Plan 551 PROGRAMA DE ESTUDIOS CONJUNTO DE GRADO EN ESTADÍSTICA Y DE GRADO EN INGENIERÍA INFORMÁTICA (INdat)
Asignatura 47102 ANÁLISIS DE DATOS CATEGÓRICOS

Tipo de asignatura (básica, obligatoria u optativa)

Obligatoria

Créditos ECTS

6

Competencias que contribuye a desarrollar

2.1
Generales

Comprensión y capacidad para la puesta en práctica del complejo proceso del análisis estadístico de datos, desde la formulación del problema, el diseño y recogida de datos, hasta el ajuste, análisis y validación de modelos estadísticos, así como la interpretación de resultados y presentación de los mismos; todo ello en muy diversos contextos de aplicación, como pueden ser las ciencias sociales, la epidemiología, las ciencias de la salud o la industria. Esta capacitación general se alcanzará como resultado de las distintas actividades y de las aplicaciones que en campos diversos y de forma coordinada se llevarán a cabo en las asignaturas de este bloque.
Capacidad para la aplicación práctica de modelos paramétricos de regresión de índole muy diversa: modelos de regresión lineal, ANOVA y ANCOVA; modelos log-lineales, logísticos y de poisson, así como otros modelos lineales generalizados y no lineales; modelos de Cox, de tiempo de fallo acelerado y de riesgo aditivo en el contexto del análisis de supervivencia o de la fiabilidad; modelos lineales para respuesta multivariante.
Capacidad para hacer una valoración del ajuste y diagnóstico, así como de la comparación de modelos con los procedimientos adecuados.
Capacidad para interpretar los resultados del ajuste de cualquiera de los modelos en el contexto de cada problema real de aplicación, ya sea a través de las inferencias sobre coeficientes, sobre las predicciones o sobre otros parámetros de interés.
Capacidad para manejar las técnicas adecuadas para la resolución de problemas específicos en el ajuste de modelos como pueden ser las transformaciones de Box-Cox, la heterocedasticidad, la multicolinealidad o la sobredispersión.
Capacidad para la utilización de programas de estadística a un nivel avanzado y para el desarrollo de métodos no implementados en los programas estándar de estadística. Capacidad para la resolución de problemas mediante técnicas de simulación y computación intensiva.
2.2
Específicas

Capacidad para el análisis de datos categóricos, ya sea la valoración del tipo de asociación en tablas de contingencia en diferentes condiciones, con el ajuste de modelos log-lineales, el ajuste de modelos logísticos u otros procedimientos específicos para respuesta catégorica, así como la interpretación de resultados.

Objetivos/Resultados de aprendizaje

Generales

- Aprender a reconocer problemas de respuesta discreta y a formular modelos estadísticos adecuados para su resolución.
- Aprender el manejo de paquetes de programas estadísticos, como R o SAS, para el Análisis de Datos Categóricos.
- Interpretar los resultados del ajuste de modelos para datos categóricos en problemas aplicados.
- Aprender a seguir los diferentes pasos del proceso que va desde la formulación del problema real por otros profesionales, hasta la solución estadística y su comunicación.

Específicos

- Manejar los métodos estadísticos más usuales en tablas de contingencia 2x2, especialmente la comparación de proporciones, riesgo relativo, razón de ventajas, test exacto de Fisher, test de McNemar.
- Conocer e interpretar los tipos de muestreo básicos asociados al estudio de tablas de contingencia, junto a las verosimilitudes asociadas y a los procedimientos de estimación y contraste subyacentes al ajuste de diferentes modelos.
- Conocer, aplicar e interpretar el test CMH en el análisis de la independencia condicional en tablas 2x2xK, así como calcular los estimadores de la OR común bajo asociación homogénea.
- Conocer la teoría básica del ajuste de modelos log-lineales en tablas de contingencia de diferentes dimensiones y sus aplicaciones al análisis de la asociación de variables categóricas.
- Conocer los fundamentos del ajuste de modelos logísticos para una respuesta dicotómica cuando se tienen variables explicativas de diferente índole, interpretando los parámetros del modelo, estimando probabilidades y otras cantidades de interés como la ED50, la sensibilidad o la especificidad de una prueba diagnóstica.
- Conocer, para una respuesta multinomial, la aplicación de modelos logit para respuesta nominal y de logits acumulativos para respuesta ordinal.
- Conocer el uso de modelos de regresión de Poisson: la verosimilitud, la estimación de parámetros y su interpretación, estimación de medias y valoración del ajuste del modelo.

Contenidos

1. Introducción a los problemas con respuesta categórica

- Reconocimiento de problemas diversos cuya solución requiere del ADC, mediante la observación de diferentes ejemplos.
- Lectura y manejo de diferentes tipos de datos categóricos mediante R y SAS. Creación de tablas de frecuencias y porcentajes.
- El método de Wald para obtener intervalos de confianza y contrastar hipótesis, y su aplicación a la estimación de una probabilidad.
- Aplicación de métodos de estimación basados en el TRV (o Deviance) y en el Score a la estimación de una probabilidad. Test chi-cuadrado.
- Problemas multiparamétricos.

2. Comparación de Proporciones y Tablas de Contingencia 2x2

- Diferentes tipos de estudios. Estimación en estudios prospectivos y retrospectivos. Causalidad y asociación.
- Estimación de la diferencia de dos probabilidades (en muestras independientes) y del "Riesgo Relativo" (RR) utilizando distribuciones asintóticas.
- La "Odds Ratio" (OR) o razón de ventajas y su relación con el RR.
- Interpretación de la OR e inferencias asintóticas sobre la misma.
- Utilidad de un diseño de muestras apareadas. Comparación de dos probabilidades (test de diferencia nula y estimación de la diferencia). Homogeneidad y Simetría en una tabla 2x2. Test de simetría de McNemar.
- Comparación de dos o más proporciones: Chi2 y TRV.

3. Tablas de Contingencia

- Tablas de Contingencia. Muestreos de Poisson, Multinomial y Multinomial producto. Otros tipos de muestreo.
- La función de verosimilitud y la estimación máximo verosímil.
- Relaciones entre las distribuciones al condicionar por las marginales.
- Interpretación del modelo de no asociación en los tres tipos de muestreo básicos y su expresión formal. Presentación de otros modelos: cuasi independencia, simetría, homogeneidad marginal…, modelo saturado y modelo nulo.
- Estimación máximo verosímil bajo no asociación.
- Tests de ajuste de un modelo: Test Chi2 y TRV (o Deviance). El AIC.
- Inferencias condicionales. Test exacto de Fisher.
- La paradoja de Simpson. ix. Tablas 2x2xK. Asociación condicional y marginal. OR condicional. Asociación homogénea.
- Test CMH de independencia condicional. Estimador MH de la asociación homogénea. Test de asociación homogénea.

4. Modelos Log-lineales en tablas IxJ

- Introducción a los modelos log-lineales.
- Diferentes codificaciones y su interpretación.
- El modelo log-lineal como un modelo lineal generalizado.
- Inclusión de efectos dependiendo del tipo de muestreo.
- Procedimientos para el ajuste de modelos log-lineales.
- Estimación de parámetros del modelo.
- Valoración del ajuste de modelos log-lineales. Cambio en la deviance para modelos anidados y el AIC.
- Ajuste de modelos adecuados a diferentes problemas: independencia, cuasi-independencia, simetría, cuasi-simetría, asociación uniforme, topológicos, efectos fila y/o columna,…

5. Modelos Log-lineales en tablas multidimensionales

- Modelos log-lineales en tablas tridimensionales.
- Diferentes tipos de asociación en una tabla IxJxK: Independencia, independencia parcial, independencia condicional, asociación homogénea. Modelos log-lineales asociados.
- Estimación máximo verosímil.
- Inclusión de efectos de las marginales fijadas.
- Ajuste de modelos log-lineales jerárquicos. Ruptura condicional de la deviance en modelos anidados.
- Alternativa al test CMH para contrastar la independencia condicional en tablas 2x2xK. Estimación de la OR común bajo asociación homogénea.
- Selección de un modelo log-lineal. Análisis secuencial de la deviance y eliminación de efectos. El AIC.

6. Modelos Logísticos

- Problemas de respuesta binaria y predictores categóricos. Modelos logit y su relación con los modelos log-lineales.
- Ajuste de modelos logísticos.
- La tolerancia en problemas de respuesta-dosis: modelos logístico, probit y clog-log. Relación con los modelos lineales generalizados.
- Interpretación de los parámetros del modelo logístico. Interacciones.
- Inferencias sobre los parámetros: EMV y su distribución asintótica, intervalos de confianza.
- Valoración del ajuste de modelos logísticos. Análisis de la deviance. El AIC. Análisis de residuos. vii. Calibración (estimación de la dosis efectiva).
- Predicción. Reglas de clasificación (sensibilidad, especificidad,…curva ROC).
- Ajuste de modelos logísticos en estudios retrospectivos (caso-control).
- Sobredispersión.
- Métodos exactos: inferencia condicional.
- Modelos para respuesta politómica.

7. Modelos de Poisson

- Regresión de Poisson.
- Estimación de parámetros.
- Ajuste y selección de un modelo.
- Sobredispersión. Alternativa binomial negativa.

Principios Metodológicos/Métodos Docentes

Clases: El profesor presentará problemas de distintos ámbitos de aplicación en los que se precisa la utilización de los métodos que el estudiante aprenderá a manejar en la asignatura. La teoría básica necesaria será expuesta en clase por el profesor de la asignatura y se ilustrará su aplicación mediante ejemplos. Esto hace difícil diferenciar claramente entre clases de teoría y clases prácticas. No obstante, se puede estimar que la "teoría" ocupará un 25% del tiempo total dedicado a las clases. Los estudiantes realizarán prácticas de ordenador en el Laboratorio de Estadística para familiarizarse con el manejo de R y SAS, y tendrán a su disposición los resultados de los análisis de diversos casos reales, cuya interpretación ocupará buena parte del tiempo dedicado a las clases.

Criterios y sistemas de evaluación

Realización de pruebas escritas en las que se evaluarán los contenidos prácticos de los temas desarrollados en las clases presenciales. 80%
Realización de trabajos y asistencia y participación en las clases, tutorías y seminarios. 20%
En total el estudiante debe alcanzar un mínimo de 5 puntos de un máximo de 10
El sistema de calificaciones a emplear será el establecido en el Real Decreto 1125/2003, de 5 de septiembre.

Responsable de la docencia (recomendable que se incluya información de contacto y breve CV en el que aparezcan sus lineas de investigación y alguna publicación relevante)

Agustín Mayo Iscar Agustin Mayo Iscar.
Tfno 983184170 email:agustinm@eio.uva.es
Licenciado en Ciencias Matemáticas. Doctor en Estadística e Investigación Operativa.
Profesor titular del departamento de Estadística e Investigación Operativa.
Línea de Investigación: Aplicación de procedimientos basados en recorte y restricciones para robustificar la obtención de clusters y la estimación de parámetros en modelos de mezcla de distribuciones y en modelos de localización y escala.
Alguna publicación
Cuesta-Albertos JA, Matrán C, Mayo-Iscar A. (2008) Trimming and likelihood: Robust Location and Dispersion in the Elliptical Model. Annals of Statistics, 36(5): 2284-2318.
Cuesta-Albertos JA, Matrán C, Mayo-Iscar A (2008) Robust estimation in the normal mixture model based on robust clustering. Journal of the Royal Statistical Society -Series B-Statistical methodology, 70:779-802.
Garcia-Escudero LA, Gordaliza A, Matrán C. Mayo Iscar A.(2008) A general trimming approach to robust cluster analysis. Annals of Statistics, 36(3): 1324-1345.
Garcia-Escudero LA; Gordaliza A; Matran C, Mayo Iscar A. (2011). Exploring the number of groups in model based

clusteringStatistics and Computing, 21, 585-599.
Fritz, H.; García-Escudero, L.A. y Mayo-Iscar; A. (2012) tclust: An R package for a trimming approach to Cluster Analysis. Journal of Statistical Software, Vol. 47, Issue 12.
García Escudero, L.A.; Gordaliza, A.; Matrán, C. y Mayo-Iscar, A. (2014). Avoiding spurious local maximizers in mixture modeling. Statistics and Computing, DOI 10.1007/s11222-014-9455-3.
García Escudero, L.A.; Gordaliza, A.; Matrán, C. y Mayo-Iscar, A. (2015). Avoiding spurious local maximizers in mixture modeling. Statistics and Computing, Vol. 25, Pag. 619-633.
García-Escudero, L.A.; Mayo Iscar, A; Sánchez-Gutiérrez, C.I. (2017). Fitting parabolas in noisy images. Computional Statistics and Data Analysis, Vol. 112, Pag. 80-87

Idioma en que se imparte

Castellano