Intervalles de confiance

Les probabilités s'attachent à décrire le comportement (souvent asymptotique) de fonctionnelles de variables aléatoires dont on connaît la loi. Une des deux grandes questions auxquelles s'intéresse la statistique est de décrire une loi de probabilité à partir d'observations supposées être des réalisations i.i.d. de cette loi inconnue. Le statisticien est une sorte de détective qui, face à de multiples individus, doit sélectionner un ou des suspects au vu d'indices dont aucun n'est une preuve.

Conseils de lecture. Ce petit condensé comporte beaucoup d'informations et dépasse le programme de tronc commun. Voici donc quelques pistes qui vous aideront à naviguer dans le texte.

- section 1 : à lire pour l'exemple 1. La définition abstraite des IdC peut être perturbante en première lecture mais elle sera claire lors de la seconde.
- section 2 : c'est un catalogue des IdC pour moyenne et variance dans le cas gaussien. Il faut retenir que, dans ce cadre, tout marche bien car on connaît plein de lois explicitement. Retenez juste l'apparition de la loi du χ^2 et apprenez à utiliser les tables de distributions se trouvant dans les livres.
- section 3 : retenir l'utilisation du TLC et la notion de probabilité de confiance asymptotique. On peut se contenter de la section 4 en première approximation.
- section 4 : à connaître sur le bout des doigts. C'est le truc le plus utilisé après le cas gaussien. Les non probabilistes peuvent oublier l'exercice 13.
- section 5 : réservé aux probabilistes.
- section 6 : très bien pour réviser l'inégalité de Markov et la transformée de Laplace, ça peut parfois servir à l'écrit...

1 Définitions et premier exemple

On se placera souvent dans un cadre paramétrique : soit $(\Omega, \mathcal{A})$ un espace mesurable et $(\mathbb{P}_\theta)_{\theta\in\Theta}$ une famille de probabilités sur $(\Omega, \mathcal{A})$ indexée par $\theta \in \Theta \subset \mathbb{R}^d$. La plupart du temps d vaudra 1 ou 2. Donnons tout de suite des exemples archi-classiques de telles familles :

$$(\mathcal{B}(\theta))_{\theta\in[0,1]}, \quad \{p = (p_1, \dots, p_k),\, p_i \geq 0,\ p_1 + \cdots + p_k = 1\}, \quad (\mathcal{E}(\theta))_{\theta\in\mathbb{R}_+}, \quad (\mathcal{N}(m, \sigma^2))_{(m,\sigma)\in\mathbb{R}\times\mathbb{R}_+}.$$

Ces familles sont respectivement associées à un sondage sur l'abstention, un premier tour d'élections présidentielles, des temps de connexion à un serveur informatique et une mesure entachée d'erreurs.

Étant donné un nombre $\alpha \in]0, 1[$ et un échantillon $X_1, \dots, X_n$ de loi $\mathbb{P}_\theta$, un intervalle (ou une région) de confiance pour le paramètre θ de probabilité de confiance $1 - \alpha$ est un intervalle (ou une région) qui dépend de l'échantillon (il est aléatoire) tel que la probabilité que cet intervalle contienne θ soit égale à $1 - \alpha$.

Exemple 1 (Échantillon gaussien de variance connue)**.** L'exemple le plus simple est le suivant : soit $X_1, \dots, X_n$ i.i.d. de loi $\mathcal{N}(\theta, 1)$ avec θ inconnu. Alors

$$\forall \theta \in \mathbb{R}, \quad \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \theta\right) \sim \mathcal{N}(0,1).$$

Or, on sait que si $Y \sim \mathcal{N}(0,1)$, alors $\mathbb{P}(|Y| \leq 1,96) = 0,95$. Ainsi,

$$\forall \theta \in \mathbb{R}, \quad \mathbb{P}_\theta\left(\theta \in \left[\overline{X}_n - \frac{1,96}{\sqrt{n}}, \overline{X}_n + \frac{1,96}{\sqrt{n}}\right]\right) = \mathbb{P}_\theta\left(\sqrt{n}\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \theta\right| \leq 1,96\right) = 0,95.$$

L'intervalle $\left[\overline{X}_n - \frac{1,96}{\sqrt{n}}, \overline{X}_n + \frac{1,96}{\sqrt{n}}\right]$ est donc un intervalle de confiance pour θ de niveau de confiance 0,95.

Voici à présent la définition mathématique d'un intervalle de confiance telle qu'on peut la trouver dans [Tas85] par exemple.

Définition 2. Soit $\alpha \in]0,1[$ donné ; on appelle région de confiance pour le paramètre θ, de niveau de confiance $1-\alpha$, la famille non vide de parties de Θ $C_{x_1,\dots,x_n}$ telle que

$$\forall \theta \in \Theta, \quad \mathbb{P}_\theta(\theta \in C_{X_1,\dots,X_n}) = 1-\alpha.$$

Exercice 3. Montrer que, dans l'exemple 1, l'intervalle obtenu est l'intervalle de confiance pour θ (de probabilité de confiance 0,95) le moins long.

Remarque 4. Très souvent, lorsque le paramètre θ est réel, la région construite se trouvera être un intervalle. On parlera alors d'intervalle de confiance.

Dans l'exemple 1, on a utilisé, pour construire l'intervalle de confiance, une variable aléatoire qui dépend de l'échantillon et du paramètre inconnu mais dont la loi ne dépend pas du paramètre. C'est ce que l'on appelle une fonction pivotale. Cette recherche de fonction pivotale sera l'une des clés pour déterminer des intervalles de confiance.

Souvent la situation ne sera pas aussi simple que dans l'exemple 1 et il faudra se contenter par exemple de fonctions asymptotiquement pivotales (c'est-à-dire que la loi de la fonction converge, quand la taille de l'échantillon tend vers l'infini, vers une loi qui ne dépend pas de θ). Les autres outils dont nous aurons besoin sont très variés : théorèmes limites (LFGN, TLC, convergence des quantiles empiriques), propriétés de lois classiques, inégalités de déviation à la Chernov...

Pour une définition complète des intervalles de confiance, des méthodes d'estimation etc... on pourra consulter [Tas85] et [Sap90]. Pour ceux qui ont déjà fait des statistiques, citons aussi [Mon82].

2 Le monde merveilleux des lois gaussiennes

L'exemple le plus commun en pratique est le cas d'un échantillon gaussien. Avec un peu de connaissance de lois dites classiques, on peut donner des intervalles de confiance pour estimer les paramètres de façon exacte (c'est-à-dire non asymptotique).

Notons $\overline{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ et $S_n^2 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$. Alors $\overline{X}_n$ est l'estimateur sans biais de variance minimale de m. Il converge presque sûrement vers m. D'autre part, S_n^2 converge aussi presque sûrement vers σ^2 mais S_n^2 est biaisé : $\mathbb{E}(S_n^2) = \dfrac{n-1}{n}\sigma^2$. On lui préfère parfois l'estimateur sans biais $\dfrac{n}{n-1}S_n^2$.

2.1 Estimation de la moyenne

2.1.1 si l'écart-type est connu

On utilise la statistique pivotale $\dfrac{\sqrt{n}}{\sigma}(\overline{X}_n - m) \sim \mathcal{N}(0,1)$. L'intervalle

$$\left[\overline{X}_n - \frac{\sigma x_{1-\alpha/2}}{\sqrt{n}} \ ; \ \overline{X}_n + \frac{\sigma x_{1-\alpha/2}}{\sqrt{n}}\right]$$

est un intervalle de confiance pour m une probabilité de confiance $1-\alpha$ où x_α est défini par la relation $\displaystyle\int_{-\infty}^{x_\alpha} e^{-x^2/2}\,\frac{dx}{\sqrt{2\pi}} = \alpha$ ou encore $x_\alpha = \Phi^{-1}(\alpha)$ en notant Φ la fonction de répartition de la loi gaussienne centrée réduite.

Exemple 5. Pamela est un mannequin célèbre dont le poids est strictement surveillé par Cruella. Cette charmante dame a investi un jour dans l'achat d'une balance *Harmonia* afin de connaître précisément le poids de sa protégée. Horreur : elle a constaté sur l'emballage de la balance que les fabriquants (d'honnêtes artisans suisses) admettaient que leur outil de mesure (nul n'est parfait) pouvait commettre des erreurs de mesure dont l'écart-type valait $0,1$ kg. En effet les pièces détachées ne sont pas toutes exactement identiques, leur montage n'est jamais parfait et le transport à travers les Alpes endommage parfois les balances. Ne faisant ni une ni deux, Cruella a, dès le lendemain, dévalisé le magasin en investissant dans l'achat de 99 nouvelles balances *Harmonia* et a forcé Pamela à sauter sur les 100 balances pendant que Cruella relevait scrupuleusement les 100 mesures. Résultat moyen des pesées : 55,4 kg. Donner à Cruella un intervalle de confiance pour le poids de Pamela de probabilité de confiance 0,95.

2.1.2 si l'écart-type est inconnu

On utilise le fait que $T = \dfrac{\overline{X}_n - m}{S_n}\sqrt{n-1}$ suit une loi de Student à $n-1$ degrés de liberté. Pour mémoire, la densité de la loi de Student à n degrés de liberté possède la densité :

$$f_{St(n)}(t) = \frac{1}{\sqrt{n}\mathrm{B}(1/2, n/2)}\left(1 + \frac{t^2}{n}\right)^{-(n+1)/2} \quad \text{où} \quad B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}.$$

Cette loi est tabulée : on peut trouver $t_{\alpha/2}$ tel que $\mathbb{P}(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) = 1-\alpha$ et l'intervalle de confiance pour m s'écrit alors

$$\left[\overline{X}_n - t_{\alpha/2}\frac{S_n}{\sqrt{n-1}} \ ; \ \overline{X}_n + t_{\alpha/2}\frac{S_n}{\sqrt{n-1}}\right].$$

Exemple 6. On a mesuré le poids de raisin par souches sur 10 souches prises au hasard dans une vigne. On a obtenu les résultats suivants (en kg) :

$$2,4\ ;\ 3,2\ ;\ 3,6\ ;\ 4,1\ ;\ 4,3\ ;\ 4,7\ ;\ 5,4\ ;\ 5,9\ ;\ 6,5\ ;\ 6,9.$$

Le vigneron se demande quel est le poids moyen de raisin par cep. Voici une question concrète de sa part. Aux mathématicien(nes) d'apporter une réponse claire et argumentée. Pour cela, on commence par ″modéliser″ le problème. Si les 10 ceps appartiennent à une même vigne, ils auront été traités aux pesticides, exposés au soleil et à la pluie et entretenus avec amour par le vigneron de façon équivalente. Cependant, des petits écarts sont inévitables : un peu d'ombre d'un côté, un pied de vigne dans un trou d'eau, un pied à l'hérédité plus solide que ses voisins... Ceci nous conduit à modéliser les mesures effectuées sur les 10 ceps par la réalisation de 10 variables aléatoires identiquement distribuées (les conditions sont globalement les mêmes) et indépendantes (les ceps ne sont pas sous le même arbre, dans le même trou d'eau...) de loi $\mathcal{N}(m,\sigma^2)$, où m représente le poids moyen de raisin par cep et σ son écart-type. Ces deux quantités sont a priori inconnues et l'on veut estimer m. Donner une estimation et un intervalle de confiance pour m.

2.2 Estimation de l'écart-type

2.2.1 si la moyenne est connue

La statistique $T=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i-m)^2$ est l'estimateur sans biais de variance minimale de σ^2 et $\dfrac{nT}{\sigma^2}$ suit une loi du χ^2 à n degrés de liberté. Pour mémoire, une variable aléatoire de loi du χ^2 à n degrés de liberté notée $\chi^2(n)$ admet pour densité

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

Les lois du χ^2 ne sont pas symétriques : les bornes de l'intervalle de confiance ne se déterminent pas tout à fait comme dans les cas précédents. On procède en général de la façon suivante : on détermine k_1 et k_2 tels que

$$\mathbb{P}(X\leq k_1)=\frac{\alpha}{2}\quad\text{et}\quad\mathbb{P}(X\geq k_2)=\frac{\alpha}{2}\quad\text{où}\quad X\sim\chi^2(n).$$

Ainsi $\mathbb{P}\left(k_1<\dfrac{nT}{\sigma^2}<k_2\right)=1-\alpha$. Un intervalle de confiance pour σ de probabilité de confiance $1-\alpha$ peut être choisi de la forme $\left[\sqrt{\dfrac{nT}{k_2}}\ ;\ \sqrt{\dfrac{nT}{k_1}}\right]$.

Exercice 7 (Pamela suite)**.** Expliquer comment les constructeurs de balance ont pu fournir une valeur de σ. Fournir aussi un intervalle de confiance.

2.2.2 si la moyenne est inconnue

On utilise le fait que $\frac{nS_n^2}{\sigma^2}$ suit une loi $\chi^2(n-1)$ (attention au nombre de degrés de liberté) et on procède comme dans le cas précédent.

Exercice 8 (Raisins suite)**.** Donner un intervalle de confiance de probabilité de confiance 0,95 pour le poids de raisin par pied de vigne dans l'exemple 6.

On pourra se reporter à [Tas85] pour les résultats et [DRV01] et [CDD99] pour des exemples d'utilisation en biologie notamment.

3 Les grands échantillons

Lorsque la loi des variables aléatoires n'est plus gaussienne, les calculs explicites de loi ne fonctionnent plus. On a alors recours aux deux théorèmes limites les plus célèbres : la loi forte des grands nombres et le théorème limite central. Si l'on dispose d'un échantillon $X_1, \dots, X_n$ de loi inconnue $\mathbb{P}_\theta$ indexée par sa moyenne θ et de variance connue, on utilise la LFGN et le TLC pour dire que l'intervalle

$$\left[\overline{X}_n - \frac{1,96\sigma}{\sqrt{n}} \ ; \ \overline{X}_n + \frac{1,96\sigma}{\sqrt{n}}\right]$$

est un intervalle de confiance de probabilité de confiance ASYMPTOTIQUE 0,95, c'est-à-dire que

$$\forall \theta \in \Theta, \quad \mathbb{P}_\theta\left(\theta \in \left[\overline{X}_n - \frac{1,96\sigma}{\sqrt{n}} \ ; \ \overline{X}_n + \frac{1,96\sigma}{\sqrt{n}}\right]\right) \xrightarrow[n\to\infty]{} 0,95.$$

Remarque 9. Le point important est que l'on ne sait pas à quelle vitesse a lieu cette convergence. Il existe des résultats théoriques pour contrôler cette erreur mais ils sont en général beaucoup trop pessimistes (car très généraux) et font intervenir des paramètres de la loi de l'échantillon, comme par exemple le moment d'ordre trois, dont on ne dispose pas. En pratique les livres conseillent d'utiliser cette approximation pour $n \geq 30$.

Dans le cas très fréquent où la variance est elle aussi inconnue (mais que l'on ne s'intéresse qu'à la moyenne) il faut utiliser un résultat supplémentaire connu sous le nom de lemme de Slutsky (voir [Tas85] ou [Mon82]) dont l'une des conséquences est que, si $\mathbb{E}(X_1^2) < \infty$,

$$\sqrt{n}\frac{\overline{X}_n - \mathbb{E}(X_1)}{S_n} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

On obtient donc un intervalle de confiance de la forme

$$\left[\overline{X}_n - \frac{1,96S_n}{\sqrt{n}} \ ; \ \overline{X}_n + \frac{1,96S_n}{\sqrt{n}}\right].$$

On introduit ainsi une deuxième approximation (en plus de celle du TLC) : la probabilité de confiance est doublement asymptotique...

4 Intervalles de confiance pour une proportion

Les méthodes employées sont les mêmes que dans le cas de grands échantillons de loi quelconque mais comme il s'agit d'un cas particulier archi-fréquent, il est bon de le détailler à part. On observe $X_1, \dots, X_n$ i.i.d. de loi $\mathcal{B}(\theta)$ avec $\theta \in [0,1]$ inconnu. On note toujours $\overline{X}_n$ la moyenne empirique de l'échantillon. Nous présentons les constructions de deux intervalles de confiance pour θ, toutes deux basées principalement sur l'utilisation du théorème limite central :

$$\sqrt{n}\frac{\overline{X}_n - \theta}{\sqrt{\theta(1-\theta)}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Exercice 10 (Méthode 1)**.** On considère la statistique asymptotiquement pivotale

$$\sqrt{n}\frac{\overline{X}_n - \theta}{\sqrt{\theta(1-\theta)}}.$$

Montrer que

$$\left[\frac{\overline{X}_n + \frac{r^2}{2n} - \frac{r}{\sqrt{n}}\sqrt{\frac{r^2}{4n} + \overline{X}_n(1-\overline{X}_n)}}{1+\frac{r^2}{n}} \; ; \; \frac{\overline{X}_n + \frac{r^2}{2n} + \frac{r}{\sqrt{n}}\sqrt{\frac{r^2}{4n} + \overline{X}_n(1-\overline{X}_n)}}{1+\frac{r^2}{n}}\right]$$

est un intervalle de confiance pour θ de probabilité de confiance asymptotique $1-\alpha$ si $r = \Phi^{-1}(1-\alpha/2)$ (où Φ est la fonction de répartition de la loi gaussienne centrée réduite).

Exercice 11 (Méthode 2)**.** Montrer que la statistique

$$\sqrt{n}\frac{\overline{X}_n - \theta}{\sqrt{\overline{X}_n(1-\overline{X}_n)}}.$$

est asymptotiquement pivotale. En déduire que

$$\left[\overline{X}_n - r\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}} \; ; \; \overline{X}_n + r\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right]$$

est un intervalle de confiance pour θ de probabilité de confiance asymptotique $1-\alpha$ si $r = \Phi^{-1}(1-\alpha/2)$ (où Φ est la fonction de répartition de la loi gaussienne centrée réduite).

Exemple 12 (taux de germination)**.** On dispose 40 graines de tournesol issues d'un lot de plusieurs tonnes sur du papier buvard humide. Au bout de huit jours, on compte 36 germes normaux (c'est-à-dire ayant évolué favorablement). Que dire du taux de germination du lot complet ? La deuxième méthode fournit l'intervalle $[0,807 \; ; \; 0,993]$ pour une probabilité de confiance (asymptotique) de 0,95.

On trouvera les résultats et l'exemple ci-dessus dans [DRV01].

Exemple 13. Nicolas S., Ségolene R. et François B. se voient tous en haut de l'affiche. Pour savoir quelles sont leurs chances, ils ont commandé un sondage (à trois c'est moins cher) au neveu de Nicolas (encore moins cher). Celui-ci a attrapé son annuaire et appelé 1000 personnes pour leur demander si leur choix se portait sur Nicolas (taper 1), Ségolene (taper 2), François (taper 3) ou un autre y compris blanc et nul (taper 4). Il a obtenu le résultat suivant :

$$N_1 = 200 \ ; \ N_2 = 180 \ ; \ N_3 = 20 \ ; \ N_4 = 600.$$

Donner des intervalles de confiance de pour chacune des proportions. Est-ce gagné pour Nicolas avec une probabilité de confiance 0,95 ? Montrer que l'on peut donner une région de confiance pour (p_1, p_2, p_3, p_4) de la forme d'un ellipsoïde. On utilisera pour cela le lemme de Slutsky et le résultat de convergence

$$n \sum_{i=1}^{4} \frac{(N_i/n - p_i)^2}{p_i} \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(3),$$

qui découle du TLC multidimensionnel pour la loi multinomiale.

5 Utilisation des quantiles empiriques

Cette partie fait appel à des notions délicates de probabilité. Il peut être omis par les lecteurs non spécialistes.

Définition 14. Soit X une variable aléatoire réelle. Pour $q \in [0,1]$, on dit que x_q est *un* quantile d'ordre q de X si

$$\mathbb{P}(X \leq x_q) \geq q \quad \text{et} \quad \mathbb{P}(X \geq x_q) \geq 1 - q.$$

Pour $q = 1/2$, on parle de médiane. Pour $q = 1/4$ et $q = 3/4$ on parle de premier et troisième quartiles...

Exercice 15. À quelle condition le quantile d'ordre p est-il unique ? Déterminer la fonction quantile de la loi exponentielle de paramètre λ et la médiane de loi gaussienne. Que dire des médianes de la loi de Bernoulli $\mathcal{B}(p)$?

Définition 16. Soit $(X_1, \ldots, X_n)$ un échantillon de loi μ, de fonction de répartition F continue. On appelle statistique d'ordre de l'échantillon le n-uplet $(X_{(1)}, \ldots, X_{(n)})$ des variables aléatoires ordonnées par ordre croissant. En particulier $X_{(1)}$ est appelée première statistique d'ordre...

L'hypothèse F continue assure que, presque sûrement, le n-uplet $(X_1, \ldots, X_n)$ appartient à l'ensemble

$$C = \{(x_1, \ldots, x_n) \in \mathbb{R}^n, \quad i \neq j \Longrightarrow x_i \neq x_j\},$$

ou encore que, presque sûrement, il existe une unique permutation de $\{1, \ldots, n\}$ qui envoie $(X_1, \ldots, X_n)$ dans l'ensemble

$$D = \{(x_1, \ldots, x_n) \in \mathbb{R}^n, \quad x_1 < \cdots < x_n\}.$$

Exercice 17 (Densité des statistiques d'ordre)**.** On suppose que la loi μ admet une densité f strictement positive.

1. Montrer que la loi de la permutation ci-dessus est la loi uniforme sur S_n.
2. Déterminer la densité de $X_{(1)}$ et $X_{(n)}$.
3. Montrer que la variable aléatoire $X_{(i)}$ (dans un échantillon de taille n) admet pour densité

$$f_{(i)}(x) = iC_n^i f(x)F(x)^{i-1}(1-F(x))^{n-i}.$$

Théorème 18 (Convergence p.s. des quantiles empiriques)**.** *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires rélles i.i.d. de fonction de répartition avec un unique quantile en p. Alors*

$$X_{([nq])} \xrightarrow[n\to\infty]{} x_q \quad p.s.$$

C'est une conséquence du théorème de Glivenko-Cantelli.

Théorème 19 (Glivenko-Cantelli)**.** *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires réelles i.i.d. de loi μ de fonction de répartition F. Notons $F_n(x) = Card\big(X_i \leq x,\ i = 1,\dots,n\big)/n$ la fonction de répartition empirique de l'échantillon. Alors*

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n\to\infty]{} 0 \quad p.s.$$

Théorème 20 (Convergence en loi des quantiles empiriques)**.** *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires réelles i.i.d. de fonction de répartition continue et de densité f. Alors, pour $q \in]0,1[$,*

$$\sqrt{n}\big(X_{([nq])} - x_q\big) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{q(1-q)}{f(x_q)^2}\right).$$

Exercice 21 (Convergence en loi de la médiane)**.** Soit μ une probabilité admettant une densité f strictement positive. On note m la médiane de μ. On suppose que la taille de l'échantillon observé n est impaire et on l'écrit $n = 2p-1$.

1. À l'aide du théorème de Glivenko-Cantelli, montrer que $X_{(p)}$ converge presque sûrement vers m.
2. Déterminer la densité de $\sqrt{2p-1}(X_{(p)} - m)$.
3. En déduire que $\sqrt{2p-1}(X_{(p)} - m)$ converge en loi vers une loi normale dont on précisera les paramètres.

Exemple 22. Les traders de Wall Street ont remarqué que les évolutions boursières étaient très irrégulières et mettaient en évidence des fluctuations d'amplitude colossales. En notant $s_n = \ln(p_{n+1}/p_n)$ le logarithme du rapport du prix d'une action entre le jour n et le jour $n+1$, ils ont proposé la modélisation de l'évolution de s par des variables aléatoires i.i.d. $S_1,\dots,S_n$ de loi de Cauchy translatée, c'est-à-dire que l'on suppose que la loi de S_1 admet pour densité

$$\forall x \in \mathbb{R}, \quad f_\theta(x) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2},$$

avec $\theta \in \mathbb{R}$ inconnu. Proposer un intervalle de confiance pour θ.

Exercice 23. On considère un échantillon $X_1, \dots, X_n$ de variables aléatoires i.i.d. de loi $\mathcal{E}(\lambda)$.

1. Déterminer la loi de $X_{(1)}$.
2. En déduire que le théorème 20 est un peu faux sur les bords (*i.e.* pour $q = 0$ ou $q = 1$) !

On trouvera la correction de l'exercice 17 dans [CGCDM99][p. 53]. Pour les théorèmes 18, 19 et 20 on pourra consulter [Tas85].

6 Inégalité de Chernov

Ce chapitre utilise des notions de probabilités qui font partie du tronc commun et les techniques employées sont très classiques. Elles seront réemployées à de nombreuses occasions.

On souhaite à présent être en mesure de fournir des intervalles de confiance non asymptotiques, c'est-à-dire qui n'utilisent pas un théorème limite pour lequel on ne contrôle pas la vitesse de convergence. Pour cela, on va minorer la probabilité de confiance par une quantité exacte dépendant de n. On assure ainsi ses arrières en prenant une marge de sécurité (espérons que c'est ce que font les constructeurs de centrales nucléaires et les fabriquants de médicaments). La méthode qui repose sur les propriétés d'intégrabilité de la loi de l'échantillon est une amélioration de l'inégalité de Tchebychev qui fournit la majoration suivante :

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X_1)\right| \geq r\right) \leq \frac{\mathbb{V}(X_1)}{nr^2}.$$

Ceci peut encore s'interpréter dans notre contexte en terme d'intervalle de confiance pour l'espérance :

$$\mathbb{P}\big(\mathbb{E}(X_1) \in \big[\overline{X}_n - r \ ; \ \overline{X}_n + r\big]\big) \geq 1 - \frac{\mathbb{V}(X_1)}{nr^2}.$$

On a ainsi construit un intervalle de confiance pour $\mathbb{E}(X_1)$ de probabilité de confiance minorée par $\mathbb{V}(X_1)/(nr^2)$. Cette inégalité, valable dès que X_1 est de carré intégrable peut être fortement améliorée sous des conditions d'intégrabilité plus fortes.

Exercice 24. Soit $\mu = (1/2)\delta_{-1} + (1/2)\delta_1$ et $X_1, \dots, X_n$ un échantillon de loi μ.

1. Calculer la transformée de Laplace de μ définie par $L_\mu(t) = \mathbb{E}(e^{tX})$ et montrer que $\ln L_\mu(t) \leq t^2/2$.
2. Montrer que, pour tout $\lambda > 0$,
$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X_1)\right| \geq r\right) = 2\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq r\right) = 2\mathbb{P}\Big(e^{\lambda \sum_i X_i} \geq e^{\lambda n r}\Big) \leq 2e^{-n(\lambda r - \ln L_\mu(\lambda))}.$$
3. Optimiser en $\lambda > 0$.
4. En déduire un intervalle de confiance non asymptotique pour θ dans le modèle suivant : pour tout $\theta \in \mathbb{R}$, $\mathbb{P}_\theta = (1/2)\delta_{-1+\theta} + (1/2)\delta_{1+\theta}$.

Le principe général, qui suit le raisonnement de l'exercice 24, fonctionne dès que la mesure μ considérée admet une transformée de Laplace $L_\mu(t) = \mathbb{E}(e^{tX})$ définie sur un voisinage de l'origine. Ceci assure que μ a tous ses moments finis et donc de bonnes qualités de décroissance que l'on va chercher à utiliser au mieux.

Exercice 25 (Un peu de transformée de Laplace)**.**

1. Montrer que le domaine de définition de la transformée de Laplace d'une mesure μ est un intervalle (un convexe) qui contient l'origine. Peut-il être réduit à $\{0\}$?
2. Montrer que L_μ est une fonction convexe et log-convexe (son logarithme est une fonction convexe).
3. Soit $X_1, \dots, X_n$ un échantillon de loi μ. Montrer que, pour tous $r \geq 0$ et $\lambda > 0$,
$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}(X_1) \geq r\right) \leq \exp\left(-n \sup_{\lambda>0}(\lambda r + \lambda \mathbb{E}(X_1) - \ln L_\mu(\lambda))\right),$$
et que le supremum est fini et atteint dès que $L_\mu(\lambda_0)$ est fini pour un certain $\lambda_0 > 0$.
4. Calculer les transformées de Laplace des lois $\mathcal{N}(m, \sigma^2)$, $\mathcal{B}(p)$, $\mathcal{P}(\lambda)$ et $\mathcal{E}(\lambda)$ en précisant leurs domaines de définition.
5. Déterminer une inégalité de déviation pour la moyenne empirique pour chacun de ces cas particuliers.
6. Déduire de la question précédent que dans tous les cas particuliers, il existe une constante c telle que
$$\limsup_{n\to\infty} \mathbb{P}\left(\sqrt{n}\left|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}(X_1)\right| \geq u\right) \leq 2e^{-cu^2}.$$
Que vous inspire ce résultat ?

Remarque 26. La majoration de la probabilité de déviation via la transformée de Laplace est le premier pas de ce que les probabilistes appellent la théorie des grandes déviations. Nous venons de montrer en particulier que
$$\lim_{n\to\infty} \frac{1}{n} \ln \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}(X_1) \geq r\right) \leq -\sup_{\lambda>0}(\lambda r + \lambda \mathbb{E}(X_1) - \ln L_\mu(\lambda)).$$
On peut en fait établir que cette majoration est en fait une égalité, c'est-à-dire que l'inégalité de Chernov donne le bon ordre de grandeur (dans l'exponentielle).

Pour retrouver tous les résultats des exercices proposés ici, on pourra consulter le premier chapitre [DZ98] qui, bien qu'écrit en anglais, est très facile à lire.

Références

[CDD99] F. Couty, J. Debord et F. Daniel – *Probabilités et statistiques*, Dunod, 1999.

[CGCDM99] M. Cotrell, V. Genon-Catalot, C. Duhamel et T. Meyre – *Exercices de probabilités*, Cassini, 1999.

[DRV01] J.-J. Daudin, S. Robin et C. Vuillet – *Statistique inférentielle*, Presses Universitaires de Rennes, 2001.

[DZ98] A. Dembo et O. Zeitouni – *Large deviations techniques and applications*, second éd., Applications of Mathematics (New York), vol. 38, Springer-Verlag, 1998.

[Mon82] A. Monfort – *Cours de statistique mathématique*, Économia, 1982.

[Sap90] G. Saporta – *Probabilités, analyse de données et statistique*, Éditions Technip, 1990.

[Tas85] P. Tassi – *Méthodes statistiques*, Économia, 1985.