Probabilidad y Estadística (C)
Trabajo en laboratorio

Resolver los siguientes items utlizando **R**. Debe entregarse un archivo word o pdf con una breve introducción al tema y una descripción del trabajo realizado, sus resultados y conclusiones, que pueda entenderse sin necesidad de leer el presente documento ni el código de R. Además, adjuntar el script de R creado para resolver el ejercicio.

1. A orillas del río Reconquista yacen numerosas industrias, de las cuales el %70 no cumple con alguna de las normas establecidas para arrojar residuos al río. Un inspector visita 30 de ellas.
   (a) Estimar la probabilidad de que más de 18 estén en infracción. Estimar cuántas industrias debe visitar el inspector para que tal probabilidad sea mayor a 0.95.
   (b) Resolver el item anterior usando R y calcularlo de manera exacta.
   (c) Graficar la convergencia de la primera estimación.
2. Para apreciar aún un poco más la Ley de los Grandes Números, realizar el siguiente experimento:
   (a) Considerar dos observaciones x_1 y x_2 de variables aleatorias X_1 y X_2 independientes con distribución $\mathcal{E}(\lambda)$ (para algún λ a elección) y guardar el promedio de ambas, es decir, $\overline{x_2}$. Repetir 1000 veces y a partir de los valores obtenidos realizar un histograma, un boxplot y un QQ-plot. ¿Qué características tienen?
   (b) Aumentar a cinco las variables promediadas, es decir, considerar ahora $n = 5$ observaciones de variables aleatorias independientes con la misma distribución del ítem anterior y guardar x_5. Repetir 1000 veces y realizar un histograma, un boxplot y un QQ-plot para los valores obtenidos. Comparar con los obtenidos en el ítem anterior. ¿Qué se observa?
   (c) Aumentar a $n = 30$ el número de observaciones de v.a.i.i.d. y repetir el ítem anterior. Repetir con $n = 500$.
   (d) ¿Qué pasaría si se siguiera aumentando el tamaño de la muestra?
   (e) Por último, hacer un boxplot de los 4 conjuntos de datos en el mismo gráfico (es decir, "boxplots paralelos").
3. El teorema central del ímite nos dice que cuando hacemos la siguiente transformación con los promedios:

$$\frac{\overline{X_n} - E(X_1)}{\sqrt{\frac{\mathrm{Var}(X_1)}{n}}}$$

   la distribución de esta variable aleatoria se aproxima a la de la normal estándar si n es suficientemente grande. Comprobaremos mediante una simulación este resultado.
   (a) Calcular la esperanza y varianza de X_1 donde X_1 es la misma distribución que en el ejercicio 2.
   (b) Realizar la transformación mencionada en los 4 conjuntos de datos del ítem 2 y graficar boxplots paralelos y QQ-plots.
   (c) Realizar 4 histogramas y a cada uno de ellos superponerle la densidad de la normal estándar.
   (d) Explicar los resultados obtenidos.
4. Sea $(U_i)_{i\in\mathbb{N}}$ una sucesión de variables aleatorias uniformes en $[0, 1]$. Definimos $N = \inf\{n \in \mathbb{N} : \sum_{i=1}^{n} U_i \geq 1\}$. Realizar simulaciones de la variable aleatoria N y estimar $\mathbb{E}(N)$.
5. Se compararon tres dietas respecto al control de azúcar en la sangre en pacientes diabéticos. En el archivo `estad descriptiva.txt` se encuentran los valores de glucosa para las tres dietas consideradas (A, B, C), que contienen las lecturas de glucosa en la sangre de los pacientes. Es deseable que el paciente tenga valores entre 80 — 110 mg/dl.

(a) Cargue los datos al R.

(b) Para cada una de las tres dietas calcule medidas de centralidad: la media, la mediana, la media α-podada para $\alpha = 0.1, 0.2$. Para cada dieta compare los valores obtenidos de las cuatro medidas de posición, si observa una notable diferencia ¿a que podría deberse?

(c) Calcule medidas de dispersión: el desvío estándar, la distancia intercuartil (o intercuartos) y la MAD en cada una de las dietas. Compare los valores de dispersión obtenidos, si observa una notable diferencia ¿a que podría deberse? ¿Cuál de las dietas parece ser la más estable?

(d) Obtenga los percentiles 10, 25, 50, 75 y 90. Compare los valores de los percentiles obtenidos entre las distintas dietas.

(e) Construya histogramas que permitan visualizar los valores de glucosa para cada dieta. Compare la distribución de glucosa. En alguna de ellas parece haber valores alejados? ¿Las dietas mantienen a los pacientes en los valores deseados? ¿La distribución de glucosa es asimétrica en alguno de los grupos? ¿En algún caso el ajuste normal parece razonable? Realice los diagramas de tallo-hoja correspondientes.

(f) Grafique los box-plots correspondientes. ¿Cómo se compara la información que dan estos gráficos con la obtenida con los histogramas? En base a los gráficos obtenidos, discuta simetría, presencia de outliers y compare dispersiones nuevamente.

(g) Grafique los qqplots correspondientes. ¿En algún caso el ajuste normal parece razonable?

(h) ¿En base al análisis anterior, cuál le parece la dieta más aconsejable?