Guía Docente

ANÁLISIS DE DATOS CATEGÓRICOS (47102)

Grado en Estadística 2016/2017

Curso: **4º** **1º Cuatrimestre**
Carácter: **Obligatorio** Créditos: **6**
Profesores:
Agustín Mayo Iscar (oficina A218)
email: agustinm@eio.uva.es; Tfno: +34 983 42 30 00 ext. 4170
José A. Menéndez Fernández (oficina A230)
e-mail: josan@eio.uva.es; web personal: http://www.eio.uva.es/~josan/
Tfno: +34 983 42 30 00 ext. 4169
Departamento responsable: **Estadística e Investigación Operativa** (web)
Facultad de Ciencias (web). Campus Miguel Delibes. Paseo de Belén, nº 7.
47011 Valladolid. España
Universidad de Valladolid (UVa)

INTRODUCCIÓN

¿Qué es el ADC? "Análisis de Datos Categóricos" es ya un término acuñado dentro de la Estadística Aplicada que describe una gran cantidad de modelos estadísticos que explican estructuras de datos en los que las variables respuesta son discretas, ya sean estas numéricas, nominales u ordinales.

¿Por qué el ADC? Porque es preciso dar respuestas adecuadas, basadas en criterios científicos, a preguntas como las siguientes:

¿Cuál es la proporción de individuos de una población que padece SIDA?, ¿Es el AZT efectivo en el desarrollo de los síntomas de SIDA?, ¿Tiene la aspirina un efecto protector sobre el infarto de miocardio?, ¿Fumar produce cáncer de pulmón?, ¿Cuál es el grado de satisfacción de los consumidores de Mahou?, ¿Qué relación existe entre el nivel de ingresos y el nivel de estudios?, ¿Cuál es la relación entre el consumo de alcohol, cigarrillos y marijuana?, ¿Qué dosis de cypermetrina debemos aplicar para reducir a la tercera parte la población de heliotis virescens?, ¿Cambia el status ocupacional de padres a hijos?, ¿Por qué ocurrió la catástrofe de la nave espacial Columbia?, ¿Qué variables, y en qué medida, determinan la gravedad de un paciente ingresado en la UCI?, … El estudiante aprenderá la metodología estadística básica necesaria para dar respuesta a preguntas como las anteriores, y a otras muchas que de forma similar se plantean en todas las ramas de la actividad humana.

La asignatura está orientada a las aplicaciones del ADC, y por ello una buena parte del trabajo que el estudiante tendrá que realizar será de índole práctico, mediante la utilización de herramientas informáticas y la interpretación de los resultados de los análisis que lleve a cabo, contribuyendo de ese modo a la adquisición del bagaje de "pensamiento estadístico" que todo profesional debe poseer.

OBJETIVOS

Generales

- Que el estudiante aprenda 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

- Que el estudiante aprenda a 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.
- Que el estudiante conozca 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.

Conocimientos previos requeridos: Es recomendable conocer los elementos básicos de Probabilidad e Inferencia Estadística, así como de Álgebra y Cálculo Infinitesimal. Asimismo, es recomendable la capacidad de leer inglés técnico.

BIBLIOGRAFÍA

Básica:

- Agresti, A. (2013). *An Introduction to Categorical Data Analysis*. Third Edition. Wiley.
- Collett, D. (2003). *Modelling Binary Data* (second edition). Chapman & Hall.
- Simonoff, J. S. (2003). *Analyzing Categorical Data*. Springer-Verlag.

Complementaria:

- Agresti, A (2002). *Categorical Data Analysis* (2nd. edition). Wiley.
- Hosmer, D. W. and Lemeshow, S. (1989). *Applied Logistic Regression*. Wiley.
- Le, C.T. (2010). *Applied Categorical Data Analysis and translational research* (2nd. edition). Wiley.

DEPARTAMENTO DE ESTADÍSTICA E I.O. - FACULTAD DE CIENCIAS- 47071 VALLADOLID - ESPAÑA
tfno.: 983 42 31 11 – fax: 983 42 30 13 – http://www.eio.uva.es/

- Tang, W., He, H. and Tu, X.M. (2012). *Applied Categorical and Count Data Analysis*. CRC Press.
- Tutz Gerhard (2012). *Regression for Categorical Data*. Cambridge University Press.
- Zelterman, D. (1999). *Models for Discrete Data*. Clarendon Press. Oxford.
- Manual de SAS: http://www.eio.uva.es/sasdoc/
- R Development Core Team (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.r-project.org/.

En la página web de la asignatura (http://www.eio.uva.es/~josan/) se dispondrá de parte del material que será utilizado a lo largo del curso.

PROGRAMA

1. Introducción a los problemas con respuesta categórica

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

2. Comparación de Proporciones y Tablas de Contingencia 2x2

i. Diferentes tipos de estudios. Estimación en estudios prospectivos y retrospectivos. Causalidad y asociación.
ii. Estimación de la diferencia de dos probabilidades (en muestras independientes) y del "Riesgo Relativo" (RR) utilizando distribuciones asintóticas.
iii. La "Odds Ratio" (OR) o razón de ventajas y su relación con el RR.
iv. Interpretación de la OR e inferencias asintóticas sobre la misma.
v. 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.
vi. Comparación de dos o más proporciones: Chi2 y TRV.

DEPARTAMENTO DE ESTADÍSTICA E I.O. - FACULTAD DE CIENCIAS- 47071 VALLADOLID - ESPAÑA
tfno.: 983 42 31 11 – fax: 983 42 30 13 – http://www.eio.uva.es/

3. Tablas de Contingencia

i. Tablas de Contingencia. Muestreos de Poisson, Multinomial y Multinomial producto. Otros tipos de muestreo.
ii. La función de verosimilitud y la estimación máximo verosímil.
iii. Relaciones entre las distribuciones al condicionar por las marginales.
iv. 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.
v. Estimación máximo verosímil bajo no asociación.
vi. Tests de ajuste de un modelo: Test Chi2 y TRV (o Deviance). El AIC.
vii. Inferencias condicionales. Test exacto de Fisher.
viii. La paradoja de Simpson.
ix. Tablas 2x2xK. Asociación condicional y marginal. OR condicional. Asociación homogénea.
x. Test CMH de independencia condicional. Estimador MH de la asociación homogénea. Test de asociación homogénea.

4. Modelos Log-lineales en tablas IxJ

i. Introducción a los modelos log-lineales.
ii. Diferentes codificaciones y su interpretación.
iii. El modelo log-lineal como un modelo lineal generalizado.
iv. Inclusión de efectos dependiendo del tipo de muestreo.
v. Procedimientos para el ajuste de modelos log-lineales.
vi. Estimación de parámetros del modelo.
vii. Valoración del ajuste de modelos log-lineales. Cambio en la deviance para modelos anidados y el AIC.
viii. 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

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

DEPARTAMENTO DE ESTADÍSTICA E I.O. - FACULTAD DE CIENCIAS- 47071 VALLADOLID - ESPAÑA
tfno.: 983 42 31 11 – fax: 983 42 30 13 – http://www.eio.uva.es/

6. Modelos Logísticos

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

7. Modelos de Poisson

i. Regresión de Poisson.
ii. Estimación de parámetros.
iii. Ajuste y selección de un modelo.
iv. Sobredispersión. Alternativa binomial negativa.

DEPARTAMENTO DE ESTADÍSTICA E I.O. - FACULTAD DE CIENCIAS- 47071 VALLADOLID - ESPAÑA
tfno.: 983 42 31 11 – fax: 983 42 30 13 – http://www.eio.uva.es/

METODOLOGÍA

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.

Trabajos: Los estudiantes realizarán dos trabajos propuestos por el profesor, en los plazos que se indicarán oportunamente. El informe de cada trabajo deberá ir firmado por su autor o autores. En el caso de que haya varios autores, cada uno de ellos deberá presentar en uno o dos folios un resumen personal explicando el trabajo realizado y sus aportaciones principales al mismo. El informe será revisado y valorado por el profesor, tanto en contenidos como en presentación, pudiendo ser requeridas de los alumnos cuantas explicaciones se consideren oportunas. Cada estudiante tendrá acceso a su informe, debidamente revisado y valorado.

Exámenes Parciales: Se realizarán dos exámenes parciales de una hora de duración.

Examen Final: 16 de Enero de 2017. (recuperación el 3 de Febrero)

Tutorías: Las tutorías individualizadas podrán ser atendidas los lunes, martes y jueves de 16:00 a 18:00, dentro del período lectivo, en el Departamento de Estadística. Fuera del horario anterior podrá consultarse al profesor previa cita con el mismo.

Calendario de Actividades: ver última página del documento.

Horario reservado: Lunes de 13 a 14 h.; Miércoles de 10 a 12 h.; Viernes de 9 a 10 h.

EVALUACIÓN

La evaluación se hará de la siguiente forma:

Denotemos por **T1** y **T2** las notas en cada uno de los dos trabajos, por **P1** y **P2** las notas en cada uno de los exámenes parciales, y por **EF** la nota en el examen final de la convocatoria ordinaria. Las notas se darán en una escala de 0 a 10.

Consideremos las ponderaciones: $\mathbf{T}=0.10*\mathbf{T1}+0.15*\mathbf{T2}$; $\mathbf{P}=0.10*\mathbf{P1}+0.15*\mathbf{P2}$; $\mathbf{F}=0.5*\mathbf{EF}$;

Si $\mathbf{T}\geq 1$, $\mathbf{P}\geq 1$ y $\mathbf{EF}\geq 3$, entonces la calificación final será $\mathbf{C}=\max(\mathbf{T}+\mathbf{P}+\mathbf{F}, \mathbf{EF})$.

En otro caso, la calificación final será $\mathbf{C}=\mathbf{EF}$.

Caso de suspender en la convocatoria ordinaria, será posible la recuperación en el examen extraordinario. En este caso, la calificación final de la asignatura será la del examen extraordinario.

DEPARTAMENTO DE ESTADÍSTICA E I.O. - FACULTAD DE CIENCIAS- 47071 VALLADOLID - ESPAÑA
tfno.: 983 42 31 11 – fax: 983 42 30 13 – http://www.eio.uva.es/

ANÁLISIS De DATOS CATEGÓRICOS (47102) Grado en ESTADÍSTICA

Cronograma de Actividades

Horario: Lunes de 13 a 14 h.; Miércoles de 10 a 12 h.; Viernes de 9 a 10 h.
Laboratorio: 308
Fechas de entrega de los trabajos: **T1, T2**.
Exámenes Parciales: **P1, P2**.

2016 SEPTIEMBRE

semana	Lunes	Martes	Miércoles	Jueves	Viernes
1	5	6	7	8	9
2	12	13	14	15	16
3	19	20	21	22	23
4	26	27	28	29	30

2016 OCTUBRE

semana	Lunes	Martes	Miércoles	Jueves	Viernes
5	3	4	5	6	7
6	10	11	12	13	14
7	**17 P1**	18	19	20	21
8	**24 T1**	25	26	27	28

2016 NOVIEMBRE

semana	Lunes	Martes	Miércoles	Jueves	Viernes
9	31	1	2	3	4
10	7	8	9	10	11 S Alberto
11	14	15	16	17	18
12	21	22	23	24	25
13	28	29	30	1	2

2016 DICIEMBRE

semana	Lunes	Martes	Miércoles	Jueves	Viernes
14	5	6	**7 P2**	8	9
15	**12 T2**	13	14	15	16
16	19	20	21	22	23
	26	27	28	29	30

2017 ENERO

	Lunes	Martes	Miércoles	Jueves	Viernes
	2	3	4	5	6
17	9	10	11	12	13
18	**16 ADC-exafin**	17	18	19	20
	23	24	25	26	27
	30	31	1	2	**3 ADC-extra**

Las fechas propuestas en el cronograma pueden verse alteradas por la necesidad de realizar algún cambio no previsto en el momento de su elaboración.

DEPARTAMENTO DE ESTADÍSTICA E I.O. - FACULTAD DE CIENCIAS- 47071 VALLADOLID - ESPAÑA
tfno.: 983 42 31 11 – fax: 983 42 30 13 – http://www.eio.uva.es/