Estudio de caso # 3

Universidad Externado de Colombia

Departamento de Matemáticas

Estadística 1

Juan Sosa, Ph. D.

October 9, 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 **viernes 5 de octubre 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!**

1 Créditos

La base de datos `creditos.txt` contiene información personal y financiera de personas y compañías a las que un banco muy reconocido otorgó algún tipo de crédito en Colombia durante 2017. Dado que los datos corresponden a información muy sensible de los clientes, la base de datos ha sido anonimizada, es decir, los registros no tienen ninguna clase de identificador. El objetivo principal de este trabajo es estudiar la distribución de los ingresos de los clientes de este banco.

Consolidación de la información

Originalmente esta base de datos contiene un total 81,536 registros y 21 variables.

```
# importar datos
creditos <- read.delim("C:/Users/Juan Camilo/Dropbox/UE/Estadistica 2  II-2018/data/creditos.txt")

# dimension de la base de datos
dim(creditos)
```

El primer paso consiste en conformar una nueva base de datos con todos los registros que tengan información completa correspondientes a las variables ingresos, género y monto otorgado, nivel de estudios y estrato. Solamente nueve registros tenían algún dato faltante en alguna de estas variables.

```
# nueva base de datos
creditos <- creditos[ , c("INGRESOS_DECLARADOS_TOTA", "SEXO", "MONTO_TOTAL_OTORGADO", "NIVEL_ESTUDIOS", "ESTRATO")]

creditos <- creditos[complete.cases(creditos), ]

dim(creditos)
```

Para manejar de forma más eficiente la información se recomienda cambiar el nombre de las variables y además transformar la escala de medición de los ingresos y el monto otorgado dividiendo todos los valores por 1000000 de modo que la variables queden medidas en millones.

```
# renombrar variables
colnames(creditos) <- c("ingresos", "genero", "monto", "estudios", "estrato")

# redefinir escala de los ingresos para trabajar en millones
```

```
creditos$ingresos <- creditos$ingresos/1000000

creditos$monto <- creditos$monto/1000000
```

Como se quiere estudiar la distribución de los ingresos de aquellos clientes a los que se les otorgó un crédito por menos de 100 millones, es necesario implementar el filtro correspondiente. Adicionalmente, con el propósito de descartar datos extremos que eventualmente puedan sesgar los resultados, en este estudio se descartan aquellos clientes con ingresos superiores a 25 millones de pesos. Así, la base de datos final que se usará para el análisis tiene 61,839 registros con información completa.

```
# filtrar por ingresos inferiores a 100M y credito hipotecario o de vehículo
creditos <- creditos[(creditos$monto < 100) & (creditos$ingresos < 25), ]

# numero de registros
nrow(creditos)
```

De otro lado, se recomienda adjuntar la base de datos para trabajar con las variables directamente sin necesidad de hacer referencia a la base de datos `creditos` todo el tiempo.

```
# adjuntar la base datos
attach(creditos)
```

Análisis exploratorio de los datos

Antes de hacer inferencia estadística sobre los parámetros de los ingresos, siempre es preciso describir numérica y gráficamente la información.

1. Hacer una tabla de frecuencias relativas para la variable género.
2. Hacer una tabla de frecuencias relativas y un gráfico de barras para el nivel educativo de los hombres.
3. Hacer una tabla de frecuencias relativas y un gráfico de barras para el estrato de los hombres.
4. Hacer un histograma y un diagramama de caja para el ingreso de los hombres.
5. Repetir los númerales 2. a 4. para las mujeres.
6. Completar la siguiente tabla acerca de los ingresos para hombres y mujeres:

Medida	Hombres	Mujeres
Mínimo		
Máximo		
Cuartíl 1		
Mediana		
Cuartíl 3		
Media		
Rango		
Rango Intercuartílico		
Desv. Estándar		
Coef. Variación		
Indice Yule-Bowley		
Coef. Asim. Fisher		

7. Comentar los resultados obtenidos en los numerales anteriores. ¿Parece haber diferencias en el nivel educativo y el estrato entre hombres y mujeres? Con base en los resultados, ¿existe algún indicio de una brecha salarial entre hombres y mujeres?

Nota 1: Para acceder a la información de las varaibles por género es preciso hacer filtros. Por ejemplo, los datos de los ingresos de los hombres se pueden obtener por medio del siguiente filtro:

```
ingresos.hombres <- ingresos[genero == "H"]
```

Es preciso que la base de datos `cretidos` se haya adjuntado previamente para que este filtro funcione correctamente.

Nota 2: La codificación de la variable nivel educativo es como sigue: BAS = básica primaria; MED = bachillerato; TEC = ténico; PRF = entrenamiento especializado en otros oficios; NOG = pregrado incompleto; UNIV = pregreado completo; POS = especialización o maestria; DOC = doctorado.

Cálculo de probabilidades

Con el fin de evaluar algunos factores que puedan determinar el ingreso, a continuación se proponen algunos cálculos de probabilidades (desde la perspectiva frecuentista de probabilidad).

1. Categorizar (construir intervalos) la variable ingreso usando los siguientes intervalos: menos de 1 SMMLV, entre 1 y 3 SMMLV, entre 3 y 5 SMMLV, entre 5 y 10 SMMLV, más de 10 SMMLV, donde "SMMLV" indica Salarios Mínimos Mensuales Legales Vigentes.

   Nota 1: usando la función `cut` utilice la opción `include.lowest = TRUE` para incluir los límites inferiores de cada intervalo.

   Nota 2: un SMMLV equivale a $ 781,242.00.

2. Hacer una tabla de bidimensional de frecuencias relativas para ingreso categorizado (columnas) frente a género (filas). A partir de esta tabla construir los perfiles files fila. Hacer el diagrama de barras compuesto correspondientes.

3. Comentar los resultados obtenidos en los numerales anteriores. ¿Parecen haber diferencias importantes entre hombres y mujeres respecto a los ingresos y el nivel educativo?

Teorema de Bayes

Sean H_1, H_2, H_3 y H_4 los eventos que relacionan los ingresos de una persona seleccionada al azar en esta muestra respecto a los cuartiles; esto es:

- H_1 : "los ingresos del cliente son menores al cuartíl 1".
- H_2 : "los ingresos del cliente son menores al cuartíl 2, pero mayores al cuartíl 1".
- H_3 : "los ingresos del cliente son menores al cuartíl 3, pero mayores al cuartíl 2".
- H_4 : "los ingresos del cliente son menores al cuartíl 4, pero mayores al cuartíl 3".

1. Por definición, se tiene que $\Pr[H_1] = \Pr[H_2] = \Pr[H_3] = \Pr[H_4] = 0.25$. ¿Por qué?

2. Observe que H_1, H_2, H_3 y H_4 son una partición del espacio muestral. ¿Por qué?

3. Sea E el evento "el cliente tiene educación universitaria". **Usando la base de datos**, **calcular e interpretar** las siguientes probabilidades: $\Pr[E \mid H_1]$, $\Pr[E \mid H_2]$, $\Pr[E \mid H_3]$ y $\Pr[E \mid H_4]$. ¿Estas probabilidades deben sumar 1? ¿Por qué?

4. Considere la distribución de los ingresos de las personas con educación universitaria. **Usando el teorema de Bayes**, **calcular e interpretar** las siguientes probabilidades: $\Pr[H_1 \mid E]$, $\Pr[H_2 \mid E]$, $\Pr[H_3 \mid E]$ y $\Pr[H_4 \mid E]$. ¿Estas probabilidades deben sumar 1? ¿Por qué? Observe que las probabilidades $\Pr[H_1 \mid E]$, $\Pr[H_2 \mid E]$, $\Pr[H_3 \mid E]$ y $\Pr[H_4 \mid E]$ difieren marcadamente de $\Pr[H_1]$, $\Pr[H_2]$, $\Pr[H_3]$, y $\Pr[H_4]$, respectivamente. ¿Por qué?

5. ¿Qué se puede concluir de la distribución de los ingresos de los clientes con grado universitario?

Independencia

Considere los eventos:

- A: "el cliente es hombre".
- B: "el cliente es mujer".
- F: "el cliente tiene educación superior".
- G: "el cliente tiene ingresos superiores a 3 SMMLV".

1. ¿Los eventos A y F son independientes? ¿Por qué?
2. ¿Los eventos B y F son independientes? ¿Por qué?
3. ¿Los eventos A y G son independientes? ¿Por qué?
4. ¿Los eventos B y G son independientes? ¿Por qué?
5. ¿Que implicaciones tienen estos resultados sobre la educación superior y los ingresos?