Programme détaillé des journées de l'équipe Systèmes dynamiques, Probabilités et statistique du LMBA (17 et 18 juin 2025 - UBS Vannes)

Mardi 17 juin

11h-11h30
Arrivée des participants, café & thé de bienvenus

11h30 - 12h30
Record linkage and analysis of linked data with application in French national health data system
Valérie Garès (INRIA Rennes)

The French National Health Data System is the national health data system which collects all the longitudinal health records and insurance information of most of the French population. These data can be used to enrich other existing databases (cohorts, health registries...), which allows to get a more comprehensive medical information on each patient, and thus, to improve the subsequent statistical analysis. However, patients in the SNDS and health databases are usually anonymized, and no unique patient identifier is available to match the databases. Fellegi and Sunter (1969) proposed a probabilistic record linkage method, based on the fact that we usually access some "matching variables" which serve as partial identifiers common to both databases (e.g., gender, postal codes, dates of the treatment…). They allow to calculate "matching probabilities" for each pair of patients taken in the SNDS and the health registry of interest. The Fellegi and Sunter model is limited to simple binary comparison between matching variables. In our first work, we proposed an extension of this model for mixed-type comparison vectors. We developed a mixture model for handling comparison values of low prevalence categorical matching variables, and a mixture of hurdle gamma distribution for handling comparison values of continuous matching variables. In a second work, we proposed models for survival analysis with matched data. Indeed, perfect matching is never achieved, and neglecting associated errors can lead to biased estimates. In this work, we proposed an adjusted estimating equation for secondary Cox regression analysis, where linked data have been prepared by someone else and no information on matching variables are available to the analyst. Finally, we may access the matching probabilities which convey some uncertainty on the matching process, and this uncertainty must be taken into account in any subsequent statistical analysis. We proposed a new method in order to take account of these errors in a survival analysis based on the Cox model. This method is based on the well-known EM algorithm for estimation in a missing-data context. The proposed models are applied to perform a survival analysis of linked data between a registry of patients suffering from venous thromboembolism in the Brest and

the SNDS. *Joint work with Vanessa Chezeu, Huan Vo Tanh, Guillaume Chauvet, J-François Dupuy.*

12h30 - 14h : déjeuner

14h - 14h45
Christophe Cuny
Titre et résumé à venir

14h45 – 15h15
Evaluation Experimental Design to Sample Mosquitoes in Cambodia
Heloise Rozier

We evaluated four aspects of experimental design: (i) efficiency of trapping devices - BG-Sentinel (BG) vs. CDC light traps (LT)), (ii) temporal efficiency over a three-day sampling period, (iii) the impact of sampling duration on vector presence and abundance assessments, and (iv) site visit frequency for biodiversity surveys. Both traps were deployed across 10 Cambodian provinces, with collections conducted every 24 hours over three-day periods from 2019 to 2021. A total of 1,992 collections yielded 181,798 mosquitoes spanning 153 identified species. Using a hurdle generalized linear mixed model to address overdispersion and zero inflation, we found that trap type significantly influenced 21 of 54 studied species: LT captured 15 species more efficiently, while 6 species preferred BG. Anthropophilic vectors such as Aedes aegypti and Aedes albopictus were more attracted to BG, whereas LT captured higher species richness and zoophilic vectors. Capture efficiency declined over time for 6 of 13 primary vector species. However, annual trends in presence and abundance remained consistent regardless of whether sampling lasted one, two, or three days. Finally, we quantified the probability of collecting a new species during subsequent missions, showing that this probability declines with visit sampling effort at rates dependent on trap type and sampling days duration. These findings provide empirical guidelines for optimizing mosquito surveillance strategies based on study objectives.

15h15 - 15h45 : café & thé

15h45 – 16h30
Un théorème local limite conditionnel via l'approximation du noyau de la chaleur
Ion Grama

Nous étudions le comportement asymptotique d'une marche aléatoire réelle à incréments indépendants et identiquement distribués, de moyenne nulle et de variance finie. Plus précisément, nous analysons le premier temps de sortie de la marche hors d'une demi-droite positive. Dans un travail antérieur, nous avions obtenu des approximations gaussiennes pour la probabilité de persistance et la distribution conjointe de la marche

avant sa sortie. Le présent article étend ces résultats en établissant un nouveau théorème local limite conditionnel pour des marches à valeurs entières.
Notre approche fournit des approximations uniformes valables pour tout point de départ et toute position finale. Ces résultats généralisent substantiellement les théorèmes classiques existants et permettent d'obtenir de nouvelles formules asymptotiques pour les probabilités locales de sortie. Le cas des marches non-lattices est également traité.

16h30 – 17h

Limit theorems for estimation of the offspring mean of a branching process in a random environment

Dianni Wang

Abstract. Let (Z_n) be a single type branching process in an independent and identically distributed random environment. We study asymptotic properties of the Lotka-Nagaev type estimator for the mean of the offspring distribution. First, we refine existing results on the estimator's consistency and asymptotic normality by relaxing moment conditions. Then, we establish large and moderate deviation principles, as well as a Berry-Esseen type bound. (Joint work with Quansheng Liu and Yanqing Wang)

17h-17h45

Discussion de l'équipe (rapport HCERES, réorganisation)

20h : diner au Piano Barge (All. Loïc Caradec, 56000 Vannes)

Mercredi 18 juin

8h30h-9h30

Modèle de croissance polynucléaire : anciens et nouveaux résultats

Alessandra Occelli (LAREMA, Angers)

Dans cet exposé, je discuterai du modèle de croissance polynucléaire avec différentes symétries et conditions aux bords. Je soulignerai les liens avec des problèmes de combinatoire (problème d'Ulam), d'analyse (équation de Painlevé II), de physique mathématique (modèles de croissance KPZ) et de théorie des matrices aléatoires. J'esquisserai la stratégie d'étude du modèle dans le cadre du demi-espace avec deux sources externes, stratégie qui s'appuie sur des identités polynomiales algébriques et orthogonales, et sur un problème de Riemann--Hilbert, et qui a conduit à une distribution limite formulée en termes de la solution de l'équation de Painlevé II. Ce résultat prouve une conjecture de Barraquand--Krajenbrink--Le Doussal '22 sur la distribution de l'équation KPZ stationnaire sur la demi-droite. Basé sur un travail en collaboration avec M. Cafasso, D. Ofner, H.Walsh.

9h30h - 10h15

Moments and large deviations for multitype branching processes in random environments

Quansheng Liu

Consider a d-type supercritical branching process Z_n =(Z_n(1),...,Z_n(d)) in an independent and identically distributed random environment, whose offspring distributions of generation n depend on the environment at time n. We present precise asymptotics of the moments and a Bahadur-Rao type large deviation asymptotic expansion for the population size | Z_n | = \sum_{j=1}^d Z_n(j) of generation $n. In the approach, we use Cramer type measure change, under which we establish a Perron-Frobenius type theorem and stable convergence for products of random positive matrices, as well as L^p convergence for the multitype branching process. (Joint work with Ion Grama and Thi Trang NGUYEN)

10h15-10h45 : café & thé

10h45 - 11h30

Convex hull peeling

Gauthier Quilan

Le convex hull peeling d'un nuage de points est obtenu en construisant l'enveloppe convexe de ces points, puis en retirant les points extrémaux du nuage et en construisant la nouvelle enveloppe convexe des points restants et ainsi de suite. On appelle couche d'ordre n la frontière de l'enveloppe convexe obtenue à l'étape n de la procédure. Dans cet exposé, on s'intéresse à l'étude de fonctions (nombre de points extrémaux, de faces k-dimensionnelles et volume défaut) des couches successives du convex hull peeling d'un ensemble de points indépendants et uniformes dans la boule unité et dans un polytope simple. On rappellera dans un premier temps des théorèmes limites classiques sur le nombre de k-faces de la première couche, c'est-à-dire de l'enveloppe convexe. Ensuite nous présenterons nos résultats qui étendent ces théorèmes limites à toutes les premières couches.

11h30-12h15

Le champ libre gaussien : deux ou trois choses que je sais de lui

Jean-Marc Derrien

S'il existait, le pont brownien sur le disque unité **D** serait une incarnation du champ libre gaussien sur **D** lui-même représentation isométrique de l'espace de Sobolev $H_0^1(\mathbf{D})$. Pour définir ce dernier, on peut choisir n'importe quelle métrique conformément équivalente à la métrique euclidienne. Cet exposé est une introduction au champ libre gaussien sur **D** qui met l'accent sur quelques commodités qu'offre la géométrie hyperbolique sur le disque unité pour en décrire les premières propriétés. Notre approche fournit des approximations uniformes valables pour tout point de départ et toute position finale. Ces résultats généralisent substantiellement les théorèmes classiques existants et permettent

d'obtenir de nouvelles formules asymptotiques pour les probabilités locales de sortie. Le cas des marches non-lattices est également traité.

12h15-13h45 : déjeuner

13h45-14h45

Implementation of a multistate modeling in oncology studies

Pierre Colin (Bristol Myers Squibb)

In oncology early phase, the usual decision rules to promote a compound to the next phase of clinical development is based on tumor response rates. These decision rules are implicitly assuming some form of surrogacy between tumor response and long-term endpoints as progression-free survival (PFS) or overall survival (OS). However, when exploring new therapies, the response rate may not be a surrogate endpoint. And the usual decision rules may be misleading.
The multistate model allows to account for more information than the simple RECIST response status, namely, the time to get to response, the duration of response, time to disease progression, and time to death. Some of these endpoints may be usually considered as secondary (or even exploratory) endpoints. But they are critical to understand which step of the patient's outcome is improved by a new therapy. The multistate model is composed of time-to-event models to represent each transition from one state to another (e.g. from positive tumor response to disease progression).
The multistate model Is also an in"eres'Ing alternative method to analyze survival data without the proportional hazard assumption. Since the multistate allows one to estimate a different treatment effect for each transition, the derived hazard ratio for a long-term endpoint (as Overall Survival) may not be constant over time.
Such outcomes can be used to support critical decision-making before starting a large pivotal study, by projecting what long-term drug effect one should expect based on the drug effect observed from each transition.

14h45 – 15h15

Forêts d'isolation basées sur la densité

Nathan Lévêque

Nous proposons une nouvelle méthode de forêt d'isolation que nous appelons la forêt d'isolation basée sur la densité (dbiForest). Comme les forêts d'isolation (iRF) introduites par Liu et al. (2008), une forêt dbiForest se compose de plusieurs arbres d'isolation basés sur la densité (notés dbiTree). Chaque dbiTree cherche à identifier les anomalies au moyen de divisions binaires et récursives de l'espace des données. L'originalité de cette nouvelle approche repose sur l'utilisation d'estimations de densités pour estimer les coupures. L'algorithme s'appuie sur l'hypothèse que, les anomalies correspondant à des observations à la fois différentes et rares par rapport aux données normales, ces observations seraient présentes dans des espaces de plus faibles densités. Chaque dbiTree cherche donc à isoler les observations présentes dans les régions de plus faible densité. Nous montrons les bonnes performances de notre méthode au travers de

plusieurs expériences numériques dans lesquelles dbiForest est comparé à d'autres méthodes, notamment l'algorithme iRF.

15h15 : Fin de l'évènement