MATH 60604
Modélisation statistique
§ 4e - Tableaux de contigence

Léo Belzile

HEC Montréal
Département de sciences de la décision

Tableaux de contingence bidimensionnels

Le format le plus commun pour les données de dénombrement sont les **tableaux de contingence**, dans lesquels les dimensions sont les modalités des variables catégorielles et les cellules le décompte par sous-catégorie.

Soit X_1 et X_2 deux variables catégorielles avec respectivement J et K niveaux. Le nombre d'événement pour chaque sous-catégorie est représenté à l'aide du tableau de contingence

	$X_2 = 1$	$X_2 = 2$	$\cdots$	$X_2 = K$
$X_1 = 1$	Y_{11}	Y_{12}	$\cdots$	Y_{1K}
$X_1 = 2$	Y_{21}	Y_{22}	$\cdots$	Y_{2K}
$\vdots$	$\vdots$	$\ddots$	$\ddots$	$\vdots$
$X_1 = J$	Y_{J1}	Y_{J2}	$\cdots$	Y_{JK}

Test d'indépendance dans un tableau de contingence

- On considère deux modèles de Poisson: sous $\mathscr{H}_0$, le modèle avec les deux variables explicatives catégorielles, mais **sans interaction**. Pour $(j = 1, \ldots, J; k = 1, \ldots, K)$, le nombre moyen dans la cellule (j,k)

$$\mu_{jk} = \exp\left(\beta_0 + \alpha_j \mathbf{1}_{X_1=j} + \gamma_k \mathbf{1}_{X_2=k}\right).$$

  où $\alpha_1 = 0$ et $\gamma_1 = 0$ pour des raisons d'identifiabilité des paramètres.
- Le modèle sous l'alternative est le modèle saturé qui inclut en plus une interaction entre X_1 et X_2. L'hypothèse nulle d'indépendance revient à tester si les paramètres additionnels pour l'interaction sont zéros.

Modèle saturé et modèle avec effets principaux pour tableaux de contingence

- Sous $\mathscr{H}_0$, le modèle inclut seulement X_1 et X_2 (effets principaux).
  - On peut montrer que la valeur ajustée de ce modèle nul est simplement le produit des proportions par ligne/colonne.
  - On dénote la valeur ajustée pour la cellule (j,k) par $\widehat{\mu}_{jk}$.
- Le modèle saturé, sous l'hypothèse alternative $\mathscr{H}_1$, inclut des paramètres pour l'interaction.
  - le modèle saturé a $n = JK$ paramètres et la valeur ajustée est Y_{jk}.

Statistiques du test d'indépendance dans les tableaux de contingence

- La statistique du rapport de vraisemblance est la déviance

$$D = 2\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{jk} \ln\left(\frac{Y_{jk}}{\widehat{\mu}_{jk}}\right)$$

  qui suit une loi $\chi^2_{(J-1)(K-1)}$ sous l'hypothèse nulle d'indépendance.
- On peut également utiliser un test du score (avec la même loi asymptotique),

$$X^2 = \sum_{j=1}^{J}\sum_{k=1}^{K} \frac{(Y_{jk} - \widehat{\mu}_{jk})^2}{\widehat{\mu}_{jk}}.$$

Affiliation politique aux États-Unis

On considère un tableau de contingence deux par trois de l'affiliation politique d'Américain(e)s en fonction de leur sexe (données de 2000).

	Affiliation			
sexe	démocrate	indépendant	républicain	Total
femmes	762	327	468	1557
	(703.7)	(319.6)	(533.7)	
hommes	484	239	477	1200
	(542.3)	(246.4)	(411.3)	
total	1246	566	945	2757

Les valeurs entre parenthèses dénotent les valeurs ajustées du modèle Poisson avec les deux variables catégorielles, mais sans interactions (effets principaux).

Tableau 2,5 de Agresti (2007), *An Introduction to Categorical Data Analysis*, Wiley.

Résultats du test d'indépendance pour l'affiliation politique

- En ajustant le modèle sans interaction, on obtient la statistique du X^2 de Pearson et la déviance à même la sortie (30.07 et 30.02, respectivement).
- Si l'affiliation politique ne dépendait pas du sexe, ces statistiques devraient suivre approximativement une loi χ^2_2.
- Les valeurs-p sont inférieures à 10^{-4} et on rejette l'indépendance; l'affiliation politique dépend du sexe de l'individu.

Critères d'évaluation de l'adéquation

Critère	DDL	Valeur	Valeur/DDL
Ecart	2	30.0167	15.0083
Déviance normalisée	2	30.0167	15.0083
Khi2 de Pearson	2	30.0701	15.0351
Pearson normalisé X2	2	30.0701	15.0351