



**UNIVERSIDAD DE BUENOS AIRES
FACULTAD DE FILOSOFÍA Y LETRAS**

DEPARTAMENTO: LETRAS

CÓDIGO N°:

SEMINARIO: ESTADÍSTICA PARA LINGÜISTAS

MODALIDAD DE DICTADO: VIRTUAL ajustado a lo dispuesto por REDEC-2021-2174-UBA-DCT#FFYL.

PROFESOR: MARAFIOTI, PABLO EZEQUIEL

CUATRIMESTRE: 1°

AÑO: 2022

**UNIVERSIDAD DE BUENOS AIRES
FACULTAD DE FILOSOFÍA Y LETRAS
DEPARTAMENTO DE LETRAS**

CÓDIGO N°:

SEMINARIO: ESTADÍSTICA PARA LINGÜISTAS

MODALIDAD DE DICTADO: VIRTUAL ajustado a lo dispuesto por REDEC-2021-2174-UBA-DCT#FFYL ¹

CARGA HORARIA: 64 HORAS

1° CUATRIMESTRE DE 2022

PROFESOR: MARAFIOTI, PABLO EZEQUIEL

a. **Fundamentación y descripción**

Muchas sub-disciplinas de la lingüística requieren la correcta aplicación de herramientas cuantitativas para el análisis de datos experimentales y de corpus. El análisis estadístico permite explorar datos, identificar patrones y refutar hipótesis.

Este curso propone una introducción a un amplio espectro de métodos estadísticos para regresión desde una perspectiva práctica. Se pondrá el énfasis en aplicaciones y casos de estudio. No se requieren conocimientos previos de programación o matemática. Únicamente es necesaria la dedicación por parte de los/as alumno/as para aprender a usar **R**, un lenguaje de programación y entorno para el análisis estadístico [software libre].

b. **Objetivos:**

- Proveer un conjunto de técnicas para el análisis de respuestas gaussianas y no gaussianas, usando modelos de regresión.
- Lograr manejar datos correlacionados.
- Poder explorar datos mediante análisis descriptivo y gráficas.
- Introducir a los/as estudiantes en el entorno de programación **R**.
- Brindar habilidades para la crítica metodológica de artículos científicos.
- Proporcionar a los/las estudiantes herramientas para poder realizar un proyecto de investigación en lingüística cuantitativa.

c. **Contenidos:** [organizados en unidades temáticas]

Unidad 1: Análisis descriptivo y nociones de probabilidad e inferencia.

¹ De acuerdo con dispuesto por REDEC-2021-2174-UBA-DCT#FFYL, la cual establece para el dictado de las asignaturas de grado durante la cursada del 1° y 2° cuatrimestre de 2022 las pautas complementarias a las que deberán ajustarse aquellos equipos docentes que opten por dictar algún porcentaje de su asignatura en modalidad virtual.

Introducción a **R**. Tipos de variables. Estadística descriptiva. Gráficas. Pre-procesamiento. Nociones de probabilidades. Función de probabilidad puntual, de densidad y acumulada. Algunas distribuciones usuales. Distribuciones muestrales. Estimación por intervalo. Nociones de: p-valor, hipótesis, potencia, tamaño muestral, tamaño de efecto. Tests de normalidad e igualdad de varianzas.

Unidad 2: El modelo lineal [LM].

Correlación. Modelo de regresión lineal simple. Supuestos. Estimación de los parámetros por cuadrados mínimos. Distribución de los estimadores. Test de hipótesis. Intervalo de confianza e intervalo de predicción. Regresión lineal múltiple. Predictoras binarias y categóricas. Comparaciones múltiples. Contrastes. Análisis de residuos. Transformaciones. Colinealidad, heteroscedasticidad, no linealidad y falta de independencia. Selección de modelos.

Unidad 3: El modelo lineal generalizado [GLM].

La familia exponencial. Funciones de enlace. Regresión logística. Estimación de los parámetros por máxima verosimilitud. Test de cociente de máxima verosimilitud. Residuos y bondad de ajuste. Test de Hosmer-Lemeshow. Curva ROC y otras medidas de desempeño para predicción. Regresión de Poisson. Sobre-dispersión.

Unidad 4: Métodos para datos agrupados y medidas repetidas.

Modelo lineal mixto [LMM]. Efectos fijos y aleatorios. Estructuras de covarianzas. Modelos lineales generalizados mixtos [GLMM]. Modelos con ecuaciones estructurales generalizadas [GEE].

d. Bibliografía, filmografía y/o discografía obligatoria, complementaria y fuentes, si correspondiera:

Unidad 1:

Bibliografía obligatoria

- Marafioti, P. E. (2021). *Estadística para lingüistas: notas de clase*. Manuscrito. Cap. 1 (secciones: 1.1 a 1.8, 1.10, 1.11) y cap. 2 (secciones: 2.1, 2.2).
- Chan et al. (2019), caps. 2, 4.1.
- Winter, B. (2019). *Statistics for linguists. An Introduction using R*. NY: Routledge. Caps. 3, 9, 10.

Bibliografía complementaria

- Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press. Caps. 1 - 3.
- Gries, S. T. (2013). *Statistics for Linguistics with R*. Berlin: Walter de Gruyter. Caps. 1 a 3.
- Levshina, N. (2015). *How to do Linguistics with R*. Amsterdam: John Benjamins. Caps. 1 - 4.
- Levy, R. (2012). *Probabilistic Models in the Study of Language*. Manuscrito. Caps. 2 a 4.
- Larson-Hall, J. (2016). *A guide to doing statistics in second language research using spss and r*. Routledge. Cap. 3.

Unidad 2:

Bibliografía obligatoria

- Marafioti, P. E. (2021). *Estadística para lingüistas: notas de clase*. Manuscrito. Cap. 3 (secciones: 3.1 a 3.8).
- Winter, B. (2019). *Statistics for linguists. An Introduction using R*. NY: Routledge. Caps. 4 a 8.

Bibliografía complementaria

- Burnham, K. P. & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: a practical information theoretic approach*. Springer. Caps. 1 a 3.
- Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press. Caps. 6.1, 6.2.
- Gries, S. T. (2013). *Statistics for Linguistics with R*. Berlin: Walter de Gruyter. Caps. 5.1, 5.2.
- Larson-Hall, J. (2016). *A guide to doing statistics in second language research using spss and r*. Routledge. Caps. 6, 7.
- Levshina, N. (2015). *How to do Linguistics with R*. Amsterdam: John Benjamins. Caps. 6, 7.
- Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed-Effects Models and Extensions in Ecology with R*. Springer. Caps. 2, 4, 6.

Unidad 3:

Bibliografía obligatoria

- Marafioti, P. E. (2021). *Estadística para lingüistas: notas de clase*. Manuscrito. Cap. 4.
- Winter, B. (2019). *Statistics for linguists. An Introduction using R*. NY: Routledge. Caps. 12, 13.

Bibliografía complementaria

- Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press. Cap. 6.3.
- Gries, S. T. (2013). *Statistics for Linguistics with R*. Berlin: Walter de Gruyter. Caps. 5.3, 5.4.
- Levshina, N. (2015). *How to do Linguistics with R*. Amsterdam: John Benjamins. Caps. 12, 13.
- Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed-Effects Models and Extensions in Ecology with R*. Springer. Caps. 8 a 11.

Unidad 4:

Bibliografía obligatoria

- Marafioti, P. E. (2021). *Estadística para lingüistas: notas de clase*. Manuscrito. Cap. 5.
- Winter, B. (2019). *Statistics for linguists. An Introduction using R*. NY: Routledge. Caps. 14, 15.

Bibliografía complementaria

- Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press. Cap. 7.
- Gries, S. T. (2013). *Statistics for Linguistics with R*. Berlin: Walter de Gruyter. Caps. 5.5, 5.6.
- Speelman, D., Heylen, K., & Geeraerts, D. [Eds.] (2018). *Mixed-Effects Regression Models in Linguistics*. Springer. Caps. 1, 4.
- Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed-Effects Models and Extensions in Ecology with R*. Springer. Caps. 5, 12, 13.

Bibliografía general

Agresti, A. (2019). *An Introduction to Categorical Data Analysis*. John Wiley & Sons.

- Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.
- Burnham, K. P. & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: a practical information theoretic approach*. Springer.
- Chan, D., Badano, C., & Rey, A. (2019). *Análisis inteligente de datos con lenguaje R*. Editorial de la Universidad Tecnológica Nacional.
- Chatterjee, S. & Hadi, A. S. (2012) *Regression Analysis by Example*. John Wiley & Sons.
- Desagulier, G. (2017). *Corpus Linguistics and Statistics with R*. Springer.
- Faraway, J. J. (2015) *Linear Models with R*. CRC Press.
- Gareth, J., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.
- Gries, S. T. (2013). *Statistics for Linguistics with R*. Berlin: Walter de Gruyter.
- Hedeker, D. & Gibbons, R. D. (2006). *Longitudinal Data Analysis*. John Wiley & Sons.
- Hilbe, J. M. (2009). *Logistic Regression Models*. CRC Press.
- Hilbe, J. M. (2014). *Modeling Count Data*. Cambridge University Press.
- Hilbe, J. M. (2015). *Practical Guide to Logistic Regression*. CRC Press.
- Hosmer, D. W. (Jr.), Lemeshow, S. & Sturdivant, R. X. (2013). *Applied Logistic Regression*. John Wiley & Sons.
- Johnson, K. (2008). *Quantitative Methods in Linguistics*. Singapore: Blackwell.
- Kracht, M. (2005). *Introduction to Probability Theory and Statistics for Linguistics*. Manuscrito.
- Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models*. McGraw-Hill.
- Larson-Hall, J. (2016). *A guide to doing statistics in second language research using spss and r*. Routledge.
- Levshina, N. (2015). *How to do Linguistics with R*. Amsterdam: John Benjamins.
- Levy, R. (2012). *Probabilistic Models in the Study of Language*. Manuscrito.
- Marafioti, P. E. (2021). *Estadística para lingüistas: notas de clase*. Manuscrito.
- Montgomery, D. C. (2017). *Design and Analysis of Experiments*. John Wiley & Sons.
- Rasinger, S. (2013). *Quantitative Research in Linguistics. An Introduction*. UK: Bloomsbury Academic.
- Speelman, D., Heylen, K., & Geeraerts, D. [Eds.] (2018). *Mixed-Effects Regression Models in Linguistics*. Springer.
- Stroup, W. W. (2013). *Generalized Mixed Models*. CRC Press.
- Tan, P-N., Steinbach, M., & Kumar, V. (2014). *Introduction to Data Mining*. Pearson.
- West, B. T., Welch, K. B., & Galecky, A. T. (2007). *Linear Mixed Models. A practical guide using statistical software*. Chapman & Hall / CRC.
- Winter, B. (2019). *Statistics for linguists. An Introduction using R*. NY: Routledge.
- Wood, S. N. (2017) *Generalized Additive Models. An introduction with R*. CRC Press.
- Wickham, H., & Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data*. O'Reilly Media, Inc.
- Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed-Effects Models and Extensions in Ecology with R*. Springer.

e. **Organización del dictado de seminario**

El seminario / proyecto se dicta atendiendo a lo dispuesto por REDEC-2021-2174-UBA-DCT#FFYL la cual establece pautas complementarias para el dictado de las asignaturas de grado durante la cursada del 1° y 2° cuatrimestre de 2022.

Seminario cuatrimestral
<p>El seminario se dictará completamente en modalidad virtual.</p> <p>El dictado de clases se realiza a través del campus virtual de la Facultad de Filosofía y Letras y de otros canales de comunicación virtual que se consideren pertinentes para favorecer el intercambio pedagógico con los/las estudiantes.</p> <p>Las clases tendrán una modalidad teórico-práctica. Los alumnos contarán con notas de clase, scripts de R ya preparados, ejercicios y su resolución. El curso es auto contenido, de modo que cada alumno pueda reproducir en casa los análisis y la programación que se hará en las clases.</p>

Seminario cuatrimestral
<p>La carga horaria mínima es de 64 horas (sesenta y cuatro) y comprenden un mínimo de 4 (cuatro) y un máximo de 6 (seis) horas semanales de dictado de clases.</p>

f. **Organización de la evaluación**

El sistema de regularidad y aprobación del seminario se rige por el Reglamento Académico (Res. (CD) N° 4428/17)

Regularización del seminario:

Es condición para alcanzar la regularidad del seminario aprobar una evaluación con un mínimo de 4 (cuatro) durante la cursada. Para ello los/las docentes a cargo dispondrán de un dispositivo definido para tal fin.

Aprobación del seminario:

La evaluación del seminario constara de las siguientes instancias: (i) un trabajo guiado de análisis estadístico con datos provistos por el docente; (ii) un trabajo final integrador de análisis estadístico sobre una base de datos recabada por los/las estudiantes o bien proporcionada por el docente.

Los/as estudiantes que cumplan el requisito mencionado podrán presentar el trabajo final integrador que será calificado con otra nota. La calificación final resultará del promedio de la nota de cursada y del trabajo final integrador.

Si el trabajo final integrador fuera rechazado, los/as interesados/as tendrán la opción de presentarlo nuevamente antes de la finalización del plazo de vigencia de la regularidad. El/la estudiante que no presente su trabajo dentro del plazo fijado, no podrá ser considerado/a para la aprobación del seminario.

VIGENCIA DE LA REGULARIDAD: El plazo de presentación del trabajo final de los seminarios es de 4 (cuatro) años posteriores a su finalización.

RÉGIMEN TRANSITORIO DE ASISTENCIA, REGULARIDAD Y MODALIDADES DE EVALUACIÓN DE MATERIAS: El cumplimiento de los requisitos de regularidad en los casos de estudiantes que se encuentren cursando bajo el Régimen Transitorio de Asistencia, Regularidad y Modalidades de Evaluación de Materias (RTARMEM) aprobado por Res. (CD) N° 1117/10 quedará sujeto al análisis conjunto entre el Programa de Orientación de la SEUBE, los Departamentos docentes y los/las Profesores a cargo del seminario.

g. **Recomendaciones**

Se requiere que el/la estudiante: (i) tenga acceso a una computadora portátil personal; (ii) sea capaz de leer en inglés.



Aclaración Dr. Pablo Ezequiel Marafioti



Miguel Vedda
Director del Departamento de Letras