

# Estudio de caso # 1

Universidad Externado de Colombia

Departamento de Matemáticas

Estadística 1

Juan Sosa, Ph. D.

August 13, 2018

## Instrucciones generales

- Puede hacer el examen solo o puede asociarse con otra persona, entendiendo que la calificación del examen será la misma para ambas personas.
- El reporte final se debe enviar a más tardar el **lunes 13 de agosto de 2018** a las **11:59 p.m.** a la cuenta de correo:  
`juan.sosa@uexternado.edu.co`.
- Reportar las cifras utilizando la **cantidad adecuada de decimales**, dependiendo de lo que se quiera mostrar y las necesidades del problema.
- Numerar figuras y tablas (<http://unilearning.uow.edu.au/report/1fi.html>) y proporcionarles un **tamaño adecuado** que no distorsione la información que estas contienen.
- El archivo del reporte final debe ser un **archivo pdf** con el siguiente formato: Letra Calibri, tamaño 12, interlineado sencillo con espacio entre párrafos y texto justificado. Márgenes: Normal. Tamaño: Carta. Orientación: Vertical.

- Especificar el software donde se llevó a cabo el computo e **incluir el código** correspondiente como un anexo al final del reporte con el siguiente formato: Letra Courier New, tamaño 10, interlineado sencillo.
- El objetivo principal de este trabajo es la claridad lógica y la interpretación de los resultados. **El informe no necesita ser extenso.** Recuerde ser minimalista escribiendo el reporte. Se deben incluir solo aquellos gráficos y tablas (¡y valores en la tabla!) que son relevantes para la discusión.
- Hacer el informe ya sea en inglés o español. No ambos!
- **Cualquier evidencia de plagio o copia se castigará severamente** tal y como el reglamento de la Universidad Externado de Colombia lo estipula.

Si está claro que (por ejemplo) dos grupos han trabajado juntos en una parte de un problema que vale 20 puntos, y cada respuesta habría ganado 16 puntos (si no hubiera surgido de una colaboración ilegal), entonces cada grupo recibirá 8 de los 16 puntos obtenidos colectivamente (para una puntuación total de 8 de 20), **y me reservo el derecho de imponer penalidades adicionales a mi discreción.**

Si un grupo resuelve un problema por su cuenta y luego comparte su solución con cualquier otro grupo (porque rutinariamente Usted hace esto, o por lástima, o bondad, o por cualquier motivo que pueda creer tener; no importa!), Usted es tan culpable de colaboración ilegal como la persona que tomó su solución, y ambos recibirán la misma penalidad. Este tipo de cosas es necesario hacerlas ya que muchas personas no hacen trampa, y debo asegurarme de que sus puntajes son obtenidos de manera genuina. En otras clases, personas perdieron la clase debido a una colaboración ilegal; **no deje que le suceda a Usted!**

## Ingresos y nivel educativo

Considere la base de datos en `TWINS.txt` correspondiente a la información acerca del ingreso, educación y otras variables socioeconómicas de gemelos monocigóticos (gemelos provenientes de un mismo ovulo). Estos datos fueron recolectados por Ashenfelter and Krueger con el fin de investigar la relación entre el nivel educativo y el ingreso.

Cada par de gemelos fue denominado aleatoriamente como “gemelo 1” y el otro como “gemelo 2”. Algunas variables también son dadas en pares. Por ejemplo, las variables `EDUCL` y `EDUCH` contienen la cantidad de años de estudio reportada por el gemelo 1 y el gemelo 2, respectivamente. La descripción completa de las variables y del diseño de la investigación se encuentra en <http://www.stat.ucla.edu/projects/datasets/twins-explanation.html>.

1. Importar esta base de datos en R.
2. ¿Cuántos registros en la base de datos? ¿Cuántas variables?
3. Remover todos los registros con algún dato faltante. ¿Cuántos registros con información completa?
4. Clasificar las variables `EDUCL` y `HRWAGEL` según su naturaleza y escala de medición.

**Trabajar los siguientes items utilizando únicamente los registros con información completa.**

5. Hacer una tabla de frecuencias (relativas) y un gráfico de barras para `WHITEH`, `WHITEL`, `MALEH`, y `MALEL`.
6. Las variables `EDUCL`, `EDUCH`, `HRWAGEL` y `HRWAGEH` contienen los años de educación y el salario por hora reportado por el gemelo 1 y gemelo 2, respectivamente. Describir gráfica y numéricamente cada variable por separado, es decir, para cada variable hacer un boxplot (dos gráficos son necesarios: uno con los boxplots de años de educación y otro con los boxplots de salario), calcular las medidas de localización y de dispersión. Una forma de presentar las medidas estadísticas organizadamente es por medio de la siguiente tabla:

Medida	Educ. 1	Educ. 2	Salario 1	Salario 2
Mínimo				
Máximo				
Cuartíl 1				
Mediana				
Cuartíl 3				
Media				
Rango				
Rango Intercuartílico				
Desv. Estándar				
Coef. Variación				

¿Cuál es el porcentaje de datos atípicos en cada caso? Comparar e interpretar los resultados obtenidos.

7. Hacer un dispersograma de:

- (a) HRWAGEL frente a AGE.
- (b) HRWAGEH frente a AGE.
- (c) HRWAGEL frente a EDUCL.
- (d) HRWAGEH frente a EDUCH.
- (e) HRWAGEL frente a HRWAGEH.

Calcular e interpretar el coeficiente de correlación de Pearson en cada caso (ver por ejemplo Sosa et al. 2012, Sección 6.5, p. 120). Puede presentar los gráficos en un sólo prisma de dispersogramas y todos los coeficientes de correlación completando la siguiente tabla:

Variables	Edad	Educ. 1	Educ. 2	Salario 1	Salario 2
Edad					
Educ. 1					
Educ. 2					
Salario 1					
Salario 2					