Introduccion al analisis de datos en salud con R

Conceptos, abordajes y aplicaciones para el estudio de enfermedades no transmisibles en Salud Pública

Instructor:
Dr. Kevin Martinez-Folgar, MD PhD(c)
Drexel University
km3785@drexel.edu

Asistente:
TBD

Modalidad: Presencial
Lugar: Ciudad Guatemala, Guatemala
Horario: Plan diario de 9:00 am a 4:30 pm

Introducción

Este curso, con una duración de 3 días, permite a los profesionales de la salud con conocimientos en salud pública, epidemiología y bioestadística a mejorar sus habilidades en el análisis de datos utilizando el software gratuito "R". R es un lenguaje de programación y ambiente estadístico muy popular por su capacidad de procesar grandes cantidades de datos y generar visualizaciones de alta calidad, siendo una herramienta esencial para aquellos que buscan analizar y comunicar datos de salud de manera efectiva. Durante el curso, se brindarán las habilidades fundamentales necesarias para utilizar R, desde la importación y manipulación de datos hasta la visualización y análisis estadístico básico.

Objetivo general

El objetivo general del curso es mejorar las habilidades de análisis de datos de los profesionales de la salud con conocimientos en salud pública, epidemiología y bioestadística, a través del uso del software estadístico R, brindando las habilidades fundamentales necesarias para utilizar R, desde la importación y manipulación de datos hasta la visualización y análisis estadístico básico.

Requisitos y/o perfil del alumno: El curso está dirigido a profesionales con conocimientos básicos sobre salud pública, epidemiología y/o bioestadística. Aunque el curso se impartirá en español es necesario contar con manejo de lectura y comprensión del idioma inglés. Es indispensable que los participantes cuenten con una computadora personal (laptop) todos los días del taller, y disponibilidad de tiempo para participar del taller todos los días en todos los horarios. No se aceptarán ausencias totales.

Metodología

El curso será dictado en forma presencial y utilizará una combinación de exposición teórica sobre los conceptos básicos de programacion y analisis de datos seguido de una sesión práctica.

El curso iniciará explicando los conceptos de programacion basica orientada a objetos, introduccion a bases de datos, y un primer contacto con R a traves de Rstudio. Tambien se incluira el uso de funciones que servirán de soporte para entender y usar el software estadistico. En la segunda parte del curso, se abordará la importacion, limpieza y exploracion de datos en preparacion de la base de datos para un analisis futuro. Se presentarán distintas técnicas para importar y preparar la base de datos según el formato de variables y de base de datos importada. Posteriormente se realizara el

analisis descriptivo y la generacion de tablas listas para publicacion en reportes y articulos cientificos.

En la tercera parte, se profundizará en el análisis estadístico descriptivo de la base de datos, utilizando técnicas de visualizacion de datos como histogramas, boxplots y scatterplots para visualizar la distribucion de datos y sus posibles relaciones entre variables.
Finalmente, en la última parte del curso, se mostrará los diferentes mecanismos para resolver dudas de programacion en R de manera autonoma y poder fomentar el autoaprendizaje futuro utilizando herramientas como la documentacion de R, comunidades en linea o tutoriales. gráficas simples bivariadas para describir los resultados obtenidos en el análisis estadístico, utilizando diferentes paquetes y funciones de R. Con esto, los estudiantes tendrán las herramientas necesarias para presentar de manera clara y concisa los resultados obtenidos en sus análisis estadísticos en el contexto de la investigación en epidemiología y salud pública.

Contenido temático

1. **Introducción a la programacion y bases de datos.**
    a. Modulo teórico: Introduccion a la programacion computacional, tipos de programacion (funcional y orientada a objetos), Tipos de datos recolectados en investigacion y formatos para almacenarlo (foto, video, audio, senales electricas, texto, etc.); datos tabulares y formato "tidy"; bases de datos transversales y longitudinales (wide-long).

    b. Módulo de discusión: Ejercicios de discusión grupales luego de cada módulo teórico

2. **Introduccion a R y Rstudio**
    a. Módulo teórico: Introduccion a funciones y tipos de objetos en R. Presentacion de la interfaz de Rstudio. Introduccion a la sintaxis de R basico (matrices de datos), funciones basicas e introduccion a las librerias (instalacion y uso). Tipos de archivos propios de R (R markdown, .R, .Rdata)
    b. Módulo de discusión: ejercicios de discusión grupales luego de cada módulo teórico
    c. Módulo resolución de problemas: Laboratorio 1: Introduccion a R markdown y uso de la interfaz de Rstudio

3. **Importacion y limpieza de bases de datos**
    a. Modulo teorico: Introduccion al tidyverse (programacion funcional y anidado de funciones con pipe %>%); Carga de datos desde diferentes fuentes en R (haven, foreign, xlsx, data.table); Selección de filas y columnas para la importacion adecuada de diferentes tipos de archivos (csv, xlsx, dta, sas7bdat, sav). Inspeccion de datos; Cambio de tipo de columnas según el tipo de dato (string, integer, time); introduccion a las librerias janitor, stringr y lubridate.
    b. Modulo discusión: ejercicios de discusión grupales luego de cada módulo teórico.
    c. Modulo resolución de problemas: Laboratorio 2 en R: Abrir una base de datos, inspeccionar los datos y adecuar las columnas según el tipo de base variable.

4. **Analisis descriptivo de bases de datos y generacion de tablas para reportes.**
    a. Módulo teórico: Análisis descriptivo de datos en general y por subgrupos (media, mediana, moda, desviación estándar, tabulación de frecuencias y generación de tablas en R markdown usando knitr.
    b. Modulo discusión: ejercicios de discusión grupales luego de cada módulo teórico.

c. Modulo resolución de problemas: Laboratorio 3 en R: Generación de una 'Tabla 1' en knitr en PDF y Word.

5. **Visualizacion de datos.**
   a. Modulo teórico: Introduccion a ggplot2, Tipos de graficas (scatterplot, boxplot, histogramas, etc); personalizacion de las graficas (colores, tamaños) y sobre posicion de capas.
   b. Modulo discusión: ejercicios de discusión grupales
   c. Modulo resolución de problemas: Laboratorio 4 en R: Generación de una 'Figura 1' en ggplot2 y exportarla en PDF o PNG.

6. **Como y donde buscar ayuda sobre programacion en R.**
   a. Modulo teórico: Revisión general del contenido del curso a modo de resumen. Documentación y manuales de R y Rstudio; Comunidades y foros en línea: Stack Overflow, RStudio Community, etc.; Tutoriales y cursos en línea: Coursera, etc.; ChatGPT.
   b. Modulo resolución de problemas: Laboratorio 5 en R: --

Criterios de acreditación

- Los asistentes deberán cursar un total de 15 horas en salón de clase,
- Se estima además aproximadamente 15 horas de trabajo individual (fuera del aula) para completar las lecturas y resolver los ejercicios señalados.
- El alumno deberá obtener un puntaje mínimo de 85/100 en el curso para acreditarlo.

Criterios de evaluación:

- Participación creativa, analítica, propositiva y reflexiva en todas las discusiones del curso y los talleres de trabajo (10%)
- Compleción de ejercicios laboratorio individual (50%)
- Trabajo grupal (30%)
- Asistencia (10%). Se requiere la asistencia al menos al 90% de las sesiones

Reconocimiento

Se otorgará constancia de acreditación a los alumnos que cumplan con los requisitos antes señalados.

Bibliografía sugerida