1 Miniprojet du cours 4M074

L'objectif de ce mini-projet est l'illustration d'un résultat théorique, d'une notion ou d'une méthode de la probabilité ou de la statistique par une application web interactive, à l'image de ce qui est fait sur le site web du laboratoire LPSM `http://simulations.lpsm.paris`. Dans ce projet, il faut, d'une part, faire preuve de votre maîtrise de R, et d'autre part, faire preuve de beaucoup de pédagogie.

Choix du sujet

A priori, vous êtes libre de travailler sur le théorème ou la méthode probabiliste ou statistique de votre choix. Si vous êtes à court d'idées, vous pouvez choisir un sujet de la liste ci-dessous. Les sujets proposés dans cette liste ont de degrés de difficulté très variés, qui sera pris en compte dans la note de projet. Néanmoins, il faut que votre choix soit validé par votre chargé de TP, ou pour les étudiants en télé-enseignement, par T. Rebafka (`tabea.rebafka@upmc.fr`). Vous devez choisir votre sujet et le faire valider au plus tard le **5 avril**.

Création d'un Rmarkdown ou d'un notebook

Avant de vous lancer dans le développement d'une application web, vous devez élaborer un simple Rmarkdown ou un notebook Jupyter dans lequel vous présentez votre théorème/méthode/problème. Cette présentation du sujet doit être très soignée et très pédagogue, accessible à un large public d'étudiants. Nous accorderons beaucoup d'importance à cette partie du projet. Ensuite, vous rajoutez dans ce notebook des simulations pertinentes qui illustrent bien votre théorème/méthode/problème et qui aideront le lecteur bien comprendre le résultat.

Création d'une application web

On utilisera Shiny disponible dans RStudio : cliquer sur File → New File → Shiny Web App pour voir un exemple. Familiarisez-vous avec Shiny et consulter les sites d'aide sur Shiny :

— `http://shiny.rstudio.com`
— `https://superstatisticienne.fr/`
— `https://stackoverflow.com/`

Ensuite, créez votre propre application web en adaptant le contenu de votre Rmarkdown ou notebook.

Consignes du projet

- Vous avez le choix entre l'anglais et le français pour votre application web.
- Pour les étudiants en présentiel, vous devez préparer le projet **en binôme**.
- Faites valider votre choix de sujet au plus tard le **5 avril**.
- Le projet (le fichier Rmarkdown ou le notebook ainsi que le fichier Shiny) est à rendre sur Moodle au plus tard le soir du **11 avril**.
- Les étudiants en présentiel présenterons leur travail le **jeudi 12 avril** entre 14h et 18h devant tous les étudiants. La présentation ne durera que 5 minutes par projet. Les autres détails seront données ultérieurement.
- Les projets les plus réussis seront mis en ligne sur la page `http://simulations.lpsm.paris` (avec l'accord des étudiants concerncés).
- La note de projet prendra en compte la qualité pédagogique et l'originalité de l'application web, la difficulté du sujet et la présentation orale (pour les étudiants en présentiel).

2 Sujets

Méthodes et résultats du cours 4M074

1. Illustrer la méthode de rejet pour la simulation des pseudo-variables aléatoires p. ex. pour la loi Gamma.
2. Illustrer la méthode de simulation d'un processus de Poisson homogène ou inhomogène par la méthode de thinning. Dans le cas homogène illustrer les propriétés de convergence de N_t/t par exemple (cf. Guyader (2007)).
3. Illustration du théorème de Glivenko-Cantelli.
4. Illustrer l'approximation de la loi normale par la loi de Student. Plus précisément, on a
$$t_q \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \qquad q \to \infty,$$
où t_q désigne la loi de Student à q degrés de liberté (cf. TD 6, Ex. 2) Ce résultat implique aussi la convergence de certaines caractéristiques de la loi de Student (comme la moyenne, variance, le coefficient d'asymétrie, le coefficient d'aplatissement (kurtosis), des quantiles) vers la moyenne/ variance/... de la loi normale standard.
5. Expliquer et illustrer le principe du bootstrap et les erreurs d'approximation associées.
6. Expliquer et illustrer la différence entre le bootstrap paramétrique et non paramétrique.
7. Illustrer des cas où le bootstrap ne marche pas et expliquer pourquoi.
8. Illustrer le résultat sur l'approximation de toute densité continue par un mélange gaussien (cf. Partie 2, Chapitre 3).

Du cours Statistique 4M015

9. Soient $X_i, \sim U[0,\theta], i = 1, 2, \ldots$ i.i.d. Illustrer les différentes vitesses de convergence de l'estimateur de maximum de vraisemblance et de l'estimateur par la méthode des moments de θ (cf. cours *Statistique*, TD 3, Exercice 3 disponible sur le site `http://www.lsta.upmc.fr/guyader/statM1.html`).
10. Illustrer les performances des différents intervalles de confiance pour le paramètre de la loi de Bernoulli (IC asymptotique, par l'inégalité de Hoeffding et de Tchebychev) (Guyader (2017), Chapitre 1.3.2).
11. Illustrer la convergence des quantiles empiriques vers la quantile théorique quand la fonction de répartition F des données est strictement croissante (cf. Guyader (2017), Chapitre 2). Montrer qu'on n'a pas convergence, si F n'est pas strictement croissante.
12. Montrer que lorsque r est grand, la loi $\Gamma(r, \lambda)$ ressemble à une loi normale (Guyader (2017), p.51).
13. Expliquer la notion de la puissance d'un test. De quel facteurs dépendent la puissance d'un test ? Montrer l'allure typique de la puissance d'un test $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$ (bilatéral) ou $H_1 : \theta > \theta_0$ (unilatéral).
14. Illustrer la "compétition" du risque de première et de seconde espèce d'un test. Par exemple, pour tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$ avec $\theta_0 < \theta_1$ considérer une région de test de la forme $\{\hat{\theta} > c\}$. Etudier comment le risque de première et de seconde espèce dépendent de la constante c.
15. Pour tester la significativité de plusieurs coefficients dans un modèle linéaire gaussien, il existe le test de Fisher et une procédure de test qui repose sur la procédure de Bonferroni (cf. Rebafka (2017), Chapitre 5). Comparer la qualité de ces deux tests (niveau et puissance des tests) dans différents scénarios. Lequel est meilleur ?
16. La théorie de l'analyse de la variance est développée sous des hypothèses fortes (loi gaussienne, homoscédasticité etc.) qui ne sont pas toujours vérifiées en pratique. Illustrer la robustesse (ou non robustesse) du test par rapport aux hypothèses du modèle (différentes variances par groupes ; des observations de loi discrète etc.)

Du cours statistique bayésienne 4M072

17. Illustrer le théorème de Bernstein-von Mises par exemple dans le cas de l'expérience de Bayes-Laplace : vraisemblance binomiale, a priori Beta(a, b) ; a posteriori beta qui s'approche d'une $\mathcal{N}(\hat{\theta}_{MV}, \frac{1}{n} I(\theta_0)^{-1})$. On peut jouer avec n et les paramètres a, b, (et possiblement θ_0) et en sortie la densité a posteriori et la distance L^1 (éventuellement approchée) entre les densités.

Modèles probabilistes avancés

18. Simuler un processus de branchement Z_t (en temps continu par exemple), dont l'espérance vaut $e^{\alpha t}$ pour un certain $\alpha > 0$ (à calculer) et étudier la limite de la martingale $W_t := e^{-\alpha t} Z_t$. Quand on simule plusieurs trajectoires indépendantes de Z, on voit bien que le comportement asymptotique est $e^{\alpha t}$ mais avec un préfacteur différent à chaque fois (on peut passer en log aussi pour le voir encore mieux). La distribution empirique de ces préfacteurs doit converger vers la loi de W_∞ (Lambert, 2008).
19. Simulation de la loi Beta comme loi limite d'une chaîne de Markov (cf. Annexe pour les détails).

Méthodes statistiques avancées

20. Le k-means (Hastie et al. (2001), pp. 509) est une des méthodes de classification des observations en un petit nombre de groupes les plus utilisées en machine learning. Illustrer le k-means p. ex. sur des données bidimensionnelles d'un mélange gaussien

Matrices aléatoires j'y connais pas grande chose ; le dernier m'a l'air dur ; j'ai demandé à Thierry d'ajouter des réf

20. Si on tire une grande matrice hermitienne à coefficients gaussiens, la mesure empirique de ses valeurs propres est (avec grande probabilité presque) égale à la loi du demi-cercle de Wigner ;
21. Si on tire une grande matrice unitaire uniformément (par exemple en prenant une matrice à coefficients i.i.d. gaussiens et en prenant la partie unitaire de sa décomposition polaire, ou une des deux matrices unitaires de sa décomposition en valeurs singulières), ses valeurs propres se répartissent uniformément sur le cercle unité du plan complexe ;
22. Si on tire encore une matrice unitaire uniformément, mais pas forcément très grande, l'ensemble de ses valeurs propres forme un processus ponctuel déterminantal sur le cercle unité du plan complexe : c'est un processus ponctuel dont l'intensité est la mesure uniforme sur le cercle, mais avec une répulsion entre les points, qui fait qu'ils sont beaucoup plus également répartis qu'un échantillon i.i.d. de la même taille ; la probabilité d'avoir un gros paquet de points serrés dans un coin est infime ;

3 Annexe

Détails du sujet n°19

Une particule se déplace sur une droite de la manière suivante : Soient deux réels positifs $\gamma > 0$ et $\delta > 0$. À l'instant $t = -1$, elle se trouve à la position $X_{-1} = 0$ et en $t = 0$ elle se trouve en $X_0 = 1$. À l'instant $t \geq 1$, elle passe à la position X_t telle que la v.a. $\dfrac{X_t - X_{t-1}}{X_{t-1} - X_{t-2}}$ suit la loi Beta de paramètres $(\delta, 1)$ si t est impair et de paramètre $(\gamma, 1)$ si t est pair.

Simuler les déplacements de la particule. Montrer que la suite X_t est convergente (p.s.). Donner un histogramme de la loi de la variable limite $X_{+\text{ß}}$.

Que se passe-t-il lorsqu'on remplace $X_0 = 1$ par $X_0 = x_0$ (x_0 est un réel positif quelconque) ?

Un calcul mathématique (difficile, à ne pas faire) montre que $\frac{X_\infty}{X_0}$ suit la loi Beta$(\gamma + 1, \delta)$ et qu'elle est indépendante de X_0 (on pourra montrer l'indépendance). Illustrer ces résultats par la simulation.

Calculer la loi de X_∞ lorsque X_0 suit la loi Gamma $\Gamma(\gamma + \delta + 1, 1)$.

Montrer que si X_0 suit une loi de la forme loi $a\delta_1 + bt^h$ pour des constantes a, b, h bien choisies alors X_∞ suit la loi Beta(γ, δ)

Illustrer ces résultats par la simulation. Montrer que cela fournit une méthode de simulation de la loi Beta pour des paramètres quelconques. Comparer cette méthode à celle du livre de Bouleau (1986), page 380.

Références

Bouleau, N. (1986). *Probabilités de l'ingénieur : Variables aléatoires et simulation*. Actualités scientifiques et industrielles. Hermann, Éd. des Sciences et des Arts.

Guyader, A. (2007). Processus markoviens de sauts. Polycopié, disponible sur le site `http://www.lsta.upmc.fr/guyader/files/teaching/Sauts.pdf`.

Guyader, A. (2017). Statistique – Partie 1. Polycopié, disponible sur le site `http://www.lsta.upmc.fr/guyader/statM1.html`.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc.

Lambert, A. (2008). Population dynamics and random genealogies. *Stoch. Models 24*, 45–163. disponible à `https://www.lpsm.paris/pageperso/amaury.lambert/index_fichiers/Guanajuato.pdf`.

Rebafka, T. (2017). Statistique – Partie 2. Polycopié, disponible sur le site `https://www.lpsm.paris//pageperso/rebafka`.

Gilles Pagès *Vincent pourrais-tu développer un peu et ajouter des réf ? ?*

1. Simuler le problème du collectionneur de coupons avec ou sans échange.
2. La ruine du joueur à pile ou face avec ou sans calcul approché de la proba que $S_{2n} = 0 \ldots$
3. Simule un problème de pricing de billets d'avion avec surbooking (cf. dans mon vieux livre "En passant par hasard") en y ajoutant une approximation de type TCL.
4. Simuler un arbre de Galton-Watson avec extinction des noms de famille, généralement avant celle de la population elle-même.
5. La recherche de quantile par algorithme stochastique et par inversion de la répartition empirique.
6. Box-Muller versus Marsaglia (polar method) pour la gaussienne ? Faire un course. Après il faut l'illustrer...
7. Simuler des processus gaussiens à temps discret par Choleski de la matrice de covariance. Application aux marches aléatoires browniennes et browniennes fractionnaires (pas besoin de savoir ce que sont les processus en question mais ils peuvent en profiter pour apprendre un peu). But : dessiner des marches en fonction de la constante de Hurst H.

Ismael Castillo

1. méthode MCMC par exemple random walk metropolis Hastings. là je vous laisse jouer toi et Vincent, mais cela doit bien marcher j'imagine dans pas mal d'exemples simples (cf. au besoin livre de Christian Robert)

Damien Simon *Vincent pourrais-tu développer un peu et ajouter des réf ? ?*

1. modèle de Wright-Fisher de populations
2. ruine du joueur
3. un Monte-Carlo sur n'importe quel modèle de méca stat (Potts en champ moyen ? En dimension 2 ?)
4. Erdös Rényi et visualisation du graphe
5. amas de percolation et transition de phase
6. un exemple de file d'attente
7. statistiques de longueurs de cycles dans une permutation aléatoire uniforme (c'est simple à simuler malgré l'intitulé)

Tabea

1. comparer différents tests nonparamétriques d'adéquation à une loi
2. Test de Kruskal-Wallis (comparé à ANOVA)
3. convergence du spectre de la matrice de covariance...