**Estadística (M)**
Segundo Trabajo Práctico
Un análisis mediante intervalos de confianza

**Miembros del grupo:**
**Cantidad Total de Hojas:**

*La resolución del TP incluirá*

1. *El script que permite efectuar los cálculos que deberá enviarse por mail para verificar el programa.*
2. *Un documento con las respuestas a las preguntas planteadas que puede entregarse por alguna de las siguientes vías*
   - *Personalmente, la copia impresa que incluya todos los gráficos necesarios.*
   - *Por mail, enviando un doc, LATEX o pdf que contenga los gráficos necesarios.*

---

**- Justifique todas sus respuestas -**

---

Sea $\mathbf{Z} = (X, Y)$, nos interesa estimar el coeficiente de correlación $\rho = \rho_{(X,Y)}$ entre $X$ e $Y$. Sean $\mathbf{Z}_i = (X_i, Y_i)$ i.i.d. tales que $\mathbf{Z}_i \sim \mathbf{Z}$. El estimador usual de $\rho$ es

$$\widehat{\rho} = \widehat{\rho}((X_1, Y_1), \ldots, (X_n, Y_n)) = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \ \sum_{i=1}^n (Y_i - \overline{Y})^2}} = \frac{\widehat{c}_{X,Y}}{\widehat{\sigma}_X \, \widehat{\sigma}_Y}$$

donde

$$\widehat{c}_{X,Y} \ = \ \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y}) \qquad \widehat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2 \qquad \widehat{\sigma}_Y^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})^2$$

Sea $\sigma_X^2 = \textsc{Var}(X_1)$ y $\sigma_Y^2 = \textsc{Var}(Y_1)$ y defina $\widetilde{X}_1 = (X_1 - \mathbb{E}(X_1))/\sigma_X$ y $\widetilde{Y}_1 = (Y_1 - \mathbb{E}(Y_1))/\sigma_Y$.

1. Muestre que $\widehat{\rho}$ y $\rho$ son invariantes por la siguiente transformación $h : \mathbb{R}^2 \to \mathbb{R}^2$, $(v, w) = h(x, y) = (ax + b, cy + d)$ con $a > 0$ y $c > 0$.

2. Deduzca que basta encontrar la distribución asintótica de $\widehat{\rho}$ cuando $\sigma_X^2 = \sigma_Y^2 = 1$.

3. Se sabe que si $\mathbb{E}(X_1^4) < \infty$ y $\mathbb{E}(Y_1^4) < \infty$, entonces $\sqrt{n}(\widehat{\rho} - \rho) \xrightarrow{D} N(0, v_\rho = \widetilde{\mathbf{a}}^{\mathrm{T}} \widetilde{\mathbf{\Sigma}} \widetilde{\mathbf{a}})$, donde

$$\begin{aligned} \widetilde{\mathbf{a}} \ &= \ \left(-\frac{\rho}{2}, -\frac{\rho}{2}, 1\right)^{\mathrm{T}} \\ \widetilde{\mathbf{\Sigma}} \ &= \ \begin{pmatrix} \mathrm{Cov}(\widetilde{X}_1^2, \widetilde{X}_1^2) & \mathrm{Cov}(\widetilde{X}_1^2, \widetilde{Y}_1^2) & \mathrm{Cov}(\widetilde{X}_1^2, \widetilde{X}_1\widetilde{Y}_1) \\ \mathrm{Cov}(\widetilde{X}_1^2, \widetilde{Y}_1^2) & \mathrm{Cov}(\widetilde{Y}_1^2, \widetilde{Y}_1^2) & \mathrm{Cov}(\widetilde{Y}_1^2, \widetilde{X}_1\widetilde{Y}_1) \\ \mathrm{Cov}(\widetilde{X}_1^2, \widetilde{X}_1\widetilde{Y}_1) & \mathrm{Cov}(\widetilde{Y}_1^2, \widetilde{X}_1\widetilde{Y}_1) & \mathrm{Cov}(\widetilde{X}_1\widetilde{Y}_1, \widetilde{X}_1\widetilde{Y}_1) \end{pmatrix} \end{aligned}$$

de donde se deduce que

$$v_\rho \;=\; \frac{\rho^2}{4}\left[\mathbb{E}\widetilde{X}_1^4 + 2\mathbb{E}\widetilde{X}_1^2\widetilde{Y}_1^2 + \mathbb{E}\widetilde{Y}_1^4\right] - \rho\left[\mathbb{E}\widetilde{X}_1^3\widetilde{Y}_1 + \mathbb{E}\widetilde{X}_1\widetilde{Y}_1^3\right] + \mathbb{E}\widetilde{X}_1^2\widetilde{Y}_1^2 \;.$$

Muestre que en el caso normal $v_\rho = (1-\rho^2)^2$.

**Sugerencia**: *Defina* $(U,V) = (\widetilde{X}_1 - \rho\widetilde{Y}_1, \widetilde{Y}_1)$ *entonces,* $(U,V)$ *tiene distribución normal bivariada. Obtenga* $\mathbb{E}(U)$, $\mathbb{E}(U^2)$, $\mathbb{E}(U^3)$ *y* $\mathbb{E}(U^4)$ *así como* $\mathbb{E}(V)$, $\mathbb{E}(V^2)$, $\mathbb{E}(V^3)$ *y* $\mathbb{E}(V^4)$.

*Calcule* $\textsc{Cov}(U,V)$ *y úselo para obtener* $\mathbb{E}\widetilde{X}_1^2\widetilde{Y}_1^2$, $\mathbb{E}\widetilde{X}_1^3\widetilde{Y}_1$ *y* $\mathbb{E}\widetilde{X}_1\widetilde{Y}_1^3$.

4. ¿Cómo estimaría $v_\rho$ en el caso normal y en el caso en que las observaciones no sean normales?

5. Obtenga un intervalo de confianza asintótico de nivel $1-\alpha$ para $\rho$ cuando las observaciones son normales.

   ¿Puede dar uno en el caso general?

6. Sea

$$\theta = g(\rho) = \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right)$$

   y $\widehat{\theta} = g(\widehat{\rho})$. A partir de 3), obtenga la distribución asintótica de $\sqrt{n}(\widehat{\theta}-\theta)$. ¿Qué observa en el caso normal?

7. Obtenga un intervalo de confianza de nivel $1-\alpha$ para $\theta$ en el caso normal y a partir de él deduzca uno para $\rho$.

8. En un mismo gráfico, grafique las longitudes de ambos intervalos como función de $\widehat{\rho} \in (-1;1)$ para distintos valores de $n$, por ejemplo, $n = 50$, $n = 100$ y $n = 200$ y $\alpha = 0.05$ y $\alpha = 0.01$. Es decir, para cada $n$ y $\alpha$ haga un gráfico donde en rojo grafica la longitud del intervalo dado en 5) y en azul la del dado en 7). ¿Qué observa?

9. Considere el siguiente estudio de simulación. Fijados $\alpha$, el tamaño muestral $n$ y el coeficiente de correlación $\rho$,

   (a) Utilizando la función `rmvnorm` de la librería `mvtnorm`, genere $n$ observaciones $\mathbf{Z}_i = (X_i, Y_i)$ i.i.d. tales que $\mathbf{Z}_i \sim N(\mathbf{0}, \mathbf{\Sigma})$ donde $\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

   (b) Calcule los intervalos obtenidos en 5) y 7), sus longitudes respectivas $L_5$ y $L_7$ y reporte si el intervalo contiene a $\rho$.

   (c) Repita a) y b) $NR = 1000$ veces y guarde:

      i. el promedio de los valores $L_5$ y $L_7$ obtenidos en cada replicación que indicaremos $EL_5$ y $EL_7$.

ii. la cobertura $C_5$ y $C_7$ de los intervalos obtenidos en 5) y 7). Más precisamente, calcule el porcentaje de replicaciones para las cuales el intervalo obtenido en 5) contiene a $\rho$ que llamaremos $C_5$ y en forma similar, $C_7$.

iii. el porcentaje de veces que el extremo inferior del intervalo es menor que -1 que indicaremos $M_{-1,5}$ y $M_{-1,7}$.

iv. el porcentaje de veces que el extremo superior del intervalo es mayor que 1 que indicaremos $M_{1,5}$ y $M_{1,7}$.

Los promedios $EL_5$ y $EL_7$ obtenidos en c) aproximan la longitud esperada de los intervalos obtenidos en 5) y 7), respectivamente, mientras que la cobertura da una aproximación al nivel de confianza. Las cantidades $M_{-1,5}$, $M_{-1,7}$, $M_{1,5}$ y $M_{1,7}$ son de interés ya que la correlación pertenece al intervalo $(-1, 1)$.

Al inicio de la simulación fije la semilla en 123 y en cada una de las situaciones que se describe a continuación utilice la misma semilla. Fije $\alpha = 0.05$ o sea, considere intervalos de confianza de nivel 0.95.

Reporte en dos tablas como las Tablas **??** y **??** los valores obtenidos cuando: $n = 20, 100$ y $\rho = -0.75, -0.5, 0, 0.5, 0.75$. Qué conclusiones puede sacar de los resultados reportados en la Tablas **??** y **??**? Recomendaría uno de los dos intervalos por sobre el otro?

| $\rho$ | $n = 20$ | | | | $n = 100$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $EL_5$ | $C_5$ | $EL_7$ | $C_7$ | $EL_5$ | $C_5$ | $EL_7$ | $C_7$ |
| $-0.75$ | | | | | | | | |
| $-0.5$ | | | | | | | | |
| 0 | | | | | | | | |
| 0.5 | | | | | | | | |
| 0.75 | | | | | | | | |

Table 1: Valores de longitud media y cobertura para los intervalo de confianza de nivel 0.95 dados en 5) y 7).

| $\rho$ | $n = 20$ | | | | $n = 100$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $M_{-1,5}$ | $M_{1,5}$ | $M_{-1,7}$ | $M_{1,7}$ | $M_{-1,5}$ | $M_{1,5}$ | $M_{-1,7}$ | $M_{1,7}$ |
| $-0.75$ | | | | | | | | |
| $-0.5$ | | | | | | | | |
| 0 | | | | | | | | |
| 0.5 | | | | | | | | |
| 0.75 | | | | | | | | |

Table 2: Porcentaje de veces que el extremo inferior y superior del intervalo es menor que -1 o mayor que 1 para los intervalo de confianza de nivel 0.95 dados en 5) y 7).