Taller evaluado 02-L 2025-2

CÓDIGO DE HONOR

"Como miembro de la comunidad de la Pontificia Universidad Católica de Chile, me comprometo a respetar los principios y normativas que la rigen. Asimismo, me comprometo a actuar con rectitud y honestidad en las relaciones con los demás integrantes de la comunidad y en la realización de todo trabajo, particularmente en aquellas actividades vinculadas a la docencia, al aprendizaje y la creación, difusión y transferencia del conocimiento."

INDICACIONES

Carácter: INDIVIDUAL.
Tiempo: 40 minutos.
Si el resultado numérico genera 4 o más decimales, aproximar hasta el cuarto decimal.
Está prohibido el uso de formularios y de archivos físicos personales. Solo está permitido estar en la interfaz de R (PositCloud o RStudio) o archivos de Canvas.

Verificar que estén instalados los paquetes:

```
library(mvtnorm)
library(TeachingDemos)
library(fitdistrplus)
```

ENUNCIADO

Parte 1

Se tiene la base de datos de la Encuesta Nacional de Salud (ENS), herramienta que utiliza el Ministerio de Salud para saber qué enfermedades y qué tratamientos están recibiendo hombres y mujeres de 15 años y más que viven en Chile.

La información que arroja esta encuesta es de vital importancia para formular los planes de prevención, atención y políticas de salud para las personas que las necesitan.

La base de datos está en el archivo `ENS.xlsx` (también disponible como `ENS.txt` y `ENS.rdata`, con la misma información) y contiene las siguientes variables:

EDAD: Edad en años
NEDU: Años de estudios 1: <8 años, 2: 8-12 años, 3: >12 años
ZONA: Zona 1: Urbano, 2: Rural
REGION: 1: I. Tarapacá, 2: II. Antofagasta, 3: III. Atacama, 4: IV. Coquimbo, 5: V. Valparaíso, 6: VI. L. Bdo. O'Higgins, 7: VII. Maule, 8: VIII. Bíobío, 9: IX. La Araucanía, 10: X. Los Lagos, 11: XI. Aysén, 12: XII. Magallanes y Antártica, 13: XIII. Metropolitana, 14:

XIV. Los Ríos, 15: XV. Arica y Parinacota
SEXO: Sexo 1: Hombre, 2: Mujer
GLUBASAL: Nivel de azúcar mg/dL
COLES: Colesterol Total mg/dL
HDL: Nivel de colesterol HDL mg/dL
LDL: Nivel de colesterol LDL mg/dL
TGD: Nivel de triglicéridos mg/dL
PAD: Presión arterial diastólica
PAS: Presión arterial sistólica
HTA: Hipertenso (1: Si, 0: No)
SM: Síndrome Metabólico (1: Si, 0: No)
FUMADOR: Fumador actual (1: Si, 0: No)
DIABETES: Diabético (1: Si, 0: No)
PESO: Peso en kilos
TALLA: Talla en cm
CINTURA: Circunferencia de cintura en cm
CUELLO: Circunferencia de cuello en cm

Pregunta 1 (1.5 puntos)

Seleccione las variables $\mathbf{X}$ = **PESO**, $\mathbf{Y}$ = **COLES** y $\mathbf{Z}$ = **GLUBASAL**.
Luego, si se asume que las variables $\mathbf{X}$ e $\mathbf{Y}$ se distribuyen Normal Bivariado, es decir:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_{\in}(\mu, \Sigma)$$

donde μ = **mu** es el vector de medias y $\mathbf{\Sigma}$ = **Sigma** la matriz de covarianza. Obtenga en primera instancia el vector de medias muestral, es decir, **mu = c(mean(X), mean(Y))** y la matriz de covarianza muestral, es decir, **Sigma = cov(cbind(X, Y))**.
Calcule las siguientes probabilidades:

a) (0.5 puntos)

La probabilidad que las personas que tienen peso menor a 75 kg dado que su nivel de colesterol es igual a 200 mg/dL, es decir: $P(X < 75|Y = 200)$

b) (0.5 puntos)

La probabilidad que las personas que tienen peso menor a 75 kg dado que su nivel de colesterol es menor a 200 mg/dL, es decir: $P(X < 75|Y < 200)$

c) (0.5 puntos)

Si se asume que las variables **X, Y, Z** se distribuyen Normal Trivariado. Obtenga en primera instancia el vector de medias muestral, es decir, **mu = c(mean(X), mean(Y), mean(Z))** y la matriz de covarianza muestral, es decir, **Sigma = cov(cbind(X, Y, Z))**. Obtener la probabilidad de las personas que tienen peso menor a 75 kg y nivel de colesterol menor a 200 mg/dL dado que su nivel de azúcar es mayor a 100 mg/dL, es decir:
$P(X < 75, Y < 200 | Z > 100)$

Pregunta 2 (1.5 puntos)

Se desea estimar la distribución adecuada de los niveles de colesterol, $\mathbf{Y} = \mathbf{COLES}$, como es una variable con soporte positivo y asimetría a la derecha, lo cual se puede verificar descriptivamente si graficamos el histograma. Es así que, se puede asumir que la variable proviene de una distribución Gamma(α, β); se pide estimar ambos parámetros, para esto puede utilizar la función `fitdist()` de R:

a) (0.50 puntos)

¿Cuál es la estimación por el método de momentos de α, donde se sabe que β es desconocido?

b) (0.50 puntos)

¿Cuál es la estimación por el método de máxima verosimilitud de α si en este caso se sabe por información previa que el parámetro de tasa es conocido, es decir, $\beta = 0.12$?

c) (0.50 puntos)

Determine cuál estimador de α es más eficiente entre el estimador de momentos (mme) y el de máxima verosimilitud (mle).

Pregunta 3 (0.75 puntos)

Suponiendo que se asume que el $\mathbf{X} = \mathbf{PESO}$ sigue una distribución Normal. Se busca verificar si el promedio poblacional de la variable PESO es menor que 72.3 kg. Es así que, se pide realizar el siguiente test de hipótesis, con nivel de significancia de $\alpha = 0.02$:

a)(0.25 puntos)

¿Cuál es el estadístico asociado al test?

b)(0.25 puntos)

¿Cuál es el p-valor asociado al test?

c)(0.25 puntos)

¿Qué decisión se toma para la hipótesis nula Ho?

Pregunta 4 (0.75 puntos)

Por otro lado, si suponemos que $\mathbf{Y} = \mathbf{COLES}$ sigue una distribución Normal. Se quiere verificar si el promedio poblacional de la variable **COLES** es distinto de 193 mg/dL. Es así que, se pide realizar el siguiente test de hipótesis, con nivel de significancia de $\alpha = 0.05$:

a) (0.25 puntos)

¿Cuál es el estadístico asociado al test?

b) (0.25 puntos)

¿Cuál es el p-valor asociado al test?

c) (0.25 puntos)

¿Qué decisión se toma para la hipótesis nula Ho?

Pregunta 5 (1.5 puntos)

Para esta pregunta no se necesita trabajar con ninguna base de datos. Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de tamaño $\mathbf{n = 200}$, donde $X_i \sim \text{Lognormal}(\lambda, \zeta^2)$, con $\lambda = 1$ y $\zeta = 0.8$. Se sabe que $\mathbb{E}[x] = \exp(\lambda + \zeta/2)$ y $\text{Var}(X) = \mathbb{E}[X]^2 \cdot (\exp(\zeta^2)1)$. Además, se define que $\bar{X}$ es el promedio muestral.
Aplicando el Teorema del Límite Central (TLC), se pide hallar $P(3.5 < \bar{X} < 4)$: