STT 6516: Données catégorielles - Hiver 2024

Ce cours est une introduction aux modèles statistiques utilisés pour l'étude et l'analyse de données catégorielles nominales ou ordinales, ainsi que les données de comptage. Le cours couvre des sujets telles que les tableaux de contingence à plusieurs dimensions, le modèle log-linéaire, les modèles de régression logistique, de Poisson et multinomiale. Il adresse aussi des méthodes d'estimation et d'inférence basés sur la théorie asymptotique ou bayésienne. Les notions théoriques seront appliquées sur des exemples pratiques avec l'aide de l'ordinateur et du logiciel R.

Horaire:

Mardi 12h30–13h50	5183 Pav. André-Aisenstadt
Jeudi 12h30–13h50	5183 Pav. André-Aisenstadt

Professeur:

Alexandro Murua	4221 André-Aisenstadt	(514) 343–6987	alejandro.murua@umontreal.ca
Disponibilités:	Mardi	11h30 à 12:20	
	Jeudi	14h00 à 14:40	
	par rendez-vous	sur Zoom	

Livre à utiliser

[AA] Alan Agresti (2013). *Categorical Data Analysis.* 3ème édition, John Wiley & Sons. *(fortement recommandé).*

Des autres livres d'interêt

[AAI] Alan Agresti (2007). *An Introduction to categorical Data Analysis.* 2ème édition. John Wiley & Sons. *(lecture supplémentaire. Non requis.)*

[BFH] Yvonne M. Bishop, Stephen E. Fienberg et Paul W. Holland (2007). *Discrete Multivariate Analysis: theory and applications.* Springer. *(lecture supplémentaire. Non requis.)*

[LDL] Lafaye de Micheaux, P., Drouilhet, R. et Liquet, B. (2010). *Le logiciel R - Maitriser le langage - Effectuer des analyses statistiques*, Springer. `http://www.springerlink.com/content/978-2-8178-0114-8`. *(lecture supplémentaire. Non requis.)*

Sujets à voir

1. Semaine 1 : introduction, tableaux de contingence
2. Semaine 2 : tableaux de contingence à deux variables catégorielles
   Distribution des données (Poisson, multinomiale, multinomiale par lignes ou colonnes)
   Estimation par maximum de vraisemblance; distribution asymptotique
   Intervalles de confiance pour une probabilité
3. Semaine 3 : tests d'association
4. Semaine 4 : inférence conditionnelle; test exact de Fisher;
5. Semaine 5 : modèle de Rasch; valeur-p exact approximative par Monte-Carlo ou Monte-Carlo par chaînes de Markov
6. Semaine 6 : modèles log-linéaires (deux ou plusieurs variables catégorielles);
   modèles hiérarchiques

7. Semaine 7 : semaine d'activités libres
8. Semaine 8 : modèles log-linéaires, modèles hiérarchiques
9. Semaine 9 : modèles graphiques
10. Semaine 10 : modèle linèaire par linèaire; théorie asymptotique;
11. Semaine 11 : sélection d'un modèle (critères d'information AIC, BIC); sélection par lasso, lasso de fusion.
12. Semaine 12 : régression logistique; théorie asymptotique; rapport de cotes
13. Semaine 13 : régression logistique; sélection de variables; test d'adéquation
14. Semaine 14 : régression multi-logit; modèle multinomiale logistique cumulatif (variables ordinales)
15. Semaine 15 : modèles à effets aléatoires ou mixtes.

Évaluation

La note finale a trois composantes:

1. Les devoirs (30%) seront assignés, rassemblés, évalués, et retournés. Tout le travail sera dû pendant la journée de la date assignée avant 23h59. Le travail en retard ne sera pas accepté.

Les devoirs seront distribués selon le programme suivant:

Dvr#1	Jeudi 18 janvier	échéance : le jeudi 1er février
Dvr#2	Jeudi 1er février	échéance : le jeudi 15 février
Dvr#3	Jeudi 14 mars	échéance : le jeudi 28 mars

Chaque devoir aura le même poids dans l'evaluation final et sera évalué sur une échelle de 0 à 100 points.

2. L'Examen Intra (30%) sera annoncé le jeudi 22 février. Il consistera à des applications particulières de l'analyse de données catégorielles à des problèmes pratiques réels. Il exigera l'exploration de données et une compréhension claire des matières presentées dans la classe et des tâches de lecture. CHAQUE ÉTUDIANT(E) DOIT RÉPONDRE AUX QUESTIONS DE L'EXAMEN, INCLUANT L'ANALYSE DES DONNÉES, DE MANIÈRE INDIVIDUELLE. TOUTE COLLABORATION DE N'IMPORTE QUELLE SORTE EST STRICTEMENT INTERDITE. **Soumettez votre rapport avec l'analyse des données avant ou pendant la journée du jeudi 29 février.**

3. Les projets et présentations (40%). Chaque étudiant(e) devra présenter un projet sur un des sujets associés à l'analyse de données catégorielles (qui peut provenir d'un sujet vu ou à voir dans le cours, ou d'un article). Chaque projet consistira à

1. travailler avec une application aux données réelles.
2. faire un résumé écrit du projet (dans un maximum de six pages).
3. faire une présentation orale devant la classe pour exposer le projet.

Remarques:

(a) avant de choisir un projet, il faudra présenter une proposition écrite (d'environ une page) du sujet et l'envoyer au professeur avant **le 20 février.** Les dates des présentations seront fixées pour les dernières deux semaines du trimestre.

(b) Des copies numériques du rapport et de la présentation devront être soumis au professeur.

(c) Ce projet implique une étude très détaillée du sujet choisi, une compréhension élevée du sujet, et une étude qui va au delà d'une présentation et d'une application aux données: il faut que vous fassiez votre propre recherche dans le sujet choisi, c'est-à-dire, regarder d'autres alternatifs, et proposer et tester vos propres idées.

(d) Si nécessaire, je vous donnerai de travail extra requis pour finir le projet après la présentation et le rapport initial. Un rapport final sera aussi requis avant la fin du trimestre.

Présentation des devoirs

Les conditions suivantes simplifieront considérablement l'évaluation des devoirs et sont obligatoires.

1. En soumettant chaque devoir, mettez votre nom, le numéro du cours *et le numéro du devoir* sur la première page, comme suit:

   Votre Nom
   STT6516 - Hiver 2024
   Devoir #

2. Seul les devoirs lisibles seront acceptés et évalués.
3. Soumettez chaque devoir dans un document PDF au format lettre directement à l'adresse courriel du professeur.
4. **Les sorties d'un logiciel sans aucune annotation ne sont pas acceptable. Vous devez clarifier quels aspects des sorties d'ordinateur sont appropiées et vous devez montrer comment ils répondent aux questions posées dans le devoir. Des parties non pertinentes ou incorrectes des sorties d'ordinateur devraient être éliminées ou bien clairement biffées.**
5. **Soumettez les problèmes dans l'ordre donné.**
6. Organisez chaque devoir de sorte que les graphiques et leur discussion soient ensemble. **NE METTEZ PAS tous le graphiques à la fin du devoir.** Marquez quels graphiques sont assortis à quels problèmes.

Ces conditions aident à s'assurer que votre devoir soit évalué efficacement et dans le meilleur délai. Les règles qui ne sont pas suivis peuvent vous faire perdre des points.

Dates importantes

mardi 23 janvier	Date limite pour modifier le choix de cours
mardi 23 janvier	Date limite pour annuler un cours sans frais
Du vendredi 29 mars au lundi 1er avril	congé universitaire (férié)
lundi 27 février au dimanche 5 mars	Période d'activités libres
vendredi 15 mars	Date limite pour abandonner un cours (avec frais)
jeudi 22 février	Examen Intra et Projet pratique (à emporter)
jeudi 29 février	Écheance du rapport associé au projet pratique (examen intra)
mardi 16 avril	dernier jour de cours
Du lundi 28 novembre au lundi 12 décembre	Présentations
mardi 30 avril	Fin du trimestre

Veuillez lire ces messages importants

1. Les devoirs ne sont pas facultatives. Si vous manquez la date-limite pour soumettre le devoir, votre devoir recevra zero (0) points.
2. Le plagiat: attention, c'est sérieux! Vous êtes invité à consulter le site `www.integrite.umontreal.ca`
3. Bien que la discussion des problèmes des devoirs soit autorisée, chaque étudiant(e) est requis(e) de préparer et soumettre ses propres solutions (travail d'ordinateur y compris) aux devoirs. Des solutions préparées "en comité" ne sont pas acceptables. **La duplication des solutions des devoirs et des sorties d'ordinateur préparé entièrement ou partiellement par quelqu'un d'autre ne sont pas acceptables et sont considerées plagiats.** Si vous recevez l'aide de n'importe qui, vous devez dûment lui (leur) rendre reconnaissance dans votre rapport (example: "puisque les données sont toutes positives et leur distribution est asymétrique, une transformation logarithmique est clairement approprié dans la prochaine étape. Je remercie David Cox de m'indiquer ceci."). **La collaboration de n'importe quelle sorte sur des examens ou des projets est interdite.**
4. Vous avez l'obligation de motiver une absence prévisible à une évaluation dès que vous êtes en mesure de constater que vous ne pourrez pas être présent. Il appartiendra à l'autorité compétente de déterminer si le motif est acceptable (article 9.9).
5. Nous faisons bon accueil à des commentaires ou à des suggestions au sujet du cours à tout moment, soit par courriel, ou par visioconférence.
6. **Ce programme est prévu pour fournir une vue d'ensemble de STT6516. Vous ne pouvez revindiquer aucun droit de lui. En particulier, les dates d'examen peuvent changer. Tandis que le programme devrait être un guide assez fiable pour la session présente, les annonces officielles sont toujours ceux qu'on fait dans la classe.**

Ressources d'aide au DMS et à l'UdeM

N'hésitez pas à aller chercher de l'aide au besoin. Voici des ressources disponibles à l'Université de Montréal.

1. Le centre de santé et de consultation psychologique (CSCP) de l'Université de Montréal (`http://www.cscp.umontreal.ca/`). La prise de rendez-vous et l'inscription à un premier rendez-vous se font entièrement en ligne à l'adresse suivante : `https://monudem.umontreal.ca/.../Consultation`
2. Le Programme Mieux-être de l'ASEQ.
   Ligne téléphonique ouverte 24 heures / 7 jours : 1 833 851-1363
   Pour plus d'informations: `http://www.aseq.ca/.../FA%C3%89CUM_Programmedaide`
3. N'hésitez pas à contacter votre TGDE (tgdesup@dms.umontreal.ca) ou votre association étudiante (aemsum@dms.umontreal.ca) qui pourront vous guider.