Introducción al Aprendizaje Estadístico - Actividades de Clase - Unidad 4

Lucía Babino - Paula Spano

Ejercicio 1: Análisis de datos Gamma-ray busts

Los datos observados corresponden al decaimiento de los rayos-X de GRB 050525a obtenidos por el Telescopio de rayos-X (XRT) a bordo del satélite Swift (A. J. Blustin y 64 coautores, Astrophys. J. 637, 901-913 2006.)

Este conjunto de observaciones consiste en 63 mediciones del brillo en la escala espectral 0.4-4.5 keV a tiempos que van entre 2 minutos y 5 días después de la explosión. Las columnas registran las siguientes variables: 1) el tiempo de observación en segundos, 2) X-ray flux (en unidades de 10-11 ergcm2s, 2-10 keV) y 3) mediciones del error del flux basadas en la relación señal-ruido. En el análisis que sigue ignoraremos la variable **tiempo**.

1. Acceder a los datos con el comando:

   ```
   read.table("http://astrostatistics.psu.edu/datasets/GRB_afterglow.dat", header=T, skip=1)
   ```

   Nos enfocaremos en la variable `f` correspondiente a **flux**, que se halla en la segunda columna y la llamaremos X.

2. Estimar $P(X \leq 40)$.

3. Graficar la función de distribución empírica (que de acá en más llamaremos "la empírica") asociada a los datos **flux** utilizando el comando `ecdf`.

4. Realizar un histograma para los datos de **flux**. ¿Identifica alguna distribución conocida? ¿Podría identificarla en base a "la empírica"?

5. Superponer al histograma la curva de la función de densidad de la distribución identificada en el punto anterior, estimando sus parámetros en base a los datos.

Ejercicio 2: Análisis de datos de Buffalo

Los datos del archivo `buffalo.txt` corresponden a mediciones de la cantidad de nieve caída (en pulgadas) en Buffalo en los inviernos de 1910/1911 a 1972/1973.

6. Realizar un histograma para estos datos utilizando los parámetros por default.

   Repetir pero ahora eligiendo como puntos de corte las siguientes secuencias:

   a) de 20 a 140 con paso 10,

   b) de 20 a 140 con paso 5.

Comparar los tres histogramas obtenidos. ¿Tiene algún efecto el refinamiento de los bins?

7. Realizar histogramas para estas observaciones según las siguientes indicaciones

   a) utilizar como puntos de corte la secuencia de 20 a 140 con paso 10,

   b) correr el punto de inicio de los bins en 2 unidades,

   c) repetir el ítem anterior.

   Comparar los tres histogramas obtenidos. ¿Tiene algún efecto la elección del punto inicial en este ejemplo?

8. Si X es la v.a. cantidad de nieve caída en un invierno en Buffalo, implementar una función que permita estimar la probabilidad

$$\mathbb{P}\Big(X \in [x_0 - h, x_0 + h]\Big) = \mathbb{P}\Big(x_0 - h \leq X \leq x_0 + h\Big)$$

   para un valor x_0 en base a un conjunto de datos provenientes de la v.a. X, para un h dado.

   Es decir, implementar una función `proba_est(x, x_0, h)` que devuelva la proporción de observaciones entre los datos disponibles `x` que están a distancia menor o igual que `h` de `x_0`, siendo `h` el tamaño de la ventana elegida.

9. Utilizar la función definida en el ítem anterior para estimar $\mathbb{P}\Big(x_0 - h \leq X \leq x_0 + h\Big)$ para $x_0 = 80$ y para $h = 10, 15$ y 20, en base a los datos de `buffalo.txt`. Repetir para $x_0 = 40$. Comparar las probabilidades estimadas para $h = 10$ entre $x_0 = 40$ y $x_0 = 80$ ¿Cuál de esas dos probabilidades estimadas es mayor? ¿Podría haber anticipado su respuesta mirando el histograma?

10. Implementar una función `densidad_est_parzen(x, x_0, h)` que tenga por argumento un conjunto de datos `x`, un punto x_0 y una ventana h y devuelva $\widehat{f_h}(x_0)$, el valor de la estimación de Parzen de la densidad f de la cual provienen los datos, en el punto x_0.

11. Utilizar la función definida en el ítem anterior para estimar $f(80)$ utilizando un ancho de venta $h = 10$. Comparar esta estimación con lo que se puede observar desde el histograma.

12. Con la función `densidad_est_parzen` implementada, estimar la densidad f en el intervalo (25, 126.4) (mínimo y máximo de las observaciones) sobre una grilla de 200 puntos equiespaciados, con $h = 10$. Graficar el estimador $\widehat{f_h}(x)$ obtenido.

13. Realizar un histograma para los datos de Buffalo y superponer las densidades estimadas mediante la función `densidad_est_parzen` utilizando `h` = 10, 20 y 50. Observar cómo varía la rugosidad de los estimadores de f obtenidos.

Ejercicio 3: Datos Simulados: Peso de niños

Considerar los datos correspondientes al peso de $n = 100$ niños de 5 años (en Kg.) de una ciudad determinada que se encuentran disponibles en el archivo **datos_sim_ninos.csv**.

14. Con la función `density` de `R`, estimar la densidad f usando el núcleo **normal**. Elegir el valor de la ventana h que parezca dar la mejor estimación.

    Graficarla configurando `ylim = c(0, 0.3)`, que el título incluya el h elegido y pegar el gráfico aquí.

15. Siguiendo con el ítem anterior, a partir de la mejor estimación utilizada para la densidad, calcular las estimaciones de f en los puntos $x = 16, 18, 20, 22$ y volcar sus resultados en el siguiente archivo.

16. Con la función `density` de `R`, estimar la densidad f usando el núcleo normal y los siguientes valores para la ventana: h =0.15, h =0.5 y $h = 10$. Representar en un mismo gráfico las tres estimaciones.

17. ¿Se observan diferencias en las estimaciones obtenidas en el ítem 14 al recorrer los resultados pegados por todos los grupos?

18. Hallar las ventanas óptimas por la regla de Silverman y por Validación Cruzada $\widehat{h}_{rot}$ y $\widehat{h}_{cv}$, respectivamente. Compararlas entre sí y con la que eligió en el ítem 14, ¿son parecidas? Estimar la función de densidad con cada ventana usando el núcleo normal y graficar las respectivas estimaciones, junto con la que eligió en el ítem 14.