Contents • Table des Matières

1 Welcome • Bienvenue

The Department of Mathematics and Statistics and Dalhousie University welcomes you to Halifax for the 31st Annual Meeting of the Statistical Society of Canada. We trust that you will find the meetings both intellectually stimulating and socially enjoyable. The Dalhousie participants can be identified by their blue nametags. We are happy to try to answer any questions about the local arrangements and Halifax in general.

Please take some time to wander around the city enjoying its history and sites. The restored Historic Properties on the waterfront are worth a visit, as are the Citadel and the Maritime Museum, for historical perspectives. Within one hour of Halifax, you can visit scenic Peggy's Cove or find other delightful fishing villages. The town of Lunenburg (a UNESCO World Heritage Site) is renowed for its interesting architecture and historic harbour. The highest tides in the world occur in the Bay of Fundy, again about 90 minutes from Halifax. Farther afield is Cape Breton with the scenic Cabot Trial and the restored French fortress at Louisburg.

Le Département de mathématique et de statistique et l'Université Dalhousie vous souhaitent la bienvenue à Halifax pour le 31e Congrès annuel de la Société Statistique du Canada. Nous espérons que ces réunions vous seront à la fois profitables et agréables, sur un plan intellectuel et social. Vous pourrez facilement identifier les participants qui travaillent à Dalhousie grâce à leur porte-noms bleus. Ils se feront un plaisir de répondre à vos questions sur les arrangements locaux et sur Halifax en général.

Prenez le temps de flâner dans la ville et de découvrir ses sites historiques. Les Propriétés Historiques restaurées du bord de l'eau valent le détour, tout comme la Citadelle et le Musée Maritime pour une leçon d'histoire. Á une heure de route de Halifax, vous pouvez visiter le pittoresque village de Peggy's Cove et d'autres villages de pêche tout aussi charmants. La ville de Lunenburg (classée site du patrimoine mondial par l'UNESCO) est célèbre pour son architecture intéressante et son port historique. La baie de Fundy, où l'on peut voir les marées les plus hautes du monde, est à quelque 90 minutes de Halifax. Si vous avez le temps d'aller plus loin, nous vous conseillons le Cap Breton et la Piste Cabot, ainsi que la forteresse française restaurée à Louisbourg.

2 Sponsors • Commanditaires

SSC 2003 thanks the sponsors of the meeting for their kind contributions. In particular, we thank the Centre de recherches mathématiques, the Fields Institute and the Pacific Institute for the Mathematical Sciences for their financial support. We also thank Dalhousie University, and the Department of Mathematics and Statistics for the use of resources and financial support.

SSC 2003 remercie les commanditaires du congrès pour leurs aimables contributions. En particulier, nous remercions le Centre de Recherches Mathématiques, le Fields Institute et le Pacific Institute for the Mathematical Sciences pour leur aide financière. Nous remercions également Dalhousie University, et le Département de mathématiques et de statistique pour l'utilisation de leurs ressources et leur appui financier.

3 Organizers • Organisateurs

Local Arrangements Committee • Comité organisateur local

Chair/*Président* :	Chris Field		
Members/*Membres* :	Wade Blanchard	Kit Bowen	Hong Gu
	RP Gupta	David Hamilton	Christophe Herbinger
	Harold Rennie	Bruce Smith	Ed Susko

The Committee has had substantial help from the graduate students in Statistics and the office staff of the Department of Mathematics and Statistics.
Le Comité remercie les étudiants de deuxième cycle en statistique et le personnel clérical du Département de mathématique et de statistique pour toute leur aide.

Programme Committee • Comité du programme

Chair/*Président* : Doug Wiens (University of Alberta)
Members/*Membres* :
Biostatistics Section •*Groupe de biostatistique*
K.C. Carriere (U. of Alberta), Salomon Minkin (U. Toronto)
Business and Industrial Statistics Section • *Groupe de statistique industrielle et de gestion*
John Brewster (U. Manitoba), Roman Viveros-Aguilera (McMaster U.)
Survey Methods Section • *Groupe de méthodologie d'enquête*
Don Royce (Statistics Canada), Patricia Whitbridge (Statistics Canada)
IMS Sessions
Mary Meyer (University of Georgia)

Translation • Traduction

Chair/*Président* :	Denis Larocque (Université de Montréal)
Members/*Membres* :	François Bellevance (HEC)
	Sorana Froda (Université du Québec à Montréal)
	Peter Macdonald (McMaster University)
	Louis-François Poirier (Université de Montréal)

Web Page Maintenance • Gestion de page web

Peter Macdonald (McMaster University)

4 Exhibitors • Exposants

John Wiley & Sons Canada Limited
Nelson Thomson Learning
Pearson Education Canada

The book displays are located in the Great Hall, University Club and will be available for viewing and purchasing from 8:30 to 17:00 from Monday through Wednesday.

Le salon des exposants est au Great Hall, University Club. Les livres pourront être consultés ou achetés de 8h30 à 17h00 de lundi à mercredi.

5 General Information • Information générale

Registration • Inscription

The Registration Desk will be open as follows:

Date	**Time**	**Location**
Saturday June 7	19:00 – 22:00	Lord Nelson lobby
Sunday June 8	8:00 – 12:00	Lord Nelson lobby
Sunday June 8	13:00 – 21:00	Great Hall, University Club
June 9-11	8:30 – 16:30	Room 219, Chase Building

L'inscription aura lieu aux heures suivantes :

Jour	**Heure**	
Samedi 7 juin	*19h00 – 22h00*	*foyer, Lord Nelson*
Dimanche 8 juin	*8h00 – 12h00*	*foyer, Lord Nelson*
Dimanche 8 juin	*13h00 – 21h00*	*Great Hall, University Club*
9-11 juin	*08h30 – 16h30*	*salle 219, Chase Building*

Directions • Directions

To Dalhousie from Lord Nelson: turn right on Spring Garden Road and continue 1.1 km to Dalhousie campus. Spring Garden changes to Coburg Road as you pass Robie Street.
To Lord Nelson from Dalhousie: reverse the above.

De l'hôtel Lord Nelson vers Dalhousie: prendre à droite sur la rue Spring Garden et continuer vers le campus de Dalhousie (1.1 km). La rue Spring Garden devient la rue Coburg après le croisement avec la rue Robie.
De Dalhousie vers l'hôtel Lord Nelson: suivre les directions d'au dessus en sens opposé.

Opening Mixer/Poster Session • Soirée d'ouverture / Session d'affichage

The Opening Mixer will be held from 18:30 to 22:00 on Sunday June 8 in the Great Hall of the University Club. Hors d'oeuvres and a cash bar will be available. One free drink ticket is included in your registration package. All conference attendees and companions are warmly invited to come and meet old friends and make new acquaintances. The opening poster session will be in this location starting at 14:00. The Registration Desk will also operate in this location on Sunday afternoon.

La réception d'accueil se tiendra de 18:30 heures à 22 heures, dimanche 8 juin dans le grand hall du Club de l'université. Nous aurons des hors d'œuvres et un bar payant à vous offrir. Votre trousse d'inscription contient un ticket pour une boisson gratuite. Les participants au congrès et leurs compagnons sont tous invités à venir retrouver de vieux amis et faire de nouvelles connaissances. La séance par affichage se tiendra dans ce même lieu à partir de 14 heures. Le bureau des inscriptions y sera également ouvert tout l'après-midi du dimanche.

Workshops • Ateliers

Workshops organized by the three sections of the SSC will be held on Sunday June 8 from 9:00 to 17:00 in the Lord Nelson Hotel. The rooms are as follows.

Les ateliers organisés par les trois sections de la SSC auront lieu dimanche le 8 juin de 9h00 à 17h00 dans l'hôtel Lord Nelson . Les salles pour ces ateliers sont les suivantes.

Workshop • Atelier	**Place • Endroit**
Biostatistics • *Biostatistique*	Regency Ballroom
Business and Industrial Statistics • *Statistique industrielle et gestion*	Britannia Room
Survey Methods • *Méthodologie d'enquête*	Imperial Ballroom

Check the scientific programme and abstract sections for further details.
Vérifiez le programme scientifique et la section des résumés pour plus de détails.

Rooms • Salles

The rooms used for the conference are LSC 234, 238, 240, 242, 332 and 338 in the Life Science Center (C204 on the campus map, see back cover) and the Ondaatje Theatre in the McCain Arts and Social Science Building (D420 on campus map).

Les salles mises à la disposition du congrès sont les suivantes : LSC 234, 238, 240, 242, 332 et 338 dans le Centre des sciences de la vie (C204 sur la carte du campus, voir en quatrième de couverture) et l'amphithéâtre Ondaatje dans l'édifice McCain des arts et sciences sociales (D420 sur la carte du campus).

Food Services on Campus • Restauration sur le campus

Lunches will be available at several locations on campus. The easiest and the recommended option is to have lunch at the cafeteria in Howe Hall (C520 on campus map). There are a number of serving stations within the cafeteria offering a variety of food. The cost is $7.38 payable in cash at the door. All committee meetings on Monday through Wednesday will be held in the Great Hall of the University Club where breakfast or lunch will be provided. There are fast food outlets in the Student Union Building and the Life Sciences Center.

Vous pourrez déjeuner à midi à divers endroits. La solution la plus simple, que nous vous recommandons, est de déjeuner à la cafétéria de Howe Hall (C520 sur la carte du campus). Celle-ci offre une variété de plats différents. Il vous en coûtera 7,38 $, payable en liquide à la caisse. Toutes les réunions de comités se tiendront du lundi au mercredi dans le grand hall du Club de l'université; le petit-déjeuner ou déjeuner, selon le cas, y sera offert aux participants. Vous trouverez également des comptoirs de prêt-à-manger dans le bâtiment du syndicat d'étudiants (SUB) et dans le Centre des sciences de la vie (LSC).

Barbecue • Barbecue

There will be a barbecue on Monday evening from 18:30 to 22:00 in the dining room of Shirreff Hall (C220 on campus map). All graduate students will have received a barbecue ticket with their registration. All other participants who checked this off on the registration form will also have a ticket. Both meat and vegetarian options will be available and there will also be a cash bar. We would like to invite those people attending the Women in Statistics dinner (see following) to come to join us after their dinner to mix with the graduate students.

Lundi soir, nous organisons un barbecue de 18:30 heures à 22 heures dans la salle à manger de Shirreff Hall (C220 sur le plan du campus). Les étudiants de deuxième cycle trouveront un ticket d'invitation au barbecue dans leur trousse d'inscription. Les autres participants qui ont choisi cette option sur leur formulaire d'inscription recevront également un ticket. Des options végétariennes et non végétariennes sont proposées, ainsi qu'un bar payant. Nous invitons les participants au dîner des Femmes en statistique (voir le paragraphe suivant) à venir se joindre à nous après leur repas pour rencontrer les étudiants de deuxième cycle.

Restaurants off Campus • Restauration hors campus

T: Member of the "Taste of Nova Scotia" Program • *Membre du programme "Taste of Nova Scotia"*
L: Licensed • *Avec license pour servir des boissons alcoolisées*
R: Reservations recommended • *Réservations recommandées*

Within walking distance from Dalhousie University • *Proche de l'Université Dalhousie:*

- 10 min on Spring Garden Road • *À10 mn à pied sur Spring Garden Road*

  • Dandelion Cafe (422 4116) 5986 Spring Garden Rd – Healthy snacks and lunches • *En-cas et dîners légers*

  • Maki Maki Japanese Food (422 3818) 5974 Spring Garden Rd • *Cuisine japonaise*

- 15 min on Quinpool Road • *À15 mn à pied sur Quinpool Road*

  • Athens Restaurant L (422 1595) 6303 Quinpool Rd – Greek and Italian • *Cuisine grecque et italienne*

  • Chicken Tandoor L R (423 7725) 6285 Quinpool Rd – North Indian and Thai - Dinner only • *Cuisine indienne et thailandaise - Souper seulement*

  • China Classic L (529 2828) 6311 Quinpool Rd – Szechuan, Cantonese and Chinese • *Cuisine chinoise, cantonaise et szechuan*

  • Heartwood Bakery and Cafe (425 2808 closed on Sunday) 6250 Quinpool Rd Vegetarian • *Restaurant végétarien*

- Hogie's Steak and Seafood House L R (422 4414) 6273 Quinpool Rd – Steak and seafood • *Steaks et fruits de mer*
- King Wah L (423 2587) 6430 Quinpool Rd – Szechuan and Cantonese specialities • *Spécialités szechuan et cantonaises*
- Spartan Restaurant L (429 6858) 6403 Quinpool Rd – Friendly home Greek cooking • *Cuisine familiale grecque*

- Within 10 min walking distance from Lord Nelson hotel • *À10 mn à pied de l'hôtel Lord Nelson*

- Anatolia L R (492 4568 – closed on Sunday) 1518 Dresden Row – Authentic Turkish cuisine • *Cuisine authentique turque – fermé le dimanche*
- Le Bistro T L R (423 8428) 1333 South Park Street
- Birmingham Bar and Grill L R (420 9622) 5637 Spring Garden Rd (Park Lane Mall) – Casual International cuisine • *Cuisine internationale*
- Curry Village L R (429 5010) 5677 Brenton Place – Authentic Indian cuisine • *Cuisine authentique indienne*
- Danube Cafe L R (431 9477) 5680 Spring Garden Rd
- Doraku L R (425 8888 – closed Monday) 1579 Dresden Row – Japanese fine dining • *Dîners japonais – fermé le lundi*
- Fid L R (422 9162 – closed on Monday) 1569 Dresden Row - Classical French with Asian accents • *Cuisine classique française avec touches asiatiques*
- The Fireside L (423 5995) 1500 Brunswick Street – Casual International cuisine • *Cuisine internationale*
- Il Mercato L (422 2866 – closed on Sunday) 5475 Spring Garden Rd – Northern Italian cuisine European–style tratorria • *Cuisine italo-européenne–style tratorria*
- Mexicali Rosa's L (422-7672) 5680 Spring Garden Rd – Californian–style Mexican food • *Cuisine mexicaine-californienne*
- Ryan Duffy's L R (421-1116) 5640 Spring Garden Rd – Steak and Seafood • *Steak et fruits de mer*

For more restaurants in downtown and waterfront areas approximately 10-15 minutes from the Lord Nelson, check Halifax Visitor Guide 2003.

Pour plus de restaurants en ville et sur le front de mer à 10-15 minutes à pied de l'hôtel Lord Nelson, se référer au Halifax Visitor Guide 2003.

Women in Statistics Dinner • Diner pour les femmes en statistique

The Caucus for Women in Statistics in collaboration with the Committee on Women in Statistics of the SSC will have an informal get-together and dinner on Monday June 9, 2003, at 6:30 pm at the Turkish restaurant Anatolia's (1518 Dresden Row, off Spring Garden) in Halifax. The restaurant is about 20 minutes walk from Dalhousie University campus and less than 5 minutes from the Lord Nelson Hotel. Interested women and men are welcome. (Afterwards dinner participants can attend the SSC sponsored barbecue/student social without charge.) If you are planning to come, please leave a message for Susana Rubin-Bleuer at the Conference Message Center as soon as possible , so we can make (informal) reservations. We can meet directly at the restaurant or at 6:10 pm in front of the University Club and walk together.

Le Caucus for Women in Statistics, en collaboration avec le Comité sur les femmes en statistique de la SSC, se réunira à l'occasion d'un dîner informel lundi 9 juin 2003, à 18 heures 30 au restaurant turc Anatolia's (1518 Dresden Row, au coin de Spring Garden) à Halifax. Le restaurant est à une vingtaine de minutes à pied du campus de l'université Dalhousie et à moins de 5 minutes de l'hôtel Lord Nelson. Toutes les intéressées et tous les intéressés sont les bienvenus. (Les participants à ce dîner sont invités à se joindre gratuitement au barbecue SSC / soirée rencontre des étudiants après leur repas). Si vous prévoyez de venir, veuillez laisser un message à l'intention de Susana Rubin-Bleuer au Centre des messages du congrès dès que possible, afin que nous puissions faire une réservation (informelle). Vous pouvez nous rencontrer au restaurant ou vous joindre à nous à 18:10 heures devant le Club de l'université pour faire le chemin ensemble.

Waterloo Statistics Alumni Reception • Reception des anciens diplômés en statistique de Waterloo

University of Waterloo Math alumni attending the 2003 SSC Meeting are warmly invited to a Reception in the Victorian Lounge of Shirreff Hall (C220 on campus map), Dalhousie University in Halifax on Monday June 9th.

Light refreshments and a cash bar will be available 4:30 - 6:30 p.m. as you mingle with fellow alumni and UW faculty.

Les anciens étudiants en mathématique de l'université de Waterloo qui participent au Congrès 2003 de la SSC sont invités à une Réception dans le salon Victorien de Shirreff Hall (C220 sur le plan du campus), à l'université Dalhousie à Halifax, lundi 9 juin.

Un repas léger et un bar payant sera disponible de 16:30 heures à 18:30 heures. Venez retrouver d'autres anciens étudiants et des professeurs de l'université de Waterloo.

Banquet • Banquet

All participants are cordially invited to attend the Conference Banquet on Tuesday evening, June 10 at Pier 21. Banquet tickets have been included in your registration material and are colour coded to indicate your food preference. Pier 21 is a restored immigration shed on the Halifax waterfront (directions below). We plan to start the evening at 6:30 with hors d'oeuvres and a cash bar. The museum will be open so you'll have a chance to look at the exhibits. As a country of many immigrants, a number of us will have come through Pier 21 or had ancestors who did. The lobster banquet will start at 8:00 with chicken and vegetarian options.

To get to Pier 21, there are several options. If you'd like to walk from Howe Hall, the distance is 2.9 km. Turn right on Coburg Road and continue until it ends at Barrington (by this time it has become Spring Garden Road). For those starting at the Lord Nelson, the distance is 1.8 km. Turn left on Spring Garden and continue until it ends at Barrington. Turn right on Barrington until South St (4 way stop). Turn left on South St for one block and then continue (ahead and slightly left) on Terminal Road which curves around the Westin Hotel. Pier 21 will be on your left. By bus you will take a number 1 from outside Howe Hall or the Lord Nelson (fare is $1.65) and transfer to a number 7 at the corner of Spring Garden and Barrington (ask for a transfer when you pay your fare). The number 7 will take you to the corner of Barrington and South Street and you follow the walking directions for about a 5 minute walk. Of course, you can always take a taxi.

Tous les participants sont cordialement invités à participer au buffet du Congrès mardi soir (le 10 juin) à Pier 21. Vous trouverez vos tickets pour le banquet dans votre trousse d'inscription. La couleur de votre ticket indique votre préférence de menu. Pier 21 est un ancien hangar d'immigration sur le bord de l'eau à Halifax (voir ci-dessous pour vous y rendre). Nous prévoyons de commencer la soirée dès 18:30 heures avec des hors d'œuvres et un bar payant. Vous aurez la possibilité de visiter le musée, qui restera ouvert pour nous. Le Canada est un pays d'immigrants - peut-être avez-vous transité par Pier 21, ou vos ancêtres ont-ils débarqué ici. Le banquet de homards sera ouvert à 20 heures (poulet ou plats végétariens en option).

Pour vous rendre à Pier 21, plusieurs possibilités s'offrent à vous. Si vous souhaitez y aller à pied depuis Howe Hall, prévoyez une promenade de 2,9 km. Prenez à droite sur Coburg Road, continuez jusqu'à ce que cette route (qui change de nom pour devenir Spring Garden Road) s'arrête à Barrington. Pour ceux qui partent du Lord Nelson, la distance est de 1,8 km. Prenez à gauche sur Spring Garden et continuez jusqu'à ce que la route s'arrête à Barrington. Prenez ensuite à droite sur Barrington jusqu'à South St (arrêt multi-sens). Tournez à gauche sur South St, suivez cette rue jusqu'à l'intersection suivante, puis continuez (devant vous, légèrement à gauche) sur Terminal Road qui fait le tour de l'hôtel Westin. Pier 21 est à votre gauche. En bus, prenez la ligne 1 devant Howe Hall ou devant le Lord Nelson (le trajet coûte 1,65 $) et prenez la correspondance sur la ligne 7 au coin de Spring Garden et Barrington (demandez un ticket de transfert lorsque vous montez dans le premier bus). La ligne 7 vous déposera au coin de Barrington et South Street; suivez ensuite les instructions de marche (cinq dernières minutes). Vous pouvez également prendre un taxi.

Job Fair • Salon de l'emploi

Interviews for those participating in the Job Fair will be conducted in various rooms located in the Chase building (C280 on the campus map). Check with the office of the Department of Mathematics and Statistics (Chase 219) for schedules and locations for interviews.

Les entretiens pour le Salon de l'emploi se tiendront dans diverses salles de l'édifice Chase (C280 sur le plan du campus). Contactez le bureau du Département de mathématique et de statistique (Chase 219) pour connaître l'horaire et le lieu de vos entretiens.

Transportation and Parking • Transport public et stationnement

For those staying on campus, parking is available for $4 per day and permits can be purchased at the front desk of your residence. For those staying off campus, daily parking permits can be purchased for $7 per day from the Security Office in the basement of the McCain Building (D420 on campus map).

For those needing transportation to the Halifax airport, there is regular bus service from a number of hotels including the Lord Nelson (Airbus, 873-2091, $12 one-way, $20 return). It is also possible to book Share-a-Cab (429-5555) which costs approximately $25 for a single rider, with a reduced rate for multiple riders. They require advance booking of 24 hours.

Pour ceux d'entre vous qui logent sur le campus, le stationnement coûte 4 $ par jour. Les permis s'achètent à la réception de la résidence d'étudiants. Si vous logez hors campus, vous pouvez vous procurer des permis de stationnement pour 7 $ par jour au bureau de la sécurité dans le sous-sol de l'édifice McCain (D420 sur le plan du campus).

Pour vous rendre à l'aéroport de Halifax, vous pouvez prendre le bus qui dessert les hôtels, dont le Lord Nelson (Airbus, 873-2091, 12 $ aller simple, 20 $ aller retour). Vous pouvez également réserver chez Share-a-cab (429-5555) : 25 $ environ pour un passager, tarif réduit pour les groupes. Réservez au moins 24 heures à l'avance.

Internet Access • Accès à l'internet

Your conference bag contains a page with a userid, password, and instructions for machines located in Chase 007 (C280 on campus map) and in the Learning Commons on the main floor of the Killam library (C580 on campus map).

Votre trousse d'inscription contient une page avec numéro d'usager, mot de passe, et instructions d'accès pour les machines situées dans la salle 007 de l'édifice Chase (C280 sur le plan du campus) et dans le "Learning Commons" à l'étage principal de la bibliothèque Killam (C580 sur le plan du campus).

6 Committees and Meetings • Comités et réunions

Day	Time	Place	Meeting
Saturday	18:00-23:00	Belleisle Room II, Lord Nelson	SSC Executive
Sun	9:00–11:00	Vanguard Room II, Lord Nelson	Finance
Sun	11:00–12:00	Vanguard Room II, Lord Nelson	Publications
Sun	12:00–17:00	Admiral Room, Lord Nelson	SSC Board of Directors
Sun	12:00–17:00	Board Room, University Club	Statistics Chairs
Sun	12:00–17:00	Vanguard I, Lord Nelson	NPCDS Committee
Mon	7:15–8:15	Great Hall, University Club	Biostatistics Executive I
Mon	7:15–8:15	Great Hall, University Club	Survey Methods Executive
Mon	7:15–8:15	Great Hall, University Club	BISS Executive
Mon	12:15–13:15	Great Hall, University Club	CJS Editorial Board
Mon	12:15–13:15	Great Hall, University Club	Implementation of Accreditation I
Mon	12:15-13:15	Great Hall, University Club	Public Relations
Mon	17:00–18:00	LSC 242	Biostatistics Section AGM
Mon	17:00–18:00	LSC 240	BISS AGM / BISS Executive
Mon	17:00–18:00	LSC 238	Survey Methods Section AGM
Tue	7:15–8:15	Great Hall, University Club	Implementation of Accreditation II
Tue	7:15–8:15	Great Hall, University Club	Research Committee
Tue	12:15–13:15	Great Hall, University Club	Bilingualism
Tue	12:15–13:15	Great Hall, University Club	Statistical Education
Tue	12:15–13:15	Great Hall, University Club	Committee for Women in Statistics
Tue	17:00–18:30	LSC 242	SSC AGM
Wed	7:15–8:15	Great Hall, University Club	Biostatistics Executive II
Wed	7:15–8:15	Great Hall, University Club	Professional Development
Wed	7:15–8:15	Great Hall, University Club	SORA AGM
Wed	12:15–13:15	Great Hall, University Club	Programme
Wed	17:30–19:30	Board Room, University Club	SSC Board of Directors
Wed	19:30–21:00	Board Room, University Club	SSC Executive

Jour	**Heure**	**Endroit**	**Réunion**
Samedi	18h00-23h00	Belleisle Room II, Lord Nelson	Comité exécutif
Dimanche	9h00–11h00	Vanguard Room II, Lord Nelson	Finances
Dimanche	11h00–12h00	Vanguard Room II, Lord Nelson	Publications
Dimanche	12h00–17h00	Admiral Room, Lord Nelson	Conseil d'administration
Dimanche	12h00–17h00	Board Room, University Club	Directeurs de statistique
Dimanche	12h00–17h00	Vanguard I, Lord Nelson	Comité PNSDC
Lundi	7h15–8h15	Great Hall, University Club	Exécutif, Biostatistiques I
Lundi	7h15–8h15	Great Hall, University Club	Exécutif, Méthodes d'enquêtes
Lundi	7h15–8h15	Great Hall, University Club	Exécutif, Statistique industrielle et de gestion
Lundi	12h15–13h15	Great Hall, University Club	Comité éditorial de la Revue CJS
Lundi	12h15–13h15	Great Hall, University Club	Implémentation de l'accréditation I
Lundi	12h15–13h15	Great Hall, University Club	Relations publiques
Lundi	17h00–18h00	LSC 242	Assemblée générale, Biostatistiques
Lundi	17h00–18h00	LSC 240	Assemblée générale et exécutif, Statistique industrielle et de gestion
Lundi	17h00–18h00	LSC 238	Assemblée générale, Méthodes d'enquête
Mardi	7h15–8h15	Great Hall, University Club	Implémentation de l'accréditation II
Mardi	7h15–8h15	Great Hall, University Club	Recherche
Mardi	12h15–13h15	Great Hall, University Club	Bilinguisme
Mardi	12h15–13h15	Great Hall, University Club	Formation statistique
Mardi	12h15–13h15	Great Hall, University Club	Promotion de la femme en statistique
Mardi	17h00–18h30	LSC 242	Assemblée générale, SSC
Mercredi	7h15–8h15	Great Hall, University Club	Exécutif, Biostatistiques II
Mercredi	7h15–8h15	Great Hall, University Club	Perfectionnement professionnel
Mercredi	7h15–8h15	Great Hall, University Club	Assemblée générale, SORA
Mercredi	12h15–13h15	Great Hall, University Club	Programme
Mercredi	17h30–19h30	Board Room, University Club	Conseil d'administration
Mercredi	19h30–21h00	Board Room, University Club	Comité exécutif

7 Outline of Events • Survol des événements

Building abbreviations • *Abbréviation des édifices*

LNH = Lord Nelson Hotel
LSC = Life Sciences Centre, Dalhousie campus (C204 on Studley map)
UC = University Club, Dalhousie campus (C440 on Studley map)
MM = Marion McCain Building, Dalhousie campus (D420 on Studley map)

SUNDAY • JUNE 8 JUIN • DIMANCHE

9:00 - 5:00 • 9h00 - 17h00

Regency Ballroom, LNH	Biostatistics Workshop *Atelier de biostatistique*
Imperial, LNH	Survey Methods Workshop *Atelier de méthodes d'enquête*
Britannia, LNH	Business and Industrial Statistics Workshop *Atelier de statistique en affaires et dans l'industrie*

5:00 - 6:00 • 17h00 - 18h00

Ondaajte Theatre, MM	***** SPECIAL LATE BREAKING SESSION *****

2:00 - 6:00 • 14h00 - 18h00

Great Hall, UC	Poster Session *Séance par affichage*

6:30 - 10:00 • 18h30 - 22h00 Mixer • Réception d'accueil
Great Hall, University Club

MONDAY • JUNE 9 JUIN • LUNDI

8:30 - 10:00 • 8h30 - 10h00

Ondaajte Theatre, MM	Session 1: Welcome and SSC Presidential Invited Address *Séance 1: Mot de bienvenue du président et allocution de son invité d'honneur*

10:00 - 6:00 • 10h00 - 18h00

Great Hall, UC	Poster Session *Séance par affichage*

10:30 - 12:00 • 10h30 - 12h00

LSC 240	Session 2: Case Study I - Blood Pressure (10:30 - 12:30) *Séance 2: Étude de cas I - Pression artérielle (10h30 - 12h30)*
LSC 332	Session 3: Longitudinal Data Analysis in Biostatistics (Biostatistics section) *Séance 3: Analyse de données longitudinales en biostatistique (Groupe de biostatistique)*
LSC 338	Session 4: Machine Learning Methods From a Statistical Perspective (IMS) *Séance 4: Méthodes d'apprentissage automatique d'un point de vue statistique (IMS)*
LSC 242	Session 5: Statistical Inference I: Inference in Partially Linear Models (Invited papers) *Séance 5: Inférence statistique I: Inférence pour des modèles partiellement linéaires (présentations sur invitation)*
LSC 238	Session 6: Environmetrics (Invited papers) *Séance 6: La mésométrie (présentations sur invitation)*
LSC 234	Session 7: Survey Methods Contributed Session I: Estimation - Applied (Contributed papers) *Séance 7: Méthodes d'enquête I: Estimation - applications (présentations régulières)*

MONDAY • JUNE 9 JUIN • LUNDI

1:30 - 3:00 • 13h30 - 15h00

LSC 338 Session 8: NSERC Open Meeting
Séance 8: Rencontre avec le CRSNG

LSC 240 Session 9: Isobel Loutit Invited Address on Business and Industrial Statistics (Business and Industrial Statistics Section)
Séance 9: Présentation sur invitation Isobel Loutit sur la statistique en affaires et dans l'industrie (Groupe de statistique industrielle et de gestion)

LSC 242 Session 10: Survival Analysis for Complex Data Structures (Survey Methods Section)
Séance 10: Analyse de survie pour des structures de données complexes (Groupe sur les méthodes d'enquête)

LSC 238 Session 11: Rank Methods for Time Series Analysis (Invited papers)
Séance 11: Méthodes basées sur les rangs pour l'analyse de séries chronologiques (présentations sur invitation)

LSC 332 Session 12: Applications of Spatial Statistics (Invited papers)
Séance 12: Applications des statistiques spatiales (présentations sur invitation)

LSC 234 Session 13: Biostatistics Contributed Session I: Estimation and Testing in Biostatistics (Contributed papers)
Séance 13: Biostatistique I: Estimation et tests d'hypothèses en biostatistique (présentations régulières)

MONDAY • JUNE 9 JUIN • LUNDI

3:30 - 5:00 • 15h30 - 17h00

LSC 240 Session 14: Statistics and Information Complexity of Probability Models (Bernoulli Society)
Séance 14: Statistique et complexité de l'information des modèles probabilistes (Société Bernoulli)

LSC 242 Session 15: Data Mining (Invited papers)
Séance 15: Fouille de données (présentations sur invitation)

LSC 238 Session 16: Sequential Methods (Invited papers)
Séance 16: Méthodes séquentielles (présentations sur invitation)

LSC 338 Session 17: Survey Methods Contributed Session II: Applications of Administrative Data (Contributed papers)
Séance 17: Méthodes d'enquête II: Applications avec des données administratives (présentations régulières)

LSC 234 Session 18: Applications of Statistics (Contributed papers)
Séance 18: Applications de la Statistique (présentations régulières)

LSC 332 Session 19: Statisticians in Action I (Committee on Professional Development)
Séance 19: Statisticiens en action I (Comité sur le perfectionnement professionnel)

6:30 - 10:00 • Barbecue • 18h30 - 22h00
Sheriff Hall Dining Room

TUESDAY • JUNE 10 JUIN • MARDI

8:30 - 10:00 • 8h30 - 10h00

Ondaajte Theatre, MM — Session 20: SSC Gold Medal Address
Séance 20: Allocution du récipiendaire de la médaille d'or de la SSC

10:30 - 12:00 • 10h30 - 12h00

LSC 240 — Session 21: Shape-Restricted Inference (IMS)
Séance 21: Inférence avec restrictions sur la forme (IMS)

LSC 338 — Session 22: Analysis of Mixed Discrete and Continuous Outcome Data (Biostatistics Section)
Séance 22: Analyse de mélanges de variables discrètes et continues (Groupe de biostatistique)

LSC 242 — Session 23: Recent Developments in Small Area Estimation (Special Invited Session of the Survey Methods Section)
Séance 23: Développements récents en estimation sur de petits domaines (Présentation sur invitation spéciale de la section sur les méthodes d'enquête)

LSC 238 — Session 24: Special Session of the Pacific Institute for the Mathematical Sciences on Robustness (Invited papers)
Séance 24: Session spéciale du Pacific Institute for the Mathematical Sciences en robustesse (présentations sur invitation)

LSC 332 — Session 25: Time Series Methods for Fitting Dynamical Models (Invited papers)
Séance 25: Méthodes de séries temporelles pour l'ajustement de modèles dynamiques (présentations sur invitation)

LSC 234 — Session 26: Decision Theory and Bayesian Methods and Resampling (Contributed papers)
Séance 26: Théorie de la décision, méthodes bayésiennes et rééchantillonnage (présentations régulières)

TUESDAY • JUNE 10 JUIN • MARDI

1:30 - 3:00 • 13h30 - 15h00

LSC 240 Session 27: Statistical Issues in Modern Biology (Caucus for Women in Statistics -Canadian Section- and Women in Statistics Committee)
Séance 27: Considérations statistiques en biologie moderne (Caucus pour les femmes en statistique -section canadienne- et comité sur les femmes en statistique)

LSC 242 Session 28: Resampling Methods (Invited papers)
Séance 28: Méthodes de rééchantillonnage (présentations sur invitation)

LSC 238 Session 29: Experimental Design (Invited papers)
Séance 29: Plans d'expérience (présentations sur invitation)

LSC 338 Session 30: Survey Methods Contributed Session III: Methods for Health Surveys (Contributed papers)
Séance 30: Méthodes d'enquête III: Méthodes pour les enquêtes sur la santé (présentations régulières)

LSC 332 Session 31: Robust Methods I (Contributed papers)
Séance 31: Méthodes robustes I (présentations régulières)

LSC 234 Session 32: Stochastic Processes and Finance and Applied Probability (Contributed papers)
Séance 32: Processus stochastique et finance et probabilités appliquées (présentations régulières)

TUESDAY • JUNE 10 JUIN • MARDI

3:30 - 5:00 • 15h30 - 17h00

LSC 240 Session 33: Comparative Research (Survey Methods Section)
Séance 33: Études comparatives (Groupe sur les méthodes d'enquête)

LSC 242 Session 34: Process Monitoring (Business and Industrial Statistics Section)
Séance 34: Suivi de processus
(Groupe de statistique industrielle et de gestion)

LSC 238 Session 35: Special Session of the Fields Institute on Matrices and Statistics (Invited papers)
Séance 35: Session spéciale de l'Institut Fields sur les matrices en statistique (présentations sur invitation)

LSC 338 Session 36: Statistics in Genomics and Proteomics (Invited papers)
Séance 36: Statistique en génomique et protéomique (présentations sur invitation)

LSC 234 Session 37: Statistical Computing and Computationally Intensive Methods (Contributed papers)
Séance 37: Statistique informatique et méthodes informatiques intensives (présentations régulières)

LSC 332 Session 38: Statisticians in Action II (Committee on Professional Development)
Séance 38: Statisticiens en action II (Comité sur le perfectionnement professionnel)

5:00 - 6:30 • 17h00 - 18h30 Annual General Meeting • Réunion générale annuelle
LSC 240

6:30 - 10:00 • Banquet • 18h30 - 22h00
Pier 21

WEDNESDAY • JUNE 11 JUIN • MERCREDI

8:30 - 10:00 • 8h30 - 10h00

LSC 238 Session 39: Statistical Methods for Health Services and Outcomes Research (Biostatistics Section)
Séance 39: Méthodes statistiques pour les services de santé et la recherche sur les outcomes (Groupe de biostatistique)

LSC 242 Session 40: Statistics and Climate Change (Invited papers)
Séance 40: Statistique et changement climatique (présentations sur invitation)

LSC 338 Session 41: Bayesian Analysis (Invited papers)
Séance 41: Analyse bayésienne (présentations sur invitation)

LSC 332 Session 42: Statistical Inference (Contributed papers)
Séance 42: Inférence statistique (présentations régulières)

LSC 234 Session 43: Survey Methods Contributed Session IV: Measuring the Quality of Survey Operations (Contributed papers)
Séance 43: Méthodes d'enquête IV: Mesure de la qualité des opérations d'enquêtes (présentations régulières)

LSC 240 Session 44: 8:30 - 9:15: Pierre Robillard Award Winner Lecture
Séance 44: 8h30 - 9h15: Allocution du lauréat du prix Pierre Robillard

LSC 240 Session 45: 9:15 - 10:00: Canadian Journal of Statistics Award Winner Lecture
Séance 45: 9h15 - 10h00: Allocution du lauréat du prix de la Revue canadienne de statistique

WEDNESDAY • JUNE 11 JUIN • MERCREDI

10:30 - 12:00 • 10h30 - 12h00

LSC 240 Session 46: Case Study II - Neighbourhood Factors and Children: Hierarchical Linear Models and Small Area Statistics (10:30 - 12:30)
Séance 46: Étude de cas II - Facteurs de voisinage et enfants: Modèles linéaires hiérarchiques et statistiques sur des petits domaines (10h30 - 12h30)

LSC 238 Session 47: Nonparametric Analysis in Natural Resources Surveys (Survey Methods Section)
Séance 47: Analyse non paramétrique pour les enquêtes sur les ressources naturelles (Groupe sur les méthodes d'enquête)

LSC 242 Session 48: Statistical Inference II: Inference Problems with Missing Data or Measurement Errors (Invited papers)
Séance 48: Inférence statistique: Problèmes d'inférence avec des données manquantes et des erreurs de mesure (présentations sur invitation)

LSC 338 Session 49: Biostatistics Contributed Session II: Epidemiological and Clinical Studies (Contributed papers)
Séance 49: Biostatistique II: Études épidémiologiques et cliniques (présentations régulières)

LSC 332 Session 50: Design and Analysis of Experiments (Contributed papers)
Séance 50: Planification et analyse d'expériences (présentations régulières)

LSC 234 Session 51: Robust Methods II and Statistical Education (Contributed papers)
Séance 51: Méthodes robustes II et éducation statistique (présentations régulières)

WEDNESDAY • JUNE 11 JUIN • MERCREDI

1:30 - 3:00 • 13h30 - 15h00

LSC 238 Session 52: Stochastic Aspects of Forestry (Canadian Operational Research Society)
Séance 52: Aspects stochastiques de la foresterie (Société canadienne de recherche opérationnelle)

LSC 338 Session 53: Estimation of Fish Stock Mixtures
(Biostatistics Section)
Séance 53: Estimation des mélanges de stocks de poissons
(Groupe de biostatistique)

LSC 240 Session 54: Applied Probability (Invited papers)
Séance 54: Probabilité appliquée (présentations sur invitation)

LSC 242 Session 55: Special Session of the Centre de Recherches Mathématiques on Statistics and Finance (Invited papers)
Séance 55: Session speciale du Centre de Recherches Mathématiques en statistique et finance (présentations sur invitation)

LSC 332 Session 56: Survey Methods Contributed Session V: Survey Sampling
(Contributed papers)
Séance 56: Méthodes d'enquête V: sondages (présentations régulières)

LSC 234 Session 57: Distributions and Multivariate Methods (Contributed papers)
Séance 57: Distribution et méthodes multidimensionnelles (présentations régulières)

WEDNESDAY • JUNE 11 JUIN • MERCREDI

3:30 - 5:00 • 15h30 - 17h00

LSC 240 Session 58: Business and Economic Statistics
(Business and Industrial Statistics Section)
Séance 58: Statistique en affaires et en économie
(Groupe de statistique industrielle et de gestion)

LSC 242 Session 59: Variable Selection (Invited papers)
Séance 59: Sélection de variables (présentations sur invitation)

LSC 238 Session 60: Inference for Time Series and Other Models of Dependence
(Contributed papers)
Séance 60: Inférence pour séries chronologiques et autres modèles de dépendance
(présentations régulières)

LSC 338 Session 61: Survey Methods Contributed Session VI: Estimation - Theoretical
(Contributed papers)
Séance 61: Méthodes d'enquête VI: Estimation - théorie
(présentations régulières)

LSC 234 Session 62: Biostatistics Contributed Session III**:** Survival and Clustered Data
(Contributed papers)
Séance 62: Biostatistique III: Données de survie et corrélées en grappes
(présentations régulières)

LSC 332 Session 63: Statisticians in Action III (Committee on Professional Development)
Séance 63: Statisticiens en action III (Comité sur le perfectionnement professionnel)

NOTE: When multiple contributors to presentations are listed, the first named contributor is the presenter. The others are grouped according to their affiliation, rather than to any presumed priority.

NOTE: Lorsque plusieurs co-auteurs sont mentionnés, le premier est le présentateur. Les autres sont regroupés selon leur affiliation et non pas selon un autre ordre quelconque.

8 Scientific Programme • Programme Scientifique

Sunday June 8 • Dimanche 8 juin

9:00 - 5:00 • 9h00-17h00 Biostatistics Workshop • Atelier de biostatistique Regency,LNH

Edward. F. Vonesh, Baxter Healthcare Corporation
Mixed-effects models for longitudinal data • Modèles à effets mixtes pour données longitudinales

9:00 - 5:00 • 9h00-17h00 Survey Methods Workshop • Atelier de méthodologie d'enquête Imperial, LNH

Pierre Lavallée, Statistics Canada/Statistique Canada
Panel surveys • Atelier sur les enquêtes longitudinales

9:00 - 5:00 • 9h00-17h00 Business and Industrial Statistics Workshop • Atelier de statistique en affaires et dans l'industrie Britannia, LNH

Doug Montgomery, Arizona State University
Response Surface Methodology: Process and Product Optimization Using Designed Experiments • Méthodologie de surface de réponse: Optimisation des processus et des produits à l'aide d'expériences planifiées

5:00 - 6:00 • 17h00-18h00 Special Late Breaking Session Ondaajte Theatre, MM

2:00 - 6:00 • 14h00-18h00 Poster Session • Séance par affichage (also Monday 10:00 - 6:00 • aussi Lundi 10h00 - 18h00) UC Great Hall

Eshetu Atenafu, Paul N. Corey, University of Toronto
Statistical properties of different ratio estimators of recommended daily food allowance estimates • Propriétés statistiques de différents estimateurs du rapport pour l'estimation de la quantité journalière d'aliments recommandée

Jeffrey Bakal, J.T. Smith, G. Takahara, Queen's University
Issues in clustering biomechanical data • Problèmes dans le regroupement de données biomécaniques

David Brillinger, University of California, Berkeley
Regression analysis and mutual information • Analyse de régression et information mutuelle

Noel Cadigan, Department of Fisheries and Oceans; Patrick Farrell, Carleton University
Local influence diagnostics for the retrospective problem in sequential population analyses of fishery data • Diagnostique d'influence locale pour le problème rétrospectif dans l'analyse séquentielle de la population sur des données sur les pêches

Judy-Anne Chapman, University of Waterloo & University of Toronto; Jiaming Sun, University of Toronto; Richard Gordon, Radhika Sivaramakrishna, University of Manitoba; Marilyn Link, Edward B. Fish, University of Toronto
Use of location-scale (log-normal) survival analysis to model survival from primary breast cancer after routine clinical use of mammography • L'utilisation de l'analyse de survie localisation-échelle (log-normale) pour modéliser la survie avec le cancer du sein primaire après avoir passer des mammographies de manière routinière.

Gemai Chen, University of Calgary; Jinhong You, The Hong Kong Polytechnic University
Moving block delete-1 jackknifing in partially linear regression models with m-dependent errors • Jackknife delete-1 à blocs mobiles dans les modèles de régression partiellement linéaire avec erreurs m-dépendantes.

Laura Cowen, Carl Schwarz, Simon Fraser University
Comparing survival estimates from a radio-tag mark-recapture study • Comparaison d'estimateurs de survie à partir d'une étude de marquage par reprise avec des étiquettes radio

Sandra Gardner, University of Toronto
Change point models for modeling discontinuation rates of Pneumocystis Carinii Pneumonia Prophylaxis in an Ontario HIV patient population • Modèles de changement ponctuel pour modéliser les taux de discontinuité de la prophylaxie pour la pneumonie à Pneumocystis Carinii dans une population de patients atteint du VIH en Ontario.

Paramjit Gill, Michael Treschow, Okanagan University College
A Stylometric Analysis of King Alfred's Literary Works • Analyse stylométrique du travail littéraire du Roi Alfred

Cristina Goia, Ontario HIV Treatment Network; A.M. Bayoumi, St. Michael's Hospital, Toronto
Imputation of missing date values in medication records data • Imputation des dates manquantes pour des données des dossiers sur la médication

Cristina Goia, Ontario HIV Treatment Network; A.M. Bayoumi, St. Michael's Hospital, Toronto
Using a matched-cohort in order to validate the effectiveness of an intervention in HIV routine practice • Utiliser une cohorte appariée pour valider l'efficacité d'une intervention dans les pratiques courantes sur le VIH

Wenqing He, Shelley B. Bull, Nalan Gokgoz, Irene Andrulis, Jay Wunder, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, University of Toronto
Statistical Issues in Microarray Data Analysis of Sarcoma Tumors • Problèmes statistiques dans l'analyse de micro réseaux de données sur les tumeurs sarcoma

Sohee Kang, John Hsieh, University of Toronto
Quality of life and survival analysis-an alternative approach to Q-TWiST • Analyse de survie et de qualité de vie: une approche alternative au Q-TWiST

Melanie Laframboise, J. Gough, D.A. Marshall, B. Jaszewski, Innovus Research Inc.
Methods used to determine the impact of administrative restrictions for antibiotic use and expenditures • Méthodes utilisées pour déterminer l'impact des restrictions administratives sur l'utilisation et les dépenses en médicaments antibiotiques

Yi Liu, Alwell J. Oyet, Memorial University of Newfoundland
Minimax designs for discrimination between competing wavelet regression models • Le design minimax pour la discrimination entre des modèles concurrents de régression par ondelettes

Ahmed Lmoudden, J. Vaillancourt, K. Ghoudi, Université de Sherbrooke
Partial convergence of Kendall's process to the Brownian Bridge: test of independence • La convergence partielle du processus de Kendall vers le pont Brownien: test d'indépendance

Cyr Emile M'lan, Hospital for Sick Children, Toronto
Bayesian sample size calculation for case-control studies • Méthodes bayésiennes de calcul de taille d'échantillon pour les études cas-témoins

Sandra Olfert, P. Pahwa, J.A. Dosman, University of Saskatchewan
Longitudinal analysis of pulmonary dysfunction in the initial years of employment in the grain industry • Analyse longitudinale des problèmes pulmonaires pour les premières années d'emploi dans l'industrie du grain

Jennifer Prokop, W.J. Braun, University of Western Ontario; V. Rousson, University of Zurich; W.A. Simpson, Glasgow Caledonian University
Statistical inference for reaction time experiment data • Inférence statistique pour des données d'expériences de temps de réaction

Jin Qian, Judy-Anne W. Chapman, Yuejiao Fu, Yan Yuan, University of Waterloo; David E. Axelrod, Rutgers University; Naomi A. Miller, Princess Margaret Hospital; William A. Christens-Barry, Equipose Imaging LLC; H. Lavina Lickley, Wedad M. Hanna, Sunnybrook and Women's
Sample size implications for nuclear assessment of non-invasive breast cancer (DCIS) • Les implications de la taille d'échantillon sur l'évaluation nucléaire non-envahissante du cancer du sein (DCIS)

Marylène Troupe, Université des Antilles-Guyane (Guadeloupe); Samuel Barclay, Jean Vaillant, Université des Antilles-Guyane; Petr Lansky, Academy of Sciences of the Czech Republic
Statistic of time point process associated with a stochastic trajectory in heterogeneous environment • Statistique d'un processus ponctuel temporel dirigé par une trajectoire stochastique en milieu hétérogène

Anjela Tzontcheva, University of Toronto
Application of models for interval-censored survival data with informative examination time to the Polaris HIV seroconversion study data in Ontario • Application de modèles pour des données de survie censurées par intervalles avec temps d'examen informatifs avec les données de l'étude de séroconversion du VIH Polaris en Ontario

Zilin Wang, David R. Bellhouse, University of Western Ontario; Jamie Stafford, University of Toronto
Generalized additive models for complex survey data • Modèles additifs généralisés pour des données de sondages complexes

Machelle Wilson, William McCormick, Tom Hinton, University of Georgia
Monte Carlo comparisons of maximum likelihood estimation of high quantiles to the extreme value estimate of maximal exposure • Comparaisons par méthode de Monte Carlo de l'estimateur du maximum de vraisemblance des quantiles élevés à l'estimateur de valeur extrême de l'exposition maximale

Zhang Ying, Ian McLeod, Hao Yu, University of Western Ontario
Experiments in multiple-choice randomized exams • Expériences pour les examens à choix multiples randomisés

Zhang Ying, Ian McLeod, University of Western Ontario
Visualization on subset autoregressive admissible boundaries • Visualisation sur des frontières admissibles d'un sous ensemble autorégressif

Ahmad Zoghoul, Mu'tah University
Estimation based on sample records versus the whole sample • Estimation basée sur des observations d'un échantillon versus tout l'échantillon

Monday June 9 • Lundi 9 juin

8:30 - 10:00 • 8h30 - 10h00 Session/Séance 1 **MM Ondaajte Theatre**

Welcome and SSC Presidential Invited Address • Mot de bienvenue du président et allocution de son invité d'honneur
Special Session • Conférence spéciale
Organizer and Chair • Responsable et président: Jim Ramsay, McGill University

Robert Gentleman, Harvard School of Public Health
Modern statistical computing • Calcul statistique moderne

10:00 - 6:00 • 10h00-18h00 Poster Session • Séance par affichage (also Sunday 2:00 - 6:00 • aussi Dimanche 14h00 - 18h00) **UC Great Hall**

10:30 - 12:30 • 10h30-12h30 Session/Séance 2 **LSC 240**

Case Study I - Blood Pressure • Étude de cas I - Pression artérielle
Organizer and Chair • Responsable et président: Peggy Ng, York University

10:30 • 10h30 Introduction: Raymond Lam, GlaxoKlineSmith

10:35 • 10h35 Pingzhao Hu, Dong Song, Dalhousie University

10:50 • 10h50 Zainab Abdurrahman, Bethany Giddings, Sofia Mosesova, University of Waterlooo

11:05 • 11h05 Louis-François Poirier, Javier Oyarzun, Université de Montréal

11:20 • 11h20 Ana-Maria Staicu, Mark Kane, Hadas Moshonov, University of Toronto

11:35 • 11h35 Hossein Yazdi, University of Guelph

11:50 • 11h50 Sophia Lee, Hanna Jankowksi, Joanna Biernacka, University of Toronto

12:05 • 12h05 Christina Frisina, Cathlin McNally, McMaster University

12:20 • 12h20 Discussion

10:30 - 12:00 • 10h30-12h00 Session/Séance 3 **LSC 332**

Longitudinal Data Analysis in Biostatistics • Analyse de données longitudinales en biostatistique
Invited Paper Session • présentations sur invitation: **Biostatistics Section • Groupe de biostatistique**
Organizer and Chair • Responsable et président: Gary Sneddon, Memorial University of Newfoundland

10:30 • 10h30 Brajendra Sutradhar, Gary Sneddon, Memorial University of Newfoundland
Analyzing two-way correlated familial longitudinal data • Analyse de la corrélation double dans des données longitudinales familiales

11:00 • 11h00 Raymond Carroll, Texas A&M University
Longitudinal data analysis in biostatistics • Données longitudinales et en grappes et régression non paramétrique et semi-paramétrique

11:30 • 11h30 Georgia Roberts, Milorad Kovacevic, Yves Lafortune, Owen Phillips, David Binder, Statistics Canada/Statistique Canada
Using an estimating equation bootstrap approach for obtaining variance estimates when modelling complex health survey data • L'emploi d'une approche bootstrap aux équations d'estimation pour obtenir des estimations de variance lors de la modélisation de donnés provenant des enquêtes sur la santé à plan complexe

10:30 - 12:00 • 10h30-12h00 Session/Séance 4 **LSC 338**

Machine Learning Methods From a Statistical Perspective • Méthodes d'apprentissage automatique d'un point de vue statistique
Invited Paper Session • présentations sur invitation: **Institute of Mathematical Statistics**
Organizer • Responsable: Yi Lin, University of Wisconsin
Chair • Président: Hao Zhang, North Carolina State University

10:30 • 10h30 George C. Tseng, Wing Hung Wong, Harvard University; Xiaotong Shen, Ohio State University; Xuegong Zhang, Tsinghua University
From margin-based classification to ψ-learning • De la classification basé sur les marges aux " ψ-learning"

11:00 • 11h00 Peter Bühlmann, ETH Zurich
Boosting methods: why they can be useful for high-dimensional data • Les méthodes de boosting: pourquoi peuvent-elles être utiles avec des données de haute dimension

11:30 • 11h30 Yi Lin, Yongho Jeon, University of Wisconsin
Random forests and adaptive nearest neighbors • Les forêts aléatoires et les voisins les plus proches adaptatifs

10:30 - 12:00 • 10h30-12h00 Session/Séance 5 **LSC 242**

Statistical Inference I: Inference in Partially Linear Models • Inférence statistique I: Inférence pour des modèles partiellement linéaires
Invited Paper Session • présentations sur invitation
Organizer • Responsable: Nancy Reid, University of Toronto
Chair • Président: Jamie Stafford, University of Toronto

10:30 • 10h30 Aidan McDermott, Francesca Dominici, Trevor Hastie, Scott L. Zeger, Jonathan Samet, Johns Hopkins University
Issues in semiparametric regression • Sujets en régression semi-paramétrique

11:00 • 11h00 Tim Ramsay, Daniel Krewski, University of Ottawa/Universite d'Ottawa; Richard Burnett, Health Canada
Concurvity-induced bias in the generalized additive model • Biais de concurvité induit dans les modèles additifs généralisés

11:30 • 11h30 Isabella Ghement, University of British Columbia
Smoothing parameter selection in partially linear models with serially correlated errors • Choix du paramètre de lissage dans les modèles partiellement linéaires avec erreurs corrélées en série

10:30 - 12:00 • 10h30-12h00 Session/Séance 6 **LSC 238**

Environmetrics • La mésométrie
Invited Paper Session • présentations sur invitation
Organizer and Chair • Responsable et président: Abdel El-Shaarawi, National Water Research Institute

10:30 • 10h30 Sylvia Esterby, University College of the Okanagan
Models for biological productivity in lakes • Modèles pour la productivité biologique dans les lacs

11:00 • 11h00 Montserrat Fuentes, North Carolina State University
Statistical assessment of geographic areas of compliance with air quality standards • Évaluation statistique des secteurs géographiques en conformité avec les standards sur la qualité de l'air

11:30 • 11h30 Lawrence H. Cox, National Center for Health Statistics/Centre national pour les statistiques en santé
On properties of multi-dimensional statistical tables • Les propriétés des tables statistiques multidimensionnelles

10:30 - 12:00 • 10h30-12h00 Session/Séance 7 **LSC 234**

Survey Methods Contributed Session I: Estimation - Applied • Méthodes d'enquête I: Estimation - applications
Contributed Paper Session • présentations régulières
Organizer • Responsable: Patricia Whitridge, Statistics Canada/Statistique Canada
Chair • Président: François Brisebois, Statistics Canada/Statistique Canada

10:30 • 10h30 Pat Newcombe-Welch, Statistics Canada/Statistique Canada, University of Waterloo; Tim Seifert, Memorial University
Big country, small sample - what can we safely conclude? • Grand pays, petit échantillon: que pouvons nous conclure de manière sûre?

10:45 • 10h45 André Cyr, Catalin Dochitoiu, Statistics Canada/Statistique Canada
Latent variable modelling in the longitudinal context: The case of the National Longitudinal Survey of Children and Youth in Canada • Modélisation de variables latentes dans le contexte longitudinal:Le cas pour l'Enquête Longitudinale Nationale sur les Enfants et les Jeunes au Canada

11:00 • 11h00 Susie Fortier, Hélène Bérard, Statistics Canada/Statistique Canada
Producing historical data according to a new classification: the experience of the Monthly Wholesale and Retail Trade Survey • La conversion de données historiques selon un nouveau système de classification: le cas de l'Enquête mensuelle sur le commerce de gros et de détail

11:15 • 11h15 Hélène Bérard, Statistics Canada/Statistique Canada
Dealing with misclassified units in repeated business surveys: the experience of the redesigned Canadian Monthly Wholesale and Retail Trade Survey • Traiter les unités classées de façon erronée dans un contexte d'enquêtes répétées: L'expérience de la nouvelle Enquête mensuelle sur le commerce de gros et de détail (EMCGD) au Canada

11:30 • 11h30 Richard Belcher, Statistics Canada/Statistique Canada
Application of the Hidiroglou-Berthelot method of outlier detection for periodic business surveys • L'application de la méthode de Hidiroglou et Berthelot pour la détection des valeurs aberrantes pour les enquêtes-entreprises

11:45 • 11h45

1:30 - 3:00 • 13h30-15h00 Session/Séance 8 **LSC 338**

NSERC Open Meeting • Rencontre avec le CRSNG

Bruce Smith, Dalhousie University
Workshop: How to prepare a winning NSERC proposal • Atelier: Comment préparer une demande gagnante au CRSNG

Judie Foster, NSERC • CRSNG
Reallocations exercise: Recommendations and impact for Statistical Sciences • Exercice de réallocation: recommandations et impact pour les sciences statistiques

Jamie Stafford, University of Toronto
National Program on Complex Data Structures • Programme national sur les structures de données complexes

1:30 - 3:00 • 13h30-15h00 Session/Séance 9 **LSC 240**

Isobel Loutit Invited Address on Business and Industrial Statistics • Présentation sur invitation Isobel Loutit sur la statistique en affaires et dans l'industrie
Special Session • Conférence spéciale: **Business and Industrial Statistics Section • Groupe de statistique industrielle et de gestion**
Organizer and Chair • Responsable et président: John Brewster, University of Manitoba

Doug Montgomery, Arizona State University
The modern practice of statistics in business and industry • La pratique moderne des statistiques en affaires et dans l'industrie

1:30 - 3:00 • 13h30-15h00 Session/Séance 10 **LSC 242**

Survival Analysis for Complex Data Structures • Analyse de survie pour des structures de données complexes
Invited Paper Session • présentations sur invitation: **Survey Methods Section • Groupe sur les méthodes d'enquête**
Organizer • Responsable: Susana Rubin-Bleuer, Statistics Canada/Statistique Canada
Chair • Président: Georgia Roberts, Statistics Canada/Statistique Canada

1:30 • 13h30 Jerry Lawless, University of Waterloo
Censoring and weighting in survival estimation from survey data • Troncation et pondération dans l'estimation de survie pour des données de sondages

2:00 • 14h00 Susana Rubin-Bleuer, Statistics Canada/Statistique Canada
Tightness of survival processes in a joint design-model space • Tension des processus de survie dans un espace conjoint modèle-design

2:30 • 14h30 Y. Peng, Memorial University of Newfoundland
Semiparametric cure models and some computational issues • Modèles de traitements semi-paramétriques et quelques problèmes informatiques

1:30 - 3:00 • 13h30-15h00 Session/Séance 11 **LSC 238**

Rank Methods for Time Series Analysis • Méthodes basées sur les rangs pour l'analyse de séries chronologiques
Invited Paper Session • présentations sur invitation
Organizer and Chair • Responsable et président: Christian Genest, Université Laval

1:30 • 13h30 Jean-Marie Dufour, Université de Montréal
Finite-sample distribution-free inference in regression models under general forms of dependence • Inférence non-paramétrique exacte dans des modèles de régression avec dépendance générale

2:00 • 14h00 Marc Hallin, Davy Paindaveine, Université libre de Bruxelles
Optimal rank-based procedures for testing multivariate elliptic white noise against VARMA dependence • Tests de rangs multivariés optimaux pour l'hypothèse de bruit blanc elliptique

2:30 • 14h30 Bruno Rémillard, École des hautes études commerciales, Montréal; Christian Genest, Université Laval
Tests of independence based on the empirical copula process • Tests d'indépendance basés sur le processus de copule empirique

1:30 - 3:00 • 13h30-15h00 Session/Séance 12 **LSC 332**

Applications of Spatial Statistics • Applications des statistiques spatiales
Invited Paper Session • présentations sur invitation
Organizer • Responsable: Subhash Lele, University of Alberta
Chair • Président: Gary Umphrey, University of Guelph

1:30 • 13h30 Lance Waller, Traci Leong, Andrew Barclay, Emory University; Bud Howard, Palm Beach County
A spatial analysis of Sea Turtle nesting patterns in Palm Beach County, Florida • Une analyse spatiale des patrons de nidification des tortues de mer dans le comté de Palm Beach en Floride

2:00 • 14h00 Carol Gotway Crawford, Center for Disease Control; Linda J. Young, University of Nebraska
A geostatistical approach to combining incompatible spatial data • Une approche géostatistique pour la combinaison de données spatiales incompatibles

2:30 • 14h30 Richard Hoskins, Washington State Department of Public Health and Epidemiology
A comparison of boundary detection, spatial scan and cluster detection methods applied to infant deaths in Washington State • Comparaison des méthodes de détection des frontières, de scan spatial et de détection des grappes appliquées à la mortalité infantile dans l'État de Washinton

1:30 - 3:00 • 13h30-15h00 Session/Séance 13 **LSC 234**

Biostatistics Contributed Session I: Estimation and Testing in Biostatistics • Biostatistique I: Estimation et tests d'hypothèses en biostatistique
Contributed Paper Session • présentations régulières
Chair • Président: Xiaoming Sheng, University of Utah

1:30 • 13h30 Abbas Khalili, University of Waterloo; Noel Cadigan, Department of Fisheries and Oceans
Inference for sequential population analysis using penalized likelihood methods • Inférence pour l'analyse séquentielle de la population (ASP) en utilisant des méthodes de vraisemblance pénalisée

1:45 • 13h45 Andrea Benedetti, Michal Abrahamowicz, McGill University & Montreal General Hospital
Smoothing parameter selection and impact on inference in generalized additive models • La sélection des paramètres de lissage et son impact sur l'inférence des modèles additifs généralisés

2:00 • 14h00 Keyue Ding, W. J. Hall, Queen's University
Sequential tests and estimates after overrunning based on p-value combination • Tests séquentiels et estimés après renvoi basé sur des combinaisons de seuils expérimentaux

2:15 • 14h15 Yang Zhao, J.F. Lawless, D.L. McLeish, University of Waterloo
Efficient estimation in regression analysis with missing data in two-phase sampling designs • Estimation efficace en analyse de régression avec des données manquantes pour des designs d'échantillonnage à deux étapes

2:30 • 14h30 François Bellavance, HEC Montréal and Centre for Research on Transportation; Mustapha Bourhattas, Stéphane Messier, CRT; Sophie Lapierre, École Polythecnique & CRT; Claire Laberge-Nadeau, Université de Montréal & CRT
Misclassification bias in the case-crossover design applied to wireless telephones and the risk of road crashes • Le devis cas chassé-croisé et le biais de mauvaise classification dans l'estimation du risque d'accident en utilisant le téléphone mobile au volant

2:45 • 14h45 Karelyn Davis, Chu-In Charles Lee, Memorial University of Newfoundland; Jianan Peng, Acadia University
Step-down testing procedure for dose-response studies • Procédure de test step-down pour des études dose-réponse

---

3:30 - 5:00 • 15h30-17h00 Session/Séance 14 **LSC 240**

Statistics and Information Complexity of Probability Models • Statistique et complexité de l'information des modèles probabilistes
Invited Paper Session • présentations sur invitation: **Bernoulli Society • Société Bernoulli**
Organizer and Chair • Responsable et président: Boris Levit, Queen's University

3:30 • 15h30 V. Koltchinskii, University of New Mexico
Complexities and margins in binary classification problems •Les complexités et les marges dans les problèmes de classification binaire

4:00 • 16h00 Yuri Golubev, University of Marseilles
Oracle inequalities and estimation of sparse vectors • Inégalités d'oracle et estimation de vecteurs "sparse"

4:30 • 16h30

3:30 - 5:00 • 15h30-17h00 Session/Séance 15 **LSC 242**

Data Mining • Fouille de données
Invited Paper Session • présentations sur invitation
Organizer and Chair • Responsable et président: Peter Hooper, University of Alberta

3:30 • 15h30 Ruben Zamar, University of British Columbia
Data mining, data quality and robustness • Fouille de données, qualité des données et robustesse

4:00 • 16h00 Hugh Chipman, University of Waterloo; Edward I. George, University of Pennsylvania; Robert E. McCulloch, University of Chicago
Boosting Bayesian tree models • Application du boosting aux arbres bayésien

4:30 • 16h30 Wayne Oldford, University of Waterloo
Structuring interactive cluster analysis • Structurer l'analyse de groupement interactive

3:30 - 5:00 • 15h30-17h00 Session/Séance 16 **LSC 238**

Sequential Methods • Méthodes séquentielles
Invited Paper Session • présentations sur invitation
Organizer and Chair • Responsable et président: Edit Gombay, University of Alberta

3:30 • 15h30 Richard Cook, University of Waterloo
Robust methods for monitoring trials with multiple treatment periods and recurrent events • Méthodes robustes pour surveiller des essais avec plusieurs périodes de traitement et avec événements récurrents

4:00 • 16h00 John Petkau, University of British Columbia
A simple model for drug screening programs • Un modèle simple pour les programmes de dépistage des drogues

4:30 • 16h30 Edit Gombay, University of Alberta
Sequential testing strategies • Stratégies pour effectuer des tests séquentiels

3:30 - 5:00 • 15h30-17h00 Session/Séance 17 **LSC 338**

Survey Methods Contributed Session II: Applications of Administrative Data • Méthodes d'enquête II: Applications avec des données administratives
Contributed Paper Session • présentations régulières
Organizer and Chair • Responsable et président: Wesley Yung, Statistics Canada/Statistique Canada

3:30 • 15h30 Richard Dorsett, Policy Studies Institute
Refreshment samples, matching and attrition bias • Échantillon de remplacement, association et biais d'attrition

3:45 • 15h45 Jason Sutherland, C.J. Schwarz, Simon Fraser University
Multi-list methods using incomplete lists in closed populations • Méthodes de listes multiples en utilisant des listes incomplètes dans des populations fermées

4:00 • 16h00 Guylaine Dubreuil, Mike Hidiroglou, Louis Pierre, Statistics Canada/Statistique Canada
Use of administrative data in the modelling of monthly survey data • Utilisation de données administratives dans la modélisation des données d'une enquête mensuelle

4:15 • 16h15 Daniel Hurtubise, Statistics Canada/Statistique Canada
Variance estimation in the context of complex surveys using administrative data • Estimation de la variance dans le cadre d'enquêtes complexes utilisant des données administratives

4:30 • 16h45

4:45 • 16h45

3:30 - 5:00 • 15h30-17h00 Session/Séance 18 **LSC 234**

Applications of Statistics • Applications de la Statistique
Contributed Paper Session • présentations régulières
Chair • Président: Rhonda Rosychuk, University of Alberta

3:30 • 15h30 James Wasiloff, Eric Wasiloff, Michigan State University
Improving the reliability and robustness of a pneumatic paint ball marker system through the practical application of parameter design optimization methodologies • Amélioration de la fiabilité et de la robustesse d'un système de marqueur pneumatique de balles de peintures par l'entremise d'une application pratique de méthodes d'optimisation de design de paramètres

3:45 • 15h45 Asokan Mulayath Variyath, Bovas Abraham, University of Waterloo
Estimation of vendor's process capability based on submitted lots • Estimation de la capacité des processus des fournisseurs basées sur des pièces fournies

4:00 • 16h00 Ejaz Ahmed, University of Windsor
Comparing multiple process capability indices in non-normal distributions • Comparaison des indices de processus multiples de capacité pour des distributions non normales

4:15 • 16h15 Michael MacLeod, St. Francis Xavier University; René F. Reitsma, Oregon State University; Lehana Thabane, McMaster University
Spatialization of web sites using a weighted frequency model of navigation data • Spatialisation des sites web en utilisant un modèle de fréquences pondérées des données sur la navigation

4:30 • 16h30 Rolf Turner, Pradeep Banerjee, University of New Brunswick
A differential equations approach to some asset selling problems • Une approche par les équations différentielles pour certains problèmes de vente d'actifs

4:45 • 16h45 Theodoro Koulis, University of Waterloo
A stochastic model for sea ice • Un modèle stochastique pour la glace de mer

3:30 - 5:00 • 15h30-17h00 Session/Séance 19 **LSC 332**

Statisticians in Action I • Statisticiens en action I

Video presentation • Présentation vidéo: **Committee on Professional Development • Comité sur le perfectionnement professionnel**

Chair • Président: Jon Baskerville

Tuesday June 10 • Mardi 10 juin

8:30 - 10:00 • 8h30 - 10h00 Session/Séance 20 **MM Ondaajte Theatre**

SSC Gold Medal Address • Allocution du récipiendaire de la médaille d'or de la SSC
Special Session • Conférence spéciale
Organizer and Chair • Responsable et président: Jack KALBFLEISCH, University of Michigan

Muni SRIVASTAVA, University of Toronto
Multivariate analysis with fewer observations than the dimension • Analyse multivariée avec moins d'observations que la dimension des données

---

10:30 - 12:00 • 10h30-12h00 Session/Séance 21 **LSC 240**

Shape-Restricted Inference • Inférence avec restrictions sur la forme
Invited Paper Session • présentations sur invitation: **Institute of Mathematical Statistics • l'Institute of Mathematical Statistics**
Organizer and Chair • Responsable et président: Mary MEYER, University of Georgia

10:30 • 10h30 Michael WOODROOFE, University of Michigan; Mary MEYER, University of Georgia
Estimating a unimodal density • Estimation d'une densité unimodale

11:00 • 11h00 Richard DYKSTRA, University of Iowa; Chris CAROLAN, East Carolina University
Characterization of the least concave majorant of Brownian Motion with application to construction • La caractérisation du majorant le moins concave du mouvement brownien avec application à la construction

11:30 • 11h30 Mary MEYER, University of Georgia
Improving the power of tests with shape-restricted alternatives via projections onto subcones • Amélioration de la puissance des tests par des alternatives de restrictions sur la forme via des projections sur des sous-cônes

10:30 - 12:00 • 10h30-12h00 Session/Séance 22 **LSC 338**

Analysis of Mixed Discrete and Continuous Outcome Data • Analyse de mélanges de variables discrètes et continues
Invited Paper Session • présentations sur invitation: **Biostatistics Section • Groupe de biostatistique**
Organizer • Responsable: A. DE LEON, University of Calgary
Chair • Président: K.C. CARRIÈRE, University of Alberta

10:30 • 10h30 Ming-yi HU, Yamanouchi Pharma America, Inc., Thomas R. BELIN, University of California at Los Angeles
Evaluating imputation model choice in a study with incomplete continuous and categorical data and follow-up of initial nonrespondents • Évaluation du choix d'un modèle d'imputation dans une étude avec des données continues et catégorielles incomplètes et des données de suivi des non-répondants initiaux

11:00 • 11h00 A. de Leon, University of Calgary; K.C. Carrière, University of Alberta
General mixed-data model: Extension of general location and grouped continuous models • Modèle général de données mixtes: une extension du modèle général de localisation et du modèle groupé continu

11:30 • 11h30 Avner Bar-Hen, University of Aix-Marseille III; F. Mortier, CIRAD-Foret
Estimation of Mahalanobis distance with continuous and discrete variables • Estimation de la distance de Mahalanobis à l'aide de variables continues et discrètes

10:30 - 12:00 • 10h30-12h00 Session/Séance 23 **LSC 242**

Recent Developments in Small Area Estimation • Développements récents en estimation sur de petits domaines
Special Invited Session of the Survey Methods Section • Présentation sur invitation spéciale de la section sur les méthodes d'enquête
Organizer • Responsable: Jack Gambino, Statistics Canada/Statistique Canada
Chair • Président: David Binder, Statistics Canada/Statistique Canada

J.N.K. Rao, Carleton University
Some new developments in small area estimation • Nouveaux développements dans le domaine de l'estimation sur de petits domaines
Discussants: Donald J. Malec, United States Bureau of the Census and
Wayne A. Fuller, Iowa State University

10:30 - 12:00 • 10h30-12h00 Session/Séance 24 **LSC 238**

Special Session of the Pacific Institute for the Mathematical Sciences on Robustness • Session spéciale du Pacific Institute for the Mathematical Sciences en robustesse
Invited Paper Session • présentations sur invitation
Organizer and Chair • Responsable et président: Debbie Dupuis, University of Western Ontario

10:30 • 10h30 Elvezio Ronchetti, University of Geneva
Future directions in robust statistics • Développements futurs en statistique robuste

11:15 • 11h15 David Tyler, Rutgers University
Multivariate M-estimation: concepts and applications • M-estimation multivariée: concepts et applications

10:30 - 12:00 • 10h30-12h00 Session/Séance 25 **LSC 332**

Time Series Methods for Fitting Dynamical Models • Méthodes de séries chronologiques pour l'ajustement de modèles dynamiques
Invited Paper Session • présentations sur invitation
Organizer and Chair • Responsable et président: Bruce Smith, Dalhousie University

10:30 • 10h30 Keith Thompson, Kassiem Jacobs, Dalhousie University
The assimilation of observations into ocean models • L'assimilation des observations dans des modèles océaniques

11:00 • 11h00 Pierre Gauthier, Environment Canada; Mark Buehner, Stéphane Laroche, Monique Tanguay, Meteorological Service of Canada/Service Météorologique du Canada
Operational implementation of variational assimilation • Mise en oeuvre opérationnelle de l'assimilation variationnelle

11:30 • 11h30 Chris Jones, L. Kuznetsov, University of North Carolina at Chapel Hill; K. Ide, UCLA
An assimilation scheme for Lagrangian data • Un schème d'assimilation pour des données lagrangiennes

10:30 - 12:00 • 10h30-12h00 Session/Séance 26 **LSC 234**

Decision Theory and Bayesian Methods and Resampling • Théorie de la décision, méthodes bayésiennes et rééchantillonnage
Contributed Paper Session • présentations régulières
Chair • Président: Duncan Murdoch, University of Western Ontario

10:30 • 10h30 Sohee Kang, Michael Escobar, University of Toronto
Nonparametric Bayesian curve estimation for logistic regression • Estimation non paramétrique bayésienne de la courbe en régression logistique

10:45 • 10h45 Lin Xue, Liqun Wang, University of Manitoba
Bayesian finite mixture model with unknown components • Le modèle de mélanges finis bayésien avec composantes inconnues

11:00 • 11h00 Tao Tan
The minimax admissibility characterization of linear estimates • La caractérisation d'admissibilité minimax des estimations linéaires

11:15 • 11h15 Wenyu Jiang, University of Waterloo; J.D. Kalbfleisch, University of Michigan
Resampling methods for estimating functions with U-statistic structure • Méthodes de rééchantillonnage pour la fonction d'estimation avec la structure de la statistique U

11:30 • 11h30 Abderazzak Mouiha, Lycée Al Wahda, Taounate, Maroc
Bootstrapping a general statistic for dependent observations • Application du bootstrap à une statistique générale avec données dépendantes

11:45 • 11h45

1:30 - 3:00 • 13h30-15h00 Session/Séance 27 **LSC 240**

Statistical Issues in Modern Biology • Considérations statistiques en biologie moderne
Special Session • Conférence spéciale: **Caucus for Women in Statistics -Canadian Section- and Women in Statistics Committee (SSC) • Caucus pour les femmes en statistique -section canadienne- et comité sur les femmes en statistique**
Organizer and Chair • Responsable et président: Jeanette O'Hara Hines, University of Waterloo

1:30 • 13h30 Jenny Bryan, University of British Columbia
Gene classification and clustering with time course data • Classification des gènes et groupement avec des données de temps de course

2:00 • 14h00 Jinko Graham, Brad McNeney, Simon Fraser University; Francoise Seillier-Moiseiwitsch, Bioinformatics Research Centre, University of Maryland
Finding recombination breakpoints in HIV molecular sequences from an individual • Trouver les points de rupture de recombinaison dans une séquence moléculaire du VIH chez un individu

2:30 • 14h30 Julie Horrocks, University of Guelph
Joint models for longitudinal data and time-to-event data with multiple outcomes • Modèles conjoints pour des données longitudinales et de temps jusqu'à un événement avec plusieurs résultats possibles

1:30 - 3:00 • 13h30-15h00 Session/Séance 28 **LSC 242**

Resampling Methods • Méthodes de rééchantillonnage
Invited Paper Session • présentations sur invitation
Organizer and Chair • Responsable et président: N.G.N. Prasad, University of Alberta

1:30 • 13h30 Michael Sherman, Texas A&M University
Nonparametric resampling for spatial data • Rééchantillonnage non paramétrique pour des données spatiales

2:00 • 14h00 Subhash Lele, University of Alberta
Impact of bootstrap on estimating functions • L'effet du bootstrap sur les fonctions d'estimations

2:30 • 14h30 Angelo Canty, A. R. Padmanabhan, McMaster University
A robust bootstrap test for the equality of several medians • Un test bootstrap robuste pour l'égalité de plusieurs médianes

1:30 - 3:00 • 13h30-15h00 Session/Séance 29 **LSC 238**

Experimental Design • Plans d'expérience
Invited Paper Session • présentations sur invitation
Organizer and Chair • Responsable et président: Weng Kee Wong, University of California at Los Angeles

1:30 • 13h30 Holger Dette, S. Biedermann, V.B. Melas, Ruhr-Universitaet Bochum
Efficient designs for regression models in microbiology • Plans d'expérience efficaces pour les modèles de régression en microbiologie

2:00 • 14h00 Rainer Schwabe, Otto-von-Guericke-Universität
Efficient and adaptive design and analysis in forced choice experiments • Design efficace et adaptatif et l'analyse d'expériences avec choix forcés

2:30 • 14h30 Julie Zhou, University of Victoria; Hongtu Zhu, Yale University
Robust experimental designs for the random effects model • Plans d'expériences robustes pour le modèle à effets aléatoires

1:30 - 3:00 • 13h30-15h00 Session/Séance 30 **LSC 338**

Survey Methods Contributed Session III: Methods for Health Surveys • Méthodes d'enquête III: Méthodes pour les enquêtes sur la santé
Contributed Paper Session • présentations régulières
Organizer and Chair • Responsable et président: Colleen Clark, Statistics Canada/Statistique Canada

1:30 • 13h30 Larry MacNabb, Statistics Canada/Statistique Canada
Application of cluster analysis towards the development of health region peer groups • Application de l'analyse de regroupement vers le développement de la santé des groupes de pairs en région

1:45 • 13h45 Yves Béland, Johane Dufour, Larry MacNabb, Statistics Canada/Statistique Canada
Sample Design of the 2004 Canadian Nutrition Survey • Plan d'échantillonnage du sondage canadien sur la nutrition de 2004

2:00 • 14h00 Amanda Lafontaine, Lehana Thabane, Aaron Childs, McMaster University
Determining the level of statisticians' participation in Canadian based research ethics committees • Détermination du niveau de participation des statisticiens dans les comités d'éthiques de recherches canadiens

2:15 • 14h15 François Brisebois, Patrice Mathieu, Statistics Canada/Statistique Canada
Creation of a new longitudinal weight for the Canadian National Population Health Survey: providing data users with greater analytical flexibility • Création d'un nouveau poids longitudinal pour l'Enquête nationale sur la santé de la population canadienne: Une plus grande flexibilité analytique pour les utilisateurs de données

2:30 • 14h30 Dany Faucher, Éric Langlet, Statistics Canada/Statistique Canada
An application of the bootstrap variance estimation method to the Participation and Activity Limitation Survey • Une application du boostrap pour l'estimation de la variance dans le cadre de l'Enquête sur la Participation et les Limitations d'Activités

2:45 • 14h45

1:30 - 3:00 • 13h30-15h00 Session/Séance 31 **LSC 332**

Robust Methods I • Méthodes robustes I
Contributed Paper Session • présentations régulières
Chair • Président: Maureen Tingley, University of New Brunswick

1:30 • 13h30 Adeniyi Adewale, Douglas P. Wiens, University of Alberta
Robust designs for approximate regression models with two interacting regressors • Design robuste pour les modèles de régression approximatif avec deux régresseurs interagissant

1:45 • 13h45 Joanna Flemming, Elvezio Ronchetti, Eva Cantoni, University of Geneva
Model selection for marginal longitudinal generalized linear models • Sélection de modèle pour les modèles linéaires généralisés logitudinaux marginaux

2:00 • 14h00 Pierre Duchesne, HEC Montréal
Robust and powerful serial correlation tests with new robust estimates in ARX models • Tests de corrélation sérielle robustes basés sur de nouveaux estimateurs dans les modèles ARX

2:15 • 14h15 Sanjoy Sinha, University of Winnipeg
Robust inference in generalized linear mixed models •

2:30 • 14h30 Fatemah Alqallaf, Ruben Zamar, University of British Columbia
Scalable robust covariance and correlation estimates • Estimations robustes et calculables de la covariance et de la corrélation

2:45 • 14h45 Anthony Almudevar, Acadia University
On the exact form for the density of multivariate M-estimators • Sur la forme exacte pour la densité d'estimateurs-M multivariés

1:30 - 3:00 • 13h30-15h00 Session/Séance 32 **LSC 234**

Stochastic Processes and Finance and Applied Probability • Processus stochastique et finance et probabilités appliquées
Contributed Paper Session • présentations régulières
Chair • Président: Smiley Cheng, University of Manitoba

1:30 • 13h30 Eric Marchand, University of New Brunswick; Anatole Joffe, Francois Perron, Université de Montréal; Paul Popadiuk, Concordia University
On a particular sum of dependent Bernouilli and its relationship to a matching type problem • À propos d'une somme particulière de Bernouilli dépendantes et du problème des rencontres

1:45 • 13h45 Mahmoud Zarepour, Mohammad Taghi Jahandideh, University of Ottawa
Option pricing formula with infinite variance innovations • La formule de pricing des options avec des innovations de variance infinie

2:00 • 14h00 René Ferland, Université du Québec à Montréal; Simon Lalancette, HEC-Montréal
Forecasting of realized volatility and correlations: an empirical study • Prédiction de volatilités et corrélations réalisées: une étude empirique

2:15 • 14h15 Nickolai Khodusov, Novosibirsk State Technical University, Novosibirsk, Russia
New Modifications of NFV Criterion • Nouvelles modifications du critère NFV

2:30 • 14h30 Yung-Ming Chang, National University of Kaohsiung, Taiwan; James C. Fu, University of Manitoba
On later waiting time distributions in a sequence of Markov dependent multistate trials • Sur les distributions des derniers temps d'attente dans une séquence d'essai multi-états dépendants de Markov

2:45 • 14h45 Rachel MacKay, University of British Columbia
Hidden Markov models for multiple processes • Les chaînes de Markov cachées pour des processus multiples

---

3:30 - 5:00 • 15h30-17h00 Session/Séance 33 **LSC 240**

Comparative Research • Études comparatives
Invited Paper Session • présentations sur invitation: **Survey Methods Section • Groupe sur les méthodes d'enquête**
Organizer • Responsable: Dianne Lievesely, UNESCO Institute for Statistics
Chair • Président: Diane Stukel, UNESCO Institute for Statistics

3:30 • 15h30 Scott Murray, Statistics Canada/Statistique Canada
Quality control in an international comparative context: experience from the International Adult Literacy survey • Le contrôle de la qualité dans un contexte de comparaisons internationales: l'expérience de l'étude internationale sur l'analphabétisation

4:00 • 16h00 Albert Motivans, UNESCO Institute for Statistics/Institut de statistiques de l'UNESCO
Investing in education: towards a cross-national perspective • Investir dans l'éducation: vers une perspective multinationale

4:30 • 16h30 Tim HOLT, University of Southampton
Methodological issues in the development of statistical indicators and their use in international comparisons • Considérations méthodologiques dans le développement d'indices statistiques et leurs utilisations pour des comparaisons internationales

3:30 - 5:00 • 15h30-17h00 Session/Séance 34 **LSC 242**

Process Monitoring • Suivi de processus
Invited Paper Session • présentations sur invitation: **Business and Industrial Statistics Section• Groupe de statistique industrielle et de gestion**
Organizer and Chair • Responsable et président: Roman VIVEROS - AGUILERA, McMaster University

3:30 • 15h30 Fred SPIRING, Pollard Banknote Ltd & University of Manitobaa
Assessing process capability: a user's view • L'évaluation de la capacité d'un processus: la vision d'un utilisateur-chercheur

4:00 • 16h00 Bovas ABRAHAM, Asokan Mulayath VARIYATH, University of Waterloo
A look at the Mahalanobis-Taguchi system • Un regard sur le système de Mahalanobis-Taguchi

4:30 • 16h30 J. B. François BOUDREAU, Mike DUDZIC, Dofasco Inc.
On-line multivariate statistical monitoring at Dofasco Inc. • La surveillance statistique multivariable en ligne à Dofasco Inc.

3:30 - 5:00 • 15h30-17h00 Session/Séance 35 **LSC 238**

Special Session of the Fields Institute on Matrices and Statistics • Session spéciale de l'Institut Fields sur les matrices en statistique
Invited Paper Session • présentations sur invitation
Organizer and Chair • Responsable et président: George STYAN, McGill University

3:30 • 15h30 Jerzy K. BAKSALARY, University of Zielona Gora, Poland
Algebraic properties and statistical applications of rank-one- modified matrices • Les propriétés algébriques et les applications statistiques des matrices modifiées de rang un

4:00 • 16h00 Simo PUNTANEN, University of Tampere, Finland; George STYAN, McGill University
Matrix tricks for teaching linear statistical models – our personal Top Ten • Astuces matricielles pour enseigner les modèles statistiques linéaires: notre " top dix " personnel

4:30 • 16h30 Hans Joachim WERNER, University of Bonn, Germany
In the year of the matrix - prediction techniques in the general Gauss- Markov model • Dans l'année de la matrice — Techniques de prédiction dans le cadre du modèle de Gauss-Markov

3:30 - 5:00 • 15h30-17h00 Session/Séance 36 **LSC 338**

Statistics in Genomics and Proteomics • Statistique en génomique et protéomique
Invited Paper Session • présentations sur invitation
Organizer • Responsable: Jenny BRYAN, University of British Columbia
Chair • Président: Harry JOE, University of British Columbia

3:30 • 15h30 Jenny BRYAN, University of British Columbia
Resolving gene expression profiles with tag-based technologies • Résoudre les profiles d'expression des gènes avec des technologies Tag

4:00 • 16h00 Peter Hooper, University of Alberta
Statistical pattern recognition methods for protein secondary structure • Modèles statistiques de reconnaissance de forme pour la structure secondaire des protéines

4:30 • 16h30 Robert Gentleman, Harvard School of Public Health
Graphs and EDA in computational biology • Les graphes et l'analyse exploratoire (EDA) de données en biologie computationnelle

3:30 - 5:00 • 15h30-17h00 Session/Séance 37 **LSC 234**

Statistical Computing and Computationally Intensive Methods • Statistique informatique et méthodes informatiques intensives
Contributed Paper Session • présentations régulières
Chair • Président: Judy-Anne Chapman, University of Waterloo and University of Toronto

3:30 • 15h30 Peter Macdonald, Juan Du, McMaster University
An R package for finite mixture distributions • Un package en R pour des mélanges finis de distributions

3:45 • 15h45 Alain Desgagné, Jean-François Angers, Université de Montréal
Computational aspect of the generalized exponential power density • Considérations quantitatives de la famille de puissances d'exponentielle généralisée

4:00 • 16h00 Daniel Lemire, National Research Council of Canada
Constant time polynomial range sums for dynamic OLAP • Sommations polynomiales calculées en temps constant pour applications OLAP dynamiques

4:15 • 16h15 Genghui Wu, University of Manitoba
A random-discretization based Monte Carlo sampling method for numerical integration • Une méthode d'échantillonnage de Monte-Carlo basée sur une discrétisation aléatoire pour l'intégration numérique

4:30 • 16h30 Francois Perron, Yves Atchade, Université de Montréal
Monte Carlo simulations via control variates • Methodes de simulations utilisant des variables de contrôle

4:45 • 16h45 Caryn Thompson, University of New Brunswick (Saint John); Leah Passmore, Liliana Gonzalez, University of Rhode Island
Linear regression in the presence of spatially correlated errors: a computer intensive approach • Régression linéaire avec présence d'erreurs corrélées spatialement: une approche computationnelle intensive

3:30 - 5:00 • 15h30-17h00 Session/Séance 38 **LSC 332**

Statisticians in Action II • Statisticiens en action II
Video presentation • Présentation vidéo: **Committee on Professional Development • Comité sur le perfectionnement professionnel**
Chair • Président: Jon Baskerville

Wednesday June 11 • Mercredi 11 juin

8:30 - 10:00 • 8h30 - 10h00 Session/Séance 39 **LSC 238**

Statistical Methods for Health Services and Outcomes Research • Méthodes statistiques pour les services de santé et la recherche sur les outcomes
Invited Paper Session • présentations sur invitation: **Biostatistics Section • Groupe de biostatistique**
Organizer and Chair • Responsable et président: Wendy Lou, University of Toronto

8:30 • 8h30 Jamie Stafford, University of Toronto; John Braun, University of Western Ontario; Thierry Duchesne, Universite Laval
A kernel density estimate for interval censored data • Un estimateur par le noyau de la densité pour des données censurées par intervalles

9:00 • 9h00 K.K. Gordon Lan, Yuhwen Soo, Zhenming Shun, Aventis Pharmaceuticals, NJ
Two-stage winner design • Design à deux étapes gagnant

9:30 • 9h30 John Koval, University of Western Ontario
The intraclass Kappa • La statistique Kappa intra-classse

8:30 - 10:00 • 8h30 - 10h00 Session/Séance 40 **LSC 242**

Statistics and Climate Change • Statistique et changement climatique
Invited Paper Session • présentations sur invitation
Organizer and Chair • Responsable et président: Francis Zwiers, Canadian Centre for Climate Modelling and Analysis

8:30 • 8h30 Peter Stott, Gareth Jones, Hadley Centre for Climate Research; Myles Allen, Oxford University
Optimal detection of anthropogenic climate change • Détection optimale des changements climatiques anthropogéniques

9:00 • 9h00 Myles Allen, Hugo Lambert, Daithi Stone, Oxford University
Will it ever be possible to attribute apparently anomalous weather events to anthropogenic climate change? • Sera-t-il possible d'attribuer des événements météorologiques extraordinaires au réchauffement global?

9:30 • 9h30 Chris Forest, Andrei P. Sokolov, Peter H. Stone, Massachusetts Institute of Technology; Myles R. Allen, Oxford University
PDFs of climate system properties including natural and anthropogenic historical climate forcings • Fonctions de densités des propriétés des systèmes climatiques incluant des forçages climatiques historiques anthropogènes et naturels

8:30 - 10:00 • 8h30 - 10h00 Session/Séance 41 **LSC 338**

Bayesian Analysis • Analyse bayésienne
Invited Paper Session • présentations sur invitation
Organizer • Responsable: Michael Newton, University of Wisconsin
Chair • Président: Duncan Murdoch, University of Western Ontario

8:30 • 8h30 Michael Newton, Hyuna Yang, University of Wisconsin; David Hastie, Bristol University
Statistical methods to analyze genomic aberrations in cancer cells: the case of overlapping ensembles • Méthodes statistiques pour analyser les aberrations génomiques dans les cellules cancéreuses: le cas des ensembles qui se chevauchent

9:00 • 9h00 David Higdon, Los Alamos National Laboratory; Herbie Lee, University of California at Santa Cruz
Characterizing uncertainty in inverse problems • Caractérisation de l'incertitude pour des problèmes inverses

9:30 • 9h30 Jean-François Angers, Stéphane Courchesne, Louis-François Poirier, Université de Montréal; Claire Laberge-Nadeau, (CRT)
Link between cell-phone and car crashes • Lien entre le téléphone cellulaire et les accidents de la route

8:30 - 10:00 • 8h30 - 10h00 Session/Séance 42 **LSC 332**

Statistical Inference • Inférence statistique
Contributed Paper Session • présentations régulières
Chair • Président: Adeniyi Adewale, University of Alberta

8:30 • 8h30 Paul Marriott, National University of Singapore and Duke University
Mixture models and geometry • Modèles de mélange et géométrie

8:45 • 8h45 Thomas O'Gorman, Northern Illinois University
Adaptive statistical methods • Méthodes statistiques adaptatives

9:00 • 9h00 Jianan Peng, Acadia University; C.I.C. Lee, L. Liu, Memorial University of Newfoundland
Cone order monotonicity of tests for treatments versus a control • L'ordonnancement monotone dans un cône des tests de traitements versus un groupe contrôle

9:15 • 9h15 Yogendra Chaubey, Concordia University
Measures of overlap for inverse Gaussian populations • Mesures de chevauchement pour des populations de densité gaussienne inverse

9:30 • 9h30 Peiming Wang, Nanyang Technological University, Singapore
A score test for testing a bivariate zero-inflated Poisson regression model against bivariate zero-inflated negative binomial alternative • Un test de score pour tester un modèle de régression de Poisson zéro-augmenté bivarié contre une alternative binomiale négative zéro-augmenté bivariée

9:45 • 9h45 Regina Nuzzo, Jim Ramsay, McGill University
Functional data analysis of continuous judgments in music cognition • Analyse fonctionnelle de données des jugements continus en cognition musicale

8:30 - 10:00 • 8h30 - 10h00 Session/Séance 43 **LSC 234**

Survey Methods Contributed Session IV: Measuring the Quality of Survey Operations • Méthodes d'enquête IV: Mesure de la qualité des opérations d'enquête
Contributed Paper Session • présentations régulières
Organizer • Responsable: Don Royce, Statistics Canada/Statistique Canada
Chair • Président: Patricia Whitridge, Statistics Canada/Statistique Canada

8:30 • 8h30 Stuart Pursey, Statistics Canada/Statistique Canada
Use of the score function to optimize data collection resources in the Unified Enterprise Survey • L'utilisation de la fonction de caractérisation pour optimiser les ressources de la collecte des données dans l'Enquête unifiée auprès des entreprises

8:45 • 8h45 Robert Philips, Statistics Canada/Statistique Canada
The theory and applications of the score function for determining the priority of follow up in the Annual Survey of Manufactures • La théorie et les applications de la fonction de score pour déterminer la priorité de suivi pour le Sondage annuel des manufactures

9:00 • 9h00 Jennifer Ali, Statistics Canada/Statistique Canada
Quality monitoring of large surveys using the Blaise audit trail • Surveillance de la qualité de sondage à grande échelle en utilisant la vérification rétrospective de Blaise

9:15 • 9h15 Fred Hazelton, Stuart Pursey, Statistics Canada/Statistique Canada
The route to the final datapoint • Le chemin vers la résultat final

9:30 • 9h30 Colleen Clark, Mark Armstrong, Christian Thibualt, Statistics Canada/Statistique Canada
Measurement and innovation in the 2001Canadian Census coverage studies • Mesures et innovations dans les études de 2001 sur la couverture des recensements

9:45 • 9h45

8:30 - 9:15 Session/Séance 44 **LSC 240**

Pierre Robillard Award Winner Lecture • Allocution du lauréat du prix Pierre Robillard
Organizer and Chair • Responsable et président: Hugh Chipman, University of Waterloo

9:15 - 10:30 Session/Séance 45 **LSC 240**

Canadian Journal of Statistics Award Winner Lecture •Allocution du lauréat du prix de la Revue canadienne de statistique
Organizer and Chair • Responsable et président: Louis-Paul Rivest, Université Laval

10:30 - 12:30 • 10h30-12h30 Session/Séance 46 **LSC 240**

Case Study II - Neighbourhood Factors and Children: Hierarchical Linear Models and Small Area Statistics • Étude de cas II - Facteurs de voisinage et enfants: Modèles linéaires hiérarchiques et statistiques sur des petits domaines
Organizer and Chair • Responsable et président: Peggy Ng, York University

10:30 • 10h30 Introduction: Patricia Whitridge, Statistics Canada/Statistique Canada

10:35 • 10h35 Jean-Francois Plante, Lawrence McCandless, Mike Danilov, Mushfiqur Rahman, University of British Columbia

10:55 • 10h55 Sigfrido Iglesias-Gonzalez, Zheng Zheng, Xiaobin Yuan, University of Toronto

11:15 • 11h15 Ouyang Jangman, Jady Lui, Hanqiuzi Wen, Sevina Chung, York University

11:35 • 11h35 Xianlin Ma, Xu Wang, Longyang Wu, University of Waterloo

11:55 • 11h55 Vaneeta Grover, McMasterUniversity

12:15 • 12h15 Discussion

10:30 - 12:00 • 10h30-12h00 Session/Séance 47 **LSC 238**

Nonparametric Analysis in Natural Resources Surveys • Analyse non paramétrique pour les enquêtes sur les ressources naturelles
Invited Paper Session • présentations sur invitation: **Survey Methods Section • Groupe sur les méthodes d'enquête**
Organizer and Chair • Responsable et président: Patrick Farrell, Carleton University

10:30 • 10h30 Noel Cadigan, Department of Fisheries and Oceans; Jiahua Chen, University of Waterloo
Improved kernel regression methods for inference about the population mean from sample surveys, with application to fishery surveys • Méthode de régression par le noyau amélioré pour l'inférence sur la moyenne de la population à partir d'un échantillon et applications en échantillonnage halieutique

11:00 • 11h00 Jay Breidt, Colorado State University; Jean D. Opsomer, Iowa State University
Nonparametric model-assisted estimation for surveys of natural resources • Estimations non paramétriques assistées par modèles pour des sondages sur les ressources naturelles

11:30 • 11h30 Changbao Wu, Jiahua Chen, Mary E. Thompson, University of Waterloo
Estimation of fish abundance indices based on scientific research trawl surveys • Estimation de l'indice d'abondance de poissons basée sur des sondages de chalut de recherches scientifiques

10:30 - 12:00 • 10h30-12h00 Session/Séance 48 **LSC 242**

Statistical Inference II: Inference Problems with Missing Data or Measurement Errors • Inférence statistique: Problèmes d'inférence avec des données manquantes et des erreurs de mesure
Invited Paper Session • présentations sur invitation
Organizer and Chair • Responsable et président: Jerry Lawless, University of Waterloo

10:30 • 10h30 Don McLeish, University of Waterloo; C.A. Struthers, St Jeromes University and University of Waterloo
Estimation of regression parameters in missing data problems • L'estimation des paramètres de régression dans des problèmes avec données manquantes

11:00 • 11h00 Bruce Turnbull, Cornell University; Wenxin Jiang, Northwestern University
The indirect method for repeated events regression analysis with covariate measurement error • La méthode indirecte pour l'analyse de régression d'événements récurrents avec erreur de mesure sur les covariables

11:30 • 11h30 Paul Gustafson, University of British Columbia
Bayesian adjustment for mismeasured explanatory variables • L'ajustement bayésien pour des variables explicatives avec erreur de mesure

10:30 - 12:00 • 10h30-12h00 Session/Séance 49 **LSC 338**

Biostatistics Contributed Session II: Epidemiological and Clinical Studies • Biostatistique II: Études épidémiologiques et cliniques
Contributed Paper Session • présentations régulières
Chair • Président: Keyue Ding, Queen's University

10:30 • 10h30 Gordon Fick, University of Calgary
Modelling the odds of disease using data from case-control studies • Modéliser le risque de maladie en utilisant des données d'études cas-témoins

10:45 • 10h45 Cyr Emile M'lan, Hospital for Sick Children, Toronto
Bayesian sample size calculation for case-control studies • Méthodes bayésiennes de calcul de taille d'échantillon pour les études cas-témoins

11:00 • 11h00 Lehana Thabane, McMaster University
A Bayesian look at the number needed to treat • Une vue bayésienne pour le nombre requis pour traitement (NNT)

11:15 • 11h15 Nandini Dendukuri, J. Hanley, R. Platt, M.-H. Mayrand, McGill University
Design and data-analysis options for clinical trials of assisted reproductive technologies • Alternatives de devis et de méthodes d'analyse pour études cliniques de technologies de reproduction assistée

11:30 • 11h30 Nicholas Barrowman, Manchun Fang, Margaret Sampson, David Moher, Chalmers Research Group
When do meta-analyses need to be updated? • À quel moment les méta-analyses doivent-elles être mises à jour?

11:45 • 11h45 Juan Pablo Lewinger, Shelley B. Bull, University of Toronto
Better tests to find susceptibility genes for complex diseases via randomization • Meilleurs tests pour trouver les gènes de susceptibilité aux maladies complexes par randomisation

10:30 - 12:00 • 10h30-12h00 Session • Séance 50 **LSC 332**

Design and Analysis of Experiments• Planification et analyse d'expériences
Contributed Paper Session • présentations régulières
Chair • Président: Stephen Smith, Department of Fisheries and Oceans

10:30 • 10h30 Arden Miller, University of Auckland
The analysis of unreplicated factorial experiments using all possible comparisons • L'analyse d'expériences factorielles non-répliquées en utilisant toutes les comparaisons possibles

10:45 • 10h45 Glen Takahara, Hwashin H. Shin, Queen's University; Duncan J. Murdoch, University of Western Ontario
Optimal designs for orientation regression experiments • Designs optimaux pour des expériences de régression sur l'orientation

11:00 • 11h00 Xin Gao, Mayer Alvo, University of Ottawa
Nonparametric tests for interaction in unbalanced design with application in QTL analysis • Des tests non-paramétriques pour l'interaction dans le plan déséquilibré et applications aux analyses de QTL

11:15 • 11h15 Paul Cabilio, Acadia University; Mayer Alvo, University of Ottawa
General scores statistics on ranks in the analysis of unbalanced designs • Statistiques de scores générales basées sur les rangs pour l'analyse de plans déséquilibrés

11:30 • 11h30 Anatoly Naumov, Novosibirsk State Technical University, Novosibirsk, Russia
From optimal design to optimal control of Experiments • Du plan optimal au contrôle optimal d'expérience

11:45 • 11h45 Vitaly Senitch, Anatoly Naumov, Novosibirsk State Technical University, Novosibirsk, Russia
Sequential optimal control of experiments • Contrôle séquentiel optimal des expériences

10:30 - 12:00 • 10h30-12h00 Session/Séance 51 **LSC 234**

Robust Methods II and Statistical Education • Méthodes robustes II et éducation statistique
Contributed Paper Session • présentations régulières
Chair • Président: Sanjoy Sinha, University of Winnipeg

10:30 • 10h30 Ivan Mizera, University of Alberta; Benoit Laine, Université libre de Bruxelles
Autoregression depth • La profondeur autorégressive

10:45 • 10h45 Shoja'eddin Chenouri, University of Waterloo
Data depth and a multivariate robust nonparametric multisample test • Profondeur des données et un test multivarié non paramétrique robuste pour échantillons multiples

11:00 • 11h00 Howard Wainer, National Board of Medical Examiners; Eric Bradlow, University of Pennsylvania; Xiaohui Wang, University of North Carolina
Testlet response theory • Théorie des réponses testlet

11:15 • 11h15 B.M. Golam Kibria, Florida International University
The predictive distributions of regression and sum of squares and products matrices for the multivariate elliptically contoured distributions • Les distributions prédictives de la régression, des

sommes de carrés et des matrices produits pour les distributions de contours elliptiques multivariées

11:30 • 11h30 Adrian MacKenzie, Dalhousie University; Lehana Thabane, McMaster University; Joe Apaloo, St. Francis Xavier University
What motivates students to work hard? • Qu'est-ce qui motive les étudiants à travailler fort?

11:45 • 11h45 Ehsanes Saleh, Carleton University; P.K. Sen, University of North-Carolina, Chapel Hill
Robust estimation of slope parameter in a simple linear model with measurement errors • Estimation robuste de la pente dans un modèle linéaire simple avec des erreurs de mesure

---

1:30 - 3:00 • 13h30-15h00 Session/Séance 52 **LSC 238**

Stochastic Aspects of Forestry • Aspects stochastiques de la foresterie
Invited Paper Session • présentations sur invitation: **Canadian Operational Research Society • Société canadienne de recherche opérationnelle**
Organizer and Chair • Responsable et président: David Martell, University of Toronto

1:30 • 13h30 Eldon Gunn, Dalhousie University; Tim McGrath, N.S. Dept. of Natural Resources
Why doesn't thinning alter the diameter: a simple simulation model • Pourquoi l'amincissement n'affecte pas le diamètre: un modèle de simulation simple

2:00 • 14h00 John Braun, Marc Vincelli, University of Western Ontario
A northwestern Ontario forest fire weather simulator • Un simulateur climatique des feux de forêts du Nord-Ouest de l'Ontario

2:30 • 14h30 David Martell, B.M. Wotton, University of Toronto; K.A. Logan, Canadian Forest Service
The development and use of daily people-caused forest fire occurrence models in Ontario • Le développement et l'utilisation des modèles d'occurrences journalières pour les feux de forêts causés par erreurs humaines en Ontario

1:30 - 3:00 • 13h30-15h00 Session/Séance 53 **LSC 338**

Estimation of Fish Stock Mixtures • Estimation des mélanges de stocks de poissons
Invited Paper Session • présentations sur invitation: **Biostatistics Section • Groupe de biostatistique**
Organizer and Chair • Responsable et président: Noel Cadigan, Department of Fisheries and Oceans

1:30 • 13h30 John Candy, T. D. Beacham, Department of Fisheries and Oceans
To Bayes or not to Bayes: MLE vs Bayesian analysis for mixed stock fisheries using highly polymorphic DNA markers in Pacific Salmon • To Bayes or not to Bayes: l'analyse bayésienne vs les EMV pour la pêche à espèces variées utilisant des marqueurs très polymorphes d'ADN chez les saumons du Pacifique

2:00 • 14h00 Daniel Ruzzante, Dalhousie University; M.M. Hansen, K. Ebert, D. Meldrup, Danish Institute for Fisheries Research
Individual and population level approaches to the analysis of stocking impact in an anadromous brown trout (Salmo trutta) complex • Une approche au niveau de l'individu et de la population pour l'analyse de l'impact du peuplement dans un complexe de truite brune anadrome (trutta de Salmo)

2:30 • 14h30

1:30 - 3:00 • 13h30-15h00 Session/Séance 54 **LSC 240**

Applied Probability • Probabilité appliquée
Invited Paper Session • présentations sur invitation
Organizer and Chair • Responsable et président: Ernest Enns, University of Calgary

1:30 • 13h30 Gordon Willmot, University of Waterloo; David Dickson, University of Melbourne
The Gerber-Shiu discounted penalty function in the stationary renewal risk model • La fonction escomptée de pénalité de Gerber et Shiu dans le modèle de risque de renouvellement stationnaire

2:00 • 14h00 Reg Kulperger, University of Western Ontario; Zengjing Chen, University of Western Ontario and Shandong University, China
Stochastic prey predator system • Système proie-prédateur stochastique

2:30 • 14h30 Chris Small, University of Waterloo; Huiling Le, University of Nottingham
Modelling the shapes of random curves • Modélisation de la forme de courbes aléatoires

1:30 - 3:00 • 13h30-15h00 Session/Séance 55 **LSC 242**

Special Session of the Centre de Recherches Mathématiques on Statistics and Finance • Session speciale du Centre de Recherches Mathématiques en statistique et finance
Invited Paper Session • présentations sur invitation
Organizer and Chair • Responsable et président: Bruno Rémillard, École des hautes études commerciales, Montréal

1:30 • 13h30 Eric Renault, Université de Montréal
Dynamic factor models in finance • Modèles factoriels dynamiques en finance

2:00 • 14h00 Jin-Chuan Duan, Geneviève Gauthier, Jean-Guy Simonato, Sophia Zaanoun, University of Toronto
Estimating structural credit risk models with consideration of survivorship • Estimation du modèle structural du risque de crédit avec des considérations de survie

2:30 • 14h30 Francois Watier, Université de Sherbrooke; Jean Vaillancourt, Université du Québec en Outaouais
Multiperiod and continuous-time mean-variance analysis in portfolio management • Analyse moyenne-variance en gestion de portefeuille dans un contexte multipériodique et en temps continu

1:30 - 3:00 • 13h30-15h00 Session/Séance 56 **LSC 332**

Survey Methods Contributed Session V: Survey Sampling • Méthodes d'enquête V: sondages
Contributed Paper Session • présentations régulières
Organizer • Responsable: Hélène Bérard, Statistics Canada/Statistique Canada
Chair • Président: André Cyr, Statistics Canada/Statistique Canada

1:30 • 13h30 Lenka Mach, Ioana Schiopu-Kratina, Jean-Marc Fillion, Statistics Canada/Statistique Canada; Phil Reiss, Columbia University
Maximizing the overlap of two business surveys • Maximiser le chevauchement de deux enquêtes entreprises

1:45 • 13h45 Rebecca Morrison, Claude Julien, Suzelle Giroux, Statistics Canada/Statistique Canada
Redesign of the agriculture surveys • Remaniement des enquêtes agricoles

2:00 • 14h00 Wilson Lu, Randy R. Sitter, Simon Fraser University
Multi-way stratification by linear programming made practical • Rendre pratique la stratification à plusieurs étapes par la programmation linéaire

2:15 • 14h15 Owen Phillips, Statistics Canada/Statistique Canada; Avi Singh, Research Triangle Institute
Calibration allocation of sample for multiple characteristic surveys under stratified random sampling • Répartition d'échantillon par calage pour les enquêtes à plusieurs variables avec échantillonnage stratifié simple

2:30 • 14h30 Joseph Duggan, Elisabeth Neusy, Yves Bélanger, Statistics Canada/Statistique Canada
Sample design issues in a large-scale multi-frame national survey: the Canadian component of the International Adult Literacy and Life- skills survey (ALL) • Problèmes de design d'expérience dans un sondage national à étapes multiples à grande échelle: la composante canadienne du sondage international sur l'alphabétisation des adultes et sur les compétences de vie

2:45 • 14h45

1:30 - 3:00 • 13h30-15h00 Session/Séance 57 **LSC 234**

Distributions and Multivariate Methods • Distribution et méthodes multidimensionnelles
Contributed Paper Session • présentations régulières
Chair • Président: Alwell Oyet, Memorial University of Newfoundland

1:30 • 13h30 Louis Doray, Université de Montréal
Estimation for the discrete generalized Linnik distribution • Estimation pour la loi de Linnik généralisée discrète

1:45 • 13h45 Abdel El-Shaarawi, National Water Research Institute
Exact and approximate expressions for the tail of Student's t and F distributions • Expressions exactes et par approximations pour la queue des distributions t de Student et F de Fisher

2:00 • 14h00 Denis Larocque, Mélanie Labarre, HEC Montréal
A one-sided (positive orthant) conditionally distribution-free sign test for multivariate data •

Un test du signe conditionnellement "distribution-free" pour contre-hypothèses unilatérales avec données multidimensionnelles

2:15 • 14h15 Mouna Fallaha, Aleppo University
The asymptotic normality of the maximum pseudo-likelihood estimator of the parameters of Markov random fields • La normalité asymptotique des estimateurs du maximum de la pseudo-vraisemblance conditionnelle des paramètres de champs de Markov

2:30 • 14h30 Abdeljelil Farhat, Centre for Interuniversity Research and Analysis on Organizations; Jean-Marie Dufour, Université de Montréal
Exact k-sample goodness-of-fit tests for continuous and discrete distributions • Tests d'ajustement de K distributions continues ou discrètes

2:45 • 14h45

3:30 - 5:00 • 15h30-17h00 Session • Séance 58 **LSC 240**

Business and Economic Statistics • Statistique en affaires et en économie
Invited Paper Session • présentations sur invitation: **Business and Industrial Statistics Section • Groupe de statistique industrielle et de gestion**
Organizer and Chair • Responsable et président: Leonard MacLean, Dalhousie University

3:30 • 15h30 Talan Iscan, Dalhousie University; Fabio Ghironi, Boston College; Allessandro Rebucci, International Monetary Fund
Productivity shocks and consumption smoothing in the international economy • Les chocs de productivité et le lissage de la consommation dans l'économie internationale

4:00 • 16h00 Michael Foster, Canmac Economics; Leonard MacLean, Dalhousie University; William Ziemba, University of British Columbia
Empirical Bayes estimation with portfolio models • Estimation de Bayes empirique pour des modèles de portfolios

4:30 • 16h30 Horand Gassman, Dalhousie University; I. Deak, T. Szantai, Technical University of Budapest
Generating multivariate normal probabilities • Générer des probabilités multinormales

3:30 - 5:00 • 15h30-17h00 Session/Séance 59 **LSC 242**

Variable Selection • Sélection de variables
Invited Paper Session • présentations sur invitation
Organizer and Chair • Responsable et président: Hugh Chipman, University of Waterloo

3:30 • 15h30 Derek Bingham, University of Michigan
Bayesian screening designs • Design de discrimination bayésien

4:00 • 16h00 Mu Zhu, Hugh Chipman, University of Waterloo
Combinatorial optimization by parallel Darwinian evolution • L'optimisation combinatoire par l'évolution darwinienne parallèle

4:30 • 16h30 John Dziak, Richard Li, Pennsylvania State University
Characterization and New Algorithm for Nonconvex Penalized Least Squares • Caractérisation et nouvel algorithme pour les moindres carrés pénalisés non convexes

3:30 - 5:00 • 15h30-17h00 Session/Séance 60 **LSC 238**

Inference for Time Series and Other Models of Dependence • Inférence pour séries chronologiques et autres modèles de dépendance
Contributed Paper Session • présentations régulières
Chair • Président: Paul Cabilio, Acadia University

3:30 • 15h30 Gülhan Alpargu, University of Massachusetts; Pierre Dutilleul, McGill University
Efficient estimation and valid testing for stepwise linear regression with autocorrelated errors • Estimation efficace et test valide pour la régression linéaire pas à pas croissante avec erreurs autocorrélées

3:45 • 15h45 Pierre Dutilleul, Bernard Pelletier, McGill University; Gülhan Alpargu, University of Massachusetts
A simple modified F-test for multiple linear regression with autocorrelated random regressors and errors • Un simple test F modifié pour régression linéaire multiple avec régresseurs et erreurs aléatoires autocorrélés

4:00 • 16h00 Mostafa Filali, Jarrar Oulidi, fsdm-Fès-Morocco
Determining the order and the differentiation coefficient of an ARI using resampling method • Déterminations de l'ordre et du coefficient de différenciation d'un ARI en utilisant la méthode de rééchantillonnage

4:15 • 16h15 Florin Cristian Gheorghe, Panait Andreea Mihaela, Ghita Constantin, Valahia University of Targoviste
Reliable intervals in the case of the depended observations • Les intervalles de confiance dans le cas des observations dependantes

4:30 • 16h30 Anwer Sager, Garian University, Lybia
Theories in linear regression • Théories en régression linéaire

3:30 - 5:00 • 15h30-17h00 Session/Séance 61 **LSC 338**

Survey Methods Contributed Session VI: Estimation - Theoretical • Méthodes d'enquête VI: Estimation - théorie
Contributed Paper Session • présentations régulières
Organizer and Chair • Responsable et président: Pat Newcombe-Welch, Statistics Canada/Statistique Canada, University of Waterloo

3:30 • 15h30 Sarjinder Singh, St. Cloud State Universityy
On Farrell and Singh's penalized chi-square distance function in survey sampling • Sur la distance du khi-carré pénalisée de Farrell et Singh en sondage

3:45 • 15h45 Thierno Aliou Baldé, Norma Chhab-Alperin, Benoit Quenneville, Statistics Canada/Statistique Canada
A study on the predictive power of the Help Wanted Index • Étude sur le pouvoir de prévision de l'Indice d'Offre d'Emploi

4:00 • 16h00 Yong You, Jack Gambino, Statistics Canada/Statistique Canada
Hierarchical Bayes small area estimation with model determination and applications • Estimation de Bayes hiérarchique pour des petits domaines avec détermination du modèle et applications

4:15 • 16h15 Roberto Gismondi, Italian National Statistical Institute
Optimal provisional estimation in longitudinal surveys • Estimation optimale de provision dans les sondages longitudinaux

4:30 • 16h30 Murlidhar Jutti, Statistics Canada/Statistique Canada
A two phase sampling approach for variance and design effect estimation in studying brain-drain from Canada to the U.S. • Une approche d'échantillonnage à deux étapes pour l'estimation de la variance et de l'effet de design pour étudier l'exode des cerveaux du Canada vers les États-Unis

4:45 • 16h45

3:30 - 5:00 • 15h30-17h00 Session/Séance 62 **LSC 234**

Biostatistics Contributed Session III: Survival and Clustered Data • Biostatistique III: Données de survie et corrélées en grappes
Contributed Paper Session • présentations régulières
Chair • Président: Peter Macdonald, McMaster University

3:30 • 15h30 Arusharka Sen, Concordia University; Winfried Stute, Justus-Liebig University, Giessen, Germany
Efficient estimation under bivariate random censoring: independent components • Estimation efficace avec censure aléatoire bivariée: composantes indépendantes

3:45 • 15h45 Xuewen Lu, University of Calgary; R.S. Singh, University of Guelph
On a partially linear single-index survival model • Sur un modèle de survie à index simple partiellement linéaire

4:00 • 16h00 M. Tariqul Hasan, Brajendra C. Sutradhar, Gary Sneddon, Memorial University of Newfoundland
Analysing longitudinal failure time data: generalised estimating equations approach • Analyse de données longitudinales de temps de bris: approche basée sur les équations d'estimations généralisées

4:15 • 16h15 Renjun Ma, University of New Brunswick, Fredericton
A random effects modelling approach to clustered ordinal outcomes with random cluster sizes • Une approche de modélisation à effets aléatoires pour des résultats ordinaux avec des grappes de tailles aléatoires

4:30 • 16h30 Shenghai Zhang, Mary E. Thompson, University of Waterloo
Estimators of variances and confidence intervals from clustered data • Estimateurs de la variance et intervalles de confiance à partir de données en grappes

4:45 • 16h45 Guangyong Zou, Allan Donner, University of Western Ontario
The asymptotic variance of the intraclass correlation coefficient in the case of arbitrary class sizes • La variance asymptotique du coefficient de corrélation intra-groupe dans le cas où les groupes sont de taille arbitraire

3:30 - 5:00 • 15h30-17h00 Session/Séance 63 **LSC 332**

Statisticians in Action III • Statisticiens en action III

Video presentation • Présentation vidéo: **Committee on Professional Development • Comité sur le perfectionnement professionnel**

Chair • Président: Jon Baskerville

9 Abstracts • Resumés

Workshops/Ateliers

Sunday June 8 • Dimanche 8 juin **9:00 - 5:00 • 9h00-17h00**
Biostatistics Workshop • Atelier de biostatistique **Regency, LNH**

Edward. F. Vonesh, Baxter Healthcare Corporation

Mixed-effects models for longitudinal data • Modèles à effets mixtes pour données longitudinales

This workshop presents different methodologies associated with the analysis of generalized linear and nonlinear mixed-effects models for longitudinal data. Such fields as population pharmacokinetics (PK) and population pharmacodynamics (PD), bioassay, studies of biological or agricultural growth, and epidemiology all require fitting continuous and/or discrete data to generalized linear or nonlinear mixed models. In this workshop, we consider the use of both population-averaged (PA) and subject-specific (SS) models for such applications. A class of mixed-effects models is presented which one can use to jointly model continuous and/or discrete longitudinal data. Methods for estimating the parameters of interest, for evaluating model goodness-of-fit and for handling missing data will be discussed. These techniques will be illustrated using numerical examples from a variety of disciplines. Participants are expected to have a working knowledge of linear models and matrix algebra.

Cet atelier présente différentes méthodologies associées à l'analyse de modèles à effets mixtes linéaires et non linéaires généralisés pour données longitudinales. Dans les domaines tels que la pharmacocinétique de population et la pharmacodynamique de population, le biodosage, les études de croissance biologique ou agricole et l'épidémiologie, les données continues et/ou discrètes doivent être adaptées à des modèles mixtes linéaires ou non linéaires généralisés. Dans cet atelier, nous explorerons l'utilisation de modèles moyennés au niveau de la population (PA) et spécifiques au sujet (SS) pour de telles applications. Nous présenterons une classe de modèles à effets mixtes qui peuvent être utilisés pour modéliser des données longitudinales continues et/ou discrètes conjointement. Nous discuterons de méthodes permettant de juger quels paramètres sont utiles, d'évaluer la qualité d'ajustement du modèle et de gérer les données manquantes. Ces techniques seront illustrées à l'aide d'exemples numériques tirées de plusieurs disciplines. Les participants doivent avoir une connaissance pratique des modèles linéaires et de l'algèbre des matrices.

Sunday June 8 • Dimanche 8 juin **9:00 - 5:00 • 9h00-17h00**
Survey Methods Workshop • Atelier de méthodologie d'enquête **Imperial, LNH**

Pierre Lavallée, Statistics Canada/Statistique Canada

Panel surveys • Atelier sur les enquêtes longitudinales

This workshop will allow the participants to understand the basic concepts underlying longitudinal surveys, either for social or economic studies. The participants will briefly look at themes such as the advantages and disadvantages of panels, typical sampling designs, parameters related to the sampling design, longitudinal units, the use of registers and other sampling frames, sample selection, questionnaire design, data collection, non-response and estimation. With this workshop, the participants will understand the basic concepts needed to help them to design, conduct and analyse longitudinal surveys.

Cet atelier permettra aux participants de comprendre les concepts de base des enquêtes longitudinales que ce soit au niveau études sociales ou économiques. Les participants survoleront des thèmes tels que les avantages et désavantages des panels, les plans de sondage typiques, les paramètres reliés aux plans de sondage, les unités longitudinales, l'utilisation de répertoires et autres bases de sondage, la sélection de l'échantillon, conception des questionnaires, la collecte des données, la non-réponse et l'estimation. Avec cet atelier, les participants auront des notions leur permettant de concevoir, de conduire et d'analyser des enquêtes longitudinales.

Sunday June 8 • Dimanche 8 juin **9:00 - 5:00 • 9h00-17h00**

Business and Industrial Statistics Workshop • Atelier de statistique en affaires et dans l'industrie **Britannia, LNH**

Doug Montgomery, Arizona State University

Response Surface Methodology: Process and Product Optimization Using Designed Experiments • Méthodologie de surface de réponse: Optimisation des processus et des produits à l'aide d'expériences planifiées

Response surface methodology (RSM) is a combination of statistical experimental design, empirical modeling, and mathematical optimization techniques used for improving process and product performance. First developed in the chemical industry in the late 1940s, RSM has subsequently been applied in many industrial settings, including discrete parts manufacturing, and the electronics, semiconductor, and processing industries. The first part of this workshop is a comprehensive overview of the primary techniques of RSM, including the method of steepest ascent for moving a process to the vicinity of the optimum, second-order model fitting and analysis to determine appropriate operating conditions, and the basic experimental designs employed (such as the central composite design). The second part of the workshop is an overview of recent research and development in the field, including new information about experimental designs, multiple response optimization, and applications of RSM to robust product and process design. Examples of modern computer software for implementation of RSM will be included.

This course is based on: Myers, R. H. and Montgomery, D. C. (2002), Response Surface Methodology: Process and Product Optimization using Designed Experiments, 2nd edition, John Wiley & Sons, New York.

La méthodologie de surface de réponse (MSR) est une combinaison de plan d'expérience statistique, de modélisation empirique et de techniques d'optimisation mathématique employée pour améliorer la performance des processus et des produits. Développée dans l'industrie chimique à la fin des années 1940, la MSR s'est vu par la suite appliquer à de nombreuses situations industrielles, y compris dans la fabrication de composants discrets, l'industrie électronique, l'industrie des semi-conducteurs et l'industrie de transformation. La première partie de cet atelier sera une présentation globale des techniques de base de la MSR, dont la méthode de la plus grande ascension utiliser pour optimiser les processus, l'ajustement et l'analyse de modèles de deuxième ordre pour déterminer les conditions d'utilisation appropriées, ainsi que les principaux schémas expérimentaux employés (tels que le plan composite central). Dans la deuxième partie de cet atelier, nous présenterons l'évolution récente de la recherche dans le domaine, et notamment de nouvelles informations sur les plans d'expérience, l'optimisation des réponses multiples et les applications de

la MSR à la conception de produits et de processus robuste. Nous présenterons également des exemples de logiciels modernes permettant la mise en œuvre de la MSR.

Ce cours se fonde sur l'ouvrage suivant: Myers, R. H. et Montgomery, D. C. (2002), Response Surface Methodology: Process and Product Optimization using Designed Experiments, 2ème édition, John Wiley & Sons, New York.

Poster Session • Séance par affichage

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Eshetu Atenafu, Paul N. Corey, University of Toronto

Statistical properties of different ratio estimators of recommended daily food allowance estimates • Propriétés statistiques de différents estimateurs du rapport pour l'estimation de la quantité journalière d'aliments recommandée

An important problem in public health nutrition brought to our attention by Dr. Paul Pencharz of the Hospital for Sick Children in Toronto is the estimation of the recommended daily allowance (RDA) for the amino acid lysine. An interval estimate of the RDA is very important for public health application. The lower bound of the interval is useful in estimating the seriousness of inadequate nutrition. In our approach the RDA is a ratio of regression coefficients of a multiple regression model within which the RDA problem is embedded. This approach leads to exact confidence limits for the RDA and easily incorporates heterogeneity of subject response within a mixed model framework. Simulation and algebra are used to compare the statistical properties of our method to other well-known approaches to the problem. Bias, variance and confidence interval coverage are estimated and compared for the:
(a) Ratio of two correlated normal variables
(b) Application of Fieller's theorem
(c) Embedding the RDA in a multiple regression model
(d) The bootstrap method
The correct use of the bootstrap involved a two stage sampling process that reflected the within and between subject variance. Our analyses included an investigation of the inherent bias of the one stage bootstrap method.

Un problème important en nutrition de santé publique porté à notre attention par Dr. Paul Pencharz de l'Hospital for Sick Children à Toronto est l'évaluation de la dose journalière recommandée (DJR) de lysine d'acide aminé. Une estimation par intervalle de la DJR est très importante pour les applications en santé publique. La borne inférieure de l'intervalle est utile pour estimer la sévérité d'une nutrition inadéquate.

Dans notre approche, la DJR est un rapport des coefficients de régression d'un modèle de régression multiple dans lequel le problème de la DJR est inclus. Cette approche mène à des limites de confiances exactes pour la DJR et incorpore facilement l'hétérogénéité de la réponse des sujets dans un cadre de mélange de modèles. Nous utilisons des simulations et de l'algèbre pour comparer les propriétés statistiques de notre méthode à d'autres approches bien connues pour le problème. Le biais, la variance et les intervalles confiance sont estimés et comparés pour:

(a) Le rapport de deux variables normales corréléesm
(b) L'application du théorème de Fieller (c) mm
(c) Inclure la DJR dans un modèle de régression multiple
(d) La méthode du bootstrap
L'utilisation appropriée du bootstrap implique un processus d'échantillonnage à deux étapes qui reflètent la variance intra et inter sujets. Nos analyses incluent une recherche sur les biais inhérents de la méthode bootstrap à une étape.

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall
Hall

Jeffrey Bakal, J.T. Smith, G. Takahara, Queen's University

Issues in clustering biomechanical data • Problèmes dans le regroupement de données biomécaniques

Functional data arising from the study of biomechanical data presents an ideal application for representation by basis function coefficients following curve registration. The basis representation gives a faithful representation of the function and a high degree of dimension reduction for the analysis of smooth functions. The set of coefficients along with the warping curves can then be used for clustering functional motion data. We present results demonstrating the advantages over the use of non-functional methods, and how these advantages apply to the analysis of biomechanical data.

Les données fonctionnelles provenant de l'étude des données biomécaniques présentent une application idéale de représentation par des coefficients de fonction de base suite à l'identification de la courbe. La représentation de base donne une représentation fidèle de la fonction et réduit la dimension pour l'analyse des fonctions lisses. L'ensemble des coefficients et les courbes de déformations peuvent alors être utilisés pour grouper des données fonctionnelles de mouvement. Nous présentons des résultats démontrant les avantages par rapport à l'utilisation de méthodes non fonctionnelles, et comment ces avantages s'appliquent à l'analyse des données biomécaniques.

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

David Brillinger, University of California, Berkeley

Regression analysis and mutual information • Analyse de régression et information mutuelle

Some basic properties of the coefficient of mutual information (MI) are indicated as are some estimates. Then applications are made to data on wildfire occurrence and prediction indices, to the home field advantage affecting goals in soccer games and lastly to flow rates at dams along the Mississippi River. MI is studied as a replacement for the coefficient of determination in various practical situations.

Nous indiquons quelques propriétés de base du coefficient d'information mutuelle (IM) de même que quelques estimateurs. Ensuite, nous faisons des applications à des données sur l'occurrence feux de forêts et des indices de prédiction, sur l'avantage d'être à domicile pour les buts au soccer et pour des taux de débits aux barrages le long du fleuve Mississippi. L'IM est étudié comme remplacement du coefficient de détermination dans diverses situations pratiques.

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Noel Cadigan, Department of Fisheries and Oceans; Patrick Farrell, Carleton University

Local influence diagnostics for the retrospective problem in sequential population analyses of fisherydata • Diagnostique d'influence locale pour le problème rétrospectif dans l'analyse séquentielle de la population sur des données sur les pêches

The retrospective problem involves systematic differences in sequential population analysis (SPA) estimates of fish stock size in a reference year. The differences occur as successively more annual data are used for estimation. The differences appear to be structural biases that result from a mis-specification of the SPA. In some cases the retrospective problem is so severe that the SPA is considered to be too unreliable for stock assessment purposes. There are many possible causes of retrospective patterns, and it is difficult in practise to determine which causes are more likely. We utilize local influence diagnostics to find small changes or perturbations to SPA input components such as catches or natural mortality rates that remove or reduce retrospective patterns. The plausibility of the perturbations can be used to assess the likelihood that the component is the source of the retrospective pattern. We apply the methods to an example SPA that has a severe retrospective pattern. We show that potentially reasonable errors in some of the inputs could cause the retrospective pattern, but that other inputs are not a likely source of the pattern.

Le problème rétrospectif implique des différences systématiques dans les estimés d'analyse séquentielle de la population (ASP) de la taille des stocks halieutiques dans une année de référence. Les différences se produisent lorsque de plus en plus de données sont utilisés pour l'estimation. Les différences semblent être des biais structuraux qui résultent d'une spécification erronée de l'ASP. Dans certains cas, le problème rétrospectif est si grave que l'ASP est considérée comme trop incertaine pour l'évaluation des stocks. Il y a beaucoup de causes possibles des patrons rétrospectifs et il est difficile en pratique de déterminer quelles causes sont les plus probables. Nous utilisons un diagnostique d'influence local pour trouver des petits changements ou des perturbations aux variables de l'ASP tels que les prises ou les taux de mortalité naturelle, qui enlèvent ou réduisent les patrons rétrospectifs. La plausibilité des perturbations peut être utilisée pour évaluer la probabilité que la composante est la source du patron rétrospectif. Nous appliquons les méthodes à un exemple d'ASP qui a un patron rétrospectif important. Nous montrons que des erreurs raisonnables potentielles dans certaines des variables pourraient causer le patron rétrospectif, mais que d'autres variables ne sont pas des sources probables du patron.

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Judy-Anne Chapman, University of Waterloo & University of Toronto; Jiaming Sun, University of Toronto; Richard Gordon, Radhika Sivaramakrishna, University of Manitoba; Marilyn Link, Edward B. Fish, University of Toronto

Use of location-scale (log-normal) survival analysis to model survival from primary breast cancer after routine clinical use of mammography • L'utilisation de l'analyse de survie localisation-échelle (log-normale) pour modéliser la survie avec le cancer du sein primaire après avoir passer des mammographies de manière routinière.

Most clinical trials have indicated that the detection of smaller tumours with mammography leads to improved survival from breast cancer. However, there is a need to demonstrate

that the benefit extends to clinical practice, and to elucidate the extent to which reduction in population mortality is attributable to regular screening mammography, or to adjuvant systemic therapy. Mammography was used routinely at Women's College Hospital across a broad age range, in an era when most patients received no adjuvant therapy. We used a location-scale survival analysis to model breast cancer survival for a cohort of 678 stage I-III primary invasive breast cancer patients accrued from 1971 to 1990, and followed to 1996; 18 percent received adjuvant hormonal therapy and 15 percent received adjuvant chemotherapy. There were 61 women <40; 136, 40-49; 341, 50-69; 140, >69. Factors available for multivariate investigations were age (years), tumour size (cm), nodal status (N-,Nx,N+), Estrogen Receptor (ER; fmol/mg protein), Progesterone Receptor (PgR; fmol/mg protein), adjuvant hormonal therapy (no,yes), adjuvant chemotherapy (no,yes). Forward step-wise multivariate regression was used to examine the effects of these factors on disease-specific survival. Ten-year survival by tumour size was adjusted for the effects of other significant factors. For women <40 years of age, 10-year survival at the T1a, T1b, T1c, and T2 cut-points for tumour size is respectively 0.77, 0.74, 0.67, 0.44; for 40-49, it is 0.92, 0.90, 0.85, 0.62; for 50-69, it is 0.81, 0.79, 0.75, 0.62; for >69, it is 0.84, 0.81, 0.73, 0.44. With routine use of clinical mammography and up to 26 years of follow up, we found breast cancer survival to be significantly better ($p<0.05$) for all women with smaller tumours that survival indicated a change in natural history with early detection. The Canadin National Breast Screening Study (NBSS) controls had significantly smaller tumours ($p<0.001$) than our patients which may indicate substantive access to mammography outside of the NBSS that reduced the apparent survival benefit for clinical trial mammography.

La plupart des tests cliniques ont indiqué que la détection de petites tumeurs par la mammographie mène à une amélioration de la survie du cancer de sein. Cependant, il y a un besoin de démontrer que l'avantage se prolonge à la pratique clinique, et d'illustrer le fait que la réduction de la mortalité dans la population est attribuable au dépistage régulier par des mammographies, ou à la thérapie auxiliaire systémique. La mammographie a été utilisé de manière routinière au Women's College Hospital sur un grand intervalle d'âges, à une époque où la plupart les patients ne recevaient aucune thérapie auxiliaire. Nous utilisons une analyse de survie de localisation-échelle pour modéliser la survie du cancer du sein pour une cohorte de 678 patients atteint du cancer invasif du sein primaire d'étape I III amassés de 1971 à 1990, et suivi jusqu'en 1996; 18 pourcent ont reçu une thérapie hormonale auxiliaire et 15 pourcent ont reçu une chimiothérapie auxiliaire. Il y avait 61 femmes < 40 ans; 136 entre 40 et 49 ans; 341 entre 50 et 69 ans et 140 > 69 ans. Les facteurs disponibles pour des analyses multivariées sont l'âge (années), la taille de la tumeur (centimètre), le statut nodal (N, Nx, N+), le récepteur d'.strogène (HEU; protéine de fmol/mg), le récepteur de progestérone (PgR; protéine de fmol/mg), la thérapie hormonale auxiliaire (non, oui) et la chimiothérapie auxiliaire (non, oui). La régression multivariée foward pas-à-pas est utilisé pour examiner les effets de ces facteurs sur la survie spécifique de la maladie. La survie sur dix ans par rapport à la taille de la tumeur est ajustée aux effets d'autres facteurs significatifs. Pour les femmes de moins de 40 ans, la survie sur 10 ans aux points de coupure T1a, T1b, Tc, et T2 par rapport à la taille de la tumeur est respectivement 0,77, 0,74, 0,67, 0,44; pour les 40-49, elle est 0,92, 0,90, 0,85, 0,62; pour 50-69 ans, elle est 0,81, 0,79, 0,75, 0,62 et pour les plus de 69 ans, elle est 0,84, 0,81, 0,73, 0,44. Avec l'utilisation courante de la mammographie clinique et de jusqu'à 26 ans de suivi, nous avons trouvé que la survie au cancer du sein est significativement meilleure

($p<0.05$) pour toutes les femmes avec des plus petites tumeurs que la survie a indiqué un changement à l'histoire naturelle avec la détection préventive. Les contrôles du National Breast Screening Study (NBSS) canadien ont eu des tumeurs significativement plus petites ($p<0.001$) que nos patients qui peut indiquer un accès substantif à la mammographie en dehors du NBSS qui a réduit l'avantage apparent de la survie des mammographies des tests cliniques.

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Gemai Chen, University of Calgary; Jinhong You, The Hong Kong Polytechnic University

Moving block delete-1 jackknifing in partially linear regression models with m-dependent errors • Jackknife delete-1 à blocs mobiles dans les modèles de régression partiellement linéaire avec erreurs m-dépendantes.

Relaxing the usual i.i.d. assumption on the errors, we study a partially linear regression model with m-dependent errors. An asymptotic theory is developed, which includes the asymptotic normality of the least squares estimators of the linear component of the model, and a moving block delete-1 jackknife estimator of the asymptotic variance of the above estimators is proposed and shown to be consistent. As a result, asymptotically valid inference on this model can be performed.

En laissant tomber l'hypothèse que les erreurs sont iid, nous étudions un modèle de régression partiellement linéaire avec erreurs m-dépendantes. Une théorie asymptotique est développée, qui inclut la normalité asymptotique des estimateurs des moindres carrés de la composante linéaire du modèle. Un estimateur jackknife delete-1 à blocs mobiles de la variance asymptotique des estimateurs mentionnés ci-dessus est proposé et nous montrons qu'il est efficace. En conséquence, nous pouvons exécuter de l'inférence asymptotique sur ce modèle.

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Laura Cowen, Carl Schwarz, Simon Fraser University

Comparing survival estimates from a radio-tag mark-recapture study • Comparaison d'estimateurs de survie à partir d'une étude de marquage par reprise avec des étiquettes radio

Survival rates of animal populations are often estimated using mark-recapture techniques. Animals are marked with distinctive tags and released. Information for the estimation of survival and catchability is obtained through recaptures. Radio-tags are often preferred because of their high detectability rates. However, tags can go missing due to battery failure. Survival estimates are then conservative as those animals experiencing battery failure will not be recaptured. Adjustments to survival estimates can be made by using additional information about tag-life. These estimates have been compared with Kaplan-Meier estimates (Pollock et al, 2002). This method was applied using 100% and 80% detectability to see how the survival estimates would perform under varying detectability conditions.

Les taux de survie des populations animales sont souvent estimés en utilisant des techniques de marquage et de reprise. Les animaux sont marqués par des étiquettes personnalisées et ensuite libérés. L'information requise pour estimer la survie et la capacité de reprise est obtenue par la reprise des animaux.

Des étiquettes radio sont souvent préférées en raison de leurs taux élevés de détectabilité. Cependant, les étiquettes peuvent manquer dû aux batterie qui peuvent manquer de charge. Les estimations de survie sont alors conservatrices car ces animaux dont la batterie est tombée à plat ne sont pas repris.

Des ajustements aux estimations de survie peuvent être faits en utilisant de l'information additionnelle sur la vie des étiquettes. Ces estimations ont été comparées aux estimations de Kaplan-Meier (Pollock et autres, 2002). Cette méthode a été appliquée en utilisant 100% et 80% de détectabilité pour observer la performance des estimations de survie sous des conditions variables de détectabilité.

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Sandra Gardner, University of Toronto

Change point models for modeling discontinuation rates of Carinii Pneumonia Prophylaxis in an Ontario HIV patient population • Modèles de changement ponctuel pour modéliser les taux de discontinuité de la prophylaxie pour la pneumonie à Pneumocystis Carinii dans une population de patients atteint du VIH en Ontario

HIV/AIDS patients usually undertake complicated treatment regimens involving many drugs. In early 1999, there was published clinical evidence that patients could safely discontinue prophylaxis for Pneumocystis Carinii Pneumonia (PCP, an AIDS defining disease) if clinical markers indicated a sustained improvement in the immune system while undergoing treatment for HIV disease. An extract from the HIV Ontario Observational Database recently acquired from the Ontario HIV Treatment Network will be analyzed to determine if changes in guidelines for prescribing PCP prophylaxis have been adopted by physicians and patients in Ontario. The time when there is a notable increase in the discontinuation rate of PCP prophylaxis is referred to as a change point. Initial analyses using Poisson regression and Cox regression models with predetermined fixed change points will be presented. Competing parametric and non parametric models which estimate the change point from the data will also be presented. The latter models will also attempt to identify if subgroups of HIV/AIDS patients are associated with different change points.

Les patients atteints du VIH/SIDA suivent habituellement des régimes de traitement compliqués impliquant plusieurs médicaments. Au début de 1999, des évidences cliniques ont été publiées indiquant que les patients pourraient, sans risque, discontinuer la prophylaxie pour la pneumonie Pneumocystis Carinii (PCP défini de la maladie du SIDA) si les marqueurs cliniques indiquent une amélioration soutenue du système immunitaire tout en étant traité pour la maladie du VIH. Un extrait de la base de données d'observation du VIH d'Ontario récemment acquise du Ontario HIV Treatment Network est analysé pour déterminer si les changements dans les directives pour prescrire la prophylaxie de PCP ont été adoptés par les médecins et les patients en Ontario. Le moment où il y a une augmentation notable du taux de discontinuation de prophylaxie de PCP est défini comme un point de changement. Des analyses initiales utilisant des modèles de régressions de Poisson et de régression de Cox avec les points fixes prédéterminés de changement sont présentés. Des modèles concurrents paramétriques et non paramétriques qui estiment le point de changement dans les données sont également présentés. Les derniers modèles tentent également d'identifier si des sous-groupes de patients atteints du VIH/SIDA sont associés à différents points de changement.

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Paramjit Gill, Michael Treschow, Okanagan University College

A Stylometric Analysis of King Alfred's Literary Works • Analyse stylométrique du travail littéraire du Roi Alfred

For many centuries, Alfred the Great was judged to have translated several Latin texts into Old English: Gregory the Great's Pastoral Care, Boethius's Consolation of Philosophy, Bede's Ecclesiastical History of the English People, Augustine's Soliloquies, Orosius's History of the World, the First Fifty Prose Psalms. However, in the mid-twentieth century it was determined on the basis of dialect that Alfred could not have been the translator of Bede. Philological considerations soon also ruled Orosius out of Alfred's literary works. Even so, the other translations generally remain attributed to his hand. Some scholars have tugged at the thread of Alfred's scholarly reputation and sought to unravel it further. They have expressed doubt whether Alfred could have done such work. He was, after all, a busy king, taken up with several Viking invasions, extensive infrastructure programs, financial and legal reform, the building of a navy, and so on. On the basis of statistical stylometric analysis, we hope to bring more certainty to the question of authorship and shed light on the nature of authorship in this period.

Pendant plusieurs siècles, on a cru qu'Alfred le Grand avait traduit plusieurs textes latins en ancien anglais: Soins pastoraux de Gregory le Grand, Consolation philosophique de Boethius, l'Histoire ecclésiastique du peuple Anglais de Bede, Soliloques d'Augustine, l'Histoire du monde d'Orosius et les cinquante premiers psaumes en prose. Cependant, au milieu du vingtième siècle il a été déterminé sur la base du dialecte qu'Alfred ne pouvait pas avoir été le traducteur de Bede. Des considérations philosophiques ont également jugées le texte d'Orosius hors des travaux littéraires d'Alfred. Néanmoins, les autres traductions demeurent généralement attribuées à sa main. Quelques disciples s'en sont prit aux brèches de la réputation savante d'Alfred et ont cherché à résoudre l'énigme. Ils ont exprimé des doutes à savoir si Alfred pouvait avoir effectué un tel travail. Il était après tout un roi occupé, pris avec plusieurs invasions des Vikings, des programmes d'infrastructures importants, une réforme financière et légale, la construction d'une marine de guerre et ainsi de suite. Sur la base de l'analyse stylométrique statistique, nous espérons apporter plus de certitude à la question des droits d'auteur et jeter la lumière sur la nature des droits d'auteur de cette période.

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Cristina Goia, Ontario HIV Treatment Network; A.M. Bayoumi, St. Michael's Hospital, Toronto

Imputation of missing date values in medication records data •Imputation des dates manquantes pour des données des dossiers sur la médication

Objective: Although methods for dealing with missing continuous or categorical values have been elaborated, similar methods for missing date values are not well developed. Such methods may be particularly important for studies where several date values interact, such as the construction of antiretroviral drug regimens for human immunodeficiency virus (HIV) positive patients. We developed a set of imputation rules for estimating missing dates and evaluated the results.

Methods: We studied antiretroviral medication records from the HIV Ontario Observational Database, an observational longitudinal cohort. In each drug record the year, month,

and day of initiating and discontinuing medication were recorded separately and each could be missing. Each record was characterized by the pattern of missing date values. Potential correlates of missing values (drug class, study site, timing of therapy initiation, current medication or not) were examined by crosstabulation. In this way an empirical distribution function of the missing values was determined. We used this distribution on data with complete dates in order to drop some values and obtain a dataset with simulated missing values. The imputation rules were applied to these missing values; the rules imputed missing dates by taking the mid-point of the earliest possible start date and the latest possible stop date. Finally, we compared the imputed dates to the real dates. Differences in start and stop dates between the two datasets were analyzed.

Results: Of 16,954 antiretroviral records, 9,050 records had complete start and stop dates. The distribution of missing values was not related to the class of antiretroviral drug or study site, but missing values were more frequent when the start or stop year was before 1996. After synthesizing the dataset of missing values, 160 records (1.8 9 were dropped because the start or stop year was missing. The median difference between imputed and real start dates was 0 days (inter-quartile range [IQR] 0 to 0) and for stop dates was 0 days (IQR 0 to 0). The bivariate difference distribution resembles a bivariate normal distribution except for a few outliers.

Conclusions: Date imputation rules result in small deviations from real values in an observational dataset. Future work is needed to explore how imputation of missing dates influences assumptions regarding use of antiretroviral medication regimens.

Objectif: Bien que des méthodes pour traiter les valeurs manquantes continues ou catégoriques aient été élaborées, des méthodes semblables pour des dates manquantes ne sont pas bien développées. De telles méthodes peuvent être particulièrement importantes pour des études où plusieurs valeurs de date interagissent, comme la construction des régimes de médicament anti-rétroviral pour les patients atteint du virus d'immunodéficience humains (VIH). Nous avons développé un ensemble de règles d'imputation pour estimer les dates manquantes et nous avons évalué les résultats.

Méthodes: Nous avons étudié les dossiers de médication anti-rétroviral de la base de données d'observation du VIH de l'Ontario, une cohorte d'observations longitudinales. Dans chaque dossier médical, l'année, le mois et le jour du début et de la cessation de la médication sont enregistrés séparément et chacune de ces données peut être manquante. Chaque dossier est caractérisé par son patron de dates manquantes. Les variables potentiellement corrélées aux valeurs manquantes (classe du médicament, emplacement de l'étude, synchronisation du déclenchement de la thérapie, médicament courant ou non) sont examinées par des tableaux croisés. De cette manière, une fonction de distribution empirique des valeurs manquantes est déterminée. Nous utilisons cette distribution sur des données dont nous avons tous les dates pour ensuite éliminer quelques valeurs et ainsi obtenir un ensemble simulé de données avec des valeurs manquantes. Les règles d'imputation sont appliquées à ces valeurs manquantes; les règles imputent les dates manquantes en prenant le point médian de la première date de début et la dernière date d'arrêt. Finalement, nous comparons les dates imputées aux vraies dates. Les différences entre les dates de début et d'arrêt des deux jeux de données sont analysées.

Résultats: De 16 954 dossiers anti-rétroviral, 9050 dossiers ont des dates complètes de début et d'arrêt. La distribution des valeurs manquantes n'est pas reliée à la classe de médicament anti-rétroviral ou à l'emplacement de l'étude, mais elles sont plus fréquentes

lorsque l'année de début ou d'arrêt a lieu avant 1996. Après avoir synthétisé le jeu de données des valeurs manquantes, 160 dossiers sont mis à part parce que l'année de début ou d'arrêt est absente. La différence médiane entre les dates de début imputées et les vraies est de 0 jours (la gamme interquartile [IQR] 0 à 0) et de même pour les dates d'arrêt (différence interquartile 0 à 0). La distribution bivariée de la différence ressemble à une distribution normale bivariée à l'exception de quelques valeurs aberrantes.

Conclusions: Les règles d'imputation des dates résultent dans de petites déviations par rapport aux valeurs réelles dans un jeu de données d'observation. Des travaux futurs sont nécessaires pour explorer comment l'imputation des dates manquantes influence les hypothèses par rapport à l'utilisation des régimes de médicament anti-rétroviral.

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Cristina Goia, Ontario HIV Treatment Network; A.M. Bayoumi, St. Michael's Hospital, Toronto

Using a matched-cohort in order to validate the effectiveness of an intervention in HIV routine practice • Utiliser une cohorte appariée pour valider l'efficacité d'une intervention dans les pratiques courantes sur le VIH

Background: HIV-specific interventions are sometimes less effective in practice than in clinical trials. We examined how a matched-cohort study could be used to evaluate the usefulness of genotypic resistance testing (GRT) in a non-experimental setting. Our results may be important in assessing how the GRT is used in routine practice.

Methods: We used data from the HIV Ontario Observational Database, a voluntary longitudinal cohort, and from the Ontario's Public Health Laboratory. Tested participants were matched with not-tested participants based on viral load levels and antiretroviral medication both at the time of the testing and in the past. The two groups were compared at the time of the GRT using conditional logistic regression and ANOVA. We evaluated viral load outcomes 3 to 9 months after the GRT using conditional logistic regression and generalized estimating equations (GEE) with logit link function.

Results: We found 128 tested subjects matched to 184 controls. At baseline tested subjects and controls had similar viral load levels (3.95 vs. 3.70, p=0.37), antiretroviral medication scores (27.76 vs. 24.24, p=0.70), CD4 counts (330.61 vs. 368.41, p=0.56), prior use of protease inhibitor containing regimens (89.84Conclusions: Our results indicate that GRT may be less effective in practice than in clinical trials. But residual confounding, matching and sample size issues may have affected the efficiency of our study.

Introduction: Les interventions spécifiques au VIH sont parfois moins efficaces en pratique que dans les essais cliniques. Nous avons examiné comment une étude de cohorte appariée peut être utilisée pour évaluer l'utilité du test de résistance génotype (TRG) dans une situation non-expérimentale. Nos résultats peuvent être importants pour évaluer comment le TRG est utilisé en pratique courante.

Méthodes: Nous utilisons des données provenant de la base de données d'observation du VIH de l'Ontario, une cohorte d'observations longitudinales volontaires, et du laboratoire sur la santé publique de l'Ontario. Des participants examinés sont appariés aux participants non-examinés basés sur les niveaux de charge virale et la médicamentation anti-rétroviral au moment du test et dans le passé. Les deux groupes sont comparés à l'heure du TRG en utilisant la régression logistique conditionnelle et l'ANOVA. Nous évaluons les résultats

de charge virale durant 3 à 9 mois après le TRG en utilisant la régression logistique conditionnelle et les équations d'estimations généralisées (EEG) avec le logit comme fonction de lien.

Résultats: Nous avons examinés 128 sujets testés appariés à 184 témoins. À la base, les sujets testés et les témoins ont un niveau de charge virale semblable (3,95 contre 3,70, p=0.37), des scores de médicamentation anti-rétroviral (27,76 contre 24,24, p=0.70), des comptes de CD4 (330,61 contre 368,41, p=0.56), utilisation antérieure de régimes contenant des inhibiteurs de protéase (89.84ˆs. 82.61ˆBC p=0.08), les états définissant le SIDA (36.72ˆs. 29.89ˆBC p=0.27). 40.63% des sujets testés ont eu un changement dans leur régime de médicamentation après le TRG par opposition à 19.02% des sujets non-testés (p=0.0003). Même si ces différences ne sont pas statistiquement significatives, elles peuvent être importantes et ainsi nous avons exploré des ajustements pour chacune d'elles dans nos analyses finales. Les participants testés et non testés ont des résultats semblables (ajustés OR=0.76, p=0.39). La régression logistique conditionnelle et l'analyse des EEG donnent des résultats semblables, le EEG étant plus sensible. Finalement, nous avons limité notre échantillon de sujets testés seulement à ceux qui ont changé de régimes de médicamentation anti-rétroviral après le TRG, comme il est normalement prévu. Nous trouvons 52 sujets testés appariés à 78 témoins. Les analyses sont répétées et donnent des conclusions semblables.

Conclusions: Nos résultats indiquent que TRG peut être moins efficace en pratique que dans les essais cliniques. Mais les problèmes de confusion résiduelle, de l'appariement et de la taille de l'échantillon peuvent avoir affecté l'efficacité de notre étude.

Statistical Issues in Microarray Data Analysis of Sarcoma Tumors •Problèmes statistiques dans l'analyse de micro réseaux de données sur les tumeurs sarcoma

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall

Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Wenqing He, Shelley B. Bull, Nalan Gokgoz, Irene Andrulis, Jay Wunder, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, University of Toronto

A number of statistical issues arise in microarray data analysis, including questions about gene filtering, differentially expressed gene selection, multiple testing, and classification methods. We applied microarray techniques to measure gene expression in Malignant Fibrous Histiocytoma (MFH) sarcoma tumors. MFH is the most common type of soft tissue sarcoma but is poorly understood. There are few accurate predictors of outcome to guide treatment decisions. We conducted multiple two-class comparisons in the high-dimensional microarray data to select prognostic markers for MFH, and used those markers to predict the outcomes. The data set includes 35 MFH patients with clinical information and expression information for over 19,000 genes. The objectives of the study are: to describe the application of alternative methodologies for finding differentially expressed genes and predicting outcome, to identify profiles of genetic alternations that may be important during tumor progression, and to classify patient outcome using the selected markers and prediction rule. The performance of the predictors is investigated by cross-validation procedures, and tested by a new set of 12 MFH expression data. We found the more "honest" cross-validation classification error estimate to be substantially higher than the apparent error rate, but similar to that in the independent validation sample.

Un certain nombre de problèmes statistiques surgissent en analyse de données de micro réseaux, incluant des questions au sujet du filtrage de gènes, de la sélection de gènes

exprimés différentiellement, des tests multiples, et des méthodes de classification. Nous appliquons des techniques de micro réseaux pour mesurer l'expression des gènes dans les tumeurs fibreuses malignes de sarcome Histiocytome (MFH). Le MFH est le type le plus commun de tissu de sarcome mou cependant il est mal connu. Il y a peu de prédicteurs précis des résultats pour aider à guider les décisions au sujet du traitement. Nous conduisons des comparaisons multiples à deux-classe avec des données à grande dimension de micro réseaux pour choisir des marqueurs pronostiques pour le MFH, et nous utilisons ces marqueurs pour prédire les résultats. Le jeu de données comprend 35 patients atteints de MFH avec des informations cliniques et sur l'expression de plus de 19 000 gènes. Les objectifs de l'étude sont: de décrire l'application de méthodes alternatives pour trouver des gènes exprimés différentiellement et prévoir les résultats, d'identifier des profils d'alternances génétiques qui peuvent être importants durant la progression de la tumeur, et de classifier les résultats des patients en utilisant les marqueurs sélectionnés et les règles de prédiction. La performance des prédicteurs est étudiée par des procédures de validation croisée, et testé par un nouvel ensemble de données sur 12 expressions de MFH. Nous avons trouvé que l'estimation de l'erreur de classification par validation croisée la plus conservatrice est sensiblement plus élevée que le taux d'erreur apparent, mais semblable à celui dans l'échantillon de validation indépendant.

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Sohee Kang, John Hsieh, University of Toronto

Quality of life and survival analysis-an alternative approach to Q-TWiST • Analyse de survie et de qualité de vie: une approche alternative au Q-TWiST

In Quality of Life studies, the traditional methods, such as the Q-TWiST methods, for treatment comparisons are based on the linear combination of the expected sojourn times spent in the transient health states, derived as the areas under the survival curves. This approach tends to introduce biases which result in invalid comparisons when the follow-up periods among the comparison groups are of different lengths. Furthermore, the linear combination formulation lacks the ability to distinguish the magnitude of Quality of Life measure from that of the expected sojourn times. To overcome such difficulties, we propose a different approach for treatment comparisons based on the hazard rate and a newly defined second-order hazard rate. These hazard rates are more sensitive measures than the expected survival time, and unlike the latter, the former are point functions unaffected by the length of the follow-up period. We plot the second-and third order distribution functions respectively, against the first-order distribution function of the survival time to obtain two results: (1) the shape of the plotted curves and their positions relative to the diagonal line provide the changing patterns of the first and second order hazard rates from one health state to the next. (2) the area between the plotted curve and the diagonal line can be used to compare the treatments statistically, as it has an asymptotic Gaussian distribution, with test statistic calculated by using bootstrap variance. We applied this method on the dataset from an Eastern Cooperative Oncology group trial for advanced non-small-cell lung cancer patients. To account for censoring and/or covariates, it employs the stratified inter-arrival time multi-state Cox proportional hazard model to obtain survival functions for construction of the functions used in the plots and the test statistics.

Dans les études sur la qualité de vie, les méthodes traditionnelles, telles que les méthodes Q-TWiST, pour la comparaisons des traitements sont basées sur la combinaison linéaire

des temps prévus de séjour passés dans les états transitoires de santé, dérivés comme les aires sous les courbes de survie. Cette approche tend à introduire des biais qui ont comme conséquence des comparaisons erronées lorsque les périodes de suivi parmi les groupes de comparaison sont de différentes longueurs. De plus, le cadre de la combinaison linéaire restreint la capacité de distinguer la mesure d'importance de la qualité de vie de celle des temps prévus de séjour. Pour surmonter de telles difficultés, nous proposons une approche différente pour la comparaison de traitements, basée sur le taux de panne et un taux de panne de second ordre nouvellement défini. Ces taux de panne sont des mesures plus sensibles que le temps de survie prévu, et contrairement au dernier, les premiers sont des fonctions ponctuelles invariantes par la longueur de la période de suivi. Nous traçons les fonctions de distribution de deuxième et troisième ordre par rapport à la fonction de distribution de premier ordre du temps de survie respectivement afin d'obtenir deux résultats: (1) la forme des courbes tracées et de leurs positions par rapport à la ligne diagonale fournissent les caractères changeants des taux de panne de premier et second ordre pour un état de santé au prochain. (2) l'aire entre les courbes tracées et la ligne diagonale peut être utilisé pour comparer statistiquement les traitements, puisqu'elle a une distribution gaussienne asymptotique, et la statistique du test est calculée en utilisant la variance calculée par bootstrap. Nous appliquons cette méthode sur les données d'une épreuve de groupe de la Coopérative de l'Est des oncologues pour des patients avancés du cancer de poumon à cellules non petites. Pour tenir compte de la censure et/ou les covariables, nous utilisons le modèle proportionnel de taux de panne de Cox multi-états des temps stratifiés inter-arrivés pour obtenir des fonctions de survie pour la construction des fonctions utilisées dans les graphes et les statistiques du test.

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Melanie Laframboise, J. Gough, D.A. Marshall, B. Jaszewski, Innovus Research Inc.

Methods used to determine the impact of administrative restrictions for antibiotic use and expenditures • Méthodes utilisées pour déterminer l'impact des restrictions administratives sur l'utilisation et les dépenses en médicaments antibiotiques

In March 2001, the Ontario Drug Benefit program implemented administrative restrictions for the reimbursement of most fluoroquinolones, a class of antibiotics. The policy was intended to address recent increasing trends in antibiotic resistance and expenditures. The purpose of this study was to determine if overall antibiotic use and expenditures were reduced in the period after the introduction of the restrictive reimbursement policy. We examined weekly costs and weekly number of prescriptions for 29 individual antibiotics, including 6 fluoroquinolones. These series of data were found to be serially correlated and therefore, ARIMA modeling was chosen as a basis for the analysis. The data demonstrated linear and seasonal trends. Seasonality was controlled in the model by adding sine and cosine variables and adding a linear time variable controlled for temporal trends. An indicator variable accounted for whether or not the administrative restriction had been implemented. The 6 fluoroquinolones were examined as a group for the use and expenditure data. Five fluoroquinolones were restricted as a result of the administrative policy change and one was not restricted. The group of 6 fluoroquinolones resulted in a differenced model for both use and expenditure data with all inputs being significant. Models were chosen based on the minimum Akaike's information criterion (AIC). It was considered important to examine possible shifts prior to the policy change that might distort the results of any

change in use or expenditures at March 2001 already tested in the model. A "Breakpoint Analysis" was performed. This analysis comprised of sequential testing of indicator variables for every 5-week time period. During this testing, if any variables were found significant after a Bonferroni adjustment, they were added as input variables in the model and tested for significance against the March 2001 variable. The group of fluoroquinolones did not demonstrate any significant "Breakpoint" variables in either dataset, therefore the March 2001 variable was kept in the model.

The final conclusions drawn from this study were that the restrictive reimbursement policy at March 2001 was associated with a statistically significant decrease in both use and expenditures for the group of 6 fluoroquinolones. However, no statistically significant change was observed for total use or expenditures on all antibiotics. Thus, there was an apparent offsetting increase in the use of other antibiotics.

This project was sponsored by Bayer Inc.

En mars 2001, le Programme de Médicaments de l'Ontario a mis en place des restrictions administratives affectant le remboursement de la plupart des fluoroquinolones, une classe d'antibiotiques. Cette mesure visait à faire face à la résistance croissante aux antibiotiques et à l'augmentation des dépenses associées. L'objectif de cette étude était de déterminer si dans l'ensemble, l'utilisation et les dépenses d'antibiotiques ont diminué dans la période qui a suivi l'introduction de la mesure de restriction des remboursements. Nous avons observé les coûts et les nombres de prescriptions par semaine pour 29 antibiotiques individuellement, dont 6 fluoroquinolones. Ayant établi que ces séries de données étaient autocorrélées, nous avons opté pour une analyse basée sur le modèle ARIMA. Nous avons constaté la présence de tendances linéaires et saisonnières. La saisonnalité a été prise en compte dans le modèle en introduisant les fonctions sinus et cosinus et une fonction linéaire du temps représentait les tendances temporelles. Une variable indicatrice identifiait les observations postérieures à la limitation des remboursements. Les 6 fluoroquinolones ont été groupées pour l'analyse des données relatives aux volumes utilisés et aux dépenses d'antibiotiques. Cinq fluoroquinolones étaient soumises à la restriction des remboursements et une ne l'était pas. Le processus décrivant le groupe des 6 fluoroquinolones était intégré d'ordre 1, pour les deux séries de données, volumes et dépenses, et tous les facteurs introduits étaient significatifs. La sélection des modèles était basée sur le minimum du critère d'information d 'Akaiké (AIC). Il était important de rechercher d'éventuels changements de tendance antérieurs à la nouvelle mesure qui pourraient influencer les résultats du test de l'existence d'un impact sur le volume ou les dépenses en mars 2001. Une " analyse de ruptures " a été réalisée. Cette analyse a consisté en un test séquentiel de variables indicatrices pour tous les intervalles de 5 semaines. Si le seuil du test, déterminé selon la méthode de Bonferroni, était dépassé pour une variable, celle-ci était introduite comme exogène dans le modèle et testée face à la variable " mars 2001 ". Aucune " rupture " significative n'a été détectée dans les deux séries de données relatives au groupe des fluoroquinolones, en conséquence la variable " mars 2001 " a été gardée dans le modèle.

Cette étude démontre que la mesure de limitation des remboursements introduite en mars 2001 était associée à une diminution statistiquement significative de l'utilisation et des dépenses pour le groupe des 6 fluoroquinolones. Cependant, aucun changement significatif de l'utilisation ou des dépenses totales d'antibiotiques n'a été constaté. Ainsi l'utilization d'antibiotiques s'est apparament décalé sur d'autre antibiotiques, aux coûts similaries.

Ce projet était financé par Bayer Inc.

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Yi Liu, Alwell J. Oyet, Memorial University of Newfoundland

Minimax designs for discrimination between competing wavelet regression models • Le design minimax pour la discrimination entre des modèles concurrents de régression par ondelettes

A problem that commonly arises when using wavelets to represent the mean response function in nonparametric regression models is that of determining the number of wavelet terms that should be included in the wavelet representation. One possible solution to this problem is to choose designs that will maximize, in some sense, the difference between the better model and the other models. We adopt this approach in constructing exact minimax designs for discrimination between competing wavelet regression models. Sequential and nonsequential designs will be discussed along with some examples.

Un problème qui surgit fréquemment lorsqu'on utilise des ondelettes pour représenter la fonction de réponse moyenne dans les modèles non paramétriques de régression est celui de déterminer le nombre d'ondelettes qui devraient être incluses dans la représentation d'ondelettes. Une solution possible à ce problème est de choisir les designs qui maximisent, dans un certain sens, la différence entre le meilleur modèle et les autres. Nous adoptons cette approche pour construire des designs minimax exacts pour la discrimination entre les modèles concurrents de régression par ondelettes. Des designs séquentiels et non séquentiels sont discutées avec quelques exemples.

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Ahmed Lmoudden, J. Vaillancourt, K. Ghoudi, Université de Sherbrooke

Partial convergence of Kendall's process to the Brownian Bridge: test of independence • La convergence partielle du processus de Kendall vers le pont Brownien: test d'indépendance

Genest, Goudi and Rémillard showed that, if $Z_1, ..., Z_n$ is a random sample of size $n > 1$ from a d-variate continous distributions function H, and if $V_{i,n}$ stands for the proportion of observation Z_j, $j \leq i$, such that $Z_j \leq Z_i$ componentwise, then the limiting behavior of the empirical distribution function K_n may be derived from the (dependent) pseudo-observations $V_{i,n}$. This random quantity is a natural non-parametric estimator of K, the distibution function of the random variable $V = H(Z)$, whose expectation is an affine transformation of the population version of Kendall's tau in the case $d = 2$. Since the sample version of au is related in the same way to the mean of K_n, Genest and Rivest (1993, J. Amer. Statist. Assoc.) suggested that $\sqrt{n(K_n(t) - K(t))}$ be referred to as Kendall's process. Weak regularity conditions on K and H are found under which this centred process is asymptotically Gaussian, and an explicit expression for its limiting covariance function is given. The purpose of this paper is to examine, for the small sample, the limiting behavior of the Kendall's process $\sqrt{k_n(K_{k_n}(t) - K(t))}$under the same conditions on k_n.

Genest, Goudi et Rémillard ont montré, si on a $Z_1, ..., Z_n$ un échantillon de $n > 1$ vecteurs aléatoires composés de d-variables aléatoires et de fonction de distribution H continue, alors le processus de Kendall $\alpha_n = \sqrt{n(K_n(t) - K(t))}$, où $K_n(t)$ est la distribution empirique de la pseudo-observation $V_{n,i}$($V_{n,i}$ représente le nombre d'observations Z_j, $j \leq i$ telles que $Z_j \leq Z_i$) et $K(t)$ est la distribution de $V = H(Z)$, est asymptotiquement

Gaussien; ils ont aussi donné une expression explicite pour la fonction de covariance de la limite. Donc, nous allons montrer que si on prend un échantillon de taille k_n telle que k_n vérifie certaines conditions, alors notre processus $\alpha_{k_n} = \sqrt{k_n(K_{k_n}(t) - K(t))}$ converge faiblement vers le pont Brownien dont on connait explicitement sa fonction de covariance, ce que nous permet de faire le test d'hypothèse sur l'independance des composantes en basant sur la statistique de Cramer-von Miss.

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Cyr Emile M'LAN, Hospital for Sick Children, Toronto

Bayesian sample size calculation for case-control studies • Méthodes bayésiennes de calcul de taille d'échantillon pour les études cas-témoins

One of the most important statistical issues at the planning stage of a case-control study is the choice of sample size. For example, one might wish to select a sample size that ensures sufficient accuracy in estimating the odds ratio. Sample size determination for the odds ratio has been investigated from a frequentist viewpoint. While most of the proposed methods have been based on power, it is well known that high power does not necessarily guarantee accurate estimation of important parameters. Therefore, if one chooses to analyse a study by interval estimation rather than p-values, the design of the study should reflect this choice. Two previous frequentist approaches (O'Neill, 1984, Kupper and Halfner, 1990) were based on ensuring sufficiently small confidence interval widths for a given coverage. Recently, however, numerous papers have addressed the advantage of using a Bayesian approach to the estimation of an odds ratio from case-control studies based on the posterior distribution of the odds ratio by using hpd intervals (Hora and Kelly; 1983, Marshall 1988; Zelen and Parker, 1986; Hashemi et al. ,1997). In addition, there is an increasing literature on the advantages of Bayesian sample size determination, which better incorporates prior information into the calculations, and more fully accounts for the uncertainty of the eventual data which will be collected. In this talk, we show how sample size determination for estimating the odds ratio can be addressed within the Bayesian paradigm. Although we have thoroughly investigated a large number of Bayesian sample size criteria, including the development of novel criteria, here we focus on the average length criterion (ALC). Basically, this method proposes that the sample size be selected that guarantees a pre-specified length for a marginal posterior credible interval of predetermined coverage, averaged over the predictive distribution of the data. The solution, while easy to define, is technically challenging to carry out in practice. We discuss three different methods for finding the optimal sample size, including exact, approximate, and Monte Carlo. We compare the sample sizes derived from this criterion to those from frequentist power and confidence interval methods.

Le choix de la taille d'échantillon est l'un des points les plus sensibles de la planification d'une étude cas-témoins; de ce choix dépend notamment la précision avec laquelle on pourra estimer le rapport de cotes. Deux approches classiques antérieures (O'Neill 1984; Satten & Kupper 1990) ont été basées sur des critères garantissant des intervalles de confiance suffisamment courts pour une couverture donnée. Cependant, de nombreux auteurs ont récemment fait valoir les avantages d'une approche bayésienne, qui permet a la fois d'incorporer une information à priori et de mieux prendre en compte l'incertitude concernant la variation inhérente aux données dans l'estimation du rapport de cotes dans les études de cas-témoins. Le conférentier présentera un survol de ces travaux et fera état de

ses propres recherches dans le domaine. Il s'attardera plus particulièrement aux résultants portant sur le critère de la longueur moyenne, qui consiste à choisir une taille d'échantillon permettant de garantir une longueur moyenne préspecifiée pour un intervalle de crédibilité à posteriori de couverture préfixée, la moyenne étant faite par rapport a la distribution de prévision des données. Comme il le fera valoir, la solution est facile à concevoir mais ne peut être déterminée en pratique qu'au moyen de méthodes numériques exactes, approximatives ou de Monte-Carlo. Il présentera en outre une étude visant à comparer les tailles d'échantillon obtenues par ce critère à celle déduites des méthodes de puissance et d'intervalle de confiance dans l'approche classique.

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Sandra Olfert, P. Pahwa, J.A. Dosman, University of Saskatchewan

Longitudinal analysis of pulmonary dysfunction in the initial years of employment in the grain industry • Analyse longitudinale des problèmes pulmonaires pour les premières années d'emploi dans l'industrie du grain

A longitudinal study of 313 newly hired male grain industry workers was conducted between 1980 and 1985. The objective was to determine the effects of employment in the grain industry on pulmonary function and respiratory symptoms; and to determine if study completers (subjects who were measured at each of 4 time points) experience similar changes in pulmonary function and respiratory symptoms as study drop outs (subjects who were measured at three or fewer time points). Pre-employment physical examination, pulmonary function tests and allergy skinprick tests were conducted on the grain industry workers at the Division of Respiratory Medicine, Department of Medicine, Royal University Hospital, University of Saskatchewan. Using Hedeker's approach for handling missing data, study completers were compared with study drop outs in both grain industry workers and controls. SPSS was used to conduct independent samples t-tests and chi-squared analyses to examine demographic characteristics, pulmonary function test values, respiratory symptoms and other irritations. The number of non-smokers and ever-smokers in the control group was statistically significant. Of the controls, 89.5% of the non-smokers were available for complete follow-up, while 66.7% of non-smokers dropped out. Of the control subjects who were current smokers, 10.5% were followed-up for the complete study, while 33.3% dropped out. There were no other differences between study completers and study drop outs among control subjects; and no differences were found between study completers and study drop out subjects among grain industry workers. The maximum likelihood approach will be used to fit random effects models, and generalized estimating equations will be used to fit transitional models, to determine the effects of grain dust exposure on pulmonary function and respiratory symptoms. Results from these models, using subjects exposed to grain dust in the initial years of employment in the grain industry, will be compared with results from Labour Canada's Grain Dust Medical Surveillance Program, which used subjects with long term employment in the grain industry.

Une étude longitudinale sur 313 ouvriers de sexe masculin nouvellement employés dans l'industrie du grain a été entreprise entre 1980 et 1985. L'objectif était de déterminer les effets de travailler dans l'industrie du grain sur les fonctions pulmonaire et les symptômes respiratoires; et pour déterminer si les sujets qui ont complétés l'étude (sujets qui ont été mesurés à chacun de 4 temps) éprouvent des changements semblables au niveau des fonctions pulmonaires et des symptômes respiratoires que les sujets qui ont sorti de l'étude (les

sujets qui ont été mesurés à trois temps ou moins). Un examen physique de préemploi, des tests de fonctions pulmonaires et des tests d'allergie de la peau ont été effectués sur les ouvriers de l'industrie du grain à la Division de la médecine respiratoire, Département de médecine, Royal University Hospital, université de Saskatchewan. En utilisant l'approche de Hedeker pour les données manquantes, les sujets qui ont complétés l'étude ont été comparés à ceux qui ont sortie de l'étude pour les ouvriers de l'industrie du grain et un groupe contrôle. Le progiciel SPSS a été utilisé pour conduire les test-t et les tests du khi-carrée pour des échantillons indépendants pour examiner les caractéristiques démographiques, les valeurs aux tests de fonctions pulmonaires, les symptômes respiratoires et d'autres irritations. Le nombre de non-fumeurs et de jamais-fumeurs dans le groupe contrôle était statistiquement significatif. Du groupe contrôle, 89.5% des non-fumeurs étaient disponibles pour un suivi complet, alors que 66.7% des non-fumeurs ont quitté. Des sujets du groupe témoins qui étaient des fumeurs courants, 10.5% ont été suivi pour l'étude au complète, alors que 33.3% ont quitté. Il n'y avait pas d'autre différences entre les sujets qui ont complété l'étude et ceux qui ont sortie parmi les sujets témoins; et aucune différence n'a été trouvée entre les sujets qui ont complété l'étude et ceux qui ont quitté parmi les ouvriers de l'industrie du grain. L'approche du maximum de vraisemblance est employée pour ajuster les modèles à effets aléatoires, et les équations estimatrices généralisées sont utilisées pour ajuster les modèles transitoires, pour déterminer les effets de l'exposition à la poussière de grain sur la fonction pulmonaire et les symptômes respiratoires. Les résultats de ces modèles en utilisant des sujets exposés à la poussière de grain pour les premières années dans l'industrie du grain, sont comparés aux résultats du programme médical de surveillance de la poussière de grain de travail Canada, qui a utilisé des sujets qui travaillent à long terme dans l'industrie du grain.

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Jennifer Prokop, W.J. Braun, University of Western Ontario; V. Rousson, University of Zurich; W.A. Simpson, Glasgow Caledonian University

Statistical inference for reaction time experiment data • Inférence statistique pour des données d'expériences de temps de réaction

A reaction time experiment is studied in which a Poisson process of flashes is presented to an observer who responds by pushing and releasing a button. The data have the character of an M/G/infinity queue. Our interest is in understanding the server(s), i.e. the eye-brain-hand system of the observer. Using nonparametric point process techniques developed in a series of papers by Brillinger in conjunction with a simple parametric model, we can elucidate some of this structure. In particular, we study the behaviour of certain point process intensity functions under our simple model assumptions. This, in turn, tells us what kinds of features to look for in nonparametric estimates based on the real data.

Nous étudions une expérience de temps de réaction dans laquelle des " flashes " sont présentés à un observateur selon un processus de Poisson et celui-là répond en appuyant et en relâchant un bouton. Les données ont les caractéristiques d'une file d'attente M/G/infini. Notre intérêt est de comprendre le(s) serveur(s), c.-à-d. le système oeil-cerveau-main de l'observateur. En utilisant des techniques de processus ponctuels non paramétriques développées dans une série d'articles par Brillinger, conjointement avec un modèle paramétrique simple, nous pouvons élucider une partie de cette structure. En particulier, nous étudions le comportement de certaines fonctions d'intensité de processus ponctuels sous les hypothèses

de notre modèle simple. Alternativement, ceci nous indique quels types de caractéristiques rechercher dans les estimations non paramétriques basées sur les données réelles.

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Jin Qian, Judy-Anne W. Chapman, Yuejiao Fu, Yan Yuan, University of Waterloo; David E. Axelrod, Rutgers University; Naomi A. Miller, Princess Margaret Hospital; William A. Christens-Barry, Equipose Imaging LLC; H. Lavina Lickley, Wedad M. Hanna, Sunnybrook and Women's

Sample size implications for nuclear assessment of non-invasive breast cancer (DCIS) • Les implications de la taille d'échantillon sur l'évaluation nucléaire non-envahissante du cancer du sein (DCIS)

In DNA assessments of tumors with microarrays, there is a trend to reduce the number of cells on which the assessment is based, towards a few, and perhaps only one or two, cells. We have been working on the image analysis of non-invasive breast cancer [Ductal Carcinoma In Situ (DCIS)] nuclei for a cohort of 82 DCIS patients. We previously found there was heterogeneity of pathologic grading, and investigated here the heterogeneity of 39 image features, measured for approximately 20 nuclei from each of five ducts. The data generated from each duct were considered a replicate. There was substantive evidence of heterogeneous variance between replicates by Levene's test; therefore, we adjusted the image feature means of the replicates by the standard error of the mean. These adjusted means were used for tests of homogeneity between replicates with like grading. T-tests indicated that the replicates differed significantly for many image features. Two possible reasons for these differences are that the very precise image analysis is identifying real differences in nuclear grading between replicates and/or that 20 nuclei are insufficient to capture the natural variability in grading within a replicate/duct. In either instance, there are sample size implications about how many nuclei may be required for stable estimates of nuclear features. Therefore, because of intratumor heterogeneity analysis of only a single cell or a few cells from a tumor by microarray or other techniques is not sufficient to characterize the tumor for diagnostic or prognostic purposes.

Lors de l'évaluation des tumeurs par l'ADN avec des microréseaux, il y a une tendance à réduire à quelques unes le nombre de cellules sur lesquelles l'évaluation est basée, voir peut-être seulement une ou deux cellules. Nous avons travaillé sur l'analyse d'image de noyaux non invasifs de cellules de cancers du sein [Ductal Carcinoma In Situ (DCIS)] pour une cohorte de 82 patients de DCIS. Nous avons trouvé précédemment une hétérogénéité de l'évaluation pathologique, et étudié ici l'hétérogénéité de 39 caractéristiques de l'image, mesurée pour environ 20 noyaux de chacun des cinq conduits. Les données générées par chacun des conduits sont considérées comme une réplique. Le tests de Levene laisse voir une hétérogénéité substantielle de la variance entre les répliques; donc, nous avons ajusté les moyennes des caractéristiques de l'image des répliques par l'écart type de la moyenne. Ces moyennes ajustées sont utilisées pour des tests d'homogénéité entre les répliques avec évaluation. Les test-t indiquent que les répliques diffèrent de manière significative pour beaucoup de caractéristiques de l'image. Il y a deux raisons possibles pour expliquer ces différences, premièrement l'analyse très précise de l'image identifie de vraies différences dans l'histopronostic nucléaire entre les répliques et deuxièmement les 20 noyaux sont insuffisants pour capturer la variabilité naturelle de l'histopronostic dans une réplique/conduit. Dans l'une ou l'autre des situations, la taille de l'échantillon est im-

pliquée dans le nombre de noyaux requis pour obtenir des estimations stables des caractéristiques nucléaires. Par conséquent, à cause de l'hétérogénéité intra-tumeur, l'analyse d'une seule ou peu de cellules d'une tumeur par microréseaux ou autres techniques n'est pas suffisante pour caractériser la tumeur dans le but de diagnostiquer ou faire des pronostiques.

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Marylène TROUPE, Université des Antilles-Guyane (Guadeloupe); Samuel BARCLAY, Jean VAILLANT, Université des Antilles-Guyane; Petr LANSKY, Academy of Sciences of the Czech Republic

Statistic of time point process associated with a stochastic trajectory in heterogeneous environment • Statistique d'un processus ponctuel temporel dirigé par une trajectoire stochastique en milieu hétérogène

We present a model describing a sequence of events concerning a mobile moving in a heterogeneous environment. This model associates a time point process M_t with a stochastic trajectory W_t whose states space X includes several random effect zones. These random effects play a role on the point process intensity, according to the mobile position with respect to the zones. Likelihood-based techniques for statistical inference are developed according to the kind of information available, for instance the absence or presence of repetitions. Application to different areas are discussed.

Nous présentons un modèle décrivant une séquence d'événements concernant un mobile se déplaçant en milieu hétérogène. Ce modèle associe un processus ponctuel temporel M_t et une trajectoire stochastique W_t dont l'espace d'états X comporte des zones à effets aléatoires. Ces derniers influent sur l'intensité du processus ponctuel selon la position du mobile par rapport aux zones. Des techniques d'inférence basées sur la vraisemblance des observations sont développées selon le type d'information disponible, notamment l'absence ou pas de répétitions. Divers domaines d'application sont discutés.

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Anjela TZONTCHEVA, University of Toronto

Application of models for interval-censored survival data with informative examination time to the Polaris HIV seroconversion study data in Ontario • Application de modèles pour des données de survie censurées par intervalles avec temps d'examen informatifs avec les données de l'étude de séroconversion du VIH Polaris en Ontario

For some specific diseases, such as the human immunodeficiency virus (HIV) interval-censored survival data naturally arise because the exact seroconversion time is not known exactly but only to lie in an interval of time. Also with this type of data there is usually some subpopulation of individuals at greater risk who might visit more frequently. Therefore, heterogeneity in HIV incidence will likely be reflected in the frequencies of examination visits. If inter-examination times are informative then a common modeling approach is to include individual-level random effects (frailties) which capture the correlation between individual's event risk and examination rate. Such methodology as developed by Farrington and Gay (1999) will be applied to data from the Polaris HIV seroconversion study in Ontario.

Pour quelques maladies spécifiques telles que le virus d'immunodéficience humain (VIH), les données de survie censurées par intervalles surgissent naturellement parce que le temps exact de séroconversion n'est pas connu exactement, nous savons seulement qu'il se situe dans un intervalle de temps. De plus, avec ce type de données il y a habituellement une certaine sous-population des individus plus à risque qui pourrait visiter plus fréquemment. Par conséquent, l'hétérogénéité dans l'incidence du VIH est plus ou moins reflétée dans la fréquence des visites aux examens. Si les temps entre les examens sont instructifs, alors une approche commune pour modéliser consiste à inclure des effets aléatoires au niveau individuel ("frailties") qui capturent la corrélation entre le risque de l'événement de l'individu et le taux d'examen. Une telle méthodologie semblable à celle développée par Farrington et Gay (1999) est appliquée aux données à partir de l'étude de séroconversion du VIH Polaris en Ontario.

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Zilin Wang, David R. Bellhouse, University of Western Ontario; Jamie Stafford, University of Toronto

Generalized additive models for complex survey data • Modèles additifs généralisés pour des données de sondages complexes

Generalized additive models are an extension of the generalized linear models where the linearity of the regression function is relaxed to a sum of functions of covariates that need to be estimated. Although the generalized additive models are useful tools for data analysis and flexible modelling, they have been under-utilized in the analysis of survey data. Bellhouse et al.(2002) adopt the idea of the additive models to complex survey. Following on their results, we attempt to extend the generalized additive model to the survey data in this paper. A back fitting algorithm, called local scoring procedure, is used to carry out the estimations. The complete establishment of the additive model in Bellhouse et al.(2002) makes the further research in the generalized additive model possible for the reason that the most essential estimation involved in the local scoring procedure is the estimation of a weighted additive model. Inferential issues and smoothing parameter selections are considered as well. A diagnostic checking of the model using the offset function plots is conducted. The developed generalized additive model is illustrated by an empirical example with data from the 1990 Ontario Health survey.

Les modèles additifs généralisés sont une prolongation des modèles linéaires généralisés où la linéarité de la fonction de régression est réduite à une somme de fonctions des covariables qui doivent être estimés. Bien que les modèles additifs généralisés soient des outils utiles pour l'analyse de données et sont flexible à modéliser, ils sont insuffisamment utilisés pour l'analyse de données de sondages. Bellhouse et al. (2002) adoptent l'idée des modèles additifs pour les sondages complexes. En continuant d'après leurs résultats, nous essayons de prolonger le modèle additif généralisé aux données de sondages de cet article. Un algorithme back fitting, appelé la procédure de score local, est utilisé pour effectuer les estimations. L'établissement complet du modèle additif dans Bellhouse et al. (2002) rend possible les recherches ultérieures sur le modèle additif généralisé puisque l'estimation la plus essentielle impliqué dans le processus de score local est l'estimation d'un modèle additif pondéré. Des problèmes d'inférence et la sélection des paramètres de lissage sont aussi considérés. Une vérification par diagnostiques du modèle en utilisant les graphes de

la fonction décentrée est conduite. Le modèle additif généralisé développé est illustré par un exemple empirique avec des données de l'enquête sur la santé de l'Ontario de 1990.

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Machelle Wilson, William McCormick, Tom Hinton, University of Georgia

Monte Carlo comparisons of maximum likelihood estimation of high quantiles to the extreme value estimate of maximal exposure • Comparaisons par méthode de Monte Carlo de l'estimateur du maximum de vraisemblance des quantiles élevés à l'estimateur de valeur extrême de l'exposition maximale

Anthropogenic radiation in the environment, and the protection of natural populations, is of continuing global interest. The current paradigm argues that if the dose of the maximally exposed individual in a population is below the limit considered safe for an individual, then the population is adequately protected. Based on data sampled from natural populations, investigators need to be able to test the hypothesis that the maximally exposed individual's is acceptable. Thus, effective regulation of radiation in the environment requires not only sound scientific knowledge to establish regulatory criteria, but novel statistical approaches for estimating maximal exposure. Currently, there is no consensus in the regulatory community as to the best statistical approach. Statistics currently being used include the 95th percentile of the mean and the sample maximum. Additionally, some risk assessors have changed the internationally accepted paradigm and applied recommended dose limits to representatively, rather than the maximally exposed individuals. Recently, investigators have proposed the use of the maximum likelihood estimate of a very high percentile, such as the 99.99th, as an estimate of dose to the maximally exposed individual. In this study we compare all of the above mentioned statistics to an estimate based on Extreme Value Theory. To determine and compare the bias and variance of these statistics, we use Markov Chain/Monte Carlo (MC/MC) simulation techniques, in a procedure similar to a parametric bootstrap to estimate the bias associated with the above mentioned statistics for determining compliance with dose limits.

Les radiations anthropogéniques dans l'environnement et la protection des populations naturelles sont d'un intérêt global continu. Le paradigme actuel argumente le fait que si la dose de l'individu le plus exposé parmi une population est au-dessous de la limite considérée sécuritaire pour un individu, alors la population est protégée adéquatement. Basé sur des données échantillonnées sur des populations naturelles, les investigateurs doivent pouvoir tester l'hypothèse que les individus les plus exposés le sont d'une manière acceptable. Ainsi, une réglementation efficace des radiations dans l'environnement exige non seulement la connaissance scientifique pour établir des critères de régulations, mais des nouvelles approches statistiques pour estimer l'exposition maximale. Actuellement, il n'y a aucun consensus dans la communauté de régulation quant à la meilleure approche statistique. Les statistiques actuellement utilisées incluent le 95e percentile de la moyenne et du maximum de l'échantillon. De plus, quelques évaluateurs du risque ont changé le paradigme admis internationalement et ont appliqué des doses limites recommandées de façon représentative, plutôt que les individus les plus exposés. Récemment, les investigateurs ont proposé l'utilisation d'estimateurs du maximum de vraisemblance d'un percentile très élevé, tel que le 99.99ème, comme une évaluation de la dose de l'individu le plus exposé. Dans cette étude nous comparons toutes les statistiques mentionnées ci-dessus à une estimation basée sur la théorie des valeurs extrêmes. Pour déterminer et comparer

le biais et la variance de ces statistiques, nous utilisions des techniques de simulation de Monte Carlo avec chaînes de Markov (MC/MC), dans une procédure semblable au bootstrap paramétrique pour estimer le biais associé aux statistiques mentionnées ci-dessus pour déterminer la conformité aux doses limites.

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Zhang Ying, Ian McLeod, Hao Yu, University of Western Ontario

Experiments in multiple-choice randomized exams • Expériences pour les examens à choix multiples randomisés

Multiple-choice randomized (MCR) examinations in which the order of the items or questions as well as the order of the possible responses is randomized independently for every student are discussed. This type of design greatly reduces the possibility of cheating and has no serious drawbacks. We briefly describe how these exams can be conveniently produced and marked. We report on an experiment we conducted to examine the possible effect of such MCR randomization on student performance and conclude that no adverse effect was detected even in a quite large sample.

Nous discutons des examens à choix multiples randomisés (CMR) dans lesquels l'ordre des items ou des questions et l'ordre des réponses possibles est randomisé indépendamment pour chaque étudiant. Ce type de design réduit considérablement la possibilité de tricherie et ne possède aucun inconvénient sérieux. Nous décrivons brièvement comment ces examens peuvent être commodément produits et notés. Nous rendons compte d'une expérience que nous avons conduite pour examiner l'effet possible d'une telle randomisation de CMR sur la performance des étudiants et nous concluons qu'aucun effet nuisible n'a été détecté, même dans un grand échantillon.

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Zhang Ying, Ian McLeod, University of Western Ontario

Visualization on subset autoregressive admissible boundaries • Visualisation sur des frontières admissibles d'un sous ensemble autorégressif

The boundary of the admissible region for autoregressiv model is derived. The Barndorff-Neilsen and Schou (1973) transformation is extended to the subset autoregression case. It is shown how these results may be applied to obtain and visualize the subset autoregressive admissible boundaries in AR(4) process.

Nous dérivons la frontière de la région admissible au modèle autorégressif. Les transformations de Barndorff-Neilsen et Schou (1973) sont prolongées au cas de sous-ensemble autorégressif. Nous montrons comment ces résultats peuvent être appliqués pour obtenir et visualiser les frontières admissibles de sous-ensemble autorégressif dans les processus AR(4).

Sunday June 8 • Dimanche 8 juin 2:00 - 6:00 • 14h00-18h00 UC Great Hall
Monday June 9 • Lundi 9 juin 10:00 - 6:00 • 10h00-18h00 UC Great Hall

Ahmad Zoghoul, Mu'tah University

Estimation based on sample records versus the whole sample • Estimation basée sur des observations d'un échantillon versus tout l'échantillon

Let X1, X2, ... be a sequence of independent and identically distributed random variables from an absolutely continuous distribution F(x;theta) . Let X(1),X(2),..., be their corresponding sequence of record values. Assume that, for some reasons, the original data is not available, but a set of record values. And suppose that estimates of theta are needed. The question is how good are these estimates based on record values compared to those based on the whole data?. Several distributions like uniform , exponential , Pareto, power function, and Weibull are considered.

Soient X1, X2... une séquence de variables aléatoires indépendantes et identiquement distribuées provenant d'une distribution F(x;theta). Soient X(1), X(2)... la séquence des valeurs observées correspondantes. Supposons que pour des raisons quelconques, les données originales ne sont pas disponibles à part un ensemble de valeurs observées. Supposons qu'une estimation de thêta soit nécessaire. La question est la suivante: à quel point ces estimations basées sur les valeurs observées sont bonnes comparativement à celles basées sur toutes les données? Plusieurs distributions comme la loi uniforme, la loi exponentielle, la fonction de puissance de Pareto et la loi Weibull sont considérés.

Session/Séance 1 • Welcome and SSC Presidential Invited Address • Mot de bienvenue du président et allocution de son invité d'honneur

Monday June 9 • Lundi 9 juin 8:30 - 10:00 • 8h30 - 10h00 MM Ondaajte Theatre

Robert Gentleman, Harvard School of Public Health

Modern statistical computing • Calcul statistique moderne

Statistical computing has evolved substantially from its roots in numerical computation and algorithm development. In this talk I will provide an overview of some of the more recent developments. These include reproducible research, visualization, computational inference, intersystems interfaces and software development. The tools required are varied and appropriate use requires a broad understanding of computation as well as of statistics. In this talk I will give examples of these different applications of statistical computing. The availability of high quality software is likely to have a larger impact on the practice of statistics (in other disciplines) than any specific methodological developments. Some concerns regarding training of students, funding of software projects and publication will also be addressed.

Le calcul statistique a évolué substantiellement de ses racines de calcul numérique et de développement d'algorithmes. Dans cette présentation, nous fournissons une vue d'ensemble sur certains des développements les plus récents. Ceux-ci incluent la recherche reproductible, la visualisation, l'inférence informatique, les interfaces inter-systèmes et le développement des logiciels. Les outils exigés sont variés et une utilisation appropriée exige une large compréhension du calcul informatique et des statistiques. Nous donnons des exemples de ces différentes applications du calcul statistique. La disponibilité de logiciels de haute qualité est susceptible d'avoir un plus grand impact sur la pratique des statistiques (dans d'autres disciplines) que tous les développements méthodologiques spécifiques. Quelques soucis concernant la formation des étudiants, le financement des projets d'amélioration des logiciels et de la publication sont également adressés.

Session/Séance 2 • Case Study I - Blood Pressure • Étude de cas I - Pression artérielle

Monday June 9 • Lundi 9 juin 10:30 • 10h30 LSC 240

Raymond LAM, GlaxoKlineSmith **Introduction**

Monday June 9 • Lundi 9 juin 10:35 • 10h35 LSC 240

Pingzhao Hu, Dong Song, Dalhousie University

Monday June 9 • Lundi 9 juin 10:50 • 10h50 LSC 240

Zainab Abdurrahman, Bethany Giddings, Sofia Mosesova, University of Waterloo

Monday June 9 • Lundi 9 juin 11:05 • 11h05 LSC 240

Louis-François Poirier, Javier Oyarzun, Université de Montréal

Monday June 9 • Lundi 9 juin 11:20 • 11h20 LSC 240

Ana-Maria Staicu, Mark Kane, Hadas Moshonov, University of Toronto

Monday June 9 • Lundi 9 juin 11:35 • 11h35 LSC 240

Hossein Yazdi, University of Guelph

Monday June 9 • Lundi 9 juin 11:50 • 11h50 LSC 240

Sophia Lee, Hanna Jankowksi, Joanna Biernacka, University of Toronto

Monday June 9 • Lundi 9 juin 12:05 • 12h05 LSC 240

Christina Frisina, Cathlin McNally, McMaster University

Session/Séance 3 • Longitudinal Data Analysis in Biostatistics • Analyse de données longitudinales en biostatistique

Monday June 9 • Lundi 9 juin 10:30 • 10h30 LSC 332

Brajendra Sutradhar, Gary Sneddon, Memorial University of Newfoundland

Analyzing two-way correlated familial longitudinal data • Analyse de la corrélation double dans des données longitudinales familiales

In many biomedical fields, there are situations where repeated data are collected from a large number of families/clusters. This type of familial longitudinal data exhibit two-way correlations: first, the responses of the family members are correlated as the members share the common family effect; secondly, these correlated responses are repeatedly observed over a period of time and the repeated responses are also correlated. In this talk, we introduce a generalized least square (GLS) approach to analyze a semiparametric linear model based familial longitudinal data, whereas we exploit a generalized quasilikelihood (GQL) approach to analyze a generalized linear model based familial longitudinal data. The GLS and GQL approaches take the two-way familial and longitudinal correlations into account and they produce consistent and efficient estimates for the parameters involved. The methods will be illustrated by numerical examples.

Dans beaucoup de domaines biomédicaux, il y a des situations où des données répétées sont rassemblées à partir d'un grand nombre de familles ou de groupes. Ce type de données longitudinales familiales comporte des corrélations doubles: premièrement, les réponses des membres de la famille sont corrélées par leur partage de l'effet commun de la famille; deuxièmement, les réponses sont observées dans le temps et les réponses répétées sont corrélées. Dans cette présentation, nous introduisons une approche généralisé des moindres carrés (GLS) pour analyser des données longitudinales familiales basées sur un modèle linéaire semi-paramétrique, tandis que nous exploitons une approche généralisée de quasi-vraisemblance (GQL) pour analyser des données longitudinales familiales basées sur un modèle linéaire. Les approches de GLS et de GQL prennent en considération la corrélation double et elles produisent des estimations efficaces pour les paramètres impliqués. Les méthodes sont illustrées par des exemples numériques.

Monday June 9 • Lundi 9 juin 11:00 •11h00 LSC 332

Raymond Carroll, Texas A&M University

Longitudinal data analysis in biostatistics • Données longitudinales et en grappes et régression non paramétrique et semi-paramétrique

I will review the problem of nonparametric and semiparametric regression for longitudinal/clustered data. In the independent data case, kernel methods and spline methods are essentially asymptotically equivalent for these problems. The same is not true for the longitudinal/clustered data case, with standard kernel methods failing to account for correlation in any sort of effective form. Indeed, such kernel methods and spline methods have different local influence functions ("effective kernels"). A method due to Naisyin Wang resolves the dilemma. We indicate its use in semiparametric regression, and also its asymptotic equivalence with splines.

Je passerai en revue le problème de la régression non paramétrique et semi paramétrique pour des données longitudinales et en grappes. Dans le cas de données indépendantes, les méthodes du noyau et les méthodes utilisant des splines sont asymptotiquement équivalentes.

L'analogie au cas des données longitudinales et en grappes n'est pas possible, les méthodes du noyau usuelles n'expliquent pas efficacement la corrélation. En effet, les méthodes du noyau et par splines ont différentes fonctions d'influence locales ("noyau opérant"). Une méthode due à Naisyin Wang résout le problème: nous décrivons son utilisation dans la régression semi-paramétrique et également son équivalence asymptotique avec les splines.

Monday June 9 • Lundi 9 juin 11:30 • 11h30 LSC 332

Georgia Roberts, Milorad Kovacevic, Yves Lafortune, Owen Phillips, David Binder,
Statistics Canada/Statistique Canada

Using an estimating equation bootstrap approach for obtaining variance estimates when modelling complex health survey data • L'emploi d'une approche bootstrap aux équations d'estimation pour obtenir des estimations de variance lors de la modélisation de donnés provenant des enquêtes sur la santé à plan complexe

Whether survey data are being used for estimating descriptive statistics about the population from which the sample was drawn or in a design-based approach for more complex analyses, sufficient information about the design is required in order to produce adequate design-based variance estimates. For the Statistics Canada health surveys being used for analysis, such as the National Population Health Survey and the Canadian Community Health Survey, the design information is being provided in the form of a final weight variable and a set of bootstrap weight variables, each one corresponding to a bootstrap sample of psu's from the sampled psu's (see Rust and Rao, 1996, for approach). Much of the software available for using the design information in this form takes a "direct" approach to producing variance estimates for an estimated quantity - using the average of the squared differences between the bootstrap estimates and the final-weight estimate. While straightforward, this approach has been found to have drawbacks, particularly in the case of iterative model-fitting in a domain containing a small sample such as in the fitting of a logistic model to the sample from a specialized subpopulation. One option would be to use an estimating equation bootstrap approach for such situations. Justification for the approach and some examples of its application will be discussed. Also to be discussed will be possibilities for extending currently-available software that were not intended for bootstrap methods of variance estimation, to make use of bootstrap weight variables.

Que ce soit pour faire l'estimation de statistiques décrivant la population à partir de laquelle l'échantillon a été tiré ou encore pour faire des analyses plus complexes fondées sur le plan de sondage, de l'information appropriée quant au plan de sondage est requise afin de produire des estimations de variance adéquates. Lorsque des enquêtes sur la santé, produites par Statistique Canada, sont utilisées à des fins analytiques, enquêtes telles que l'Enquête nationale sur la santé de la population (ENSP) et l'Enquête sur la santé dans les collectivités canadiennes (ESCC), l'information relative au plan de sondage est transmise sous la forme d'une variable contenant le poids final et un ensemble de variables de poids bootstrap, chacune d'elle correspondant à un échantillon bootstrap d'unités primaires d'échantillonnage (upé), prélevé parmi les upé échantillonnées (voir Rust et Rao, 1996, pour une discussion de cette approche). La plupart des logiciels utilisant l'information reliée au plan de sondage prennent une approche 'directe' pour produire les estimations de variance pour les quantités estimées - c'est-à-dire la moyenne des différences au carré entre les estimations bootstrap et l'estimation avec le poids final. Bien que directe, cette approche présente tout de même certains inconvénients, en particulier dans le cas d'une modélisation avec composante itérative pour un domaine contenant une faible taille d'échantillon, comme

lors de l'utilisation d'un modèle de régression logistique construit à partir d'une sous-population spécialisée d'un échantillon. Une option consisterait à utiliser une approche bootstrap aux équations d'estimation dans ces situations. Une justification de l'approche et des exemples de son application seront présentés. Les possibilités d'étendre les capacités de certains logiciels, initialement non conçus pour faire l'estimation de variance via la méthode du bootstrap, seront également discutées.

Session/Séance 4 • Machine Learning Methods From a Statistical Perspective • Méthodes d'apprentissage automatique d'un point de vue statistique

Monday June 9 • Lundi 9 juin 10:30 • 10h30 LSC 338

George C. Tseng, Wing Hung Wong, Harvard University; Xiaotong Shen, Ohio State University; Xuegong Zhang, Tsinghua University

From margin-based classification to ψ-learning • De la classification basé sur les marges aux " ψ-learning"

The concept of large margins plays an important role in analyzing learning methodologies such as Boosting, Neural Networks, and Support Vector Machine (SVM). In this talk, I will present a new learning methodology called ψ-learning as well as the associated computational tools. While retaining the interpretation of large margins, ψ-learning delivers high performance in generalization, especially in nonseparable cases, as it is derived from a direct consideration of generalization errors. In order to realize its potential, we tackle a difficult nonconvex minimization problem, utilizing the global optimization techniques. In particular, we propose two computational strategies based on d.c. (differenced convex) programming, yielding a sequential quadratic program for the minimization problem. Numerical experiments are performed using simulated and benchmark examples, which suggest that the computational strategies can realize the theoretical advantages of ψ-learning.

Le concept des grandes marges joue un rôle important pour analyser les méthodes d'apprentissage telles que le boosting, les réseaux neuronaux et les machines à vecteurs de suppports (SVM). Dans cette présentation, je présenterai une nouvelle méthode d'apprentissage appelée le " ψ-learning " ainsi que les outils informatiques associés. Tout en retenant l'interprétation des grandes marges, le "ψ-learning" fournit des performances élevées lors de la généralisation, particulièrement pour des cas non séparable car ils sont dérivés en considérant directement les erreurs de généralisation. Afin de réaliser tout son potentiel, nous abordons un problème difficile de minimisation non convexe en utilisant les techniques d'optimisation globales. En particulier, nous proposons deux stratégies informatiques basées sur la programmation convexe différentié, donnant un programme quadratique séquentiel pour le problème de minimisation. Des expériences numériques sont effectué en utilisant des simulations et des exemples repères pour montrer que les stratégies informatiques peuvent réaliser les avantages théoriques du " ψ-learning".

Monday June 9 • Lundi 9 juin 11:00 • 11h00 LSC 338

Peter Bühlmann, ETH Zurich

Boosting methods: why they can be useful for high-dimensional data • Les méthodes de boosting: pourquoi peuvent-elles être utiles avec des données de haute dimension

Boosting (Freund and Schapire, early 1990's) has been proposed as a technique for generating and aggregating multiple classifiers. In a number of interesting cases, it has empirically been found to produce excellent predictions. As pointed out first by Breiman (1999), Boosting is a gradient descent technique in function space. This view turns out to be very fruitful to modify Boosting as a general technique which can also be used in other settings than classification such as regression or survival analysis. In the regression case, (modified) Boosting is essentially the same as Matching Pursuit (Mallat and Zhang, 1993) in signal processing and more generally as a so-called Weak Greedy Algorithm (deVore and Temlyakov, late 1990's) in computational mathematics. The talk will build up from understanding of Boosting to various facts and results such as: (i) new asymptotic adaptivity results about Boosting; (ii) Boosting as a method which simultaneously does variable selection and assigns different amount of degrees of freedom among selected predictors; (iii) computational attractiveness. We will demonstrate that the issues mentioned in (ii) and (iii) are particularly useful for data with very high-dimensional predictors. Partially joint work with Bin Yu, UC Berkeley.

La technique du boosting a été proposé (Freund et Schapire, début des années 90) pour générer et agréger des classificateurs multiples. Dans un certain nombre de cas intéressants, on a empiriquement montré qu'elle peut produire d'excellentes prédictions. Tel qu'il a été montré par Breiman (1999), le boosting est une technique du gradient sur l'espace des fonctions. De cela, il s'avère être fructueux d'adapter le boosting, comme technique générale, à d'autres situations que la classification, telle que la régression ou l'analyse de survie. Dans le cas de la régression (modifié), le boosting est essentiellement la même chose que le " Matching Pursuit " (Mallat et Zhang, 1993) dans le traitement des signaux et plus généralement, la même chose que le " Weak Greedy Algorithm " (deVore et Temlyakov, fin des années 90) en mathématiques informatiques. La présentation se développe par la compréhension du boosting à divers faits et résultats comme: (i) des nouveaux résultats d'adaptativité asymptotique au sujet du boosting; (ii) le boosting comme méthode qui choisi simultanément les variables et assigne une quantité différente de degrés de liberté parmi les prédicteurs choisis et (iii) les avantages informatiques. Nous démontrons que les cas mentionnés en (ii) et (iii) sont particulièrement utiles pour des données avec des prédicteurs de très haute-dimension. Travail partiellement conjoint avec Bin Yu, UC Berkeley.

Monday June 9 • Lundi 9 juin 11:30 • 11h30 LSC 338

Yi Lin, Yongho Jeon, University of Wisconsin

Random forests and adaptive nearest neighbors • Les forêts aléatoires et les voisins les plus proches adaptatifs

In this talk random forests are studied through their connection with a new framework of adaptive nearest neighbor methods. We first introduce a concept of potential nearest neighbors and show that random forests can be seen as adaptively weighted potential nearest neighbor methods. Various aspects of random forests are then studied from this perspective. We show that random forests with adaptive splitting schemes assign weights to potential nearest neighbors in a desirable way: for the estimation at a given target point, these random forests assign voting weights to the potential nearest neighbors of the target point according to the local importance of different input variables. We propose a new simple splitting scheme that achieves desirable adaptivity in a straightforward fashion. This simple scheme can be combined with existing algorithms. The resulting algorithm is computationally faster, and gives comparable results. Other possible aspects of random

forests, such as using linear combinations in splitting, are also discussed. Simulations and real datasets are used to illustrate the results.

Dans cette présentation, nous étudions les forêts aléatoires par leur lien avec un nouveau cadre de méthodes adaptatives du voisins le plus proche. Tout d'abord, nous présentons le concept du voisins potentiellement le plus proche et nous montrons que les forêts aléatoires peuvent être vues comme une version adaptative pondérée de la méthode du voisin potentiellement le plus proche. Différents aspects des forêts aléatoires sont alors étudiés sous cet angle. Nous montrons que les forêts aléatoires avec des division adaptatives assignent des poids aux voisins potentiellememt les plus proches d'une manière souhaitable: pour l'estimation à un point cible donné, ces forêts aléatoires assignent les poids aux voisins potentiellement les plus proches du point cible selon l'importance locale de différentes variables d'entrée. Nous proposons une nouvelle méthode de division simple qui réalise une adaptabilité souhaitable de manière directe. Cet arrangement simple peut être combiné avec des algorithmes existants. L'algorithme résultant est plus rapide lors du traitement informatique, et donne des résultats comparables à d'autres méthodes existantes. Nous discutons aussi d'autres aspects des forêts aléatoires, tels que l'utilisation de divisions basées sur des combinaisons linéaires. Des simulations sur des jeux de données réels sont présentées pour illustrer les résultats.

Session/Séance 5 • Statistical Inference I: Inference in Partially Linear Models • Inférence statistique I: Inférence pour des modèles partiellement linéaires

Monday June 9 • Lundi 9 juin 10:30 • 10h30 LSC 242

Aidan McDermott, Francesca Dominici, Trevor Hastie, Scott L. Zeger, Jonathan Samet,
Johns Hopkins University

Issues in semiparametric regression • Sujets en régression semi-paramétrique

Evidence from time series studies of air pollution and health is central to major policy decisions concerning the risk of death associated with air pollution exposure. The nature and characteristics of time series data make risk estimation challenging, requiring development of complex statistical methods able to detect effects that are very small relative to the combined effects of confounders and residual variation. Using the National Mortality Morbidity Air Pollution Study, which includes time series data from the 90 largest US locations for the period 1987-1994, we discuss: 1) approaches for control of confounding factors in absence of a strong biological model; 2) generalizations of the S-plus software package "gam" which allows to calculate asymptotically exact standard errors of the air pollution effects; and 3) hierarchical models for estimating national-average pollution effects as function of the degree of adjustment for confounding factors. Identification of the air pollution effect (a weak signal) in presence of several potential confounders call for a systematic assessment of model choice and the development of new methods. Importantly, the heavy policy weight placed on these findings requires a level of confidence that is difficult to attain, regardless of the sophistication of the statistical approach.

Les études de séries chronologiques sur la pollution atmosphérique et sur la santé fournissent des indices cruciaux pour les décisions sur les grandes politiques concernant le risque de mourir lié à l'exposition à la pollution atmosphérique. La nature et les caractéristiques des données de séries chronologiques rendent l'estimation du risque difficile, exigeant l'élaboration de méthodes statistiques complexes capables de détecter des effets

très petits relativement aux effets combinés de la variation des facteurs de confusion et de la variation résiduelle. En utilisant " l'étude nationale sur la morbidité et la mortalité de la pollution atmosphérique " qui inclut des données de série chronologique des 90 plus grandes localisations des États-Unis pour la période de 1987-1994, nous discutons: 1) des approches pour contrôler les facteurs de confusion en absence d'un modèle biologique fort; 2) des généralisations du package " gam " de S-plus qui permet de calculer des écarts types asymptotiquement exacts pour les effets de la pollution atmosphérique; et 3) des modèles hiérarchiques pour estimer les effets nationaux moyens de la pollution comme une fonction du degré d'ajustement des facteurs de confusion. L'identification de l'effet de la pollution atmosphérique (un signal faible) en présence de plusieurs facteurs de confusions potentiels réclame une évaluation systématique du choix du modèle et de l'élaboration de nouvelles méthodes. Cependant, le lourd poids des politiques qui pèse sur ces résultats exige un niveau de confiance difficile à atteindre, indépendamment de la sophistication de l'approche statistique.

Monday June 9 • Lundi 9 juin 11:00 • 11h00 LSC 242

Tim Ramsay, Daniel Krewski, University of Ottawa/Universite d'Ottawa; Richard Burnett, Health Canada

Concurvity-induced bias in the generalized additive model • Biais de concurvité induit dans les modèles additifs généralisés

The fact that concurvity, or nonparametric multicollinearity, can introduce bias into partially linear generalized additive models (GAMs) has recently caused quite a stir among environmental epidemiologists studying the health effects of air pollution. While the unfortunate fact that some contentious policy decisions limiting airborne emissions have relied on GAM analyses which are now believed to be biased has led air pollution researchers to be particularly concerned with concurvity-induced bias, no-one performing inference on GAMs can afford to ignore this problem. Although parametric bootstrap bias estimators are useful for demonstrating that a given dataset may be subject to concurvity-induced bias, certain types of dependency between the bias and the unknown functional relationship generating the observed data can result in these bias estimators themselves being biased. As a consequence, parametric bootstrap analysis can indicate the presence of bias but should not be used to draw conclusions about the direction and magnitude of the bias. We argue that the results of an unashamedly ad-hoc exploration of the range of biases that might plausibly affect a GAM fitted to fifteen years of daily air pollution measurements for the city of Toronto indicate that, at least for now, the best way to avoid biased fits is to only use GAMs with no nonparametric components. This is equivalent to GLMs (generalized linear models) rather than nonparametric GAMs. This solution is admittedly somewhat less than satisfactory, since the appeal of GAMs is precisely their ability to fit models without parametric assumptions, and the search for a better solution is ongoing. There is some hope that a modification of the backfitting algorithm currently used to fit nonparametric GAMs may eliminate the bias.

Le fait que le concurvité, ou la multi-colinéarité non paramétrique, peut introduire un biais dans les modèles additifs généralisés partiellement linéaires (MAG) a récemment causé tout une agitation parmi les épidémiologues environnementaux étudiant les effets de la pollution atmosphérique sur la santé. Malheureusement, puisque que des décisions de politique controversées limitant les émissions dans l'air se sont fondées sur les analyses de MAG qui sont maintenant soupçonnées d'être biaisées, cela a mené les chercheurs sur la

pollution atmosphérique à être concernés par le biais induit par la concurvité. Personne qui fait de l'inférence sur des modèles additifs généralisés ne peut ignorer ce problème. Bien que les estimateurs paramétriques du biais par bootstrap sont utile pour démontrer qu'un jeu de données peut être sujet au biais induit par concurvité, certains types de dépendance entre le biais et la relation fonctionnelle inconnue qui génère les données observées peut faire que ces estimateurs du biais soient eux-mêmes biaisés. Par conséquent, l'analyse de bootstrap paramétrique peut indiquer la présence de biais mais ne doit pas être utilisée pour tirer des conclusions au sujet de la direction et de l'importance du biais.

Monday June 9 • Lundi 9 juin 11:30 • 11h30 LSC 242

Isabella Ghement, University of British Columbia

Smoothing parameter selection in partially linear models with serially correlated errors • Choix du paramètre de lissage dans les modèles partiellement linéaires avec erreurs corrélées en série

Partially linear models with serially correlated errors provide a flexible framework for analyzing time series air pollution data. They allow the practitioner to make easily interpretable inferences about the pollutant effect, assumed to be linear, while accounting for possibly nonlinear nuisance seasonal or weather effects and potential error dependence. The quality of these inferences can however be severely affected by the specification of the amount of smoothing controlling the seasonal cycles or weather patterns. Although complicated by the dependence of errors and relation between the parametric and nonparametric predictors in the model, the choice of an appropriate amount of smoothing is vital to carrying out valid inferences about the pollutant effect. Since traditional bandwidth selection methods such as cross-validation or modified cross-validation are unable to yield the appropriate amount of smoothing, we propose using an empirical bias smoothing parameter selection method. The method finds the amount of smoothing that minimizes the estimated conditional mean squared error of a modified backfitting estimate of the parametric component of the model given the predictors. The performance of the method is tested on a Mexico City air pollution dataset and on simulated data. Our results are a generalization of those contained in Opsomer and Ruppert (1999), who studied a similar model with i.i.d. errors.

Les modèles partiellement linéaires avec des erreurs corrélées en série fournissent un cadre flexible pour analyser des données de série chronologique sur la pollution atmosphérique. Ils permettent au praticien de faire facilement de l'inférence interprétables sur l'effet des polluants, que l'on suppose linéaires, en tenant compte de la présence potentielle d'effets de nuisances saisonniers ou météorologiques non-linéaires et de la dépendance des erreurs. La qualité de ces inférences peut cependant être sévèrement affectée par les spécifications de la quantité de lissage contrôlant les cycles saisonniers ou les patrons météorologiques. Bien que compliqué par la dépendance des erreurs et de la relation entre les prédicteurs paramétriques et non paramétriques dans le modèle, le choix d'une quantité appropriée de lissage est essentiel pour faire le l'inférence valide au sujet de l'effet du polluant. Puisque les méthodes traditionnelles de choix de taille de fenêtre telles la validation croisée ou la validation croisée modifiée ne peuvent pas donner la quantité appropriée de lissage, nous proposons d'utiliser une méthode de sélection des paramètres par lissage empirique du biais. La méthode trouve la quantité de lissage qui minimise l'erreur quadratique moyen conditionnelle estimée avec une estimation " backfitting " modifiée de la composante paramétrique du modèle étant donné les prédicteurs. La performance de la méthode est examinée par un

jeu de données sur la pollution atmosphérique de Mexico et sur des données simulées. Nos résultats sont une généralisation de ceux contenu dans Opsomer et Ruppert (1999), qui ont étudié un modèle semblable avec des erreurs indépendantes et identiquement distribuée.

Session/Séance 6 • Environmetrics • La mésométrie

Monday June 9 • Lundi 9 juin 10:30 • 10h30 LSC 238

Sylvia Esterby, University College of the Okanagan

Models for biological productivity in lakes • Modèles pour la productivité biologique dans les lacs

The principle of limiting factors provides the rationale for lake-management practices that control the level of the nutrient considered to be the limiting factor for growth. Data sets obtained from large-scale studies on many lakes are generally collected to establish the factor or factors that limit productivity, where mean values for each lake are used. The most common approach is to model the mean response as a function of the potentially limiting factor. Alternative models have been proposed which better account for the principle of limiting factors, where it is the maximum response that is limited. Monitoring programs that provide data collected over space and time for the same lake are a source of individual measurements of both response and potentially limiting variables for a number of different productivity scenarios. The commonly used models and models that incorporate the principle of limiting factors will be examined for the latter type of data set.

Le principe des facteurs limitants fournit le raisonnement pour les pratiques en matière de gestion des lacs qui contrôlent le niveau du nutriment considéré comme le facteur limitant pour la croissance. Des jeux de données obtenus à partir d'études à grande échelle sur plusieurs lacs sont généralement rassemblés pour établir le ou les facteurs qui limitent la productivité, où les valeurs moyennes pour chaque lacs sont utilisées. L'approche la plus commune est de modéliser la réponse moyenne comme une fonction du facteur limitant potentiel. Nous avons proposé des modèles alternatifs qui gèrent mieux le principe des facteurs limitant, où c'est la réponse maximale qui est limitée. Les programmes de contrôle qui fournissent des données sur un espace donné et sur une période de temps pour un même lac sont une source de mesures individuelles pour la variable réponse et les variables limitantes potentielles pour une panoplie de scénarios de productivité différents. Les modèles généralement utilisés et les modèles qui incorporent le principe de facteurs limitants sont examinés pour le dernier type jeux de données.

Monday June 9 • Lundi 9 juin 11:00 • 11h00 LSC 238

Montserrat Fuentes, North Carolina State University

Statistical assessment of geographic areas of compliance with air quality standards • Évaluation statistique des secteurs géographiques en conformité avec les standards sur la qualité de l'air

A statistical method is developed to classify geographical regions according to the air quality standards for criteria pollutants. A geographic location is designated by the US Environmental Protection Agency as an area of non-attainment when it does not meet the air quality standard for one of the criteria pollutants. We statistically model air pollution and interpolate ground measurements of pollution levels at locations where there are no air quality monitors to determine the geographic areas of non-attainment. Our approach for interpolation takes into account that spatial patterns of air pollutants change with location.

We also provide estimates of probabilities of non-attainment, that are used to classify the non-attainment areas by degree of severity. The approach is applied using available data from 513 sites throughout the eastern USA where ground-level ozone is measured hourly.

Une méthode statistique est développée pour classifier les régions géographiques selon les standards sur la qualité de l'air pour certains polluants. Un emplacement géographique est indiqué par l'agence de protection de l'environnement des États-Unis comme un secteur non conforme lorsqu'il ne répond pas aux standards sur la qualité de l'air pour un des polluants types. Nous modélisons la pollution de l'air et nous interpolons les mesures du niveau de pollution au sol aux endroits où il n'y a aucun moniteur de la qualité de l'air pour déterminer les secteurs géographiques non conformes. Notre approche pour l'interpolation prend en compte la non-homogénéité spatiale des modèles. Nous fournissons également des évaluations des probabilités de non conformité, qui sont utilisées pour classifier les secteurs de non conformité par degré de sévérité. L'approche est appliquée en utilisant des données disponibles à partir de 513 emplacements, où l'ozone au niveau du sol est mesuré à chaque heure, dans l'Est des États-Unis.

Monday June 9 • Lundi 9 juin 11:30 • 11h30 LSC 238

Lawrence H. Cox, National Center for Health Statistics/Centre national pour les statistiques en santé

On properties of multi-dimensional statistical tables • Les propriétés des tables statistiques multidimensionnelles

Statistical data often can be conveniently organized in tabular form for display and analysis. Counts are nonnegative integers, and often magnitude data take nonnegative integer values. Two-dimensional tables enjoy mathematical properties on which important statistical methods depend, e.g., for stratified sampling, imputation, disclosure limitation,and sampling and fitting log-linear models to contingency tables. We demonstrate that many desirable mathematical properties, and consequently their associated statistical methods, are not extendible in all cases to three or higher dimensions. We demonstrate that ill-behaved examples are ubiquitous, abundant and consequently not mathematical anomalies. To address these shortcomings, we provide necessary and sufficient conditions and an empirical test for the existence of an n-dimensional table with prescribed (n-1)-dimensional marginal totals (feasibility) and a characterization of n-dimensional tables with prescribed (n-1)-dimensional marginals for which continuous bounds on integer-valued entries exist and are integer (integrality).

Les données statistiques peuvent souvent être organisées en forme de tableaux pour les présenter et les analyser. Les comptes sont des nombres entiers non négatifs et souvent les données de grandeur prennent aussi des valeurs entières non négatives. Les tables à deux dimensions ont des propriétés mathématiques sur lesquelles dépendent certaines méthodes statistiques importantes, par exemple pour l'échantillonnage stratifié, l'imputation, la limitation des divulgations et pour échantillonner et ajuster un modèle log-linéaire aux tableau de contingences. Nous démontrons que beaucoup de propriétés mathématiques intéressantes, et par conséquent leurs méthodes statistiques associées, ne se généralisent pas au cas à trois dimensions ou plus. Nous démontrons que les exemples problématiques sont omniprésents et abondants, et par conséquent, ne sont pas des anomalies non mathématiques. Pour adresser ces imperfections, nous fournissons des conditions nécessaires et suffisantes et un test empirique pour l'existence d'une table à n dimensions associé aux totaux marginaux à (n-1)-dimensions (faisabilité) et une caractérisation des tables à n dimensions associé

aux marginales de (n-1)-dimensions pour lesquelles des bornes continues sur les entrées à valeur entière existent et sont entière (intégralité).

Session/Séance 7 • Survey Methods Contributed Session I: Estimation - Applied • Méthodes d'enquête I: Estimation - applications

Monday June 9 • Lundi 9 juin 10:30 • 10h30 LSC 234

Pat Newcombe-Welch, Statistics Canada/Statistique Canada, University of Waterloo; Tim Seifert, Memorial University

Big country, small sample - what can we safely conclude? • Grand pays, petit échantillon: que pouvons nous conclure de manière sûre?

Statistics Canada provides survey weights along with its data files and advises that the weights be used in order to produce unbiased estimates. The survey weights are constructed so that known subtotals, such as provincial populations and gender totals are preserved. One aspect of this method is that the sample respondents from a small province such as Prince Edward Island will on average have much smaller weights than the sample respondents from a large province such as Ontario. If a statistic is based on a subsample which contains relatively few respondents from each province, although the weighted estimate will be theoretically unbiased, the accompanying variance may be expected to be large due to the large variability amongst the weights. Researchers who experience this phenomenon may feel justified in using an unweighted analysis. An example of this type from the National Longitudinal Survey of Children and Youth (NLSCY) is presented, in which it is demonstrated that it is possible to conduct a weighted analysis and minimize the effect of variance inflation by using provincial groupings as a variable in the model.

Statistique Canada fournit des poids de sondages avec ses fichiers de données et conseille que les poids soient utilisés afin de produire des estimations non biaisées. Les poids de sondages sont construits de sorte que les totaux partiels connus, tels que les populations provinciales et les totaux des sexes soient préservés. Une caractéristique de cette méthode est que l'échantillon des répondants d'une petite province telle que l'Île du Prince Édouard a en moyenne des poids beaucoup plus petits que l'échantillon des répondants d'une grande province comme l'Ontario. Si une statistique est basée sur un sous-échantillon qui contient relativement peu de répondants de chaque province, bien que l'estimation pondérée soit théoriquement non biaisée, nous pouvons prévoir que la variance associée sera grande dû à la grande variabilité des poids. Les chercheurs qui voient ce phénomène peuvent se sentir justifié d'utiliser une analyse non pondérée. Un exemple de ce type de phénomène provenant de l'Enquête longitudinale nationale sur les enfants et la jeunesse (NLSCY) est présenté, dans lequel nous démontrons qu'il est possible de conduire une analyse pondérée et de minimiser l'effet de l'inflation de la variance en utilisant des groupements provinciaux comme variable dans le modèle.

Monday June 9 • Lundi 9 juin 10:45 • 10h45 LSC 234

André Cyr, Catalin Dochitoiu, Statistics Canada/Statistique Canada

Latent variable modelling in the longitudinal context: The case of the National Longitudinal Survey of Children and Youth in Canada • Modélisation de variables latentes dans le contexte longitudinal: Le cas pour l'Enquête Longitudinale Nationale sur les Enfants et les Jeunes au Canada

A unique study of Canadians from birth to adulthood, the National Longitudinal Survey of Children and Youth (NLSCY) provides a single source of data for the examination of child development in context, including the diverse life paths of normal development. The NLSCY is designed to follow a representative sample of Canadian children from 0 to 25 years of age, with data collection occurring at two-year intervals. The current sample of NLSCY children is large enough for analysis by cohorts, sub-populations and provinces. Starting In 1994, the first year of data collection, the sample of about 20,000 children then aged 0 to 11, were selected and information about their home and school environment is being collected.

Item response theory (IRT) offers many advantages to researchers who need to quantify children's reading and writing abilities, and for this reason, IRT methods have been adopted in Statistics Canada's National Longitudinal Survey for Children and Youth. IRT methods have a long history in the field of psychometrics, and provide a model-based method for characterising both test items and subject abilities, and for generating predictions of individual abilities. For the most part, IRT methods implicitly assume i.i.d. observations, so that the application of these methods to complex surveys raises a number of issues. Questions arise as to the use or otherwise of the sample weights, appropriate methods for predicting individual abilities (for inclusion on public use datasets), and appropriate methods for estimating parameters and their variances.

The presentation will outline the many steps taken by Statistics Canada to apply psychometric techniques to derive appropriately measured items parameters in the construction of assessment tools and to derive relevant ability scores for children to support longitudinal analysis. This will be an opportunity to showcase some of the research efforts and findings from the last few years of the issues of psychometric testing in a complex setting of longitudinal data and surveys with complex designs. Specific empirical results based on NLSCY data will be provided, including: (1) the potential for biases due to ignoring survey weights; (2) the impact of survey design on variances of parameter estimates; (3) biases in the distribution of ability predictors, and the dependence of this bias on test length. Issues requiring further research will be identified.

L'Enquête longitudinale nationale sur les enfants et les jeunes au Canada (ELNEJ) est une étude unique sur les Canadiens de la naissance à l'âge adulte et fournit une source unique de données pour l'évaluation du développement dans le contexte des enfants, y compris les divers chemins de vie d'un développement normal. L'ELNEJ est conçu pour suivre un groupe représentatif d'enfants canadiens de 0 à 25 ans, avec la collecte des données à des intervalles de deux ans. L'échantillon actuel d'enfants de l'ELNEJ est assez grand pour produire des analyses selon des cohortes, des sous-populations et des provinces. L'étude a commencée en 1994 avec un groupe d'environ 20000 enfants sélectionnés et alors âgés de 0 à 11 ans. Des informations sur leur environnement scolaire et à la maison sont rassemblées.

La théorie des réponses aux items (IRT) offre beaucoup d'avantages aux chercheurs qui doivent quantifier les capacités de lecture et d'écriture des enfants, et pour cette raison, les méthodes IRT sont adoptées pour les statistiques de l'Enquête longitudinale nationale sur les enfants et les jeunes du Canada. Les méthodes IRT ont une longue histoire dans le domaine psychométrique et fournissent une méthode basée sur des modèles pour caractériser les items testés et les capacités des sujets et pour produire des prévisions pour les capacités individuelles. Pour la plupart, les méthodes IRT supposent implicitement l'i.i.d.

des observations, de sorte que l'application de ces méthodes aux sondages complexes soulève un certain nombre de problèmes. Des questions se posent quant à l'utilisation des poids échantillonnaux, des méthodes appropriées pour prévoir les capacités individuelles (pour l'inclusion sur des jeux de données d'utilisation publique) et des méthodes appropriées pour estimer les paramètres et leurs variances.

La présentation décrit les nombreuses étapes prises par Statistiques Canada pour appliquer les techniques psychométriques pour dériver les paramètres convenablement mesurés des items dans la construction d'outils d'évaluation et pour dériver des scores appropriés de capacités pour les enfants dans le but de supporter l'analyse longitudinale. C'est également une occasion de présenter certains des efforts et des résultats de recherches des dernières années sur les problèmes des tests psychométriques dans une situation complexe de données longitudinales et de sondages avec des designs complexes. Des résultats empiriques spécifiques basés sur des données de l'ELNEJ sont fournis, incluant: (1) le risque de biais due au fait d'ignorer les poids du sondage; (2) l'impact du design du sondage sur la variance des estimés des paramètres; (3) le biais dans la distribution des prédicteurs de capacités, et la dépendance de ce biais à la durée des tests. Nous identifions également des problèmes qui nécessitent davantage de recherche.

Monday June 9 • Lundi 9 juin 11:00 • 11h00 LSC 234

Susie Fortier, Hélène Bérard, Statistics Canada/Statistique Canada

Producing historical data according to a new classification: the experience of the Monthly Wholesale and Retail Trade Survey • La conversion de données historiques selon un nouveau système de classification: le cas de l'Enquête mensuelle sur le commerce de gros et de détail

Industrial classification systems provide a conceptual framework that allows the description of economic activities. Throughout the years, Statistics Canada has used different versions of the Standard Industrial Classification (SIC) system and the North American Industry Classification System (NAICS) for industrial classification. Changes in industrial classification systems create disruptions in associated time series of population estimates. To maintain continuity, existing series under the previous classification must be converted according to the new classification. The Monthly Wholesale and Retail Trade Survey (MWRTS), a major survey conducted by Statistics Canada, is in the midst of a redesign and faces the challenges of classification conversion. The current MWRTS was developed in the late 1980's to produce sales and inventories estimates for SIC-based industrial sectors. The redesigned survey will produce NAICS-based estimates starting in 2004. This talk summarizes the options considered for backcasting the SIC-based data into NAICS-based estimates. The pros and cons of each method will be discussed and the final strategy will be presented within the specific context of the MWRTS.

Les systèmes de classification industrielle fournissent un cadre commun de concepts permettant de décrire l'activité économique. Au cours des années, Statistique Canada a utilisé différentes versions de la Classification type des industries (CTI) et, plus récemment, le Système de classification des industries de l'Amérique du Nord (SCIAN). Les changements de système de classification créent des brisures dans les séries chronologiques. Afin de préserver la continuité de ces séries, il devient nécessaire de convertir les estimations historiques obtenues sous un ancien système en estimations historiques selon la classification plus récente. Une enquête majeure de Statistique Canada, l'Enquête mensuelle sur le commerce de gros et de détail (EMCGD), subit présentement un remaniement complet

qui inclut l'utilisation de la nouvelle classification et doit relever le défi d'une telle conversion. La version actuelle de l'EMCGD a été développée à la fin des années 1980 afin de produire principalement des estimations sur les ventes et les stocks pour des secteurs industriels définis par la CTI de 1980. L'enquête remaniée produira dorénavant des estimations selon le SCIAN à partir de 2004. Diverses options ont été considérées afin de convertir les données des secteurs de la CTI en estimations pour les secteurs du SCIAN. Les avantages et inconvénients de ces options seront discutés. L'option choisie, étant donné les contraintes reliées à l'EMCGD, sera présentée.

Monday June 9 • Lundi 9 juin 11:15 • 11h15 LSC 234

Hélène Bérard, Statistics Canada/Statistique Canada

Dealing with misclassified units in repeated business surveys: the experience of the redesigned Canadian Monthly Wholesale and Retail Trade Survey • Traiter les unités classées de façon erronée dans un contexte d'enquêtes répétées: L'expérience de la nouvelle Enquête mensuelle sur le commerce de gros et de détail (EMCGD) au Canada

Stratified sampling designs are often chosen by surveys concerned with skewed business populations. The efficiency of the stratified sampling design depends largely on the quality of the stratification variables (in terms of accuracy and timeliness) and in the case of repeated surveys, their stability through time. The Monthly Wholesale and Retail Trade Survey (MWRTS) of Statistics Canada uses a stratified sampling design that can suffer from the undesirable effects of misclassified units on estimates and variances. Many innovative solutions have been developed within the sampling design, the frame and sample update processes and at the estimation stage to reduce the impact of such units for the redesigned MWRTS that will be implemented in 2004. Within the sampling design, development of an improved size measure for stratification purposes uses data from other surveys and administrative files like the Goods and Services Tax (GST) files. New frame and sample update procedures allow the monitoring and treatment of units that are no longer part of the proper size stratum in an unbiased manner. To deal with misclassified units at the estimation stage we investigated the use of post-stratification based on counts and we developed an outlier detection and treatment strategy. In this communication, each of the new measures developed to reduce the impact of misclassified units in the MWRTS is discussed in details, as well as their implementation in the context of a monthly production cycle.

Les enquêtes économiques dont les populations sont asymétriques utilisent souvent un plan d'échantillonnage stratifié. L'efficacité du plan stratifié dépend largement de la qualité des variables de stratification (dont l'exactitude et l'actualité) et, dans le cas des enquêtes répétées, de leur stabilité dans le temps. L'Enquête mensuelle sur le commerce de gros et de détail (EMCGD) de Statistique Canada utilise un plan de sondage stratifié sensible aux effets indésirables des unités classées de façon erronée sur les estimations du niveau et de la variance. Afin de réduire l'impact de ces unités dans l'EMCGD remaniée qui débutera en 2004, plusieurs solutions innovatrices ont été développées touchant le plan d'échantillonnage, les processus de mise à jour de la base de sondage et de l'échantillon et le module d'estimation. Le plan d'échantillonnage bénéficiera d'une nouvelle mesure de taille qui a été développée en utilisant des données provenant d'enquêtes et de sources administratives dont les fichiers de données sur la Taxe sur les produits et services (TPS). Une nouvelle procédure pour la mise à jour de la base de sondage et de l'échantillon permet de mieux suivre l'évolution de la mesure de taille des unités et de traiter des unités qui

ne sont plus dans la strate appropriée d'une façon non biaisée. La possibilité d'appliquer à l'estimation une post-stratification basée sur les comptes de populations a aussi été examinée et une stratégie pour la détection et le traitement des valeurs aberrantes lors de l'estimation a été mise en place. Cette communication présentera en détail chacune des mesures développées afin de réduire l'impact des unités classées de façon erronée et leur utilisation dans le cycle de la production mensuelle de l'EMCGD.

Monday June 9 • Lundi 9 juin 11:30 • 11h30 LSC 234

Richard Belcher, Statistics Canada/Statistique Canada

Application of the Hidiroglou-Berthelot method of outlier detection for periodic business surveys • L'application de la méthode de Hidiroglou et Berthelot pour la détection des valeurs aberrantes pour les enquêtes-entreprises

Detecting outliers in business surveys can be particularly difficult due to the extreme variation in the size of respondents. A well-known method for detecting outliers in periodic business surveys was created by Hidiroglou and Berthelot (1986). The strength of this method is that it allows us to include the size of the unit as being an important factor in declaring outliers. This is done via the inclusion of parameters that allow for manipulation of the size and shape of the acceptance region. Selection of appropriate values for these parameters, however, is not straightforward. This article presents a tool that can be useful in the specification of parameters for this method. We will also show an application of the method to the General Index of Financial Information, a census of Canadian corporate financial information.

La détection des valeurs aberrantes dans les enquêtes auprès des entreprises peut être particulièrement difficile à cause de la très grande variation dans la taille des répondants. Une méthode très connue pour identifier les valeurs aberrantes dans les enquêtes-entreprise périodiques a été créée par Hidiroglou et Berthelot (1986). La force de cette méthode est qu'elle nous permet d'inclure la taille de l'unité comme un facteur important dans l'identification des valeurs aberrantes. Ceci est fait par l'inclusion de paramètres qui nous permettent d'ajuster la dimension et la forme de la région d'acceptation. Cependant, la sélection des valeurs appropriées pour ces paramètres n'est pas évidente. Le présent document propose un outil qui peut être utile dans la spécification des paramètres pour cette méthode. Nous montrons aussi une application de la méthode à l'Index général des renseignements financiers, un recensement de renseignements financiers d'entreprises canadiennes.

Session/Séance 8 • NSERC Open Meeting • Rencontre avec le CRSNG

Monday June 9 • Lundi 9 juin 1:30 • 13h30 LSC 338

Statistical Sciences Grants Selection Committee • Comité de sélection des octrois des sciences statistique

Workshop: How to prepare a winning NSERC proposal • Atelier: Comment préparer une demande gagnante au CRSNG

Judie Foster, NSERC • CRSNG

Reallocations exercise: Recommendations and impact for Statistical Sciences • Exercice de réallocation: recommandations et impact pour les sciences statistiques

Jamie Stafford, University of Toronto

National Program on Complex Data Structures • Programme national sur les structures de données complexes

Representatives from NSERC and members of the Statistical Sciences Grants Selection Committee (GSC) will make a presentation to familiarize researchers with the peer review process and the way in which Grant Selection Committees function. Advice will be given on how to prepare an application. While the workshop will be most helpful to new faculty members and those preparing applications this fall, all researchers are welcome to attend. The workshop will cover topics such as discovery grants, grant selection committees, criteria for evaluation, application forms and research tools & instruments grants. Questions will be encouraged during the presentation. Following the Discovery Grant workshop, NSERC will provide results from the Third Reallocations Exercise and the impact the recommendations will have on researchers in Statistical Science. Jamie Stafford will provide information on the progress of the National Program on Complex Data Structures,which is a fully-funded initiative resulting from the Third Reallocations Exercise. The NPCDS was conceived as a model for a national network in the statistical sciences, in partnership with the mathematics institutes. The broad goal of NPCDS is to foster nationally coordinated projects with substantial interactions with the large community of scientists involved in analysis of complex data sets, and to establish a framework for national networking of research activities in the statistical community. Information on NPCDS activities (workshops, research positions) and a call for proposals (deadline is June 15th!) may be found at www.fields.utoronto.ca/programs/scientific/NPCDS.
AGENDA
*The Discovery Grants Program
* The Notification of Intent to Apply (Form 180)
* How a Grant Selection Committee Works
* The Review Criteria
* The Application Form
* The Personal Data Form
* The Research Tools & Instrument Program
* Reallocations
* National Program on Complex Data Structures (NPCDS)

Les représentants du CRSNG et des membres du Comité de sélection des octrois des sciences statistique (CSO) feront une présentation pour familiariser les chercheurs au processus d'examen par les pairs et à la manière dont les comités de sélection des octrois fonctionnent. Des conseils seront donnés pour la préparation des demandes. Même si l'atelier sera le plus utile aux nouveaux membres du corps professoral et à ceux qui préparent des applications cet automne, tous les chercheurs sont bienvenus à être présents. L'atelier couvrira des sujets tels que les octrois pour les découvertes, les comités de sélection des octrois, les critères d'évaluation, les formulaires de demande et les octrois pour les outils et les instruments de recherches. Les questions seront acceptées durant la présentation.

Après l'atelier des octrois pour les découvertes, le CRSNG fournira des résultats du Troisième exercice de redistributions et l'impact que les recommandations auront sur les chercheurs en sciences statistiques.

Jamie Stafford fournira des informations sur le Progrès du programme national sur les structures de données complexes, qui est une initiative pleinement financée qui résulte du

Troisième exercice de redistributions. Le PNSDC a été conçu comme un modèle pour un réseau national en sciences statistiques, avec le partenariat des instituts de mathématiques. Le large but du PNSDC est de stimuler les projets coordonnés nationalement avec des interactions substantielles de la communauté des scientifiques impliqués dans l'analyse de bases de données complexes, et d'établir un cadre aux réseaux nationaux des activités de recherches dans la communauté statistique. Des informations sur les activités du PNSDC (ateliers, positions de recherches) et un appel aux propositions (la date-limite est le 15 juin!) peut être trouvé à www.fields.utoronto.ca/programs/scientific/NPCDS.
ORDRE DU JOUR
**Le programme d'octrois pour les découvertes*
* *L'avis de l'intention à appliquer (formulaire 180)*
* *Comment fonctionne un Comité de sélections des octrois*
* *Les critères de revue*
* *Le formulaire de demande*
* *Le formulaire d'informations personnelles*
* *Le programme d'outils et d'instrument de recherches*
* *Redistributions*
* *Programme national sur les structures de données complexes (PNSDC)*

Session/Séance 9 • Isobel Loutit Invited Address on Business and Industrial Statistics • Présentation sur invitation Isobel Loutit sur la statistique en affaires et dans l'industrie

Monday June 9 • Lundi 9 juin 1:30 • 13h30 LSC 240

Doug Montgomery, Arizona State University

The modern practice of statistics in business and industry • La pratique moderne des statistiques en affaires et dans l'industrie

The last 20 years have seen significant advances in the use of statistical methodology in business and industry, with applications in new product design and development, optimization and control of manufacturing processes, supply chain management, as well as non-manufacturing and service industries. The field of industrial statistics has emerged as an important branch of statistical science that focuses on this application environment. Yet as applications of statistics in industry have expanded, creating many new opportunities for the modern industrial statistician, many new challenges have arisen. Some of these challenges are technical, while others have managerial and organizational aspects. Furthermore, the success of six-sigma has resulted in the role of the modern industrial statistician changing from internal consultant to team leader, facilitator, and change agent. This has resulted in concerns about training and education of industrial statisticians. This presentation focuses on some of these issues, and identifies some potential solutions.

Les 20 dernières années ont vu des avances significatives dans l'utilisation de la méthodologie statistique dans les affaires et l'industrie, avec des applications dans la conception et le développement de nouveaux produits, l'optimisation et la commande des processus de fabrication, la gestion des chaînes d'approvisionnement, aussi bien que la non-fabrication et le secteur tertiaire. Le champ des statistiques industrielles a émergé comme branche importante de la science statistique qui met l'emphase sur cet environnement d'applications. Ainsi, comme les applications des statistiques dans l'industrie ont augmenté, créant beaucoup de nouvelles occasions pour le statisticien industriel moderne, beaucoup de nouveaux

défis ont surgi. Certains de ces défis sont techniques, alors que d'autres ont des aspects de gestion et d'organisation. En outre, le succès du six-sigma a eu comme conséquence que le rôle du statisticien industriel moderne a changé du conseiller interne en chef d'équipe, en facilitateur, et agent de changement. Ceci a eu comme conséquence de créer des soucis concernant la formation et l'éducation des statisticiens industriels. Cette présentation se concentre sur certains de ces problèmes, et identifie quelques solutions potentielles.

Session/Séance 10 • Survival Analysis for Complex Data Structures • Analyse de survie pour des structures de données complexes

Monday June 9 • Lundi 9 juin 1:30 • 13h30 LSC 242

Jerry Lawless, University of Waterloo

Censoring and weighting in survival estimation from survey data • Troncation et pondération dans l'estimation de survie pour des données de sondages

Design weights, possibly adjusted for nonresponse, are used for estimation of super-population parameters from complex survey data when the survey design is nonignorable relative to the super-population model. For many settings associated with longitudinal surveys, weights must be adjusted at discrete followup or interview times in order to allow for patterns of attrition or loss to followup. The estimation of survival distributions or other event history models is often more complicated, and weights must be time-varying in order to adjust for dependent censoring. This talk will discuss followup and censoring processes, and the type of weighting needed to provide consistent estimates. Pseudo Kaplan-Meier estimation and estimates for regression models will be considered explicitly.

Les poids dans les plans d'expérience, possiblement ajustés pour la non réponse, sont utilisés pour estimer les paramètres de super-populations à partir de données complexes de sondage quand le plan d'échantillonnage est non négligeable relativement au modèle de la super-population. Pour plusieurs situations d'études longitudinales, les poids doivent être ajustés aux temps de suivi ou d'enquête discrets afin de tenir compte des modèles d'usure ou de perte au suivi. L'estimation des distributions de survie ou d'autres modèles sur l'histoire d'un événement est souvent plus compliquée et les poids doivent être variable dans le temps afin de d'ajuster la troncation dépendante. Cette présentation discutera du suivi, des processus de troncation, et du type de poids requis pour fournir des estimations convergentes. Les pseudo estimations et estimés de Kaplan-Meier pour des modèles de régression seront considérées explicitement.

Monday June 9 • Lundi 9 juin 2:00 • 14h00 LSC 242

Susana Rubin-Bleuer, Statistics Canada/Statistique Canada

Tightness of survival processes in a joint design-model space • Tension des processus de survie dans un espace conjoint

Censored failure time data can be analyzed in a unified manner using counting process theory. These methods were developed for independent identically distributed random variables (iidrv)'s. When data is obtained from complex surveys, the i.i.d. - assumption is not valid and standard results may not apply. An usual method to account for the sampling design, is to replace the corresponding statistic (e.g. the proportional hazard regressions estimator) with its respective sample estimator. Rubin-Bleuer (2001) showed that for data obtained from a probability proportional to size design, we can apply asymptotic results developed for i.i.d data to the sample-Gehan-Wilcoxon significance test for differences in

failure time distributions. Before we could extend this result to other designs and other sample-survival statistics, we need to deal with the issue of tightness. Asymptotic theory for counting processes relies on the assumption of "tightness" of the process, which for i.i.d. data, often requires the convergence in sup norm of the average of the "number of units at risk" process (Glivenko-Cantelli). However, a Glivenko-Cantelli property for the design space does not yet exist. In this paper, we examine several sampling designs and sufficient conditions for a Glivenko-Cantelli property to hold in a "joint" design-model space.

Des données de temps de bris censurées peuvent être analysées de manière unifiée en utilisant la théorie des processus de comptage. Ces méthodes ont été développées pour des variables aléatoires indépendantes et identiquement distribuées (iid). Lorsque les données sont obtenues à partir de sondages complexes, l'hypothèse i.i.d. n'est plus valide et les résultats standard ne peuvent pas s'appliquer. Une méthode habituelle qui tient compte du plan d'échantillonnage, consiste à remplacer la statistique correspondante (par exemple l'estimateur de régression proportionnel du taux de panne) par son estimateur respectif échantillonnal. Rubin-Bleuer (2001) ont prouvé que pour des données obtenues à partir d'une probabilité proportionnelle à la taille de l'échantillon, nous pouvons appliquer des résultats asymptotiques développés pour des données i.i.d au test d'échantillon de Gehan-Wilcoxon pour des différences de distributions de temps de bris. Avant de prolonger ce résultat à d'autres designs et à d'autres statistiques de survie sur des échantillons, nous devons traiter la question de la tension. La théorie asymptotique pour des processus de comptage se base sur l'hypothèse de tension du processus, qui, pour des données i.i.d., exige souvent la convergence en norme du sup de la moyenne du "nombre d'unités en danger" dans le processus (Glivenko-Cantelli). Cependant, les propriétés de Glivenko-Cantelli pour l'espace de design n'existent pas encore. Dans cette présentation, nous examinons plusieurs plans d'échantillonnage et des conditions suffisantes pour qu'une propriété de Glivenko-Cantelli tienne dans un espace de modèle-design conjoint.

Monday June 9 • Lundi 9 juin 2:30 • 14h30 LSC 242

Y. Peng, Memorial University of Newfoundland

Semiparametric cure models and some computational issues • Modèles de traitements semi-paramétriques et quelques problèmes informatiques

There is a great recent interest in semiparametric cure models for failure time data when long-term censored observations are present. In this talk, I will review the cure models and some recent semiparametric estimation methods. Several computational issues related to the semiparametric cure models will be addressed. They include baseline estimation in the cure models and a simple computational approach for the model. A simulation study of the proposed baseline estimation method is reported and the proposed computational approach for the cure model is illustrated with a real-life data set.

Nous observons actuellement un grand intérêt pour les modèles semi-paramétriques de traitement pour des données de temps de bris lorsque des observations censurées à long terme sont présentes. Dans cette présentation, nous passerons en revue les modèles de traitement et quelques méthodes récentes d'estimations semi-paramétriques. De plus, plusieurs problèmes informatiques reliés aux modèles semi paramétrique de traitement seront adressés, comme par exemple l'estimation de la ligne de base dans les modèles de traitement et une approche informatique simple pour le modèle. Une étude de simulation de la méthode proposée pour l'estimation de la ligne de base est présentée et l'approche informatique proposée pour le modèle de traitement est illustrée avec un vrai jeu de données.

Session/Séance 11 • Rank Methods for Time Series Analysis • Méthodes basées sur les rangs pour l'analyse de séries chronologiques

Monday June 9 • Lundi 9 juin 1:30 • 13h30 LSC 238

Jean-Marie DUFOUR, Université de Montréal

Finite-sample distribution-free inference in regression models under general forms of dependence • Inférence non-paramétrique exacte dans des modèles de régression avec dépendance générale

We study the construction of finite-sample distribution-free tests and confidence sets in regression models when the observations are serially dependent, according to a nonparametric dependence scheme (such as a general Markovian process). Such nonparametric models are typically analyzed using only asymptotically justified approximate methods, which can be arbitrarily unreliable. We observe first that many nonparametric test problems do not possess valid finite-sample solutions and thus constitute ill-defined statistical problems. Then, for properly formulated problems, we review several techniques allowing one to obtain provably valid inference procedures in finite samples. In particular, we show that exact inference procedures can be obtained in regressions with dependent errors when the dependence between the errors is weakly specified according to a nonparametric scheme. We stress that finite-sample procedures can be produced in many setups by combining three techniques: (1) sign transformations to eliminate serial dependence (as opposed to their usual application to allow for non-normal and/or heteroskedastic observations); (2) Monte Carlo tests to obtain finite-sample procedures based on test statistics whose cannot be derived through analytical methods; (3) projection techniques to draw inference on individual parameters in models which involve several parameters. We show that such a "cocktail" of techniques does indeed allow one to obtain finite-sample inference in regressions models with dependent errors, under weak assumptions of serial dependence (as well as nonnormality and heteroskedasticity). In particular, this setup covers general forms of GARCH dependence and stochastic volatility. Applications to time series analysis in econometrics and finance are finally discussed.

Nous étudions la construction de tests non paramétriques et de régions de confiance avec échantillons finis dans les modèles de régression lorsque les observations sont dépendantes en série, selon un schème non paramétrique de la dépendance (tel qu'un processus de Markov généralisé). De tels modèles non paramétriques sont généralement analysés en utilisant des méthodes approximatives qui sont valides seulement asymptotiquement et qui peuvent être arbitrairement incertaines. Nous observons premièrement que beaucoup de problèmes de tests non paramétriques ne possèdent pas de solutions valides pour des échantillons finis et ainsi constituent des problèmes statistiques mal définis. Deuxièmement, pour des problèmes correctement formulés, nous passons en revue plusieurs techniques permettant d'obtenir des procédures d'inférence valides avec des échantillons finis. En particulier, nous montrons que des procédures d'inférence exactes peuvent être obtenues en régressions avec des erreurs dépendantes lorsque la dépendance entre les erreurs est faiblement définie selon un schème non paramétrique. Nous montrons que des procédures d'échantillons finis peuvent être produites dans beaucoup situations en combinant trois techniques: (1) transformations par le signe pour éliminer la dépendance périodique (par opposition à leur application habituelle consistant à tenir compte des observations non-normales et/ou hétéroscédastiques); (2) tests de Monte Carlo pour obtenir des procédures

d'échantillon finis basées sur les statistiques de test qui ne peuvent pas être obtenus analytiquement; (3) techniques de projection pour faire de l'inférence sur des paramètres individuels dans des modèles qui impliquent plusieurs paramètres. Nous prouvons qu'un tel "cocktail" de techniques permet en effet d'obtenir l'inférence pour des échantillons finis dans des modèles de régressions avec des erreurs dépendantes, sous l'hypothèse de faibles dépendance périodique (ainsi que la non normalité et l'hétéroscédasticité). En particulier, cette situation comprend les formes générales de la dépendance de GARCH et de la volatilité stochastique. Des applications à l'analyse des séries chronologiques en économétrie et en finance sont finalement discutées.

Monday June 9 • Lundi 9 juin 2:00 • 14h00 LSC 238

Marc Hallin, Davy Paindaveine, Université libre de Bruxelles

Optimal rank-based procedures for testing multivariate elliptic white noise against VARMA dependence • Tests de rangs multivariés optimaux pour l'hypothèse de bruit blanc elliptique

We propose multivariate generalizations of signed-rank tests for testing elliptically symmetric white noise against VARMA serial dependence. These tests are based on Randles (1989)'s concept of interdirections and the ranks of pseudo-Mahalanobis distances, or on the lift-interdirections ranks of Oja and Paindaveine (2003). They are affine-invariant, asymptotically equivalent to a strictly distribution-free statistic, and do not require any moment conditions. Depending on the score function considered (van der Waerden, Laplace, ...), they allow for locally asymptotically maximin tests at selected densities (multivariate normal, multivariate double-exponential, ...). Local powers and asymptotic relative efficiencies with respect to the Gaussian procedure are derived. We extend to the multivariate serial context the Chernoff-Savage result, showing that classical correlogram-based procedures are uniformly dominated by the van der Waerden version of our tests, so that correlogram methods are not admissible in the Pitman sense. We also prove an extension of the celebrated Hodges-Lehmann ".864 result", providing, for any fixed space dimension k, the lower bound for the asymptotic relative efficiency of the proposed multivariate Spearman type tests with respect to the Gaussian test. These asymptotic results are confirmed by a Monte-Carlo investigation.

Nous présentons une classe de tests pour l'hypothèse de bruit blanc indépendant elliptique. Les contre-hypothèses considérées sont des modèles ARMA multivariés. Ces tests sont fondés sur le concept d'interdirections de Randles (1989) et les rangs des distances de Mahalanobis, ou sur les rangs associés aux lift-interdirections d'Oja et Paindaveine (2003). Ils généralisent les tests de rangs signés univariés, sont invariants par transformations affines, et ne demandent aucune condition de moment; ils sont également localement et asymptotiquement optimaux, au sens de Le Cam, pour peu qu'ils soient fondés sur la fonction score adéquate. Nous calculons les efficacités asymptotiques relatives des tests proposés—par rapport à la procédure paramétrique gaussienne, et aux équivalents sériels des tests de signe de Randles (1989) et des tests de type Wilcoxon de Peters et Randles (1990). Nous établissons aussi une extension multivariée du fameux résultat de Chernoff-Savage (1958) dans le cadre sériel, en montrant que la procédure classique gaussienne (un test de portemanteau multivarié fondé sur les autocorrélations croisées) est uniformément dominée par la version de van der Waerden de nos procédures. Nous présentons enfin une généralisation du célèbre résultat ".864" de Hodges et Lehmann en donnant la borne inférieure de l'efficacité relative asymptotique de la procédure de type Spearman proposée par rapport à la procédure

gaussienne en fonction de la dimension de l'espace des observations.

Monday June 9 • Lundi 9 juin 2:30 •14h30 LSC 238

Bruno Rémillard, École des hautes études commerciales, Montréal; Christian Genest, Université Laval

Tests of independence based on the empirical copula process • Tests d'indépendance basés sur le processus de copule empirique

Exploiting an overlooked observation of Blum, Kiefer & Rosenblatt, Dugué, then followed by Deheuvels, described a decomposition of empirical distribution processes into a finite sum of asymptotically mutually independent terms whose limiting distribution is simple under the hypothesis that a multivariate distribution is equal to the product of its marginals. In this work, we revisit that idea and we propose to test independence using a combination of Cramér-von Mises statistics arising from the decomposition of the empirical copula process, which involves only the ranks of the observations. The mathematical exposition, which is based on recent work of Ghoudi, Kulperger & Rémillard, allows for a simultaneous treatment of the serial and non-serial case. It is shown, among other things, that the asymptotic distribution of rank statistics based on the empirical copula process is the same in both cases, thereby shedding new light on the theory of nonparametric tests of serial dependence initiated by Hallin, Ingenbleek & Puri. A graphical device is also proposed which helps identify the dependence structure when the null hypothesis is rejected. Monte Carlo simulations are used to illustrate the power of the proposed tests under various alternatives.

Exploitant une importante remarque de Blum, Kiefer & Rosenblatt qui a été négligée par la suite, Dugué, et par la suite Deheuvels, ont décrit une décomposition d'un processus empirique en une somme finie de termes qui sont asymptotiquement indépendants et dont la limite s'exprime simplement, sous l'hypothèse que la fonction de répartition multivariée est le produit de ses marginales. Dans ce travail, nous jetons un nouveau regard sur cette idée et nous proposons de tester l'hypothèse d'indépendance en utilisant une combinaison de statistiques du type Cramér-von Mises provenant de la décomposition de la copule empirique, qui s'écrit en fonction des rangs des observations. La méthodologie, basée sur un travail récent de Ghoudi, Kulperger & Rémillard, permet le traitement simultané des cas sériels et non sériels. En outre, il est montré que la loi asymptotique de statistiques de rangs basées sur la copule empirique est la même dans les deux cas, expliquant ainsi les résultats de la théorie des tests non paramétriques dans le cas de séries chronologiques initiée par Hallin, Ingenbleek & Puri. Un outil graphique est aussi proposé afin de permettre l'identification de structures de dépendance lorsque l'hypothèse nulle est rejetée. Des simulations Monte Carlo illustrent la puissance des tests proposés sous diverses contre-hypothèses

Session/Séance 12 • Applications of Spatial Statistics • Applications des statistiques spatiales

Monday June 9 • Lundi 9 juin 1:30 • 13h30 LSC 332

Lance Waller, Traci Leong, Andrew Barclay, Emory University; Bud Howard, Palm Beach County

A spatial analysis of Sea Turtle nesting patterns in Palm Beach County, Florida • Une analyse spatiale des patrons de nidification des tortues de mer dans le comté de Palm Beach en Floride

We explore variations in the observed spatial pattern of sea turtle nesting behavior at Juno Beach, Palm Beach County, Florida for the 1998-2000 nesting seasons. Of particular interest is an assessment of possible effects due to a 990-foot fishing pier constructed in 1998-1999. The data include approximately 8,000-10,000 emergence locations per nesting season over 6 miles of beach with locations identified by global positioning system (GPS) units with sub-meter accuracy. Typical statistical analyses (Chi-square and Kruskal-Wallis tests) suggest some significant changes between years in counts of emergences for certain marked zones but do not readily identify where significant local differences occur along the beach.

We conduct a spatial analysis by estimating the density of emergences (number of emergences per unit length of beach) as a function of beach location, and compare densities of emergences between nesting seasons, densities of nesting and non-nesting emergences within each season, and densities between species (green and loggerhead) within each season. The approach reveals significant decreases in emergence density (nesting, non-nesting, and total) near the pier in the first post construction year (1999) in contrast to 1998 and 2000. The approach also reveals a possible distributional shift in nesting locations in the second year post construction even though total emergence counts are similar. Finally, the approach suggests an impact of the pier on nesting behavior, i.e. a reduced probability of nesting per emergence in the immediate vicinity of the pier.

Nous explorons les variations dans le patron spatial observé du comportement de nidification des tortues de mer sur la plage de Juno, dans le comté de Palm Beach en Floride pour les saisons de nidification 1998-2000. Un intérêt particulier est une évaluation des effets possibles d'une jetée de pêche de 990 pieds de long construite en 1998-1999. Les données incluent approximativement 8000 à 10000 endroits de sortie de l'eau par saison de nidification sur plus de 6 milles de plage, la localisation étant identifiés par un système de positionnement global (GPS) avec une précision inférieure au mètre. Les analyses statistiques typiques (Chi-carré et tests de Kruskal-Wallis) suggèrent quelques changements significatifs entre les années pour le nombre de sorties pour certaines zones, mais n'identifient pas directement où les différences locales significatives se produisent le long de la plage.

Nous conduisons une analyse spatiale en estimant la densité des sorties (nombre des sorties par unité de longueur de plage) comme une fonction de l'endroit sur la plage, et nous comparons les densités de sorties entre les saisons de nidification. Nous comparons également les densités de sorties avec nidification et sans nidification pour chaque saison et les densités entre les espèces (verte et loggerhead) pour chaque saison. L'approche indique des diminutions significatives des densités de sorties (avec nidification, sans nidification, et total) près de la jeté pour la premières années après la construction de celle-ci (1999) en contraste à 1998 et à 2000. L'approche indique également un déplacement possible de la distribution pour les endroits de nidification dans la deuxième année après la construction même si les comptes totaux de sorties sont semblables. En conclusion, l'approche suggère un impact de la jetée sur le comportement de nidification, c.-à-d. une probabilité réduite de nidification pour chaque sortie à proximité immédiate de la jetée.

Monday June 9 • Lundi 9 juin 2:00 • 14h00 LSC 332

Carol Gotway Crawford, Center for Disease Control; Linda J. Young, University of Nebraska

A geostatistical approach to combining incompatible spatial data • Une approche géostatistique pour la combinaison de données spatiales incompatibles

The widespread availability of digital spatial data and the capabilities of geographic information systems make it possible to easily synthesize spatial data from a variety of sources. More often than not, these data have been collected at different scales, and each of the scales may be different from the one of interest. In integrating all this information, the most meaningful aspect of spatial data, its support (e.g., shape and orientation), is ignored or compromised. In this presentation, an overview of several methods for predicting outcomes on different spatial scales that explicitly incorporates changes in support will be given. A geostatistical approach that has the mass-balance property and allows predictions to be adjusted for covariates will be proposed. The methods will be illustrated using low birth weight and environmental data from southeastern Georgia. Strengths and weaknesses of each method will be discussed.

La disponibilité répandue des données spatiales numériques et les possibilités des systèmes d'information géographiques rendent possible de synthétiser facilement des données spatiales de différentes sources. Très souvent, ces données ont été rassemblées sur différentes échelles, et chacune des échelles peuvent être différentes de celle qui nous intéresse. En intégrant toute cette information, l'aspect le plus significatif des données spatiales, son support (par exemple sa forme et son orientation), est ignoré ou compromis. Dans cette présentation, nous donnons une vue d'ensemble sur plusieurs méthodes de prévision de résultats avec différentes échelles spatiales, qui incorporent explicitement les changements dans le support. Nous proposons une approche géostatistique qui a la propriété masse-équilibre et qui permet aux prévisions d'être ajustées aux covariables. Les méthodes sont illustrées en utilisant des données sur le faible poids à la naissance et les conditions environnementales en Géorgie du sud-est. Nous discutons aussi des forces et des faiblesses de chacune des méthodes.

Monday June 9 • Lundi 9 juin 2:30 • 14h30 LSC 332

Richard Hoskins, Washington State Department of Public Health and Epidemiology

A comparison of boundary detection, spatial scan and cluster detection methods applied to infant deaths in Washington State • Comparaison des méthodes de détection des frontières, de scan spatial et de détection des grappes appliquées à la mortalité infantile dans l'État de Washinton

Fortunately the rate of infant death continues to fall in Washington State, especially sudden infant death syndrome (SIDS) because of a simple intervention. However, no epidemiological studies have been done with a large number of infant deaths over an extended time period or area which focus on the geographic distribution. Classic public health assessment is almost always restricted to the presentation of rates, some trends, with no consideration for spatial distribution of disease or the analysis of clusters or temporal geographic trends. A descriptive spatial analysis over a large area and time period might lead to the development of interventions or geographically targeted interventions that could not be anticipated otherwise.

The purpose of this study to carry out a very complete descriptive study using a variety of tools, some of which are new to spatial analysis in public health. The birth certificates for all live births and death certificates for all infants (less than age 1) from 1990 until 2001 were geocoded to longitude and latitude. The infant death certificates were linked to the birth certificate with subsequent determination of rates and SMRs (standard mortality ratio). The SMRs were adjusted using empirical Bayes smoothing. Then using newly available boundary detection software, optimal areal boundaries and spatial zones

of rapid change (difference boundaries), were determined. Boundary detection and overlap methods identify the gradients of change in data, say, where health outcomes or possible explanatory variables change rapidly. Boundary overlap methods test the association between boundaries in separate variables allowing a description of infant death with respect to ecological covariates such as socioeconomic variables. Then local spatial and temporal cluster detection techniques including Anselin's Local Moran, Kulldorff's Spatial Scan test and others were applied. The result is an infant death "atlas" for the entire state over a 12 year period with the elucidation of potential clusters and regions of both high and low rates which can be at least empirically linked to socioeconomic information in the 1990 and 2000 Census. Rates and clusters were determined for several different causes of death. These areas were further characterized with respect to medical complications of the mother, the pre- and perinatal period, and the delivery. The various spatial analytical methods had the expected concordance but sometimes differed unexpectedly. Although public health interventions are often tailored using the information from assessment studies, this one may allow interventions to be targeted to specific regions or even neighborhoods.

Heureusement, le taux de mortalité infantile continue à baisser dans l'État de Washington, en particulier le syndrome de mort infantile soudaine (SIDS) grâce à une intervention simple. Cependant, aucune étude épidémiologique n'a été effectuée sur un grand nombre de décès infantiles et sur une longue période de temps ou une grande surface avec un regard particulier sur la distribution géographique. Les enquêtes sur la santé publique classiques sont presque toujours limitées à la présentation de taux et de certaines tendances sans considérations particulière sur la distribution spatiale de la maladie, d'analyse en grappes ou de tendances géographiques temporelles. Une analyse descriptive spatiale sur une grande période de temps et sur une grande surface pourrait mener au développement d'interventions globales ou bien d'interventions sur des secteurs géographiques spécifiques qui ne pourraient pas être anticipé autrement.

Le but de cette étude est d'effectuer une étude descriptive très complète en utilisant une variété d'outils, dont certains sont nouveaux à l'analyse spatiale sur la santé publique. Les actes de naissance pour chaque naissances vivantes et les certificats de décès pour toutes les morts en bas âge (moins qu'un an) de 1990 à jusqu'à 2001 sont codés géographiquement par leur longitude et leur latitude. Les certificats de décès infantiles sont liés à l'acte de naissance avec la détermination d'obtenir des taux et des RSM (rapport standard de mortalité). Les RSM sont ajustés en utilisant le lissage de Bayes empirique. Ensuite, en utilisant un nouveau logiciel de détection des frontières, des frontières régionales optimales et des zones spatiales de changement rapide (frontières de différence) sont déterminées. Les méthodes de détection et de chevauchement des frontières identifient les gradients de changements dans les données, par exemple, où les résultats de santé ou les variables explicatives changent rapidement. Les méthodes de chevauchement de frontières testent l'association entre les frontières dans les variables séparées, permettant une description de la mortalité infantile par rapport aux covariables écologiques telles les variables socio-économiques. Puis, des techniques spatiales locales et temporelles de détection des grappes comme le test d'Anselin Local Moran, le scan spacial de Kulldorff et d'autres sont appliquées. Le résultat est un " atlas " de la mortalité infantile pour l'ensemble de l'État sur une période de 12 ans avec élucidation de grappes et de régions potentielles de taux faibles et élevés qui peuvent être au moins empiriquement liés à l'information socio-économique des recensements de 1990 et de 2000. Des taux et des grappes sont déterminés pour plusieurs causes de mortalité. Ces secteurs sont caractérisés par rapport aux complications médicales de la mère, la période

pré-natale et périnatale et l'accouchement. Les diverses méthodes d'analyse spatiale ont la concordance espérée, mais diffèrent parfois de manière inattendue. Même si les interventions en santé publique sont souvent préparées en utilisant l'information des enquêtes, celle-ci peut permettre à aux interventions de viser des régions spécifiques ou même des quartiers.

Session/Séance 13 • Biostatistics Contributed Session I: Estimation and Testing in Biostatistics • Biostatistique I: Estimation et tests d'hypothèses en biostatistique

Monday June 9 • Lundi 9 juin 1:30 • 13h30 LSC 234

Abbas Khalili, University of Waterloo; Noel Cadigan, Department of Fisheries and Oceans

Inference for sequential population analysis using penalized likelihood methods • Inférence pour l'analyse séquentielle de la population (ASP) en utilisant des méthodes de vraisemblance pénalisée

Effective fisheries management decisions and biological investigations of fish populations rely on information about fish stocks. One source of information about fish stocks is the data available from trawl surveys and commercial catches. The data quite often consist of a time series of annual catches and relative abundance indices such as average catch per tow. Mathematical and statistical models are often used to analyze the fishery data. Virtual or sequential population analysis (VPA or SPS) is one of the commonly used models in studying the dynamic of fish populations. An SPA model connects historical stock sizes and sequentially estimates fish population abundance for all years under consideration. Statistically, an SPA model is in the framework of semi-parametric regression models. In this study, we investigate statistical properties of the point estimation of fish population abundance based on an SPA model . We consider penalized likelihood methods. Further, we study the coverage properties of confidence intervals for the fish population abundance.

Les décisions efficaces de gestion des pêches et les investigations biologiques sur les populations de poissons se basent sur d l'information sur les stocks halieutiques. Certaines de ces sources d'informations proviennent des données fournies par les sondages de chalut et les prises commerciales. Les données consistent souvent d'une série chronologique des prises annuelles et d'indices d'abondance relative tels que la prise moyenne par traîné. Des modèles mathématiques et statistiques sont souvent utilisés pour analyser les données des pêches. L'analyse virtuelle ou séquentielle de la population (VPA ou ASP) est un des modèles généralement utilisés pour étudier la dynamique des populations de poissons. Un modèle d'analyse séquentielle de la population relie les tailles historiques de stocks et l'estimation séquentielle de l'abondance de la population de poissons pour toutes les années à l'étude. Statistiquement, un modèle d'analyse séquentielle de la population fait partie du cadre des modèles semi-paramétriques de régression. Dans cette étude, nous étudions les propriétés statistiques de l'estimation ponctuelle de l'abondance de la population des poissons basée sur un modèle ASP. Nous considérons des méthodes de vraisemblance pénalisée. De plus, nous étudions les propriétés de couverture des intervalles de confiance pour l'abondance de population de poissons.

Monday June 9 • Lundi 9 juin 1:45 • 13h45 LSC 234

Andrea Benedetti, Michal Abrahamowicz, McGill University & Montreal General Hospital

Smoothing parameter selection and impact on inference in generalized additive models • La sélection des paramètres de lissage et son impact sur l'inférence des modèles additifs généralisés

Generalized additive models (GAM) are a non-parametric flexible modelling tool which allows the user to model a variety of non-linear dose-response curves without imposing a priori assumptions about the functional form of the relationship. In GAM, the extent of smoothing is controlled by the user-defined degrees of freedom (df). Fixing the df a priori keeps statistical inference intact. However, choosing too few df may result in missed non-linearities of higher degree. Conversely, choosing too many df may result in too bumpy a curve and reduced power. Among a posteriori approaches, graphical selection methods are subjective and hard to replicate across studies. Automatic selection methods, although objective and less likely to miss higher degree non-linearities, complicate statistical inference. In general, in non-parametric modelling, the problem is that while a posteriori model selection improves the accuracy of the estimated curve, at the same time it may invalidate standard statistical inference about the estimates. When fitting non-parametric models that require a smoothing parameter, such as smoothing splines or loess, a common approach is to vary the df and then choose the best-fitting model based on e.g. the minimum AIC as the final model. We investigated the impact of this approach on statistical inference, through simulations. In particular, we estimated the type I error of the tests of (i) no association, and (ii) linearity, conditional upon the optimal df selected. Overall empirical type I error rates of the conditional tests were two to three times greater than nominal levels. We calculated new critical values, from the empirical distribution of the test statistic, which resulted in type I error of about 5%.

Les modèles additifs généralisés (GAM) sont des outils de modélisation non paramétrique flexibles qui permettent à l'utilisateur de modéliser une variété de courbes dose-réponse non linéaires sans imposer des hypothèses a priori sur la forme fonctionnelle de la relation. Dans les modèles additifs généralisés, l'ampleur du lissage est contrôlée par les degrés de liberté (dl) définis par l'utilisateur. Fixer le nombre de degré de liberté a priori laisse intacte l'inférence statistique. Cependant, choisir un nombre de degré de liberté trop petit peut avoir comme conséquence de ne pas voir des non-linéarités à un degré plus élevé. Réciproquement, choisir trop de degrés de liberté peut avoir comme conséquence une courbe trop inégale et ainsi réduire la puissance. Parmi les approches a posteriori, les méthodes de sélection graphiques sont subjectives et difficiles à refaire à travers les études. Les méthodes de sélection automatique, bien qu'objectives et moins sujettes à manquer des non-linéarités à des degrés plus élevés, rendent l'inférence statistique plus compliquée. En général, en modélisation non paramétrique, le problème est que, même si la sélection du modèle a posteriori améliore la précision de la courbe estimée, cela peut en même temps rendre invalide l'inférence statistique standard sur les estimations. Lorsqu'on ajuste des modèles non paramétriques qui demandent un paramètre de lissage, tel un spline de lissage ou loess, une approche utilisée souvent consiste à faire varier le degré de liberté et à choisir le modèle qui s' ajuste le mieux selon un critère tel le AIC. Nous avons étudié l'impact de cette approche sur l'inférence statistique à l'aide de simulations. En particulier, nous avons estimé l'erreur de type I de tests (i) d' association et, (ii) de linéarité, conditionnellement au degré de liberté optimal choisi. Globalement, les taux d'erreurs de type I empiriques des tests conditionnels sont de deux à trois fois plus grand que les niveaux nominaux. Nous avons calculé de nouveaux points critiques, à partir de la distribution empirique de la statistique de test. Les erreurs de type I résultantes sont alors près de 5%.

Monday June 9 • Lundi 9 juin 2:00 • 14h00 LSC 234

Keyue Ding, W. J. Hall, Queen's University

Sequential tests and estimates after overrunning based on p-value combination • Tests séquentiels et estimés après renvoi basé sur des combinaisons de seuils expérimentaux

Often in sequential trials, additional data become available after a stopping boundary has been reached. A method of incorporating such information from overrunning is developed, based on the weighted Liptak method of combining p-values. This yields a combined p-value for the primary test and a median-unbiased estimate and confidence bounds for the parameter under test. When the amount of overrunning information is proportional to the amount available upon terminating the sequential test, exact inference methods are provided; otherwise, approximate methods are given and evaluated. The context is that of observing a Brownian motion with drift, with either linear stopping boundaries in continuous time or discrete-time group sequential boundaries. The method is compared with other available methods, and is exemplified with data from a sequential clinical trial.

Souvent dans des épreuves séquentielles, des données additionnelles deviennent disponibles après qu'une frontière d'arrêt ait été atteinte. Une méthode d'incorporer une telle information sans renvoi est développée, basé sur la méthode pondérée de Liptak qui consiste à combiner des seuils expérimentaux. Ceci donne un seuil expérimental combiné pour le test primaire, une estimation médiane sans biais et un intervalle de confiance pour le paramètre testé. Quand la quantité d'information qui dépasse la frontière d'arrêt est proportionnelle à la quantité disponible lors de la fin du test séquentiel, des méthodes exactes d'inférence sont montrés; si ce n'est pas le cas, nous présentons et évaluons des méthodes approximatives. Le contexte est celui d'observer un mouvement brownien avec dérive, avec des frontières d'arrêt linéaires à temps continu ou discret dans les groupes de frontière séquentielle. La méthode est comparée à d'autres méthodes disponibles, et est exemplifiée avec des données d'une épreuve clinique séquentielle.

Monday June 9 • Lundi 9 juin 2:15 • 14h15 LSC 234

Yang Zhao, J.F. Lawless, D.L. McLeish, University of Waterloo

Efficient estimation in regression analysis with missing data in two-phase sampling designs • Estimation efficace en analyse de régression avec des données manquantes pour des designs d'échantillonnage à deux étapes

Regression analysis that involves incomplete observations can utilize auxiliary information to provide more efficient estimates of regression parameters. In addition, many two-phase studies are designed to measure some variables only in a random sub-sample of all subjects. We will study the asymptotic relative efficiency of the maximum likelihood estimator for normal linear models in these situations. Then we will discuss optimal design problems in two-phase sampling using the results. Some simulation results on more general models will be provided.

L'analyse de régression avec des observations manquantes peut utiliser de l'information auxiliaire pour fournir des estimations plus efficaces des paramètres de régression. De plus, beaucoup d'études en deux étapes sont conçues pour mesurer des variables seulement pour un sous-échantillon aléatoire de tous les sujets. Nous étudions l'efficacité relative asymptotique de l'estimateur du maximum de vraisemblance pour les modèles linéaires normaux dans ces situations. Ensuite, nous discutons des problèmes de designs optimaux

dans l'échantillonnage en deux étapes en utilisant ces résultats. Nous présentons également quelques résultats de simulation sur des modèles plus généraux.

Monday June 9 • Lundi 9 juin 2:30 • 14h30 LSC 234

François Bellavance, HEC Montréal and Centre for Research on Transportation; Mustapha Bourhattas, Stéphane Messier, CRT; Sophie Lapierre, École Polythecnique & CRT; Claire Laberge-Nadeau, Université de Montréal & CRT

Misclassification bias in the case-crossover design applied to wireless telephones and the risk of road crashes • Le devis cas chassé-croisé et le biais de mauvaise classification dans l'estimation du risque d'accident en utilisant le téléphone mobile au volant

The case-crossover design was proposed a decade ago to avoid control selection bias in a study to estimate the risk of myocardial infarction after exposure to an intermittent event such as a physical exertion. More recently, this design was used to study the association between wireless telephones calls while driving and motor vehicle collisions. One important possible source of bias in this latter study is the misclassification of phone calls due to reporting errors in the exact time of the collision. Indeed, the time written in the police report is often a multiple of 5 minutes and tends to be larger than the exact time of the collision. Results of simulation studies showed that the case-crossover design gives unbiased estimates of the relative risk when there is no misclassification bias and the hazard interval is less than the average duration of phone calls. However, when we introduced random errors between the exact time of the collision and the time in the police report, we obtained estimates that were up to three times larger than the true relative risk. The bias due to exposure misclassification was larger for smaller values of the true relative risk.

Le devis cas chassé-croisé a été proposé il y a une dizaine d'années dans le but d'éviter un biais de sélection dans une étude ayant pour objectif d'estimer le risque d'un infarctus du myocarde après une exposition à un événement intermittent comme un effort physique. Plus récemment, ce devis expérimental a été utilisé pour étudier l'association entre l'utilisation du téléphone mobile au volant et les accidents de la route. Une importante source de biais possible dans cette dernière étude est la mauvaise classification d'appels téléphoniques à cause de l'inexactitude de l'heure des accidents dans les rapports de police. En effet, l'heure inscrite dans le rapport de police est souvent un multiple de cinq minutes et a tendance a être plus grande que l'heure exacte de la survenue de l'accident. Les résultats de simulations montrent que le devis cas chassé-croisé donne une estimation sans biais du risque relatif lorsqu'il n'y a aucun biais de mauvaise classification et que la période de risque considérée est moindre que la durée moyenne des appels téléphoniques. Par contre, lorsque nous introduisons une erreur aléatoire entre l'heure exacte de l'accident et l'heure inscrite dans le rapport de police, nous obtenons des estimations du risque relatif qui sont jusqu'à trois fois plus grandes que le vrai risque. Le biais occasionné par une mauvaise classification d'appels était plus grand pour les valeurs de risque plus petites.

Monday June 9 • Lundi 9 juin 2:45 • 14h45 LSC 234

Karelyn Davis, Chu-In Charles Lee, Memorial University of Newfoundland; Jianan Peng, Acadia University

Step-down testing procedure for dose-response studies • Procédure de test step-down pour des études dose-réponse

Scientific experiments are often concerned with the comparison of several treatment means with a control mean. In particular, such multiple comparisons arise in biopharmaceutical

studies in which it is desirable to give the inferences in a specified order and failure to achieve the desired inference at any step renders subsequent comparisons unnecessary. Researchers in dose-response studies desire such a method to not declare a lower dose to be efficacious if it does not declare a higher dose to be efficacious.

An important and practical dosing quantity is the minimum effective dose (MED). The MED is defined as the minimum dose such that the mean response at that dose is significantly better than the mean response of the control by a practical significant difference. In this talk, confidence lower bounds are developed for the aforementioned difference. The approaches of previous authors in relation to estimating the MED will be examined and an innovative approach using Kuhn-Tucker conditions to evaluate the optimal lower bound is derived.

Les expériences scientifiques sont souvent concernées par la comparaison des moyennes de plusieurs traitements avec une moyenne témoin. En particulier, de telles comparaisons multiples surgissent dans les études bio-pharmaceutiques dans lesquelles il est souhaitable de faire les inférences dans un ordre spécifié et le fait de ne pas réaliser l'inférence désirée à une certaine étape rend les comparaisons ultérieures inutiles. Les chercheurs dans les études de dose-réponse désirent qu'une telle méthode ne déclare pas une dose inférieure efficace si elle ne fait pas de même pour une dose plus élevée.

Une quantité de dosage importante et pratique est la dose minimum efficace (DME). La DME est définie comme étant la dose minimale telle que la réponse moyenne à cette dose est sensiblement meilleure que la réponse moyenne témoin par une différence significative pratique. Dans cette présentation, les limites inférieures de confiance sont développées pour la différence mentionnée ci-dessus. Les approches des auteurs précédents par rapport à l'estimation de la DME seront examinées et nous dérivons une approche innovatrice en utilisant des conditions de Kuhn-Tucker pour évaluer la limite inférieure optimale.

Session/Séance 14 • Statistics and Information Complexity of Probability Models • Statistique et complexité de l'information des modèles probabilistes

Monday June 9 • Lundi 9 juin 3:30 • 15h30 LSC 240

V. Koltchinskii, University of New Mexico

Complexities and margins in binary classification problems • Les complexités et les marges dans les problèmes de classification binaire

We discuss various complexity measures of function classes and of individual functions that have been frequently used in machine learning, especially in large margin methods of binary classification, such as boosting and support vector machines. We show that these data-dependent complexities can be used together with classification margins to provide tight probabilistic upper bounds on generalization error of classification algorithms. The proof of these bounds is rather involved and is based on various tools of modern probability (concentration inequalities for product measures, methods of Gaussian and empirical processes, etc.). The bounds provide a partial explanation of generalization performance of the large margin classification algorithms and suggest possible ways to enhance this performance.

Nous discutons de diverses mesures de complexité de classes de fonctions et de fonctions individuelles qui sont fréquemment utilisées dans l'apprentissage automatique, particulièrement pour les grandes méthodes classification binaire à grandes marges, telles que

le boosting et les machines à vecteurs de supports. Nous montrons que ces complexités données-dépendantes peuvent être utilisées avec les marges de classification pour fournir des bornes supérieures probabilistes étroites sur la généralisation de l'erreur des algorithmes de classification. La preuve de ces bornes supérieures est assez complexe et est basée sur divers outils de la probabilité moderne (inégalités de concentration pour des mesures de produit, méthodes de processus gaussiens et processus empiriques, etc.). Les bornes fournissent une explication partielle sur la généralisation de la performance d'algorithmes de classification de grande marge et suggèrent des manières possibles d'améliorer cette performance.

Monday June 9 • Lundi 9 juin 4:00 • 16h00 LSC 240

Yuri Golubev, University of Marseilles

Oracle inequalities and estimation of sparse vectors • Inégalités d'oracle et estimation de vecteurs "sparse"

The problem of sparse vectors recovering plays an important role in statistical analysis. In this talk I would like to discuss some statistical aspects of this problem for the white Gaussian noise model. It is assumed that the size of the model is large whereas the sparsity of the vector is small but unknown. It is well known that the model selection approach is a fruitful statistical method of recovering sparse vectors in this situation. The main idea of this approach consists of choosing the best estimator from a given family of linear estimators. Since the rule of discrimination between estimators is based on the penalized empirical risk, the performance of the method crucially depends on the chosen penalization. Therefore, in order to obtain an appropriate penalty, we need tight upper bounds for the risk of the method. In some special cases one can derive good upper bounds, but unfortunately, in general case these upper bounds are not yet known. The main difficulty is related to the fact that the family of all linear estimators is very large. In order to overcome this difficulty we consider two methods of estimation: plug-in methods and aggregation of hard thresholding smoothers. The idea behind this approach is very simple: to replace the huge family of linear estimators used in the model selection method by the very small family of hard thresholding rules. This allow us to derive oracle type upper bounds for the quadratic risk, and minimizing these bounds, to obtain almost optimal penalizations. It is shown that nearly optimal penalizations for the aggregation of hard thresholding smoothers are data dependent. The aggregation method has good statistical properties over-performing theoretically plug-in methods. In numerical experiments it works even better then the model selection rule with the optimal penalty.

Le problème de l'estimation des vecteurs "sparse" joue un rôle important dans l'analyse statistique. Dans cet exposé je voudrais discuter de certains aspects statistiques de ce problème pour le modèle du bruit blanc gaussien. On suppose que la dimension du modèle est grande tandis que la "sparsité" du vecteur est petite mais inconnue. Il est très bien connu que la sélection de modèle est une méthode fructueuse pour l'estimation des vecteurs "sparse" dans cette situation. L'idée générale de cette approche consiste à choisir le meilleur estimateur dans une famille donnée d'estimateurs linéaires. Puisque la règle de classification des estimateurs se base sur le risque empirique pénalisé, la performance de la méthode dépend fortement de la pénalité choisie. C'est pourquoi pour obtenir une pénalisation raisonnable on a besoin de bornes supérieures précises pour le risque. Dans certains cas particuliers on peut les trouver, mais malheureusement dans le cas général, les bornes précises ne sont pas encore connues. La difficulté principale est liée au fait que la

famille de tous les estimateurs linéaires est trop grande. Pour éviter cet inconvénient nous considérons deux méthodes d'estimation: méthodes de plug-in et agrégation d'estimateurs par seuillage dur. L'idée de cette approche est simple: remplacer la grande famille des estimateurs linéaires dans la méthode de sélection modèle par la très petite famille des estimateurs par seuillage dur. Cela nous permet d'obtenir des bornes supérieurs de type oracle et de trouver des pénalisations presque optimales en minimisant ces bornes. Il est démontré que la pénalisation presque optimale pour l'agrégation des estimateurs par seuillage dur dépend des données. La méthode d'agrégation a des propriétés statistiques remarquables en dominant les méthodes de plug-in. Dans les simulations elle fonctionne même mieux que la sélection de modèle avec pénalité optimale.

Session/Séance 15 • Data Mining • Fouille de données

Monday June 9 • Lundi 9 juin **3:30 • 15h30** **LSC 242**

Ruben Zamar, University of British Columbia

Data mining, data quality and robustness • Fouille de données, qualité des données et robustesse

Many data mining applications utilize statistical procedures such as regression, multivariate location and scatter matrices, principal components, etc. which are part of automatic or semiautomatic knowledge discovery algorithms. There are at least two good reasons for using robust estimates in such applications. First, since the statistical procedures are part of a larger computer aided operation, there is not much opportunity for careful scrutiny of the data. Hence we need methods that are insensitive to outliers and other data anomalies. Second, data mining applications usually involve very large datasets, rendering "statistical efficiency" consideration rather unimportant. When data quantity becomes abundant, data quality remains as the main statistical concern. This talk will address two main issues. Unfortunately, standard robust procedures are very computer intensive and do not scale well to dimensions and sample sizes usually encountered in data mining application. There is a need, then, for feasible robust statistical methods. This is the first issue addressed here. The second issue is perhaps more interesting. Standard statistical models used to represent "contaminated data" do not scale well to high dimensional settings. Hence, there is a need then for flexible "contamination models" to guide the theoretical study and the empirical testing of robust procedures.

Beaucoup d'applications en fouille de données utilisent des procédures statistiques telles que la régression, la position multivariée et les matrices de dispersion et l'analyse en composantes principales, qui font partie d'un groupe d'algorithmes automatiques ou semi-automatiques de découverte de connaissance. Au moins deux raisons justifient l'utilisation d'estimations robustes dans de telles applications. D'abord, puisque les procédures statistiques font partie d'une plus grande opération assistée par ordinateur, cela ne laisse pas beaucoup d'opportunité pour examiner minutieusement les données. Par conséquent, nous avons besoin de méthodes peu sensibles aux valeurs aberrantes et à d'autres anomalies dans le jeux de données. Deuxièmement, les applications de la fouille de données impliquent habituellement des jeux de données très grands, rendant les considérations "d'efficacité statistique" peu importante. Lorsque la quantité de données devient très grande, la qualité des données demeure le principal intérêt. Deux questions seront abordées dans cette présentation. La première fait état des procédures robustes standards demandent beaucoup de temps de calcul et ne sont pas très bien adaptés à la taille des l'échantillons et

à la dimension habituellement rencontrés en fouille de données. Nous avons donc besoin de méthodes statistiques robustes faisables lorsque les échantillons sont très grands. La deuxième question est sans doute plus intéressante. Les modèles statistiques standards utilisés pour représenter des données " contaminées " ne se généralise pas très bien aux situations avec de grande dimensions. Par conséquent, nous avons besoin de modèles de contamination flexibles pour guider l'étude théorique et les tests empiriques des procédures robustes.

Monday June 9 • Lundi 9 juin 4:00 • 16h00 LSC 242

Hugh Chipman, University of Waterloo; Edward I. George, University of Pennsylvania; Robert E. McCulloch, University of Chicago

Boosting Bayesian tree models • Application du boosting aux arbres bayésien

Boosting is a popular technique for enhancing the predictive accuracy of models such as trees. The boosting algorithm constructs a sum of trees, with each individual tree making a small contribution to the overall fit. Inspired by this, we consider a boosting-like modification of MCMC, fitting an additive model with trees as each additive component. This yields performance competitive with boosting, but it also gives more: some inference for the resultant model. A posterior distribution on the both the number of trees needed and the predictions is available at no extra computational cost.

Le boosting est une technique populaire pour améliorer l'exactitude de prédiction de modèles tels les arbres. L'algorithme boosting construit une somme d'arbres, chacun d'eux apportant une petite contribution à l'ajustement global. En nous inspirant de ceci, nous considérons une modification de type boosting aux modèles MCMC, ajustant un modèle additif avec des arbres pour chaque composante additive. Cette modification donne des résultats comparables à ceux du boosting et permet également de faire de l'inférence pour le modèle résultant. De plus, une distribution a posteriori sur le nombre d'arbres requis et les prévisions est disponible sans augmenter le temps de calcul.

Monday June 9 • Lundi 9 juin 4:30 • 16h30 LSC 242

Wayne Oldford, University of Waterloo

Structuring interactive cluster analysis • Structurer l'analyse de groupement interactive

The problem of cluster analysis, or finding groups in data, may be inherently ill-posed; hence the multitude of different methods which purport to solve "the" problem. Computational resources are often spent on increasingly complex algorithms or on dealing with ever larger datasets. In this talk, a different approach is taken. Instead of a single new method, existing methodologies (algorithmic and graphical) are assumed and the problem of how to integrate them in a highly interactive software environment is considered.

The methodological approach is to step back and consider the problem as one of looking at partitions of data as the primary objects of interest and so consider how one might navigate the space of all possible partitions. Computational resources are dedicated toward making this navigation interactive and intuitive. Proposed interaction and navigational tools will be discussed and some demonstrated via a prototype implementation in Quail (http://www.stats.uwaterloo.ca/Quail).

Le problème de l'analyse de groupement, ou bien de trouver des groupes à l'intérieur d'un jeux de données, pourrait être mal défini; d'où la multitude de méthodes qui prétendent

résoudre le problème. Des ressources informatiques sont souvent dépensées sur des algorithmes de plus en plus complexes ou pour traiter des jeux de données toujours plus grands. Dans cette présentation, une approche différente est adoptée. Au lieu d'apporter une nouvelle méthode, nous regardons comment intégrer les méthodes existantes (algorithmiques et graphiques) dans un logiciel interactif.

L'approche méthodologique consiste à prendre du recul et à considérer le problème comme étant celui d'examiner les partitions possibles des données et de considérer comment nous pouvons parcourir l'espace de toutes les partitions possibles. Les ressources informatiques sont consacrées à rendre cette navigation interactive et intuitive. L'interaction proposée et les outils de navigations seront discutés et certains seront démontrés par l'intermédiaire d'une implémentation prototype sur Quail (http://www.stats.uwaterloo.ca /Quail).

Session/Séance 16 • Sequential Methods • Méthodes séquentielles

Monday June 9 • Lundi 9 juin 3:30 • 15h30 LSC 238

Richard Cook, University of Waterloo

Robust methods for monitoring trials with multiple treatment periods and recurrent events • Méthodes robustes pour surveiller des essais avec plusieurs périodes de traitement et avec événements récurrents

Robust methods for the analysis of recurrent events have been developed based on Poisson estimating equations (Lawless and Nadeau, 1995). Here alternative robust methods are considered based on binomial or multinomial models estimating functions motivated by suitably conditioning with the class of mixed Poisson models. This approach is feasible provided trials are designed with baseline periods of observation or patients' are observed during successive treatment periods (Wei, 2002). Here we review this approach to testing for treatment effects based on recurrent events when baseline periods of observation are available. Extensions are then developed to facilitate interim monitoring. Simulation studies are reported on to assess the frequency properties of this monitoring scheme and the performance of this sequential design is compared with that described in Cook and Lawless (1997). An application is provided for illustrative purposes.

Des méthodes robustes pour l'analyse des événements récurrents ont été développées basées sur des équations d'estimations de Poisson (Lawless et Nadeau, 1995). Ici, des méthodes robustes alternatives sont considérées basées sur des fonctions d'estimations de modèles binomiales ou multinomiales obtenues en conditionnant convenablement avec la classe des modèles mélangés de Poisson. Cette approche est faisable lorsque des essais planifiés avec des périodes d'observation de ligne de base sont observés pendant des périodes successives de traitement (Wei, 2002). Dans cette présentation, nous passons en revue cette approche à tester des effets de traitement basés sur des événements récurrents quand les périodes de ligne de base de l'observation sont disponibles. Des prolongements sont alors développés pour faciliter la surveillance d'intérim. Des études de simulation sont présentées pour évaluer les propriétés de fréquence de ce patron de surveillance et la performance de ce design séquentiel est comparée à celle décrite dans Cook et Lawless (1997). Nous illustrons la méthode par une application.

Monday June 9 • Lundi 9 juin 4:00 • 16h00 LSC 238

John Petkau, University of British Columbia

A simple model for drug screening programs • Un modèle simple pour les programmes de dépistage des drogues

A scenario involving a large, and continuously augmented, number of agents available for testing is relevant to designing both drug screening programs and phase II clinical trials. The objective is to identify promising agents for further study, while eliminating less promising agents with a minimum of testing. Suppose testing an agent corresponds to observing a segment of the path of a Wiener process with unknown drift and known variance. For a sequence of drift parameters with a given prior distribution and specified reward functions, the objective is a policy maximizing the infinite horizon expected average reward per unit time. The simple version of this problem with a two-point prior can be solved explicitly and provides some general insight. We suppose a reward is received whenever the final decision concerning an agent is taken, and costs are incurred for inappropriate decisions. A direct approach yields a description of the optimal fixed sample size procedure and an indirect method based on a potential reward function yields the corresponding results for sequential procedures. The relative performance of these procedures is examined.

Un scénario impliquant un grand nombre de sujets et qui augmente avec le temps est approprié pour concevoir des programmes de dépistage des drogues et des essais cliniques de phase II. L'objectif est d'identifier les sujets les plus prometteurs pour des études ultérieures, tout en éliminant ceux qui sont le moins prometteurs avec un minimum de tests. Supposons qu'examiner un sujet correspond à observer un segment de chemin d'un processus de Wiener avec dérive inconnue et variance connue. Pour une séquence de paramètres de dérive avec une distribution a priori donnée et des fonctions de récompense fixée, l'objectif est de déterminer une politique qui maximise la récompense moyenne prévue à l'horizon infini par unité de temps. La version simple de ce problème avec une fonction a priori à deux points peut être résolue explicitement et fournit des pistes de solutions générales. Nous supposons qu'une récompense est reçus toutes les fois qu'une décision finale concernant un sujet est prise, et des coûts sont encourus pour une mauvaise décision. Une approche directe donne une description de la procédure optimale de taille d'échantillon fixe et une méthode indirecte basée sur une fonction de récompense potentielle donne les résultats correspondants pour des procédures séquentielles. Nous examinons la performance relative de ces procédures.

Monday June 9 • Lundi 9 juin 4:30 • 16h30 LSC 238

Edit Gombay, University of Alberta

Sequential testing strategies • Stratégies pour effectuer des tests séquentiels

We briefly survey the main methods currently used for testing hypotheses sequentially as the data arrive. These tests have applications in quality control, medical trials, environmental studies. A new class of tests will be introduced and some comparison will be made with existing methods. The tests will include parametric models when nuisance parameters are present and nonparametric models with very weak assumptions on the underlying distributions. The algorithms can be classified also as sequential tests and sequential change detection methods.

Nous passons brièvement en revue les principales méthodes utilisées pour tester séquentiellement des hypothèses à mesure que les données deviennent disponibles. Ces tests sont applicables dans les processus de contrôle de la qualité, lors d'essais cliniques et les études environnementales. Une nouvelle classe de test sera présentée et des comparaisons seront faites par rapport aux méthodes existantes. Les tests incluront les modèles paramétriques avec des paramètres de nuisances et des modèles non paramétriques ne nécessitant que peu

d'hypothèses concernant les distributions sous-jacentes. Les algorithmes peuvent également être classifiés en tant que des tests séquentiels et des méthodes de détection de changement séquentiel.

Session/Séance 17 • Survey Methods Contributed Session II: Applications of Administrative Data • Méthodes d'enquête II: Applications avec des données administratives

Monday June 9 • Lundi 9 juin 3:30 • 15h30 LSC 338

Richard Dorsett, Policy Studies Institute

Refreshment samples, matching and attrition bias • Échantillon de remplacement, association et biais d'attrition

Attrition in longitudinal datasets can bias inference. This paper assesses the potential for using propensity score matching to replace survey dropouts with individuals from a refreshment sample and thereby restore the representativeness of the sample. This is examined by considering longitudinal data collected over two waves of interviews with participants in a UK active labour market programme. The resulting data was characterised by significant attrition. However, combining administrative records with the survey data allows unemployment outcomes to be observed for all those responding to the first stage of survey, regardless of whether they cooperated at the second stage. This allows the success of the propensity score matching approach to be evaluated by comparing unemployment of dropouts with that of their replacements. The paper begins by applying a recently-advanced method to test the model of attrition. This shows the dropout process to be nonignorable such that the commonly-used methods of reweighting non-dropouts or of multiple imputation are ineffective in overcoming attrition bias. By constructing artificial refreshment samples, the potential of the matching approach to overcome such bias is demonstrated. The best results occur when the refreshment sample is 'targeted' in the sense that it is made up of individuals who are predisposed to dropping out. However, in the more realistic scenario where the refreshment sample is made up of those merely suspected of being more likely to drop out, representativeness is restored from the point of drop-out onwards, but not before. Hence, the approach may be most effective in restoring a longitudinal dataset to cross-sectional representativeness but may be less useful when examining individual trends over a period that includes the first wave of data.

Les pertes dans les jeux de données longitudinales peuvent biaiser l'inférence. Dans cette présentation, nous évaluons le potentiel d'utiliser des scores de propension appariés pour remplacer les sujets qui sortent des sondages par des individus d'un échantillon de rechange et ainsi retrouver la représentativité de l'échantillon. Nous examinons ceci avec des données longitudinales rassemblées sur deux séries d'entrevues avec des participants à un programme britannique sur le marché des travailleurs actifs. Les données résultantes étaient caractérisées par une perte significative. Cependant, la combinaison des dossiers administratifs avec les données du sondage permet d'observer certains résultats sur le chômage pour tous ceux qui ont répondus à la première étape du sondage, indépendamment de leur participation à la deuxième étape. Ceci permet d'évaluer le succès de l'approche des scores de propension appariés en comparant le chômage des individus qui ont quitté le sondage à celui de leurs remplacements. Dans la présentation, nous appliquons premièrement une méthode nouvellement présentée pour examiner le modèle de perte. Ceci montre que le processus des sujets qui quitte le sondage est non-ignorable au point tel que les méthodes

généralement utilisées de pondération des individus qui restent ou d'imputation multiple sont inefficaces pour éliminer le biais dû à la perte. En construisant les échantillons de remplacement artificiels, le potentiel de l'approche appariée pour éliminer un tel biais est démontré. Les meilleurs résultats se produisent lorsque l'échantillon de remplacement est ciblé dans le sens qu'il se compose des individus qui sont prédisposés à quitter. Cependant, dans le scénario plus réaliste où l'échantillon de remplacement se compose de ceux que l'on suspecte le plus de quitter, la représentativité est reconstituée à partir du moment de départ et pas avant. Par conséquent, l'approche peut être plus efficace pour reconstituer un ensemble de données longitudinales à représentativité à travers les sections mais peut être moins utile lorsque nous examinons certaines caractéristiques des individus sur une période qui inclut la première vague des données.

Monday June 9 • Lundi 9 juin 3:45 • 15h45 LSC 338

Jason Sutherland, C.J. Schwarz, Simon Fraser University

Multi-list methods using incomplete lists in closed populations • Méthodes de listes multiples en utilisant des listes incomplètes dans des populations fermées

Multi-list methods are now commonly applied to estimating the size of human populations, having been successfully applied to estimating census under-count, prevalence of diseases, such as diabetes and human immunodeficiency virus (HIV) and estimating homelessness and drug abuser populations. A key assumption in multi-list methods is that individuals have a unique "tag" that allows them to be matched across all lists. This paper develops multi-list methodology that relaxes the assumption of a single tag common to all lists. Estimates are found using estimating functions. An example illustrates its application to estimating the prevalence of diabetes and a simulation study investigates conditions under which the methodology is robust to different list and population sizes.

Les méthodes de listes multiples sont généralement appliquées pour estimer la taille des populations humaines. Elles ont été utilisées avec succès pour estimer le sous-dénombrement de recensement, la prévalence des maladies comme le diabète et le virus d'immunodéficience humain (VIH) et pour estimer les populations de sans abris et les population qui abuses de la drogue. Une hypothèse importante dans les méthodes de listes multiples est que les individus ont un " étiquette " unique, qui leur permet d'être associés parmi toutes les listes. Cette présentation développe la méthodologie des listes multiples qui relâche l'hypothèse d'une étiquette unique, commune à toutes les listes. Des estimateurs sont trouvé en utilisant des fonctions d'estimations. Un exemple illustre son application à l'estimation de la prévalence du diabète et une étude de simulation étudie les conditions dans lesquelles la méthodologie est robuste à différentes tailles de liste et de populations.

Monday June 9 • Lundi 9 juin 4:00 • 16h00 LSC 338

Guylaine Dubreuil, Mike Hidiroglou, Louis Pierre, Statistics Canada/Statistique Canada

Use of administrative data in the modelling of monthly survey data • Utilisation de données administratives dans la modélisation des données d'une enquête mensuelle

The cost and response burden associated to the survey activities have always been a major concern to Statistics Canada. These issues are especially important for monthly surveys. One option is the use of administrative data. We have to make sure that a source of monthly administrative data exists and that these data are available on a timely and reliable fashion.

Statistics Canada has been accessing the Canadian Goods and Services Tax (GST) data since 1997. These data offer the potential of modelling monthly economic survey data in the aim of decreasing the sample size. On the GST file, each unit (remitter) has a set of transactions. Each transaction reports the Sales and the corresponding Tax Paid for a reference period. Remitters are requested to report on a monthly, quarterly or annual frequency. That does not mean that they will report from the first day to the last day of a month, for instance. Moreover, there is a delay between the end of the reference period and the time that Statistics Canada receives the data and processes them. Modelling monthly survey data with administrative data (in particular GST data) is not as simple because of these particularities.

During this presentation, we will discuss the processing of the monthly GST files, namely, the editing and imputation of outliers and missing data, the calendarisation which transforms the data on a monthly basis and extrapolates the data that are not received yet because they are not due. Then, we will focus on the application of a model that uses clean calendarised GST data to model monthly survey data while respecting the timeliness issue. Finally, an application relevant to the Monthly Restaurants, Caterers and Taverns Survey will be presented.

Le coût et le fardeau de réponse associés aux activités d'enquête ont toujours été une préoccupation majeure à Statistique Canada. Ces questions sont particulièrement importantes pour les enquêtes mensuelles. Une option est l'utilisation de données administratives. Nous devons nous assurer qu'une source de données administratives mensuelles existe et que ces données sont disponibles à temps et sous un format utilisable.

Statistique Canada a accès aux données de taxe sur les produits et services (TPS) du Canada depuis 1997. Ces données offrent la possibilité de modéliser les données d'enquêtes économiques mensuelles, faisant en sorte que la taille d'échantillon puisse être réduite. Sur le fichier de TPS, chaque unité (rapporteur) a un ensemble de transactions. Pour chaque transaction, les ventes et la taxe sont rapportées pour une période de référence donnée. Les rapporteurs sont tenus de faire un rapport sur une fréquence mensuelle, trimestrielle ou annuelle. Cela ne veut pas nécessairement dire qu'ils rapporteront sur une période débutant le premier jour et se terminant le dernier jour d'un mois. De plus, il y a un délai entre la fin de la période de référence et le temps où Statistique Canada reçoit les données et les traite. La modélisation des données d'une enquête mensuelle avec des données administratives (en particulier les données de TPS) n'est pas aussi simple à cause de ces particularités.

Pendant cette présentation, nous discuterons du traitement des fichiers de TPS mensuels, à savoir, la vérification et l'imputation des données aberrantes et des données manquantes, la calendarisation qui transforme les données sur une base mensuelle et extrapole les données qui ne sont pas reçues parce qu'elles ne sont pas dues. Ensuite, nous nous concentrerons sur l'application d'un modèle qui utilise les données de TPS calendarisées et traitées pour modéliser les données d'une enquête mensuelle tout en respectant les délais prescrits. Finalement, une application faisant appel aux données de l'Enquête mensuelle sur les restaurants, traiteurs et tavernes sera présentée.

Monday June 9 • Lundi 9 juin 4:15 • 16h15 LSC 338

Daniel Hurtubise, Statistics Canada/Statistique Canada

Variance estimation in the context of complex surveys using administrative data • Estimation de la variance dans le cadre d'enquêtes complexes utilisant des données administratives

Estimating the precision of estimates has always been an important component for measuring the quality of survey data. This precision is sometimes based on approximate estimates of variance due to the complexity of the estimators and sampling plans. Hurtubise, et al. (2000) suggested a solution to this problem. This paper presents a general formula for the variance of a sampling plan based on two independent sources of data: a sample of administrative data and a survey sample. This general formula is based on Goodman ideas (1960). The main component of this variance formula is sampling variability of both sources. One assumption beneath this formula is that the variance due to classical imputation of administrative data is negligible in relation to total variance. In the context of census of administrative data, this assumption is no longer valid. Felx and Rancourt (2000) studied the variability due to imputation for some common imputation methods. These results are used to add this component to total variance. Synthetic estimators are used to produce the estimates. These synthetic estimators are functions of the administrative variables as well as of survey variables. Moreover, some variables are obtained from regression models. These models, calculated within model groups, are derived from survey data. They are defined such as they form an exhaustive partition of the survey sampling frame and that the best possible fit is obtained. The variability due to these models can not be negligible in the context of census of administrative data. In this paper, both data sources are described along with the model groups. A general variance formula is then described, including all sources of variability. Examples of synthetic estimators are used from Statistics Canada Survey on Employment, Payrolls and Hours (SEPH).

Le calcul de la précision des estimations a toujours été un élément important pour mesurer la qualité des données à l'étude. Cette précision est quelquefois basée sur des estimations approximatives de la variance dues à la complexité des estimateurs et des plans de sondage. Hurtubise, et al. (2000) ont suggéré une solution à ce problème. Cet article présente une formule générale pour la variance d'un plan de sondage basé sur deux sources indépendantes de données: un échantillon de données administratives et un échantillon tiré d'une enquête. Cette formule générale est basée sur l'approche de Goodman (1960). L 'élément principal de cette formule de variance est la variabilité due à l'échantillonnage des deux sources de données. Une hypothèse sous-jacente à cette formule est que la variance due à l'imputation classique des données administratives est négligeable par rapport à la variance totale. Cependant, dans le contexte d'un recensement des données administratives, cette hypothèse n'est plus valide. Felx et Rancourt (2000) ont étudié la variabilité due à l'imputation pour quelques méthodes courantes d'imputation. Ces résultats sont utilisés pour ajouter cette composante à la variance totale. Des estimateurs synthétiques sont utilisés pour produire les estimations. Ces estimateurs synthétiques sont fonctions des variables administratives aussi bien que des variables tirées de l'enquête. De plus, quelques variables sont obtenues à l'aide de modèles de régression. Ces modèles, calculé pour différents groupes-modèles, sont obtenus des données de l'enquête. Ces groupes-modèles sont définis tels qu'ils forment une partition exhaustive de la base de sondage de l'enquête et que le meilleur ajustement possible des modèles est obtenu. La variabilité due à ces modèles ne peut pas être négligeable dans le contexte d'un recensement des données administratives. Dans cet article, les deux sources de données sont décrites ainsi que les groupes-modèles. Une formule générale de la variance est décrite, incluant toutes les sources de variabilité. Des exemples d'estimateurs synthétiques sont tirés de l'Enquête sur l'Emploi, la Rémunération et les Heures (EERH) de Statistique Canada.

Session/Séance 18 • Applications of Statistics • Applications de la Statistique

Monday June 9 • Lundi 9 juin 3:30 • 15h30 LSC 234

James Wasiloff, Eric Wasiloff, Michigan State University

Improving the reliability and robustness of a pneumatic paint ball marker system through the practical application of parameter design optimization methodologies • Amélioration de la fiabilité et de la robustesse d'un système de marqueur pneumatique de balles de peintures par l'entremise d'une application pratique de méthodes d'optimisation de design de paramètres

Our team has developed and deployed a strategic practical parameter design optimization model for improving the precision, reliability and robustness of a pneumatic paint marker system. The global situation is such that users are continuously striving to improve performance (precision, reliability and robustness) of their paint marker systems. The trend to more advanced technologies needed to achieve this goal involves significant capital investment. No systematic or scientific method currently exists to efficiently evaluate and optimize the reliability and robustness of many forms of potential hardware component or sub-system design upgrades. We will present the successful application of elements of key contemporary quality and reliability engineering methodologies, where appropriate, including Design for 6 Sigma, Shainin, Taguchi, etc. to optimize system function and robustness in a factorial experiment with noise factors in an outer array.

Notre équipe a développé et mis en place un modèle pratique stratégique d'optimisation du design des paramètres pour améliorer la précision, la fiabilité et robustesse d'un système pneumatique de marqueur à base de peinture. La situation générale est telle que les utilisateurs tâchent sans interruption d'améliorer la performance (précision, fiabilité et robustesse) de leurs systèmes de marqueur. Les technologies avancées requises pour réaliser ce but implique des investissements importants de capitaux. L'achat des mises à niveau de composantes et de sous-systèmes est souvent basés sur des évidences faibles telle que le point de vue des utilisateurs ou les dires des fabricants de composantes et les composantes sont informellement évaluées à l'aide de tests sur un facteur à la fois. Aucune méthode systématique ou scientifique n'existe pour évaluer et optimiser efficacement la fiabilité et la robustesse de plusieurs de types de matériaux potentiels ou de mises à niveau des conceptions de sous-systèmes. Nous présentons l'application réussie des éléments des méthodologies contemporaines de génie pour la qualité et la fiabilité, là où approprié, comme le design pour 6 Sigma, Shainin, Taguchi, etc. pour optimiser les fonctions et la robustesse d'un système par une expérience d'analyse factorielle avec des facteurs de bruit dans un tableau externe.

Monday June 9 • Lundi 9 juin 3:45 • 15h45 LSC 234

Asokan Mulayath Variyath, Bovas Abraham, University of Waterloo

Estimation of vendor's process capability based on submitted lots • Estimation de la capacité des processus des fournisseurs basées sur des pièces fournies

Modern Quality Management practices such as ISO 9000 /QS 9000 stresses the need for improving the quality of all parts supplied by the vendors. To ensure this, vendors should have good quality systems and their processes should exhibit the capability to meet the specifications dictated by the customers. Organizations are continually monitoring the performance of the vendors by assessing the quality based on the submitted lots since it is

impossible or very costly to make onsite assessment. Evaluation based on the submitted lots may not give an accurate level of the vendor's process, if the submitted lots are screened with respect to the specifications by the vendor to avoid rejection of lots at the customer's end. Estimation of variability based on the screened lots are biased and it is overestimate the process capability. Two methods for assessing the vendor's process capability based on Maximum Likelihood and EM algorithm are proposed in this paper. An improvement in convergence of EM algorithm is also discussed. Simulation studies are used to compared the methods.

Les procédures modernes de qualité de la gestion telles que ISO 9000/QS 9000 soulignent le besoin d'améliorer la qualité de toutes les pièces fournies par les fournisseurs. Pour assurer ceci, les fournisseurs doivent avoir des systèmes de bonne qualité et leurs processus doivent montrer leur capacité à répondre aux spécifications dictées par les clients. Les organisations surveillent continuellement la performance des fournisseurs en évaluant la qualité basée sur les lots déjà livrés puisqu'il est impossible ou très coûteux de faire l'évaluation sur le site de production. L'évaluation basée sur les lots déjà livrés peut ne pas donner le niveau précis des processus du fournisseur, si les lots déjà livrés sont examinés par rapport aux spécifications du fournisseur pour éviter le rejet des lots du côté du client. L'estimation de la variabilité basée sur les lots examinés est biaisée et elle surestime la capacité du processus. Nous proposons deux méthodes pour évaluer la capacité des processus du fournisseur basées sur le maximum de vraisemblance et l'algorithme EM. Une amélioration de la convergence de l'algorithme EM est également discutée. Des études de simulation sont utilisées pour comparer les méthodes.

Monday June 9 • Lundi 9 juin 4:00 • 16h00 LSC 234

Ejaz Ahmed, University of Windsor

Comparing multiple process capability indices in non-normal distributions • Comparaison des indices de processus multiples de capacité pour des distributions non normales

The problem of comparing multiple process indices is an important one. In this talk, we consider the large-sample inference of a capability process index and investigate its statistical properties in a multi-sample set up when random samples are drawn from arbitrary populations. This communication will show some lower and upper confidence limits for the actual capability index. Further, asymptotic statistical procedures are developed for testing the homogeneity of these indices. A log-transformation is also proposed. A simulation study is carried out to appraise the performance of the suggested methods for a moderate sample. The study indicates that the proposed methods are comparable in terms of coverage probabilities and average length of confidence intervals, size of the tests and power functions. We provide a total inferential package for the problem under study. Further, some realistic examples of the application of the methods will be given.

Le problème de comparer les index de processus multiples est important. Dans cette présentation, nous considérons l'inférence sur de grands échantillons d'un index de processus de capacité et nous étudions ses propriétés statistiques dans un cadre d'échantillons multiples lorsque les échantillons aléatoires sont tirés de populations arbitraires. Cette présentation montre quelques limites de confiance inférieures et supérieures pour l'index réel de capacité. De plus, des procédures statistiques asymptotiques sont développées pour tester l'homogénéité de ces index. Nous proposons également une transformation logarithmique. Une étude de simulation est effectuée pour évaluer la perfor-

mance des méthodes suggérées pour un échantillon de taille moyenne. L'étude indique que les méthodes proposées sont comparables en termes de probabilités de couverture, de longueur moyenne des intervalles de confiance, de niveau des tests et des fonctions de puissance. Nous fournissons un package d'inférence complet pour le problème étudié. De plus, quelques exemples réalistes d'application des méthodes sont donnés.

Monday June 9 • Lundi 9 juin **LSC 234**

Michael MacLeod, St. Francis Xavier University; René F. Reitsma, Oregon State University; Lehana Thabane, McMaster University

Spatialization of web sites using a weighted frequency model of navigation data • Spatialisation des sites web en utilisant un modèle de fréquences pondérées des données sur la navigation

A common problem in the spatialization of information systems is the determination of geometry, i.e., dimensionality and metric. Such geometry is either chosen apriori or is inferred aposteriori from secondary data. Recent work emphasizes the use of geometric information latent in a system's navigational record. Resolving this information from its noisy background, however, requires an unambiguous criterion of selection. In this paper we use a previously published, statistically robust method for resolving a web-based information system's geometry from navigational data. However, because of the method's (theoretical) sensitivity to data selection, a weighted frequency correction approach based on empirical probability distributions is applied. The effect of this correction on the determination of a web-space geometry is investigated. Results indicate that the inferred geometry remains robust, i.e. it does not significantly change under this probabilistic correction.

Un problème commun dans la spatialisation des systèmes d'information est la détermination de la géométrie, c.-à-d., la dimensionnalité et la métrique. Une telle géométrie est choisie a priori ou est supposé a posteriori à partir des données secondaires. Les travaux récents soulignent l'utilisation d'information géométrique latente dans l'enregistrement d'un système de navigation. Cependant, l'obtention de cette information à partir d'un background bruité exige un critère de sélection non ambigu. Dans cette présentation, nous utilisons une méthode qui a déjà été publiée et statistiquement robuste pour résoudre la géométrie d'un système d'information internet pour des données de navigation. Cependant, en raison de la sensibilité (théorique) de la méthode à la sélection des données, une approche de correction de fréquence pondérée basée sur les distributions empiriques de probabilité est appliquée. L'effet de cette correction sur la détermination de la géométrie d'un espace internet est étudié. Les résultats indiquent que la géométrie supposée demeure robuste, c.-à-d. elle ne change pas de manière significative avec cette correction probabiliste.

Monday June 9 • Lundi 9 juin 4:30 • 16h30 **LSC 234**

Rolf Turner, Pradeep Banerjee, University of New Brunswick

A differential equations approach to some asset selling problems • Une approche par les équations différentielles pour certains problèmes de vente d'actifs

It is a well-known phenomenon that airline passengers travelling on the same flight (same origin and same destination) and in the same class (cabin) will often have paid substantially different fares. It is also well known that in order to obtain a cheap fare one has to purchase one's ticket well in advance of the departure date. This latter phenomenon is at first blush

somewhat counter-intuitive: One might expect that, in order to dispose of unsold seats, the airline might offer substantial discounts just prior to departure time. The explanation of the apparent paradox can be expressed by saying that the sales are subject to a time-varying elasticity of demand. In this talk I shall indicate how, given some assumptions about purchase probabilities and the process of customer arrivals, it is possible to work out a pricing policy which will maximize expected revenue for the airline. The solution of the problem involves setting up a coupled system of differential equations which is readily amenable to numerical solution. The ideas are essentially elementary but appear to be new; they could have substantial practical application.

C'est un phénomène bien connu que des passagers de compagnies aériennes voyageant sur le même vol (la même origine et la même destination) et dans la même classe (cabine) auront souvent payé des prix substantiellement différents pour leurs billets. Il est également bien connu qu'afin d'obtenir le meilleur prix, on doit acheter son billet bien avant la date de départ. Ce dernier phénomène peut, à première vue, sembler aller à l'encontre de l'intuition: quelqu'un pourrait penser qu'afin de se départir des sièges non vendus, la compagnie aérienne pourrait offrir des escomptes substantiels juste avant la date de départ. L'explication du paradoxe apparent peut être donnée en disant que les ventes sont sujettes à une élasticité de la demande qui varie dans le temps. Dans cette présentation, j'indiquerai comment, étant donné quelques hypothèses au sujet des probabilités d'achat et du processus d'arrivé des clients, il est possible d'établir une politique des prix qui maximisera le revenu prévu pour la compagnie aérienne. La solution du problème implique de mettre en place un système couplé d'équations différentielles qui mène aisément à la solution numérique. Les idées sont essentiellement élémentaires mais semblent être nouvelles; elles pourraient avoir des applications pratiques substantielles.

Monday June 9 • Lundi 9 juin 4:45 • 16h45 LSC 234

Theodoro Koulis, University of Waterloo

A stochastic model for sea ice • Un modèle stochastique pour la glace de mer

The role of seasonal sea ice formation at the poles is complex and closely linked to the Earth's climate. It is thought that the amount of sea ice can have a significant effect on the energies transferred between the atmosphere and the ocean. Understanding the seasonal sea ice process at the poles is therefore of great interest to scientists. Sea ice concentration data sets derived from Earth orbiting satellites are readily available and contain observations that span several decades. This data, which is both spatial and temporal in nature, can be quite difficult to analyze. The methods of analysis for this type of data can be computationally intensive. We present a spatial nearest-neighbor model as a candidate for describing the sea ice process. The model is a Markov process on a lattice and can be controlled through two parameters. These parameters give some insight on the long term behavior of the process. We will discuss various methods for estimating these parameters. The methods are based on differential equations associated with the biased voter model. It is hoped that these methods will be helpful in analyzing multi-temporal spatial data and to make inferences on global climate change.

Le rôle de la glace de mer saisonnière aux pôles est complexe et étroitement lié au climat de la Terre. On pense que la quantité de glace de mer peut avoir un effet significatif sur l'énergie transféré entre l'atmosphère et l'océan. La compréhension du processus saisonnier des glaces de mer aux pôles est donc de grand intérêt pour les scientifiques. Les bases de données sur la concentration en glace de mer obtenues des satellites qui orbitent autour de

la Terre sont aisément disponibles et contiennent des observations sur plusieurs décennies. Ces données spatiales et temporelles sont très difficiles à analyser. Les méthodes d'analyse pour ce type de données sont très intensives au niveau informatique. Nous présentons un modèle spatial du plus proche voisin pour décrire le processus des glaces de mer. Le modèle est un processus de Markov sur un treillis et peut être contrôlé par deux paramètres. Ces paramètres donnent certaines informations sur le comportement à long terme du processus. Nous discuterons de diverses méthodes pour estimer ces paramètres. Ces méthodes sont basées sur des équations différentielles liées au modèle biaisé de décision. Nous espérerons que ces méthodes seront utiles pour analyser des données spatiales multi-temporelles et pour faire de l'inférence sur les changements climatiques.

Session/Séance 19 • Statisticians in Action I • Statisticiens en action I

Monday June 9 • Lundi 9 juin 3:30 • 15h30 LSC 332

Video presentation • Présentation vidéo

Committee on Professional Development • Comité sur le perfectionnement professionnel

Session/Séance 20 • SSC Gold Medal Address • Allocution du récipiendaire de la médaille d'or de la SSC

Tuesday June 10 • Mardi 10 juin 8:30 • 8h30 MM Ondaajte Theatre

Muni Srivastava, University of Toronto

Multivariate analysis with fewer observations than the dimension • Analyse multivariée avec moins d'observations que la dimension des données

In this paper we develop multivariate theory for analyzing multivariate datasets with fewer observations than the dimension. Such data arise, for example, in DNA microarrays where there are observations on thousands of genes but only on few patients. Methods of drawing inference such as testing of hypotheses and confidence intervals are presented for one-sample, two sample and Manova. A sample measure of distance between two populations is defined. This sample (squared) distance is used in classifying an individual with p-vector observation into one of the several multivariate populations by minimum distance rule. Methods for detecting outliers and imputing missing observations are also presented.

Dans cette présentation, nous développons la théorie multivariée pour analyser des ensembles de données multivariées avec moins d'observations que la dimension des données. De telles données surgissent par exemple dans des microréseaux d'ADN où il y a des observations sur des milliers de gènes mais sur quelques patients seulement. Les méthodes d'inférence telle que les tests d'hypothèses et des intervalles de confiance sont présentés pour un échantillon, deux échantillons et pour des Manova. Nous définissons une mesure échantillonnale de la distance entre deux populations. Cette distance échantillonnale au carré est utilisée pour la classification d'un individu avec un vecteur de p observations dans une des multiples populations multivariées par la règle de la distance minimale. Des méthodes pour détecter des valeurs aberrantes et imputer des valeurs aux observations manquantes sont également présentées.

Session/Séance 21 • Shape-Restricted Inference • Inférence avec restrictions sur la forme

Tuesday June 10 • Mardi 10 juin 10:30 • 10h30 LSC 240

Michael Woodroofe, University of Michigan; Mary Meyer, University of Georgia

Estimating a unimodal density • Estimation d'une densité unimodale

Maximum likelihood estimators of a unimodal or non-increasing density are severely affected by the spiking problem, even if the location of the mode is known apriori: The estimated modal value is too big! Some existing methods for circumventing or ameliorating the spiking problem will be reviewed and some new ones explored. The existing methods include, grouping the data, requiring piecewise linearity, penalizing large modal values, and combinations thereof. The new methods all involve making additional assumptions about the unknown density f for the purposes of estimation. One of these is to suppose that the density if concave in a neighborhood of the mode. This approach has worked well in simulations when the mode is known. An intriguing possibility is to suppose that f is a Polya frequency function in the computation of the maximum likelihood estimator. This approach does not require the mode to be specified but always produces an estimator in the (fairly small) class of Polya frequency functions. A related approach is to suppose that 1/f is convex. This approach does not require the mode to be specified either and uses a larger subclass.

Les estimateurs du maximum de vraisemblance d'une densité unimodale ou non-croissante sont sévèrement affectés par les pointes dans les données, et ce, même si l'endroit du mode est connu a priori: la valeur estimée du mode est trop grande! Nous passons en revue quelques méthodes existantes pour éviter ou améliorer le problème des pointes et nous explorons quelques nouvelles méthodes. Les méthodes existantes incluent le groupage des données, l'exigence de la linéarité par morceaux, la pénalisation des grandes valeurs modales et la combinaison. Toutes les nouvelles méthodes impliquent des hypothèses additionnelles sur la densité inconnue f dans le but de l'estimation. Une de ces hypothèses est de supposer que la densité est concave dans un voisinage du mode. Cette approche a bien fonctionné dans les simulations lorsque le mode est connu. Une possibilité intéressante est de supposer que f est une fonction de poly-fréquence dans le calcul de l'estimateur du maximum de vraisemblance. Cette approche n'exige pas la connaissance a priori du mode, mais produit toujours un estimateur dans la classe (assez petite) des fonctions de poly-fréquence. Une autre approche reliée est de supposer que 1/f est convexe. Cette dernière approche n'exige pas la connaissance a priori du mode et utilise une plus grande classe sous jacente.

Tuesday June 10 • Mardi 10 juin 11:00 • 11h00 LSC 240

Richard Dykstra, University of Iowa; Chris Carolan, East Carolina University

Characterization of the least concave majorant of Brownian Motion with application to construction • La caractérisation du majorant le moins concave du mouvement brownien avec application à la construction

The characterization of the least concave majorant of Brownian motion by Pitman (1983) is tweaked, conditional on the value of a vertex point. The joint distribution of the vertex point is derived and shown that it can be generated very easily. A procedure is then outlined by which one can construct the least concave majorant of a standard Brownian

motion over any finite, closed subinterval of (0, infinity). This construction is exact in distribution. A discussion of how to translate the aforementioned construction to the least concave majorant of a Brownian bridge is also presented.

Nous apportons quelques changements à la caractérisation du majorant le moins concave du mouvement brownien de Pitman (1983), conditionnellement à la valeur d'un point de sommet. La distribution conjointe du point de sommet est dérivée et il est prouvé qu'elle peut être générée très facilement. Une procédure est alors décrite pour construire le majorant le moins concave d'un mouvement brownien standard sur n'importe quel sous-intervalle fini et fermé de (0, infini). Cette construction est exacte en distribution. Nous présentons également une discussion sur la manière de traduire la construction mentionnée ci-dessus au majorant le moins concave d'un pont brownien.

Tuesday June 10 • Mardi 10 juin 11:30 • 11h30 LSC 240

Mary Meyer, University of Georgia

Improving the power of tests with shape-restricted alternatives via projections onto subcones • Amélioration de la puissance des tests par des alternatives de restrictions sur la forme via des projections sur des sous-cônes

Unbiased tests for the constant versus monotone regression function, as well as the linear versus convex regression function, are known to have null distributions equal to those of mixtures of beta random variables. Both monotone and convex regression estimators exhibit "spiking" at the endpoints of the data range, where the estimator is inconsistent. Consistent estimators for both shape-restricted alternatives are proposed, for which the test statistic using the consistent estimator has again the form of a mixture of beta densities. The power of the test is shown through simulations to improve with the consistent estimators in both examples, for many true underlying regression functions.

Les tests sans biais pour la constante contre la fonction monotone de régression, comme la fonction de régression linéaire contre la fonction convexe, sont connus pour avoir des distributions nulles égales à celles des mélanges de variables aléatoires bêtas. Les estimateurs de régression monotones et convexes montrent des pics aux extrémités de la plage des données, où l'estimateur est inconsistant. Nous proposons des estimateurs consistants pour les deux alternatives forme-restreintes, pour lesquelles la statistique du test basée sur l'estimateur consistant a encore la forme d'un mélange de densités bêtas. Des simulations montrent que la puissance du test s'améliore avec l'utilisation des estimateurs consistants dans les deux exemples, et ce pour plusieurs fonctions de régression sous jacentes réelles.

Session/Séance 22 • Analysis of Mixed Discrete and Continuous Outcome Data • Analyse de mélanges de variables discrètes et continues

Tuesday June 10 • Mardi 10 juin 10:30 • 10h30 LSC 338

Ming-yi Hu, Yamanouchi Pharma America, Inc., Thomas R. Belin, University of California at Los Angeles

Evaluating imputation model choice in a study with incomplete continuous and categorical data and follow-up of initial nonrespondents • Évaluation du choix d'un modèle d'imputation dans une étude avec des données continues et catégorielles incomplètes et des données de suivi des non-répondants initiaux

In an outcomes research study on mental health services, a mixture of continuous and categorical data reflecting successful community adaptation were collected on mental health patients. Outcome data were missing for many patients, but follow-up data were available for a number of initial non-respondents. We performed multiple imputation for missing values based on two statistical models, each applied to 18 continuous variables and 16 categorical variables: a general location (GL) model with cell means assumed to have a multivariate regression on the main effects of the categorical factors, and a multivariate normal (MVN) model adapted to produce imputed binary values based on MVN imputations. Observed follow-up values were compared to imputed values on patients who had missing data on the last observation period. Although neither statistical model produced imputed values that were perfectly consistent with the follow-up data, the multivariate normal model outperformed the general location model. We discuss the results and highlight issues of model selection with incomplete continuous and categorical data.

Dans une étude d'efficacité sur les services de santé mentale, un mélange de données continues et catégorielles qui reflètent le degré d'adaptation dans la communauté ont été rassemblés pour des patients psychiatriques. Les données à la fin de l'étude étaient inconnues pour beaucoup de patients, mais des données de suivi étaient disponibles pour un certain nombre de non-répondants. Nous avons effectué une imputation multiple pour les données manquantes basées sur deux modèles statistiques chacun ayant 18 variables continues et 16 variables catégorielles. Le premier est un modèle général de position (GL) où nous supposons que les moyennes des cellules ont un effet de régression multiple sur les effets principaux des facteurs catégoriels, et le second, un modèle normal multivarié (MVN) adapté pour produire des valeurs imputées binaires, basées sur des imputations de MVN. Les valeurs de suivi observées ont été comparées aux valeurs imputées sur les patients dont nous avions des données manquantes pour la dernière période d'observation. Bien que ni l'un ni l'autre des modèles statistiques ne produit des valeurs imputées parfaitement conforment aux données de suivi, la performance du modèle normal multivarié surpasse celle du modèle général de position. Nous discutons des résultats et du choix de modèle avec des données continues et catégorielles incomplètes.

Tuesday June 10 • Mardi 10 juin 11:00 • 11h00 LSC 338

A. de Leon, University of Calgary; K.C. Carrière, University of Alberta

General mixed-data model: Extension of general location and grouped continuous models • Modèle général de données mixtes: une extension du modèle général de localisation et du modèle groupé continu

In this paper, a general model for multivariate data with mixtures of nominal, ordinal and continuous variables called the general mixed-data model is proposed. The approach adopted in developing the model is motivated by the need to account for the various levels of measurement in the data, which many conventional approaches fail to incorporate in the analysis. The general mixed-data model includes as special cases the general location model of Olkin and Tate (1961) and the mixed-data models studied by Bedrick et al. (2000), Poon and Lee (1987) and Anderson and Pemberton (1985). A full likelihood-based approach that yields maximum likelihood estimates of the model parameters is outlined, and algorithms to implement it are provided. An alternative estimation method based on the pairwise likelihood approach is also presented. Statistical inference is also discussed for testing the population means, polychoric and polyserial correlation parameters among the variables across and within states based on both approaches. To illustrate the utility

of the general mixed-data model, it is used to develop a distance measure that can be used for mixed data with ordinal, in addition to nominal and continuous, variables. This application of the model extends earlier work by Bedrick et al. (2000) and Bar-Hen and Daudin (1995).

Dans cette présentation, nous proposons un modèle général pour des données multivariées constitué d'un mélanges de variables nominales, ordinales et continues appelées le modèle général de données mixtes. L'approche utilisée pour développer le modèle est motivée par la nécessité d'expliquer les divers niveaux de mesure parmi les données, ce que beaucoup d'approches conventionnelles n'incorporent pas dans leur analyse. Le modèle général de données mixtes inclut, comme un cas spécial, le modèle général de localisation d'Olkin et Tate (1961) et les modèles de données mixtes étudié par Bedrick et al. (2000), Poon et Lee (1987) et Anderson et Pemberton (1985). Nous décrivons une approche basée sur la vraisemblance qui donne des estimations du maximum de vraisemblance des paramètres du modèles, et nous donnons des algorithmes pour implémenter la méthode. Nous présentons également une méthode alternative d'estimation basée sur la vraisemblance par paire. Nous discutons aussi de l'inférence statistique pour la moyenne des populations, la corrélation polychorique et polysériale parmi les variables basés sur les deux approches. Pour illustrer l'utilité du modèle général de données mixtes, il est utilisé pour développer une mesure de distance qui peut être utilisé pour des données mixtes avec variables ordinales, en plus de variables nominales et de continues. Cette application du modèle prolonge les premiers travaux par Bedrick et al. (2000) et Barre-Bar-Hen et Daudin (1995).

Tuesday June 10 • Mardi 10 juin 11:30 • 11h30 LSC 338

Avner Bar-Hen, University of Aix-Marseille III; F. Mortier, CIRAD-Foret

Estimation of Mahalanobis distance with continuous and discrete variables • Estimation de la distance de Mahalanobis à l'aide de variables continues et discrètes

Mahalanobis distance is a common tool in discriminant analysis. We present two extensions to the the case of mixed continuous and discrete variables. The first approach is based on Kullback-Leibler divergence. We study statistical properties of these distance estimator, variables selection and atypicality index (possibility that an individual is not coming from one of the pre-determined populations). The second approach consider that discrete variables are realizations of non-observed continuous variables. This modelling is based on generalized probit models. We consider the case where the parameter of the model is estimated with exogeneous variables. Quality os the estimation is study through simulations. Both approaches are compared and results are applied to data for varietal distinctivness.

Dans les problèmes de discrimination la distance de Mahalanobis est fréquemment utilisée. Nous présentons deux généralisations au cas du mélange de variables continues et discrètes. La première approche est basée sur la divergence de Kullback-Leibler. Nous étudions les propriétés statistiques de l'estimateur de la distance, la question de la sélection de variables ou de possibilité de non classement d'un individu (atypicalité) La deuxième approche consiste à mimer la situation quantitative en considérant que les variables discrètes sont une discrétisation d'une variable continue sous-jacente. Ceci est réalisé à l'aide d'un modèle probit généralisé. Nous montrons comment les paramètres du modèle probit peuvent être estimés à l'aide de variables exogènes. La qualité de l'estimation est étudiée à l'aide de simulations. Les deux approches seront comparées et les résultats seront appliquées au problème de la protection variétale.

Session/Séance 23 • Recent Developments in Small Area Estimation • Développements récents en estimation sur de petits domaines

Tuesday June 10 • Mardi 10 juin 10:30 • 10h30 LSC 242

J.N.K. Rao, Carleton University

Some new developments in small area estimation • Nouveaux développements dans le domaine de l'estimation sur de petits domaines

Discussants: Donald J. Malec, United States Bureau of the Census and
Wayne A. Fuller, Iowa State University

Small area estimation has received a lot of attention in recent years due to growing demand for reliable small area statistics. Traditional area-specific direct estimators may not provide adequate precision because sample sizes in small areas are seldom large enough. This makes it necessary to employ indirect estimators that borrow strength from related areas, in particular model-based indirect estimators based on explicit linking models. Basic area level and unit level models and their extensions have been extensively studied in the literature to derive empirical best linear unbiased prediction (EBLUP), empirical Bayes (EB) and hierarchical Bayes(HB) small area estimators and associated measures of variability. In this presentation, I will cover several important new developments related to model-based small area estimation including the following topics: unmatched sampling and linking area level models, use of sampling weights in unit level models, jackkinfe methods for MSE estimation, triple-goal estimators, and MSE estimation for area level models when the sampling variances are estimated. Some applications will also be presented.

L'estimation sur de petit domaines a suscité beaucoup d'attention ces dernières années dû à la demande croissante de statistiques fiables sur ces domaines. Les estimateurs directs traditionnels pour des secteurs spécifiques peuvent ne pas avoir une bonne précision car la taille de l'échantillon lorsqu'on travail sur des petits domaines est rarement grande. De là la nécessité d'utiliser les estimateurs indirects qui empruntent la force aux domaines apparentés, en particulier les estimateurs indirects basés sur les modèles explicites de liaison. Des modèles de niveau de secteur et de niveau d'unité standard et leurs extensions ont été intensivement étudiés dans la littérature pour dériver la meilleure prévision linéaire sans biais empirique (EBLUP), l'estimateur de Bayes de petits domaines empirique (EB) et l'estimateur hiérarchique de petits domaines de Bayes (HB) ainsi que les mesures de variabilité qui leur sont associées. Dans cette présentation, nous couvrons plusieurs nouveaux développements importants liés à l'estimation sur des petits domaines basée sur les modèles. Les sujets suivants seront traités: l'échantillonnage non pairé et les modèles de liaison de niveau de secteur, l'utilisation de poids échantillonnaux dans des modèles de niveau d'unité, les méthodes du jackkinfe pour l'estimation des estimateurs des moindres carrées, les estimateurs de triple-but, et l'estimation des moindres carrées pour des modèles de niveau de secteur quand les variances échantillonnales sont estimées. Nous présentons également quelques applications.

Session/Séance 24 • Special Session of the Pacific Institute for the Mathematical Sciences on Robustness • Session spéciale du Pacific Institute for the Mathematical Sciences en robustesse

Tuesday June 10 • Mardi 10 juin 10:30 • 10h30 LSC 238

Elvezio Ronchetti, University of Geneva

Future directions in robust statistics • Développements futurs en statistique robuste

Robust statistics deals with deviations from the assumptions on the model and develops statistical procedures which are still reliable and reasonably efficient in a small neighborhood of the model. It can be viewed as the statistics of approximate parametric models and it builds a bridge between the Fisherian parametric approach and the fully nonparametric approach. In the first part of the talk we will review some basic ideas developed in robust statistics which have become standard tools in modern statistics. In the second part we will try to outline some future directions starting from recent work in fairly complex models in two broad areas: model selection and estimation and inference in financial models.

Les statistiques robustes conjuguent avec les déviations provenant des hypothèses du modèle et développent des procédures statistiques fiables et raisonnablement efficaces sur un petit voisinage du modèle. Elles peuvent être vues comme des statistiques de modèles approximativement paramétriques et elles forment un pont entre l'approche paramétrique de Fisher et l'approche non paramétrique complète. Dans la première partie de cette présentation, nous allons passer en revue les idées de bases développées par les statistiques robustes, qui sont devenues des outils standards en statistique moderne. Dans la deuxième partie, nous essayons de souligner des directions futures commençant par des travaux récents sur des modèles relativement complexes dans deux champs: la sélection de modèles et l'estimation et l'inférence dans les modèles financiers.

Tuesday June 10 • Mardi 10 juin 11:15 • 11h15 LSC 238

David Tyler, Rutgers University

Multivariate M-estimation: concepts and applications • M-estimation multivariée: concepts et applications

In this talk, a general review of redescending M-estimates of multivariate location and scatter is presented. It is argued that the proper way to view such M-estimates is to partition the scatter matrix into a structural or "shape" component and a nuisance "scale" component. This leads to the introduction of the class of redescending M-estimates of multivariate location and shape with auxiliary scale. This class then provides a general framework for studying some of the high breakdown point methods developed for multivariate statistics and a unifying framework for comparing these methods. The multivariate M-estimates with auxiliary scale include as special cases the minimum volume ellipsoid estimates, the multivariate S-estimates, the multivariate constrained M-estimates, and the recently introduced multivariate MM-estimates. Various interpretation and applications of these multivariate M-estimates are explored. In particular, their general relationship to multivariate density estimation and cluster analysis problems in statistics, as well as their relationship to the scale space analysis and multiple structure problems in computer vision are discussed.

Dans cette présentation, nous présentons une revue générale des M-estimateurs de redescente pour la localisation et la dispersion multivariée. Nous argumentons sur le fait que la manière appropriée de voir de tels M-estimateurs est de diviser la matrice de dispersion en une composant structurale ou " de forme " et une composante d'échelle de nuisible. Ceci mène à l'introduction de la classe de M-estimateurs de redescente de localisation et de formes multivariées avec échelle auxiliaire. Cette classe fournit un cadre général pour étudier certaines des méthodes de point de panne élevées, développées pour des statistiques multivariées et un cadre unifié pour comparer ces méthodes. Le M-estimeur multivarié avec

échelle auxiliaire inclut comme un cas spécial les estimations du volume de l'ellipsoïde minimal, le S-estimeur multivarié, le M-estimeur multivarié sous contrainte et le MM-estimeur multivarié qui a été récemment présenté. Diverses interprétations et applications de ces M-estimeurs multivariés sont explorées. Nous discutons en particulier des liens généraux qu'ils ont avec les estimations de densités multivariées, des problèmes d'analyse en grappes en statistiques ainsi qu'avec l'analyse de l'espace d'échelle et des problèmes de structure multiple en vision informatique.

Session/Séance 25 • Time Series Methods for Fitting Dynamical Models • Méthodes de séries chronologiques pour l'ajustement de modèles dynamiques

Tuesday June 10 • Mardi 10 juin 10:30 • 10h30 LSC 332

Keith Thompson, Kassiem Jacobs, Dalhousie University

The assimilation of observations into ocean model • L'assimilation des observations dans des modèles océaniques

Oceanographers are using a wide range of techniques to assimilate data into dynamical models with the goal of improving predictions of ocean conditions on scales ranging from global to that of small tidal inlets. An overview is presented of the main types of data assimilation currently used by oceanographers. Where possible an interpretation is given in Bayesian terms. Two examples will be used for illustration: (i) Estimating of the seasonal state of the North Atlantic from seasonal means of observed temperature and salinity (ii) Forecasting the evolution of fields of rotating vortices from the observed motion of passive drifters. For (i) the dynamical model is based on the discretized form of a set of coupled, nonlinear partial differential equations corresponding to accepted physical principles. For (ii) the dynamical model is highly idealized and describes the stochastic motion of a set of vortices and the drifters advected by them. The dimension of the state vector for (i) is many orders of magnitude greater than that of (ii). For both examples the Kalman Filter is shown to form the basis of the assimilation scheme. The presentation will conclude with a discussion of some of the statistical problems that may arise as oceanographers start to develop operational models of the deep ocean and couple them with models of the atmosphere and marine ecosystems.

Les océanographes utilisent un éventail de techniques pour assimiler des données dans des modèles dynamiques dans le but d'améliorer des prévisions sur les conditions océaniques sur des échelles s'étendant de global à celle de petits goulets. Nous présentons une vue d'ensemble des principaux types d'assimilation des données actuellement utilisées par les océanographes. Dans les situations où c'est possible, nous interprétons les données dans un cadre bayésien. Deux exemples sont utilisés pour illustrer: (i) l'estimation de l'état saisonnier de l'Atlantique Nord à partir des moyennes saisonnières de la température observée et de la salinité (ii) la prévision de l'évolution des champs des vortex rotatifs à partir des mouvements observés des dériveurs passifs. Pour (i), le modèle dynamique est basé sur la forme discrétisée d'un ensemble couplé d'équations différentielles non-linéaires correspondant aux principes physiques admis. Pour (ii), le modèle dynamique est fortement idéalisé et décrit le mouvement stochastique d'un ensemble de vortex et des dériveurs affectés par eux. La dimension du vecteur d'état pour (i) est d'ordre de grandeur beaucoup plus grand que celle de (ii). Pour les deux exemples, nous montrons que le filtre de Kalman forme la base du schème d'assimilation. Nous concluons la présentation avec une discussion sur

certains des problèmes statistiques qui peuvent surgir pendant que les océanographes commencent à développer des modèles opérationnels pour les profondeurs de l'océan et à les coupler avec des modèles des écosystèmes atmosphériques et marins.

Tuesday June 10 • Mardi 10 juin 11:00 • 11h00 LSC 332

Pierre Gauthier, Environment Canada; Mark Buehner, Stéphane Laroche, Monique Tanguay, Meteorological Service of Canada/Service Météorologique du Canada

Operational implementation of variational assimilation • Mise en oeuvre opérationnelle de l'assimilation variationnelle

The variational form of the statistical estimation problem has been implemented at many operational numerical weather prediction centres, including the Canadian Meteorological Centre (CMC). The 3D variational assimilation (or 3D-Var) has been motivated initially to provide a better framework for the direct assimilation of satellite radiance data which are nonlinearly related to the analysis variables. The 3D-Var can naturally be extended to include the time dimension and lead to a quadri-dimensional variational assimilation system (4D-Var) that makes it possible to assimilate observations at the exact observation time with implicitly defined flow-dependent error statistics. This paper will introduce the variational formulation from a Bayesian perspective that allows the inclusion of nonlinearities in the observation operators. Practical aspects of the operational implementation of a variational analysis will be discussed including the incremental approach and the use of non-Gaussian probability distributions that better represent the error statistics of observations with gross errors. This will be supported with results obtained with the operational 3D-Var assimilation system of the CMC and from the preoperational 4D-Var system currently under development.

La forme variationnelle du problème d'estimation statistique a été mise en oeuvre dans plusieurs centres opérationels de prévision numérique du temps, incluant le Centre Météorologique Canadien (CMC). L'assimilation variationnelle 3D (ou 3D-Var) fut initialement motivée par la nécessité d'avoir un cadre plus approprié pour l'assimilation de mesures de radiance satellitales qui sont reliés non-linéairement aux variables d'analyse. Le 3D-Var peut être tout naturellement étendu pour inclure la dimension temporelle ce qui conduit à un système d'assimilation variationnel quadri-dimensionnel (4D-Var) qui rend possible l'assimilation d'observations au temps exact où la mesure est prise tout en utilisant des covariances d'erreur de prévision qui dépendent implicitement de l'écoulement. Cette présentation introduira la formulation variationnelle d'un point de vue Bayésien qui permet l'inclusion de non-linéarités dans les opérateurs d'observation. Des aspects pratiques de l'implémentation opérationnelle d'une analyse variationnelle seront discutés incluant l'approche incrémentale et l'utilisation de distributions non Gaussiennes pour décrire les statistiques d'erreur d'observation lorsque des erreurs grossières sont présentes. Ceci sera supporté par des résultats obtenus avec le 3D-Var opérationnel du CMC et avec le 4D-Var préopérationnel actuellement en développement au CMC.

Tuesday June 10 • Mardi 10 juin 11:30 • 11h30 LSC 332

Chris Jones, L. Kuznetsov, University of North Carolina at Chapel Hill; K.Ide, UCLA

An assimilation scheme for Lagrangian data • Un schème d'assimilation pour des données lagrangiennes

Ocean drifters and floats gather velocity field information along their trajectories. Difficulties arise in the assimilation of Lagrangian data because the state of the prognostic model

is usually described in terms of Eulerian variables. There is no direct connection between the model variables and Lagrangian observations that carries time-integrated information. We present a method, based on the extended Kalman filter, for assimilating drifter/float positions, observed at discrete times, directly into the model. The technique is tested on point vortex flows. Its performance is compared to an alternative indirect approach in which the flow velocity, estimated from two (or more) consecutive drifter observations, is assimilated. The influence of flow features, such as saddle points of the velocity field, on the performance of the scheme is analyzed.

Les dérives et les flotteurs dans les océans recueillent de l'information sur les champs de vitesse le long de leur trajectoire. Les difficultés surgissent dans l'assimilation des données lagrangiennes parce que l'état du modèle pronostique est habituellement décrit en termes de variables eulériennes. Il n'y a aucunes connections directes entre les variables du modèle et les observations lagrangiennes qui contiennent de l'information intégrée au temps. Nous présentons une méthode basée sur le filtre prolongé de Kalman pour assimiler les positions des dérives/flotteurs, observées à des temps discrets, directement dans le modèle. La technique est examinée sur des flux ponctuels de vortex. Nous comparons sa performance à une approche alternative indirecte dans laquelle la vitesse de flux, estimée à partir de deux (ou plus) observations consécutives de dérives, est assimilée. L'influence des caractéristiques du flux, telles que des points de selle du champ de flux, sur la performance du schème est analysée.

Session/Séance 26 • Decision Theory and Bayesian Methods and Resampling • Théorie de la décision, méthodes bayésiennes et rééchantillonnage

Tuesday June 10 • Mardi 10 juin 10:30 • 10h30 LSC 234

Sohee Kang, Michael Escobar, University of Toronto

Nonparametric Bayesian curve estimation for logistic regression • Estimation non paramétrique bayésienne de la courbe en régression logistique

We propose a new method to perform a nonparametric Bayesian regression for data with binary outcomes. This is done by using a mixture of Dirichlet process as a prior distribution. This allows us to produce a density estimation of the joint distribution of the outcome variable and the covariate variables. This results in a fitted regression function which is a k mixtures of logistic regression functions weighted by functions of marginal distribution of covariates. Computations are based on Markov chain simulation analysis of Dirichlet mixture models. Simulated datasets will demonstrate the curve estimation compared with Generalized Additive model (Hastie and Tibshirani 1987).

Nous proposons une nouvelle méthode pour exécuter une régression bayésienne non paramétrique pour des données avec des résultats binaires. Ceci est fait en utilisant un mélange de processus de Dirichlet comme distribution a priori. Cela nous permet de produire une estimation de la densité de la distribution conjointe des variables réponses et des covariables. Ceci donne une fonction de régression ajustée qui est un mélange de k fonctions de régression logistiques pondérées par les fonctions des distributions marginales des covariables. Les calculs sont basés sur l'analyse de simulation de chaîne de Markov des modèles de mélange de Dirichlet. Les jeux de données simulés comparent l'estimation de la courbe au modèle additif généralisé (Hastie et Tibshirani 1987).

Tuesday June 10 • Mardi 10 juin 10:45 • 10h45 LSC 234

Lin Xue, Liqun Wang, University of Manitoba

Bayesian finite mixture model with unknown components • Le modèle de mélanges finis bayésien avec composantes inconnues

Finite mixture models are very flexible parametric models for statistical inferences. Bayesian theory is more coherent than classical theory in finite mixture models. It is applied when data come from several sub-population of a large population. In real practical world, challenge is often arisen when data set has unknown components (k). Fu and Wang(2002) have developed a numerical sampling algorithm, which can be used for unknown components of Bayesian mixture model. We apply this algorithm to Yangzu river chemical elements data set, two populations exist in the data set, various marginal posterior distributions of the parameters are also given.

Les modèles de mélange finis sont des modèles paramétriques très flexibles pour l'inférence statistique. La théorie bayésienne est plus cohérente que la théorie classique dans les modèles de mélanges finis. Elle est appliquée quand les données proviennent de plusieurs sous-populations d'une grande population. Dans le monde pratique réel, le défi apparaît lorsque les jeux de données ont des composantes inconnues (k). Fu et Wang(2002) ont développé un algorithme numérique d'échantillonnage qui peut être utilisé pour des composantes inconnues de modèle de mélange bayésien. Nous appliquons cet algorithme aux données sur les éléments chimiques de fleuve Yangzu. Deux populations existent dans le jeu de données, plusieurs distributions marginales à posteriori des paramètres sont également données.

Tuesday June 10 • Mardi 10 juin 11:00 • 11h00 LSC 234

Tao Tan

The minimax admissibility characterization of linear estimates • La caractérisation d'admissibilité minimax des estimations linéaires

Admissibility, which has been discussed by many statisticians, is one of the difficult problems. It lacks of a general and effective solution. There are many admissible estimates of an unknown parameter. According to some standards, we can choose a special one from them. In this paper, we consider the minimax admissibility characterization of linear estimates with respect to restricted multivariate regression coefficient under matrix loss function. The necessary and sufficient conditions are given for a linear estimate AY+C to be Minimax admissible in the class of nonhomogeneous linear estimates.

L'admissibilité, qui a été discutée par beaucoup de statisticiens, est un problème difficile par son manque d'une solution générale et efficace. Il existe beaucoup d'estimations de l'admissibilité d'un paramètre inconnu. Selon certaines normes, nous pouvons en choisir une spéciale. Dans cette présentation, nous considérons l'admissibilité minimax, une caractérisation des estimations linéaires par rapport au coefficient de régression multivarié sous la matrice de fonction de perte. Les conditions nécessaires et suffisantes sont données pour une estimation linéaire AY+C pour être admissible minimax dans la classe des estimations linéaires non-homogènes.

Tuesday June 10 • Mardi 10 juin 11:15 • 11h15 LSC 234

Wenyu Jiang, University of Waterloo; J.D. Kalbfleisch, University of Michigan

Resampling methods for estimating functions with U-statistic structure • Méthodes de rééchantillonnage pour la fonction d'estimation avec la structure de la statistique U

Suppose that inference about parameters of interest are to be based on unbiased estimating functions that are U-statistics of degree two. We discuss the pros and cons of various studentizations for this class of estimating functions with U-statistic form and propose studentized versions suitable for interval estimation. Normal approximations to such studentized U-statistics are directly applicable in defining confidence regions. We also consider bootstrap methods based on resampling the studentized estimating functions. These methods are compared in examples and simulations with each other and with other suggestions (Jin, Ying and Wei, Biometrika, 2001) for inference for U statistics based on resampling.

Supposons que l'inférence au sujet des paramètres d'intérêt doit être basée sur les fonctions d'estimations non biaisées qui sont des statistiques U de degré deux. Nous discutons le pour et le contre de diverses studentisations pour cette classe de fonctions d'estimations avec la forme de la statistique U et nous proposons des versions studentisées appropriées à l'estimation d'un intervalle. Les approximations normales de telles statistiques U studentisées sont directement applicables pour définir des régions de confiance. Nous considérons également des méthodes bootstrap, basées sur le rééchantillonnage, des fonctions d'estimations studentidées. Ces méthodes sont comparées entre elles dans des exemples et des simulations et aussi à d'autres méthodes (Jin, Ying et Wei, Biometrika, 2001) pour l'inférence sur des statistiques U basées sur le rééchantillonnage.

Tuesday June 10 • Mardi 10 juin 11:30 • 11h30 LSC 234

Abderazzak Mouiha, Lycée Al Wahda, Taounate, Maroc

Bootstrapping a general statistic for dependent observations • Application du bootstrap à une statistique générale avec données dépendantes

For bootstrapping dependent observations, K"unsch (1989) has proposed the moving blocks bootstrap, noted "MBB". But this method does not reproduce dependence structure presented in the original data. To remedy this problem, we propose a new bootstrap procedure, called bootstrap for stationary processes noted "BSP", this method is based on approximating a stationary process by a Markov chain of order p, where p is estimated by Akaike criterion. We give some simulations to compare the performance of our method with the K"unsch's method.

Pour faire du bootstrap sur des observations dépendantes, Kunsch (1989) a proposé le bootstrap avec blocs mobiles, noté "MBB". Cependant, cette méthode ne reproduit pas la dépendance que l'on retrouve dans les données initiales. Pour remédier à ce problème, nous proposons une nouvelle procédure de bootstrap appelée bootstrap pour processus stationnaire. Noté "BSP", cette méthode est basée sur l'approximation d'un processus stationnaire par une chaîne de Markov d'ordre p, où p est estimé par le critère d'Akaike. Nous présentons les résultats de quelques simulations comparant la performance de notre méthode à la méthode de Kunsch.

Session/Séance 27 • Statistical Issues in Modern Biology • Considérations statistiques en biologie moderne

Tuesday June 10 • Mardi 10 juin 1:30 • 13h30 LSC 240

Jenny Bryan, University of British Columbia

Gene classification and clustering with time course data • Classification des gènes et groupement avec des données de temps de course

We will talk about statistical approaches to gene classification and clustering based on time course data from two high throughput experimental platforms: (1) gene expression profiling with DNA microarrays and (2) phenotypic studies of strains in a deletion set. Both technologies provide powerful tools for gaining insight into the functional roles of individual genes and for identifying putative targets of test compounds. We propose the application of appropriate regression analysis methods to first summarize intrinsic properties of each gene with a vector of regression parameters. Gene clustering and/or classification can then be carried out based on these regression parameters. By acknowledging the inherent biological and experimental variation in the data and, therefore, in the estimated regression parameters, we can then provide appropriate measures of statistical significance for certain features of the gene clustering and classification results. This work draws on collaborations with Kristin Baetz and Phil Hieter (UBC, Centre for Molecular Medicine and Therapeutics, Phil Hieter Lab), Michel Roberge (UBC, Biochemistry), and Virginia Marks and Hennie van Vuuren (UBC, Centre for Wine Research, van Vuuren Lab).

Dans cette présentation, nous parlons des approches statistiques pour la classification et le groupage des gènes basé sur des données de temps de course de deux plate formes expérimentales à sorties élevées: (1) expression profilé des gènes avec des microréseaux d'ADN et (2) études phénotypiques des contraintes dans un ensemble de suppression. Les deux technologies fournissent des outils puissants pour trouver des indices du rôle fonctionnel de différents gènes et pour identifier les cibles putatives des composés de test. Nous proposons l'application de méthodes de régression appropriées pour résumer les propriétés intrinsèques de chaque gène avec un vecteur des paramètres de régression. Le groupage et/ou la classification des gènes peut maintenant être effectué, basé sur ces paramètres de régression. En reconnaissant la variation biologique et expérimentale inhérente dans les données et en conséquence dans les paramètres estimés de la régression, nous pouvons alors fournir des mesures appropriées de signification statistique pour certains aspects des résultats de groupage et de classification des gènes. Ce travail est basé sur une collaboration avec Kristin Baetz et Phil Hieter (UBC, centre pour la médecine moléculaire et la thérapeutique, Phil Hieter Lab), Michel Roberge (UBC, biochimie), Virginie Marks et Hennie van Vuuren (UBC, Centre pour la recherche sur le vin, van Vuuren Lab).

Tuesday June 10 • Mardi 10 juin 2:00 • 14h00 LSC 240

Jinko Graham, Brad McNeney, Simon Fraser University; Francoise Seillier-Moiseiwitsch, Bioinformatics Research Centre, University of Maryland

Finding recombination breakpoints in HIV molecular sequences from an individual • Trouver les points de rupture de recombinaison dans une séquence moléculaire du VIH chez un individu

Retroviral recombination is an issue in studies of HIV-1 evolution within individuals who harbor a genetically diverse virus population. The nature of retroviral replication is such that recombinant HIV-1 genomes are expected to be a complex mosaic of parental sequences. However, given the comparatively low levels of virus diversity within an individual, relatively few recombination breakpoints are expected to be detectable. Under these circumstances, a sequential search for breakpoints is expected to be effective. We consider a stepwise procedure in which additional recombination breakpoints are simultaneously tested for and estimated, given the location of previously known breakpoints. The procedure can be used to assess the number and location of recombination breakpoints in a sequence alignment from an individual. We apply the procedure to an alignment of

HIV-1 env gene sequences sampled from a patient harbouring a heterogeneous population of subtype A virus.

La recombinaison rétrovirale est un problème dans les études de l'évolution du VIH-1 chez les individus qui hébergent une population génétiquement diverse du virus. La nature de la réplication rétrovirale est telle qu'on s'attend à ce que les génomes VIH-1 recombinés soient une mosaïque complexe des séquences parentales. Cependant, donné les niveaux comparativement bas de la diversité du virus chez un individu, on s'attend à ce que relativement peu de points de rupture de recombinaison soient discernables. Dans ces circonstances, on s'attend à ce qu'une recherche séquentielle des points soit efficace. Nous considérons un procédé pas à pas où des points de rupture recombinaison additionnels sont simultanément testés et estimés, étant donné l'endroit des points de rupture connus précédemment. Le procédé peut être employé pour évaluer le nombre et l'endroit des points de rupture de recombinaison dans un alignement de séquences moléculaires provenant d'un individu. Nous appliquons le procédé à un alignement de séquences du gène env VIH-1 échantillonnées chez un patient hébergeant une population hétérogène du virus de sous-type A.

Tuesday June 10 • Mardi 10 juin 2:30 • 14h30 LSC 240

Julie Horrocks, University of Guelph

Joint models for longitudinal data and time-to-event data with multiple outcomes • Modèles conjoints pour des données longitudinales et de temps jusqu'à un événement avec plusieurs résultats possibles

In this talk, I discuss joint models for longitudinal and time-to-event data with multiple outcomes. This work has applications for instance to length of stay in an Intensive Care Unit (ICU). Here the outcomes of interest are discharge, transfer or death, and the longitudinal data may represent body temperature, cytokine levels, or presence of encephalopathy. Note that only one type of event can be observed for each individual; in other words, the individual is subject to competing risks. Previous work in this area includes Horrocks and Thompson (in press), who proposed a regression model for time-to-event data with multiple outcomes, based on a latent Wiener process with drift. Only time-fixed covariates were considered in this work. On another tack, Wang and Taylor (JASA 2001) proposed a joint model, incorporating an IOU process for the longitudinal data and a Cox model for time-to-event data with a single outcome. In this talk we will compare joint modelling to other common approaches, such as separate modelling of the longitudinal and time-to-event data, and joint modelling of longitudinal data and time to a single outcome, treating the other outcomes as censored.

Dans cette présentation, nous discutons des modèles conjoints pour des données longitudinales et de temps de survie avec des résultats multiples. Un exemple d'applications est la durée du séjour dans une unité de soins intensifs (USI). Ici, les résultats d'intérêt sont la décharge, le transfert ou la mort, et les données longitudinales peuvent représenter la température corporelle, le niveau de cytokine ou la présence de l'encéphalopathie. Notons que nous pouvons observer seulement un type d'événement pour chaque individu; en d'autres termes, l'individu est sujet à des risques de concurrence. Les travaux précédents dans ce domaine incluent Horrocks et Thompson (sous presse), qui ont proposé un modèle de régression pour des données de survie avec des résultats multiples, basés sur un processus latent de Wiener avec dérive. Seulement les covariables à temps fixes ont été considérés dans ce travail. D'un autre point de vue, Wang et Taylor (JASA 2001) ont proposé un modèle conjoint, incorporant un processus IOU pour les données longitudinales et un modèle

de Cox pour des données de survie avec des résultats simples. Dans cette présentation, nous comparons les modèles conjoints à d'autres approches courantes, telles que la modélisation séparé des données longitudinales et de survies et la modélisation conjointe des données longitudinales et des temps de survie à résultats simples, traitant les autres résultats comme des temps censurés.

Session/Séance 28 • Resampling Methods • Méthodes de rééchantillonnage

Tuesday June 10 • Mardi 10 juin 1:30 • 13h30 LSC 242

Michael Sherman, Texas A&M University

Nonparametric resampling for spatial data • Rééchantillonnage non paramétrique pour des données spatiales

In spatial statistics data are typically correlated which makes the usual methods of analysis inappropriate. Nonpamametric resampling is proposed as a method to assess distributional properties of estimators computed from spatial data. We discuss results for both equally spaced and unequally spaced data. In the latter case the data form a realization of a marked point process. We formulate subsampling estimators of the moments of general statistics computed from marked point process data, and establish their large sample properties. In particular, the variance estimator can be used for the construction of confidence intervals for estimated parameters. We discuss a data-based method of choosing a subsampling parameter and illustrate methods on several data sets. In a specific application we develop a resampling based test of spatial isotropy.

En statistiques spatiales, les données sont typiquement corrélées ce qui rend les méthodes d'analyse habituelles inadéquates. On propose le rééchantillonnage non paramétrique comme méthode pour évaluer certaines propriétés distributionnelles d'estimateurs calculés à partir de données spatiales. Nous discutons des résultats pour des données équidistantes ou inégalement espacées. Dans le dernier cas, les données forment une réalisation d'un processus ponctuel marqué. Nous formulons des estimateurs de sous-échantillonnage des moments de statistiques générales calculées à partir des données d'un processus ponctuel marqué, et nous établissons leurs propriétés pour des grands échantillons. En particulier, l'estimateur de la variance peut être utilisé pour la construction d'intervalles de confiance pour les paramètres estimés. Nous discutons une méthode basée sur les données pour choisir un paramètre de sous-échantillonnage et nous illustrons la méthode à partir de plusieurs jeux de données. Nous développons, pour un cas spécifique, un test de rééchantillonnage d'isotropie spatiale.

Tuesday June 10 • Mardi 10 juin 2:00 • 14h00 LSC 242

Subhash Lele, University of Alberta

Impact of bootstrap on estimating functions • L'effet du bootstrap sur les fonctions d'estimations

Estimating functions form an attractive statistical methodology because of their dependence only on a few features of the underlying probabilistic structure. It also puts premium on developing methods that obtain model robust confidence intervals. Bootstrap and Jackknife ideas can be fruitfully used towards this purpose. In this article, I review, compare and contrast various approaches for bootstrapping estimating functions.

Les fonctions d'estimations forment une méthodologie statistique attrayante en raison de leur dépendance à peu d'aspect de la structure probabiliste sous jacente. Elles mettent également l'accent sur le développement de méthodes qui obtiennent des intervalles de confiances robustes sur le modèle. Des idées provenant du bootstrap et du jackknife peuvent aussi être utilisées vers ce but. Dans cette présentation, nous passons en revue, comparons et contrastons diverses approches pour faire du bootstrap sur des fonctions d'estimations.

Tuesday June 10 • Mardi 10 juin 2:30 • 14h30 LSC 242

Angelo Canty, A. R. Padmanabhan, McMaster University

A robust bootstrap test for the equality of several medians • Un test bootstrap robuste pour l'égalité de plusieurs médianes

Babu and Padmanabhan (2002) proposed a robust bootstrap solution to the Behrens-Fisher problem. We show that this method can be extended to the test of equality of several medians. The basic assumption is that the data come from a location-scale family but no assumptions on the shape of the underlying distribution are made. An extensive Monte Carlo simulation study will be presented to illustrate the power of the test against "umbrella alternatives" (Mack & Wolfe, 1981) and "pattern alternatives" (Hettmansperger & Norton, 1987). We shall also give some theoretical results to justify the method asymptotically. These results constitute a multi-sample generalization of those in Babu and Padmanabhan (2002) and the asymptotics follow along the same lines as in that paper.

Babu et Padmanabhan (2002) ont proposé une solution bootstrap robuste au problème de Behrens-Fisher. Nous prouvons que cette méthode peut être appliquée pour tester l'égalité de plusieurs médianes. L'hypothèse de base est que les données proviennent d'une famille d'échelle locale, mais aucunes hypothèses ne sont faites sur la forme de la distribution sous jacente. Une étude de simulation de Monte Carlo est présentée pour illustrer la puissance du test contre les "alternatives de parapluie" (Mack et Wolfe, 1981) et les "alternatives de pattern" (Hettmansperger et Norton, 1987). Nous donnons également quelques résultats théoriques pour justifier asymptotiquement la méthode. Ces résultats constituent une généralisation multi-échantillonnale de ceux dans Babu et Padmanabhan (2002) et les résultats asymptotiques suivent le même déroulement que dans cet article.

Session/Séance 29 • Experimental Design • Plans d'expérience

Tuesday June 10 • Mardi 10 juin 1:30 • 13h30 LSC 238

Holger Dette, S. Biedermann, V.B. Melas, Ruhr-Universitaet Bochum

Efficient designs for regression models in microbiology • Plans d'expérience efficaces pour les modèles de régression en microbiologie

In this talk the estimation problem and the problem of designing experiments in a nonlinear regression model, used in microbiology, are studied. The model is called Monod model, defined imlicitly by a differential equation for the regression function and has numerous applications in microbial growth kinetics, water research, pharmacokinetics and plant physiology. The asymptotic covariance matrix of the least squares estimator is the basis for the construction of efficient designs of experiments. In particular locally D-, E-and c-optimal designs are determined and their properties are studied. Moreover the performance of the designs (determined by the asymptotic theory) is confirmed in simulation experiments for realistic sample sizes. If certain intervals for the nonlinear parameters can be specified

based on microbiological background, locally optimal designs can be constructed, which are robust with respect to misspecification of the initial parameters and which allow efficient estimation of the parameters in the Monod model. The results indicate that parameter variances can be decreased by a factor two by simply sampling at optimal times during the experiment.

Dans cette présentation, nous étudions le problème d'estimation et de planification des expériences avec un modèle non-linéaire de régression, souvent utilisé en microbiologie. Le modèle s'appel le modèle de Monod et est défini implicitement par une équation différentielle pour la fonction de régression. Ce modèle a de nombreuses applications en cinétique de croissance microbiennes, en recherche sur l'eau, en pharmacocinétique et en physiologie des plantes. La matrice de covariance asymptotique de l'estimateur des moindres carrés sert de base à la construction de plans d'expériences efficaces. En particulier des plans D-, E- et C-optimaux localement sont déterminées et nous étudions leurs propriétés. De plus, la performance des plans (déterminées par la théorie asymptotique) est confirmée par des simulations pour des échantillons de taille réaliste. Si certains intervalles pour les paramètres non-linéaires peuvent être spécifiés, basés sur le cadre micro biologique, des plans localement optimaux peuvent être construits en étant robustes à la mauvaise spécification des paramètres initiaux et qui permettent l'estimation efficace des paramètres dans le modèle de Monod. Les résultats indiquent que la variance des paramètres peut être diminué par un facteur deux, tout simplement en échantillonnant à des temps optimaux pendant l'expérience.

Tuesday June 10 • Mardi 10 juin 2:00 • 14h00 LSC 238

Rainer Schwabe, Otto-von-Guericke-Universität

Efficient and adaptive design and analysis in forced choice experiments • Design efficace et adaptatif et l'analyse d'expériences avec choix forcés

Psychophysical properties like visual acuity, visual or auditory ability etc. can often be evaluated by indirect measurements of sensitivity only. In such situations dichotomous answers are available whether a stimulus has been correctly observed or not. Stimuli are presented at various well-defined intensities and thresholds have to be determined from those answers. Based on the comparison of individual thresholds with normal values pathologies are to be detected. In forced choice experiments, where the experimental subject has to choose between two or more alternatives, an additional problem occurs, since it is possible to obtain a "correct" decision by pure guessing. In ophthalmological examinations of the visual acuity, for example, tasks are offered in which the experimental subject is forced to decide between two different shapes. This leads to a "correct" decision in half of the cases even if the shapes are indistinguishable. The aim is to develop a procedure which leads to reliable thresholds with a reasonable number of stimulus presentations even for unfavorable intitial values which are far from the true threshold. We propose an adaptive procedure for determining the stimuli to be presented based on the concept of information maximization. The performance of this procedure will be illustrated by a simulation study.

Les propriétés psychophysiques comme l'acuité visuelle et les capacités auditives ou visuelles peuvent souvent être évaluées seulement par des mesures indirectes de sensibilité. Dans de telles situations, les réponses dichotomiques sont disponibles, qu'on ait correctement observé un stimulus ou non. Des stimuli sont présentés à diverses intensités bien définies et des seuils doivent être déterminés à partir de ces réponses. Basés sur la comparaison de différents seuils avec des valeurs normales, des pathologies doivent être détectées. Dans des

expériences avec choix obligatoires, c'est-à-dire où le sujet soumit à l'expérience doit absolument choisir entre deux alternative ou plus, un problème additionnel se produit puisqu'il est possible d'obtenir la bonne réponse par hasard. Dans les examens ophtalmologiques de l'acuité visuelle par exemple, les questions sont de la forme où le sujet est forcé de décider entre deux formes différentes. Ceci mène à une bonne réponse dans la moitié des cas, et ce, même si les formes ne sont pas distinguables. Notre but est de développer un procédé qui mène à des seuils fiables avec un nombre raisonnable de stimulus présentés, même pour des valeurs initiales défavorables loin du vrai seuil. Nous proposons une procédure adaptative pour déterminer les stimulus à présenter basés sur le concept de la maximisation de l'information. Nous illustrons ce procédé par une étude avec simulation.

Tuesday June 10 • Mardi 10 juin 2:30 • 14h30 LSC 238

Julie Zhou, University of Victoria; Hongtu Zhu, Yale University

Robust experimental designs for the random effects model • Plans d'expériences robustes pour le modèle à effets aléatoires

Experimental designs for estimating variance components in the random effects model are often based on non-robust estimation methods, such as the traditional analysis of variance and maximum likelihood estimation. However, in the presence of outliers in the data, we need to apply robust methods to obtain reliable estimates for variance components. Thus it is natural to study experimental designs based on robust estimation methods. In this talk, one robust estimator will be discussed, and its finite-sample breakdown point will be explored. A new criterion for selecting optimal robust designs is proposed based on the mean squared errors for the robust estimates and the finite-sample breakdown point. Several examples will be given to show optimal robust designs.

Les plans d'expériences pour estimer des composantes de variance dans le modèle à effets aléatoires sont souvent basés sur des méthodes d'estimation non-robustes, telles que l'analyse de la variance traditionnelle et l'estimation par le maximum de vraisemblance. Cependant, en présence de valeurs aberrantes dans les jeux de données, nous devons appliquer des méthodes robustes pour obtenir des évaluations fiables pour les composantes de variance. Ainsi, il est normal d'étudier des plans d'expériences basés sur des méthodes d'estimation robustes. Dans cette présentation, un estimateur robuste sera présenté, et son point de rupture pour des échantillons finis sera exploré. Nous proposons un nouveau critère pour choisir des plans robustes optimaux basés sur l'erreur carré moyen pour les estimations robustes et le point de rupture pour des échantillons finis. Plusieurs exemples seront présentés pour illustrer des plans robustes optimaux.

Session/Séance 30 • Survey Methods Contributed Session III: Methods for Health Surveys • Méthodes d'enquête III: Méthodes pour les enquêtes sur la santé

Tuesday June 10 • Mardi 10 juin 1:30 • 13h30 LSC 338

Larry MacNabb, Statistics Canada/Statistique Canada

Application of cluster analysis towards the development of health region peer groups • Application de l'analyse de regroupement vers le développement de la santé des groupes de pairs en région

The inception of the Canadian Community Health Survey in conjunction with the expansion of existing data products for the provision of health region level information has

necessitated the need to develop a method of comparing regions with similar socio-economic determinants of health. After the effects of the various social and economic characteristics known to influence health status have been removed it then becomes possible to compare the health status of regions in the same "peer group" and measure the relative effectiveness of the many health promotion and prevention strategies employed across regions. Cluster analysis was used to place 139 health regions into 10 distinct groupings or "peer groups" possessing similar socio-economic characteristics. Health regions were defined using 24 variables chosen to cover as many of the known social and economic determinants of health as possible. This paper will outline the data sources used to define the peer groups as well as the practical constraints and criteria placed on the analysis. Methods used will be outlined and results will be presented along with the obstacles encountered in the actual application of the methodology.

Le commencement de l'enquête canadienne sur la santé de la communauté en conjonction avec l'expansion des sources existantes de données pour la santé au niveau régional a rendu nécessaire le développement d'une méthode pour comparer les régions avec des déterminants socio-économiques de la santé. Après que les effets des diverses caractéristiques sociales et économiques connues pour influencer l'état de santé ont été enlevés, il devient alors possible de comparer l'état de santé des régions dans le même groupe de pairs (peer group) et de mesurer l'efficacité relative des nombreuses stratégies de promotion et de prévention de la santé utilisées à travers les régions. L'analyse de regroupement est utilisée pour regrouper 139 régions de santé dans 10 groupes de pairs qui possède des caractéristiques socio-économiques semblables. Les régions de santé sont définies en utilisant 24 variables choisies pour couvrir le plus de causes sociales et économiques connues déterminantes pour la santé. Cette présentation décrit les sources de données utilisées pour définir les groupes de pairs ainsi que les contraintes pratiques et les critères faits sur l'analyse. Les méthodes utilisées sont décrites et les résultats sont présentés avec les obstacles encourus lors de l'application réelle de la méthode.

Tuesday June 10 • Mardi 10 juin 1:45 • 13h45 LSC 338

Yves Béland, Johane Dufour, Larry MacNabb, Statistics Canada/Statistique Canada

Sample Design of the 2004 Canadian Nutrition Survey • Plan d'échantillonnage du sondage canadien sur la nutrition de 2004

As part of the Canadian Community Health Survey (CCHS) biennial strategy, the provincial survey component of the second cycle of the CCHS will focus on nutrition related issues of Canadians of all ages living in private dwellings. This new survey will collect information using a 24-hour dietary recall, with repeats, approach in order to estimate the usual dietary intake of the Canadian population. This in turn will allow for the estimation of distribution patterns in regard to the consumption of key nutrients required for optimal health. As well, data on selected psycho-social determinants of eating behaviour, food insecurity and some anthropometric measurements for body weight measurement will also be collected. All this will be rounded out with the collection of a series of health status (chronic conditions, general health, etc.), health determinants (smoking, alcohol, physical activity, etc.) and socio-demographic characteristics. To meet the requirement of producing intake distributions for 15 key age-sex domains of interest a sample of 30,000 respondents will be selected from two sample frames. The main part of the sample of individuals will be chosen from households selected from an area frame under a multistage stratified cluster design where the dwelling is the final sampling unit. In some provinces the households selected

as part of the regional component of the CCHS (conducted in 2003) will also be used as a sample frame. Only one person (aged 0 and up) per household will be selected at random with varying age-based probabilities of selection and a 45-minute face-to-face interview including the 24-hour recall will be conducted using a computer-assisted application. Between three to ten days after the interview there will be a second 24-hour recall interview conducted over the phone in order to estimate the intra-subject variability for adjusting the intake distributions for a sub-sample of 10,000 individuals. Data collection will begin in January 2004 and will extend over 12 months to eliminate seasonal effects. This paper will describe several key aspects of the sample design of this new and exciting survey as well as some challenges encountered in the testing and validation of the questionnaire.

Faisant partie de la stratégie bi-annuelle du sondage canadien sur la santé des communautés (CCHS), les composantes provinciales du deuxième cycle du sondage du CCHS se concentrent sur les questions reliées à la nutrition des Canadiens de tous les âges vivant dans des logements privés. Ce nouveau sondage va rassembler de l'information en utilisant une approche de rappel diététique des dernières 24 heures, avec répétitions, afin d'estimer l'apport alimentaire habituelle de la population canadienne. Ceci va permettre d'estimer les modèles de distribution par rapport à la consommation des aliments principaux requis pour une santé optimale. De plus, des données sur certaines causes psychosociales déterminantes des habitudes de consommation, de l'insécurité à la nourriture et de quelques mesures anthropométriques pour la mesure du poids corporel seront également rassemblées. Le tout sera agrémenté par la collection d'une série d'états de santé (conditions chroniques, santé générale, etc.), de déterminants de la santé (cigarette, alcool, activité physique, etc...) et de caractéristiques socio-démographiques. Pour répondre à l'exigence de produire des distributions de l'apport alimentaire pour 15 principaux domaines d'intérêt âge-sexe, un échantillon de 30 000 répondants sera choisi à partir de deux cadres d'échantillonnage. La grande partie des individus échantillonés sera sélectionné à partir de ménages choisis dans un cadre de secteur sous un design stratifiée en grappe à plusieurs étapes où le logement est l'unité de sélection finale. Dans quelques provinces, les ménages choisis en tant qu'élément de la composante régionale du CCHS (conduit en 2003) seront également utilisés comme cadre échantillonnale. Seulement une personne (âgée de 0 et plus) par ménage sera choisie au hasard avec des probabilités de sélections qui changent avec l'âge. Une entrevue de 45 minutes en tête à tête incluant le rappel de 24 heures seront conduits en utilisant une application assistée par ordinateur. De trois à dix jours suite à l'entrevue, une deuxième entrevue avec un rappel sur les 24 dernières heures sera faite par téléphone afin d'estimer la variabilité intra-sujet pour ajuster les distributions d'apport alimentaire à un sous-échantillon de 10.000 individus. La collecte de données commence en janvier 2004 et se prolongera sur plus de 12 mois pour éliminer les effets saisonniers. Cette présentation décrit plusieurs aspects principaux du design d'échantillonnage de ce nouveau et passionnant sondage, ainsi que quelques lors du test et de la validation du questionnaire.

Tuesday June 10 • Mardi 10 juin 2:00 • 14h00 LSC 338

Amanda Lafontaine, Lehana Thabane, Aaron Childs, McMaster University

Determining the level of statisticians' participation in Canadian based research ethics committees • Détermination du niveau de participation des statisticiens dans les comités d'éthiques de recherches canadiens

The role of Research Ethics Committees (RECs) includes the assessment of ethical and scientific validity of research and guarding the safety and protection of research partici-

pants. In determining the scientific validity of the research, input from a statistician can be very helpful to assess such things as research designs, determination of sample sizes and methods of analysis. However, there is little or no information on the level of participation of statisticians in Canadian based RECs. In this presentation, we report on the findings of a national survey of Canadian based RECs. The primary outcome of this study is to determine the level of participation of statisticians in RECs as measured by the proportion of RECs with statisticians in their membership.

Le rôle des comités d'éthique de recherches (CER) comprend l'évaluation de la validité éthique et scientifique de la recherche et le rôle de gardien de la sécurité et la protection des participants de recherches. Pour déterminer la validité scientifique de la recherche, l'implication d'un statisticien peut être très utile pour évaluer des aspects tels le design de la recherche, la détermination de la taille de l'échantillon et les méthodes d'analyse. Cependant, il y a peu ou pas d'information sur le niveau de participation des statisticiens dans les CER canadiens. Dans cette présentation, nous rendons les résultats d'un sondage national sur les CER canadiens. Le but principal de cette étude est de déterminer le niveau de participation des statisticiens dans les CER qui est mesuré par la proportion de CER avec des statisticiens membres.

Tuesday June 10 • Mardi 10 juin 2:15 • 14h15 LSC 338

François Brisebois, Patrice Mathieu, Statistics Canada/Statistique Canada

Creation of a new longitudinal weight for the Canadian National Population Health Survey: providing data users with greater analytical flexibility • Création d'un nouveau poids longitudinal pour l'Enquête nationale sur la santé de la population canadienne: Une plus grande flexibilité analytique pour les utilisateurs de données

The National Population Health Survey (NPHS) is a longitudinal survey designed to collect information on the health of the Canadian population and related socio-demographic characteristics. The NPHS panel of 17,276 people was first interviewed in 1994-95, and it is to be recontacted every second year for 20 years. Although nonresponse has been considerably low up until now, the panel faces a certain attrition that reduces, cycle after cycle, the number of records with fully completed data for all cycles. Until 2000-01, two longitudinal weights were provided to data users for their analysis. One was attributed to all 17,276 panel members, no matter their interview outcome, and is used mainly to study nonresponse patterns. A second weight was attributed only to the subset of panel members who had fully responded to all cycles of the survey, which is more convenient for data analyses since it excludes all nonrespondents. Since the attrition constantly increases, the size of this subset decreases with each cycle, making the sample available for analysis smaller and smaller. To overcome this problem, a third weight was created in 2000-01; the idea is to attribute this weight to all panel members who fully completed the 1994-95 interview, as well as the most recent one, which in this case was in 2000-01. This new subset allows data users to gain in sample size when their analyses focus exclusively on variables collected during these specific cycles. The justification and methodology used to create this weight will be presented as well as some results obtained from the comparisons of all three NPHS weights now available.

L'Enquête nationale sur la santé de la population (ENSP) est une enquête longitudinale visant à recueillir des renseignements sur la santé de la population canadienne ainsi que des renseignements socio-démographiques connexes. Le panel de l'ENSP, composé de 17 276 personnes, a été interviewé pour la première fois en 1994-1995 et est recontacté à tous les

deux ans pour une période de vingt ans. Malgré que la non-réponse ait été considérablement faible jusqu'à maintenant, le panel fait quand même face à une certaine érosion qui, cycle après cycle, réduit le nombre d'enregistrements avec des données complètes à tous les cycles. Jusqu'en 2000-2001, deux poids longitudinaux étaient fournis aux utilisateurs de données pour faire leurs analyses. Le premier était attribué aux 17 276 membres du panel, peu importe leur statut d'interview, et est principalement utile pour l'analyse du profil des non-répondants. Un second poids était attribué seulement au sous-ensemble de membres du panel ayant fourni une réponse complète à tous les cycles; ce poids est plus pratique pour l'analyse de données puisqu'il exclut tous les non-répondants. Puisque l'érosion augmente constamment, la taille de ce sous-ensemble diminue à chaque cycle et rend donc l'échantillon disponible à des fins d'analyse de plus en plus petit. Pour surmonter ce problème, un troisième poids a été créé en 2000-2001; l'idée est d'attribuer ce poids à tous les membres du panel ayant fourni une réponse complète à l'interview de 1994-1995, de même qu'à la plus récente interview, qui dans ce cas était en 2000-2001. Ce nouveau sous-ensemble permet aux utilisateurs de données d'augmenter leurs tailles d'échantillon lorsque leurs analyses réfèrent exclusivement aux variables recueillies aux cycles en question. La justification de même que la méthodologie utilisée pour créer ce poids seront présentées, ainsi que quelques résultats obtenus lors de la comparaison des trois poids maintenant disponible avec l'ENSP.

Tuesday June 10 • Mardi 10 juin 2:30 • 14h30 LSC 338

Dany Faucher, Éric Langlet, Statistics Canada/Statistique Canada

An application of the bootstrap variance estimation method to the Participation and Activity Limitation Survey • Une application du boostrap pour l'estimation de la variance dans le cadre de l'Enquête sur la Participation et les Limitations d'Activités

The bootstrap method is more and more commonly used to estimate the variance of estimates obtained from complex survey designs. This method offers the advantage of being applicable to virtually any type of estimates. Also, the use of bootstrap weights in microdata files eliminates the need of keeping strata and Primary Sampling Units (PSU's) identifiers, which reduces the risk of disclosure. The sampling plan of the Participation and Activity Limitations Survey (PALS) is a stratified two-stage design in which PSU's are selected without replacement with probability proportional to their estimated sizes. The survey presents specific challenges to the application of the bootstrap method. For instance, the sampling fraction for PALS is relatively high in many strata, which causes the bootstrap method to over-estimate the variance. Also, a logistic regression response propensity model is used for the non-response adjustment in PALS. Should a logistic regression model be fitted on each bootstrap sample for the non-response adjustment? This paper will address these issues as well as other particular problems.

Le bootstrap est une méthode de plus en plus utilisée pour l'estimation de la variance pour des enquêtes avec des plans de sondage complexes. Cette méthode offre l'avantage d'être applicable pour le calcul de n'importe quel type d'estimation. De plus, l'utilisation des poids bootstrap dans les fichiers de micro-données permet d'éviter le besoin de fournir les identificateurs de strate et d'Unité Primaire d'Échantillonnage (UPÉ), ce qui réduit le risque de divulgation d'information confidentielle. Le plan de sondage de l'Enquête sur la Participation et les Limitations d'Activités (EPLA) est un plan stratifié à deux degrés dans lequel les UPÉ sont sélectionnées sans remise avec probabilité proportionnelle à leur taille estimée. Cette enquête offre des défis particuliers dans l'application de la

méthode du bootstrap. D'abord, dans plusieurs strates de l'EPLA, la fraction de sondage est relativement élevée, ce qui fait en sorte que la méthode bootstrap aura tendance à surestimer la variance. Ensuite, un modèle de régression logistique est utilisé pour l'ajustement de la non-réponse pour l'EPLA. Doit-on ajuster un modèle de régression logistique à chacun des échantillons bootstrap pour l'ajustement de la non-réponse? Cet article traitera entre autres de ces questions ainsi que d'autres problèmes particuliers.

Session/Séance 31 • Robust Methods I • Méthodes robustes I

Tuesday June 10 • Mardi 10 juin 1:30 • 13h30 LSC 332

Adeniyi Adewale, Douglas P. Wiens, University of Alberta

Robust designs for approximate regression models with two interacting regressors • Design robuste pour les modèles de régression approximatif avec deux régresseurs interagissant

Classical experimental design theory hinges on the assumption that the regression mean response is exactly correct and that errors are uncorrelated and have a constant variance. The former assumption is often a far cry from reality. Box and Draper (1959) highlighted the inherent dangers in holding to this assumption when it is not valid. In this work, we construct designs under the assumption that the mean response may be approximated by a linear combination of terms X1, X2, X1*X2 (two interacting regressors) and that the difference between the best such linear combination, and the true response, is bounded in L2 space. We measure the quality of estimates of the mean response by using functions of the Mean Squared Error matrix. In particular,we examine the Integrated Mean Squared Error, Determinant of Mean Squared Error and Trace of Mean Squared Error as overall measures of loss. Using the minimax approach we construct optimal designs which are robust against model misspecification. We employ Lagrange multipliers to determine the form of the optimal density corresponding to each measure of overall loss, and using numerical methods we derive explicit expressions for the optimal densities. Finally, we illustrate the implementations of the optimal designs.

*La théorie de design expérimental classique s'articule sur l'hypothèse que la réponse moyenne de régression est exactement correcte et que les erreurs sont non-corrélées et ont une variance constante. L'hypothèse précédente est souvent loin de la réalité. Box et Draper (1959) ont fait ressortir les dangers inhérents à se tenir à cette prétention lorsqu'elle est inadmissible. Dans cette présentation, nous construisons des designs sous l'hypothèse que la réponse moyenne peut être estimée par une combinaison linéaire des termes X1, X2, X1*X2 (deux régresseurs interagissant) et que la différence entre le meilleur d'une de ces combinaisons linéaires, et la vraie réponse, est borné dans l'espace L2. Nous mesurons la qualité des estimations de la réponse moyenne en utilisant des fonctions de la matrice de l'erreur quadratique moyenne. En particulier, nous examinons l'erreur quadratique moyenne intégrée, le déterminant de l'erreur quadratique moyenne et la trace de l'erreur quadratique moyenne comme des mesures de pertes globales. En utilisant l'approche du minimax nous construisons des designs optimaux qui sont robustes à la mauvaise spécification du modèle. Nous utilisons des multiplicateurs de Lagrange pour déterminer la forme de la densité optimale correspondant à chaque mesure de pertes globales, et en utilisant des méthodes numériques, nous dérivons des expressions explicites pour les densités optimales. En conclusion, nous illustrons l'implémentation des designs optimaux.*

Tuesday June 10 • Mardi 10 juin 1:45 • 13h45 LSC 332

Joanna Flemming, Elvezio Ronchetti, Eva Cantoni, University of Geneva

Model selection for marginal longitudinal generalized linear models • Sélection de modèle pour les modèles linéaires généralisés logitudinaux marginaux

Model selection is an essential part of any statistical analysis and yet has been somewhat neglected in the context of longitudinal data analysis. Mallow's C_p is an effective model selection procedure in regression. Here we propose a generalized version of C_p (GC_p) suitable for use with both parametric and nonparametric models. GC_p provides an estimate of the measure of adequacy of a model for prediction. We examine it's performance with popular marginal longitudinal models (fitted using GEEs) and contrast results with what is typically done in practice: variable selection based on Wald-type tests. An application to real data further demonstrates the merits of our approach while at the same time emphasizing some important robust features inherent to GC_p.

La sélection du modèle est une partie essentielle de n'importe quelle analyse statistique, mais a été légèrement négligé dans le contexte de l'analyse de données longitudinales. Le C_p de Mallow est une procédure de sélection du modèle efficace en régression. Dans notre situation, nous proposons une version généralisée du C_p (GC_p), approprié pour les modèles paramétriques et non paramétriques. Le GC_p fournit une estimation de la mesure de qualité d'ajustement d'un modèle pour la prévision. Nous examinons sa performance avec des modèles logitudinaux marginaux populaires (adaptés en utilisant GEE) et nous mettons en contraste les résultats avec ce qui est typiquement fait en pratique: la sélection des variables basé sur des tests de type Wald. Une application à des données réelles démontre les mérites de notre approche tout en soulignant quelques caractéristiques robustes importantes au sujet du GC_p.

Tuesday June 10 • Mardi 10 juin 2:00 • 14h00 LSC 332

Pierre Duchesne, HEC Montréal

Robust and powerful serial correlation tests with new robust estimates in ARX models • Tests de corrélation sérielle robustes basés sur de nouveaux estimateurs dans les modèles ARX

We consider robust serial correlation tests in autoregressive models with exogenous variables (ARX). Since the least squares estimators are not robust when outliers are present,

a new family of estimators is introduced. They provide resistant estimators that are less sensitive to abnormal observations in the output variable of the dynamic model. We show that the new robust estimators are consistent and we can consider robust and powerful tests of serial correlation in ARX models based on these estimators. The new one-sided tests of serial correlation are obtained in extending Hong's (1996) approach in a framework resistant to outliers. They are based on a weighted sum of robust squared residual autocorrelations and on any robust and sqrt(n)-consistent estimators. Our approach generalizes Li's (1988) test statistic, that can be interpreted as a test using the truncated uniform kernel. However, many kernels deliver a higher power. This is confirmed in a simulation study, where we investigate the finite sample properties of the new robust serial correlation tests in comparison to some commonly used robust and non-robust tests.

Nous considérons des tests de corrélation sérielle robustes dans les modèles autorégressifs avec variables exogènes (ARX). Puisque les estimateurs des moindres carrés ne sont pas

robustes, une nouvelle famille d'estimateurs est introduite. Ces nouveaux estimateurs sont plus résistants et ils sont moins sensibles aux observations aberrantes qui pourraient survenir dans la variable dépendante du modèle dynamique. Nous montrons que les nouveaux estimateurs sont convergents et nous pouvons avec ces derniers contruire des tests robustes et puissants de corrélation sérielle dans le cadre des modèles ARX. Les nouveaux tests unilatéraux sont obtenus en généralisant l'approche de Hong (1996) dans un cadre de travail robuste aux valeurs aberrantes. Les nouveaux tests reposent sur une somme pondérée d'autocorrélations carrées basées sur les résidus, ainsi que sur des estimateurs racine(n)-convergents. Notre approche généralise la statistique de test de Li (1988), qui peut être interprétée en utilisant un noyau particulier, le noyau uniforme tronqué. Cependant, bien des noyaux procurent une puissance plus élevée. Ceci est confirmé dans une étude de simulation, où nous comparons les nouveaux tests robustes avec d'autres tests robustes et non-robustes.

Tuesday June 10 • Mardi 10 juin 2:15 • 14h15 LSC 332

Sanjoy Sinha, University of Winnipeg

Robust inference in generalized linear mixed models • Inférence robuste pour les modèles linéaires mixtes généralisés

The method of maximum likelihood (ML) is widely used for analyzing generalized linear mixed models (GLMM's). A full maximum likelihood analysis requires numerical integration techniques for calculation of the log-likelihood, and to avoid the computational problems involving irreducibly high-dimensional integrals, several maximum likelihood algorithms had been proposed in the literature to estimate the model parameters by approximating the log-likelihood function. While these likelihood algorithms are useful in fitting the GLMM's efficiently under strict model assumptions, these can be highly influenced by the presence of unusual data points. In this paper, the author develops a robust Monte Carlo Newton-Raphson (RMCNR) algorithm for fitting GLMM's, which appears to be useful in downweighting the influential data points when estimating the model parameters. The asymptotic properties of the RMCNR estimators are investigated under some regularity conditions. Small simulations are carried out to study the behavior of the robust estimates in the presence of outliers, and these estimates are also compared to the ordinary classical estimates. The method is illustrated in an analysis of data from a clinical experiment described in the biometrical journal.

La méthode du maximum de vraisemblance (MV) est largement utilisée pour analyser les modèles linéaires mixtes généralisés (GLMM). Une analyse complète du maximum de vraisemblance exige des techniques d'intégration numérique pour le calcul de la fonction log-vraisemblance. Pour éviter des problèmes informatiques impliquant des intégrales irréductibles de grande dimension, plusieurs algorithmes de maximum de vraisemblance ont été proposés dans la littérature pour estimer les paramètres du modèle en estimant la fonction de log-vraisemblance. Tandis que ces algorithmes de probabilité sont utiles pour ajuster le GLMM efficacement dans des hypothèses strictes du modèle, ceux-ci peuvent être fortement influencés par la présence de points à l'écart. Dans cette présentation, nous développons un algorithme robuste de Monte Carlo Newton-Raphson (RMCNR) pour ajuster le modèle linéaire mixte généralisé, qui semble être utile pour enlever du poids aux points influents lors de l'estimation des paramètres du modèle. Les propriétés asymptotiques des estimateurs de RMCNR sont étudiées sous quelques conditions de régularité. De petites simulations sont également effectuées pour étudier le comportement des esti-

mations robustes en présence de valeurs aberrantes, et ces estimations sont comparées aux estimateurs classiques. La méthode est illustrée par une analyse des données d'une expérience clinique décrite dans Biometrical Journal.

Tuesday June 10 • Mardi 10 juin 2:30 • 14h30 LSC 332

Fatemah Alqallaf, Ruben Zamar, University of British Columbia

Scalable robust covariance and correlation estimates • Estimations robustes et calculables de la covariance et de la corrélation

Covariance and correlation matrix estimates have important applications in data analysis, including but not limited to routine reporting of correlations between variables, principal components analysis and dimensionality reduction and detection of multi-dimensional outliers. Classical covariance and correlation estimates are not reliable estimates for such purposes because they are not robust in that they are adversely influenced by outliers. A small fraction of outliers, in some cases even a single outlier, can distort the value of the classical covariance and correlation estimates so much that they are virtually useless for their intended purposes: correlations for the vast majority of the data can be very erroneously reported, principal components transformations can be quite misleading, and multidimensional outlier detection via Mahalanobis distances based on sample covariance matrix estimates can completely fail to detect outliers. There is a large statistical literature on robust covariance and correlation matrix estimates, with an emphasis on affine equivariant estimators that possess high breakdown points and small worst case biases. Unfortunately, all such estimators have unacceptable exponential complexity in the number of variables. And one of the more attractive of these estimators, the Stahel-Donoho estimator, has an unacceptable quadratic complexity in the number of observations. In this paper we focus on several variants of robust covariance and correlation matrix estimates with quadratic complexity in the number of variables and linear complexity in the number of observations. These estimators are based on several forms of pairwise robust covariance and correlation estimates, and are designed to preserve positive definiteness of the covariance matrix. The estimators studied include two extremely fast estimators based on coordinate-wise robust transformations (the quadrant correlation estimator, and the coordinate-wise Huber estimator), as well as a fully bivariate estimator embedded in an overall procedure recently proposed by Maronna and Zamar (2001). We show that the estimators have attractive robustness properties toward outliers, and give several applications examples. The new estimators are not affine equivariant, but this is often an unnecessary property in large data applications.

Les estimations des matrices de covariance et de corrélation ont des applications importantes en analyse de données, incluant, mais sans s'y limiter, au rapport routinier de la corrélation entre les variables, l'analyse en composantes principales et la réduction de la dimension et la détection des valeurs aberrantes multidimensionnelles. Les estimations classiques de la covariance et de la corrélation ne sont pas des estimations fiables pour de tels buts parce qu'elles ne sont pas robustes dans le sens où elles sont défavorablement influencées par les valeurs aberrantes. Une petite fraction des valeurs aberrantes, même une seule dans certain cas, peut modifier la valeur des estimés classiques de la covariance et de la corrélation, tellement qu'elles sont pratiquement inutiles pour leurs buts prévus; les corrélations pour la grande majorité des données peuvent être très incorrectement rapportées, les transformations des composantes principales peuvent être erronées et la détection de valeurs aberrantes multidimensionnelles par la distance de Mahalanobis

basée sur des estimations de la matrice de covariance échantillonnale peut échouer à détecter des valeurs aberrantes. Il y a beaucoup de littérature statistique sur les estimés robustes des matrices de covariance et de corrélation, avec une emphase sur les estimateurs affines-équivariants qui possèdent des points de rupture élevés et des biais fiables dans les pires des cas. Malheureusement, tous ces estimateurs ont une complexité exponentielle inacceptable par rapport au nombre de variables. Un des estimateurs les plus attrayant, l'estimateur de Stahel-Donoho, a une complexité quadratique inacceptable par rapport au nombre d'observations. Dans cette présentation, nous nous concentrons sur plusieurs variantes des estimations robustes des matrices de covariance et de corrélation avec complexité quadratique par rapport au nombre de variables et complexité linéaire par rapport au nombre d'observations. Ces estimateurs sont basés sur plusieurs formes d'estimations robustes par paires de la covariance et de la corrélation, et sont conçus pour préserver la propriété d'être définie positive. Les estimateurs étudiés incluent deux estimateurs extrêmement rapides basés sur des transformations robustes par rapport aux coordonnées (l'estimateur de corrélation de quadrant, et l'estimateur de Huber par rapport aux coordonnées), ainsi qu'un estimateur complètement bivarié incorporé dans une procédure globale récemment proposé par Maronna et Zamar (2001). Nous montrons que les estimateurs ont des propriétés intéressantes de robustesse envers les valeurs aberrantes, et nous donnons plusieurs exemples d'applications. Les nouveaux estimateurs ne sont pas affines-équivariants, mais c'est souvent une propriété inutile pour des applications sur de grands jeux de données.

Tuesday June 10 • Mardi 10 juin 2:45 • 14h45 LSC 332

Anthony Almudevar, Acadia University

On the exact form for the density of multivariate M-estimators • Sur la forme exacte pour la densité d'estimateurs-M multivariés

Recently, a general form for the density of estimators which solve an estimating equation $\Psi(X,\theta) = 0$ has been obtained by Skovgaard (1990), Jensen and Wood (1998) and Almudevar, Field and Robinson (2000). Under general conditions the density of the estimator at a point θ is equal to the density of $\Psi^*(X,\theta)$ evaluated at zero, where $\Psi^*(X,\theta)$ equals $\Psi(X,\theta)$ multiplied by the inverse of its derivative matrix. In this talk I will discuss conditions under which this holds. Also, I show how accurate approximations of the density of an estimator can be obtained directly from the density of $\Psi(X,\theta)$. Applications to GLMs and nonlinear regression will be discussed.

Récemment, une forme générale pour la densité des estimateurs qui résolvent une équation d'estimation $\Psi(X,\theta) = 0$ a été obtenue par Skovgaard (1990), Jensen et Wood (1998) et Almudevar, Field et Robinson (2000). Sous des conditions générales, la densité de l'estimateur à un point heta est égale à la densité de $\Psi^*(X,\theta)$ évaluée à zéro, où $\Psi^*(X,\theta)$ égale $\Psi(X,\theta)$ multiplié par l'inverse de sa matrice des dérivées. Dans cette présentation je discuterai des conditions dans lesquelles ceci est vérifié. De plus, je montre comment des approximations précises de la densité d'un estimateur peuvent être obtenues directement à partir de la densité de $\Psi(X,\theta)$. Des applications aux GLM et à la régression non-linéaire sont abordées.

Session/Séance 32 • Stochastic Processes and Finance and Applied Probability • Processus stochastique et finance et probabilités appliquées

Tuesday June 10 • Mardi 10 juin 1:30 • 13h30 LSC 234

Eric Marchand, University of New Brunswick; Anatole Joffe, Francois Perron, Université de Montréal; Paul Popadiuk, Concordia University

On a particular sum of dependent Bernouilli and its relationship to a matching type problem • À propos d'une somme particulière de Bernouilli dépendantes et du problème des rencontres

Let $S_n = \sum_{k=1}^{n} X_k X_{k+1}$; and $S = \lim_{n\to\infty} S_n$ where $\{X_k\}_{k=1}^{\infty}$ are independent Bernoulli random variables with mean p_k. We study the distributions of S and S_n by establishing a recurrence for the probability generating functions of S_n. For the cases when $p_k = k^{-1}+B$ with $B \geq 0$, we show that the distribution of S is a Beta mixture of Poisson distributions. In particular, when $p_k = k^{-1}$, S follows a Poisson distribution with mean 1. We also give an interesting connection with a matching type problem, giving an independent derivation of the above results when $p_k = k^{-1}+B$ with B a nonnegative integer. The talk involves elementary probability and should be accessible to all.

Soit $S_n = \sum_{k=1}^{n} X_k X_{k+1}$; *et* $S = \lim_{n\to\infty} S_n$ *où* $\{X_k\}_{k=1}^{\infty}$ *sont des variables aléatoires indépendantes de Bernoulli avec moyenne* p_k. *Nous étudions les distributions de* S *et* S_n *en établissant une récurrence pour la fonction génératrice de probabilité de* S_n. *Pour les cas où* $p_k = k^{-1} + B$ *avec* $B \geq 0$, *nous montrons que la distribution de* S *est un mélange bêta de loi de Poisson. En particulier, lorsque* $p_k = k^{-1}$, S *suit une loi de Poisson de moyenne 1. Nous donnons également une relation intéressante avec un problème " matching type ", donnant une dérivation indépendante des résultats ci-dessus lorsque* $p_k = k^{-1} + B$ *avec* B *un nombre entier non négatif. La présentation implique des notions de probabilités élémentaires et devrait être accessible à tous.*

Tuesday June 10 • Mardi 10 juin 1:45 • 13h45 LSC 234

Mahmoud Zarepour, Mohammad Taghi Jahandideh, University of Ottawa

Option pricing formula with infinite variance innovations • La formule de pricing des options avec des innovations de variance infinie

The randomized discounted option pricing formula when the daily return is in the domain of attraction of a nongaussian stable law presents some challenging problems in practice. We will study some of these problems and to some extent we will present some methods that can partially answer some related questions.

La formule randomisée de pricing des options escompté quand le retour quotidien est dans le domaine d'attraction d'une loi stable non gaussienne présente quelques problèmes important en pratique. Nous étudions certains de ces problèmes et dans une certaine mesure, nous présentons des méthodes qui peuvent répondre partiellement à quelques questions reliées.

Tuesday June 10 • Mardi 10 juin 2:00 • 14h00 LSC 234

René Ferland, Université du Québec à Montréal; Simon Lalancette, HEC-Montréal

Forecasting of realized volatility and correlations: an empirical study • Prédiction de volatilités et corrélations réalisées: une étude empirique

This study empirically examines the competitiveness of different forecasting sets of realized volatilities and correlations using linear and nonlinear specifications of time series based on high frequency data. The empirical performance of those specifications is compared

to a GARCH diagonal-BEKK model in the context of a trader who would simultaneously quote a call spread option price based on the forecasted parameters and delta-hedge her position with a replicating portfolio. More traditional loss functions based on the absolute forecasting error are also used. The forecasting methodologies based on time series of realized volatilities and correlations generally (but not unanimously) dominate the GARCH approach. General performance ranking for the approaches based on realized volatilities and correlations is not robust to the chosen loss function.

Nous évaluons la capacité prévisionnelle de spécifications linéaire et non linéaire appliquées à des séries de volatilités et corrélations dites "réalisées" obtenues à partir de données à très hautes fréquences. Aux fins de comparaisons, nous estimons aussi un modèles de type GARCH BEKK-diagonal. L'évaluation des prévisions pour chaque modèle s'effectue en supposant qu'une arbitragiste affiche des prix d'options de type écarts sur taux d'intérêt, à partir des prévisions et, simultanément, gère un portefeuille de réplication périodiquement révisé par "delta-hedging". Une approche plus traditionnelle, qui dépend directement des erreurs de prévisions, est également utilisée. En général, les prévisions qui émanent d'un traitement des volatilités et corrélations réalisées supplantent celles dégagées par le modèle GARCH. Cependant, la performance relative des différentes approches apparaît sensible au choix du critère d'évaluation.

Tuesday June 10 • Mardi 10 juin 2:15 • 14h15 LSC 234

Nickolai Khodusov, Novosibirsk State Technical University, Novosibirsk, Russia

New Modifications of NFV Criterion • Nouvelles modifications du critère NFV

In the paper the main classical criteria of investment effectiveness evaluation are presented and analyzed. These criteria are NPV, IRR, PP, NFV and other. The new criteria based on modification of NFV and IRR criteria are proposed. The comparison of classical and proposed criteria is made. When calculating NPV discount factor is used to discount ingoing and outgoing cash flows to the beginning of the project. In real life discount factor is not well known function of time and depends on the dynamics of interest rate and index of inflation. The investigation of possible contradictions between NPV and IRR was made. It was appeared that classical NFV gives the same as NPV recommendations and cannot be called as its alternative. In the paper the modifications of NFV and IRR proposed (MNFV and MIRR respectively). These modifications allow taking into account the possibility to reinvest funds earned in the project and differences in values of crediting and depositing rates. Various models of MNFV were examined. As a way of subsequent investigations the modification with optimization of MNFV or MIRR is proposed.

Dans cette présentation, les principaux critères d'évaluation de l'efficacité de l'investissement classiques sont présentés et analysés. Ces critères sont le NPV, IRR, PP, NFV et certains autres. Nous proposons des nouveaux critères basés sur la modification des critères NFV et IRR. Nous comparons les critères classiques à ceux proposés lorsque le calcul du facteur escompté de NPV est utilisé pour escompter des flux monétaire entrant et sortant au commencement du projet. Dans des situations réelles, le facteur escompté n'est pas une fonction connue du temps et dépend de la dynamique du taux d'intérêt et de l'index d'inflation. La recherche de contradictions possibles entre le NPV et le IRR a été faite. Il apparaît que le NFV classique donne les mêmes résultats que les recommandations de NPV et ne peut pas s'exprimer comme son alternative. Dans la présentation, nous proposons des modifications du NFV et du IRR (MNFV et MIRR respectivement). Ces modifications tiennent compte de la possibilité de réinvestir des fonds gagnés dans le

projet et des différences en valeurs des taux de crédits et de débits. Divers modèles de MNFV sont examinés. Pour des investigations futures, nous proposons la modification avec optimisation du MNFV ou du MIRR.

Tuesday June 10 • Mardi 10 juin 2:30 • 14h30 LSC 234

Yung-Ming Chang, National University of Kaohsiung, Taiwan; James C. Fu, University of Manitoba

On later waiting time distributions in a sequence of Markov dependent multistate trials • Sur les distributions des derniers temps d'attente dans une séquence d'essai multi-états dépendants de Markov

The sooner and later waiting time problems have been extensively studied and successfully applied in various areas of statistics and applied probability. In this talk, we introduce three different ways of obtaining the probability generating functions for later waiting time distributions of l ($l \geq 2$) simple patterns with respect to nonoverlapping and overlapping counting schemes in a sequence of Markov dependent multistate trials. Our work is based on the finite Markov chain imbedding technique introduced by Fu and Koutras (1994). Examples are given for illustrating our results.

Les problèmes des premiers et derniers temps d'attente ont été étudiés intensivement et appliqués avec succès dans divers secteurs des statistiques et des probabilités appliquées. Dans cette présentation, nous introduisons trois manières différentes d'obtenir la fonction génératrice de probabilité pour les distributions des derniers temps d'attente de l (l geq 2) patrons simple par rapport à des schème de comptage qui se chevauchent ou non dans séquence d'essais multi-états dépendants de Markov. Notre travail est basé sur la technique qui inclut la chaîne de Markov finie présentée par Fu et Koutras (1994). Des exemples sont donnés pour illustrer nos résultats.

Tuesday June 10 • Mardi 10 juin 2:45 • 14h45 LSC 234

Rachel MacKay, University of British Columbia

Hidden Markov models for multiple processes • Les chaînes de Markov cachées pour des processus multiples

Hidden Markov models (HMMs) are a useful tool for capturing the behaviour of overdispersed, autocorrelated data. These models have been applied to many different problems, including speech recognition, precipitation modelling, and gene finding and profiling. Typically, HMMs are applied to individual stochastic processes; HMMs for simultaneously modelling multiple processes have not been widely studied. In this context, random effects may be a natural choice for capturing differences among processes. In this talk, we discuss the theory required for implementing and interpreting these models in a general setting, using the framework of generalized linear mixed models. We address the topics of parameter estimation and hypothesis testing, including testing of the variance components. We then apply these models to data on lesion counts in multiple sclerosis patients.

Les chaînes de Markov cachées (CMCs) sont un outil utile pour capturer le comportement des données surdispersées et autocorrélées. On a appliqué ces modèles à plusieurs problèmes différents, y compris la reconnaissance de la parole, la modélisation de la précipitation, et la découverte des gènes et de leurs profiles. Typiquement, on applique les CMCs aux processus stochastiques individuels; peu d' étude a été réalisée sur les CMCs pour la modélisation simultanée des processus multiples. Dans ce contexte, les effets aléatoires pourraient être un choix naturel pour capturer des différences entre les processus. Dans cet exposé, nous

discutons la théorie requise pour implémenter et interpréter ces modèles dans un cadre général, en se basant sur la procédure des modèles linéaires généralisés à effets mixtes. Nous abordons les sujets de l'estimation des paramètres ainsi que les tests d'hypothèse, y compris les tests des composantes de la variance. Puis, nous appliquons ces modèles aux données des comptes de lésions sur des patients de sclérose en plaques.

Session/Séance 33 • Comparative Research • Études comparatives

Tuesday June 10 • Mardi 10 juin 3:30 • 15h30 LSC 240

Scott Murray, Statistics Canada/Statistique Canada

Quality control in an international comparative context: experience from the International Adult Literacy survey • Le contrôle de la qualité dans un contexte de comparaisons internationales: l'expérience de l'étude internationale sur l'analphabétisation

Statistics Canada, in collaboration with the U.S. National Center for Education Statistics (NCES) the OECD, UNESCO and the Educational Testing Service, has led an international effort to collect comparative data on adult literacy skills. Combining the methods of educational measurement with household survey methods the program has administered a test to a representative samples of adults in some 35 countries. Known as either the International Adult Literacy Survey (IALS) or the Adult Literacy and Life Skills Survey (ALL), the study has had a significant impact on educational policies in many countries. Elaborate quality assurance procedures have been developed to ensure results are valid, reliable and comparable. This paper describes the extent and nature of these procedures and provides a critical review of their success in containing bias.

Statistiques Canada, en collaboration avec le Centre national des américain pour les statistiques en éducation (NCES), l'OCDE, l'UNESCO et le Service de tests éducationnels, a maintenu un effort international pour rassembler des données comparatives sur les niveau d'éducation des adultes. En combinant les méthodes de mesures éducatives avec les méthodes de sondage sur les foyers, le programme a administré un test à un échantillon représentatif d'adultes dans environ 35 pays. Connu comme l'Enquête internationale sur l'éducation des adultes (IALS) ou Enquête sur l'éducation des adultes et des qualifications de la vie (ALL). L'étude a eu un impact significatif sur les politiques en éducation dans beaucoup de pays. Des procédures élaborées de qualité ont été développées pour assurer que les résultats sont valides, fiables et comparables. Cette présentation décrit l'ampleur et la nature de ces procédures et fournit une revue critique de leur succès pour gérer le biais.

Tuesday June 10 • Mardi 10 juin 4:00 • 16h00 LSC 240

Albert Motivans, UNESCO Institute for Statistics/Institut de statistiques de l'UNESCO

Investing in education: towards a cross-national perspective • Investir dans l'éducation: vers une perspective multinationale

Cross-national comparisons of financial indicators can help policy makers assess whether they are adequately funding education and using resources in the most effective, efficient and equitable manner. Comparing different processes and mechanisms used to finance education systems also shows how national policy makers respond to different contexts in order to achieve national goals and aspirations. Despite their wide use in OECD countries, international financial statistics elswhere are often criticized for shortcomings in comparability.

To explore these issues, the World Education Indicators Programme (WEI) coordinated jointly by UNESCO-UIS and the OECD, conducted expert site visits in 11 middle-income countries in 2001/2002. The chief goal was to map public and private flows of financial resources to educational institutions. In addition, the study documented data underlying finance indicators and identified definitional problems, data gaps and areas that require further development. This presentation compares key education finance indicators across this set of countries and discusses some of the constraints associated with their interpretation.

Les comparaisons multinationales des indicateurs financiers peuvent aider les décisionnaires à évaluer s'ils financent adéquatement l'éducation et utilisent les ressources de la manière la plus efficace et équitable. Comparer différents processus et mécanismes utilisés pour financer les systèmes d'éducation montre également comment les décisionnaires nationaux répondent aux différents contextes afin de réaliser les buts et les aspirations nationaux. En dépit de leur utilisation large dans des pays de l'OCDE, les statistiques financières internationales sont souvent critiquées pour certaines imperfections dans la comparabilité. Pour explorer ces issues, le Programme mondial d'indicateurs en éducation (WEI) coordonné conjointement par l'UNESCO-UIS et l'OCDE, a conduit des visites d'experts dans 11 pays à revenu moyen en 2001-2002. Le but principal était de tracer les flux de ressources financières publics et privés aux établissements éducatifs. De plus, les l'étude à documenté des données sous jacentes aux indicateurs financiers et a identifié certains problèmes de définition, des lacunes dans les données et certains secteurs qui exigent des développements ultérieurs. Cette présentation compare les principaux indicateurs financiers en éducation à travers cet ensemble de pays et discute de certaines des contraintes associées à l'interprétation de ces indicateurs.

Tuesday June 10 • Mardi 10 juin 4:30 • 16h30 LSC 240

Tim Holt, University of Southampton

Methodological issues in the development of statistical indicators and their use in international comparisons • Considérations méthodologiques dans le développement d'indices statistiques et leurs utilisations pour des comparaisons internationales

There is a growing emphasis on the international use of statistical indicators to develop and monitor global policies and to support the international comparability of statistics derived from nation states. This is not completely new. Economic statistics and the System of National Accounts have been used for 50 years to monitor economic development and to make national comparisons. Organisations such as the IMF and the OECD depend heavily on the availability of internationally comparable statistics from countries. However this type of use has accelerated in recent years and increasingly statistics and statistical indicators are being used to set and monitor global policy. For example a review of UN Summits and major Conferences during the 1990's identified over 280 statistical indicators needed to monitor UN policies made through conference decisions. The Millennium Development Goals, for example, subscribed to by 164 Heads of State or their representatives have resulted in statistical indicators and targets that will be monitored over the coming decades. Hence the need for internationally comparable statistics has never been greater. This paper has two purposes: (i) to describe the current need for internationally comparable statistical indicators for UN and related agency purposes, and (ii) to suggest that despite the huge investment in methodological research and development to support national statistical needs, there has not been quite as much emphasis on methodological

issues supporting the need for international comparability. Examples will illustrate some of the methodological issues that arise.

Il y a présentement une croissance de l'utilisation internationale des indicateurs statistiques pour développer et surveiller les politiques globales et du soutien de la comparabilité internationale des statistiques provenant des États. Cette réalité n'est pas complètement nouvelle, des statistiques économiques et le Système des Comptes Nationaux ont été utilisés pendant 50 ans pour surveiller le développement économique et pour faire des comparaisons nationales. Les organismes tels que le FMI et l'OCDE dépendent fortement de la disponibilité de statistiques internationalement comparables provenant des pays. Cependant, ce type d'utilisation est en croissance depuis quelques années et, de plus en plus, les statistiques et les indicateurs statistiques sont employés pour mettre en place des politiques globales et analyser leur impact. Par exemple, lors des sommets et des principales conférences de l'ONU des années 90, 280 indicateurs statistiques ont été identifiés pour aider à surveiller les politiques de l'organisme. Les Buts de Développement du Millénium, par exemple, souscrits par près de 164 chefs d'État ou leur représentant, ont eu comme conséquences l'établissement d'indicateurs et de buts à atteindre qui seront surveillés durant les prochaines décennies. Par conséquent, le besoin de statistiques comparables internationalement n'a jamais été plus grand. Cette présentation a deux buts: premièrement de décrire le besoin courant d'indicateurs statistiques comparables internationalement pour les besoins de l'ONU et des agences reliés et deuxièmement, pour suggérer qu'en dépit des investissements énorme en recherche méthodologique et pour le développement et le soutient des besoins statistiques nationaux, il n'y a pas eu autant d'emphase sur les questions méthodologiques soutenant le besoin de comparabilité internationale. Nous allons présenter des exemples pour illustrer certains problèmes méthodologiques que l'on peut encourir.

Session/Séance 34 • Process Monitoring • Suivi de processus

Tuesday June 10 • Mardi 10 juin 3:30 • 15h30 LSC 242

Fred Spiring, Pollard Banknote Ltd & University of Manitoba

Assessing process capability: a user's view • L'évaluation de la capacité d'un processus: la vision d'un utilisateur-chercheur

Research efforts in the area of process capability have largely been devoted to finding a better process capability index (PCI) and to a lesser extent on the stochastic behaviour/properties of the estimated PCIs. Much of this development has gone unused for many reasons including a) interpretation, b) software support, c) standards and d) dissemination. The addition of the Cp alphabet appears to have had little impact on practical use. Cp and Cpk (including Cpl and Cpu) continue to be the most heavily used indices with Cpm and Cpmk occurring occasionally. The addition of stochastic assessments for estimated PCIs is a positive development, however statistical developments have frequently lacked background knowledge, hindering implementation by practitioners. Using a case study involving the printing of a lottery ticket, we will attempt to draw attention to areas impacting the practical use of PCIs. Concepts including a) establishing effective tolerance limits, b) the ongoing assessment and interpretation of PCIs and c) ongoing improvement will be presented. We will use the case study a) to illustrate the difficulties encountered by practitioners in using PCIs, b) to draw attention to gaps that exist in the practical use of PCIs, c) to illustrate the approach we used to overcome some of these difficulties and d) to highlight research areas in the practical use of PCIs.

Les efforts de recherches dans le secteur de la capacité des processus ont été en grande partie consacrés à trouver un meilleur index de capacité de processus (PCI) et à un degré inférieur, sur le comportement et les propriétés stochastiques du PCI estimé. Beaucoup de ces développements furent inutilisés pour plusieurs raisons dont a) l'interprétation, b) le support technique pour les logiciels, c) les normes et d) la diffusion. L'addition de l'alphabet de Cp semble avoir eu peu d'impact sur son utilisation pratique. Le Cp et le Cpk (y compris le Cpl et le Cpu) continuent d'être les index les plus fortement utilisés avec le Cpm et le Cpmk. L'addition des évaluations stochastiques pour les PCI estimé est un développement positif, toutefois les développements statistiques ont fréquemment manqué de connaissances de base, gênant son utilisation par des praticiens. En utilisant une étude de cas impliquant l'impression d'un billet de loterie, nous essayerons de mettre l'accent sur des secteurs suggérant l'utilisation des PCI . Des concepts comprenant a) l'établissement de limites de tolérance efficaces, b) l'évaluation et l'interprétation continue de PCI et c) l'amélioration continue seront présentés. Nous utiliserons l'étude de cas pour a) l'illustration des difficultés rencontrées par des praticiens en employant les PCI, b) pour attirer l'attention sur les lacunes qui existent dans l'utilisation pratique des PCI, c) pour illustrer l'approche que nous avons pris pour surmonter certaines de ces difficultés et d) pour souligner des domaines de recherches dans l'utilisation pratique des PCI.

Tuesday June 10 • Mardi 10 juin 4:00 • 16h00 LSC 242

Bovas Abraham, Asokan Mulayath Variyath, University of Waterloo

A look at the Mahalanobis-Taguchi system • Un regard sur le système de Mahalanobis-Taguchi

In a recent paper Wooddall(2003) (to appear in Technometrics) reviewed and commented on the so called Mahalanobis Taguchi System which was proposed and named by G. Taguchi and his co-authors. They use a slightly modified version of Mahalanobis D^2 as a diagnostic measure as well as a tool for classification. In this paper we review this procedure and compare it with the multivariate T^2 statistic used in process monitoring and with other multivariate procedures. We find that the traditional multivaraiate procedures are more attractive from a statistical point of view.

Dans un papier récent (à paraître dans Technometrics), Wooddall (2003) a passé en revue et a commenté le prétendu système de Mahalanobis-Taguchi, proposé et nommé par G. Taguchi et ses co-auteurs. Ils emploient une version légèrement modifiée de la statistique D^2 de Mahalanobis comme mesure de diagnostique ainsi que comme un outil de classification. Dans cet présentation, nous passons en revue cette technique et nous la comparons à la statistique T^2 multivariée utilisée en contrôle des processus et avec d'autres procédures multivariées. Nous constatons que les procédures multivariées traditionnelles sont plus attrayantes d'un point de vue statistique.

Tuesday June 10 • Mardi 10 juin 4:30 • 16h30 LSC 242

J. B. François Boudreau, Mike Dudzic, Dofasco Inc.

On-line multivariate statistical monitoring at Dofasco Inc. • La surveillance statistique multivariable en ligne à Dofasco Inc.

In the steel industry, the need to ensure product quality and process efficiency has been a driving force for instrumentation and automation. With the implementation of extensive data infrastructures required to support automation systems, overwhelming amounts of

data are now available in real time and for historical analysis. Over the last decade, a number of applications based on multivariate analytical methods such as Principal Components Analysis (PCA) and Partial Least Squares (PLS) have been successfully implemented at Dofasco to meet the challenge of extracting useful information from process data. Classes of application include off-line data analysis, on-line prediction, and on-line process monitoring. Practical aspects of the implementation of real-time systems will be discussed, focusing on the on-line monitoring of Dofasco's #1 continuous casting machine.

Le besoin d'assurer la qualité des produits et l'efficacité des procédés dans l'industrie sidérurgique a été une force motrice pour l'instrumentation et l'automatisation. Comme d'imposantes infrastructures de données sont nécessaires au fonctionnement des systèmes automatisés, des quantités considérables de données sont maintenant disponibles pour l'analyse en temps réel ou a posteriori. Dans cette dernière décennie, un nombre d'applications basées sur les méthodes multivariables telles que l'analyse en composantes principales et la régression partielle des moindres carrés ont été implantées à Dofasco afin de relever le défi que représente l'extraction d'information utile des donnés d'opération. Les classes d'applications incluent l'analyse hors ligne, la prédiction en ligne et la surveillance en ligne des procédés. On discutera de certains aspects pratiques de l'implantation de systèmes en temps réel en prenant pour exemple la surveillance en ligne de la machine de coulée continue No 1 à Dofasco.

Session/Séance 35 • Special Session of the Fields Institute on Matrices and Statistics • Session spéciale de l'Institut Fields sur les matrices en statistique

Tuesday June 10 • Mardi 10 juin 3:30 • 15h30 LSC 238

Jerzy K. Baksalary, University of Zielona Gora, Poland

Algebraic properties and statistical applications of rank-one-modified matrices • Les propriétés algébriques et les applications statistiques des matrices modifiées de rang un

For given real matrix A and nonzero real vectors b, c, relationships between generalized inverses of A and generalized inverses of its rank-one-modification $M = A + bc'$ (with c' being the transpose of c) are investigated. The formulae of C.D. Meyer, Jr. [SIAM J. Appl. Math. 24 (1973) 315-323] expressing the Moore-Penrose inverse M^+ of M as modifications of A^+ are revisited from the view-point of a result concerning comparison of the ranks of M and A. Three further problems considered are: (i) when does a given generalized inverse A^- belongs to the set $\mathcal{M}$ of all generalized inverses of M, (ii) when does $A^- \in \mathcal{A}$ exist such that simultaneously $A^- \in \mathcal{M}$, and (iii) when does the set $\mathcal{A} \subset \mathcal{M}$. Some applications of the results obtained to problems in mathematical statistics and theory of experimental designs are pointed out.

Pour une matrice réelle A donnée et des vecteurs réelles non nuls b, c, nous investiguons les relations entre l'inverse généralisé de A et l'inverse généralisé de sa modification de premier rang $M = A + bc'$ (avec c' qui se trouve à être la transposé de c). Nous révisons la formule de C.D. Meyer, Jr. [SIAM J. Appl. Math. 24 (1973) 315-323] qui exprime l'inverse de Moore-Penrose M^+ of M comme une modification de A^+ d'un point de vue d'un résultat concernant la comparaison du rang de M et A. Trois autres problèmes aussi considérés sont: i) lorsqu'un certain inverse généralisé A^- appartient à l'ensemble M de tous les inverses généralisés de M, ii) lorsque $A^- \in A$ existe de sorte que simultanément $A^- \in M$ et iii) lorsque l'ensemble $A \subset M$. Nous indiquons quelques applications des

résultats obtenus aux problèmes en statistiques mathématiques et en théorie des designs expérimentaux.

Tuesday June 10 • Mardi 10 juin 4:00 • 16h00 LSC 238

Simo PUNTANEN, University of Tampere, Finland; George STYAN, McGill University

Matrix tricks for teaching linear statistical models – our personal Top Ten • Astuces matricielles pour enseigner les modèles statistiques linéaires: notre " top dix " personnel

In teaching a course in linear regression models to second or third year undergraduate students, say, there is no way to proceed smoothly without matrices and related concepts of linear algebra. Their use is really not questioned. One interesting signal of the increasing importance of matrix methods for statistics is that recently quite a few new books have been published in this area.

Our experience is that making some particular matrix tricks familiar to students can increase their insight into linear statistical models (and also multivariate analysis). In matrix algebra, there are handy, sometimes even very simple, "tricks" to simplify and clarify the problem treatment – both for a student and a researcher. In this paper we collect together our Top Ten favourite matrix tricks.

Losqu'on enseigne un cours sur les modèles de régression linéaire à des étudiants de deuxième ou troisième année d'études de baccalauréat par exemple, il est impossible de procéder convenablement sans une bonne base théorique sur les matrices et les concepts d'algèbre linéaire. L'utilisation de l'algèbre linéaire n'est pas du tout remise en cause dans cette présentation. Un indicateur intéressant qui montre l'importance croissante des méthodes matricielles en statistiques est l'accroissement des livres récemment édités dans ce secteur.

Notre expérience montre que de rendre familier aux étudiants quelques astuces particulières sur les matrices peut augmenter leur perspicacité avec les modèles statistiques linéaires (également avec l'analyse multivariée). Avec l'algèbre matricielle, il y a parfois des astuces très simple et utiles pour simplifier et clarifier le traitement d'un problème, pouvant être utile à un étudiant comme à un chercheur. Dans cette présentation, nous rassemblons notre " top dix " d'astuces sur les matrices.

Tuesday June 10 • Mardi 10 juin 4:30 • 16h30 LSC 238

Hans Joachim WERNER, University of Bonn, Germany

In the year of the matrix - prediction techniques in the general Gauss- Markov model • Dans l'année de la matrice — Techniques de prédiction dans le cadre du modèle de Gauss-Markov

In the framework of the general (possibly singular) Gauss-Markov model, we will consider two powerful prediction techniques. We will investigate their basic properties, explain the motivations behind, and discuss the connections between them. Most observations may be obtained by employing rather elementary, yet powerful, matrix algebra. The relations to the estimation techniques BLUE and BLIMBE, which are discussed in detail in Schönfeld and Werner (1986), will also be mentioned. Reference: Schönfeld, P. and Werner, H. J. (1986), Beiträge zur Theorie und Anwendung linearer Modelle. In Ökonomische Prognose-, Entscheidungs- und Gleichgewichtsmodelle (W. Krelle, ed.), 251–262. VCH Verlagsgesellschaft, Weinheim.

Dans le cadre du modèle général de Gauss-Markov, nous considérerons deux techniques puissantes de prédiction. Nous étudierons leurs propriétés de base, en expliquerons les motivations et discuterons leurs interconnexions. La plupart des observations peuvent être obtenues en utilisant des résultats plutôt élémentaires, quoique puissants, de l'algèbre des matrices. Les relations avec les deux méthodes d'estimation BLUE et BLIMBE, méthodes discutées en détail par Schönfeld et Werner (1986), seront également mentionnés. Référence: Schönfeld, P. et Werner, H. J. (1986), Beiträge zur Theorie und Anwendung linearer Modelle. Dans: Ökonomische Prognose-, Entscheidungs- und Gleichgewichtsmodelle (W. Krelle, éd.), 251–262. VCH Verlagsgesellschaft, Weinheim.

Session/Séance 36 • Statistics in Genomics and Proteomics • Statistique en génomique et protéomique

Tuesday June 10 • Mardi 10 juin 3:30 • 15h30 LSC 338

Jenny Bryan, University of British Columbia

Resolving gene expression profiles with tag-based technologies • Résoudre les profiles d'expression des gènes avec des technologies Tag

We will present methods for estimating gene expression profiles from a universal array, tag-based profiling technique that is under development. This technique has features in common with Serial Analysis of Gene Expression (SAGE) and with microarrays, although it is distinct from both. Issues that have implications for estimation and inference include tag degeneracy (when multiple transcripts cannot be distinguished from one another) and various substantial sources of noise and bias in sample preparation (for example, PCR bias) that affect both new and current expression profiling techniques. This work is part of a collaboration with Dr. Charles Haynes, in the Biotechnology Laboratory and Chemical & Biological Engineering Dept. at UBC.

Nous présentons des méthodes pour estimer les profils d'expression des gènes par une technique en cours de développement, profilé sur tableau universel et basée sur des étiquettes. Cette technique a des aspects en commun avec l'analyse en série de l'expression des gènes (SAUGE) et avec les microréseaux, bien qu'elle soit distincte de chacune d'elles. Les problèmes qui impliquent l'estimation et l'inférence incluent la dégénérescence de l'étiquette (quand les transcriptions multiples ne peuvent pas être distinguées les unes des autres) et diverses sources substantielles de bruit et de biais dans la préparation de l'échantillon (par exemple, le biais PCR) qui affectent la l'expression des techniques profilées nouvelles et courantes. Cette présentation fait partie d'une collaboration avec le Dr. Charles Haynes, dans le laboratoire de biotechnologie et de service chimique et de génie biologique à UBC.

Tuesday June 10 • Mardi 10 juin 4:00 • 16h00 LSC 338

Peter Hooper, University of Alberta

Statistical pattern recognition methods for protein secondary structure • Modèles statistiques de reconnaissance de forme pour la structure secondaire des protéines

A protein is a polymer made of amino acids. There are several levels of protein structure. The primary structure is the linear sequence of amino acids. The secondary structure represents recurring structural patterns: alpha-helices and beta strands (which assemble into beta-sheets), interspersed with coil regions. The tertiary structure (or fold) is the overall packing of secondary structure elements into a compact 3D architecture. The prediction of

secondary structure from primary structure is an important intermediate step in predicting tertiary structure. This talk will review some of the literature on secondary structure prediction and describe new approaches based on "reference point logistic" classification models and hidden Markov models.

Une protéine est un polymère fait d'acides aminés. Il y a plusieurs niveaux de structure de protéine. La structure primaire est l'ordre linéaire des acides aminés. La structure secondaire représente les schèmes de récurrence structurale: alpha-spirales et brins bêta (qui s'assemble dans des feuillet bêta), entremêlées avec des régions spirales. La structure tertiaire (ou le pli) est l'emballage global des éléments de la structure secondaires dans une architecture 3D compacte. La prédiction de la structure secondaire à partir de la structure primaire est une étape intermédiaire importante pour prédire la structure tertiaire. Cette présentation passe en revue une partie de la littérature sur la prédiction secondaire de la structure et décrit de nouvelles approches basées sur les modèles de classification de " point de référence logistiques " et les modèles cachés de Markov.

Tuesday June 10 • Mardi 10 juin 4:30 • 16h30 LSC 338

Robert Gentleman, Harvard School of Public Health

Graphs and EDA in computational biology • Les graphes et l'analyse exploratoire (EDA) de données en biologie computationnelle

Graphs provide a unique data structure for exploring biological data. There are many different graphs that can be constructed based on biologic data. These include metabolic pathways, protein-protein interactions as well as co-citation of genes in the scientific literature. In this talk I will consider various methods of using graphs and their properties to perform exploratory data analysis (EDA) on data from a microarray experiment using different graphs based on biological meta-data.

Les graphes fournissent une structure de données unique pour explorer des données biologiques. Il y a beaucoup de différents graphes qui peuvent être construits basé sur des données biologiques. Ceux-ci incluent les voies étaboliques, les interactions de types protéine-protéine et aussi la co-citation des gènes dans la littérature scientifique. Dans cette présentation je considérerai diverses méthodes pour utiliser les graphes et leurs propriétés pour effectuer l'analyse exploratoire de données (EDA) sur les données d'une expérience microréseau en utilisant différents graphes basés sur des méta-données biologiques.

Session/Séance 37 • Statistical Computing and Computationally Intensive Methods • Statistique informatique et méthodes informatiques intensives

Tuesday June 10 • Mardi 10 juin 3:30 • 15h30 LSC 234

Peter Macdonald, Juan Du, McMaster University

An R package for finite mixture distributions • Un package en R pour des mélanges finis de distributions

We have developed a comprehensive R package for fitting finite mixture distributions to grouped data. The component distributions may be normal, lognormal, gamma, Weibull, binomial, negative binomial or Poisson. The grouped data may have missing or censored bins and may include conditional as well as mixed data. Maximum likelihood estimates

are calculated by a combination of EM and quasi-Newton algorithms. Ill-conditioned problems can be resolved by constraints on the parameters. Graphical methods include the rootogram as a diagnostic tool. This talk will illustrate the logic and functionality of Rmix with a number of examples from different areas of application.

Nous avons développé un package R complet pour ajuster des mélanges finis de distributions à des données groupées. Les distributions des composantes peuvent être normales, log-normales, gamma, Weibull, binomiales, binomiales négatives ou Poisson. Les données groupées peuvent avoir des sections absentes ou censurées et peuvent inclure des données conditionnelles ou mixtes. Les estimés du maximum de vraisemblance sont calculées par une combinaison d'algorithmes EM et quasi-Newton. Les problèmes mal définis peuvent être résolus par des contraintes sur les paramètres. Les méthodes graphiques incluent le rootogram comme outil de diagnostique. Cette présentation illustre la logique et la fonctionnalité de Rmix avec un certain nombre d'exemples dans différents domaines d'application.

Tuesday June 10 • Mardi 10 juin 3:45 • 15h45 LSC 234

Alain Desgagné, Jean-François Angers, Université de Montréal

Computational aspect of the generalized exponential power density • Considérations quantitatives de la famille de puissances d'exponentielle généralisée

In this paper, the generalized exponential power (GEP) distribution is studied. This family encompasses a vast majority of the usual distributions and some others using a simple transformation. The rich variety of its tails behavior makes the GEP density a natural benchmark to characterize and order tails of a large class of densities. Analytic formulas and numerical methods are proposed to evaluate its normalizing constant, its moments and its cdf. Furthermore, two methods to simulate observations from the GEP distribution are proposed. Some examples of simulations are presented. A numerical method using the GEP density is also given for the estimation of posterior moments when the prior and the likelihood are symmetric densities defined on the real line. Some examples of the behavior of the posterior density when an observation is an outlier are presented. It can be seen that the use of heavy-tailed distributions is a valuable tool in developing robust Bayesian procedures, limiting the influence that extreme information sources can have on posterior inferences.

Nous voulons étudier dans cet article la famille de puissances d'exponentielle généralisée (GEP). Cette famille comprend la grande majorité des densités usuelles et quelques autres à une transformation près. La riche diversité du comportement des queues de la densité GEP en fait une densité de référence naturelle pour caractériser et ordonner les queues d'une grande classe de densités. Des méthodes analytiques et numériques sont proposées afin d'évaluer sa constante de normalisation, ses moments et sa fonction de répartition. De plus, deux méthodes sont proposées pour simuler des observations provenant de la densité GEP. Quelques exemples de simulations sont présentées. Une méthode numérique basée sur la densité GEP est aussi présentée pour l'estimation des moments a posteriori lorsque la loi a priori et la vraisemblance sont des densités symétriques définies sur les réels. Des exemples du comportement de la densité a posteriori lorsqu'on est en présence d'une valeur aberrante est présentée. On peut voir que l'utilisation de distributions à queues épaisses est un outil précieux pour développer des procédures bayésiennes robustes, limitant l'influence qu'une source d'information extrême peut avoir sur l'inférence a posteriori.

Tuesday June 10 • Mardi 10 juin 4:00 • 16h00 LSC 234

Daniel Lemire, National Research Council of Canada

Constant time polynomial range sums for dynamic OLAP • Sommations polynomiales calculées en temps constant pour applications OLAP dynamiques

Businesses and governments increasing rely on On-Line Analytical Processing (OLAP) engines for their data mining needs. OLAP aims to accelerate common queries so that experts can very quickly analyze very large data sets. Many common OLAP queries such as count, mean, variance, and covariance can be expressed as "polynomial sums". We present a novel high performance approach offering constant time $O(1)$ exact queries for all polynomial range sums and small storage requirements of $n^{d/\eta}$ components where n^d is the size of the data cube and η is an arbitrarily large integer parameter. The approach used is inspired by wavelets and subdivision schemes.

Les entreprises et les gouvernements utilisent de plus en plus les systèmes OLAP (On-Line Analytical Processing) pour l'exploration des données. Les systèmes OLAP visent à accélérer les requêtes typiques de telle manière à ce que des experts puissent rapidement analyser de très grands ensembles de données. Plusieurs requêtes OLAP telles que les sommes, les moyennes, les variances, et les covariances peuvent être écrites en termes de "sommes polynomiales". Nous présentons une nouvelle approche très performante permettant de calculer de façon exacte et en temps constant $O(1)$ et utilisant au plus un tampon de $n^{d/\eta}$ composantes où n^d est la taille du cube de données et η est un entier arbitrairement grand. L'approche utilisée est inspirée des ondelettes et des schémas de subdivision.

Tuesday June 10 • Mardi 10 juin 4:15 • 16h15 LSC 234

Genghui Wu, University of Manitoba

A random-discretization based Monte Carlo sampling method for numerical integration • Une méthode d'échantillonnage de Monte-Carlo basée sur une discrétisation aléatoire pour l'intégration numérique

Importance sampling is a popular tool for numerical integration. The accuracy of its result, however, depends on the chosen importance density; an inappropriate choice can cause devastating loss in accuracy. Moreover, in practice there is no guaranteed way to choose a good importance density. This problem becomes more severe when the dimension of the integral increases. Recently, Fu and Wang (2002) have developed a new Monte Carlo method for multivariate sampling. In this paper, we extend their algorithm to numerical integration problems. We demonstrate that this approach has many advantages over importance sampling method. In particular, this algorithm is non-iterative, dimension-free and easy to implement. Some benchmark examples are used to demonstrate this algorithm and to compare with importance sampling method.

L'échantillonnage d'importance est un outil populaire en intégration numérique. Cependant, la précision de ses résultats dépend de la densité d'importance choisie; un choix inadéquat peut causer une perte dévastatrice précision. De plus, dans la pratique il n'y a aucune procédure qui garantie un bon choix de la densité d'importance. Ce problème devient plus grave lorsque la dimension de l'intégrale augmente. Récemment, Fu et Wang (2002) ont développé une nouvelle méthode de Monte Carlo pour l'échantillonnage multivarié. Dans cette présentation, nous prolongeons leur algorithme aux problèmes d'intégration numérique. Nous démontrons que cette approche a beaucoup d'avantages par rapport aux

méthodes d'échantillonnage d'importance. En particulier, cet algorithme est non-itératif, indépendant de la dimension et facile à implémenter. Quelques exemples repères sont utilisés pour démontrer l'usage de cet algorithme et pour comparer avec les méthodes d'échantillonnage d'importance.

Tuesday June 10 • Mardi 10 juin 4:30 • 16h30 LSC 234

Francois Perron, Yves Atchade, Université de Montréal

Monte Carlo simulations via control variates • Methodes de simulations utilisant des variables de contrôle

This talk proposes methods to improve Monte Carlo estimates of a mean μ using a covariate. During the talk, we shall propose a new unbiased estimator when the correlation with the covariate is unknown. We shall apply our results to the importance sampling and the accept-reject algorithm. We shall also work on Rao-Blackwellizations. Numerical results will be given.

Cet exposé propose des méthodes pour améliorer l'estimateur standard de la moyenne par une simulation de Monte Carlo en utilisant des covariables. Nous allons proposer un nouvel estimateur sans biais lorsque les covariances entre la variables d'intérêt et les covariables est inconnue. Nous allons appliquer nos résultats aux algorithmes d'échantillonnage pondéré et à la méthode d'acceptation-rejet. Nous utiliserons également une Rao-Balckwellisation pour améliorer encore plus nos estimateurs. Des résultats numériques seront présentés.

Tuesday June 10 • Mardi 10 juin 4:45 • 16h45 LSC 234

Caryn Thompson, University of New Brunswick (Saint John); Leah Passmore, Liliana Gonzalez, University of Rhode Island

Linear regression in the presence of spatially correlated errors: a computer intensive approach • Régression linéaire avec présence d'erreurs corrélées spatialement: une approche computationnelle intensive

Computer intensive statistical techniques, such as randomization, bootstrapping and jackknifing, have been used for testing hypotheses concerning linear regression model parameters (Edgington, 1995; Manly, 1997; Anderson, 2001). These methods are appealing, in that they assume only that the model errors are identically and independently distributed. Little is known about the effect of correlated errors on performance of these methods. Computer intensive procedures allowing for dependent data are available. In the time series setting, most of these involve dividing the series into either overlapping (Carlstein, 1986) or non-overlapping (Kunsch, 1989) blocks. The blocks are then randomly resampled with replacement, and joined together to produce a version of the original series. Hall (1985) proposed a similar blocking approach for spatial data. Several variants of this technique have been developed for a number of purposes, including prediction of spatial cumulative distribution functions (Lahiri et al, 1999), and comparison of cumulative distribution functions over subregions (Zhu and Morgan, 2003). However, the performance of block resampling techniques has not been evaluated in the context of linear regression analysis. This study compares several standard computer intensive procedures (randomization, bootstrapping and jackknifing of observations and residuals) and spatial resampling methods (overlapping and non-overlapping blocks) for significance testing of regression coefficients when model errors are spatially correlated. Simulation experiments were performed under several spatial covariance structures (including SAR(1), spherical, exponential, and

Gaussian) to evaluate the size and power of tests achieved using each method. Recommendations are given concerning practical application of computer intensive techniques for regression models with spatially dependent errors.

Des techniques statistiques computationnelles intensive, telles la randomisation, le bootstrap et le jackknife, ont été utilisées pour tester des hypothèses sur les paramètres du modèle de régression linéaire (Edgington, 1995; Manly, 1997; Anderson, 2001). Ces méthodes sont intéressantes du fait qu'elles supposent seulement que les erreurs du modèle sont indépendantes et identiquement distribuées. Cependant, peu est connu sur l'effet des erreurs corrélées sur la performance de ces méthodes. Des procédures computationnelles intensives qui tiennent compte des données dépendantes sont disponibles. Dans le cadre des séries chronologiques, la plupart de ces dernières impliquent de séparer la série soit en blocs qui se chevauchent (Carlstein, 1986) ou non (Kunsch, 1989). Les blocs sont ensuite aléatoirement rééchantillonnés avec remplacement, et regroupés ensemble pour produire une nouvelle version de la série originale. Hall (1985) a proposé une approche par blocs semblable pour des données spatiales. Plusieurs variantes de cette technique ont été développées pour un certain nombre de buts, dont la prévision des fonctions de distribution cumulatives spatiales (Lahiri et autres, 1999), et la comparaison des fonctions de distribution cumulatives sur une sous région (Zhu et Morgan, 2003). Cependant, la performance des techniques de rééchantillonnage en blocs n'a pas été évaluée dans le cadre de l'analyse de régression linéaire.

Session/Séance 38 • Statisticians in Action II • Statisticiens en action II

Tuesday June 10 • Mardi 10 juin **3:30 • 15h30** **LSC 332**

Video presentation • Présentation vidéo

Committee on Professional Development • Comité sur le perfectionnement professionnel

Session/Séance 39 • Statistical Methods for Health Services and Outcomes Research • Méthodes statistiques pour les services de santé et la recherche sur les outcomes

Wednesday June 11 • Mercredi 11 juin **8:30 • 8h30** **LSC 238**

Jamie Stafford, University of Toronto; John Braun, University of Western Ontario; Thierry Duchesne, Universite Laval

A kernel density estimate for interval censored data • Un estimateur par le noyau de la densité pour des données censurées par intervalles

We propose a method for kernel smoothing data that is either interval-censored or has been aggregated into bins. The method relies on weights computed using a proposed kernel density estimate for such data. This estimate retains the simplicity and intuitive appeal of the usual kernel density estimate for complete data and is easy to compute. It results from an algorithm where conditional expectations of a kernel are computed at each iteration. These conditional expectations are computed with respect to the density estimate from the previous iteration as part of a Monte Carlo scheme. The estimator is applied to HIV data where interval censoring is common and to the Ontario health survey where data has been aggregated into bins. The proposed estimator is embedded in a rich framework for

this datatype. In terms of the cumulative distribution function the algorithm is shown to coincide with the self-consistency algorithms of Efron (1967), Turnbull (1976), and Li et al. (1997), as the window size of the kernel shrinks to zero. It modifies these algorithms sensibly by introducing kernel smoothing at each iteration resulting in an estimate that is continuous in appearance rather than involving a step function. Alternatively, it may be viewed as a special case of the local likelihood approach to density estimation given in Hjort and Jones (1996) that has been embedded in a multiple imputation scheme. Convergence within this class is assessed through the examination of the spectral radius of a fixed point equation and "leave one (inner-most interval) out" likelihood cross validation is proposed as a method for choosing the smoothing parameter. "Obvious" extensions to intensity estimation are suggested.

Nous proposons une méthode par le noyau pour lisser des données qui sont soient censurées par intervalles ou agrégées dans des casiers. La méthode se fonde sur des poids calculés en utilisant une estimation par le noyau de la densité proposé pour de telles données. Cette estimation conserve la simplicité et l'intuitivité de l'estimation par le noyau habituelle de la densité pour des données complètes et il est facile à calculer. Elle résulte d'un algorithme où des espérances conditionnelles d'un noyau sont calculées à chaque itération. Ces espérances conditionnelles sont calculées par rapport à l'estimation de la densité à l'itération précédente comme élément d'un schème de Monte Carlo. L'estimateur est appliqué à des données sur le VIH, où la censure par intervalles est commune et à l'enquête sur la santé en Ontario, où les données ont été agrégées dans des casiers. L'estimateur proposé fait parti d'un cadre riche pour ce type de données. En termes de fonction de répartition, nous montrons que l'algorithme coïncide avec les algorithmes de consistance individuel d'Efron (1967), de Turnbull (1976), et de Li et autres (1997), quand la taille de la fenêtre du noyau tend vers zéro. Il modifie ces algorithmes sensiblement en lissant le noyau à chaque itération ayant pour résultat une estimation d'aspect continue plutôt que d'impliquer une fonction en escalier. De manière alternative, l'estimateur peut être vu comme cas spécial de l'approche de la vraisemblance locale à l'estimation de la densité donnée dans Hjort et Jones (1996) qui a été imbriqué dans un schème d'imputations multiples. La convergence parmi cette classe est évaluée par l'examen du rayon spectral d'une équation de point fixe et la validation croisée " leave one out (l'intervalle le plus à l'intérieur)" de la vraisemblance est proposé comme méthode pour choisir le paramètre de lissage. Des prolongements évidents pour l'estimation de l'intensité sont suggérés.

Wednesday June 11 • Mercredi 11 juin 9:00 • 9h00 LSC 238

K.K. Gordon Lan, Yuhwen Soo, Zhenming Shun, Aventis Pharmaceuticals, NJ

Two-stage winner design • Design à deux étapes gagnant

We study a two-stage decision-process design where, in addition to the "control", there are two competing interventions being considered for a randomized trial. For example, following hospital discharge for patients with congestive heart failure, transitional care plans A and B both were designed to possibly improve hospital-to-home transition, in terms of health-related quality of life, rates of readmission, and/or emergency room use. The effectiveness of these two interventions (A and B) is to be compared to the usual care (C). During recruitment, patients are randomized in two stages. In stage I, patients are randomized to receive care plans A, B, or C (i.e. three randomization groups in stage I). The outcome responses of A and B are compared, a "winner" W between A and B is determined, and the losing strategy is dropped from further study. In stage

II, newly recruited patients are now allocated to groups W and C. At the end of the study, the combined responses of all patients in plans W and C from stages I and II will be evaluated to determine whether W is better than C. Such a design involves early decision on a favorable intervention, and the combined samples from stages I and II increase the power for hypothesis testing. We will examine the distribution of the test statistic associated with this design. For the purpose of hypothesis testing, we will provide a very simple approximation method, as well as modifications for the exact and the almost-exact approaches. Extensions of this design will also be discussed.

Nous étudions un plan de processus de décision à deux étapes où, en plus du contrôle, il y a deux interventions concurrentes considérés pour une épreuve randomisée. Par exemple, à la suite de la sortie de l'hôpital pour des patients avec un insuffisance cardiaque congestive, les plans de soins transitionnels A et B ont été conçus pour améliorer la transition de hôpital vers la maison en termes de qualité de vie relative à la santé, de taux de réadmission, et/ou le recours à l'urgence. L'efficacité de ces deux interventions (A et B) doit être comparée au soins habituels (C). Pendant le recrutement des sujets, les patients sont randomisés en deux étapes. Dans l'étape I, les patients sont randomisés pour recevoir les plans de soin A, B, ou C (c.-à-d. trois groupes de randomisation dans étape I). Les résultats de A et de B sont comparées et on détermine un gagnant W entre les deux et la stratégie perdante est abandonnée pour les études subséquentes. Dans l'étape II, les patients nouvellement recrutés sont assignés aux groupes W ou C. À la fin de l'étude, les réponses combinées de tous les patients des plans W et C des étapes I et II seront évalués pour déterminer si W est meilleur que C. Un tel design implique une décision rapide sur une intervention favorable, et les échantillons combinés des étapes I et II augmentent la puissance pour effectuer des tests d'hypothèses. Nous allons examiner la distribution de la statistique du test associée à ce design. Dans le but de faire de tests d'hypothèses, nous montrons une méthode d'approximation très simple, ainsi que des modifications pour les apporches exacte et quasi exacte. Nous discuttons également des prolongements de ce design.

Wednesday June 11 • Mercredi 11 juin 9:30 • 9h30 LSC 238

John Koval, University of Western Ontario

The intraclass Kappa • La statistique Kappa intra-classse

The kappa statistic was first proposed by Scott (1955) and Cohen (1960) as a "chance corrected" measure of agreement between two raters. Kramer (1979) instead defined it as an intraclass measure of reliability, making it a parameter in a unique probabiliity distribution so that a maximum likelihood estimator and asymptotic standard errors of various estimators, including Cohen's original "kappa", could be determined, and comparisons made between estimators (see, for example, Blackman and Koval, 2001). For this same model, Donner and Eliasziw (1992) proposed a "goodness of fit" method for computing significance tests and asymmetric confidence intervals which had excellent statistical properties, particularly for small to moderate sample sizes. Investigations have been made into inferential procedures for generalizations of the binary outcome, two rater case, including more outcomes (both nominal and ordinal), and more that two raters. Although Kramer (2002) despairs of a meaningfuil interpretation of reliability when the outcome is nominal, there have been several recent extensions, one of the "goodness of fit" method to the multinomial outcome, two rater case, and others, of both the "goodness of fit" and of a

related "modified Wald" method, to the binary outcome, multirater case. The statistical properties of these methods of estimation are being investigated.

La statistique kappa a été spécifiée la première fois par Scott (1955) et Cohen (1960) comme une mesure de " corrigé pour le hasard " d'accord entre deux juges. Kramer (1979) l'a défini comme une mesure de fiabilité intra-classe, en faisant ainsi un paramètre d'une distribution de probabilité unique de sorte qu'un estimateur du maximum de vraisemblance et des écarts types asymptotiques de divers estimateurs, y compris le "kappa" original de Cohen, peuvent être déterminés. Cette dernière définition donne aussi la possibilité de faire des comparaisons entre les estimateurs (voir par exemple, Blackman et Koval, 2001). Pour ce même modèle, Donner et Eliasziw (1992) ont proposé une méthode de qualité d'ajustement pour calculer les tests de signification et les intervalles de confiance asymétriques qui ont d'excellentes propriétés statistiques, en particulier pour des petits et moyens échantillons. Des investigations ont été faites dans les procédures d'inférence pour des généralisations des résultats binaires, des cas de deux juges avec plus de résultats (nominal et ordinal), et de plus de deux juges. Bien que Kramer (2002) semble désespéré de trouver une interprétation de fiabilité complète et sensée quand les résultats sont nominaux, il y a eu plusieurs prolongements récents, un sur la méthode de qualité d'ajustement aux résultats multinomiaux, et d'autres au cas de deux juges, pour la qualité d'ajustement et de la méthode modifié de Wald reliée aux résultats binaires, et au cas de plusieurs juges. Les propriétés statistiques de ces méthodes d'estimations sont étudiées.

Session/Séance 40 • Statistics and Climate Change • Statistique et changement climatique

Wednesday June 11 • Mercredi 11 juin 8:30 • 8h30 LSC 242

Peter Stott, Gareth Jones, Hadley Centre for Climate Research; Myles Allen, Oxford University

Optimal detection of anthropogenic climate change • Détection optimale des changements climatiques anthropogéniques

We describe optimal detection techniques and how they have been applied to understanding recent climate change. Using fingerprints of climate change estimated from ensembles of simulations of the HadCM3 ocean-atmosphere climate model we estimate the contributions, with their uncertainties, of different climate forcing agents, both anthropogenic and natural, to observed 20th century temperature changes. Recently we have modified the standard optimal fingerprint algorithm to take account of uncertainty in model-simulated responses to climate change due to internal chaotic variability. Our results show that greenhouse gas increases explain most of the global warming observed in the second half of the twentieth century with increases in solar output being potentially a major contributor to warming observed in the first half of the 20th century.

Nous décrivons des techniques optimales de détection et expliquons comment elles ont été appliquées pour comprendre les changements climatiques récents. En utilisant des empreintes digitales de changements climatiques estimées à partir d'un ensemble de simulations du modèle climatique océan-atmosphère HadCM3, nous estimons les contributions, avec leurs incertitudes, de différents agents de forçage climatiques anthropogéniques et naturels pour observer les changements de températures aux XXe siècle. Récemment, nous avons modifié l'algorithme optimal standard d'empreinte digitale pour tenir compte de l'incertitude des réponses aux changements climatiques dans les simulations modélisées à cause d'une variabilité interne chaotique. Nos résultats prouvent que l'augmentations des

gaz à effet de serre expliquent la majeure partie du réchauffement planétaire observé dans la deuxième moitié du vingtième siècle et l'augmentations du rendement solaire est potentiellement un contribuant important du réchauffement de la première moitié du XXème siècle.

Wednesday June 11 • Mercredi 11 juin 9:00 • 9h00 LSC 242

Myles Allen, Hugo Lambert, Daithi Stone, Oxford University

Will it ever be possible to attribute apparently anomalous weather events to anthropogenic climate change? • Sera-t-il possible d'attribuer des événements météorologiques extraordinaires au réchauffement global?

The flood waters of the Thames are 30cm from our kitchen door, and the view from the UK Met Office is that "while this is the kind of event we might expect to become more likely in a changing climate, it is impossible to attribute this particular event to anthropogenic climate change." Is this an impossibility in principle, or simply a problem of inadequate knowledge? In a highly chaotic system, will it ever be possible to make meaningful attribution statements about the role of external drivers in specific weather events? I will argue that it is possible to make such statements, and while attribution claims will always be partial and uncertain, it is nevertheless possible to formulate a rigorous and quantitative process of "probabilistic attribution" with strong methodological links to the problem of probabilistic forecasting. The practical challenges are considerable, but the case for further research in this area is strong, since there is clearly a high level of public interest in this question (our neighbour with the flooded kitchen would like to know the answer). If successful, it also opens up the possibility of new approaches to redistributing of the costs of climate change.

Les eaux d'inondation de la Tamise sont à 30cm de notre porte de cuisine, telle est la vue du bureau du UK Met Office, et c'est le genre d'événement que nous pourrions voir devenir plus probables à cause du changement climatique. Il est impossible d'attribuer cet événement au changement climatiques anthropogène. Est-ce une impossibilité en principe, ou simplement le problème d'une connaissance inadéquate? Dans un système fortement chaotique, sera-t-il jamais possible de faire des rapports significatifs d'attribution au sujet du rôle des facteurs externes dans des événements climatiques spécifiques? J'argumenterai le fait qu'il est possible de faire de tels rapports, et tandis que les attributions seront toujours partielles et incertaines, il est néanmoins possible de formuler un processus rigoureux et quantitatif d'attributions probabilistes avec des liens méthodologiques forts au problème des prévisions probabilistes. Les défis pratiques sont considérables, mais le désir d'augmenter le nombre de recherches dans ce secteur est fort, puisqu'il y a clairement un niveau d'intérêt public élevé pour cette question (notre voisin avec la cuisine inondée voudrait savoir la réponse). Si ça fonctionne, ça ouvre également la possibilité de nouvelles approches pour redistribuer les coûts du changement climatique.

Wednesday June 11 • Mercredi 11 juin 9:30 • 9h30 LSC 242

Chris Forest, Andrei P. Sokolov, Peter H. Stone, Massachusetts Institute of Technology; Myles R. Allen, Oxford University

PDFs of climate system properties including natural and anthropogenic historical climate forcings • Fonctions de densités des propriétés des systèmes climatiques incluant des forçages climatiques historiques anthropogènes et naturels

Previous results for probability density functions (PDFs) of climate system properties treated uncertainty both in the climate forcings and in the response to the total anthropogenic forcing over the 1860-1995 period. The treatment of the uncertainty in the climate forcings was accomplished by varying the net aerosol forcing and thus included the uncertainty due to all forcings that were not explicitly included (e.g., solar and volcanic forcings). Primarily, this treatment was dictated by the computational requirements of including additional uncertain parameters as well as maintaining a level of complexity in the forcings commensurate with the zonally averaged climate model. At this time, we are improving our estimates of the PDFs by including additional forcings, extending the observational data records (up to 2001), and improving the climate model resolution (from 7.86 to 4 degrees latitude). In addition to changes in concentrations of greenhouse gases, sulfate aerosols, and ozone (1979-1995 only) (aka, GSO) for 1860-1995, we include changes in: the stratospheric aerosols from volcanic eruptions (Sato et al., 1993), solar irradiance (Lean, 2000), and land-use vegetation (Ramunkutty & Foley, 1999) and apply all forcings (GSOVSV) for 1860-2001. As in previous work (Forest et al., 2002), we compare the modelled climate changes to multiple observational diagnostic datasets and compute goodness-of-fit statistics as derived from climate change detection algorithm. We explicitly vary the climate system properties, P, over a wide range to locate regions of parameter space that are inconsistent with the observed record of climate change. Here, P is the set of uncertain climate system properties (climate sensitivity, deep-ocean heat uptake, and net aerosol forcing.) Then, based on the goodness of fit statistics, for each diagnostic, we estimate the likelihood functions $p(D_i|P)$ where D_i are individual climate change diagnostics (surface, upper-air, or deep-ocean temperature changes for late 20th century.) The likelihood functions $p(D_i|P)$ are then used to update the posterior PDF, $p(P|D\_i)$ according to Bayes theorem where assumed priors are required (see Forest et al., 2002.) The main focus of this talk will be to compare the updated PDFs with the previous results given the updated climate model resolution, longer observational records, and additional climate forcings. We will also compare the climate change detection results for GSO forcings with those for GSOVSV as climate sensitivity and other parameters vary. The individual impacts of the solar and volcanic forcings will be explored as well on the detection results.

Les résultats existant sur les fonctions de densité (PDF) des propriétés des systèmes climatiques traitent conjointement de l'incertitude dans les forçages climatiques et de la réponse à tous les forçages anthropogènes au cours de la période 1860-1995. Le traitement de l'incertitude dans les forçages de climat a été accompli en changeant le forçage aérosol net et a ainsi inclus l'incertitude due à tous les forçages qui n'ont pas été explicitement inclus (par exemple, des forçages solaires et volcaniques). Ce traitement a principalement été dicté par les besoins informatiques d'inclure des paramètres incertains additionnels aussi bien que maintenir un niveau de complexité dans les forçages proportionnés au modèle climatique de moyenne des zones. Actuellement, nous améliorons nos estimations de la fonction de densité en incluant des forçages additionnels, en augmentant le nombres d'observations enregistrées (jusqu'en 2001) et en améliorant la résolution du modèle climatique (de 7,86 à 4 degrés de latitude). En plus des changements de concentrations des gaz à effets de serres, des sulfates en aérosols, et de l'ozone (pour 1979-1995 seulement) (aussi appelés GSO) pour 1860-1995, nous incluons des changements dans: les aérosols stratosphériques provenant des éruptions volcaniques (Sato et autres, 1993), de l'irradiance solaire (Lean, 2000), et de la végétation d'utilisation des sols (Ramunkutty et

Foley, 1999) et nous appliquons tous les forçages (GSOVSV) pour 1860-2001. Comme dans les travaux précédents (Forest et autres, 2002), nous comparons les changements du climat modélisés à l'ensembles des données d'observation multiples et nous calculons des statistiques d'adaptation au modèle comme dérivé de l'algorithme de détection de changements climatiques. Nous modifions explicitement les propriétés du système climatique, appelées P, sur un large éventail pour localiser les régions de l'espace paramétrique qui sont contradictoires avec les observation de changements climatiques. Ici, P est l'ensemble des propriétés incertaines de systèmes climatiques (sensibilité du climat, prise de la chaleur provenant des fonds océaniques, et les forçages aérosols nets.) Puis, basé sur la qualité des statistiques d'ajustement, pour chaque diagnostic, nous estimons la fonction de vraisemblance $p(D_i|P)$ *où* D_i *sont les diagnostics individuels de changements climatiques (la surface, l'air supérieure, ou les changements de température en profondeur des océans pour la fin du 20ème siècle.) Les fonctions de vraisemblances* $p(D_i|P)$ *sont alors utilisées pour mettre à jour la fonction de densité a posteriori,* $p(D_i|P)$ *par le théorème de Bayes où les densités a priori supposées sont requises (voir Forest et autres, 2002.) Dans cette présentation, nous comparons la fonction de densité a posteriori aux résultats précédents donnés la mise à jour de la résolution du modèle climatique, de plus grosses banques observations, et des forçages climatiques additionnels. Nous comparons également les résultats de détection de changement de climat pour des forçages de GSO à ceux de GSOVSV avec le changement de la sensibilité climatique et d'autres paramètres. Les différents impacts des forcings solaires et volcaniques seront aussi explorés sur les résultats de détection.*

Session/Séance 41 • Bayesian Analysis • Analyse bayésienne

Wednesday June 11 • Mercredi 11 juin 8:30 • 8h30 LSC 338

Michael Newton, Hyuna Yang, University of Wisconsin; David Hastie, Bristol University

Statistical methods to analyze genomic aberrations in cancer cells: the case of overlapping ensembles • Méthodes statistiques pour analyser les aberrations génomiques dans les cellules cancéreuses: le cas des ensembles qui se chevauchent

Biomolecular technologies allow oncologists to survey genomic damage in cancer cells and to record the many and varied aberrations presented in a sample of tumors. In a recent article (Newton, 2002, JASA, 97:931-942), a statistical approach is described for the analysis of such data where the goal is to identify collections of aberrations having the property that their co-occurrence in a progenitor cell is somehow beneficial to the tumor. These ensembles of genomic aberration were restricted in that work to be non-overlapping. Here we describe efforts to deal with with the general setting of overlapping ensembles. Statistically, the problem is a model-based cluster analysis in which we have potentially overlapping clusters. Details of prior and posterior calculations will be discussed in the context of data measured by comparative genomic hybridization.

Les technologies biomoléculaires donne la possibilité aux oncologistes de sonder le dommage génomique dans des cellules cancéreuses et d'enregistrer les aberrations présentent dans un échantillon de tumeurs. Dans un article récent (Newton, 2002, JASA, 97:931-942), une approche statistique est décrite pour l'analyse de telles données où le but est d'identifier les collections d'aberrations qui ont la propriété que leurs co-occurrence dans une cellule progéniteure est quelque peu bénéfique pour la tumeur. Dans l'article mentionné précédemment, les ensembles d'aberrations génomiques ont été définis de sorte qu'ils ne se chevauchent pas. Dans cette présentation, nous décrivons les efforts requis pour travailler avec des ensembles qui se chevauchent. Statistiquement, le problème est un modèle d'analyse en grappe dans lequel nous avons potentiellement des grappes qui se chevauchent. Nous discutons des détails du calcul de la fonction a priori et de la fonction a posteriori dans le contexte de données mesurées par hybridation génomique comparative.

Wednesday June 11 • Mercredi 11 juin 9:00 • 9h00 LSC 338

David Higdon, Los Alamos National Laboratory; Herbie Lee, University of California at Santa Cruz

Characterizing uncertainty in inverse problems • Caractérisation de l'incertitude pour des problèmes inverses

Simulation based inference in computationally intensive inverse problems. A typical setup for many inverse problems is that one wishes to update beliefs about a spatially dependent set of inputs x given rather indirect observations y. Here, the inputs and observed outputs are related by complex physical relationship y = F(x) + e. Applications include medical and geological tomography, hydrology, and the modeling of physical and biological systems. We consider applications where the physical relationship F(x) can be well approximated by detailed simulation code f(x). When the forward simulation code f(x) is sufficiently fast, Bayesian inference can, in principle, be carried out via MCMC. Difficulties arise for two main reasons: (i) Even though the code may accurately represent the physical process, there are a large number of unknown, but required, inputs that must be calibrated to match the observed data y. (ii) The computational burden of the fastest available forward simulators is often large enough that approaches for speeding up the MCMC calculations are required. This talk develops approaches for specifying effective low-dimensional representations of the inputs x along with MCMC approaches for sampling the posterior distribution. In particular we consider augmenting the basic formulation with fast, possibly coarsened, formulations to improve MCMC performance. This approach can be very easily implemented in a parallel computing environment. We give examples in single photon emission computed tomography (SPECT) and in hydrology.

Une situation typique pour beaucoup de problèmes inverses est que l'on souhaite mettre à jour la connaissance d'un ensemble d'observations spatialement dépendantes x à l'aide d'observations indirectes y. Ici, les entrées et les sorties observées sont reliées par une relation physique complexe y = F(x) + e. Les applications incluent la tomographie médicale et géologique, l'hydrologie et la modélisation de systèmes physiques et biologiques. Nous considérons les applications où le rapport physique F(x) peut être bien estimé par le code de simulation détaillé f(x). Lorsque le code de simulation vers l'avant f(x) est suffisamment rapide, l'inférence bayésienne peut, en principe, être effectuée par l'intermédiaire des méthodes MCMC. Des difficultés surgissent pour deux raisons principales: (i) Quoique le code puisse représenter exactement le processus physique, il y a un grand nombre d'inconnus nécessaires qui doivent être calibrées pour ajuster les données observées y. (ii) Le fardeau informatique des simulations vers l'avant les plus rapides est souvent si lourd que des techniques pour accélérer les calculs MCMC sont requises. Cet entretien développe des approches pour spécifier des représentations efficaces de faibles dimensions des entrées x avec des approches MCMC.

Wednesday June 11 • Mercredi 11 juin 9:30 • 9h30 LSC 338

Jean-François Angers, Stéphane Courchesne, Louis-François Poirier, Université de Montréal; Claire Laberge-Nadeau, (CRT)

Link between cell-phone and car crashes • Lien entre le téléphone cellulaire et les accidents de la route

In this presentation, we want to see if the use of cell-phone while driving increases the risk of an accident. Moreover, we also investigate if there is a dose-response relation between the frequency of cell-phone use and the car accidents. The data for this project come from a

large epidemiological study done at the Laboratory on Transportation Safety of the Centre de recherche sur les transports (CRT) with the participation of the Société de l'assurance automobile du Québec (SAAQ) and the four major Canadian cellular phone companies. Hence, we want to estimate the instantaneous risk, a measure of association between the two variables "to use a mobile phone while driving" and "to have a car accident." Since we only observe the cases when there is an accident, we cannot estimate this risk directly. However, using Bayesian approach, we are able to model the missing information. Using the EM algorithm along with Monte Carlo simulation, we obtain an overall instantaneous risk of 1.735. Adjusting for the frequency of their use of the cell-phone, the instantaneous risk goes from 0.781 for the small users to 2.270 for the heavy ones.

Dans cette présentation, nous voulons savoir si l'utilisation du téléphone cellulaire en conduisant augmente le risque d'avoir un accident de la route. De plus, nous voulons également savoir s'il y a une relation de dose-réponse entre la fréquence d'utilisation du téléphone cellulaire et les accidents de la route. Les données pour ce projet proviennent d'une étude épidémiologique exhaustive faite par le Laboratoire sur la sécurité des transports du Centre de recherche sur les transports (CRT) avec la participation du Société de l'assurance automobile du Québec (SAAQ) et des quatre principales compagnies canadiennes de téléphonie cellulaire. Par conséquent, nous voulons estimer le risque instantané, une mesure d'association entre les deux variables ¡¡ utiliser un téléphone cellulaire en conduisant¿¿ et ¡¡ avoir un accident de voiture¿¿ . Puisque nous observons seulement les cas où il y a un accident, nous ne pouvons pas estimer ce risque directement. Cependant, en utilisant une approche bayésienne, nous pouvons modéliser l'information manquante. En utilisant l'algorithme EM et des simulations Monte-Carlo, nous obtenons un risque instantané global de 1,735. En tenant compte de la fréquence d'utilisation du téléphone cellulaire, le risque instantané va de 0,781 pour les petits utilisateurs à 2,270 pour les grands utilisateurs.

Session/Séance 42 • Statistical Inference • Inférence statistique

Wednesday June 11 • Mercredi 11 juin 8:30 • 8h30 LSC 332

Paul Marriott, National University of Singapore and Duke University

Mixture models and geometry • Modèles de mélange et géométrie

This talk takes a geometric approach to modelling with mixtures. The flexibility and interpretability of mixtures models make them very attractive tools for modelling, however they typically also have difficulties that make them unattractive inferencially. Such problems include identifiability, problems with dimensionality and the existence well defined parametrisations. The geometric approach explores these issues and helps to motivate a new class of, so called, true local mixture models which retain flexibility and interpretability while being also attractive inferentially. Applications which will be discussed include measurement error modelling.

Cette présentation adopte une approche géométrique pour modéliser avec des mélanges. La flexibilité et l'interprétabilité des modèles de mélanges en font des outils très attrayants pour modéliser. Toutefois, elles ont également certaines difficultés qui les rendent moins intéressantes pour l'inférence. De tels problèmes incluent des problèmes d'identification, des problèmes de dimensions et l'existence de paramétrisations bien définies. L'approche géométrique explore ces problèmes et aide à motiver une nouvelle classe de véritables modèles de mélange locaux qui maintiennent la flexibilité et l'interprétabilité tout en étant

intéressant pour l'inférence statistique. Les applications qui sont présentées incluent la modélisation de l'erreur de mesure.

Wednesday June 11 • Mercredi 11 juin 8:45 • 8h45 LSC 332

Thomas O'Gorman, Northern Illinois University

Adaptive statistical methods • Méthodes statistiques adaptatives

An adaptive statistical method uses the data to determine what statistical procedure would be most appropriate for the analysis. I will describe several ways of constructing adaptive tests of significance so that they do maintain the nominal level of significance. I will also show the advantages that adaptive methods have over traditional methods. Usually, if at least 20 observations are used in the analysis, the adaptive tests are more powerful than the traditional tests for non-normal error distributions. In addition, software for adaptive tests will be described. I will also demonstrate a method of computing adaptive confidence intervals and will illustrate its use by an example. By carefully constructing adaptive confidence intervals we find that they are often narrower than the traditional confidence intervals, while they maintain their nominal coverage probabilities.

Une méthode statistique adaptative utilise les données pour déterminer quel procédé statistique serait le plus approprié pour l'analyse. Nous décrivons plusieurs manières de construire les tests d'hypothèses adaptatifs de sorte qu'ils maintiennent le niveau de confiance nominal. Nous prouvons également les avantages que les méthodes adaptatives ont par rapport aux méthodes traditionnelles. Habituellement, si au moins 20 observations sont utilisées dans l'analyse, les tests adaptatifs sont plus puissants que les tests traditionnels pour des distributions non-normales des erreurs. De plus, un logiciel pour des tests adaptatifs sera présenté. Nous démontrons également une méthode permettant de calculer des intervalles de confiance adaptatifs et nous illustrons son utilisation par un exemple. En construisant soigneusement des intervalles de confiance adaptatifs nous constatons qu'ils sont souvent plus courts que les intervalles de confiance traditionnels, bien qu'ils maintiennent leurs probabilités de couverture nominales.

Wednesday June 11 • Mercredi 11 juin 9:00 • 9h00 LSC 332

Jianan Peng, Acadia University; C.I.C. Lee, L. Liu, Memorial University of Newfoundland

Cone order monotonicity of tests for treatments versus a control • L'ordonnancement monotone dans un cône des tests de traitements versus un groupe contrôle

A problem frequently encountered in the practice of statistics is comparing several treatment means with a control mean, or a standard. In many situations, experimenters may have the prior knowledge that each treatment mean is at least as large as the control mean, or each treatment mean is at least as large as the grand mean. The likelihood ratio test for testing the homogeneity of treatment means and the control mean versus the alternative restricted by the prior knowledge of treatments are at least as good as the control is not cone order monotonicity, as observed by Cohen, Kemperman, and Sackrowitz (2000). A lack of cone order monotonicity for a test procedure may be counter-intuitive and undesirable. We consider a likelihood ratio test for testing homogeneity versus the alternative restricted by the prior knowledge of each treatment mean is at least as large as the grand mean. The new likelihood ratio test is cone order monotonicity. We also offer an alternative test procedure which is cone order monotonicity for both restrictions and has competitive power performance.

Un problème que l'on rencontre fréquemment dans la pratique des statistiques est la comparaison des moyennes de plusieurs traitements à la moyenne d'un groupe contrôle, ou à une norme. Dans beaucoup de situations, les expérimentateurs peuvent savoir à l'avance que chacune des moyennes des traitements sont au moins aussi grandes que la moyenne du groupe contrôle, ou bien que chacune des moyennes des traitements sont au moins aussi grandes que la moyenne générale. Le test du rapport de vraisemblance pour tester l'homogénéité des moyennes des traitements et du groupe contrôle contre l'alternative limité par la connaissance a priori que les traitements sont au moins aussi performants que le groupe contrôle n'est pas la propriété d'ordonnancement monotone dans un cône, comme l'ont observé Cohen, Kemperman et Sackrowitz (2000). Le fait de ne pas posséder la propriété d'ordonnancement monotone dans un cône pour une procédure de test peut aller à l'encontre de l'intuition et ainsi être indésirable. Nous considérons un test du rapport de vraisemblance pour tester l'homogénéité contre l'alternative limitée par la connaissance a priori que chaque moyenne est au moins aussi grandes que le moyenne générale. Le nouveau test du rapport de vraisemblance possède la propriété d'ordonnancement monotone dans un cône. Nous présentons également une procédure de test alternative qui possède la propriété d'ordonnancement monotone dans un cône pour les deux restrictions et qui présente une performance compétitive pour la puissance.

Wednesday June 11 • Mercredi 11 juin 9:15 • 9h15 LSC 332

Yogendra Chaubey, Concordia University

Measures of overlap for inverse Gaussian populations • Mesures de chevauchement pour des populations de densité gaussienne inverse

Inference for various measures of overlap for two normal populations have been extensively studied in literature(see Mulekar and Mishra (1994), J. Japan Statist. Soc., Vol. 24, 169-180 and Mulekar and Mishra (2000), Comput. Stat. and Data Anal., vol 34, 121-137 for details and other references.) In this paper we consider the well known alternative to the Gaussian model, namely the inverse Gaussian model and study the properties of these measures with respect to point and confidence interval estimation. Some striking similarities to the Gaussian case are observed for some measures but for other measures properties may be quite different.

L'inférence pour différentes mesures de chevauchement pour deux populations normales a été intensivement étudiée dans la littérature (voir Mulekar et Mishra (1994), J. Japon Statist. Soc., vol. 24, 169-180 et Mulekar et Mishra (2000), Comput. Stat. et Data Anal., vol. 34, 121-137 pour des détails et d'autres références.) Dans cette présentation, nous considérons l'alternative bien connue au modèle gaussien, à savoir le modèle gaussien inverse et nous étudions les propriétés de ces mesures par rapport à l'estimation ponctuelle et aux intervalles de confiance. Quelques similitudes étonnantes au cas gaussien sont observées pour quelques mesures mais pour d'autres, les propriétés peuvent être tout à fait différentes.

Wednesday June 11 • Mercredi 11 juin 9:30 • 9h30 LSC 332

Peiming Wang, Nanyang Technological University, Singapore

A score test for testing a bivariate zero-inflated Poisson regression model against bivariate zero-inflated negative binomial alternative • Un test de score pour tester un modèle de régression de Poisson zéro-augmenté bivarié contre une alternative binomiale négative zéro-augmenté bivariée

For the univariate cases, the zero-inflated Poisson and negative binomial regression models are often useful for analyzing count data with a lot of zeros; and the score test (Ridout, et al., 2001) can be used to determine which of the two models is more appropriate. For the bivariate cases where the two dependent variables are positively correlated, the bivariate zero-inflated Poisson and bivariate negative binomial regression models have been used for analyzing bivariate count data with a lot of zeros in different applications (e.g., Li, et al. 1999, and Wang, 2003). However, when the nonzero counts are overdispersed in terms of the bivariate Poisson distribution, the inference about the regression parameters in the bivariate zero-inflated Poisson regression model may be misleading. This paper therefore presents a score test for testing a bivariate zero-inflated Poisson regression model against a bivariate zero-inflated negative binomial alternative. A simulation study is used to assess the performance of this test.

Pour le cas univarié, les modèles de régression de Poisson et binomiaux négatifs zéro-augmentés sont souvent utiles pour analyser des données de comptage avec plusieurs zéros. Le test de score (Ridout, et autres, 2001) peut être utilisé pour déterminer lesquels des deux modèles est le plus approprié. Pour le cas bivarié, où les deux variables dépendantes sont corrélées positivement, les modèle de Poisson et binomial négatif zéro-augmentés bivariés sont utilisés pour analyser des données de comptage bivariées avec plusieurs zéros dans plusieurs applications (par exemple, Li, et autres 1999, Wang, 2003). Cependant, lorsque les comptes de non nuls sont trop dispersés en termes de la distribution de Poisson bivariée, l'inférence sur les paramètres de régression dans le modèle de régression de Poisson zéro-augmenté bivarié peut être erronée. Cette présentation propose donc un test de score pour tester un modèle de régression de Poisson zéro-augmenté bivarié contre une alternative binomiale négative zéro-augmenté bivariée. Une étude de simulation est utilisée pour évaluer la performance de ce test.

Wednesday June 11 • Mercredi 11 juin 9:45 • 9h45 LSC 332

Regina Nuzzo, Jim Ramsay, McGill University

Functional data analysis of continuous judgments in music cognition • Analyse fonctionnelle de données des jugements continus en cognition musicale

Functional Data Analysis (FDA) is a class of tools used to analyze data whose observations consist of curves. These methods make particular use both of the smoothness of observed data and also of their derivatives. The usefulness of FDA is particularly seen when working with high dimensional data, such as a large number of measurements taken over time. In these cases traditional statistical methods can fail to capture the essential modes of variation. The unique contributions of FDA are illustrated here through a very large dataset from the field of music cognition. The observations consist of continuous measures of emotion and musical structure as experienced by audience members during a musical performance. FDA was able to help answer questions such as the following: How do audience members' judgments change throughout the performance? What sensory components of a performance affect this change? How do audience judgments relate to objective measures of the musical score, such as note density and loudness? What further information can be gained by examining the curves' derivatives? A multimedia presentation is used to demonstrate the use of FDA tools such as time-warping registration, functional linear models, graphical exploration of derivatives, and functional data transformations.

L'analyse fonctionnelle des données (AFD) est une classe d'outils utilisés pour analyser les données dont les observations se composent de courbes. Ces méthodes font l'utilisation

particulière du lissage des données observées et également de leurs dérivées. Nous voyons l'utilité de l'AFD particulièrement lorsque nous travaillons avec des données de dimension élevées, telles qu'un grand nombre de mesures prises dans le temps. Dans ces cas, les méthodes statistiques traditionnelles peuvent échouer à capturer les modes de variation essentiels. Les contributions uniques de l'AFD sont illustrées ici par un grand jeux de données du champ de la cognition musicale. Les observations se composent de mesures continues d'émotions et de structure musicale ressenties par les membres de l'assistance à une présentation musicale. L'AFD peut aider à répondre à des questions comme: comment le jugement des membres de l'assistances change-t-il au cour de la performance? Quelles composantes sensorielles d'une performance affectent ce changement? Comment les jugements des membres de l'assistance se relient-ils à des mesures objectives des scores musicaux, telles la densité des notes et le volume? Quelle autre information peut être obtenue en examinant la dérivée des courbes? Une présentation de multimédia est utilisée pour démontrer l'utilisation des outils d'AFD tels que l'enregistrement de déphasage, les modèles linéaires fonctionnels, l'exploration graphique des dérivés et les transformations fonctionnelles des données.

Session/Séance 43 • Survey Methods Contributed Session IV: Measuring the Quality of Survey Operations • Méthodes d'enquête IV: Mesure de la qualité des opérations d'enquête

Wednesday June 11 • Mercredi 11 juin 8:30 • 8h30 LSC 234

Stuart Pursey, Statistics Canada/Statistique Canada

Use of the score function to optimize data collection resources in the Unified Enterprise Survey • L'utilisation de la fonction de caractérisation pour optimiser les ressources de la collecte des données dans l'Enquête unifiée auprès des entreprises

The Unified Enterprise Survey (UES) is Statistics Canada's largest annual business survey. It is designed to measure provincial and national economic variables across a wide range of industries. Non-response and edit failure during data collection leads to significant follow-up costs. The UES Score Function identifies collection units that are most important for follow-up due to their impact on estimates by province and by industry. In this way the Score Function ensures that limited follow-up resources are used effectively to reach minimum levels of quality. In this article we examine the Score Function by describing its method, explaining why it improves the process of data collection, analyzing its impact on the quality of the final estimates, and discussing its relationship to the UES sample design.

L'Enquête unifiée auprès des entreprises (EUE) est la plus grande enquête économique annuelle de Statistique Canada. Elle est conçue pour mesurer des variables économiques provinciales et nationales pour une grande variété d'industries. Pendant la collecte des données, les coûts reliés au suivi auprès des non-répondants et à la vérification des données sont considérables. La fonction de caractérisation de l'EUE identifie les unités de collecte qui sont les plus importantes pour faire l'objet d'un suivi étant donné leur impact sur les estimations par province et par industrie. La fonction de caractérisation nous assure que les ressources consacrées au suivi sont utilisées efficacement de façon à atteindre un niveau minimum de qualité. Dans cet article, nous examinons la fonction de caractérisation en décrivant sa méthodologie, en expliquant pourquoi elle améliore le processus de collection des données, en analysant son impact sur la qualité des estimations finales et en discutant de son implication par rapport au plan de sondage de l'EUE.

Wednesday June 11 • Mercredi 11 juin 8:45 • 8h45 LSC 234

Robert Philips, Statistics Canada/Statistique Canada

The theory and applications of the score function for determining the priority of follow up in the Annual Survey of Manufactures • La théorie et les applications de la fonction de score pour déterminer la priorité de suivi pour le Sondage annuel des manufactures

Statistics Canada undertakes an Annual Survey of Manufactures (ASM), whose data is used as an indicator of the state of the country's economy. A large portion of the survey budget is spent on data collection, in particular, the follow-up of both non-respondents and respondents to confirm their data. In response to budgetary concerns and the need to lower respondent burden, it became essential to prioritize the follow-up of units in the collection process. However, as business surveys generally are such that, a small percentage of units are responsible for a large percentage of the population characteristics being estimated, it was believed that concentrating on the "larger" units would not compromise overall data quality. Given these factors, a score function was developed that would determine which units to follow up and in what order of priority. In this paper, the underlying theory behind the score function as well as its effective integration in the collection process for ASM is presented. Various simulation results are also presented.

Statistiques Canada entreprend un sondage annuel des manufactures (SAM), dont les données sont utilisées comme indicateur de l'état de l'économie du pays. Une grande partie du budget du sondage est dépensée pour la collecte des données, en particulier, pour le suivi des non-répondants et des répondants pour qu'ils confirment leurs données. En réponse aux soucis budgétaires et à la nécessité d'abaisser le fardeau des répondants, il est devenu essentiel de donner la priorité au suivi des unités au moment du processus de collection. Cependant, puisque les sondages sur les entreprises sont généralement tels qu'un petit pourcentage des unités est responsable pour un grand pourcentage des caractéristiques de la population estimée, nous avons cru que de se concentrer sur les "grandes" unités ne compromettraient pas la qualité globale des données. Étant donné ces facteurs, nous avons développé une fonction de score qui détermine pour quelles unités nous devons faire le suivi et dans quel ordre de priorité. Dans cette présentation, la théorie fondamentale derrière la fonction de score aussi bien que son intégration efficace aux processus de collection pour le SAM est présentée. Nous présentons également divers résultats de simulations.

Wednesday June 11 • Mercredi 11 juin 9:00 • 9h00 LSC 234

Jennifer Ali, Statistics Canada/Statistique Canada

Quality monitoring of large surveys using the Blaise audit trail • Surveillance de la qualité de sondage à grande échelle en utilisant la vérification rétrospective de Blaise

Statistics Canada's Canadian Community Health Survey has developed an innovative program for the ongoing evaluation of the integrity of data during data collection. This program allows for interventions during data collection to improve data quality. The program uses the Blaise audit trail which yields a richly detailed file of times and trailing within the questionnaire. It records the duration and content of each keystroke at the case level. Two primary types of indicators are duration timings and non-response. For the Canadian Community Health Survey, analysis was conducted at the national, regional, interviewer and case levels. This paper describes the process, identifies appropriate applications, discusses statistics that can be generated from the process, and outlines actions taken by Statistics Canada to improve data quality. Limitations and challenges in the

implementation of a data quality monitoring program based on the Blaise audit trail are discussed.

L'enquête canadienne sur la santé des communauté de Statistiques Canada a développé un programme innovateur pour l'évaluation continue de l'intégrité des données durant la collecte des données. Ce programme tient compte des interventions durant la collecte des données pour améliorer la qualité des données. Le programme utilise la vérification rétrospective de Blaise qui donne un rapport détaillé du temps et des retards dans le questionnaire. Il enregistre la durée et contenu de chaque frappe au niveau de chaque cas. Deux principaux types d'indicateurs sont la synchronisation de la durée et la non réponse. Pour l'enquête canadienne sur la santé des communautés, l'analyse a été conduite aux niveaux national, régional, de l'interviewer et de chaque cas. Cette présentation décrit le processus, identifie les applications appropriées, discute des statistiques qui peuvent être calculées du processus et décrit les actions prises par Statistiques Canada pour améliorer la qualité des données. Les limitations et les défis dans l'implémentation d'un programme de contrôle de qualité des données basé sur la vérification rétrospective de Blaise sont discutés.

Wednesday June 11 • Mercredi 11 juin 9:15 • 9h15 LSC 234

Fred Hazelton, Stuart Pursey, Statistics Canada/Statistique Canada

The route to the final datapoint • Le chemin vers la résultat final

The Unified Enterprise Survey (UES) is Statistics Canada's largest business survey designed to measure provincial and national economic variables across a wide range of industries. UES micro-data can be edited and changed at various stages of survey processing. Analyzing the flow of micro-data through these stages becomes difficult with multiple industries, multiple variables of interest and multiple stages of processing. The "Route to the Final Datapoint" is a visual tool that allows us to better understand the flow of data between the processing stages and to analyze the changes that are made. SAS and Visual Basic are used to create charts and graphs that are viewed in Excel. The frequency of changes, their direction (positive or negative) and the impact on the macro-data can all be viewed and analyzed in a link-based environment. As a self-evaluation tool for those processing the data or as a data quality assessment tool for analysts this tool provides a perspective to the travels of data from point A to point B not always understood while in the grind of producing final estimates, revealing tendencies previously unknown.

Le sondage unifié sur les entreprises (SUE), la plus grande enquête économiques de Statistiques Canada, est conçu pour mesurer des variables économiques provinciales et nationales à travers un éventail d'industries. Des micro-données du SUE peuvent être éditées et changées à diverses étapes du processus de sondage. L'analyse du flux des micro-données pour ces étapes devient difficile lorsqu'il y a plusieurs industries, plusieurs variables d'intérêt et des étapes multiples de processus. Le " chemin vers le résultat final " est un outil visuel qui nous permet de mieux comprendre le flux des données entre les étapes du processus et d'analyser les changements qui sont faits. SAS et Visual Basic sont utilisés pour créer des tableaux et des graphiques qui sont visualisés dans Excel. La fréquence des changements, leur direction (positive ou négative) et l'impact sur les macro-données peuvent tous être regardés et analysés dans un environnement relié. Comme outil d'auto évaluation pour ceux qui traitent les données ou comme un outil d'évaluation de la qualité des données pour les analystes, celui-ci fournit une perspective aux mouvements des données du point A au point B pas toujours compréhensibles, et dans le but de produire des estimations finales, il indique des tendances qui étaient jusqu'à maintenant inconnues.

Wednesday June 11 • Mercredi 11 juin 9:30 • 9h30 LSC 234

Colleen Clark, Mark Armstrong, Christian Thibualt, Statistics Canada/Statistique Canada

Measurement and innovation in the 2001Canadian Census coverage studies • Mesures et innovations dans les études de 2001 sur la couverture des recensements

The 2001 Canadian Census coverage studies introduced two automations in the largest study, the Reverse Record Check (RRC), that provided more and better data for classifying sampled persons as enumerated by the Census, missed, or not enumerable. Moving to Computer Assisted Telephone Interviewing gave interviewers a better tool to support tracing and allowed questionnaire additions that would not have been possible on paper. Considerable automation introduced in the clerical processing of checking Census documents for evidence of enumeration allowed staff to focus on the job of research rather than on the job of managing paper. Following an overview of the methodologies of the Reverse Record Check and the three overcoverage studies, this paper will look at RRC CATI and RRC processing, indicating how these innovations led to better data for classifying sampled persons. To evaluate the results of the coverage studies, the paper then presents comparisons of RRC estimates with estimates from other sources, and a discussion of error of closure. In conclusion, there is an introduction to the 2006 Census coverage studies.

Les études de 2001 sur la couverture des recensements canadiens ont présenté dans la plus grande d'entres elles, le Reverse Record Check (RRC, contrôle d'enregistrement inversé), deux automations qui ont fourni une plus grande quantité et de meilleures données pour classifier les personnes échantillonnées énumérées manquantes ou non énumérable par le recensement. Le changement vers les interviews téléphoniques assistés par ordinateur (ITAO) a donné aux interviewers un meilleur outil pour le rappel et a permis des additions aux questionnaires qui n'auraient pas été possibles sur papier. L'automation importante introduite dans les traitements pour examiner les documents de recensement pour assurer l'évidence d'énumération a permis au personnel de se concentrer sur le travail de la recherche plutôt que sur la gestion des papiers. Après un regard général sur les méthodologies du Reverse Record Check et des trois études de sur-couverture, cette présentation examine le traitement du RRC et des ITAO, en indiquant comment ces innovations ont mené à améliorer les données pour classifier les personnes échantillonnées. Pour évaluer les résultats des études de couverture, nous comparons les estimés du RRC avec des estimations provenant d'autres sources et une discussion sur l'erreur de fermeture. En conclusion, nous introduisons les études de 2006 sur la couverture du recensement.

Session/Séance 44 • Pierre Robillard Award Winner Lecture • Allocution du lauréat du prix Pierre Robillard

Wednesday June 11 • Mercredi 11 juin 8:30 - 9:15 LSC 240

Session/Séance 45 • Canadian Journal of Statistics Award Winner Lecture • Allocution du lauréat du prix de la Revue canadienne de statistique

Wednesday June 11 • Mercredi 11 juin 9:15 - 10:00 LSC 240

Session/Séance 46 • Case Study II - Neighbourhood Factors and Children: Hierarchical Linear Models and Small Area Statistics • Étude de cas II - Facteurs de voisinage et enfants: Modèles linéaires hiérarchiques et statistiques sur des petits domaines

Wednesday June 11 • Mercredi 11 juin 10:30 • 10h35 LSC 240

Patricia Whitridge, Statistics Canada/Statistique Canada

Introduction

Wednesday June 11 • Mercredi 11 juin 10:35 • 10h35 LSC 240

Jean-Francois Plante, Lawrence McCandless, Mike Danilov, Mushfiqur Rahman, University of British Columbia

Wednesday June 11 • Mercredi 11 juin 10:55 • 10h55 LSC 240

Sigfrido Iglesias-Gonzalez, Zheng Zheng, Xiaobin Yuan, University of Toronto

Wednesday June 11 • Mercredi 11 juin 11:15 • 11h15 LSC 240

Ouyang Jangman, Jady Lui, Hanqiuzi Wen, Sevina Chung, York University

Wednesday June 11 • Mercredi 11 juin 11:35 • 11h35 LSC 240

Xianlin Ma, Xu Wang, Longyang Wu, University of Waterloo

Wednesday June 11 • Mercredi 11 juin 11:55 • 11h55 LSC 240

Vaneeta Grover, McMasterUniversity

Session/Séance 47 • Nonparametric Analysis in Natural Resources Surveys • Analyse non paramétrique pour les enquêtes sur les ressources naturelles

Wednesday June 11 • Mercredi 11 juin 10:30 • 10h30 LSC 238

Noel Cadigan, Department of Fisheries and Oceans; Jiahua Chen, University of Waterloo

Improved kernel regression methods for inference about the population mean from sample surveys, with application to fishery surveys • Méthode de régression par le noyau amélioré pour l'inférence sur la moyenne de la population à partir d'un échantillon et applications en échantillonnage halieutique

An important source of information about the status of many fish stocks comes from annual research surveys, which involve fishing at randomly chosen sites. The survey catch at a particular site (t) is also a random measurement of local exploitable stock size, u(t). An important objective of the survey is to estimate the average exploitable abundance at all sites; that is, the average of $u(t_1), ..., u(t_N)$. This provides an index of total stock size, and such indices often form the basis for fishery management decisions such as setting the total allowable catch in the next year. We consider two problems in this talk. The first problem involves smoothing methods to estimate $u(t)$, and the second problem involves methods to estimate the average of $u(t_1), ..., u(t_N)$. We develop a new kernel regression smoother that has some improved statistical properties compared to standard smoothers, and is more suitable for making inferences about average exploitable abundance.

Une source importante d'informations sur le statut de beaucoup de stocks de poissons provient d'enquêtes annuelles, qui consiste à pêcher sur des sites choisis aléatoirement. L'échantillon prélevé à un emplacement particulier (t) est également une mesure aléatoire de la taille exploitable sur ce site, u(t). Un objectif important du sondage est d'estimer la quantité exploitable moyenne dans tous les emplacements; c'est-à-dire, la moyenne de $u(t_1), ..., u(t_N)$. *Ceci fournit un index de la taille courante totale, et de tels index forment souvent la base des décisions de gestion des pêches, telles que déterminer le quota pour*

l'année suivante. Nous considérons deux problèmes dans cette présentation. Le premier implique d'utiliser des méthodes de lissage pour estimer $u(t)$, et le deuxième problème implique des méthodes pour estimer la moyenne de $u(t_1), ..., u(t_N)$. Nous développons une nouvelle méthode de lissage par régression par le noyau qui a quelques propriétés statistiques améliorées comparées aux méthodes de lissage standards. La méthode est aussi plus appropriées pour faire de l'inférences sur la quantité exploitable moyenne.

Wednesday June 11 • Mercredi 11 juin 11:00 • 11h00 LSC 238

Jay Breidt, Colorado State University; Jean D. Opsomer, Iowa State University

Nonparametric model-assisted estimation for surveys of natural resources • Estimations non paramétriques assistées par modèles pour des sondages sur les ressources naturelles

A classical estimation strategy in surveys is to use parametric regression techniques to assist in estimation of population parameters. In natural resource surveys, covariate information used in the regression is often readily available from remotely-sensed imagery or geographic information system coverages. Nonparametric model-assisted regression estimation offers an alternative to parametric approaches. Nonparametric estimators based on linear smoothers have most of the desirable design and model properties of the generalized regression estimator, but the assumptions on the superpopulation model are much weaker. The nonparametric methodology is compared to parametric procedures analytically and via simulation, and applied to surveys of natural resources.

Une stratégie classique d'estimations dans les sondages est d'utiliser des techniques de régression paramétrique pour aider à l'estimation des paramètres de la population. Dans les sondages sur les ressources naturelles, l'information des covariables utilisée dans la régression est souvent directement fournie par l'imagerie à distance ou par les systèmes de couverture d'informations géographiques. L'estimation par la régression non paramétriques assistée par modèle offre une alternative aux approches paramétriques. Les estimateurs non paramétriques basés sur les méthodes de lissage linéaires ont pour la plupart des propriétés souhaitables du design et du modèle de l'estimateur généralisé de régression, mais les hypothèses sur le modèle de la super-population sont beaucoup plus faibles. La méthodologie non paramétrique est comparée aux procédures paramétriques analytiquement et par l'intermédiaire de simulations. Nous appliquons la méthode aux sondages sur les ressources naturelles.

Wednesday June 11 • Mercredi 11 juin 11:30 • 11h30 LSC 238

Changbao Wu, Jiahua Chen, Mary E. Thompson, University of Waterloo

Estimation of fish abundance indices based on scientific research trawl surveys • Estimation de l'indice d'abondance de poissons basée sur des sondages de chalut de recherches scientifiques

The fish abundance index over an ocean region is defined here to be the integral of catch per unit effort (CPUE), approximated by the sum of CPUE over grid squares. When trawl surveys are done within grid squares selected according to a probability sampling design, model assisted methods for estimating abundance can be a useful alternative to purely design based methods using, for example, the Horvitz-Thompson (HT) estimator. In this talk we build semiparametric and nonparametric models based on scientific research trawl survey data and develop empirical likelihood (EL) methods for estimating the fish abundance indices. The methods are applied to grid surveys of the Grand Bank region

carried out annually by Fishery Products International from 1996 through 2002. The HT and EL methods produce similar point estimates of abundance for three species of fish. The EL estimator under local linear smoothing is associated with smaller estimated standard errors.

L'index d'abondance de poissons sur une région océanique est défini ici comme étant l'intégrale du nombre de prises par unité d'effort (PPUE), estimé par la somme des PPUE sur un quadrillé. Quand les sondages de chalut sont faits selon un plan d'échantillonnage de probabilité sur un quadrillage, pour estimer l'abondance, les méthodes assistées par modèles peuvent être une alternative utile aux méthodes basées uniquement sur le design qui utilisent, par exemple, l'estimateur de Horvitz-Thompson (HT). Dans cette présentation, nous construisons des modèles semi-paramétriques et non paramétriques basés sur des données de sondages de chalut de recherches scientifiques et nous développons des méthodes de vraisemblance empiriques (EL) pour estimer les index d'abondance de poissons. Les méthodes sont appliquées aux sondages sur un quadrillé de la région de la Grande Banque effectuée annuellement par Fishery Products International de 1996 à 2002. Les méthodes HT et EL produisent des estimateurs ponctuels similaire pour trois espèces de poissons, mais l'estimateur EL sous un lissage linéaire local est associé à de plus petits écarts types estimées.

Session/Séance 48 • Statistical Inference II: Inference Problems with Missing Data or Measurement Errors • Inférence statistique: Problèmes d'inférence avec des données manquantes et des erreurs de mesure

Wednesday June 11 • Mercredi 11 juin 10:30 • 10h30 LSC 242

Don McLeish, University of Waterloo; C.A. Struthers, St Jeromes University and University of Waterloo

Estimation of regression parameters in missing data problems • L'estimation des paramètres de régression dans des problèmes avec données manquantes

It is common in applications of regression for one or more covariates to be unobserved for some of the experimental subjects, either by design (for example they are expensive or difficult to obtain) or by some random censoring mechanism. For example in a two-stage experiment, a preliminary analysis may determine for which observations we should obtain complete data. Specifically, suppose Y is a response variable, possibly multivariate, with a density function $f(y|x, v; b)$ conditional on the covariates (x, v) where x and v are vectors and b is a vector of unknown parameters. We discuss the estimation of the parameter b when data on the covariate v are available for all observations but data on the covariate x are missing for some. We assume that x is "missing at random" i.e. that the probability that x is missing depends only on the fully observed quantities (y, v). Variations on this problem have been considered by a large number of authors, including among many others Chatterjee et al., 2003; Lawless et al., 1999; Reilly and Pepe, 1995; Carrol and Wand, 1991; and Pepe and Fleming, 1991, Robins et al., 1994, 1995. We suggest and compare several estimators and algorithms for this problem when the data is discrete or continuous.

Il est commun dans les applications de la régression qu'une ou plusieurs covariables soient non observées pour certains sujets, soit par le design (par exemple elles sont difficiles ou dispendieuses à obtenir) ou par un certain mécanisme de censure aléatoire. Par exemple, dans une expérience à deux étages, une analyse préliminaire peut déterminer pour quelles

observations nous devons obtenir des données complètes. Spécifiquement, supposons que Y est une variable réponse, possiblement multivariée, avec une fonction de densité $f(y|x, v; b)$ conditionnellement sur les covariables (x, v) où x et v sont des vecteurs et où b est le vecteur des paramètres inconnus. Nous discutons l'estimation du paramètre b lorsque les données sur la covariable v sont disponibles pour toutes les observations mais certaines données pour x sont absentes. Nous supposons que x est "manquante dû au hasard" c'est-à-dire que la probabilité que x soit manquante dépend seulement des données complètes (y, v). Des variations de ce problème ont été considérées par un grand nombre d'auteurs, incluant, Chatterjee et al, 2003; Lawless et al, 1999; Reilly et Pepe, 1995; Carrol et Wand, 1991; Pepe et Fleming, 1991 et Robins et al, 1994, 1995. Nous suggérons et comparons plusieurs estimateurs et algorithmes pour ce problème avec des données discrètes ou continues.

Wednesday June 11 • Mercredi 11 juin 11:00 • 11h00 LSC 242

Bruce Turnbull, Cornell University; Wenxin Jiang, Northwestern University

The indirect method for repeated events regression analysis with covariate measurement error • La méthode indirecte pour l'analyse de régression d'événements récurrents avec erreur de mesure sur les covariables

We present use of the so-called method of indirect inference to accommodate measurement error in the covariates when analyzing recurrent event data with parametric or semiparametric regression models. The idea is borrowed from the econometric literature but recast in a likelihood-flavoured approach. We illustrate the technique with data from a randomized clinical trial for the prevention of recurrent kin tumours, where important prognostic blood biochemical levels can only be measured with error.

Nous présentons la méthode d'inférence indirecte adaptée aux erreurs de mesure sur les covariables pour l'analyse d'événements récurrents avec des modèles de régression paramétriques ou semi-paramétriques. L'idée est empruntée de l'économétrie, mais remaniée avec une approche basée sur la vraissemblance. Nous illustrons la technique avec des données tirées d'un essai clinique randomisé traitant de la prévention des tumeurs répétitives sur la peau, situation où des pronostiques importants sont basés sur le niveau d'une substance biochimique dans le sang qui peut seulement être mesuré avec erreur.

Wednesday June 11 • Mercredi 11 juin 11:30 • 11h30 LSC 242

Paul Gustafson, University of British Columbia

Bayesian adjustment for mismeasured explanatory variables • L'ajustement bayésien pour des variables explicatives avec erreur de mesure

The first part of the talk will review the basic framework for Bayesian analysis in problems involving mismeasured explanatory variables (i.e. either continuous variables subject to measurement error, or discrete variables subject to misclassification). An example will be given, and strengths and weaknesses relative to non-Bayesian methods will be discussed. Then, some curious findings relating to identifiability in such models will be presented. These relate to the pragmatic concern that sometimes investigators may lack sufficient knowledge about the mismeasurement process to yield an identifiable statistical model.

Dans la première partie de la présentation, nous passerons en revue les bases du cadre bayésien pour l'analyse de problèmes impliquant des variables explicatives avec erreur de mesure (c.-à-d. variables continues sujette à l'erreur de mesure, ou variables discrètes sujette à une erreur de classification). Un exemple est donné et nous discutons des

avantages et des désavantages par rapport aux méthodes non bayésiennes. Ensuite, nous présentons quelques résultats surprenants concernant l'identification dans de tels modèles. Ces derniers se rapportent au souci pragmatique que les investigateurs peuvent parfois manquer de connaissances suffisantes sur le processus d'erreur de mesure pour donner un modèle statistique identifiable.

Session/Séance 49 • Biostatistics Contributed Session II: Epidemiological and Clinical Studies • Biostatistique II: Études épidémiologiques et cliniques

Wednesday June 11 • Mercredi 11 juin 10:30 • 10h30 LSC 338

Gordon Fick, University of Calgary

Modelling the odds of disease using data from case-control studies • Modéliser le risque de maladie en utilisant des données d'études cas-témoins

When analyzing data from case-control studies, the investigator has to decide whether to model the odds of exposure or whether to model the odds of disease. In certain specific settings, this decision has no impact on the assessment of the primary objective of such studies. In a majority of settings, however, this decision can have a major impact on the analysis and interpretation of the results. Standard likelihood theory provides support for modelling the odds of exposure. It would appear that there is very little theoretical justification for modelling the odds of disease.

Lorsque nous analysons des données d'études cas-témoins, l'investigateur doit décider s'il modélise le risque d'exposition ou s'il modélise le risque de maladie. Dans certaines situations spécifiques, cette décision n'a aucun impact sur l'évaluation de l'objectif principal de l'étude. Cependant, dans plusieurs situations cette décision peut avoir un impact important pour l'analyse et l'interprétation des résultats. La théorie habituelle de vraisemblance fournit un support pour modéliser le risque d'exposition. Il semble y avoir peu de justifications théoriques pour modéliser le risque de maladie.

Wednesday June 11 • Mercredi 11 juin 10:45 • 10h45 LSC 338

Cyr Emile M'lan, Hospital for Sick Children, Toronto

Bayesian sample size calculation for case-control studies • Méthodes bayésiennes de calcul de taille d'échantillon pour les études cas-témoins

Bayesian sample size determination for case-control studies. One of the most important statistical issues at the planning stage of a case-control study is the choice of sample size. For example, one might wish to select a sample size that ensures sufficient accuracy in estimating the odds ratio. Sample size determination for the odds ratio has been investigated from a frequentist viewpoint. While most of the proposed methods have been based on power, it is well known that high power does not necessarily guarantee accurate estimation of important parameters. Therefore, if one chooses to analyse a study by interval estimation rather than p-values, the design of the study should reflect this choice. Two previous frequentist approaches (O'Neill, 1984, Kupper and Halfner, 1990) were based on ensuring sufficiently small confidence interval widths for a given coverage. Recently, however, numerous papers have addressed the advantage of using a Bayesian approach to the estimation of an odds ratio from case-control studies based on the posterior distribution of the odds ratio by using hpd intervals (Hora and Kelly; 1983, Marshall 1988; Zelen and

Parker, 1986; Hashemi et al. ,1997). In addition, there is an increasing literature on the advantages of Bayesian sample size determination, which better incorporates prior information into the calculations, and more fully accounts for the uncertainty of the eventual data which will be collected. In this talk, we show how sample size determination for estimating the odds ratio can be addressed within the Bayesian paradigm. Although we have thoroughly investigated a large number of Bayesian sample size criteria, including the development of novel criteria, here we focus on the average length criterion (ALC). Basically, this method proposes that the sample size be selected that guarantees a pre-specified length for a marginal posterior credible interval of predetermined coverage, averaged over the predictive distribution of the data. The solution, while easy to define, is technically challenging to carry out in practice. We discuss three different methods for finding the optimal sample size, including exact, approximate, and Monte Carlo. We compare the sample sizes derived from this criterion to those from frequentist power and confidence interval methods.

Le choix de la taille d'échantillon est l'un des points les plus sensibles de la planification d'une étude cas-témoins; de ce choix dépend notamment la précision avec laquelle on pourra estimer le rapport de cotes. Deux approches classiques antérieures (O'Neill 1984; Satten & Kupper 1990) ont été basées sur des critères garantissant des intervalles de confiance suffisamment courts pour une couverture donnée. Cependant, de nombreux auteurs ont récemment fait valoir les avantages d'une approche bayésienne, qui permet a la fois d'incorporer une information à priori et de mieux prendre en compte l'incertitude concernant la variation inhérente aux données dans l'estimation du rapport de cotes dans les études de cas-témoins. Le conférentier présentera un survol de ces travaux et fera état de ses propres recherches dans le domaine. Il s'attardera plus particulièrement aux résultants portant sur le critère de la longueur moyenne, qui consiste à choisir une taille d'échantillon permettant de garantir une longueur moyenne préspecifiée pour un intervalle de crédibilité à posteriori de couverture préfixée, la moyenne étant faite par rapport a la distribution de prévision des données. Comme il le fera valoir, la solution est facile à concevoir mais ne peut être déterminée en pratique qu'au moyen de méthodes numériques exactes, approximatives ou de Monte-Carlo. Il présentera en outre une étude visant à comparer les tailles d'échantillon obtenues par ce critère à celle déduites des méthodes de puissance et d'intervalle de confiance dans l'approche classique.

Wednesday June 11 • Mercredi 11 juin 11:00 • 11h00 LSC 338

Lehana Thabane, McMaster University

A Bayesian look at the number needed to treat • Une vue bayésienne pour le nombre requis pour traitement (NNT)

In this talk, I present a Bayesian approach to estimation of the number needed to treat (NNT). The use of NNT as a measure of clinical benefit is now becoming commonplace. Various methods of estimation have been proposed, but none of them seem to provide entirely good estimates. Much still remains to be done to understand the statistical properties of NNT. Here, I derive the posterior distribution of NNT and use simulations to investigate the general behaviour of the distribution. The posterior mode of the distribution is proposed as a point estimate and results are compared with the conventional method of estimation of NNT done by inversion.

Dans cette présentation, nous présentons une approche bayésienne pour l'estimation du nombre requis pour traiter (NNT). L'utilisation du NNT comme mesure de gain clinique

est maintenant banale. Diverses méthodes d'estimations ont été proposées, mais aucune d'elles ne semble fournir des estimations complètement adéquates. Il reste beaucoup à faire pour comprendre les propriétés statistiques du NNT. Ici, nous dérivons la distribution a posteriori du NNT et nous employons des simulations pour étudier le comportement général de la distribution. Nous proposons le mode a posteriori de la distribution comme une estimation ponctuelle et nous comparons les résultats à la méthode conventionnelle d'estimation par inversion du NNT.

Wednesday June 11 • Mercredi 11 juin 11:15 • 11h15 LSC 338

Nandini Dendukuri, J. Hanley, R. Platt, M.-H. Mayrand, McGill University

Design and data-analysis options for clinical trials of assisted reproductive technologies • Alternatives de devis et de méthodes d'analyse pour études cliniques de technologies de reproduction assistée

There is a considerable amount of literature pointing to evidence of heterogeneity of fecundability, or probability of becoming pregnant, across women. Further, there is evidence that ignoring this heterogeneity results in biased estimates of a new intervention, irrespective of the study design. Two commonly used designs for randomized trials of fertility interventions are: 1) an alternating-sequence design, where women who are unsuccessful after each treatment cross over to the opposite treatment and, 2) a parallel design, where women remain in the group they are randomly assigned to until they become pregnant. We will discuss estimation and inference for three different approaches to account for heterogeneity of fecundability in such designs: 1) modeling the first C moments of the fecundability in a study with C cycles, 2) modeling the marginal probability of success across women, 3) modeling the couple-specific probability of success. The methods will be illustrated by application to data from a clinical trial comparing pregnancy rates following insemination with frozen rather than fresh semen.

La littérature scientifique suggère que la fécondabilité (probabilité de devenir enceinte dans un cycle donné) varie d'une femme à l'autre. De plus, il semble que si cette hétérogénéité est ignorée, les évaluations de nouvelles technologies produisent des estimations biaisées, et ce, peu importe le devis utilisé. Il existe deux devis qui sont fréquemment utilisés pour les études randomisées de traitement de fertilité: 1) le plan d'étude " séquence alternante " dans lequel les femmes qui ne sont pas enceintes après un cycle reçoivent l'autre traitement au cycle suivant, et 2) le plan d'étude " parallèle " dans lequel les femmes demeurent dans le groupe de traitement auquel elles sont assignées en début d'étude jusqu'à ce qu'elles deviennent enceintes. Nous discuterons de l'impact sur le processus d'estimation et d'inférence en comparant trois approches qui permettent de tenir compte de l'hétérogénéité de la fécondabilité dans ces 2 devis de recherche: 1) en modélisant les premiers moments C de fécondabilité dans un étude avec C cycles, 2) en modélisant la probabilité marginale de succès pour le groupe, 3) en modélisant la probabilité de succès propre à chaque couple. Nous illustrerons ces différentes méthodes en les appliquant aux données d'une étude clinique.

Wednesday June 11 • Mercredi 11 juin 11:30 • 11h30 LSC 338

Nicholas Barrowman, Manchun Fang, Margaret Sampson, David Moher, Chalmers Research Group

When do meta-analyses need to be updated? • À quel moment les méta-analyses doivent-elles être mises à jour?

Particularly in the health care field, meta-analysis has become an indispensable tool for summarizing the totality of evidence on a given question. Unfortunately, if the evidence is not up to date, meta-analytic results may be incomplete or even misleading. Some authors have proposed continuous updating of meta-analyses to incorporate newly available evidence. However, in addition to raising statistical issues of multiple testing, this would require continuous monitoring and evaluation of the research literature, with considerable resource implications. An alternative approach is to somehow determine when the newly available evidence is substantial enough to warrant an update. Here we propose a quantitative approach for predicting when meta-analyses need to be updated and a related diagnostic test for determining when meta-analyses are out of date. Simulations are used to assess the performance of this approach and several published meta-analyses are used for illustration. Strengths and limitations of the approach are discussed and possible future areas of research are considered.

Particulièrement dans le domaine de santé, la méta-analyse est devenue un outil indispensable pour récapituler la totalité des faits sur une question particulière. Malheureusement, si l'information n'est pas à jour, les résultats méta-analytiques peuvent être incomplets ou même trompeurs. Quelques auteurs ont proposé une mise à jour continue des méta-analyses pour incorporer les nouveaux faits disponibles. Cependant, en plus de soulever des problèmes statistiques reliés aux tests multiples, ceci exigerait le contrôle et l'évaluation continu de la littérature de recherche, avec des implications considérables en terme de ressources. Une approche alternative est de déterminer d'une manière ou d'une autre quand est-ce que les faits nouvellement disponibles sont assez substantiels pour justifier une mise à jour. À quel moment les méta-analyses doivent-elles être mises à jour?Ici nous proposons une approche quantitative pour prédire quand est-ce que les méta-analyses doivent être mises à jour et un test de diagnostic pour déterminer si les méta-analyses sont désuètes. Des simulations sont utilisées pour évaluer la performance de cette approche et plusieurs méta-analyses publiées sont utilisées pour illustrer l'approche. Nous discutons des forces et ses limitations de l'approche et nous considérons les avenues de recherche futures possibles.

Wednesday June 11 • Mercredi 11 juin 11:45 • 11h45 LSC 338

Juan Pablo Lewinger, Shelley B. Bull, University of Toronto

Better tests to find susceptibility genes for complex diseases via randomization • Meilleurs tests pour trouver les gènes de susceptibilité aux maladies complexes par randomisation

Testing that a marker locus is linked or is in linkage disequilibrium (associated) with a susceptibility gene is key in the process of gene mapping. For multifactorial complex diseases the validity of tests of linkage and association is of concern, since spurious findings could potentially arise from population stratification or other inadequately modeled features. The Transmission Disequilibrium Test (TDT) was designed to overcome this shortcoming and is widely used among gene-hunters. However, it requires complete parental genotypes and sacrifices power by not using all the available phenotypic information. Subsequent proposals that allow for incomplete parental genotypes such as the Sibship Disequilibrium Test (SDT), have been shown to have much poorer power. We introduce new tests of linkage and association that are both distribution free and powerful, and allow for any pattern of missing parental genotypes. The distribution free property is achieved by randomization of the genes present in the parents, allowing complete freedom in the choice

of test statistic. We show how the latter can be selected to obtain asymptotically optimal power within any given parametric model of linkage and association, and how it lends itself to simple asymptotic approximations of power and p-values. Additionally, we show how to efficiently estimate exact power and p-values via Monte Carlo importance sampling. Simulation over a range of models an sample sizes show that the proposed tests have near optimal power and perform substantially better than their competitors.

Tester qu'un marqueur sur un locus est lié ou est dans le déséquilibre de lien (associé) avec un gène prédisposé est primordial pour le processus de cartographie génétique. Pour les maladies complexes multifactorielles, la validité des tests de liens et d'association est importante, puisque les faux résultats peuvent potentiellement résulter de la stratification de la population ou d'autres caractéristiques inadéquatement modelées. Le test de déséquilibre de transmission (TDT) est conçu pour surmonter cette imperfection et est largement utilisé par les chasseurs de gènes. Cependant, il exige les génotypes parentaux complets et sacrifie la puissance en n'utilisant pas toute l'information phénotypique disponible. Il est montré que des propositions subséquentes qui tiennent compte des génotypes parentaux incomplets tels que le test de déséquilibre de Sibship (TDS), ont une puissance beaucoup plus faible. Nous présentons des nouveaux tests de liens et d'association qui sont indépendants de la distribution et puissants, et qui tiennent compte de n'importe quel patron de génotypes parentaux manquants. La propriété d'indépendance de la distribution est réalisée par la randomisation des gènes présents chez les parents, permettant une liberté complète dans le choix de la statistique de test. Nous montrons comment celle-ci peut être choisie pour obtenir une puissance asymptotiquement optimale dans n'importe quel modèle paramétrique de lien et d'association, et comment elle se prête aux approximations asymptotiques simples de la puissance et des seuils expérimentaux. De plus, nous montrons comment estimer efficacement la puissance exacte et les seuils expérimentaux par l'intermédiaire de l'échantillonnage d'importance de Monte Carlo. La simulation sur une gamme de modèles et de taille d'échantillons montre que les tests proposés ont une puissance quasi optimale et performent substantiellement mieux que leurs concurrents.

Session/Séance 50 • Design and Analysis of Experiments • Planification et analyse d'expériences

Wednesday June 11 • Mercredi 11 juin 10:30 • 10h30 LSC 332

Arden Miller, University of Auckland

The analysis of unreplicated factorial experiments using all possible comparisons • L'analyse d'expériences factorielles non-répliquées en utilisant toutes les comparaisons possibles

A new procedure for analyzing small unreplicated factorial experiments is presented. This procedure is called all possible comparisons (APC) since it is based on using likelihood ratio tests to compare all pairs of competing models. An easy method of implementing the procedure is presented and then demonstrated on a real set of data. Results of a simulation study that compares the performance of APC to Lenth's method are presented.

Une nouvelle procédure pour analyser des petites expériences factorielles non-répliquées est présentée. Cette procédure s'appelle " toutes les comparaisons possibles " (TCP) puisqu'elle est basée sur le test de rapport de vraisemblance pour comparer toutes les paires de modèles en concurrence. Une méthode facile pour implémenter la procédure est présentée et nous montrons ensuite son application sur un vrai jeu de données. Nous présentons aussi les

résultats d'une étude de simulation qui compare la performance du TCP à la méthode de Lenth.

Wednesday June 11 • Mercredi 11 juin 10:45 • 10h45 LSC 332

Glen Takahara, Hwashin H. Shin, Queen's University; Duncan J. Murdoch, University of Western Ontario

Optimal designs for orientation regression experiments • Designs optimaux pour des expériences de régression sur l'orientation

For calibrating electromagnetic motion tracking equipment for use in human motion studies in biomechanics, it is a non-trivial task to place a sensor in a known orientation relative to the electromagnetic source. At a given design orientation U_i (a 3 by 3 rotation matrix), we observe the orientation $V_i = A_1^t E_i U_1 A_2$, where A_1 and A_2 are unknown matrix parameters and E_i is a random distortion. The matrices A_1 and A_2 represent systematic distortion from the true orientations U_i. For n observations, we consider optimal designs $U_1, ..., U_n$ to minimize the prediction error variance. We show that the design criterion is $U_1 + ... + U_n = 0$ and show how to construct such designs. For $n = 3, ..., 9$, we compare the theoretical gain of such designs over a non-optimal 9-point design with simulation.

Pour calibrer des équipements de détection du mouvement électromagnétique utilisés pour les études sur le mouvement humain en biomécanique, placer une sonde dans une orientation connue par rapport à la source électromagnétique n'est pas une tâche facile. Pour une orientation donnée U_i (une matrice de rotations 3 par 3), nous observons l'orientation $V_i = A_1^t E_i U_1 A_2$, où A_1 et A_2 sont des matrices inconnues de paramètres et E_i est une distorsion aléatoire. Les matrices A_1 et A_2 représentent la distorsion systématique des vraies orientations U_i. Pour n observations, nous considérons les designs optimaux $U_1, ..., U_n$ pour minimiser la variance de l'erreur de prévision. Nous prouvons que le critère de design est $U_1 + ... + U_n = 0$ et montrons comment construire de tels design. Pour n $= 3, ..., 9$, nous comparons le gain théorique de tels designs par rapport aux designs non optimaux à 9-points à l'aide de simulations.

Wednesday June 11 • Mercredi 11 juin 11:00 • 11h00 LSC 332

Xin Gao, Mayer Alvo, University of Ottawa

Nonparametric tests for interaction in unbalanced design with application in QTL analysis • Des tests non-paramétriques pour l'interaction dans le plan déséquilibré et applications aux analyses de QTL

To fully dissect complex trait, it is desirable to have methods able to test gene-gene interaction for genetic data. The genetic analysis imposes a hypothesis testing problem of interaction for a two-way layout design with unequal replicates.Nonparametric hypothesis testing for unbalanced designs has been a question of great interest but remained largely unsolved in the last three decades. We establish the asymptotic normality of correlated linear rank statistics under mild conditions. A notion of weighted rank is introduced and a new methodology is devised to solve the problem of unbalanced designs. The limiting distributions under Pitman alternatives and asymptotic relative efficiencies are studied. Consistent estimators are provided for the limiting variance-covariance matrix of an arbitrary composite linear rank statistics.

Dans le but d'étudier les caractéristiques complexes de gènes, il est désirable de disposer des méthodes pouvant tester l'effet de l'interaction entre les gènes pour les données génétiques.

L'analyse génétique mène à un problème de test pour mesurer l'effet de l'interaction lorsque les cellules ont des tailles différentes. Les tests non-paramétriques pour le plan déséquilibré demeurent une question très intéressante depuis trente ans. On établi la normalité asymptotique de statistiques corrélées linéaires de rang sous des conditions assez faibles. On introduit la définition de rang pondéré ainsi qu'une nouvelle méthodologie afin de résoudre le problème de design déséquilibré. On étudie les distributions asymptotiques ainsi que l'efficacité relative asymptotique dans le sens dde Pitman. Notre étude fourni des estimateurs consistents pour la matrice de variance-covariance.

Wednesday June 11 • Mercredi 11 juin 11:15 • 11h15 LSC 332

Paul Cabilio, Acadia University; Mayer Alvo, University of Ottawa

General scores statistics on ranks in the analysis of unbalanced designs • Statistiques de scores générales basées sur les rangs pour l'analyse de plans déséquilibrés

We consider the situation of incomplete rankings in which each of $1 \leq i \leq n$ judges is presented with $2 \leq k_{\{i\}} \leq t$ objects which are ranked independently by the judges. We wish to test the hypothesis that each judge, when presented with the specified $k_{\{i\}}$ objects, picks the ranking at random from the space of $k_{\{i\}}!$ permutations of $(1, 2, ..., k_{\{i\}})$. Our statistic is a generalization of the Friedman test in which the ranks assigned by each judge are replaced by real valued functions of the ranks,$a(j, k_{\{i\}})$, $1 \leq j <= k_{\{i\}} \leq t$. We define a measure of pairwise similarity between complete rankings based on such functions, and use averages of such similarities to construct measures of the level of concordance of the judges' rankings. In the complete ranking case, the resulting statistics coincide with those defined by Hájek and ˜Jidák (1967), and Sen (1968). We extend these measures of similarity in the complete case to the situation of incomplete rankings, and derive a statistic in this more general setting and investigate its properties.

Nous considérons la situation d'un rangement incomplet dans lequel on présente à chacun des $1 \leq i \leq n$ juges les $2 \leq k_{\{i\}} \leq t$ objets qui sont rangés indépendamment par les juges. Nous souhaitons tester l'hypothèse que chaque juge, une fois qu'on lui a présenté les $k_{\{i\}}$ objets spécifiés, sélectionne le rang au hasard dans l'espace des $k_{\{i\}}!$ permutations de $(1, 2, ..., k_{\{i\}})$. Notre statistique est une généralisation du test de Friedman dans lequel les rangements affectés par chaque juge sont remplacés par des fonctions à valeur réelle des rangements, $a(j, k_{\{i\}})$, $1 \leq j \leq k_{\{i\}} \leq t$. Nous définissons une mesure de similitudes par paires entre les rangements complets basés sur de telles fonctions, et nous utilisons les moyennes de ces similitudes pour construire des mesures du niveau de concordance des rangements des juges. Dans le cas d'un rangement complet, les statistiques résultantes coïncident avec celles définies par Hájek et Jidák (1967), et Sen (1968). Nous prolongeons ces mesures de similitude dans le cas complet à la situation des rangements incomplets, et nous dérivons une statistique pour ce cadre plus général et étudions ses propriétés.

Wednesday June 11 • Mercredi 11 juin 11:30 • 11h30 LSC 332

Anatoly Naumov, Novosibirsk State Technical University, Novosibirsk, Russia

From optimal design to optimal control of Experiments • Du plan optimal au contrôle optimal d'expérience

This paper is devoted to research of the classical optimum designs of experiments from the point of view of a posterior dispersion of an estimation of regression model. The execution of the equality in equivalence theorem is considered and investigated. The main result is

so. A distance between a prior dispersion function and a posterior one can be very essential and the equality in Kiefer's theorem can be not true for a posterior data. The difference of the designs of experiments constructed according to prior and posterior information can be very essential. According to this circumstance, it is offered to use a posterior criterion at construction of the designs (the strategy) of experiments. The algorithms, appropriate to this criterion, carry consecutive (iterative) character, and the designs are realizations of some random variable. In such case loses sense to speak about an optimality of the designs, as it is done in the classical theory of optimum design of experiments. It is offered to use the term of optimal control of experiments. The algorithms are considered for synthesis of the optimal control with the fixed spectrum. By modeling and comparative analysis is carried out their convergence. Many examples are given where an optimal control and optimal classical design are compared.

Cette présentation est consacrée à la recherche des plans optimaux d'expériences classiques du point de vue d'une dispersion a posteriori d'une estimation d'un modèle de régression. La réalisation de l'égalité dans le théorème d'équivalence est considérée et étudiée. Le résultat principal est le suivant. Une distance entre une fonction de dispersion a priori et une a posteriori peut être essentielle et l'égalité dans le théorème de Kiefer peut être fausse pour des données a posteriori. La différence entre les plans d'expériences construits selon l'information a priori et a posteriori peut être essentielle. Selon cette circonstance, nous proposons d'utiliser un critère a posteriori pour la construction des plans (la stratégie) d'expériences. Les algorithmes appropriés pour ce critère portent le caractère (itératif) consécutif et les plans sont des réalisations d'une certaine variable aléatoire. Dans un tel cas, il est inutile de discuter de l'optimalité des plans, car cela est fait dans la théorie classique des plans d'expériences optimaux. Nous proposons d'utiliser le terme de contrôle optimal d'expérience. Les algorithmes sont considérés pour la synthèse du contrôle optimal avec un spectre fixe. Par modélisation, nous effectuons une analyse comparative de leur convergence respective. Beaucoup d'exemples sont donnés où un contrôle optimal et un plan classique optimal sont comparés.

Wednesday June 11 • Mercredi 11 juin 11:45 • 11h45 LSC 332

Vitaly Senitch, Anatoly Naumov, Novosibirsk State Technical University, Novosibirsk, Russia

Sequential Optimal Control of Experiments • Contrôle séquentiel optimal des expériences

The difference of the designs of experiments constructed according to prior and posterior information can be very essential. According to this circumstance, it is offered to use a posterior criterion at construction of the sequential strategy of experiments. The algorithms, appropriate to this criterion, carry iterative character. It is offered to use the term of optimal control of experiments. The algorithms are considered for synthesis of the optimal control. The problem of constructions of the optimal control of experiments on the basis of new algorithms is considered. The classical approach to optimum designing of experiments optimal control cannot be constructed as a prior one. They can be constructed only on the recurrent basis, simultaneously with realization of experiments (nature or modeling). Such feature means last, that, first, they cannot be tabulated how it is done for the classical approach and, secondly, a set of optimal strategies of experiments at the fixed assumptions of the model (and properties of external environment) is an infinite set. Such strategies depend on results of experimenting and consequently are realizations of some random variable. Two lgorithms of construction of the D-, A- and E-optimal strategies (The algorithms

with a fixed spectrum (FS), The algorithm of uniform (parallel) experimenting AUE(FS)) are considered too.

La différence des plans d'expériences construits selon l'information a priori et a posteriori peut être très importante. Dans ce cas, nous proposons d'utiliser un critère a posteriori pour construire la stratégie séquentielle des expériences. Les algorithmes appropriés à ce critère portent un caractère itératif. Nous proposons d'utiliser le terme de contrôle optimal des expériences. Les algorithmes sont considérés pour la synthèse du contrôle optimal. Le problème de la construction du contrôle optimal des expériences sur la base de nouveaux algorithmes est considéré. L'approche classique pour un plan d'expérience optimum du contrôle optimal ne peut pas être construite comme une approche a priori. Elles peuvent être construites seulement sur la base récurrente, simultanément avec la réalisation des expériences (naturelles ou modélisées). Une telle caractéristique signifie qu'elles ne peut pas être tabulé comme cela est fait pour l'approche classique et deuxièmement, un ensemble de stratégies optimales d'expériences sous l'hypothèse fixe du modèle (et les propriétés de l'environnement externe) est un ensemble infini. De telles stratégies dépendent des résultats de l'expérimentation et sont par conséquent des réalisations d'une certaine variable aléatoire. Deux algorithmes de construction des stratégies D, A et E-optimales (les algorithmes avec un spectre fixe (SF), l'algorithme d'expérimentation uniforme (parallèle) AEU(FS)) sont aussi considérés.

Session/Séance 51 • Robust Methods II and Statistical Education • Méthodes robustes II et éducation statistique

Wednesday June 11 • Mercredi 11 juin 10:30 • 10h30 LSC 234

Ivan Mizera, University of Alberta; Benoit Laine, Université libre de Bruxelles

Autoregression depth • La profondeur autorégressive

We consider an analog of the halfspace (Tukey) depth in time series, in the vein of general methodology for extending depth notions to various statistical models. In autoregression models, the result parallels the regression depth for standard regression models. We study its possible statistical applications, in particular its potential for quantile estimation and maximum autoregression depth estimators, which happen to extend the ad hoc proposal for AR(1) models that already appeared in the literature. We explore, both empirically via simulations and examples and theoretically via asymptotics related to that for general maximum depth estimators, how autoregression depth-based procedures compare with similar median-oriented proposals - for instance L1- and sign-based procedures. A special attention is paid to the robustness properties and the behavior for heavy-tailed distributions.

Nous considérons un analogue de la profondeur halfspace (Tukey) dans les séries chronologiques, dans la veine d'une méthodologie générale pour prolonger les notions de profondeur à divers modèles statistiques. Dans les modèles autorégressifs, les résultats sont un parallèle de la profondeur de régression pour les modèles de régression standard. Nous étudions ses applications statistiques possibles, en particulier son potentiel pour l'estimation des quantiles et des estimateurs du maximum de profondeur autorégressive, qui s'avèrent à être une extension de la proposition ad hoc pour les modèles AR(1) qui sont déjà apparus dans la littérature. Nous explorons, empiriquement par l'intermédiaire de simulations et d'exemples et théoriquement par l'intermédiaire de résultats asymptotiques liés aux estimateurs généraux du maximum de profondeur, comment les procédures autorégressives basées sur la profondeur se comparent avec les propositions semblables basées sur la médiane,

par exemple les procédures L1, et celles basées sur les signes. Une attention particulière est prêtée aux propriétés de robustesse et au comportement pour des distributions avec de fortes queues.

Wednesday June 11 • Mercredi 11 juin 10:45 • 10h45 LSC 234

Shoja'eddin Chenouri, University of Waterloo

Data depth and a multivariate robust nonparametric multisample test • Profondeur des données et un test multivarié non paramétrique robuste pour échantillons multiples

Order statistics and linear ranking are powerful tools in the univariate nonparametric statistics, but because of lack of natural ordering on p dimensional Euclidian Space p>2 the extension of linear ranked based methods from univariate to multivariate cases would not be appropriate. The subject of multivariate ordering has attracted considerable interest over the years. There exist many nonparametric multivariate approaches which in essence apply univariate nonparametric methods to analyze the multivariate observations coordinate wise. However, they have difficulties in cases of dependence between coordinates of variables. A multivariate methodology based on the concept of data depth has been recently developed (eg. Liu, Parelius and Singh (1999)). A data depth is a measure of how deep or how central a given point is with respect to a multivariate distribution or data cloud. This leads to a natural center-outward ordering of the sample points. In this paper we will review briefly the existing Non-Parametric Multivariate Multisample tests and introduced a new Multivariate Robust Non-Parametric rank test for testing the equality of two or more multivariate populations using depth based ranking.

Les statistiques d'ordre et le rang linéaire sont des outils puissants en statistiques non paramétriques univariées, mais comme il n'est pas facile d'ordonner des éléments dans un espace euclidien à p dimensions (p>2), la généralisation au cas multivarié des méthodes basées sur les rangs linéaires n'est pas appropriée. Le sujet de l'ordonnance multivariée a attiré un intérêt considérable au cours des dernières années. Il existe beaucoup d'approches multivariées non paramétriques qui appliquent essentiellement des méthodes non paramétriques univariées pour analyser des observations multivariées par rapport à leurs coordonnées. Cependant, ces méthodes éprouvent des difficultés dans les cas de dépendance entre les coordonnées des variables. Une méthodologie multivariée basée sur le concept de la profondeur des données a été récemment développée (par exemple Liu, Parelius et Singh (1999)). La profondeur des données est une mesure de la profondeur ou de la centralité d'un point donné par rapport à une distribution multivariée ou un nuage de points. Ceci permet d'ordonner les points d'un échantillon du centre vers l'extérieur. Dans cette présentation, nous passons en revue brièvement les tests multivariés non paramétriques pour des échantillons multiples existants et nous introduisons un nouveau test d'ordre multivarié non paramétrique robuste pour tester l'égalité de deux populations multivariées ou plus qui utilise l'ordonnancement basée sur la profondeur.

Wednesday June 11 • Mercredi 11 juin 11:00 • 11h00 LSC 234

Howard Wainer, National Board of Medical Examiners; Eric Bradlow, University of Pennsylvania; Xiaohui Wang, University of North Carolina

Testlet response theory • Théorie des réponses testlet

Standard item response theory (IRT) models fit to dichotomous examination responses ignore the fact that sets of items (testlets) often come from a single common stimuli (e.g.

a reading comprehension passage). In this setting, all items given to an examinee are unlikely to be conditionally independent (given examinee proficiency). Models that assume conditional independence will overestimate the precision with which examinee proficiency is measured. Overstatement of precision may lead to inaccurate inferences such as prematurely ending an examination in which the stopping rule is based on the estimated standard error of examinee proficiency (e.g an adaptive test). To model examinations that may be a mixture of independent items and testlets, we modified a standard IRT model to include an additional random effect for items nested within the same testlet. We use a Bayesian framework to facilitate posterior inference via a Data Augmented Gibbs Sampler. In this talk I will describe the new model and demonstrate how some difficult problems in test theory are solved simply within this Bayesian framework.

Les modèles standard de la théorie des réponses aux items (IRT), adaptés à des réponses d'examen dichotomiques, ignorent le fait que les ensembles d'items (testlets) viennent souvent d'un seul et même stimulus (exemple la compréhension d'un passage de lecture). Dans cette situation, tous les items donnés à un candidat ne sont pas susceptibles d'être conditionnellement indépendants (étant donné la compétence du candidat). Dans cette situation, les modèles qui supposent l'indépendance conditionnelle surestiment la précision avec laquelle la compétence du candidat est mesurée. Cette surestimation de la précision peut mener à des inférences imprécises telles la fin prématurée un examen dans lequel le critère d'arrêt est basée sur l'estimation de l'écart type de la compétence du candidat (par exemple un test adaptatif). Pour modéliser les examens qui peuvent être un mélange d'items et de testlets indépendants, nous avons modifié le modèle standard d'IRT pour inclure un effet aléatoire additionnel pour des items provenant du même testlet. Nous employons la théorie bayésienne pour faciliter l'inférence a posteriori par l'intermédiaire d'un échantillonneur de Gibbs pour données augmentées. Dans cette présentation nous décrivons le nouveau modèle et démontrons comment quelques problèmes difficiles de la théorie des tests sont résolus simplement dans le cadre bayésien.

Wednesday June 11 • Mercredi 11 juin 11:15 • 11h15 LSC 234

B.M. Golam Kibria, Florida International University

The predictive distributions of regression and sum of squares and products matrices for the multivariate elliptically contoured distributions • Les distributions prédictives de la régression, des sommes de carrés et des matrices produits pour les distributions de contours elliptiques multivariées

The predictive inference of the future regression as well as sum of squares and product (SSP) matrices under the multivariate elliptically contoured error distributions are considered in this paper. It has been shown that the predictive distributions of future regression matrix and SSP matrix under the elliptical error assumption are identical to those obtained under matric normal or matric-t errors. This gives inference robustness with respect to departure from the reference case of independent sampling from the matric normal or dependent but uncorrelated sampling from matric-t distributions.

L'inférence prédictive de la régression future, des sommes des carrés et des matrices produits (SSP) sous les distributions des erreurs de contours elliptiques multivariées est considérée dans cette présentation. Il a été montré que les distributions prédictives de la matrice de régression future et de la matrice de SSP sous l'hypothèse de la distribution elliptique de l'erreur sont identiques à celles obtenues sous des erreurs matric-normales ou matric-t. Ceci donne de la robustesse à l'inférence par rapport au cas de référence

d'échantillonnage indépendant de la distribution matric-normal ou de l'échantillonnage dépendant mais non-corrélé de la distribution matric-t.

Wednesday June 11 • Mercredi 11 juin 11:30 • 11h30 LSC 234

Adrian MacKenzie, Dalhousie University; Lehana Thabane, McMaster University; Joe Apaloo, St. Francis Xavier University

What motivates students to work hard? • Qu'est-ce qui motive les étudiants à travailler fort?

It is of interest to professors, and indeed to all educators, to know how to motivate their students to work hard. One would expect that this becomes easier for an educator to do if he/she has an understanding of his/her students' personal academic motivations, i.e. what makes them want to work hard in their studies. In this presentation, we provide some insights on the subject based on the results of a survey of St. Francis Xavier University students, conducted in the fall of 2001. "Getting a satisfying career after graduation" and "getting good grades" were rated the top two motivating factors by students. We also investigated differences in motivation between students in different subgroups, including gender, ethno-cultural group and Faculty.

Il est de l'intérêt des professeurs, et par le fait même, de tous les éducateurs, de savoir motiver leurs étudiants à travailler fort. Certains peuvent croire qu'il est plus facile pour un éducateur de le faire si il/elle connaît les motivations académiques personnelles de ses étudiants, c.-à-d. ce qui leur donne le goût de travailler fort dans leurs études. Dans cette présentation, nous fournissons quelques avant goût sur le sujet en nous basant sur les résultats d'un sondage présenté aux étudiants de l'université de St-Francis Xavier à l'automne 2001. "Obtenir une carrière satisfaisante après la graduation" et "obtenir de bons résultats" ont été les deux principaux facteurs de motivation soulevés par les étudiants. Nous étudions également les différences de motivation entre les étudiants de différents sous-groupes, y compris le sexe, le groupe ethno-culturel et la faculté.

Wednesday June 11 • Mercredi 11 juin 11:45 • 11h45 LSC 234

Ehsanes Saleh, Carleton University; P.K. Sen, University of North-Carolina, Chapel Hill

Robust estimation of slope parameter in a simple linear model with measurement errors • Estimation robuste de la pente dans un modèle linéaire simple avec des erreurs de mesure

Least-squares estimators of the slope parameter in a simple linear model with measurement errors is vulnerable to gross errors and sensitive to non-normality.In this presentation,we consider a simple robust (point as well as interval) estimator of the slope parameter based on the U-statistics known as Kendall's tau.The point estimator is the median of pair-wise divided differences of the observed set of variables and the confidence interval is based on two suitably chosen order statistics of this set of observed divided differences.Various properties of the estimators are studied and compared with those of the least squares estimators.

Les estimateurs de moindres carrés du paramètre de pente dans un modèle linéaire simple avec des erreurs de mesure sont vulnérables aux erreurs brutes et sensible à la non-normalité des données. Dans cette présentation, nous considérons un estimateur robuste simple (ponctuel aussi bien que par intervalle) du paramètre de pente basé sur les statistiques U, connues sous le nom de coefficient tau de Kendall. L'estimateur ponctuel est la

médiane des différences divisées des paires de l'ensemble des variables observé et l'intervalle de confiance est basé sur deux statistiques d'ordre convenablement choisies de cet ensemble de différences divisées. Plusieurs propriétés de ces estimateurs sont étudiés et comparés à celles des estimateurs de moindres carrés.

Session/Séance 52 • Stochastic Aspects of Forestry • Aspects stochastiques de la foresterie

Wednesday June 11 • Mercredi 11 juin 1:30 • 13h30 LSC 238

Eldon Gunn, Dalhousie University; Tim McGrath, N.S. Dept. of Natural Resources

Why doesn't thinning alter the diameter: a simple simulation model • Pourquoi l'amincissement n'affecte pas le diamètre: un modèle de simulation simple

We have shown that the diameter distribution of unmanaged, fully stocked freely growing hardwood stands can be modelled as a family of Weibull distributions where the location, scale and shape parameter vary smoothly as a function of the stand total diameter. Surprisingly, we also found through hypothesis testing that the diameter distributions developed for the unmanaged stands appeared to be equally good as a model for managed stands which had been commercially or pre-commercially thinned. To understand this phenomenon, we developed a simple simulation model of the thinning process and use this to demonstrate that this phenomenon is plausible. This property of the diameter distribution is very useful for modelling the effects of cleaning and thinning. It gives a way of predicting the change in stand average diameter for any level of percent basal area removal.

Nous avons prouvé que la distribution du diamètre des troncs d'arbres à bois dur qui poussent librement sans gestion particulière peut être modélisé comme une famille de distributions Weibull où les paramètres de localisation, d'échelle et de forme changent de manière lisse en fonction du diamètre total du tronc. Étonnamment, nous avons également trouvé par des tests d'hypothèse que les distributions du diamètre développées pour les troncs non gérés semblent être équivalents au modèle pour les troncs contrôlés qui avaient été aminci commercialement ou pré-commercialement. Pour comprendre ce phénomène, nous développons un modèle simple de simulation du processus d'amincissant et employons ceci pour démontrer que ce phénomène est plausible. Cette propriété de la distribution du diamètre est très utile pour modéliser les effets du nettoyage et de l'amincissement. Elle donne une manière de prévoir le changement du diamètre moyen du tronc pour n'importe quel baisse de pourcentage du niveau basique dans le secteur.

Wednesday June 11 • Mercredi 11 juin 2:00 • 14h00 LSC 238

John Braun, Marc Vincelli, University of Western Ontario

A northwestern Ontario forest fire weather simulator • Un simulateur climatique des feux de forêts du Nord-Ouest de l'Ontario

An R package for simulating forest fire weather is described. The simulator generates realistic, relative humidity, rainfall, wind speed and wind direction at three weather stations in Northwestern Ontario for use in forest fire simulators. The simulator also computes the following seven Canadian Forest Fire Weather Index System (CFFWIS) measures: Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), Buildup Index (BUI), Fire Weather Index (FWI), and the Daily Severity Rating (DSR). The simulator is a parametric/nonparametric bootstrap of 36 years of data collected at the three weather stations.

Nous décrivons un package R pour simuler le climat des feux de forêts. Le simulateur génère de l'humidité relative, des précipitations et la vitesse et la direction des vents réalistes de vent à trois stations météorologiques du Nord-Ouest de l'Ontario pour des simulateurs de feux de forêts. Le simulateur calcule également les sept mesures suivantes du Système d'index du climat des feux de forêts canadien (CFFWIS): Code fin d'humidité de carburant (FFMC), code d'humidité de Duff (DMC), code de sécheresse (C.C), index de diffusion initiale (ISI), index d'alimentation (BUI), index de climat des feux (FWI) et l'estimation quotidienne de sévérité (DSR). Le simulateur est un bootstrap paramétrique et non-paramétrique sur 36 ans de données rassemblées aux trois stations météorologiques.

Wednesday June 11 • Mercredi 11 juin 2:30 • 14h30 LSC 238

David Martell, B.M. Wotton, University of Toronto; K.A. Logan, Canadian Forest Service

The development and use of daily people-caused forest fire occurrence models in Ontario • Le développement et l'utilisation des modèles d'occurrences journalières pour les feux de forêts causés par erreurs humaines en Ontario

The number of people-caused forest fires that occur across a forest region each day varies with the moisture content of the forest vegetation or fuels on the forest floor. Poisson regression analysis methods are used to develop predictive models for the number of fires occurring each day across Ontario. Significant predictors in the models include weather-based estimates of forest floor moisture content and the probability of sustained flaming combustion as well as periodic functions that model seasonality. Predictions are tested against 4 years of recent people-caused fire occurrence data not included in the original dataset. We illustrate how these people-caused fire occurrence prediction models are used for planning the daily deployment of fire fighting resources and combined with general circulation model data scenarios of future fire weather to generate predictions of the change in the level of people-caused fire activity across the province of Ontario in the future.

Le nombre de feux de forêts causés par des erreurs humaines qui se produisent chaque jour à travers une région forestière varie selon la teneur en humidité de la végétation de la forêt ou des carburants au sol. Des méthodes d'analyse de régression de Poisson sont utilisées pour développer des modèles de prévision du nombre de feux qui se produisent chaque jour à travers l'Ontario. Les prédicteurs significatifs dans les modèles incluent des estimations, basées sur le climat, de la teneur en humidité du sol forestier et la probabilité d'une combustion soutenue ainsi que des fonctions périodiques qui modélise la saisonnalité. Des prévisions sont examinées sur 4 ans de données d'occurrence de feux causés par erreurs humaines qui ne sont pas incluses dans le jeu de données original. Nous illustrons comment ces modèles d'occurrence des feux causés par erreur humaine sont utilisés pour planifier le déploiement quotidien des ressources pour lutter contre les incendies et comment ils sont combinés avec les scénarios de données de modèles de circulation général de climat de feu pour produire des prévisions du changement du niveau de l'activité de feux causés par erreurs humaine à travers la province d'Ontario.

Session/Séance 53 • Estimation of Fish Stock Mixtures • Estimation des mélanges de stocks de poissons

Wednesday June 11 • Mercredi 11 juin 1:30 • 13h30 LSC 338

John Candy, T. D. Beacham, Department of Fisheries and Oceans

To Bayes or not to Bayes: MLE vs Bayesian analysis for mixed stock fisheries using highly polymorphic DNA markers in Pacific Salmon • To Bayes or not to Bayes: l'analyse bayésienne vs les EMV pour la pêche à espèces variées utilisant des marqueurs très polymorphes d'ADN chez les saumons du Pacifique

Increasingly, west coast of Canada fisheries managers rely on accurate stock composition estimates to fish abundant populations while protecting those that are vulnerable to over-explotation. Depending on the species, we screen between 5-15 thousand individuals annually from mixed stock fisheries for 10-15 microsatellite/MHC loci. Using coast-wide baselines (Russia-California) of between 40-45 thousand individuals, estimates of stock composition are calculated using Maximum Likelihood Estimation (MLE) or a Bayesian algorithm. Some of this analysis occurs "in-season" with turn-around times of between 9-48 hours between the time the lab receives the tissue samples and the results provided to the managers. The MLE method provides fast analysis with both the baseline and mixture being bootstrapped to incorporate variation in the baseline and mixture. However, the highly polymorphic nature of the microsatellite markers may result in "rare alleles" occurring in the mixture and not in the baseline causing the multi-locus conditional probability matrix to be zero for some populations although it may be "stock of origin". To get around this problem data can be binned (adjacent alleles combined) to remove "rare alleles" from the data set. However, binning alleles reduces the combined information content of the markers which may reduce the accuracy of the estimate. A recently developed alternative mixture analysis method employs a Bayesian approach which is computationally intensive but does not require binned data. Accuracy of the two methods are evaluated using coded-wire tagged fish for chinook salmon and blind mixture samples made from spawning ground collections for sockeye salmon.

De plus en plus, les directeurs des pêcheries de la côte ouest du Canada se basent sur des estimations précises de la composition des stocks pour exploiter les populations abondantes tout en protégeant celles qui sont vulnérable à la sur-exploitation. Selon l'espèce, nous examinons entre 5000 et 15000 individus annuellement provenant de pêche à espèces variées pour 10 000 à 15 000 locus de microsatellite/MHC. En utilisant les lignes de base le long des côtes (Russie,Californie) d'environ 40000 à 45000 individus, des estimations de la composition des stocks sont calculées en utilisant l'estimation par le maximum de vraisemblance (EMV) ou un algorithme bayésien. Une partie de cette analyse se produit en saison avec des délais d'entre 9 à 48 heures entre le moment où le laboratoire reçoit les échantillons de tissu et le moment où les résultats sont fournis aux directeurs. La méthode du maximum de vraisemblance fournit une analyse rapide avec un bootstrap sur les espèces de la ligne de base et celles variées pour incorporer la variation entre ces dernières. Cependant, la nature très polymorphe des marqueurs de microsatellites peut avoir comme conséquence l'apparition d'allèles rares dans la ligne variée et pas dans la ligne de base causant la matrice des probabilités conditionnelles des multi-locus d'être zéro pour certaines populations bien que ce puisse être les stocks d'origines. Pour venir à bout de ce problème, les données peuvent être jumelées (les allèles adjacents combinés) pour enlever les allèles rares de la base de donnée. Cependant, le jumelage des allèles peut réduire l'information conjointe

des marqueurs ce qui peut réduire la précison de l'estimation. Une méthode alternative récemment développée pour l'analyse des espèces variées utilise une approche bayésienne qui demande beaucoup de temps de calcul mais qui n'exige pas le jumelage des données. La précision des deux méthodes est évaluée en utilisant des poissons étiquetés par des fils codés pour le saumons chinook et des échantillons d'un mélange pris à partir de frayère pour des saumons sockeyes.

Wednesday June 11 • Mercredi 11 juin 2:00 • 14h00 LSC 338

Daniel Ruzzante, Dalhousie University; M.M. Hansen, K. Ebert, D. Meldrup, Danish Institute for Fisheries Research

Individual and population level approaches to the analysis of stocking impact in an anadromous brown trout (Salmo trutta) complex • Une approche au niveau de l'individu et de la population pour l'analyse de l'impact du peuplement dans un complexe de truite brune anadrome (trutta de Salmo)

Polymorphism was examined at 7 microsatellite loci in ˜4900 Salmo trutta collected from 32 tributaries to the Limfjord in Denmark, from two hatcheries used for stocking these tributaries, and from three regions in the marine environment of the Limfjord (anadromous sea trout). Populations differ in their estimated sizes and stocking histories. Relatedness varies between sites within rivers indicating varied local dynamics at a very small geographic scale. In the western, less heavily stocked area of the Limfjord a higher proportion of the genetic variance is distributed among rivers than among locations within rivers while the reverse is true of the eastern, more heavily stocked area of the Limfjord. These and related results of assignment tests can be interpreted as reflecting stocking impact. The null hypothesis that stocking has had no impact on population structure is rejected but the relatively high proportion of locally assigned trout in populations where stocking with domestic fish no longer takes place suggests limited long term success of stocking. The populations of most likely origin of the sea trout is examined using a recently developed Bayesian method for mixed stock analysis and results are interpreted in the light of known aspects of the ecology of sea trout in the region.

Le polymorphisme est examiné à 7 locus de microsatellite dans 4900 truites brunes rassemblé de 32 tributaires au Limfjord au Danemark, de deux établissements d'incubation utilisés pour stocker ces tributaires, et de trois régions de l'environnement marin du Limfjord (truite anadrome de mer). Les populations diffèrent dans leurs tailles estimées et leur histoires d'incubation. La relation change entre les emplacements dans les fleuves indiquant une dynamique locale diverse à une échelle géographique très petite. Dans le secteur ouest, le moins fortement stocké du Limfjord, une proportion plus élevée de la variance génétique est distribué selon différents fleuves que parmi les endroits à l'intérieur de ces fleuves tandis que l'inverse est vrai pour le secteur est, le plus fortement stocké du Limfjord. Ces résultats relatifs des tests d'assignation peuvent être interprétés comme l'impact du peuplement. L'hypothèse nulle que le peuplement n'a eu aucun impact sur la structure de la population est rejetée mais la proportion relativement élevée de la truite localement assignée dans les populations où le peuplement avec des poissons domestiques n'a plus lieu suggère un succès à long terme limité du peuplement. Les populations d'origine les plus susceptible de la truite de mer est examinées en utilisant une méthode bayésienne récemment développée pour l'analyse des peuplements mixtes et les résultats sont interprétés à la lumière des aspects connus de l'écologie de la truite de mer dans la région.

Session/Séance 54 • Applied Probability • Probabilité appliquée

Wednesday June 11 • Mercredi 11 juin 1:30 • 13h30 LSC 240

Gordon Willmot, University of Waterloo; David Dickson, University of Melbourne

The Gerber-Shiu discounted penalty function in the stationary renewal risk model • La fonction escomptée de pénalité de Gerber et Shiu dans le modèle de risque de renouvellement stationnaire

The discounted penalty function introduced by Gerber and Shiu (1998) is considered in the stationary renewal risk model, where it is expressed in terms of the same discounted penalty function in the ordinary renewal risk model. This relationship unifies and generalizes known special cases. An invariance property between the stationary renewal risk model and the classical Poisson model with respect to the ruin probability is also generalized as a result.

La fonction escomptée de pénalité présentée par Gerber et Shiu (1998) est considérée dans le modèle stationnaire de renouvellement du risque, où elle est exprimée en termes de la même fonction de pénalité escomptée mais dans le modèle ordinaire de renouvellement du risque. Cette relation unifie et généralise des cas particuliers connus. Une propriété d'invariance entre le modèle stationnaire de renouvellement du risque et le modèle classique de Poisson par rapport à la probabilité de ruine est également généralisée comme une conséquence.

Wednesday June 11 • Mercredi 11 juin 2:00 • 14h00 LSC 240

Reg Kulperger, University of Western Ontario; Zengjing Chen, University of Western Ontario and Shandong University, China

Stochastic prey predator system • Système proie-prédateur stochastique

The classic prey predator differential equation is $dx_t/dt = x_t(\mu - \lambda y_t) dy_t/dt = y_t - \alpha + \eta x_t$ where the constants μ etcetera are positive. It is stable and the solution lives on the contour $F(x,y) = \eta x - \alpha \log(x) + \lambda y - \mu \log(y) = constant$. Suppose the rates become random $dx_t = x_t((\mu - \lambda y_t)dt + \sigma_1 dW_t^{\{(1)\}}) = y_t((-\alpha + \eta x_t)dt + \sigma_2 dW_t^{\{(2)\}})$ where $W^{\{(1)\}}$ and $W^{\{(2)\}}$ are independent Brownian motions. If $\min(\sigma_1, \sigma_2) > 0$ (ie some noise) then the random contour level $F(x_t, y_t)$ has the property $E(F(x_t, y_t) = F(x_0, y_0) + \frac{1}{2}(\alpha\sigma_1^2 + \mu\sigma_2^2)t \to \infty \; as \; t \to \infty$. The the random contour level tends to infinity in some sense. We study in what sense this stochastic system is unstable, that is the stochastic prey predator null recurrent or transient. Classical methods deal with functions of the Euclidean norm, but that is not helpful in this problem. Different functionals are used to give transient versus recurrent criteria.

L'équation différentielle proie-prédateur classique est défini comme suit: $x_t/dt = x_t(\mu - \lambda y_t) dy_t/dt = y_t - \alpha + \eta x_t$ *où les constantes* μ *etcetera sont positives. Ce système est stable et la solution se trouve sur le contour* $F(x,y) = \eta x - \alpha \log(x) + \lambda y - \mu \log(y) = constant$. *Supposons que les taux deviennent aléatoires* $dx_t = x_t((\mu - \lambda y_t)dt + \sigma_1 dW_t^{\{(1)\}}) = y_t((-\alpha + \eta x_t)dt + \sigma_2 dW_t^{\{(2)\}})$ *où* $W^{\{(1)\}}$ *et* $W^{\{(2)\}}$ *sont des mouvements browniens aléatoires. Si* $\min(\sigma_1, \sigma_2) > 0$ *(c'est à dire du bruit) alors le niveau de contour aléatoire* $F(x_t, y_t)$ *à la propriété* $E(F(x_t, y_t) = F(x_0, y_0) + \frac{1}{2}(\alpha\sigma_1^2 + \mu\sigma_2^2)t \to \infty \; as \; t \to \infty$. *Le niveau de contour aléatoire tend vers l'infinie en quelques sortes. Nous étudions dans quel sens ce système stochastique est instable, c'est à dire le proie-prédateur stochastique nul récurent ou transitoire. Les méthodes classiques utilisent des fonctions de la norme euclidienne, mais ce n'est pas utile pour notre problème. Différentes fonctions sont utilisées pour donné un critère transitoire versus récurent.*

Wednesday June 11 • Mercredi 11 juin 2:30 • 14h30 LSC 240

Chris Small, University of Waterloo; Huiling Le, University of Nottingham

Modelling the shapes of random curves • Modélisation de la forme de courbes aléatoires

In this talk, we shall examine the properties of a stochastic model for the shapes of random curves such as arise in diverse applications such as image analysis, medical research, growth patternsof organisms, as well as geometrically complex curves such as produced by handwriting script. We understand a curve in a general sense as a directed continuous path in the plane with clearly defined endpoints. The curve may have corners or a more complex set of locations where the tangent vector is not defined. The model that we propose for such paths has certain advantages: it is robust to digitisation of the image; it assigns simple parameter values to geometrically simple curves such as lines and circles; it permits simple "averaging" of curve shapes to produce "textbook" specimens; and it is compatible with a straightforward metric for computing the difference between curves of different shapes for classification, clustering or the fitting of idealised curve shapes. We will also show that simple moment properties, such as the moments of the vector difference of the endpoints, can be calculated from the model. This permits the use of method of moment techniques for fitting parameters to curves.

Dans cette présentation, nous examinons les propriétés d'un modèle stochastique pour les formes de courbes aléatoires qui surviennent dans plusieurs applications telles qu'en analyse d'images, en recherche médicale et en schème de croissance d'organismes, aussi bien que pour les courbes géométriquement complexes comme l'écriture manuscrite. Nous définissons une courbe de manière général comme un chemin continu et dirigé dans le plan avec un point de départ et d'arrêt bien définis. La courbe peut avoir des coins ou un ensemble de points plus complexe où le vecteur tangent n'est pas défini. Le modèle que nous proposons pour de tels chemins a certains avantages: il est robuste à la numérisation de l'image; il assigne des valeurs de paramètre simples aux courbes géométriquement simples telles que les lignes et les cercles; il permet l'estimation simple de la formes des courbes pour produire des spécimens de référence; et il est compatible avec une métrique directe pour calculer la différence entre des courbes de différentes formes pour la classification, l'analyse de grappes ou l'ajustement à certaines formes de courbe. Nous montrons également que des propriétés simples des moments, telles les moments du vecteur de la différence des points aux extrémités, peuvent être calculées à partir du modèle. Ceci permet l'utilisation de techniques par la méthode des moments pour ajuster des paramètres aux courbes.

Session/Séance 55 • Special Session of the Centre de Recherches Mathématiques on Statistics and Finance • Session speciale du Centre de Recherches Mathématiques en statistique et finance

Wednesday June 11 • Mercredi 11 juin 1:30 • 13h30 LSC 242

Eric Renault, Université de Montréal

Dynamic factor models in finance • Modèles factoriels dynamiques en finance

Factor models in finance originate both from asset pricing theory and time series analysis. These two strands of literature appeal to two different concepts of factors, which are both useful to reduce the dimension of the statistical model. In the CAPM or APT beta pricing models, the dimension reduction is cross-sectional in nature, through a postulated conditional independence between contemporaneous returns of a large number of assets given a small number of factors like in standard Factor Analysis. In time series state-space models,

dimension is reduced longitudinally by assuming conditional serial independence between consecutive returns given a small number of state variables also often called factors. In this lecture, we will provide a set of unifying principles to better integrate asset pricing and time series concepts of factors in the context of multivariate processes of conditionally heteroskedastic returns. Modeling issues will therefore take precedence over inference methods, but some specific techniques which have recently been developed for this class of models will be reviewed.

Les modèles factoriels en finance proviennent de la théorie du prix des capitaux et de l'analyse de séries chronologiques. Ces deux courants de littérature font appel à deux concepts différents de facteurs, qui sont tout les deux utiles pour réduire la dimension du modèle statistique. Dans le modèle de prix bêta CAPM ou APT, la réduction de la dimension est de nature inter-sectionnelle, par l'entremise d'une indépendance conditionnelle postulée entre les retours contemporains d'un grand nombre de capitaux donné un nombre restreint de facteurs comme dans l'analyse factorielle standard. Dans les modèles de séries chronologiques d'état et d'espace, la dimension est réduite longitudinalement en supposant l'indépendance périodique conditionnelle entre les retours consécutifs donné un petit nombre de variables d'états souvent appelés facteurs. Dans cette conférence, nous fournissons un ensemble de principes d'unification pour mieux intégrer les concepts d'évaluation des prix des capitaux et des séries chronologiques des facteurs dans le contexte des processus multivariées des retours conditionnellement hétéroscédastiques. Nous discuterons principalement des problèmes de modélisation plutôt que des méthodes d'inférences, mais quelques techniques spécifiques d'inférence récemment développées pour cette classe de modèles seront passées en revue.

Wednesday June 11 • Mercredi 11 juin 2:00 • 14h00 LSC 242

Jin-Chuan Duan, Geneviève Gauthier, Jean-Guy Simonato, Sophia Zaanoun, University of Toronto

Estimating structural credit risk models with consideration of survivorship • Estimation du modèle structural du risque de crédit avec des considérations de survie

One critical difficulty in implementing structural credit risk models is that the underlying asset value cannot be directly observed. Models require the unobserved asset value and the unknown parameter(s) as inputs; for example, asset value and volatility are in practice unknown when the model of Merton (1974) is applied. The estimation problem is further complicated by the fact that typical data samples are for the survived firms. This paper applies the maximum likelihood principle to develop an estimation procedure. The maximum likelihood estimator for parameter(s), asset value, credit spread and default probability are derived for Merton's model. A Monte Carlo study is conducted to examine the performance of the maximum likelihood method. An application to real data is also presented.

Une difficulté critique dans l'implémentation des modèles structuraux de risque de crédit est qu'on ne peut pas directement observer la valeur sous jacente des actifs. Les modèles requièrent la valeur des actifs non observés et le(s) paramètre(s) inconnu(s) comme entrées; par exemple, la valeur des actifs et la volatilité sont en pratique inconnues quand nous appliquons le modèle de Merton (1974). Le problème d'estimation est encore plus compliqué par le fait que les échantillons typiques de données sont pour les firmes qui ont survécues. Cette présentation applique le principe du maximum de vraisemblance pour développer une procédure d'estimation. L'estimateur du maximum de vraisemblance pour

le(s) paramètre(s), la valeur des actifs, le crédit a écarter et la probabilité par défaut sont dérivées pour le modèle de Merton. Une étude de Monte Carlo est conduite pour examiner la performance de la méthode du maximum de vraisemblance. Nous présentons également une application sur de vraies données.

Wednesday June 11 • Mercredi 11 juin 2:30 • 14h30 LSC 242

Francois Watier, Université de Sherbrooke; Jean Vaillancourt, Université du Québec en Outaouais

Multiperiod and continuous-time mean-variance analysis in portfolio management • Analyse moyenne-variance en gestion de portefeuille dans un contexte multipériodique et en temps continu

Modern portfolio theory was at a corner stone with the publication in 1952 of H.M. Markowitz's celebrated article. This laureate of a Nobel Prize in Economy devised a static strategy allowing an investor to achieve an expected total gain while downsizing risk. We suggest newly developped extensions to his approach in a dynamic setting in both the case of a multiperiod and continous time framework, namely by incorporating the possible use of exogeneous factors in modeling the rate of return of risky assets. Finally, the specific construction of a wide class of models here called SIMMI (stationary multiplicative market impulses) for excess rates of return will allow us to significally reduce the complexity of statistical estimation of parameters while maintaining real-time computational efficiency of the optimal solution.

La théorie moderne de la gestion de portefeuille a connu un véritable essor depuis la publication, au milieu du siècle dernier, du célèbre article de H. M. Markowitz récipiendaire d'un prix Nobel d'économie. Il a développé une stratégie statique permettant d'obtenir un gain moyen espéré de la part d'un investisseur tout en minimisant une mesure de risque exprimée sous la forme d'une variance. Nous proposons d'explorer de nouvelles extensions de cette approche dans un cadre dynamique, aussi bien en contexte multipériodique qu'en temps continu, en suggérant notamment l'incorporation de facteurs exogènes dans la modélisation du rendement des titres risqués. Enfin, l'identification d'une classe générale de modèles dits SIMMI (stationary multiplicative market impulses) pour le rendement excédentaire permettra de réduire la complexité de l'estimation de paramètres statistiques tout en assurant une efficacité numérique des calculs en temps réel de la solution optimale.

Session/Séance 56 • Survey Methods Contributed Session V: Survey Sampling • Méthodes d'enquête V: sondages

Wednesday June 11 • Mercredi 11 juin 1:30 • 13h30 LSC 332

Lenka Mach, Ioana Schiopu-Kratina, Jean-Marc Fillion, Statistics Canada/Statistique Canada; Phil Reiss, Columbia University

Maximizing the overlap of two business surveys • Maximiser le chevauchement de deux enquêtes entreprises

In this paper we describe two methods for maximizing the expected overlap of samples selected before and after a frame re-stratification and/or update for births and deaths. On both occasions, the units within strata are selected according to a simple random sampling without replacement design (SRSWOR). Just as the Kish & Scott (K-S) method, our first method attains the first-order inclusion probabilities in the new design through additional

selection or de-selection of units in the initial strata. It constitutes an improvement over the K-S method, as it actually maximizes the expected overlap. The second method maximizes the overlap and preserves the entire design within a stratum when the frame is updated for births and deaths. The second method can be generalized to minimize the expected overlap of several surveys in order to reduce response burden. We discuss the properties of these two methods as well as the methodology for their implementation. We also compare our methodology to other existing methodologies, for example the micro-strata methodology for sample co-ordination (also known as SALOMON).

Dans ce papier nous décrivons deux méthodes pour maximiser le chevauchement attendu des échantillons sélectionnés avant et après une re-stratification de la base de sondage et/ou une mise à jour pour les naissances et les décès. En les deux occasions, les unités dans la strate sont sélectionnées suivant un tirage aléatoire simple sans remise (EASSR). Comme dans le cas de la méthode de Kish & Scott (K-S), notre première méthode atteint les probabilités d'inclusion à l'ordre un dans le nouveau plan de sondage à partir d'une sélection additionnelle ou désélection d'unités dans la strate initiale. Ceci constitue une amélioration de la méthode de K-S car le chevauchement attendu est maximisé. La seconde méthode maximise le chevauchement et préserve le plan de sondage entièrement au niveau de la strate lorsque la base de sondage est mise à jour pour les naissances et les décès. La seconde méthode peut être généralisée pour minimiser le chevauchement attendu de plusieurs enquêtes pour réduire le fardeau de réponse. Nous discutons les propriétés de ces deux méthodes ainsi que de la méthodologie de la mise en œvre. Nous comparons aussi notre méthodologie à d'autres méthodologies existantes, comme par exemple la méthodologie de micro strates pour la coordination d'échantillons (connu aussi sous SALOMON).

Wednesday June 11 • Mercredi 11 juin 1:45 • 13h45 LSC 332

Rebecca Morrison, Claude Julien, Suzelle Giroux, Statistics Canada/Statistique Canada

Redesign of the agriculture surveys • Remaniement des enquêtes agricoles

Every five years, the agricultural commodity surveys are redesigned following the Canadian Census of Agriculture. Since the previous redesign, significant changes to the target population and the survey program itself have occurred. Firstly, the number of farms in Canada has decreased and, secondly, in order to alleviate respondent burden, the commodity surveys will no longer interview 'small' farms. In addition, it was decided to investigate a further reduction of the sampling population. As a result, the survey designs have undergone many changes; the most significant of which is the adoption of a simpler design with a more dynamic frame to ensure good coverage until the next redesign in 2007. The focus of the presentation will be the design used to achieve the multiple objectives of the survey program in light of the new challenges that the program now faces.

Chaque cinq ans, à la suite du recensement agricole canadien, les enquêtes agricoles sont remaniées. Depuis le dernier remaniement, des changements considérables à la population cible et au programme d'enquêtes se sont produits. Premièrement, le nombre de fermes au Canada a diminué, et deuxièmement, dans le but de diminuer le fardeau de réponse, il a été décidé de ne plus enquêter les ¡¡ petites¿¿ fermes. De plus, il a été décidé d'examiner l'impact d'une réduction encore plus grande de la population à échantillonner. En conséquence, beaucoup de changements au plan de sondage ont été effectués; le plus considérable est l'adoption d'un plan de sondage plus simple avec une base de sondage plus dynamique pour s'assurer une bonne couverture jusqu'au prochain remaniement. Le su-

jet principal de la présentation sera le plan de sondage adopté pour atteindre les objectifs multiples du programme d'enquêtes.

Wednesday June 11 • Mercredi 11 juin 2:00 • 14h00 LSC 332

Wilson Lu, Randy R. Sitter, Simon Fraser University

Multi-way stratification by linear programming made practical • Rendre pratique la stratification à plusieurs étapes par la programmation linéaire

Sitter and Skinner (1994) presented a method which applies linear programming to designing surveys with multi-way stratification. The idea in their approach is simple, easily understood and easy to apply. However, the main practical constraint of their approach is that it rapidly becomes expensive in terms of magnitude of computation as the number of cells in the multi-way stratification increases, to the extent that it cannot be used in most real situations. In this presentation, we will extend this linear programming approach and explore methods to reduce the amount of computation.

Sitter et Skinner (1994) ont présenté une méthode qui applique la programmation linéaire pour planifier des sondages avec stratification multiple. L'idée de leur approche est simple, facile à comprendre et facile à appliquer. Cependant, la principale contrainte pratique de leur approche est qu'elle devient rapidement très coûteuse en termes de calcul, à mesure que le nombre de cellules dans la stratification augmente, à un point tel où elle ne peut plus être utilisée dans la plupart des situations réelles. Dans cette présentation, nous prolongeons cette approche par programmation linéaire et explorons des méthodes pour réduire la quantité de calculs.

Wednesday June 11 • Mercredi 11 juin 2:15 • 14h15 LSC 332

Owen Phillips, Statistics Canada/Statistique Canada; Avi Singh, Research Triangle Institute

Calibration allocation of sample for multiple characteristic surveys under stratified random sampling • Répartition d'échantillon par calage pour les enquêtes à plusieurs variables avec échantillonnage stratifié simple

This project examines the problem of sample allocation for a stratified random design given multiple characteristics of interest. Existing solutions use non-linear programming to obtain a minimum cost and optimal allocation for a given set of variance constraints. If the resulting cost is not acceptable, the existing solution relaxes all variance constraints uniformly (i.e. allows uniform tolerance) such that an acceptable cost can be met. This is not reasonable, as not all of the constraints have an impact of the same order. An alternative formulation is required where there is a direct control on cost and the objective function is defined in terms of the variance constraints such that the two formulations become equivalent. An equivalent formulation is proposed using the idea of penalized distance function as in the ridge-calibration of sampling weights in Rao and Singh (1997). The proposed method is termed calibration allocation and allows for suitable, non-uniform tolerances when cost is reduced.

Ce projet examine le problème de répartition d'échantillon pour un plan d'échantillonnage stratifié simple où il y a plusieurs variables d'intérêt. Les solutions actuelles utilisent la programmation non-linéaire afin d'obtenir le coût minimum et la répartition optimale pour les contraintes de variance spécifiées. Si le coût réalisé n'est pas acceptable, la solution actuelle relaxe toutes contraintes de variance d'une manière uniforme (c.-à-d. permet une tolérance uniforme) de façon à ce qu'un coût acceptable soit atteint. Ceci n'est pas raisonnable car

toutes les contraintes n'ont pas un impact de la même magnitude. Une formulation alternative est requise là où il y a un contôle direct sur le coût et la fonction objective est définie en termes des contraintes de variance de façon à ce que les deux formulations soient équivalentes. Une formulation équivalente est proposée en utilisant l'idée d'une fonction de distance pénalisée comme dans le calage-réduction des poids d'échantillonnage dans Rao et Singh (1997). La méthode proposée s'appelle la répartition par calage et permet une tolérance non-uniforme appropriée lorsque le coût est réduit.

Wednesday June 11 • Mercredi 11 juin 2:30 • 14h30 LSC 332

Joseph Duggan, Elisabeth Neusy, Yves Bélanger, Statistics Canada/Statistique Canada

Sample design issues in a large-scale multi-frame national survey: the Canadian component of the International Adult Literacy and Life- skills survey (ALL) • Problèmes de design d'expérience dans un sondage national à étapes multiples à grande échelle: la composante canadienne du sondage international sur l'alphabétisation des adultes et sur les compétences de vie

This paper provides an overview of the design considerations in the development of the International Adult Literacy and Life-skills Survey (ALL), to be implemented in March to July of 2003. ALL was designed to provide measures of proficiency in several literacy and life-skill domains for the adult populations. The Canadian component also seeks to profile these skill sets for targetted subpopulations such as youth, urban aboriginals, immigrants, and linguistic minorities in certain sub-national regions. Each of these provincial regions has been stratified and has a two-stage (dwelling and then individual) sample for the urban portion and a three-stage sample (with geographical areas as primary sampling units) for the rural portion. A base sample of privately occupied dwellings was selected using the 2001 Canadian Census of Population and Housing was used as a frame for dwellings. Then, supplementary samples were selected sequentially for each region; a modified multi-frame weighting method was proposed to account for the dependencies in this design and the estimation of the variance will be performed using the combined jackknife technique. Other methodological challenges to be covered include adapting the design for the three northern Canadian territories, and inflating sample sizes to account for the mobility of subpopulations in terms of their characteristics of interest.

Cette présentation fournit une vue d'ensemble des considérations du design dans le développement du sondage international sur l'alphabétisation des adultes et des compétences de vie (ALL), qui sera mis en application de mars à juillet 2003. ALL a été conçu pour fournir des mesures de compétence dans plusieurs domaines touchant l'alphabétisation et les compétences de la vie pour les populations d'adulte. La composante canadienne cherche également à profiler ces ensembles de compétence pour viser des sous-populations telles que les jeunes, les aborigènes urbains, les immigrants et les minorités linguistiques dans certaines régions sous-nationales. Chacune de ces régions provinciales ont été stratifiée et a un échantillon à deux étapes (par logement et ensuite par individu) pour la partie urbaine et un échantillon à trois étapes (avec les secteurs géographiques comme principale unité d'échantillonnage) pour la partie rurale. Un échantillon de base de logements occupés par les propriétaires a été choisi en utilisant le recensement canadien de la population et de logements de 2001 comme cadre pour les logements. Puis, des échantillons supplémentaires ont été choisis séquentiellement pour chaque région; une méthode pondéré modifiée multi-trame a été proposée pour tenir compte de la dépendance dans ce design et l'estimation de la variance est effectuée par la technique combinée du jackknife. D'autres défis méthodologiques couverts

incluent l'adaptation du design pour les trois territoires canadiens nordiques, et augmenter les tailles des échantillons pour tenir compte de la mobilité des sous-populations en termes de leurs caractéristiques d'intérêt.

Session/Séance 57 • Distributions and Multivariate Methods • Distribution et méthodes multidimensionnelles

Wednesday June 11 • Mercredi 11 juin 1:30 • 13h30 LSC 234

Louis Doray, Université de Montréal

Estimation for the discrete generalized Linnik distribution • Estimation pour la loi de Linnik généralisée discrète

For this non-negative discrete distribution, no analytic expression for its probability function exists, making maximum likelihood estimation of its parameters difficult to apply. Using the probability generating function (pgf), we will develop a recursive relationship for the terms of its probability function. To estimate the parameters of the distribution, we will minimize a quadratic distance between the theoretical and the empirical pgf's. The asymptotic properties of the estimators obtained with this method will be analyzed and it will be shown how to implement the minimization algorithm using an iteratively reweighted least-squares procedure.

Pour cette loi discrète définie sur les entiers non-négatifs, il n'existe pas d'expression analytique pour sa fonction de masse, rendant l'estimation des paramètres par la méthode du maximum de vraisemblance difficile à appliquer. En utilisant la fonction génératrice des probabilités (fgp), nous développons une relation de récurrence entre les termes de sa fonction de masse. Pour estimer les paramètres de la loi, nous minimisons une distance quadratique entre les fgp théorique et empirique. Nous analysons les propriétes asymptotiques des estimateurs obtenus par cette méthode, et montrons comment implanter l'algorithme de minimisation, avec la procédure des moindres carrés pondérés itérés.

Wednesday June 11 • Mercredi 11 juin 1:45 • 13h45 LSC 234

Abdel El-Shaarawi, National Water Research Institute

Exact and approximate expressions for the tail of Student's t and F distributions • Expressions exactes et par approximations pour la queue des distributions t de Student et F de Fisher

A recursive formula will be presented for computing the exact cdf $F(x)$ of the F distribution. This is then specialized to obtain the exact cdf of the Student t distribution. It shows the slow convergence of the Student t to the normal distribution. Furthermore, a general and simple formula for approximating the tail of probability distribution for large x will be presented. Limited numerical results will demonstrate its adequacy for approximating the tail of the normal distribution.

Une formule récursive est présentée pour calculer la fonction de répartition exacte $F(x)$ de la distribution F. Cette approche est ensuite spécialisée pour obtenir la fonction de répartition exacte de la distribution t de Student. Nous montrons la convergence lente de la loi t de Student vers la distribution normale. De plus, une formule simple et générale pour estimer la queue de la distribution pour grand un grand x est présentée. Certains résultats numériques démontrent que la formule est adéquate pour estimer la queue de la distribution normale.

Wednesday June 11 • Mercredi 11 juin 2:00 • 14h00 LSC 234

Denis Larocque, Mélanie Labarre, HEC Montréal

A one-sided (positive orthant) conditionally distribution-free sign test for multivariate data • Un test du signe conditionnellement "distribution-free" pour contre-hypothèses unilatérales avec données multidimensionnelles

The one-sided (positive orthant) one sample problem with multivariate data is considered. A conditionally distribution-free sign test is proposed for that problem. This test is related to Hodges test and can be seen as a union-intersection test. Moreover, it is valid under very mild assumptions. For the bivariate case, an explicit formula for the exact null conditional distribution of the test statistic is derived. For dimensions greater than two, the exact null conditional distribution can be approximated to any desired accuracy by simulation.

Nous considérons le problème de position multidimensionnel unilatéral. Nous proposons un test du signe conditionnellement " distribution-free ". Ce test est apparenté au test de Hodges et peut être vu comme un test de type union-intersection. De plus, il est valide sous des présupposés très peu restrictifs. Dans le cas bidimensionnel, nous présentons une formule explicite pour le calcul de la loi conditionnelle exacte de la statistique de test sous l'hypothèse nulle. Cette loi peut être approximée par simulation, avec un degré de précision aussi grand que souhaité, pour les dimensions supérieures à deux.

Wednesday June 11 • Mercredi 11 juin 2:15 • 14h15 LSC 234

Mouna Fallaha, Aleppo University

The asymptotic normality of the maximum pseudo-likelihood estimator of the parameters of Markov random fields • La normalité asymptotique des estimateurs du maximum de la pseudo-vraisemblance conditionnelle des paramètres de champs de Markov

We study the asymptotic normality of the pseudo-likelihood estimator for Gibbs Markov Random Fields. Two cases are considered: that in which the energy depends linearly on the parameters, then that where the dependence is non-linear. To obtain the asymptotic normality of these estimators, we adopt two new techniques of martingale approximation for the Random Fields. A simulation study is also presented, we also calculate the approximate asymptotic variances of these estimations.

Nous étudions la normalité asymptotique des estimateurs du maximum de la pseudo-vraisemblance conditionnelle de paramètres de champs de Gibbs markoviens et stationnaires. Deux cas sont considérés, celui où l'énergie dépend linéairement des paramètres, puis celui où la dépendance est non-linéaire. Pour obtenir la normalité asymptotique de ces estimateurs, nous adoptons deux nouvelles techniques d'approximation par des martingales pour les champs aléatoires. Des simulations numériques sont présentées, nous calculons aussi les variances asymptotiques approchées de ces estimateurs.

Wednesday June 11 • Mercredi 11 juin 2:30 • 14h30 LSC 234

Abdeljelil Farhat, Centre for Interuniversity Research and Analysis on Organizations; Jean-Marie Dufour, Université de Montréal

Exact k-sample goodness-of-fit tests for continuous and discrete distributions • Tests d'ajustement de K distributions continues ou discrètes

An important statistical problem in biology, econometrics and finance consists in testing whether the observations from k different samples have the same distribution (the k-sample

homogeneity hypothesis). In this paper, we focus on cases where the number of samples is larger than two (k>2), which is the basic concern in multiple comparisons. Important tests which have been proposed for this problem include, in particular, extensions of one- and two-sample goodness-of-fit tests, such as Kolmogorov-Smirnov, Cramér-von Mises and Anderson-Darling-type tests. The statistics used by such tests typically involve complex finite-sample and large-sample distributions which are difficult to compute and tabulate. The tests suggested have been tabulated only for fairly limited sets of cases in terms of sample sizes (mainly for samples of equal sizes) and levels, which can be the source of size and power problems. In this paper, we make three main contributions. First, we propose to use the technique of Monte Carlo tests in order to control the size of k-sample homogeneity tests, with an appropriate correction for the fact that the statistics considered may follow discrete null distributions. We show both theoretically and by simulation that the tests obtained in this way are exact irrespective of sample sizes. Second, we propose new test procedures for the k-sample homogeneity hypothesis, which are based kernel-type density estimators as well as procedures based on combining several tests. We show that the size of these new tests can again be easily be controlled through the Monte Carlo test technique. Thirdly, we show in a simulation experiment that the new tests suggested can largely outperform the existing procedures, in terms of size control, power and computational convenience.

Un problème important en statistique, en économétrie, en finance et en biologie consiste à tester l'hypothèse d'homogénéité de k échantillons ; c'est à dire si les observations de k échantillons différents proviennent de la même distribution. Dans cet essai, nous nous concentrons justement sur ce problème (k > 2). Les tests qui ont été proposés pour le traiter (Kolmogorov-Smirnov, Cramèr-von Mises et Anderson-Darling) ne sont que des prolongements de ceux utilisés dans le cas de deux échantillons. Les distributions des statistiques de ces tests sont cependant très compliquées. Il est en général très difficile de calculer les points critiques nécessaires à la conduite de ces tests. Dans cet essai, nous apportons trois contributions principales. D'abord, nous proposons l'utilisation de la technique de Monte Carlo, avec une correction appropriée pour les statistiques ayant, sous l'hypothèse nulle d'homogénéité dans le cas de k (k > 2) échantillons, des distributions discrètes avec un contrôle parfait des niveaux des tests. Nous montrons, théoriquement et par simulation, que les tests obtenus de cette façon sont exacts, indépendamment des dimensions des échantillons. En second lieu, nous proposons de nouvelles méthodes de tests basées sur les estimateurs de densités, fondés sur la méthode du noyau. Nous proposons alors deux tests combinés. Nous montrons que les niveaux de ces nouveaux tests sont facilement contrôlés par la méthode des tests de Monte Carlo. Troisièmement, nous prouvons, à l'aide d'une expérience de simulation, que les puissances des nouveaux tests suggérés peuvent largement dépasser celles des tests originaux.

Session/Séance 58 • Business and Economic Statistics • Statistique en affaires et en économie

Wednesday June 11 • Mercredi 11 juin 3:30 • 15h30 LSC 240

Talan Iscan, Dalhousie University; Fabio Ghironi, Boston College; Allessandro Rebucci, International Monetary Fund

Productivity shocks and consumption smoothing in the international economy • Les chocs de productivité et le lissage de la consommation dans l'économie internationale

This paper tests the significance of the net foreign assets in the transmission of productivity fluctuations. We develop a two-country general equilibrium model which allows for different steady-state net foreign asset positions across countries, distinguishes between country-specific and world productivity, and isolates their separate influences on consumption,and the foreign assets. We calibrate a state space solution of the theoretical model, and also empirically represent it as a cointegrated structural vector error correction model. We jointly estimate the structural consumption and the net foreign asset equations using panel data from G-7. The maximum liklihood procedure we employ allows for heterogeneity across countries, and imposes the short- and long-run restrictions implied by our theoretical model.

Cette présentation teste la signification des capitaux étrangers nets dans la transmission des fluctuations de productivité. Nous développons un modèle général d'équilibre de deux-pays qui tient compte de différentes positions nettes équilibrées de capitaux étrangers à travers les pays, distingue la productivité pays-spécifique et celle du monde et isole leurs influences individuelles sur la consommation et les capitaux étrangers. Nous calibrons une solution de l'espace d'état du modèle théorique, et la représentons empiriquement comme un modèle de vecteur structural de correction d'erreurs co-intégré. Nous estimons conjointement la consommation structurale et les équations capitaux nets étrangers en utilisant des données de panel du G-7. La méthode du maximum de vraisemblance que nous utilisons tient compte de l'hétérogénéité à travers les pays et impose des restrictions à court et à long terme requises par notre modèle théorique.

Wednesday June 11 • Mercredi 11 juin 4:00 • 16h00 LSC 240

Michael Foster, Canmac Economics; Leonard MacLean, Dalhousie University; William Ziemba, University of British Columbia

Empirical Bayes estimation with portfolio models • Estimation de Bayes empirique pour des modèles de portfolios

This paper considers the estimation of parameters in a dynamic stochastic model for securities prices, where the expected rate of return is a random variable. An empirical Bayes estimator is developed from the model structure. The estimator is an improvement on other popular estimators in terms of mean squared error. The effect of reduced estimation error on accumulated wealth is analyzed for the portfolio choice model with constant relative risk aversion utility.

Cette présentation considère l'estimation des paramètres dans un modèle stochastique dynamique pour les prix de valeurs mobilières, où le taux de rendement prévu est une variable aléatoire. Un estimateur de Bayes empirique est développé à partir de la structure du modèle. L'estimateur est une amélioration, par rapport à d'autres estimateurs populaires, en termes d'erreur quadratique moyenne. L'effet de la réduction de l'erreur d'estimation sur la richesse accumulée est analysé pour le modèle de choix de portfolio avec l'utilisation d'une aversion constante face au risque.

Wednesday June 11 • Mercredi 11 juin 4:30 • 16h30 LSC 240

Horand Gassman, Dalhousie University; I. Deak, T. Szantai, Technical University of Budapest

Generating multivariate normal probabilities • Générer des probabilités multinormales

This paper describes and compares several numerical methods for finding multivariate probabilities over a rectangle. The problem has recieved considerable attention in the

literature, including hundreds of theoretical papers and a variety of different computational approaches.Statistical applications include the multivariate probit model, the multivariate ordinal response model, and multivariate paired comparisons. There are also applications in stochastic programming, water resource management, and energy management. This paper was motivated by two considerations. On the one hand, a recent paper by Szantai (2000) developed several new bounding techniques whose computational qualities have not yet been fully explored. On the other hand, previous computational studies have tended to try to establish global superiority of one method over others, without regard to specifics of the problem such as the nature of the correlation matrix and the rectangle probability. The present paper will specifically address the question of how characteristics such as these will affect the efficiency of the methods tested. The methods compared fall into a variety of classes: There are applications of several multivariate integretation techniques, Monte Carlo methods, the bounding method of Szantai, and a recursive method first described by Plackett (1954). We briefly describe each of the methods. Numerical tests were conducted on approximately 10,000 problems generated randomly in up to twenty dimensions. We carefully describe the experimental set-up and give typical results. The best method found four-digit accurate probabilities for every twenty-dimensional problem in less than five minutes on a 650 MHz Pentium IV computer.

Dans cette présentation, nous décrivons et comparons plusieurs méthodes numériques pour trouver des probabilités multivariées sur un rectangle. Ce problème a reçu une attention considérable dans la littérature, y compris des centaines d'articles théoriques et une variété de différentes approches informatiques. Les applications statistiques incluent le modèle probit multivarié, le modèle de réponses ordinales multivariées et les comparaisons multivariées pairées. Il y a également des applications en programmation stochastique, en gestion de ressource de l'eau et en gestion de l'énergie. Cette présentation est motivée par deux considérations. D'une part, un article récent de Szantai (2000) a développé plusieurs nouvelles techniques de bornes dont les qualités informatiques n'ont toujours pas été entièrement explorées. D'autre part, les études précédentes tentent d'établir la supériorité globale d'une méthode sur les autres, sans souci des spécificités du problème tels que la nature de la matrice de corrélation et de la probabilité sur le rectangle. Cette présentation questionne spécifiquement comment ces types de caractéristiques affectent l'efficacité des méthodes testées. Les méthodes comparées entrent dans une variété de classes: il y a des applications de plusieurs techniques d'intégration multivariée, de méthodes de Monte Carlo, de la méthode de borne de Szantai et d'une méthode récursive d'abord décrite par Plackett (1954). Nous décrivons brièvement chacune des méthodes. Des tests numériques ont été effectués sur approximativement 10 000 problèmes de dimension inférieure ou égale à vingt produits aléatoirement. Nous décrivons soigneusement l'approche expérimentale et donnons des résultats typiques. La meilleure méthode a trouvé des probabilités précises à quatre décimales pour chaque problème de dimension vingt en moins de cinq minutes sur un processeur Pentium IV de 650 mégahertz.

Session/Séance 59 • Variable Selection • Sélection de variables

Wednesday June 11 • Mercredi 11 juin 3:30 • 15h30 LSC 242

Derek Bingham, University of Michigan

Bayesian screening designs • Design de discrimination bayésien

In the optimal design of experiments, there are numerous criteria concerning the quality of estimated parameters from the experiment. In screening experiments one is interested

in (i) identifying the best subset of effects explaining the data; and (ii) efficient paramter estimation. This work considers a different kind of optimality, seeking designs that best facilitate efficient discrimination between competing models. The approach is Bayesian, allowing information about models and effects (e.g., main effects, interactions, etc) to be incorporated into the design selection. A new design criterion is proposed. The criterion takes into account prior distributions used in the analysis of designed experiments, thereby creating a more integrated design/analysis framework. New optimal designs will be demonstrated for several examples and computational issues discussed.

Dans plans d'expériences optimaux, il y a de nombreux critères concernant la qualité des paramètres estimés. Dans les expériences de discrimination, nous sommes intéressé à (i) identifier le meilleur sous-ensemble d'effets expliquant les données; et (ii) l'estimation efficace des paramètres. Ce travail considère un type d'optimalité différent, nous cherchons les plans d'expériences qui facilitent le mieux la discrimination entre les modèles concurrents. Nous utilisons une approche bayésienne, permettant à l'informations sur les modèles et les effets (par exemple, effets principaux, interactions, etc..) d'être incorporée au choix du design. Nous proposons aussi un nouveau critère de design. Celui-si tient compte des distributions a priori utilisées dans l'analyse des plans d'expériences, donnant ainsi un cadre plus intégré au design et à l'analyse. De nouveaux designs optimaux sont montrés pour plusieurs exemples et nous discutons de certains problèmes informatiques.

Wednesday June 11 • Mercredi 11 juin 4:00 • 16h00 LSC 242

Mu Zhu, Hugh Chipman, University of Waterloo

Combinatorial optimization by parallel Darwinian evolution • L'optimisation combinatoire par l'évolution darwinienne parallèle

The evolutionary algorithm is a powerful tool that can be used to solve some difficult combinatorial optimization problems. Variable selection is a classic combinatorial problem in statistics, one that is becoming ever more important in modern data-mining applications, where the number of possible predictors is very large. Like simulated annealing, in order for the evolutionary algorithm to perform well, the algorithm must be fine tuned with great care, often a delicate and difficult task. I present an easy way to improve the performance of the algorithm by running several evolutionary paths in parallel. For variable selection, this allows us to identify the correct variables quite easily, but more generally, the idea has significant implications on how the evolutionary algorithm can be used more effectively in practice.

L'algorithme évolutionnaire est un outil puissant qui peut être utilisé pour résoudre quelques problèmes difficiles d'optimisation combinatoires. Le choix des variables est un problème combinatoire classique en statistiques, un qui devient de plus en plus important dans les applications modernes d'analyse exploratoire de données, où le nombre de prédicteurs possibles est très grand. Comme pour la méthode du recuit simulé, pour que l'algorithme évolutionnaire performe bien, l'algorithme doit être ajusté avec grand soin, ce qui est souvent une tâche sensible et difficile. Nous présentons une manière facile d'améliorer la performance de l'algorithme en faisant plusieurs chemins évolutionnaires en parallèle. Pour le choix des variables, ceci nous permet d'identifier facilement les bonnes variables, mais plus généralement, l'idée a des implications significatives sur la façon dont l'algorithme évolutionnaire peut être utilisé plus efficacement dans la pratique.

Wednesday June 11 • Mercredi 11 juin 4:30 • 16h30 LSC 242

John Dziak, Richard Li, Pennsylvania State University

Characterization and New Algorithm for Nonconvex Penalized Least Squares • Caractérisation et nouvel algorithme pour les moindres carrés pénalisés non convexes

Variable selection is fundamental to high dimensional statistical modeling. Fan and Li (2001) proposed a class of variable selection procedures via nonconcave penalized likelihood for likelihood based models and/or nonconvex penalized least squares for linear regression models. In this talk, I first present characterizations of nonconvex penalized least squares. Based on these characterizations, a new algorithm for finding the solution of the penalized least squares is proposed. The newly proposed algorithm is easily implemented, and overcome some drawbacks of the local quadratic approximation algorithm by Fan and Li (2001). The new algorithm is tested by numerical studies, and further illustrated by real data applications.

La sélection de variable est fondamentale pour la modélisation statistique en haute dimension. Fan et Li (2001) ont proposé une classe de procédures de sélection de variable, par l'entremise de vraisemblance pénalisée non concave, pour des modèles de vraisemblance et/ou par l'entremise des moindres carrés pénalisés non convexes pour des modèles de régression linéaire. Dans cette présentation, nous allons d'abord discuter des caractérisations des moindres carrés pénalisés non convexes. Nous proposons un nouvel algorithme pour trouver la solution des moindres carrés pénalisés basé sur ces caractérisations. Le nouvel algorithme peut facilement être implémenté et surmonte les inconvénients de l'algorithme par approximation quadratique locale de Fan et Li (2001). Le nouvel algorithme est testé à l'aide d'études numériques et est illustré avec des applications sur de vraies données.

Session/Séance 60 • Inference for Time Series and Other Models of Dependence • Inférence pour séries chronologiques et autres modèles de dépendance

Wednesday June 11 • Mercredi 11 juin 3:30 • 15h30 LSC 238

Gülhan Alpargu, University of Massachusetts; Pierre Dutilleul, McGill University

Efficient estimation and valid testing for stepwise linear regression with autocorrelated errors • Estimation efficace et test valide pour la régression linéaire pas à pas croissante avec erreurs autocorrélées

The questions of efficient estimation and valid testing are crucial in the stepwise procedure of selection of explanatory variables in quantitative linear models with autocorrelated errors, especially when the explanatory variables are of different nature (i.e., fixed versus random). In stepwise linear regression with an intercept and errors following a temporal AR(1) process, we have studied the efficiency of maximum likelihood, restricted maximum likelihood and two new estimation procedures called first differences and first-difference ratios (FDR), relative to ordinary least squares. We have also studied the validity of seven testing procedures to assess the significance of the slope of variable X_p to enter the model. In particular, we propose the FDR t-test with $n-q$ df and the modified t-test with $n*-q$ df on the partial correlation of Y and X_p given $X_1, ..., X_{p-1}$ when X_p is random, where $q = p + 1$ and $n*$ is Dutilleul's (1993) effective sample size in this case. Efficiency and validity were analyzed in a Monte Carlo study with $p = 2$. Results are discussed in relation to the nature, fixed versus random (purely random or autocorrelated), of X_1 and X_2, the

sample size and the magnitude and sign of the parameter of the error AR(1) process. An illustration with the environmental data that motivated this study is presented.

Les questions portant sur l'efficacité des estimateurs et la validité des tests sont cruciales dans la procédure de sélection pas à pas croissante des variables explicatives pour les modèles linéaires quantitatifs avec erreurs autocorrélées, spécialement lorsque les variables explicatives sont de nature différente (i.e., fixe versus aléatoire). En régression linéaire pas à pas croissante avec un intercept et des erreurs suivant un processus AR(1) temporel, nous avons étudié l'efficacité des estimateurs du maximum de vraisemblance, du maximum de vraisemblance restreint et de deux nouvelles méthodes appelées méthode des premières différences et méthode des rapports des premières différences (RPD), relativement à celle de l'estimateur des moindres carrés ordinaires. Nous avons également étudié la validité de sept procédures de test afin d'évaluer la signification de la pente de la variable X_p qui est candidate à entrer dans le modèle. En particulier, nous proposons le test t RPD avec $n-q$ degrés de liberté et le test t modifié avec $n*-q$ degrés de liberté sur la corrélation partielle de Y et X_p étant donné $X_1, ..., X_{p-1}$ lorsque X_p est aléatoire, où $q = p+1$ et $n*$ est la taille d'échantillon effective de Dutilleul (1993) dans ce cas. L'efficacité et la validité ont été analysées dans le cadre d'une étude de Monte Carlo où $p = 2$. Les résultats sont discutés en fonction de la nature, fixe versus aléatoire (purement aléatoire ou autocorrélée), de X_1 et X_2, de la taille d'échantillon, et de l'ordre de grandeur et du signe du paramètre du processus AR(1) des erreurs. Une illustration avec les données environnementales qui ont motivé notre étude est présentée.

Wednesday June 11 • Mercredi 11 juin 3:45 • 15h45 LSC 238

Pierre Dutilleul, Bernard Pelletier, McGill University; Gülhan Alpargu, University of Massachusetts

A simple modified F-test for multiple linear regression with autocorrelated random regressors and errors • Un simple test F modifié pour régression linéaire multiple avec régresseurs et erreurs aléatoires autocorrélés

Let $\{Y(s); s \in S\}$ and$\{X_1(s), \ldots, X_p(s); s \in S\}$ denote $p+1$ stochastic processes for which one partial realization is available for multiple linear regression: $Y(s) = a_0 + \sum_j a_j X_j(s) + E(s)$. A number of procedures have been proposed in the literature for testing the significance of such regression models with potentially autocorrelated random regressors and errors. These model-testing procedures are essentially asymptotic and likelihood- or score-based. In this paper, we present as an alternative a simple modified F-test that we show to be valid in finite samples. The modification is in the adjustment of the number of degrees of freedom of the denominator of the F-ratio statistic, using an extension of Dutilleul's (1993) effective sample size. The mathematical proof is given. Simulation results obtained in the geostatistical context of the linear model of coregionalization are presented, together with an illustration on environmental spatial data. Reference: Dutilleul, P. 1993. Modifying the t-test for assessing the correlation between two spatial processes. Biometrics 49:305-314.

Soient $\{Y(s); s \in S\}$ et $\{X_1(s), \ldots, X_p(s); s \in S\}p+1$ processus stochastiques pour lesquelles une réalisation partielle est disponible pour régression linéaire multiple: $Y(s) = a_0 + \sum_j a_j X_j(s) + E(s)$. Un nombre de procédures ont été proposées dans la littérature pour tester la signification de tels modèles de régression avec régresseurs et erreurs aléatoires potentiellement autocorrélés. Ces procédures de test de modèles sont essentiellement asymptotiques et basées sur la vraisemblance ou le score. Dans cet article, nous présentons en

guise d'alternative un simple test F modifié que nous montrons être valide en échantillons finis. La modification est dans l'ajustement du nombre de degrés de liberté du dénominateur de la statistique du F-ratio, en utilisant une extension de la taille d'échantillon effective de Dutilleul (1993). La démonstration mathématique est donnée. Des résultats de simulation obtenus dans le contexte géostatistique du modèle linéaire de corégionalisation sont présentés, ainsi qu'une illustration avec des données spatiales environnementales. Référence: Dutilleul, P. 1993. Modifying the t-test for assessing the correlation between two spatial processes. Biometrics 49:305-314.

Wednesday June 11 • Mercredi 11 juin 4:00 • 16h00 LSC 238

Mostafa Filali, Jarrar Oulidi, fsdm-Fès-Morocco

Determining the order and the differentiation coefficient of an ARI using resampling method • Déterminations de l'ordre et du coefficient de différenciation d'un ARI en utilisant la méthode de rééchantillonnage

We consider n observations from ARI(p,d) model with unknown p and d (p is order of autoregressive and d is order of differentiation). We propose a resampling procedure for estimating p and d. The classical criteria such as AIC and BIC for estimate p and d have a penalty factor specified (for AIC and BIC the factor of penalty for n observation is (2/n) and (2(Ln ln(n))/n) respectively). A resampling scheme is proposed to estimate an improved penalty factor conditional on the observations. This procedure produces a consistent estimate of p+d, Simulation results support the effectiveness of this procedure when compared with some of the traditional order selection criteria.

Nous considérons n observations d'un modèle AR(p,d) avec p et d sont inconnu (p l'ordre d'autorégressif et d l'ordre de différentiation). Nous proposons une méthode de rééchantillonnage pour estimer p et d. Les critères de sélection d'ordre comme AIC et BIC ont un facteur de pénalité prédéfini (pour AIC et BIC le facteur de pénalité pour n observations c'est (2/n) et (2(Ln ln(n))/n) respectivement). Une schéma de rééchantillonnage est proposé pour estimer un facteur de pénalité développé conditionnelle aux observations. Cette procédure nous produit un estimateur consistent de p+d. Et à l'aide des résultats de simulation nous montrons l'efficacité de cette procédure en la comparant a quelque autre critère de sélection d'ordre classique.

Wednesday June 11 • Mercredi 11 juin 4:15 • 16h15 LSC 238

Florin Cristian Gheorghe, Panait Andreea Mihaela, Ghita Constantin, Valahia University of Targoviste

Reliable intervals in the case of the depended observations • Les intervalles de confiance dans le cas des observations dependantes

The estimation theory for the depended variables was the object of a lot of theoretical studies. In this subject P. Billingsley's monograph(1961) and Ghe. Mihoc and V. Craiu's book (1972) represent excellent studies for the statistics inference in the depended variables case. A great number of problems on this subject have been already solved but the explicit determining of the estimation functions for the certain Markov chains often seen in practical applications still presents many difficulties. In this paper we consider an arranged selection compounded of n depended variables whose theoretical repartition has a certain probability law We have three parameters of estimation. So, we have a simple and constant Markov chain. For this case we estimate parameters. Also we present the construction of the

reliable intervals for the case when 2 ot the 3 parameters are known. The paper ends with an application in statistics control where we obtain the generalized limits of control.

La theorie d'estimation pour les variables dependantes a fait l'object pour numerouse etudes theoriques. En ce matiere la monographie de P. Billingsley(1961)et celle de Ghe. Mihoc et M.Craiu(1972)sont des etudes excellents pour l'inference statistique dans le cas des variables dependants. Un grand numero des problemes en ce qui concerne ce sujet ont ete deja solves mais la determination explicite des function d'estimation pour certains chaines Markov souvent dans les applications pratique presente encore beaucoup de difficultes. Dans ce travail nous considerons une selection ordonnee compose par n variabiles dependantes avec une repartition theorique qui a une certaine expression, avec trois parameters d'estimation.C'est un chaine Markov simple et constant. Pour ce cas nous calculons les parameters. Nous realisons aussi la construction des intervalles de confiance dans le cas ou deux du ce trois parameters sont connus. Le travail se fini avec une application dans le control statistique ou nous obtenons les limites de control generalises.

Wednesday June 11 • Mercredi 11 juin 4:30 • 16h30 LSC 238

Anwer Sager, Garian University, Lybia

Theories in linear regression • Théories en régression linéaire

Theory (1): this theory for analysis unexplained variance (SSE) SSE = SSEy + SSEx; SST = SSR + SSEy + SSEx the (SST) is total variance, (SSR) is explained variance, (SSEy) is sum of square error in the model, (SSEx) is sum of square of measurement error in the independent variable. Theory (2): If (Xi,Yi)....(Xn,Yn)is random sample and Y is dependent variable and X is independent variable and the relation between X and Y is linear relation, than: the average of X and Y it will equal the average of X and Y after reflection on the line of regression, so the reflection does not change the averages in the regression linear model.

Théorie (1): cette théorie traite de la variance non expliqué par l'analyse (SSE) SSE = SSEy + SSEx; SST = SSR + SSEy + SSEx où (SST) est la variance totale, (SSR) est la variance expliquée, (SSEy) la somme des carrés des erreurs dans le modèle, (SSEx) la somme des carrés des erreurs de mesure de la variable indépendante. Théorie (2): Si (Xi,Yi) ... (Xn,Yn) est un échantillon aléatoire, Y la variable dépendante et X la variable indépendant et la relation entre X et Y est une relation linéaire, alors: la moyenne de X et Y est égale à la moyenne de X et Y après réflexion sur la droite de régression. Ainsi la réflexion ne change pas les moyennes dans le modèle de régression linéaire.

Session/Séance 61 • Survey Methods Contributed Session VI: Estimation - Theoretical • Méthodes d'enquête VI: Estimation - théorie

Wednesday June 11 • Mercredi 11 juin 3:30 • 15h30 LSC 338

Sarjinder Singh, St. Cloud State University

On Farrell and Singh's penalized chi-square distance function in survey sampling • Sur la distance du khi-carré pénalisée de Farrell et Singh en sondage

We suggest here a few new calibration techniques which improves the estimators recently suggested by Farrell and Singh (2002, JSM-NY proceedings) by proposing penalized chi-square distance functions. Although Farrell and Singh have shown that the minimization

of penalized chi-square distance functions leads to cover several estimators of population total such as that due to Searls (1964, JASA), Singh and Srivastava (1980, Biometrika), and the famous unbiased ratio estimator of Hartley and Ross (1954, Nature), but the present investigation is much more wider and includes Farrell and Singh as a special case. We also suggest here to re-calibrate the already calibrated weights of Deville and Sarndal (1992, JASA), for large enough sample size, so that all the calibrated weights are positive. Such a technique leads to a new market of estimators, which may be sold any where in the world with the help of logics given in the present investigation. Following Singh and Arnab (2003,JSM-San Francisco), a new penalized chi-square distance function to deal with random non-response will be dicussed. The five estimators of Chaubey and Crisalli (1995, SSC-proceedings),and the work of Lundstrom and Sarndal (1999, JOS) will be shown as special cases. Another plus point of the proposed methodology will be that the calibrated response weights will be shown to depend upon the value of study variable or past information, unlike Lundstom and Sarndal's calibrated response weigts depends only on the auxiliary variable.

Nous suggérons ici quelques nouvelles techniques de calibration qui améliore les estimateurs récemment suggérés par Farrell et Singh (2002, comptes rendus de JSM-NY) en proposant des fonctions de distance du khi-carré pénalisée. Bien que Farrell et Singh ont montré que la minimisation des fonctions de distance du khi-carré pénalisée mène à couvrir plusieurs estimateurs du total de la population comme dû à Searls (1964, JASA), à Singh et à Srivastava (1980, Biometrika), et au célèbre estimateur sans biais du rapport de Hartley et de Ross (1954, Nature), mais la présente recherche est beaucoup plus large et inclut les résultats de Farrell et Singh comme un cas particulier. Nous suggérons également ici de recalibrer les poids déjà calibrés de Deville et Sarndal (1992, JASA), pour des échantillons assez grands, de sorte que tous les poids calibrés soient positifs. Une telle technique mène à un nouveau marché d'estimateurs, qui en peuvent être vendus partout dans le monde avec l'aide de la logique présenté dans cette présentation. Suivant Singh et Arnab (2003, JSM-San Francisco), une nouvelle fonction de distance du khi-carré pénalisée qui gère la non réponse aléatoire sera discutée. Les cinq estimateurs de Chaubey et Crisalli (1995, en cours dans SSC), et le travail de Lundstrom et de Sarndal (1999, JOS) sont montrés en tant que cas particuliers. Un autre point positif de la méthodologie proposée est que les poids calibrés de réponse dépendent des valeurs des variables à l'étude ou d'information passée, contrairement à Lundstom et Sarndal, où les poids calibrés de réponse dépendent seulement de la variable auxiliaire.

Wednesday June 11 • Mercredi 11 juin 3:45 • 15h45 LSC 338

Thierno Aliou Baldé, Norma Chhab-Alperin, Benoit Quenneville, Statistics Canada/Statistique Canada

A study on the predictive power of the Help Wanted Index • Étude sur le pouvoir de prévision de l'Indice d'Offre d'Emploi

This study concerns the predictive power of the Help Wanted index in comparison to macroeconomic variables such as employment, unemployment, etc. The index is based on "Help Wanted" advertisements placed in newspapers by employers to attract potential employees. More precisely, it gathers job advertisements from twenty-two major newspapers from metropolitan areas across Canada. The counts of 1996 are used as benchmarks and are thus indexed to 100. The index is computed monthly and published in the first (or second) week of the following month. This study has three principal parts. The first

is a graphical analysis comparing the historical series of the Help Wanted index with the level of employment, the rate of employment and the rate of unemployment. The goal is to identify the points of reversal of these series in order to highlight the lead of the Help Wanted index (and thus it's predictive power) over the other series. The second part is a statistical analysis of causality between the Help Wanted index and the other macroeconomic series. The approach used is that of Granger (1969) and Sims (1972). This approach brings out causality, as the prediction of a variable Y is made jointly from its past and another variable X. The question that arises is to know if the prediction of Y from the joint history is significantly better than the prediction made from its past alone. The last part consists of modeling the relation that would exist between the Help Wanted index and the variables under consideration, using transfer functions. This model would make it possible to predict and quantify the behaviour of these variables from the observed behaviour of the Help Wanted index.

Cette étude porte sur le pouvoir de prévision de l'indice d'offre d'emploi par rapport à quelques variables macroéconomiques telles l'emploi, le chomage etc. L'indice a été mis au point à partir des annonces d'offres d'emploi faites dans les journaux par les employeurs dans le but d'attirer de potentiels employés. Plus précisement, il regroupe les annonces classées de vingt deux journaux majeurs de régions métropolitaines à travers le Canada. Les comptes de l'année 1996 servent de données de référence et sont donc indexés à 100. L'indice est calculé une fois par mois et publié dans la première (ou deuxième semaine) du mois suivant la compilation. L'étude comporte trois parties principales: la première consiste en une analyse graphique de comparaison entre la série historique de l'indice d'offre d'emploi et les séries du niveau d'emploi, du taux d'emploi et du taux de chomage. Le but recherché est d'identifier les points de retournements de ces séries afin de chercher à mettre en évidence ¡¡ l'avance¿¿ de l'indice d'offre d'emploi (et donc son pouvoir de prévision) sur les autres séries. La deuxième partie est une analyse statistique de de la causalité entre l'indice d'offre d'emploi et les autres séries macroéconomiques. L'approche utilisée est celle de Granger (1969) et Sims (1972). Cette approche aborde la causalité entre de prédiction d'une variable Y à partir de son passé conjoint avec une autre variable X. La question qui se pose est de savoir si la prédiction de Y à partir de l'historique conjoint est significativement meilleure à la prédiction faite à partir de son seul passé. La dernière partie consiste en une modélisation, à l'aide de fonctions de transfert, de la relation qui existerait entre l'indice d'offre d'emploi et les variables considérées. ce modèle permettrait alors de prédire et de quantifier le comportement de ces variables à partir de celui observé sur l'indice d'offre d'emploi.

Wednesday June 11 • Mercredi 11 juin 4:00 • 16h00 LSC 338

Yong You, Jack Gambino, Statistics Canada/Statistique Canada

Hierarchical Bayes small area estimation with model determination and applications • Estimation de Bayes hiérarchique pour des petits domaines avec détermination du modèle et applications

In recent years, model-based approaches have been widely used in small area estimation to obtain efficient model-based small area estimates. Many area level and unit level models including linear and nonlinear mixed models have been proposed and used for various small area problems. The hierarchical Bayes (HB) approach with Gibbs sampling makes it possible to use many complex models for small area estimation. In this paper, we consider basic area level and unit level models with some important extensions under

a HB framework. Bayesian model choice and determination methods are also studied. Applications including household survey data analysis, unemployment rate estimation and census undercoverage estimation will be presented.

Ces dernières années, les approches basées sur des modèles ont été largement utilisées dans l'estimation sur de petits domaines pour obtenir des estimations efficaces. Beaucoup de modèle de niveau de secteur et de niveau d'unité comprenant les modèles mélangés linéaires et non-linéaires ont été proposés et utilisés pour différents problèmes de petit domaine. L'approche de Bayes hiérarchique (HB) avec l'échantillonneur de Gibbs permet l'utilisation de plusieurs modèles complexes pour l'estimation sur de petits domaines. Dans cette présentation, nous considérons des modèles de niveau de secteur et de niveau d'unité de base avec quelques prolongements importants sous un cadre de Bayes hiérarchique. Le choix bayésien du modèle et les méthodes de détermination sont également étudiés. Des applications comprenant l'analyse de données d'enquête sur les ménages, l'estimation du taux de chômage et l'estimation de la sous-couverture des recensements sont présentées.

Wednesday June 11 • Mercredi 11 juin 4:15 • 16h15 LSC 338

Roberto Gismondi, Italian National Statistical Institute

Optimal provisional estimation in longitudinal surveys • Estimation optimale de provision dans les sondages longitudinaux

EUROSTAT, the European Union statistical office, actually calculates and spreads out an overall EU retail trade monthly index, based on a weighted arithmetic mean of the single EU countries indexes. The delay of publication, that is about 60 days from the end of the reference month, is considered too large for operators, researchers and decision makers. For this reason, since 2001 a task force managed by EUROSTAT is planning a statistical strategy aimed at selecting, in each EU country, a particular sub-sample from the national sample currently used, on the basis of which a provisional quick index at the EU level could be calculated with a delay of about 30 days.

A technique has been applied to the monthly data available for year 2002, according to the stratification suggested by EUROSTAT. Results are quite good, being the sample mean significantly more precise - in terms of average difference between the sample means referred, respectively, to the ISTAT sample and the smallest EUROSTAT one - when using quasi-balanced samples instead of simple random sampling or systematic sampling.

EUROSTAT, le bureau des statistiques de l'Union Européenne, a calculé et répandu un index mensuel global du commerce au détail pour l'Union Européenne, basé sur une moyenne arithmétique pondérée des index de chaque pays de UE. Le délai de publication de ces index, qui est d'environ 60 jours après la fin du mois de référence, est considéré trop grand par les opérateurs, les chercheurs et les décideurs. Pour cette raison, depuis 2001 un groupe de travail contrôlé par EUROSTAT projette une stratégie statistique qui vise à choisir un sous-échantillon particulier des pays de l'Union à partir de l'échantillon national actuellement utilisé, sur la base duquel un index rapide temporaire au niveau européen pourrait être calculé avec un délai d'environ 30 jours. Présentement, chaque pays de l'UE évalue le degré d'anomalie entre les estimations rapides temporaires et les index définitifs, afin d'évaluer la qualité du système. Selon l'attribution optimale de Neyman, EUROSTAT a calculé que l'Italie, à partir de 2003, devrait utiliser pour les estimations rapides un sous-échantillon de 1929 entreprises au détail, parmi un panel d'environ 7200 représentants l'échantillon total mensuel (EUROSTAT, 2001). Un problème non trivial auquel nous avons dû faire face est

le choix de la technique de sélection du sous-échantillon, un sujet sur lequel EUROSTAT n'a donné aucunes recommandations particulières.

Wednesday June 11 • Mercredi 11 juin 4:30 • 16h30 LSC 338

Murlidhar Jutti, Statistics Canada/Statistique Canada

A two phase sampling approach for variance and design effect estimation in studying brain-drain from Canada to the U.S. • Une approche d'échantillonnage à deux étapes pour l'estimation de la variance et de l'effet de design pour étudier l'exode des cerveaux du Canada vers les États-Unis

The Canadian National Graduate Survey (NGS) collects data on graduates from Canadian Universities. Respondents to this survey, who have since moved to the United States were also followed up in a study of certain "brain-drain" issues. This follow-up produces various total and ratio estimates. Currently, variance estimation for survey estimates adopts the assumption of stratified single-phase random sampling.

Two-phase sampling designs offer a variety of possibilities for use of auxillary information. This paper examines the effect of alternative variance estimation and estimation of domains of interest which corresponds to the two-phase sample design when viewed from the NGS. Observations will be presented by comparing the estimates of design effects between the two methods.

L'enquête canadienne nationale sur les diplômés (NGS) rassemble des données sur les diplômés des universités canadiennes. Les répondants à ce sondage qui ont déménagé aux États-Unis depuis leur graduation ont été également suivis dans une étude sur certains problèmes " d'exode des cerveaux ". Cette dernière étude a donné diverses estimations des totaux et de certains ratios. Actuellement, l'estimation de la variance pour des estimés dans les sondages suppose l'échantillonnage aléatoire stratifié à une étape.

Les designs d'échantillonnage à deux étapes offrent une variété de possibilités pour l'usage d'information auxiliaire. Cette présentation examine l'effet de d'estimations alternatives de la variance et de l'estimation de domaines d'intérêts qui correspond au design à deux étapes lorsqu'on se place d'un point de vu du NGS. Nous présentons des observations en comparant entre les deux méthodes les estimations des effets de design.

Session/Séance 62 • Biostatistics Contributed Session III: Survival and Clustered Data • Biostatistique III: Données de survie et corrélées en grappes

Wednesday June 11 • Mercredi 11 juin 3:30 • 15h30 LSC 338

Arusharka Sen, Concordia University; Winfried Stute, Justus-Liebig University, Giessen, Germany

Efficient estimation under bivariate random censoring: independent components • Estimation efficace avec censure aléatoire bivariée: composantes indépendantes

We consider efficient estimation of linear functionals of a bi-variate distribution function when each component variable is subject to random censoring. This problem has a long history, and the well-known efforts in the literature are those by Dabrowska (Ann.Statist., 1988) and Van der Laan (Ann.Statist., 1996), which are not entirely satisfactory (non-monotonic (Dabrowska) or non-explicit (van der Laan)). We focus on the special case when the two components are independent. We propose a method which does not suffer

from the above drawbacks and compare it with the Dabrowska estimator in terms of efficiency (under independence). We also present an estimator in the general case, which is monotonic but not efficient.

Nous considérons l'estimation efficace de fonctions linéaires d'une fonction de distribution bivariée lorsque chaque variable des composantes est sujette à la censure aléatoire. Ce problème a une longue histoire et les efforts qui ont été fait dans la littérature sont de Dabrowska (Ann.Statist., 1988) et Van der Laan (Ann.Statist., 1996). Cependant, les résultats ne sont pas entièrement satisfaisants, soit non-monotone dans Dabrowska ou non-explicite dans Van der Laan. Nous nous concentrons sur le cas particulier où les deux composantes sont indépendantes. Nous proposons une méthode qui ne souffre pas des inconvénients mentionnés ci-dessus et nous la comparons à l'estimateur de Dabrowska en termes d'efficacité (sous l'hypothèse d'indépendance). Nous présentons également un estimateur dans le cas général qui est monotone mais non efficace.

Wednesday June 11 • Mercredi 11 juin 3:45 • 15h45 LSC 234

Xuewen Lu, University of Calgary; R.S. Singh, University of Guelph

On a partially linear single-index survival model • Sur un modèle de survie à index simple partiellement linéaire

The proportional hazards regression model usually assumes that the covariate has a log-linear effect on the hazard function. In this paper, we consider a partially linear single-index model with flexible covariate effects. We assume the baseline hazard function can be parameterized, while the risk function associated with covariates is modeled in a additive form having a nonparametric component plus a linear combination of covariates. The model includes the parametric proportional harards model and the "single-index" model. Using the local linear method, the estimates of the unknown parameters and the unknown covariate function are determined, and their asymptotic distributions obtained.

Le modèle de régression de taux de panne proportionnel suppose habituellement que la covariable a un effet log-linéaire sur la fonction de taux de panne. Dans cette présentation, nous considérons un modèle à index simple partiellement linéaire avec des effets flexibles des covariables. Nous supposons que la fonction de taux de panne de base peut être paramétrisée, alors que la fonction de risque associée aux covariables est modélisée selon une forme additive avec une composante non paramétrique et une combinaison linéaire des covariables. Le modèle inclut le modèle paramétrique de taux de panne proportionnel et le modèle à index simple. En utilisant la méthode linéaire locale, les estimations des paramètres inconnus et la fonction inconnue des covariables sont déterminées, et leurs distributions asymptotiques sont obtenues.

Wednesday June 11 • Mercredi 11 juin 4:00 • 16h00 LSC 234

M. Tariqul Hasan, Brajendra C. Sutradhar, Gary Sneddon, Memorial University of Newfoundland

Analysing longitudinal failure time data: generalised estimating equations approach • Analyse de données longitudinales de temps de bris: approche basée sur les équations d'estimations généralisées

Longitudinal survival data may comprise repeated failure times and a set of multidimensional covariates for a large number of individuals. In this set up, it is likely that the repeated failure times will be longitudinally correlated. It is of scientific interest to obtain

consistent as well as efficient estimates for the hazard ratio parameters, i.e. the effects of the associated covariates on the failure times. As the correlation structure of the failure times is however unknown in practice, it becomes extremely difficult to obtain efficient estimators for the hazard ratio parameters. As opposed to the existing frailty model based correlation structure (suitable for familial data), we introduce the observations driven longitudinal correlation structures appropriate for the repeated failure times, which are exploited to construct certain modified weighted estimating equations to get consistent and efficient estimates for the hazard ratio parameters. The proposed estimation methodology is illustrated by analysing failure time data on multiple tumour recurrence for patients with bladder cancer.

Les données longitudinales de survie peuvent comporter des temps de bris répétés et un ensemble de covariables multidimensionnelles pour un grand nombre d'individus. Dans cette situation, il est probable que les temps de bris répétés soient longitudinalement corrélés. Il est d'intérêt scientifique d'obtenir des estimés consistants et efficaces pour les paramètres de ratio de taux de panne, c.-à-d. les effets des covariables associés sur les temps de bris. Puisque la structure de corrélation des temps de bris est inconnue en pratique, il devient extrêmement difficile d'obtenir les estimateurs efficaces pour les paramètres du ratio de taux de panne. Par opposition à la structure de corrélation existante basée sur un modèle (appropriée aux données familiales), nous présentons les structures longitudinales de corrélation basées sur les observations appropriées aux temps de bris répétés, qui sont utilisés pour construire certaines équations d'estimations pondérées modifiées pour obtenir des estimations consistantes et efficaces pour les paramètres du ratio de taux de panne. La méthodologie d'estimation proposée est illustrée en analysant des données de survies sur la récurrence de tumeurs multiples pour des patients atteints du cancer de la vessie.

Wednesday June 11 • Mercredi 11 juin 4:15 • 16h15 LSC 234

Renjun Ma, University of New Brunswick, Fredericton

A random effects modelling approach to clustered ordinal outcomes with random cluster sizes • Une approche de modélisation à effets aléatoires pour des résultats ordinaux avec des grappes de tailles aléatoires

In developmental toxicity studies, potential developmental toxicants are administered to pregnant rodents. In analysis of ordinal outcomes such as (death, malformation, low normal), the little effects have usually been accounted for by introducing litter-specific random effects; however, the extra-variation arising from the random litter sizes is often ignored. We introduce a random effects approach to address both the intra-cluster effects and extra-variation arising from the random cluster sizes in the analysis of such ordinal outcomes. This approach is illustrated with the analysis of developmental toxicity data.

Dans les études de développement sur la toxicité, les toxiques à effets potentiels sur le développement sont administrés à des rongeurs qui attendent des petits. Dans l'analyse des résultats ordinaux comme la mort, la malformation, la pré-maturation, les petits effets sont habituellement expliqués en incluant des effets aléatoires spécifiques à chaque portée; cependant, la variation supplémentaire provenant des tailles aléatoires des portées est souvent ignorée. Nous présentons une approche à effets aléatoires pour adresser les effets intra-groupe et la variation supplémentaire provenant des tailles aléatoires des grappes dans l'analyse de tels résultats ordinaux. Cette approche est illustrée avec l'analyse de données de toxicité sur le développement.

Wednesday June 11 • Mercredi 11 juin 4:30 • 16h30 LSC 234

Shenghai Zhang, Mary E. Thompson, University of Waterloo

Estimators of variances and confidence intervals from clustered data • Estimateurs de la variance et intervalles de confiance à partir de données en grappes

Variance estimation is important for statistical inference. It is well known that the robust covariance matrix estimator (sandwich estimator) has achieved increasing use in the statistical literature especially with the growing popularity of generalized estimating equations. Its virtue is that it provides consistent estimates of the covariance matrix for mean parameter estimates even when the fitted parametric model for covariances is misspecified. However, the properties of the sandwich method other than consistency had been discussed very little until the recent work by Kauermann and Carroll (2001). We will provide a new variance estimator and compare it with other variance estimators: model based estimator, sandwich estimator and corrected sandwich estimator. Methods of constructing conference intervals based on these variance estimators and the generalized estimating function will also be discussed.

L'estimation de la variance est importante pour l'inférence statistique. Il est connu que l'estimateur robuste de la matrice de covariance (estimateur en sandwich) est de plus en plus utilisé dans la littérature statistique, particulièrement avec la popularité croissante des équations d'estimantions généralisées. Son avantage est qu'il donne des estimations consistantes de la matrice de covariance pour des estimations du paramètre de moyenne même lorsque le modèle paramétrique est ajusté pour un modèle avec covariances mal spécifiées. Cependant, les propriétés de la méthode en sandwich, autre que la consistance, ont été très peu discutées avant les travaux récents de Kauermann et Carroll (2001). Nous présentons un nouvel estimateur de la variance et le comparons à d'autres estimateurs de variance tels l'estimateur basé sur le modèle, l'estimateur en sandwich et l'estimateur en sandwich corrigé. Nous discutons également des méthodes pour construire des intervalles de confiance sur ces estimateurs de variance et la fonction d'estimation généralisée.

Wednesday June 11 • Mercredi 11 juin 4:45 • 16h45 LSC 234

Guangyong Zou, Allan Donner, University of Western Ontario

The asymptotic variance of the intraclass correlation coefficient in the case of arbitrary class sizes • La variance asymptotique du coefficient de corrélation intra-groupe dans le cas où les groupes sont de taille arbitraire

The intraclass correlation coefficient (ICC), a quantitative measure of the degree of similarity between individuals within groups (or clusters), has a lengthy history of application in several different fields of research. Consequently, inference procedures for the ICC have received considerable attention in the statistical literature, with theory that is well-developed for the case of continuous outcome data (e.g., Donner, 1986 International Statistical Review 54: 67–82). However, inference procedures for the ICC in the case of binary data are much less developed. Moreover, previous work has tended to focus on special cases, mainly in the context of interrater agreement studies (Kraemer, Periyakoil, and Noda 2002 Statistics in Medicine 21: 2109–2029). We obtain close-formed asymptotic variance formulas for three point estimators of the ICC considered by Ridout, Demetrio, and Firth (1999, Biometrics 55: 137–148) for the case of binary data in classes of arbitrary size. Two of the formulas obtained include results of Bloch and Kraemer (1989, Biometrics 45: 269–287) and Altaye, Donner, and Klar (2001, Biometrics 57: 584–588) as special cases.

Le coefficient de corrélation intra groupe (ICC), une mesure quantitative du degré de similitude entre les individus chez des groupes (ou grappe), a une longue histoire d'application dans différents domaines de recherche. En conséquence, les procédures d'inférence pour l'ICC ont suscité une attention considérable dans la littérature statistique, avec la théorie qui est bien développée pour le cas des données continues de réponse (par exemple, Donner, 1986 International Statistical Review 54: 67 – 82). Cependant, les procédures d'inférence pour l'ICC dans le cas de données binaires sont beaucoup moins développées. D'ailleurs, les travaux précédents ont tendance à se concentrer sur des cas particuliers, principalement dans le contexte d'études sur l'accord inter-juges (Kraemer, Periyakoil, et Noda 2002 Statistics in Medicine 21: 2109 – 2029). Nous obtenons des formules explicites de variance asymptotiques pour trois estimateurs ponctuel de l'ICC considéré par Ridout, Demetrio, et Firth (1999, Biometrics 55: 137 – 148) pour le cas de données binaires dans des classes de taille arbitraire. Deux de des formules obtenues incluent des résultats de Bloch et Kraemer (1989, Biometrics 45: 269 – 287) et Altaye, Donner, et Klar (2001, Biometrics 57: 584 – 588) en tant que cas particuliers.

Session/Séance 63 • Statisticians in Action III • Statisticiens en action III

Wednesday June 11 • Mercredi 11 juin **3:30 • 15h30** **LSC 332**

Video presentation • Présentation vidéo

Committee on Professional Development • Comité sur le perfectionnement professionnel

Index