DOCTORADO EN BIOTECNOLOGÍA PUCV – UTFSM

DBT 845 - INVESTIGACIÓN REPRODUCIBLE Y ANÁLISIS DE DATOS BIOTECNOLÓGICOS CON R

DESCRIPCION GENERAL DEL CURSO

El curso de postgrado en ***investigación reproducible y análisis de datos biotecnológicos con r*** está orientado a que los estudiantes desarrollen habilidades para llevar a cabo investigación original y reproducible en el ámbito de la biotecnología usando el lenguaje de programación R. Este curso le proporcionará al estudiante competencias prácticas para el almacenamiento, lectura, procesamiento, análisis de datos y presentación de resultados bajo el paradigma del desarrollo de la investigación reproducible, incluyendo entre otros: el análisis exploratorio de datos; la inferencia estadística; la aplicación de modelos lineales y no lineales para el análisis de datos biotecnológicos; el análisis multivariado, entre otros. Los contenidos del curso se explican usando ejemplos de datos y casos de estudio en diferentes ámbitos de la biotecnología incluyendo áreas de salud humana y animal, bioprocesos y medioambiente.

PREREQUISITOS DE CONOCIMIENTOS Y HABILIDADES

Este curso asume un conocimiento base de bioestadística, así como competencias iniciales de programación en R y/o Shell. Los alumnos que no posean estos conocimientos y competencias deberán adquirirlos mediante trabajo individual de autoaprendizaje.

PROFESORES

Coordinador
José Gallardo Matus
jose.gallardo@pucv.cl
Laboratorio de genética y genómica aplicada
https://genomics.pucv.cl/
Pontificia Universidad Católica de Valparaíso

CONTENIDOS CENTRALES

UNIDAD 1. INVESTIGACIÓN REPRODUCIBLE Y ANÁLISIS EXPLORATORIO DE DATOS

Introducción a la unidad

Se entregan los fundamentos de la investigación reproducible y del análisis exploratorio de datos. Se discuten los aspectos beneficiosos de hacer investigación reproducible para el investigador, así como los principales criterios que determinan que una investigación sea reproducible o no. Respecto del análisis exploratorio de datos se definen y clasifican los distintos tipos de variables aleatorias, y se entregan herramientas que permiten observar y predecir el comportamiento de las variables aleatorias bajo distintos tipos de distribución de probabilidad como la distribución Normal, Bernoulli, Binomial negativa, la distribución normal multivariante, entre otros.

Resultado de aprendizaje de la unidad

Al finalizar la unidad, el estudiante será capaz de aplicar los fundamentos de la investigación reproducible y del análisis exploratorio de datos biotecnológicos usando el lenguaje de programación R.

Palabras clave

Reproducibilidad, R, Rstudio, Rmarkdown, Github, variables aleatorios, distribución de probabilidad.

Subtópicos

Subtópico 1.1.- Investigación reproducible con R, Rmarkdown y Github.
Subtópico 1.2.- Variables aleatorias y distribuciones de probabilidad.
Subtópico 1.3.- Análisis exploratorio de datos.

UNIDAD 2. CONTRASTES DE HIPÓTESIS PARAMÉTRICAS Y NO PARAMÉTRICAS

Introducción a la unidad

Se entregan los fundamentos de la inferencia estadística y de las pruebas de contraste de hipótesis. Respecto del análisis de datos biotecnológicos y sobre la base de estudios de caso se entregan herramientas para aplicar diferentes test estadísticos incluyendo: prueba de t-student para la correlación de variables continuas y para diferencia de 2 medias, prueba de F para la diferencia de más de 2 medias (análisis de varianza), prueba de chi^2 para asociación de dos o más variables categóricas (tablas de contingencia), prueba de Wilcoxon para comparación de 2 muestras independientes, entro otras pruebas no paramétricas de uso común en investigación biotecnológica.

Resultado de aprendizaje de la unidad

Al finalizar el unidad, el estudiante será capaz de aplicar los fundamentos de la inferencia estadística en el análisis de datos biotecnológicos usando el lenguaje de programación R.

Palabras clave

Parámetro, estadístico, correlación, permutación, combinación, inferencia estadística, contraste de hipótesis, análisis de supervivencia.

Subtópicos
Subtópico 2.1.- Pruebas de contraste de hipótesis paramétrica.
Subtópico 2.2.- Pruebas de contraste de hipótesis no paramétrica.
Subtópico 2.3.- Análisis de supervivencia.

Unidad 3. MODELOS LINEALES Y ANÁLISIS MULTIVARIADO

Introducción a la unidad
Se entregan los fundamentos del uso de los modelos lineales y del análisis multivariado en la investigación biotecnológica. Se discute el uso de modelos lineales y no lineales para explicar, modelar o predecir la relación de una variable respuesta *Y* con una o más variables predictoras *X*. Se revisan los supuestos de cada tipo de modelo, aplicando un diagnóstico apropiado en cada caso. Se verán los modelos lineales generalizados como una alternativa para corregir el modelo lineal cuando no se cumplen los supuestos. También se introduce el análisis de modelos mixtos y el análisis multivariado con especial énfasis en análisis de componentes principales y análisis de cluster.

Resultado de aprendizaje de la unidad
Al finalizar la unidad, el estudiante será capaz de aplicar diferentes modelos lineales y generalizados para el análisis de datos de biotecnológicos. También será capaz de aplicar algunas técnicas de análisis multivarido para el análisis de datos de biotecnológicos.

Palabras clave
Regresión lineal, regresión lineal múltiple, regresión cuadrática, regresión logística y análisis multivariado.

Subtópicos
Subtópico 3.1.- Modelos lineales.
Subtópico 3.2.- Modelos lineales generalizados.
Subtópico 3.3.- Análisis multivariado.

COMPONENTES DE EVALUACIÓN

PROYECTO INDIVIDUAL DE INVESTIGACIÓN Y ANÁLISIS DE DATOS BIOTECNOLÓGICOS CON R
La evaluación del curso consiste en el desarrollo de un proyecto personal de investigación y análisis de datos biotecnológicos con R. Se dará énfasis a que los alumnos resuelvan un problema de análisis de datos en el tema de su investigación de tesis doctoral. El trabajo se desarrolla en dos etapas, la primera pondera un 40% y la segunda un 60%.

No entregar el trabajo en el plazo establecido para ello será calificado con la nota mínima (1.0).

Es causal de reprobación de la asignatura, no cumplir con el mínimo de asistencia de un 80%. Esto es independiente de que las calificaciones parciales o totales sean mayores de 4.0.

BIBLIOGRAFÍA

1. Recursos Didácticos

Los recursos didácticos de aprendizaje a utilizar son:
a) Guías de trabajo diseñadas y elaboradas por el profesor.

2. Bibliografía Obligatoria

Rafael A Irizarry and Michael I Love. 2015. Data Analysis for the Life Sciences. Leanpub.

John Kloke and Joseph W. McKean. 2016. Nonparametric Statistical Methods Using R. Chapman & Hall/CRC Texts in Statistical Science

Faraway, Julian James. 2015. Linear models with R. CRC Press; Boca Ratón; Estados Unidos. 2a. ed.

3. Bibliografía Complementaria

Julio Sergio Santana Sepúlveda y Efraín Mateos Farfán.2014. El arte de programar en R: un lenguaje para la estadística. México : Instituto Mexicano de Tecnología del Agua. UNESCO. Comité Nacional Mexicano del Programa Hidrológico Internacional. 182 p.

4. Webgrafía

R project website
https://cran.r-project.org

R-Studio
http://www.rstudio.com