



# LES RENCONTRES DE STATISTIQUE APPLIQUÉE

## Modèles de durée : Expériences croisées

**Judi 16 octobre 2014, 9h30 - 17h00**

**Institut Henri Poincaré (amphithéâtre Hermite)**

**11 rue Pierre et Marie Curie • 75231 Paris cedex 05**



Service des Méthodes Statistiques (SMS) de l'INED, avec la collaboration du Pôle GRAB (Approche biographique et multiniveaux) de l'INED et la Société Française de Statistique

Les modèles de durée constituent des outils d'analyse précieux et performants dans de nombreuses applications mettant en jeu des données longitudinales. Utilisés dans diverses disciplines, ces modèles se heurtent souvent à la complexité des données recueillies (effet de sélection, censure à gauche) ou de la nature des événements à modéliser (événements répétés, issues multiples). Il faut alors prendre en compte ces difficultés d'analyse au sein de modèles plus complexes ou coupler ces méthodologies avec d'autres techniques.

Au travers d'expériences issues des sciences sociales et de l'épidémiologie, exposées au cours de cette séance des pratiques d'analyse qui rendent compte de cette complexité et des précautions particulières nécessaires dans la mise en œuvre des procédures d'analyse seront abordées.

### PROGRAMME

9h30 : **Arnaud Bringé (INED-SMS), Valérie Golaz (INED-GRAB) Introduction de la séance**

9h45 : **Jean-Marie Le Goff (Centre Lines, Université de Lausanne & Pôle National de recherche Lives)**

#### **Analyse des biographies en temps discret: Echéances multiples et événements répétés**

Les modèles d'analyse des biographies en temps discret consistent à modéliser la distribution de la probabilité conditionnelle, de connaître un événement durant un intervalle de temps ainsi que le rôle joué par des caractéristiques fixes ou dépendantes du temps sur ces probabilités. Ces modèles devraient être préférés aux modèles en temps continu, en premier lieu lorsque les échéances ont lieu à des moments précis, par exemple seulement au mois de septembre chaque année, et, en second lieu, lorsque les durées mesurées sont peu précises, par exemple lorsque l'unité de temps est l'année. L'approche en temps discrets est simple à mettre en œuvre, puisque après constitution d'un fichier de données personne-période, il s'agit d'estimer les paramètres d'une régression binomiale, en adoptant un lien logit, probit ou complémentaire log-log.

Après une courte présentation de ces modèles, notre communication proposera une extension de l'approche en temps discret au cas des échéances multiples. Nous souhaitons notamment indiquer les différences entre estimer une régression binomiale pour chacune des échéances ou estimer un modèle de régression multinomiale. Par ailleurs, nous évoquerons aussi le cas de l'approche en temps discret des événements. Le propos s'appuiera sur un exemple, le devenir des unions non-maritales (mariage ou séparation) vécues par les personnes enquêtées dans le cadre de la *British Cohort Study 1970*.

10h15 : **Anne-Sophie Bruno (Centre d'études de l'emploi, Université Paris 13)**

#### **Analyser des trajectoires professionnelles de migrants. Quelques outils au service d'une démarche historique**

En se fondant sur l'analyse de données de trajectoires professionnelles reconstituées au moyen des dossiers de retraite et appariées aux informations sur les employeurs fournies par les DADS, cette communication étudie la question des variations de la mobilité inter-entreprises observées dans un échantillon de migrants de Tunisie sur le marché du travail francilien des années 1950 aux années 1990 ; nous proposons de recourir à une série de modèles de Cox multiniveaux pour tenter de démêler les facteurs tenant aux caractéristiques sociodémographiques de la main-d'œuvre, notamment le genre et la nationalité, de ceux ayant trait au mode de gestion de la relation salariale mis en œuvre par les entreprises.

11h15 : **Milan Bouchet-Valat (INED, OSC-Sciences Po & CNRS, LSQ-CREST)**

#### **Modéliser un événement n'affectant qu'une partie des individus à l'aide d'un modèle log-logistique à temps de sortie accéléré : la première mise en couple**

Il arrive fréquemment qu'un événement d'intérêt n'affecte qu'une partie de la population étudiée. Si cette proportion d'individus réfractaires ou immunisés est importante, l'utilisation de modèles à temps de sortie accéléré classiques pose problème. Brüderl et Diekmann (1995) ont proposé une extension du modèle log-logistique à temps de sortie accéléré tenant compte de cette possibilité. La proportion d'individus immunisés (stayers) est estimée à partir des données ; elle peut dépendre de variables explicatives qualitatives ou quantitatives. Nous illustrons ce modèle à partir de l'étude de la première mise en couple en France : il nous permet d'estimer l'effet du diplôme aussi bien sur l'âge de première mise en couple que sur la proportion d'individus qui ne vivront jamais en couple, et son évolution au fil des cohortes, et ce malgré le fait que les cohortes les plus récentes sont loin d'avoir achevé leur vie conjugale.

11h45 : **Xavier Bry (I3M, UMR CNRS 5149, Université Montpellier II), Axel Jean-Théodore (INED-ENS Cachan)**

#### **Méthode globale d'estimation de transitions graduelles**

Jusqu'à présent, l'analyse de durée s'est intéressée à la survenue d'un type d'évènement dans une population

donnée; les transitions entre deux états (vivant/mort en analyse de survie) sont donc instantanées. Il s'avère que dans de nombreux cas d'étude, les évolutions sont plus complexes. Nous proposons donc d'étendre l'analyse de durée aux transitions graduelles entre deux états latents, appréhendées via l'accumulation au cours du temps d'un ensemble d'évènements marqueurs observés. Cette analyse trouve de nombreuses applications, notamment en analyse biographique où les trajectoires individuelles présentent des changements d'états étalés dans le temps (le passage à l'âge adulte par exemple).

D'un point de vue statistique, il s'agit de modéliser les trajectoires étudiées par un processus de comptage, processus dont on cherche à estimer la valeur moyenne à tout instant. La démarche globale d'estimation a été proposée par X.Bry (2006) : après avoir évalué l'allure de la transition étudiée avec une estimation non-paramétrique, on propose un paramétrage des fonctions de risque dont on estime les coefficients par une régression de Poisson. Les méthodes présentées permettent de s'affranchir du problème de censure à droite, et de modéliser l'hétérogénéité de la population étudiée par l'introduction d'effets aléatoires. En sortie, on obtient une série d'estimateurs permettant une description de la transition étudiée.

Dans cet exposé, nous présenterons cette modélisation ainsi que sa mise en œuvre via la librairie R contenant l'ensemble des estimateurs mentionnés. À terme, une interface graphique s'appuyant sur l'outil R Commander sera ajoutée pour faciliter l'utilisation de la librairie.

14h15 : **Aurélien Latouche (Cnam)**

#### **Analyse et interprétation des risques concurrents : consensus et problèmes ouverts**

L'analyse de survie concerne la modélisation des délais de réalisation d'un unique événement. Dans de nombreuses situations cependant, les individus sont soumis à plusieurs causes d'événement exclusives, définissant un cadre dit de risques concurrents ou compétitifs.

Nous présentons une introduction à l'analyse des risques dit concurrents. Les quantités fondamentales que sont les fonctions de risques cause-spécifiques et les incidences cumulées sont introduites ainsi que les modèles de régression qui leurs sont associés. Nous insisterons sur la présentation des résultats d'analyses et l'interprétation ainsi que quelques questions méthodologies ouvertes.

14h45 : **Karen Leffondré (ISPED, Université de Bordeaux, INSERM U897)**

#### **Censure par intervalle et compétition avec le décès : avantages du modèle illness-death**

Dans les études épidémiologiques longitudinales, il est fréquent que la maladie étudiée (par exemple la démence) ne soit diagnostiquée qu'à l'occasion des visites prévues dans le protocole. Par exemple, dans l'étude française PAQUID, 3777 personnes de la population générale âgée de plus de 65 ans ont été incluses dans la cohorte en 1988-1990, et bénéficient depuis d'un dépistage de la démence lors des visites biannuelles. L'âge de début de la démence est donc censuré par intervalle entre l'âge à la visite de diagnostic et l'âge à la visite précédente. Cette censure par intervalle complique la prise en compte du risque compétitif avec le décès qui est naturellement très fréquent dans cette population. Dans cette présentation, je présenterai une approche statistique qui permet de tenir compte de la particularité de ces données, le modèle illness-death pour données censurées par intervalle. Je comparerai les résultats de ce modèle à ceux de modèles de survie classiquement utilisés dans la pratique, pour estimer l'effet de facteurs sur le risque de développer la maladie, à la fois sur des données simulées et sur les données de la cohorte PAQUID.

15h45 : **Philippe Saint Pierre (LSTA, Université Pierre et Marie Curie)**

#### **Modèles multi-états Markoviens et semi-Markoviens**

Dans de nombreux domaines, on est confronté de plus en plus souvent à des données de type cohorte. Ces données longitudinales correspondent à des mesures répétées dans le temps d'une même variable. En épidémiologie, par exemple, ces données peuvent provenir de l'observation répétée dans le temps d'individus atteints d'une maladie chronique ou d'individus exposés à un risque. Ces données correspondant à plusieurs durées (dépendantes) pour un même individu sont sujettes aux phénomènes de censure et de troncature et doivent donc être analysées avec des méthodes d'analyse de survie multivariée. Les modèles multi-états Markoviens qui utilisent les notions d'états et de processus sont bien adaptés pour modéliser ce type de données. Ils permettent de représenter l'évolution des individus à travers différents états et d'étudier les forces de transition entre ces différents états. Dans ce contexte, on introduira les modèles Markoviens homogènes (intensités de transition indépendantes du temps) et non-homogènes (intensités de transition dépendantes de la durée du suivi ou du temps calendaire). On discutera ensuite des modèles semi-Markoviens qui permettent de faire dépendre les intensités de transition du temps passé dans un état. Une application à l'asthme illustrera l'intérêt de ces modèles en situation pratique.

16h50 : **Conclusion de la journée**

Le séminaire est ouvert à tous, sans frais de participation.

Pour une bonne organisation, nous demandons aux personnes désirant y assister de s'inscrire à l'adresse suivante : [http://www.ined.fr/fr/rendez\\_vous/rencontres\\_statistique\\_appliquee/](http://www.ined.fr/fr/rendez_vous/rencontres_statistique_appliquee/)

Pour tout autre renseignement contacter Bénédicte Garnier ([benedicte.garnier@ined.fr](mailto:benedicte.garnier@ined.fr))