Anàlisi de dades categòriques (mostreig II)

Curs 2005-2006

Joan Valls i Marsal

Programa

1. Introducció
    - Breu història de l'anàlisi de dades categòriques
    - Tipus de dades categòriques
2. Descripció estadística de taules de contingència bidimensionals
    - Taules de contingència
    - Independència
    - Comparació de proporcions: diferència de proporcions, risc relatiu, odds ratio (raó d'avantatges). Relació entre la raó d'avantatges i el risc relatiu.
    - Mesures d'associació ordinals: Gamma de Goodman i Kruskal, Q de Yule, Tau-b de Kendall, d de Sommers.
    - Mesures d'associació nominals: Tau de Goodman i Kruskal o coeficient de concentració, coeficient d'incertesa.
    - Mesures d'acord. Kappa de Cohen.
    - Anàlisi de correspondències simples. Distància entre files i columnes. Representació en biplots.
3. Inferència per a taules de contingència bidimensionals
    - Distribucions mostrals: distribució de Poisson, distribució multinomial, distribució multinomial independent. Estudis prospectius i retrospectius.
    - Funcions de versemblança i estimació màxim versemblant.
    - Sobre dispersió
    - Testos de bondat d'ajust: per a una multinomial, exemple de Mendel, bondat d'ajust amb estimació de les freqüències esperades.
    - Testos d'independència: test de la khi-quadrat de Pearson, test de la raó de versemblances, invariància de la khi-quadrat per ordenacions de categories, particions de la khi-quadrat.
    - Intervals de confiança per a mostres grans: estimació de la raó d'avantatges, diferència de proporcions i risc relatiu.
    - Testos exactes per a mostres petites. El test exacte de Fisher. Altres testos.

4. La paradoxa de Simpson
    - Estructura de taula per a tres dimensions.
    - Associació parcial i marginal
    - Exemple de la pena de mort

5. Models de regressió logística
    - El model de regressió logística simple. Ajust del model. Testos per a la significació dels coeficients. Altres mètodes d'estimació
    - El model de regressió logística múltiple. Ajust del model. Testos per a la significació del model. Altres mètodes d'estimació.
    - Interpretació dels coeficients dels models de regressió logística. Una variable explicativa dicotòmica. Una variable explicativa politòmica. Una variable explicativa contínua. Combinacions multivariants. Interaccions i confusió. Estimació de odds-ratios en presència d'interaccions.
    - Estratègies per a la construcció del model. Selecció de variables. Mètode stepwise i mètode best subset.
    - Avaluació de l'ajust del model. Mesures de bondat de l'ajust: khi-quadrat de pearson i deviança, el test de Hosmer-Lemeshow i altres.

Bibliografia

- Agresti, A. Categorical data analysis. Wiley. 1990.
- Hosmer, D.W. i S. Lemeshow. Applied Logistic Regression. Wiley. 1989.
- Simonoff, J.S. Analyzing categorical data. Springer. 2003.
- Greenacre, M.J, Correspondence Analysis in Practice. Academic Press, London. 1993
- R Development Core Team (2003). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL http://www.R-project.org.
- Dalgaard, P. Introductory Statistics with R. Springer. 2002.
- Venables, W.N. i D.M. Smith. An introduction to R. www.R-project.org. 2003.
- Everitt, B. I S. Rabe-Hesketh. Analyzing medical data using S-PLUS. Springer. 2002.

Pràctiques

Es realitzaran dues hores setmanals de pràctiques. S'emprarà, principalment, el paquet estadístic R

Projecte d'anàlisi de dades

Els alumnes hauran de cercar un conjunt de dades de temàtica lliure, amb l'objectiu doble d'aplicar les eines de l'anàlisi estadística de dades categòriques i de presentar correctament els resultats obtinguts en un informe i extreure'n conclusions. Per a avaluar-lo es tindrà en compte:

1) Aplicació correcta de les tècniques estadístiques.
2) Originalitat i interès del problema i la seva resolució.
3) Claredat i brevetat en la presentació i redacció de resultats. Es penalitzaran aquells informes que siguin innecessàriament llargs (més de 15 pàgines).

Existirà la possibilitat d'una avaluació preliminar.

Avaluació

80% examen + 20% projecte d'anàlisi de dades

L'examen inclourà una part pràctica que s'haurà de resoldre amb l'ajut del paquet estadístic R.

Consultes

A concertar: joan.valls@iconcologia.catsalut.net

Cal que els alumnes que segueixin l'assignatura virtualment es posin en contacte amb el professor. Regularment s'aniran penjant de la pàgina web de l'assignatura materials diversos.

Horari i calendari

Les classes de teoria es realitzaran els dimarts de 17:00 a 19:00. Les classes de pràctiques amb ordinador es realitzaran els dilluns de 18:00 a 20:00.

La convocatòria ordinària tindrà lloc el dia 22 de juny, i la extraordinària el 5 de setembre.