



UNIVERSIDAD DE BUENOS AIRES
FACULTAD DE FILOSOFIA Y LETRAS

DEPARTAMENTO: LETRAS

SEMINARIO: Introducción al análisis estadístico de datos
lingüísticos con R

PROFESOR: Mag. Carolina A. Gattei

CUATRIMESTRE: 1ero

AÑO: 2016

PROGRAMA N°:

UNIVERSIDAD DE BUENOS AIRES
FACULTAD DE FILOSOFÍA Y LETRAS
DEPARTAMENTO DE LETRAS
SEMINARIO: Introducción al análisis estadístico de datos lingüísticos con R
PROFESOR: Carolina A. Gattei
1er Cuatrimestre DE 2016
PROGRAMA N°

1. Fundamentación y descripción

El campo de la investigación en lingüística aplicada requiere frecuentemente la recolección y análisis de datos lingüísticos. El análisis de datos de un corpus, la existencia de diferencias en tiempos de respuesta, la cuantificación de ocurrencias de distintos fenómenos lingüísticos en una comunidad de habla, son todos fenómenos que presentan un desafío a la hora de abordar cuantitativa y cualitativamente nuestro objeto de estudio. La aplicación de un análisis estadístico confiable es necesario para corroborar o refutar las hipótesis a investigar.

R es un lenguaje y entorno de programación para análisis estadístico y gráfico. Se trata de un proyecto de software libre, resultado de la implementación del lenguaje S, uno de los lenguajes más utilizados en investigación por la comunidad estadística. A diferencia de otros programas de estadística usuales en el campo de la investigación, R no utiliza una interfaz interactiva para la realización de pruebas estadísticas, lo que genera frustración en usuarios que tienen un primer acercamiento al programa. Sin embargo, al tratarse de un lenguaje de programación, las posibilidades que ofrece en cuanto a la manipulación y análisis de los datos es mucho más vasta, ya que los usuarios pueden definir sus propias funciones.

En conjunto, el aprendizaje de las nociones elementales del análisis estadístico de datos lingüísticos y de una herramienta de análisis apropiada, permitirán al alumno establecer conclusiones más adecuadas sobre su objeto de estudio.

2. Objetivos

Este curso tiene como objetivo brindar un primer acercamiento al análisis estadístico de datos lingüísticos a través del entorno R. Se hará hincapié en la enseñanza de nociones y tipos de análisis básicos de estadística (análisis paramétricos y no paramétricos), y en el aprendizaje de la sintaxis y funciones básicas del lenguaje R que en el futuro permitirán al alumno tener un aprendizaje más autodidacta de funciones más complejas del programa.

3. Contenidos

Unidad I: Acercamiento metodológico al objeto de estudio

Teoría

- Abordaje del objeto de estudio: Testeo de hipótesis y el problema de la generalizabilidad.
- Tipos de estudios: Testeo Experimental vs. No experimental; Estudios longitudinales vs. Transversales; Estudios de grupo vs. Estudios de caso; Análisis cuantitativo vs. cualitativo; Testeo in-situ vs. Testeo en laboratorio.
- Función del análisis estadístico
- Tipos de variables y operacionalización: Variables discretas vs. Continuas; variables dependientes vs. variables independientes.
- Problemas de operacionalización: Tipos de escalas
- Tipos de análisis estadísticos: Estadística descriptiva. Media, moda y mediana. Rango, dispersión y desvío estándar.

Práctica: R

- Familiarización con R: entorno, consola, comandos para importación y exportación de archivos, búsqueda de ayuda, bibliotecas.
- Operaciones aritméticas básicas, definición de variables, creación de tablas.
- Selección de datos en una matriz de datos. Manipulación de variables.
- Importación de tablas a R.
- Trabajo con matrices de datos: datos descriptivos (cálculo de promedio, desviación estándar, moda, mediana, rango).
- Diagramas de dispersión, diagramas de cajas.

Unidad II: Comparación de medias I

Teoría

- Distribuciones de probabilidad: Teoría de distribución gaussiana, distribuciones bimodales.
- Transformación de variables: escala logarítmica e inversa.
- Estadística inferencial: contraste de hipótesis, hipótesis nula, tipos de errores, diferencia entre población y muestreo. Desvío estándar vs. error estándar.
- Prueba de significancia.
- Prueba de diferencias en las medias (t-test)

Práctica: R

Para practicar los conceptos clave de la clase teórica, en esta unidad se analizarán los datos descriptivos de distintos experimento de muestra, se graficarán los datos y se analizarán estadísticamente utilizando las pruebas de significancia aprendidas.

- Análisis de experimento de muestra (1): Exploración y análisis de frecuencias en corpus lingüístico.
- Análisis de experimento de muestra (2): Exploración y análisis de resultados de tarea de decisión léxica.

Unidad III: Comparación de medias II

Teoría

- Comparaciones múltiples, pruebas para diferencias múltiples en la media (ANOVA de un factor).
- Pruebas de diferencia en diseños de dos factores (ANOVA de dos factores)
- Análisis de datos categoriales (prueba binomial, regresión logística).
- Tests post-hoc.

Práctica R

Práctica de los conceptos vistos en la clase teórica mediante análisis de los datos de un experimento de muestra.

- Análisis de experimento de muestra (3): tarea de comprensión de oraciones,
- Análisis de datos de muestras de alumnos (en caso de que el tipo de análisis esté incluido en los contenidos de la unidad)

Unidad IV: Análisis no paramétricos

Teoría

- Análisis de frecuencias: prueba de Chi-cuadrado para una muestra
- Análisis de diferencias de medias en datos ordinales: Prueba de U-Mann Whitney
- Análisis de distribución de la población: Prueba de Kolmogorov-Smirnov para dos muestras.
- Análisis de diferencias múltiples en la media para datos ordinales: Prueba H de Kruskal-Wallis

Práctica R:

Práctica de los conceptos vistos en la clase teórica mediante análisis de los datos de un experimento de muestra.

- Análisis de datos de muestra (5): frecuencia de aparición de auxiliares en corpus del holandés.
- Análisis de datos de muestras de los alumnos (en caso de que el tipo de análisis necesario esté incluido en los contenidos de la unidad).

Unidad V: Regresiones lineales

Teoría

- Regresiones lineales: regresiones simples y múltiples. Recta de regresión y pendiente. Interpretación de los resultados.
- Crítica de modelo: bondad de ajuste y comparación de modelos.
- Modelos lineales generalizados para datos binomiales.
- Introducción a modelos lineales con efectos mixtos.

Práctica R:

Práctica de los conceptos vistos en la clase teórica mediante análisis de los datos de un experimento de muestra.

- Análisis de datos de muestra (6): Tiempos de Respuesta y respuestas correctas en tarea de comprensión de oraciones.

- Análisis de datos de muestra del alumno (en caso de que el tipo de análisis necesario esté incluido en los contenidos de la unidad).

Unidad VI: Correlaciones

Teoría

- Relaciones ente dos variables continuas: correlación de Pearson r
- Relaciones entre variables ordinales: correlación de Spearman Rho .

Práctica R:

Práctica de los conceptos vistos en la clase teórica mediante análisis de los datos de un experimento de muestra.

- Análisis de datos de muestra (4): tarea de decisión léxica.
- Análisis de datos de muestra del alumno (en caso de que el tipo de análisis necesario esté incluido en los contenidos de la unidad)

4. Bibliografía específica obligatoria

Unidad I

Baayen, R. H. (2008) *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press. Capítulos 1 y 2.

Unidad II

Baayen, R. H. (2008) *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press. Capítulos 3 y 4.

Unidad III

Baayen, R. H. (2008) *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press. Capítulo 4.

Unidad IV

Baayen, R. H. (2008) *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press. Capítulos 4 y 5.

Unidad V:

Baayen, R. H. (2008) *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press. Capítulos 6 y 7.

Unidad VI:

Baayen, R. H. (2008) *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press. Capítulo 6.

5. Bibliografía complementaria general

- Baayen, R. H., and Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, 3 (2), 12 - 28.
- Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random-effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68, 255-278.
- Bates, D. M. (2005). Fitting Linear Mixed Models in R. *R News*, 5, 27–30.
- Bresnan, J., Cueni, A., Nikitina, T. and Baayen. R.H. (2007) Predicting the Dative Alternation. In G. Bouma, I. Kraemer, and J. Zwarts, (Eds.), *Cognitive Foundations of Interpretation*, 69–94. Royal Netherlands Academy of Science.
- Pinheiro, J. C. and Bates, D.M. (2000) *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing. Springer, New York.
- Tyler, L.K., Marslen-Wilson, W.D., and Stamatakis, E.A. (2005). Differentiating Lexical Form, Meaning, and Structure in the Neural Language System. *PNAS*, 102, 8375–8380.

6. Carga horaria

Cuatro horas semanales.

7. Actividades planificadas

Las clases serán mitad teóricas y mitad prácticas. Se asignarán las primeras dos horas a la exposición de los contenidos teóricos sobre análisis estadístico de datos, y las segundas dos horas al análisis de datos correspondientes a distintas muestras. Una vez que los estudiantes cuenten con las herramientas básicas para el manejo de R, se les pedirá que analicen una muestra de datos propios o de algún corpus de dominio público. Los alumnos complementarán este análisis con la exposición gráfica de los resultados de manera oral. Además, durante el transcurso del seminario, se complementará la práctica en clase con ejercicios para resolver fuera de ella.

8. Condiciones de regularidad y régimen de promoción

El seminario se dictará en no menos de 4 (cuatro) horas semanales y los alumnos deberán asistir al 80 % de las reuniones y prácticas que se establezcan en el régimen cuatrimestral.

El profesor evaluará la participación de los alumnos con una nota. Esta nota será el resultante de la evaluación que el profesor realice sobre las presentaciones de informes parciales, exposiciones orales individuales o grupales, lecturas, síntesis e informes bibliográficos, etc. Si ésta fuera inferior a cuatro puntos, significará un aplazo en el Seminario. Separadamente, calificará el trabajo monográfico. Si éste fuera rechazado, los interesados tendrán opción en este caso y por única vez a presentarlo nuevamente antes de

la finalización del plazo fijado en el apartado IV (esto es, dentro de los cuatro años posteriores a la finalización el seminario). La calificación final resultará del promedio de ambas notas.

9. Recomendaciones

- Para la realización del curso no es necesario que el alumno cuente con conocimientos previos de estadística.
- Si bien durante el curso se utilizarán en gran parte datos obtenidos de experimentos psicolingüísticos, el curso no está limitado a alumnos que formen parte de este campo. Al tratarse de un curso introductorio, los contenidos son igualmente válidos para el análisis de datos pertenecientes a otras áreas de investigación de la lingüística aplicada.
- Para la realización del curso se debe contar con una computadora que cuente con la última versión de R correspondiente a su sistema operativo. El programa puede bajarse de la página <http://mirror.fcaglp.unlp.edu.ar/CRAN/>. También se debe bajar el programa R Studio desde la página <https://www.rstudio.com/products/rstudio/download/>. Ambos programas son de acceso libre, por lo que no se necesita la adquisición de licencia para su uso.

Firma

Aclaración

Cargo