

Aspectos generales

Título:	Manejo avanzado y análisis de datos con R
Programas de posgrado o planes de estudio en donde se ofertará adicionalmente:	-
Área del conocimiento:	Genética, genómica y bioinformática
Semestre:	2027-1
Modalidad:	Curso fundamental
Horario:	Martes de 17:00 a 19: 30h
No. sesiones:	16
Horas por sesión:	2.5
Total alumnos PDCB:	6
Total alumnos:	6
Videoconferencia:	Si
Lugar donde se imparte:	Sala de juntas del Departamento de Farmacología de la Facultad de Medicina de la UNAM.
Informes:	ad.ortega@unam.mx

Métodos de evaluación

MÉTODO	PORCENTAJE	NOTAS
Examen	30%	
Participación en clase	50%	
Trabajos y ejercicios	20%	

Contribución de este curso/tópico en la formación del alumnado del PDCB:

El análisis de inferencia es fundamental para el aprendizaje del análisis científico de los datos, sin embargo, el análisis de varias variables contribuye a un mejor entendimiento de los fenómenos en las ciencias biomédicas. Este curso contribuirá a que las y los alumnos, aumenten su criterio estadístico a partir de pruebas descriptivas y de análisis bivariados para poder elegir y ejecutar la o las pruebas adecuadas de análisis múltiple para el óptimo análisis de sus datos y la obtención de conclusiones.

Profesor (a) responsable

Nombre:	Ortega Ayala Adiel
Teléfono:	
Email:	ad.ortega@unam.mx

Profesores (as) participantes

PARTICIPANTE	ENTIDAD O ADSCRIPCIÓN	SESIONES
ORTEGA AYALA ADIEL Responsable	Facultad de Medicina	: Diagnóstico y limpieza de datos Clustering y segmentación II Combinando la regresión lineal y el análisis de varianza Estudio de las pruebas de análisis múltiple de la varianza EXAMEN La regresión logística multinomial Métodos de clasificación II: Modelos generativos de clasificación Métodos de regresión II: Modelo de riesgos proporcionales de Cox
MOLINA GUARNEROS JUAN ARCADIO Integrante	Facultad de Medicina	Análisis de medidas repetidas Análisis de supervivencia Clustering y segmentación I Introducción al análisis de datos con R Métodos clásicos en estadística Métodos de clasificación I Métodos de clasificación II: Modelos generativos de clasificación Técnicas de reducción de dimensión I

Introducción

Una vez que hemos ya dudado de los datos y hemos podido entenderlos, es momento de utilizarlos.

Este curso ofrece una inmersión en técnicas avanzadas de análisis de datos y está orientado a personas con conocimientos previos de R que desean profundizar en modelaje estadístico, manejo eficiente de datos y visualización. Como es del conocimiento de muchos alumnos del Programa de Doctorado en Ciencias Biomédicas de la UNAM, combinamos la teoría con la práctica para entender algoritmos complejos en la resolución de problemas. Este curso representa la continuación del curso de "Introducción al análisis estadístico con R" y también representa el punto medio entre la estadística clásica y la introducción a una rama de la inteligencia artificial muy utilizada en investigación biomédica y ciencias de la salud: El machine learning.

Temario

Sesión 1. TEMA 1: Introducción al análisis de datos con R (Dr. Juan Arcadio Molina Guarneros, 2.5 h) martes 04 de agosto 2026.

1.1 Módulo de estadística

1.1.1 Medidas de tendencia central y de dispersión.

1.1.1.1 Media, mediana, moda

1.1.1.2 Desviación estándar, varianza, covarianza y rangos intercuartiles.

1.1.2 Inferencia estadística en el análisis exploratorio de datos.

1.1.2.1 Normalidad y libre distribución

1.1.2.2 Pruebas de inferencia: T de Student, prueba t pareada, U de Mann Whitney, Prueba de rangos con signo de Wilcoxon.

1.1.2.3 Prueba de ANOVA de 1 vía

1.1.2.4 Prueba de Kruskal-Wallis

1.1.2.5 Pruebas post hoc.

1.1.3 Medidas de asociación: OR, RR y HR.

1.1.3.1 Pruebas de inferencia de datos cualitativos.

1.1.3.1.1 Chi cuadrada de Pearson

1.1.3.1.2 Prueba exacta de Fisher

1.1.3.1.3 Prueba de McNemar

1.1.3.1.4 ¿Qué es el Hazard ratio?

1.2 Manipulación de datos: Tidyverse.

Sesión 2. TEMA 2: Diagnóstico y limpieza de datos (Dr. En C. Adiel Ortega Ayala, 2.5 h) martes 11 de agosto 2026.

2.1 Tipos de datos faltantes.

2.2 Tratamiento de datos faltantes: Imputación simple y múltiple.

2.3 Detección de outliers.

Sesión 3. TEMA 3: Métodos clásicos en estadística (Dr. Juan Arcadio Molina Guarneros, 2.5 h) martes 18 de agosto de 2026.

3.1 Regresión lineal simple

3.1.1 Estimación de coeficientes

3.1.2 Evaluación de la precisión de los coeficientes y el modelo

3.2 Regresión lineal múltiple

3.2.1 Consideraciones de los modelos lineales múltiples.

Sesión 4. TEMA 4: ANCOVA: Combinando la regresión lineal y el análisis de varianza (Dr. En C. Adiel Ortega Ayala, 2.5 h) martes 25 de agosto de 2026.

Sesión 5. TEMA 5: Análisis de medidas repetidas (Dr. Juan Arcadio Molina Guarneros, 2.5 h) martes 01 de septiembre de 2026.

5.1 ANOVA de medidas repetidas

5.1.1 Supuestos y pruebas

5.1.2 Prueba de esfericidad

5.1.3 Test de Friedman

5.2 Prueba Q de Cochran

Sesión 6. TEMA 6: Estudio de las pruebas de análisis múltiple de la varianza (Dr. En C. Adiel Ortega Ayala, 2.5 h) martes 08 de septiembre de 2026.

6.1 ANOVA de 2 vías.

6.2 ANOVA de medidas repetidas

6.3 ANCOVA de 2 vías

Sesión 7. TEMA 7: Métodos de clasificación I (Dr. Juan Arcadio Molina Guarneros, 2.5 h) martes 22 de septiembre de 2026.

7.1 ¿Por qué no una regresión lineal?

7.2 La regresión logística

7.2.1 Estimación de coeficientes de regresión logística.

7.2.2 Elaboración de predicciones.

Sesión 8. TEMA 8: La regresión logística multinomial (Dr. En C. Adiel Ortega Ayala, 2.5 h) martes 29 de septiembre de 2026.

8.1 ¿Regresión logística múltiple o regresión logística multinomial?

Sesión 9. TEMA 9: Análisis de supervivencia (Dr. Juan Arcadio Molina Guarneros, 2.5 h) martes 06 de octubre de 2026.

9.1 Concepto de censura y evento

9.2 Media y mediana de supervivencia

9.3 Estimación de estadísticos de supervivencia

9.4 Construcción de curvas de Kaplan-Meier

9.5 Prueba de Log-Rank y Wilcoxon generalizada.

Sesión 10. TEMA 10: Métodos de regresión II: Modelo de riesgos proporcionales de Cox (Dr. En C. Adiel Ortega Ayala, 2.5 h) martes 13 de octubre de 2026.

10.1 El modelo múltiple de la supervivencia

10.2 ¿OR, RR o HR?

10.3 El concepto de Hazard ratio.

Sesión 11. TEMA 11: Métodos de clasificación II: Modelos generativos de clasificación (Dr. Juan Arcadio Molina Guarneros, 2.5 h) martes 20 de octubre de 2026.

11.1 Análisis lineal discriminante.

11.2 Análisis discriminante cuadrático

Sesión 12. TEMA 11: Métodos de clasificación II: Modelos generativos de clasificación (Dr. En C. Adiel Ortega Ayala, 2.5 h) martes 27 de octubre de 2026.

11.3 Algoritmo Naïve-Bayes

11.4 Algoritmo K-NN

11.5 Elaboración de curvas ROC

11.6 Sensibilidad, especificidad, AUC y métricas de evaluación de clasificación.

Sesión 13. TEMA 12: Clustering y segmentación I (Dr. Juan Arcadio Molina Guarneros, 2.5 h) martes 03 de noviembre de 2026.

12.1 ¿Algoritmo supervisado o no supervisado?

12.2 ¿Clasificación o clustering?

12.3 El algoritmo K-means

Sesión 14. TEMA 13: Clustering y segmentación II (Dr. En C. Adiel Ortega Ayala, 2.5 h) martes 10 de noviembre de 2026.

13.1 Clustering jerárquico

Sesión 15. TEMA 14: Técnicas de reducción de dimensión I (Dr. Juan Arcadio Molina Guarneros, 2.5 h) martes 17 de noviembre de 2026.

14.1 Introducción al álgebra lineal para el PCA.

14.2 Análisis de componentes principales.

Sesión 16: EXAMEN (Dr. En C. Adiel Ortega Ayala, 2.5 h) martes 24 de noviembre de 2026.

Bibliografía

1. Raykov, T., & Marcoulides, G. A. (2012). Basic statistics: An introduction with R. Routledge. ISBN 978-1-4422-1846-8 (cloth); ISBN 978-1-4422-1847-5 (pbk).
2. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R (Vol. 103). Springer. <https://doi.org/10.1007/978-1-4614-7138-7>.
3. Fernández Casal, R., Costa Bouzas, J., & Oviedo de la Fuente, M. (2024). Métodos predictivos de aprendizaje estadístico. Universidade da Coruña. <https://doi.org/10.17979/spudc.9788497498937>
4. Crawley, M. J. (2015). Statistics: An introduction using R (2ª ed.). John Wiley & Sons. <https://doi.org/10.1002/9781118941126>

Observaciones

El alumno deberá contar con equipo de cómputo propio y transportable (laptop). Las computadoras que los alumnos poseen suelen cubrir los requisitos mínimos para el curso, las cuales son:

1. Laptop con cualquier sistema operativo al que esté acostumbrado el alumno (Windows, Mac OS, GNU/Linux, etc).
2. Memoria RAM mínimo 16 GB.
3. Procesador: A partir de Intel Core i5, AMD Ryzen 5 o equivalente.
4. Disco duro: Al menos de 512 GB SSD o mecánico.
5. Pantalla: Idealmente superior a 14 pulgadas.

Evaluación de la participación en clase.

Durante cada clase, el profesor explica los scripts de R y asigna ejercicios del tema o de dominio del lenguaje R. Los alumnos foráneos y aquellos que acuden de forma presencial, deberán unirse a una sesión de Zoom para poder enviar en el chat, los scripts de sus ejercicios realizados en clase.

En caso de que algún alumno tenga problemas, se le facilita la acción "compartir pantalla" y presenta su problema. El profesor invita los alumnos a solucionar el problema del o la compañera que está compartiendo pantalla. La participación culmina cuando el alumno que presenta pantalla escribe el script correcto y muestra resultados frente al grupo.

La asistencia y participación se evalúan cada clase y se asigna un porcentaje proporcional a las 17 sesiones del curso.

Trabajos y ejercicios.

Durante el curso, se asignan 4 tareas:

1. Manejo de datos perdidos.
2. Elaboración de un modelo de ANOVA de medidas repetidas.
3. Elaboración de un modelo de Riesgos proporcionales de Cox.
4. Aplicación de un modelo avanzado (clustering o de reducción de dimensión).

Criterios de evaluación de los trabajos y ejercicios.

1. Presentación de estadísticos descriptivos.
2. Presentación de la inferencia estadística correspondiente.
3. Elaboración de gráficos.
4. Script utilizado en R.
5. Descripción escrita de las conclusiones.

Cada uno de estos puntos tiene un valor de 2.0 puntos.

Examen.

Se realizará un examen en la sesión 16. Se realiza en 10 minutos y es frente al profesor y/o mostrando su pantalla en zoom. La calificación es numérica entre 0 y 10. Son 10 elementos para evaluar y cada uno tiene valor de 1 punto. El script utilizado para el examen deberá ser enviado al profesor antes de finalizar el examen.