Estadística (M) - Introducción a la Estadística y Ciencia de Datos

Primer Cuatrimestre 2024

Trabajo Práctico Inicial

Condiciones formales

- Este trabajo práctico es de carácter obligatorio, con lo cual aprobarlo es imprescindible para aprobar la materia.
- La consigna debe ser resuelta usando el software R.
- Fecha de entrega: hasta el martes 02/04 a las 23:55.
- Modalidad de trabajo: grupal. Cantidad de integrantes: 3 integrantes máximo, sugerimos 2 de la carrera de Datos y 1 de Matemática.
- Modalidad de entrega: cada grupo deberá subir **2** archivos a la tarea correspondiente en el aula virtual de la materia.

Archivo 1 elaborado con R, de extensión Rmd. Contendrá el informe junto con el código implementado para resolver el ejercicio, demostrando la comprensión en los temas.

Archivo 2 en pdf: informe sobre la resolución del ejercicio, el que se genera a partir del de extensión Rmd.

Cada grupo nombrará los archivos con los apellidos de sus integrantes separados por guión bajo. Por ejemplo, si el grupo estuviese integrado por el equipo docente de la práctica de la materia, los nombres de sus archivos serían: `alen_ditella_statti.Rmd` y `alen_ditella_statti.pdf`.

En el archivo `ENNyS_menorA2.txt` se encuentra parte de la base de datos de la *Encuesta Nacional de Nutrición y Salud 2018-2019* de la República Argentina correspondiente a datos de bebés (o sea, cuya edad es menor a 2 años). Las variables que se seleccionaron para este trabajo práctico son Sexo, Edad, Peso, Perímetro Cefálico y Talla del/ de la bebé registrado/a, además del Tipo de embarazo (simple o múltiple) del que nació. Sexo y Tipo de embarazo son variables categóricas y las restantes son variables continuas.

1. Para la variable Perímetro Cefálico, construir un histograma y superponer, distinguiendo con colores, las densidades estimadas que brinda el comando `density`, utilizando la ventana que viene dada por defecto, con los siguientes núcleos:

a) rectangular.

b) de Epanechnikov.

c) gaussiano.

¿Qué puede decirse acerca de las densidades estimadas?

2. Usando la base de datos brindada, calcular la probabilidad estimada de que un/a bebé de hasta 2 años tenga un perímetro cefálico de entre 42 y 48 cm. utilizando

   a) el histograma y

   b) la densidad estimada cuyo núcleo es el de Epanechnikov,

   graficadas en el ítem . Justificar, desarrollando la deducción de esos cálculos cuando sea necesario.

3. Respecto a lo realizado en el ítem 1, indicar la ventana que brinda, por defecto, cada densidad estimada. ¿Qué pasa si se considera la mitad de esa ventana? ¿Y si se la duplica?

4. Para la variable Perímetro Cefálico, construir un histograma y superponer la densidad estimada utilizando el núcleo gaussiano y la ventana que se usa por defecto. Luego superponer las densidades estimadas por el núcleo gaussiano según la variable Sexo. Interpretar.

5. Para la variable Perímetro Cefálico, construir un histograma y superponer la densidad estimada utilizando el núcleo gaussiano. Luego superponer las densidades estimadas por el núcleo gaussiano según la variable Tipo de embarazo. Interpretar.

6. Hacer un bagplot entre las variables Perímetro Cefálico y Talla. Identificar los datos atípicos que este gráfico brinda. ¿Qué interpreta? ¿Se registraron bebés de Talla alta y Perímetro Cefálico chico? ¿Y de Talla baja y Perímetro Cefálico grande?

7. Para explorar los datos, una investigadora decide realizar un nuevo bagplot excluyendo a los datos atípicos detectados en el ítem anterior. Reproducir el gráfico realizado por la investigadora. ¿Se visualizan datos atípicos ahora? ¿Qué está sucediendo?

8. Realizar un bagplot entre las variables Perímetro Cefálico y Talla, según la variable Sexo. ¿Se observan los mismos datos atípicos que en el ítem 6? ¿Qué interpreta?