1	2	3	4	Calificación

Introducción a la Estadística y Ciencia de Datos - Segundo cuatrimestre
Recuperatorio primer examen parcial - 07/12/2023

Nombre y Apellido:
Cantidad Total de Hojas:

Por favor, numerar todas las hojas y colocar el nombre en ellas. Cada ejercicio debe realizarse en hoja separada. Se aprueba con al menos 60 puntos.

- Justificar todas las respuestas -

---

1. (30 puntos) **Primera parte** Sean $X, X_1, \ldots, X_n$ variables aleatorias independientes con densidad dada por

$$f_X(x) = \frac{2x}{\theta^2} I_{[0,\theta]}(x), \quad \theta > 0\,. \tag{1}$$

En este marco interesa estimar $q(\theta) = \theta^2$.

a) (3 puntos) Hallar el estimador de $q(\theta) = \theta^2$ basado en el segundo momento. Lo llamaremos de aquí en más $T(X)$.

b) (9 puntos) Calcular el error cuadrático medio de $T(X)$.

c) (9 puntos) Verificar si $T(X)$ es débilmente consistente. Justificar.

d) (9 puntos) Hallar la distribución asintótica de $T(X)$. Justificar.

(30 puntos) **Segunda parte** Sea $X_1, \ldots, X_n$ una muestra aleatoria donde cada variable tiene densidad dada en (1) y sea $X_{(n)}$ el estimador de máxima verosimilitud de θ.

a) (5 puntos) Mostrar que la distribución de $\frac{X_{(n)}}{\theta}$ es independiente de θ.

b) (10 puntos) En base a a), hallar una familia de intervalos de nivel $1 - \alpha$.

c) (7 puntos) De todos los intervalos anteriores, obtener el de longitud mínima.

d) (8 puntos) En un circuito eléctrico la corriente que circula tiene un amperaje que es una variable aleatoria con distribución dada por (1). Se toman 10 mediciones independientes de la corriente siendo la máxima registrada de 5.6 A. Hallar un intervalo de confianza de nivel 0.95 para θ. ¿Cómo resultaría un intervalo de confianza del mismo nivel para $q(\theta) = \theta^2$?

2. **Teórico** (15 puntos) Sea $T_n = T_n(X_1, \ldots, X_n)$ un estimador de $q(\theta)$ basado en una muestra aleatoria de tamaño n y sea $ECM_\theta(T_n) \to 0$ el error cudrático medio de T_n. Probar los siguientes resultados:

a) (5 puntos) Definir el $ECM_\theta(T_n) \to 0$ y deducir una expresión que lo relacione con el sesgo y la varianza del estimador.

b) (5 puntos) Probar que si $ECM_\theta(T_n) \to 0$ cuando $n \to \infty$, entonces T_n es un estimador débilmente consistente de $q(\theta)$.

c) (5 puntos) Si cuando $n \to \infty$ tenemos que $\mathbb{E}_\theta(T_n) \to q(\theta)$ y $\mathbb{V}ar_\theta(T_n) \to 0$, entonces T_n es débilmente consistente para $q(\theta)$.

3. (25 puntos) El archivo ranas.csv (sugerimos leer mediante read.csv("ranas.csv",head=T)) contiene 10 variables medidas en 212 lugares en Southern Corroboree en Snowy Mountains que es un área de New South Wales, Australia y corresponden a un estudio sobre la distribución de poblaciones de ranas. Las variables medidas son:

   - *pres.abs*: variable binaria (0/1) que indica la ausencia/presencia de ranas en un lugar en particular.
   - *northing*: metros al norte del punto de referencia
   - *easting*: metros al este del punto de referencia
   - *altitude*: altitud en metros
   - *distance*: distancia a la población existente más cercana, en metros
   - *NoOfPools*: número de grupos de reproducción potenciales
   - *NoOfSites*: número de posibles criaderos en un radio de 2 km
   - *avrain*: precipitación media durante la primavera
   - *meanmin*: temperatura mínima media durante la primavera
   - *meanmax*: temperatura máxima media durante la primavera

   Definir como factor las variables que así lo requieran.

   a) Realizar boxplots paralelos para la variable *northing* basado en los datos registrados según haya presencia o ausencia de ranas en el lugar. A partir de estos plots describir la distribución de esta variable en cada uno de los dos tipos de lugar (con ranas y sin ellas) y compararlas. (No más de 4 renglones).

   Muky dice que el gráfico muestra que los lugares con ranas tienden a estar a menos metros al norte del punto de referencia que los que no tienen ranas. ¿Está de acuerdo con Muky? Justificar en no más de 3 renglones.

   b) Graficar la función de distribución empírica de la variable *meanmin* para los lugares sin ranas (en azul) y la correspondiente a los lugares con presencia de ranas (en rojo) superpuestas en un mismo plot. Repetir para la variable la variable *meanmax*.

   Muky dice que los gráficos muestran que durante la primavera en los lugares con ranas las temperaturas mínimas y máximas tienden a ser superiores que en los lugares donde no hay ranas. Justificar en no más de 3 renglones porqué Muky está en lo cierto o no lo está.

c) Hallar la ventana de Silverman para estimar la densidad de la variable *avrain* usando el núcleo normal en los lugares sin ranas. Repetir para los lugares con presencia de ranas.

d) A partir de las ventanas calculadas en el ítem anterior, graficar superpuestas las densidades estimadas obtenidas con dichas ventanas usando azul para los lugares sin ranas y rojo para los que registran ranas. ¿Cuáles son las características más importantes de las estimaciones obtenidas? (No usar más de 3 renglones)