Analyse des données quantitatives : des tableaux de fréquences à la régression logistique

DESCRIPTION GÉNÉRALE

Le séminaire d'une durée de deux jours introduit les participants à l'analyse des données quantitatives principalement par le biais de l'analyse du tableau de fréquences (tableau de contingence), de l'analyse de corrélation ainsi que de la régression multiple (linéaire et logistique).

Cette formation vise principalement la compréhension des méthodes d'analyse et l'interprétation des épreuves informatisées produites à l'aide du progiciel SAS. Les considérations mathématiques seront limitées à l'essentiel. À partir d'exemples tirés de l'exploitation de l'*Enquête sur la santé des collectivités canadiennes (ESCC)* réalisée par Statistique Canada, il s'agira de mettre en relief tout au long de la formation :

- Les conditions d'utilisation des procédures;
- Les précautions qui s'imposent lors de leur utilisation;
- La lecture et l'interprétation des épreuves informatisées;
- La programmation informatique nécessaire pour produire les épreuves informatisées commentées durant le séminaire.

PLAN DE LA FORMATION

Première journée – Avant-midi – 8h30 à 10h30

MODULE 1. EN GUISE D'INTRODUCTION À L'ANALYSE DES DONNÉES

On a trop souvent tendance à se pencher sur la mise en opération de l'analyse statistique une fois le terrain terminé, après la collecte des données. Or, l'analyse statistique n'est pas sans contraintes; elle exige, pour réaliser pleinement ses promesses, des données de qualité dont la nature sera à même d'atteindre les résultats escomptés. Dans cette perspective, il ne faut pas cantonner l'analyse statistique à une série d'opérations sans liens aucuns avec les premières étapes d'une recherche. Bref, l'analyse statistique se prépare.

1.1 QUE VOULEZ-VOUS FAIRE AU JUSTE?

- Mesurer l'indépendance des observations?
- Mesurer la différence des observations?
- Mesurer l'intensité de la relation entre les observations?
- Modéliser une relation en fonction d'une prévision?

1.2 LES CONDITIONS DE L'INFÉRENCE STATISTIQUE

- Le plan de sondage aléatoire
- Le plan de sondage aléatoire complexe
- Le plan de sondage complexe et l'effet de plan
- Le plan de sondage complexe et les progiciels statistiques (SAS et SPSS)

MODULE 2. CONSIDÉRATIONS PRATIQUES SUR LA FORME DE LA DISTRIBUTION

La collecte terminée, la saisie des données complétée, les questions que se posent généralement les chercheurs sont du type : « Que dois-je faire maintenant? À quelles techniques statistiques devrais-je m'en remettre?». Dans cette perspective, il est méthodologiquement sage de retarder le recours à un processus d'analyse spécifique (par exemple, l'analyse de variance ou la régression multiple) avant d'avoir acquis une familiarité suffisante de la nature des observations sur lesquelles se portera éventuellement cette analyse. Cette deuxième partie se penche ainsi sur ces étapes qui servent de contrôle de qualité permettant au chercheur de se familiariser avec ses données d'analyse, d'en reconnaître les forces, les faibles, les limites et lui permettant ainsi d'éviter de sérieuses erreurs.

2.1 LA FORME DE LA DISTRIBUTION ET SES INCIDENCES

- De l'utilité des indices de la forme
- Les indices de symétrie et de kurtose
- Asymétrie, kurtose et taille de l'échantillon

2.2 DES MESURES DE NORMALITÉ

- De l'importance d'une distribution normale
- Les tests de normalité
- La normalité à travers les graphiques
- Que faire puisqu'elle n'est pas normale? Faut-il transformer?

Première journée – Avant-midi – 10h30 à 12h00

MODULE 3. ANALYSE DES TABLEAUX DE FRÉQUENCES

Le tableau de fréquences (tableau de contingence) est sans contredit parmi les approches analytiques les plus souvent utilisées en recherche sociale. Or son exploitation efficace renvoie à des règles précises de confection et de lecture. Accompagnant presque invariablement les tableaux de contingence, le chi-carré constitue vraisemblablement la mesure statistique la plus utilisée en recherche sociale mais trop souvent, à tort. En effet, son exploitation efficace renvoie notamment à la nature et à la taille de l'échantillon et au niveau de mesure des variables. Un regard particulier sera posé sur les notions de signification statistique et de rapport de cotes (*odds ratio*).

3.1 UN TEST D'INDÉPENDANCE : LE CHI-CARRÉ
3.2 D'AUTRES CHI-CARRÉ ET LEURS CONDITIONS D'UTILISATION
3.3 LES MESURES D'ASSOCIATION FONDÉES SUR LE CHI-CARRÉ
3.4 L'INTÉRÊT D'UN CHI-CARRÉ NON « STATISTIQUEMENT SIGNIFICATIF »
3.5 LE RAPPORT DE COTES
3.6 LES TESTS *COCHRAN-MANTEL-HAENSZEL*
3.7 L'ANALYSE STRATIFIÉE

- Qu'est-ce qu'un facteur de confusion?
- Qu'est-ce qu'un effet d'interaction?
- Vérification de l'homogénéité d'association

Première journée – Après-midi – 13h30 à 16h00

MODULE 4. L'ANALYSE DE CORRÉLATION

L'analyse de corrélation renvoie pour l'essentielle à l'analyse des données métriques. Elle constitue un chemin logique fréquemment emprunté qui se situe au carrefour de l'analyse de l'indépendance et de la régression.D'ailleurs, l'analyse de corrélation demeure inextricable de l'analyse de régression. En effet, la recherche ne s'interroge pas seulement sur le type de la relation entre les phénomènes mais aussi sur la force, l'intensité de cette relation. L'*analyse de régression* vise à répondre à la première de ces préoccupations alors que l'analyse de corrélation répond à la seconde. Une attention particulière sera accordée à la notion de variance.

4.1 CONSIDÉRATIONS GÉNÉRALES SUR LA FORME DE LA RELATION
4.2 COMMENT FAIRE POUR VÉRIFIER LE DEGRÉ DE LINÉARITÉ DE LA RELATION?
4.3 LES TESTS DE TENDANCE
4.4 UNE MESURE DE NON-LINÉARITÉ : *ETA*
4.5 LE COEFFICIENT DE CORRÉLATION (R) DE PEARSON ET SES CONDITIONS D'UTILISATION
4.6 SIGNIFICATION STATISTIQUE, SIGNIFICATION PRATIQUE ET COEFFICIENT DE DÉTERMINATION
4.7 SEUIL DE SIGNIFICATION ET INTERVALLE DE CONFIANCE
4.8 D'AUTRES TAUX DE CORRÉLATION
4.9 CONDITIONS DE SURESTIMATION ET SOUS-ESTIMATION DE L'INTENSITÉ D'UNE RELATION

Deuxième journée – Avant-midi – 8h30 à 12h00

MODULE 5. LA RÉGRESSION LINÉAIRE (SIMPLE ET MULTIPLE)

La régression linéaire est l'une des techniques statistiques les plus utiles et l'une de celles qu'on emploie de plus en plus couramment dans le cas d'une variable dépendante continue. De plus, parce qu'on peut l'étendre au-delà des données bivariées en l'appliquant à une situation multivariée, la régression se révèle un outil très utile de la recherche sociale. L'analyse de régression multiple permet de construire une équation explicative d'un phénomène donné. On identifie alors les variables indépendantes les plus significatives, ce qui permet de «prédire» les comportements non mesurés directement.

5.1 L'OBJECTIF DE LA RÉGRESSION LINAIRE

5.2 DES PRINCIPES DE MODÉLISATION

- Principe d'étalement
- Principe de pertinence théorique
- Principe de simplicité

5.3 LES CONTEXTES DE LA MODÉLISATION

- Contexte d'exploration
- Contexte d'explication
- Contexte de prédiction

5.4 L'ORDRE D'ENTRÉE DES VARIABLES DANS LE MODÈLE

- La régression standard
- La régression hiérarchique
- La régression « pas-à-pas » (*stepwise*)
- La régression setwise

5.5 LECTURE DES SORTIES INFORMATIQUES

- Premier regard: l'ajustement global du modèle
- Deuxième regard: les paramètres de régression
- Troisième regard: vérification des postulats du modèle

Deuxième journée – Après-midi – 13h30 à 16h00

MODULE 6. LA RÉGRESSION LOGISTIQUE BINAIRE

La régression logistique permet d'estimer la force de l'association entre une variable qualitative dichotomique (binaire) dépendante et des variables qualitatives ou quantitatives indépendantes. La régression logistique peut être univariée mais son intérêt réside dans son utilisation multivariée. La régression logistique est un outil qui permet de mettre en relation des variables explicatives à une variable réponse dichotomique, c'est-à-dire qui ne peut prendre que deux valeurs, le cas classique étant celui d'une variable réponse (dépendante) binaire. Cette situation est fréquente dans divers champs d'application, particulièrement dans les sciences sociales.

6.1 QUEL EST L'OBJECTIF DE LA RÉGRESSION LOGISTIQUE?
6.2 À QUEL MOMENT DEVRAIT-ON RECOURIR À LA RÉGRESSION LOGISTIQUE?
6.3 POURQUOI UTILISER LA RÉGRESSION LOGISTIQUE PLUTÔT QUE LA RÉGRESSION LINÉAIRE?
6.4 DE LA RÉGRESSION LINÉAIRE À LA RÉGRESSION LOGISTIQUE : LA TRANSFORMATION LOGISTIQUE
6.5 LES POSTULATS DE LA RÉGRESSION LOGISTIQUE
6.6 LES CONSÉQUENCES DE LA VIOLATION DES POSTULATS
6.7 LA LECTURE DES TIRAGES INFORMATIQUES

- Premier regard: Quelques préalables à l'analyse du modèle
- Deuxième regard: L'ajustement global du modèle (*Goodness-of-fit*)
- Troisième regard: Le pouvoir discriminant du modèle
- Quatrième regard: L'importance relative des paramètres et paramétrisation
- Cinquième regard: Les rapports de cotes
- Sixième regard: Vérification des postulats de la régression logistique

  - La colinéarité
  - L'influence

6.8 LES EFFETS D'INTERACTION
6.9 UNE VARIABLE POLYNOMIALE

Troisième journée – Avant-midi – 8h30 à 10h30

SESSION DE TRAVAUX PRATIQUES DANS LE LABORATOIRE DU CIQSS
*** (les participants seront divisés en 2 groupes, soit un en avant-midi et un en après-midi)**

MODULE 7. FAMILIARISATION AVEC STATA

Dans le laboratoire du CIQSS, les participants se familiariseront avec le progiciel Stata. Après une présentation de la structure de l'interface et des différents types de fichiers Stata, les participants commenceront à travailler avec Stata. On passera en revue les commandes usuelles qui servent à manipuler les données (visualisation des données, opérateurs logiques et création de variables, création de labels, fusion de bases de données).

Troisième journée – Avant-midi – 10h30 à 12h00

MODULE 8. STATISTIQUES DESCRIPTIVES ET MODÈLES DE RÉGRESSION ÉLÉMENTAIRES

Les participants apprendront à produire des statistiques descriptives (moyenne, médiane, indicateurs de dispersion). Les différents types de tableaux seront présentés avec leurs avantages et leurs inconvénients (commandes TABLE, TABSTAT, SUMMARIZE, TABULATE.). Les graphiques seront également abordés (histogrammes, diagrammes en boîte, nuages de points). Enfin, les commandes relatives aux différents types de régression linéaire et logistique seront présentées.

Nous expliquerons également aux participants comment rechercher l'information relative à l'utilisation de Stata pour aller plus loin.

Troisième journée – Après-midi – 13h00 à 15h00

SESSION DE TRAVAUX PRATIQUES DANS LE LABORATOIRE DU CIQSS
*** (les participants seront divisés en 2 groupes, soit un en avant-midi et un en après-midi)**

MODULE 7. FAMILIARISATION AVEC STATA

Dans le laboratoire du CIQSS, les participants se familiariseront avec le progiciel Stata. Après une présentation de la structure de l'interface et des différents types de fichiers Stata, les participants commenceront à travailler avec Stata. On passera en revue les commandes usuelles qui servent à manipuler les données (visualisation des données, opérateurs logiques et création de variables, création de labels, fusion de bases de données).

Troisième journée – Après-midi – 15h00 à 16h30

MODULE 8. STATISTIQUES DESCRIPTIVES ET MODÈLES DE RÉGRESSION ÉLÉMENTAIRES

Les participants apprendront à produire des statistiques descriptives (moyenne, médiane, indicateurs de dispersion). Les différents types de tableaux seront présentés avec leurs avantages et leurs inconvénients (commandes TABLE, TABSTAT, SUMMARIZE, TABULATE.). Les graphiques seront également abordés (histogrammes, diagrammes en boîte, nuages de points). Enfin, les commandes relatives aux différents types de régression linéaire et logistique seront présentées.

Nous expliquerons également aux participants comment rechercher l'information relative à l'utilisation de Stata pour aller plus loin.