

2016365-1: Estadística descriptiva multivariada

Programa del curso I-11

Martes y jueves: 2 a.m. a 4 p.m.

Campo Elías Pardo E-mail: cepardot@unal.edu.co

Web: <http://www.docentes.unal.edu.co/cepardot/>

Consultas: martes y jueves 16 a 18. Miércoles y viernes 11 a 12. Oficina: 404-325

4 créditos = 4 horas presenciales y 8 horas de trabajo del estudiante semanalmente.

Prerrequisitos

- Para estudiantes de la Carrera de Estadística: Álgebra matricial
Los estudiantes de la carrera de estadística aprovechan mejor esta asignatura si han tomado los cursos: Comunicación (español), Bases de datos, Diseño y desarrollo de encuestas y Metodología.
- Para estudiantes de otras carreras
 - Los cursos básicos de estadística propios del plan de estudios.
 - El curso de matemáticas que contenga los temas de álgebra lineal.

Descripción de la asignatura

Aborda el análisis descriptivo y exploratorio multivariado de tablas grandes de datos (muchas filas y columnas). Recurre a la representación geométrica multidimensional de las tablas de datos y a su lectura mediante proyecciones en planos, denominados factoriales, y a la conformación de grupos homogéneos en el sentido de variabilidad baja dentro de los grupos y alta entre grupos. Las representaciones geométricas permiten visualizar la información relevante contenida en las tablas de datos. El aprendizaje se consolida mediante una aplicación real siguiendo las pautas de la metodología de la investigación.

El estudiante que curse la asignatura y cumpla con las exigencias académicas, podrá:

- Aplicar los métodos empleando programas de uso libre y comercial y utilizarlos en el contexto de situaciones específicas.
- Abordar el aprendizaje de otros métodos de la estadística exploratoria multidimensional.
- Consolidar los conocimientos sobre metodología de la investigación.
- Mejorar las destrezas en redacción de textos, normas de presentación de trabajos escritos y de presentación oral de los resultados.

Contenido

1. Introducción (semana 1)

Métodos estadísticos exploratorios multidimensionales (Lectura: introducción de Lebart, Morineau & Piron (1995)). Lectura en línea del manual de introducción a R.

2. Representación multivariada de datos. (semana 2)

Repaso de espacios euclidianos multidimensionales (espacios vectoriales en \mathbb{R}^n con producto interno). Representación geométrica de tablas de datos. Significado de las estadísticas básicas y de las operaciones de centrado y reducido. Inercia y contribuciones a la inercia de filas y columnas. Proyección sobre cualquier eje. Contribución de las filas y columnas a la inercia proyectadas sobre un eje. Calidad de la representación sobre un eje.

3. Análisis en componentes principales - ACP. (semana 3 a 5)

Objetivos del ACP. Espacio de los individuos. Espacio de las variables. Relaciones entre los dos espacios. Ayudas para la interpretación de los ejes factoriales. Proyección de elementos ilustrativos. ACP generalizado. Análisis en coordenadas principales.

Primer parcial (marzo 15).

4. Análisis de correspondencias simples - ACS. (semana 6 a 8)

Objetivos del ACS Tablas obtenidas de la tabla de contingencia. Representación geométrica de las tablas de perfiles. El ACS como dos ACP.

5. Análisis de correspondencias múltiples - ACM. (semanas 9 a 11)

Objetivos del ACM. Codificaciones de las variables cualitativas. El ACM como el AC de la tabla disyuntiva completa. El ACM como el AC de la tabla de Burt. ACM y ACS en el caso de dos variables.

Segundo parcial (abril 28)

6. Métodos de clasificación (agrupamiento) (semanas 12 a 14)

Objetivos de los métodos de agrupamiento. Índices de similitud, disimilitud y distancia entre individuos. Inercia intra y entre grupos. Clasificación alrededor de centros móviles. Clasificación jerárquica aglomerativa. El método de Ward. Caracterización de las clases. Combinación de métodos factoriales y de clasificación.

Tercer parcial (mayo 24)

7. Presentación de trabajos de aplicación (semana 15 y 16, **mayo 26 y 31 y junio 1**).

Metodología

Este es un curso de aprendizaje asistido. Las clases se utilizan para aclarar dudas, presentar avances de los trabajos y controlar el aprendizaje. Durante el curso se realizan varios talleres y es recomendable hacerlos en grupo (los mismos del trabajo de curso). Los talleres no tienen calificación pero se recomienda realizarlos por escrito para adquirir destreza en la preparación de informes utilizando las herramientas de edición y las normas de presentación de trabajos. El curso cuenta con un aula virtual donde se pone el programa los talleres, documentos del curso, enlaces de interés, etc.

R. Los paquetes de R que se utilizan en este curso son: **FactoMineR**, **ade4** y **FactoClass**. Para evitar la proliferación de virus y el uso más eficiente de R se recomienda utilizar el sistema operativo Linux. En la sala esta instalada la distribución Ubuntu que es bastante amigable.

Software comercial

Este semestre disponemos del SPAD versión 7 en inglés, instalado en la sala y se utilizará para los talleres en clase (15 licencias simultáneas). Los estudiantes pueden utilizar cualquier software comercial para los talleres y trabajos, siempre que la Universidad tenga licencia, tales como: SAS, SPSS, Xstat (en Excel).

Calificación

- Tres parciales 20 % c/u : 60.
- Trabajo de investigación utilizando los métodos aprendidos 40 %: propuesta: 10 % y final: 30 %. El trabajo se debe hacer en grupo de 3 estudiantes. Se recomienda usar \LaTeX con el tipo de documento `\documentclass[report]{revcoles}`. La plantilla está disponible en:
<http://www.docentes.unal.edu.co/eccubidesg/docs/LaTeX/Report.zip>.

Textos guía

Lebart et al. (1995), Escofier & Pagès (1992) y Langrand & Pinzón (2009).

Documentos: Cabarcas & Pardo (2001), Pardo & Ortiz (2004), Pardo (2005, 2008), Bautista (1990), Bautista (1994), Morineau & Aluja (1994).

Referencias

- Bautista, L. (1990), 'Introducción al análisis multivariado de datos', Folleto Coloquio Distrital de Matemáticas y Estadística, Bogotá.
- Bautista, L. (1994), 'Métodos de clasificación', Folleto Simposio de Estadística [sobre] Análisis multivariado de datos, Bogotá.
- Benzecri, J. (1992), *Correspondence Analysis Handbook*, Marcel Dekker.
- Cabarcas, G. & Pardo, C.-E. (2001), 'Métodos estadísticos multivariados en investigación social', *Simposio de Estadística*.
*<http://www.docentes.unal.edu.co/cepardot/docs/SimposiosEstadistica/>
- Escofier, B. & Pagès, J. (1992), *Análisis factoriales simples y múltiples. Objetivos, métodos e interpretación*, Universidad del País Vasco, Bilbao.
- Greenacre, M. (2007), *Correspondence Analysis in Practice*, 2 edn, Chapman & Hall, Boca Raton, FL.
- Jambu, M. (1983), *Cluster Analysis and Data Analysis*, North-Holland, Amsterdam.
- Langrand, C. & Pinzón, L. M. (2009), *Análisis de datos. Métodos y ejemplos*, Editorial Escuela Colombiana de Ingeniería, Bogotá.
- Lebart, L., Morineau, A. & Piron, M. (1995), *Statistique exploratoire multidimensionnelle*, Dunod, Paris.
- Lebart, L., Morineau, A. & Warwick (1984), *Multivariate Descriptive Statistical Analysis*, Wiley, New York.
- Morineau, A. & Aluja, T. (1994), 'Análisis de correspondencias', Folleto Simposio de Estadística [sobre] Análisis multivariado de datos, Bogotá.
- Pardo, C. E. (2005), Análisis de correspondencias de tablas de contingencia estructuradas, in 'Memorias Coloquio Distrital de Matemáticas y Estadística', Universidad Distrital, pp. 65–90.
*<http://www.docentes.unal.edu.co/cepardot/docs/ColoquioDistritalMatEst/AnalCorresTCE.pdf>
- Pardo, C. E. (2008), 'Geometría euclidiana en estadística: métodos en ejes principales'.
*<http://www.docentes.unal.edu.co/cepardot/docs/Conferencias/ACPgeometriaEuclidiana.pdf>
- Pardo, C. E. & Ortiz, J. (2004), Análisis multivariado de datos en R, in 'Simposio de Estadística', Universidad Nacional de Colombia. Departamento de Estadística, Cartagena.
*www.docentes.unal.edu.co/cepardot/docs/SimposiosEstadistica/PardoOrtiz04.pdf
- Peña, D. (2002), *Análisis de datos multivariantes*, McGraw-Hill, Madrid.