définition 14. Soit $(\Xi,\mathcal{E},P_\theta,\theta\in\Theta)$ un modèle statistique, X une observation et une fonction $g:\Theta\to\mathbb{R}$. Soit α un réel de $[0;1]$. Un intervalle de confiance de $g(\theta)$, de niveau de confiance $1-\alpha$ est un intervalle I_α, dont les bornes sont des fonctions mesurables de l'observation, vérifiant

$$\forall\theta\in\Theta,\quad \mathbb{P}_\theta(g(\theta)\in I_\alpha) \ge 1-\alpha.$$

Lorsque cette inégalité est une égalité, on dit que le niveau de confiance est *exactement* égal à $1-\alpha$.

Pour donner une définition dans le cas asymptotique, on utilise le modèle canonique défini dans la première partie. Dans les applications qui suivent, le contexte est très clair donc vous pouvez sauter par-dessus cette définition.

Définition 15. Supposons construit le modèle canonique $(\Omega, \mathcal{F}, \mathbb{P}_\theta, \theta \in \Theta)$ de la première partie et la suite d'observations $X^{(n)}$, pour $n = 1, 2, \ldots$. Un intervalle de confiance asymptotique de $g(\theta)$ au niveau de confiance $1-\alpha$ est une suite d'intervalles $I_\alpha^{(n)}$, pour $n = 1, 2, \ldots$ où pour chaque n, les bornes de $I_\alpha^{(n)}$ sont des fonctions mesurables de l'observation $X^{(n)}$ et tels que

$$\forall \theta \in \Theta, \quad \liminf_{n\to\infty} \mathbb{P}_\theta \left(g(\theta) \in I_\alpha^{(n)} \right) \geq 1-\alpha.$$

On dit que le niveau de confiance est exactement égal à $1-\alpha$ lorsque la limite inférieure est une (vraie) limite et l'inégalité est une égalité.

Remarque 16. Dans les ouvrages [RS12] et [Tou99], sont définies les *régions de confiance*. Elles apparaissent lorsque la quantité $g(\theta)$ à estimer n'est pas dans $\mathbb{R}$.

Exemple 17. Commençons par le cas (un peu trop simple) du modèle gaussien à variance connue. On suppose donc que $\sigma > 0$ est connu à l'avance et on considère le cas d'un n-échantillon $(X_1, X_2, \ldots, X_n)$ de loi commune $\mathcal{N}(\theta, \sigma^2)$, où $\theta \in \mathbb{R}$ est le paramètre. On sait déjà que la moyenne empirique $\overline{X}_n$ est l'estimateur du maximum de vraisemblance de θ, qu'il est sans biais et fortement consistant, par la loi des grands nombres.

Dans ce cas très particulier, on connaît également *la loi de l'estimateur*. Comme les X_i sont gaussiennes indépendantes, $\overline{X}_n$ est encore gaussienne, de loi $\mathcal{N}\left(\theta, \frac{\sigma^2}{n}\right)$. On en déduit que la variable aléatoire[5]

$$Z_n \coloneqq \frac{\overline{X}_n - \theta}{\frac{\sigma}{\sqrt{n}}}$$

est normale centrée réduite. *En particulier, sa loi ne dépend plus du paramètre.* Pour obtenir un intervalle de confiance symétrique de niveau exactement égal à $1-\alpha$, on cherche la valeur z_α telle que

$$\mathbb{P}(-z_\alpha \leq Z_n \leq z_\alpha) = 1-\alpha.$$

Les valeurs $z_{0{,}05} \approx 1{,}960$ et $z_{0{,}01} \approx 2{,}576$ sont souvent utilisées. L'intervalle de confiance de niveau exactement égal à $1-\alpha$ est donc

$$I_\alpha = \left[\overline{X}_n - z_\alpha \frac{\sigma}{\sqrt{n}} \,;\, \overline{X}_n + z_\alpha \frac{\sigma}{\sqrt{n}}\right].$$

On pourrait de la même manière obtenir des intervalles de confiance dits unilatères, où l'une des bornes est infinie.

Exemple 18. Suite de l'exemple 12 Soit $\alpha \in [0\,;1]$. On souhaite construire un intervalle de confiance de θ de niveau de confiance exactement égal à α. On rappelle que $\widehat{\theta}_n \coloneqq \max_{1\leq i\leq n} X_i$ est l'estimateur du maximum de vraisemblance de θ. On cherche un réel ε qui dépend de α *mais surtout pas de* θ tel que

$$\mathbb{P}_\theta \left(\widehat{\theta}_n \leq \theta \leq \widehat{\theta}_n (1+\varepsilon) \right) = 1-\alpha.$$

La minoration est $\mathbb{P}_\theta$-presque sûre. Il suffit donc de calculer

$$\mathbb{P}_\theta \left(\frac{\widehat{\theta}_n}{\theta} \geq \frac{1}{1+\varepsilon} \right) = 1 - \mathbb{P}_\theta \left(\max_{1\leq i\leq n} \frac{X_i}{\theta} < \frac{1}{1+\varepsilon} \right) = 1 - \left(\frac{1}{1+\varepsilon} \right)^n,$$

5. appelée *z-score* dans les pays anglo-saxons, tandis que le terme $\frac{\sigma}{\sqrt{n}}$ est appelé *standard error* et est souvent abrégé en SE.

où pour la dernière égalité on a utilisé que les variables $\frac{X_i}{\theta}$ sont i.i.d. de loi uniforme sur $[0\,;1]$. On en déduit que l'intervalle $[\widehat{\theta}_n\,;\widehat{\theta}_n\alpha^{-\frac{1}{n}}]$ est un intervalle de confiance au niveau exactement $1-\alpha$ de θ. Numériquement, pour $n=25$ et $\alpha=5\%$, on trouve $\alpha^{-\frac{1}{n}}\approx 1,127$.

Dans les deux exemples précédents, la méthode a été la même. On commence par déterminer la loi de l'estimateur (ou sa loi limite dans le cas asymptotique) et on la transforme de façon à obtenir une loi de probabilité ne dépendant plus du paramètre. On cherche ensuite, de façon explicite ou en utilisant une table, le quantile approprié de cette dernière loi.

On va donner plus loin un intervalle de confiance asymptotique de l'espérance, dans le cas d'un n-échantillon où toutes les lois ont un moment d'ordre 2, en utilisant le théorème central limite. Mais avant cela nous allons faire un petit intermède probabiliste sur la convergence en loi.

3.2 Lemme de Slutsky

Commençons par une bien triste nouvelle... Il est en général *faux* que si deux suites de variables aléatoires $(X_n)_{n\geq 1}$ et $(Y_n)_{n\geq 1}$ convergent chacune en loi vers X et Y respectivement, alors le couple (X_n, Y_n) converge en loi vers (X, Y). En fait ce couple peut même ne pas converger du tout.

Exemple 19. Supposons que, pour tout $n\geq 1$, on ait

$$\mathbb{P}\left(X_{2n}=0, Y_{2n}=0\right)=\mathbb{P}\left(X_{2n}=1; Y_{2n}=1\right)=\frac{1}{2},$$
$$\mathbb{P}\left(X_{2n+1}=0, Y_{2n+1}=1\right)=\mathbb{P}\left(X_{2n+1}=1; Y_{2n+1}=0\right)=\frac{1}{2}.$$

Alors pour tout $k\geq 1$, les variables aléatoires X_k et Y_k prises individuellement ont toujours la même loi (de Bernoulli, de paramètre $\frac{1}{2}$). Cependant, *le couple* (X_k, Y_k) ne converge bien sûr pas en loi.

Heureusement, dans certains cas, le lemme de Slutsky peut nous tirer d'affaire.

Lemme 20 (Lemme de Slutsky)**.** *Soient* $(X_n)_{n\geq 1}$ *et* $(Y_n)_{n\geq 1}$ *deux suites de variables aléatoires, à valeurs, respectivement dans* $\mathbb{R}^p$ *et dans* $\mathbb{R}^q$ *On suppose que la suite* $(X_n)_{n\geq 1}$ *converge en loi vers* X *et que la suite* (Y_n) *converge en probabilité* vers une constante c. *Alors, la suite des* couples (X_n, Y_n) *converge en loi vers le couple* (X, c).

La démonstration classique, plus courte, utilise le théorème dit « Portmanteau » (voir le chapitre 10 de [RS12]). Mais celui-ci n'est pas mentionné dans le programme de l'agrégation externe... En voici une qui utilise le théorème de Lévy dont on rappelle la version suivante :

Théorème 21 (Théorème de Lévy)**.** *Soient* $X_1, X_2, \ldots, X_n, \ldots$ *et* X *des variables aléatoires à valeurs dans l'espace* $\mathbb{R}^d$ *euclidien*[6]. *Alors la suite* (X_n) *converge en loi vers* X *si et seulement si, pour tout vecteur* v *de* $\mathbb{R}^d$*,*

$$\mathbb{E}\left[e^{i<v,X_n>}\right]\xrightarrow[n\to\infty]{}\mathbb{E}\left[e^{i<v,X>}\right]$$

On peut trouver la preuve de ce résultat, par exemple, dans [LG06].

Démonstration du lemme de Slutsky. Soient u un vecteur de $\mathbb{R}^p$ et v un vecteur de $\mathbb{R}^q$. On doit montrer que

$$\mathbb{E}\left[e^{i(<u,X_n>+<v,Y_n>)}\right]\xrightarrow[n\to\infty]{}\mathbb{E}\left[e^{i(<u,X>+<v,c>)}\right].$$

6. mais pas nécessairement définies sur le même espace de probabilités

Comme c est une constante, c'est équivalent à

$$\mathbb{E}\left[e^{i(<u,X_n>+<v,Y_n-c>)}\right] \xrightarrow[n\to\infty]{} \mathbb{E}\left[e^{i<u,X>}\right].$$

On écrit ensuite

$$e^{i<v,Y_n-c>} = 1 + (\cos < v, Y_n - c > -1) + i \sin < v, Y_n - c > .$$

Il reste à montrer que le terme

$$\mathbb{E}\left[e^{i<u,X>} (\cos \langle v, Y_n - c)\rangle - 1)\right]$$

tend vers 0, et de même pour le terme en sinus. Soit $\varepsilon > 0$.

$$\begin{aligned}
&\left|\mathbb{E}\left[e^{i<u,X>} (\cos \langle v, Y_n\rangle - 1)\right]\right| \\
&\leq \mathbb{E}\left[\left|\cos \langle v, Y_n - c\rangle - 1\right|\right] \\
&= \mathbb{E}\left[\left|\cos \langle v, Y_n - c\rangle - 1\right| \mathbf{1}_{\{\|Y_n-c\|>\varepsilon\}}\right] + \mathbb{E}\left[\left|\cos \langle v, Y_n - c\rangle - 1\right| \mathbf{1}_{\{\|Y_n-c\|\leq\varepsilon\}}\right] \\
&\leq 2\mathbb{P}\left[\|Y_n - c\| > \varepsilon\right] + \|v\| \,\mathbb{E}\left[\|Y_n - c\| \,\mathbf{1}_{\{\|Y_n-c\|\leq\varepsilon\}}\right] \\
&\leq 2\mathbb{P}\left[\|Y_n - c\| > \varepsilon\right] + \varepsilon \|v\| \xrightarrow[n\to\infty]{} \varepsilon \|v\| ,
\end{aligned}$$

La majoration du terme de droite dans la troisième ligne provient de l'inégalité des accroissements finis appliquée à la fonction $x \mapsto 1 - \cos(x)$ puis de l'inégalité de Cauchy-Schwartz. La convergence vers 0 du terme en sinus est identique. □

On rappelle que la convergence en probabilité implique toujours la convergence en loi. Lorsque la convergence a lieu vers *une constante*, la réciproque est vraie.

Lemme 22. *Soit (X_n) une suite de variables aléatoires à valeurs dans un espace métrique (E, d), qui converge en loi vers une constante c de E. Alors (X_n) converge en probabilité vers c.*

Démonstration. Soit $\varepsilon \in \,]0\,;1[$. Pour tous réels x et y, on note $x \wedge y = \min\{x, y\}$.

$$\mathbb{P}(d(X_n, c) > \varepsilon) = \mathbb{P}(d(X_n, c) \wedge 1 > \varepsilon) \leq \frac{1}{\varepsilon}\mathbb{E}\left[d(X_n, c) \wedge 1\right],$$

d'après l'inégalité de Markov. Il suffit pour conclure de remarque que l'application

$$x \mapsto d(X_n, c) \wedge 1$$

est continue et bornée, donc par définition de la convergence en loi,

$$\mathbb{E}\left[d(X_n, c) \wedge 1\right] \xrightarrow[n\to\infty]{} \mathbb{E}\left[d(c, c) \wedge 1\right] = 0.$$ □

Pour conclure ces rappels de probabilité, rappelons une version faible[7] de ce que les anglo-saxons appellent le *continuous mapping theorem*.

Proposition 23 (Image continue d'une suite convergente de variables aléatoires)**.** *Soit (X_n) une suite de variables aléatoires à valeurs dans un espace métrique (E, d) et f une fonction continue de (E, d) dans un espace métrique (F, δ). Si (X_n) converge presque-sûrement (respectivement en probabilité, en loi) vers une variables aléatoire X, alors la suite $(f(X_n))$ converge presque-sûrement (respectivement en probabilité, en loi) vers $f(X)$.*

Exercice 3.1. 1. En utilisant uniquement la définition de la convergence presque-sûre et de la convergence en loi, démontrer la proposition précédente dans ces deux derniers cas.

7. la condition de continuité peut être affaiblie

2. Soit $\varepsilon > 0$. Pour tout $\alpha > 0$, on pose
$$A_{\alpha,\varepsilon} \coloneqq \{x \in E \mid \exists x' \in E,\ d(x,x') \leq \alpha \text{ et } \delta(f(x), f(x')) > \varepsilon\}.$$
Montrer que
$$\bigcap_{\alpha \downarrow 0} A_{\alpha,\varepsilon} = \emptyset.$$

3. Montrer, pour tout $\alpha > 0$, l'inclusion
$$\{\delta(f(X_n), f(X)) > \varepsilon\} \subset \{X \in A_{\alpha,\varepsilon}\} \cup \{d(X_n, X) > \alpha\}$$
et démontrer la proposition précédente dans le cas de la convergence en probabilité.

Finalement, en rassemblant les résultats de cette partie, on obtient :

Corollaire 24. *Soit (X_n) une suite de variables aléatoires à valeurs dans $\mathbb{R}$, convergeant en loi vers une variable aléatoire X et (Y_n) une suite de variables aléatoires à valeurs dans $\mathbb{R}$, convergeant en loi vers une* constante *c. Alors, les suite $(X_n + Y_n)$ et $(X_n Y_n)$ convergent en loi respectivement vers $(X + c)$ et cX. Si, de plus, $c \neq 0$, alors la suite $\left(\frac{X_n}{Y_n}\right)$ converge en loi vers $\frac{X}{c}$.*

3.3 Intervalle de confiance asymptotique dans le cas de la variance finie

On se place maintenant dans le cas d'un n-échantillon de variables aléatoires à valeurs réelles $(X_1, \ldots, X_n)$, de loi commune ν_θ. On suppose que, pour tout $\theta \in \Theta$, la variance σ_θ^2 de ν_θ est finie et non nulle. On note μ_θ son espérance. Alors d'après le théorème central limite, on a la convergence en loi
$$\sqrt{n}\frac{\overline{X}_n - \mu_\theta}{\sigma_\theta} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$
Comme, par ailleurs, la variance empirique $\widehat{s}_n^2$ définie par la formule 1 est un estimateur fortement consistant de la variance σ_θ^2, on a la convergence $\mathbb{P}_\theta$-presque-sûre (donc en probabilité, donc en loi)
$$\widehat{s}_n \xrightarrow[n\to\infty]{\mathbb{P}_\theta\text{-p.s.}} \sigma_\theta.$$
Le lemme de Slutsky permet alors d'établir la convergence en loi :
$$\sqrt{n}\frac{\overline{X}_n - \mu_\theta}{\widehat{s}_n} = \sqrt{n}\frac{\overline{X}_n - \mu_\theta}{\sigma_\theta}\frac{\sigma_\theta}{\widehat{s}_n} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$
On obtient ainsi, en reprenant les mêmes notations et la même démarche que dans l'exemple 17 un intervalle de confiance asymptotique au niveau de confiance exactement $1-\alpha$ de l'espérance μ_θ :
$$I_\alpha^{(n)} = \left[\overline{X}_n - z_\alpha\frac{\widehat{s}_n}{\sqrt{n}}\,; \overline{X}_n + z_\alpha\frac{\widehat{s}_n}{\sqrt{n}}\right].$$
Ce qui a été dit ici est encore valable en remplaçant la variance empirique par sa version « débiaisée » σ_n^2. On peut également, sans peine, construire des intervalles de confiance unilatères.

4 Échantillon de variables gaussiennes

Le modèle est maintenant celui d'un n-échantillon $(X_1, \ldots, X_n)$ dont la loi commune est la loi normale $\mathcal{N}(m, \sigma^2)$, où le paramètre est $\theta = (m, \sigma_2) \in \Theta = \mathbb{R} \times]0\,; +\infty]$. On sait déjà que le couple
$$\left(\overline{X}_n, \sigma_n^2\right)$$

est l'estimateur du maximum de vraisemblance de (m,σ^2), qu'il est fortement consistant, et que la loi de $\overline{X}_n$ est $\mathcal{N}\left(m, \frac{\sigma^2}{n}\right)$, mais pour pouvoir construire des intervalles de confiances (non asymptotiques), il va falloir encore un peu travailler.

Les définitions et propositions donnés dans cette partie sont loin d'être les plus générales, mais elles seront suffisantes pour nos applications. Dans la référence [RS12], les lois du χ^2 et de Student sont données avec un paramètre de décentrage. Le chapitre 2 de [Tou99] les exprime en terme de lois bêtas et gamma et présente également la loi du F de Fisher.

4.1 Loi du χ^2 et loi de Student

Définition 25. Soit d un entier naturel non nul. La loi du χ^2 à d degrés de libertés (notée $\chi^2(d)$) est la loi de la variable aléatoire

$$U = N_1^2 + N_2^2 + \cdots + N_d^2,$$

où les variables aléatoires $N_1, N_2, \ldots, N_d$ sont i.i.d. de même loi normale centrée réduite.

Soit Z une variable aléatoire de loi $\mathcal{N}(0,1)$, *indépendante de U*. La loi du t de Student à d degrés de liberté, notée $\mathcal{T}(d)$ est la loi de la variable aléatoire $\frac{Z}{\sqrt{U/d}}$.

La loi du χ^2 à d degrés de libertés admet pour densité la fonction

$$x \mapsto \frac{1}{2^{\frac{k}{2}}\,\Gamma\left(\frac{k}{2}\right)} x^{\frac{d}{2}-1} e^{-x/2} \mathbf{1}_{\{x\geq 0\}}, \tag{2}$$

tandis que la loi du t de Student admet pour densité la fonction

$$x \mapsto \frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)\sqrt{d\pi}} \left(1 + \frac{x^2}{d}\right)^{-\frac{d+1}{2}}. \tag{3}$$

On rappelle que la fonction Γ est définie sur $]0\,;\infty[$ par

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\,\mathrm{d}t.$$

Ici, il faut surtout regarder la vitesse de décroissance et ne pas, dans un premier temps, se soucier des constantes. On voit que la loi $\chi^2(d)$ admet des moments de tous ordres, tandis que la loi $\mathcal{T}(d)$ de Student décroît beaucoup moins rapidement et n'admet un moment d'ordre m que si $d > m$. Pour $d = 1$, on retrouve la loi de Cauchy, qui n'admet pas d'espérance.

La démonstration que nous donnons (en exercice guidé) utilise la loi gamma dont nous rappelons la définition. La loi gamma de paramètres $k > 0$ et $\theta > 0$, notée $\Gamma(k,\theta)$ est la loi de densité

$$x \mapsto \frac{1}{\theta^k \Gamma(k)} x^{k-1} e^{-x/k} \mathbf{1}_{\{x\geq 0\}} \tag{4}$$

Un simple changement de variable permet de vérifier que la fonction ci-dessus est bien d'intégrale 1.

Exercice 4.1. Quand faut y aller...

1. On commence par montrer une formule sur la fonction Γ. Soient a et b deux réels strictement positifs. On note :

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,\mathrm{d}t.$$

Montrer que

$$\Gamma(a)\Gamma(b) = \int_0^\infty u^{a-1}\left(\int_0^\infty v^{b-1}e^{-(u+v)}\,\mathrm{d}v\right)\mathrm{d}u,$$

puis que

$$\Gamma(a)\Gamma(b) = \int_0^\infty e^{-w}\left(\int_0^w (w-u)^{b-1}u^{a-1}\,\mathrm{d}u\right)\mathrm{d}w.$$

En déduire que

$$\Gamma(a)\Gamma(b) = \Gamma(a+b)B(a,b).$$

2. Soit $\theta > 0$. À l'aide du résultat précédent, montrer que si X et Y sont deux variables aléatoires *indépendantes*, de lois respectives $\Gamma(a,\theta)$ et $\Gamma(b,\theta)$, alors la variable aléatoire $X+Y$ a pour loi $\Gamma(a+b,\theta)$.
3. Montrer que la loi $\chi(1)$ est en fait la loi $\Gamma\left(\frac{1}{2},2\right)$.
4. Déduire des deux questions précédentes l'expression 2 de la densité de la loi $\chi^2(d)$.
5. Soit f une fonction mesurable positive et T une variable aléatoire de loi $\mathcal{T}(d)$. Justifier que

$$\mathbb{E}[f(T)] = C\int_{\mathbb{R}\times[0;\infty]} f\left(\frac{z}{\sqrt{u/d}}\right)e^{-z^2/2}u^{(d/2)-1}e^{-u/2}\,\mathrm{d}z\mathrm{d}u, \tag{5}$$

où la constante C vaut

$$C = \frac{1}{2^{\frac{k+1}{2}}\Gamma\left(\frac{k}{2}\right)\sqrt{\pi}}. \tag{6}$$

6. À l'aide du changement de variable $t = \frac{z}{\sqrt{\frac{u}{d}}}$ et $v = u$, montrer que l'espérance 5 vaut

$$\mathbb{E}[f(T)] = \frac{C}{\sqrt{d}}\int_{-\infty}^\infty f(t)\left(\int_0^\infty e^{-\frac{1}{2}(t^2+d)\frac{v}{d}}v^{\frac{d-1}{2}}\,\mathrm{d}v\right)\mathrm{d}t.$$

7. Exprimer en fonction de t l'intégrale

$$\int_0^\infty e^{-\frac{1}{2}(t^2+d)\frac{v}{d}}v^{\frac{d-1}{2}}\,\mathrm{d}v$$

en utilisant la définition de la loi gamma (4), puis établir la formule 3 de la densité de la loi de Student.

4.2 Théorème de Cochran

Théorème 26 (Théorème de Cochran)**.** *Soit X un vecteur gaussien de l'espace euclidien $\mathbb{R}^n$, de loi $\mathcal{N}(0,I_n)$, où I_n est la matrice identité de taille d. Soient $E_1,E_2,\ldots,E_r$, r sous-espaces vectoriels de $\mathbb{R}^d$, orthogonaux deux à deux, tels que $\mathbb{R}^d = E_1\oplus_\perp E_2\oplus_\perp\cdots\oplus_\perp E_r$. Alors, les projections orthogonales $\Pi_{E_1}X,\ldots,\Pi_{E_r}X$ de X sur ces sous-espaces sont des vecteurs gaussiens indépendants.*

Démonstration. On rappelle que les composantes d'un vecteur gaussien sont indépendantes si et seulement si sa matrice de covariance est diagonale.

Soit $\mathbf{b}$ une base orthonormale adaptée à la décomposition $\mathbb{R}^d = E_1\oplus_\perp E_2\oplus_\perp\cdots\oplus_\perp E_r$ et soit P la matrice (orthogonale) de passage de la base canonique vers $\mathbf{b}$. Dans la base $\mathbf{b}$, les coordonnées du vecteur gaussiens sont celle de $P^{\mathrm{T}}X$, dont la matrice de covariance est encore $P^{\mathrm{T}}I_nP = P^{\mathrm{T}}P = I_n$. □

Corollaire 27. *Soit* $X = (X_1, \ldots, X_n)$ *un n-échantillon dont la loi commune est* $\mathcal{N}(m, \sigma^2)$*. Alors, la moyenne empirique* $\overline{X}_n$ *a pour loi* $\mathcal{N}\left(m, \frac{\sigma^2}{n}\right)$ *et est indépendante de la variable aléatoire* $\widehat{s}_n^2 / (\sigma^2 n)$*, laquelle a pour loi* $\chi^2(n-1)$*. La variable aléatoire*

$$\frac{\overline{X} - \mu}{\sqrt{s_n^2/(n-1)}}$$

suit la loi de Student $T(n-1)$*.*

4.3 Intervalles de confiance

Soit, pour k entier naturel non nul, $T^{(k)}$ une variable aléatoire de loi $\mathcal{T}(k)$ et $U^{(k)}$ une variable aléatoire de loi $\chi^2(k)$. On note, pour tout $\alpha \in [0;1]$, $t_{k,\alpha}$ et $c_{k,\alpha}$ les uniques réels positifs tels que

$$\mathbb{P}\left(-t_{k,\alpha} \leq T^{(k)} \leq t_{k,\alpha}\right) = 1 - \alpha \quad \text{et} \quad \mathbb{P}\left(U^{(k)} \leq c_{k,\alpha}\right) = 1 - \alpha.$$

D'après ce qui précède, et en conservant les mêmes notations, l'intervalle

$$J_\alpha^{(n)} = \left[0 \; ; \; \frac{n\widehat{s}_n^2}{c_{n-1,\alpha}}\right]$$

est un intervalle de confiance de niveau exactement $1-\alpha$ de la variance σ^2, tandis que l'intervalle

$$I_\alpha^{(n)} = \left[\overline{X}_n - t_{n-1,\alpha}\sqrt{\frac{\widehat{s_n^2}}{n-1}} \; ; \; \overline{X}_n + t_{n-1,\alpha}\sqrt{\frac{\widehat{s_n^2}}{n-1}}\right]$$

très utilisé en pratique est un intervalle de confiance de niveau exactement $1-\alpha$ de l'espérance m.

5 Tests d'hypothèses

5.1 Définition

Soit $(\Xi, \mathcal{E}, P_\theta, \theta \in \Theta)$ un modèle statistique et soient Θ_0 et Θ_1 deux parties *disjointes* de Θ. Soit X l'observation. On note H_0 l'hypothèse dite « nulle » que θ appartient à Θ_0 et H_1 l'hypothèse dite « alternative » que θ appartient à Θ_1.

Un test de l'hypothèse H_0 contre l'hypothèse H_1 est une fonction du type $\Phi(X) = \mathbf{1}_{\{h(X) \in R\}}$, où h est une fonction mesurable à valeur dans un espace $\mathbb{R}^k$, appelée *la statistique du test* et R est une partie mesurable de $\mathbb{R}^k$ appelée *zone de rejet.*

Lorsque le résultat du test est 1, on dit qu'on rejette l'hypothèse nulle en faveur de l'hypothèse alternative. Lorsque le résultat est 0, on dit simplement *qu'on ne peut pas rejeter l'hypothèse nulle.*

Exemple 28. On mène une expérience de pile ou face sur une pièce de monnaie. On souhaite savoir si la pièce est bien équilibrée ou non. En codant 0 pour « pile » et 1 pour « face », le résultat de $n = 30$ lancers indépendants a donné un total de 9. Peut-on raisonnablement en conclure que cette pièce n'est pas équilibrée ? Le résultat d'un lancer suit la loi de Bernoulli de paramètre θ. On teste l'hypothèse $H_0 : \theta = 1/2$ contre l'hypothèse $H_1 : \theta \neq 1/2$. La statistique de test est la somme S_n

Sous l'hypothèse H_0, la variable aléatoire S_n suit la loi binomiale $\mathcal{B}(n, 1/2)$. Un calcul direct montre que

$$\mathbb{P}_{1/2}(10 \leq S_n \leq 30) \approx 0.957.$$

Le choix de la zone de rejet $R = [\![0\,;9]\!] \cup [\![21\,;30]\!]$ conduit ici à rejeter l'hypothèse nulle et donc à considérer la pièce comme truquée.

On voit que ce test conduit parfois (dans un peu plus de 4% des cas) à considérer comme truquées des pièces qui seraient bien équilibrées.

Définition 29. La fonction $\underline{\alpha} : \Theta_0 \to [0\,;1]$ définie par $\underline{\alpha}\,(\theta) = \mathbb{P}_\theta\,(h(X) \in R)$ est appelée risque de première espèce du test. Elle quantifie le risque de rejeter à tort l'hypothèse nulle. La *taille* du test est la borne supérieure de la fonction $\underline{\alpha}$. On dit qu'un test est de *niveau* α lorsque sa taille est inférieure ou égale à α.

La fonction $\underline{\beta} : \Theta_1 \to [0\,;1]$ définie par $\underline{\beta}\,(\theta) = \mathbb{P}_\theta\,(h(X) \notin R)$ est appelée risque de seconde espèce du test. Elle quantifie le risque de conserver l'hypothèse nulle à tort. La fonction $1 - \underline{\beta}$ est la *puissance* du test.

On peut faire un parallèle judiciaire en disant que le risque de première espèce mesure la probabilité de condamner à tort un innocent, tandis que celui de deuxième espèce mesure la probabilité de laisser un criminel en liberté.

Comme on l'a vu dans l'exemple précédent, on peut facilement construire des tests lorsque l'on dispose d'intervalles de confiance de la statistique du test.

Expliciter ce dernier point, parler de p-valeur.

5.2 Test z

Usage : (théorique) loi normale variance connue, (plus pratique) asymptotique dans le cas L^2, avec $\widehat{s}_n^2 \approx \sigma^2$ et un grand nombre de données.

5.3 Test t

Usage : nombre peu élevé de valeurs de loi à peu près normale. Plus « prudent » que le test z.

5.4 Test d'adéquation du χ^2

Exemple du dé.

Références

[LG06] Jean-François Le Gall. Intégration, probabilités et processus aléatoires. http://www.math.u-psud.fr/~jflegall/IPPA2.pdf, 2006. Polycopié de cours.

[RS12] Vincent Rivoirard and Gilles Stoltz. *Statistique mathématique en action.* Vuibert, 2012.

[Tou99] Paul S. Toulouse. *Thèmes de probabilités et statistiques.* Dunod, Paris, 1999.