Caso de estudio # 3

Estadística 2

Juan Sosa, Ph. D.

25 septiembre 2018

Instrucciones

- 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.
- 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!

Distribución de ingresos por género

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 utilizano diferetes técnicas estadísticas. Especificamente, se quiere investigar si exite una brecha salarial entre los hombres y las mujeres del banco a los cuales se les otorgó algún tipo de crédito por menos de cien millones de pesos. Por tal motivo, se quire respondiender la siguientes pregunta: ¿Existen direncias significativas entre el ingreso promedio de hombres y mujeres?

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)
```

```
## [1] 81536    21
```

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 (no es obligatorio retener las variables nivel de estudios y estrato, pero es interezante hacerlo para tener en cuenta otros aspectos en el análisis). No hay registros con algún dato faltante en alguna de estas variables.

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

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

dim(creditos)
```

```
## [1] 81536     5
```

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 61,845 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)
```

```
## [1] 61845
```

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)
```

Anális exploratorio de los datos

Antes de hacer inferencia estadística sobre los parámetros de los ingresos, siempre es preciso describir numérica y graficamente 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 un histograma y un diagramama de caja para el ingreso de los hombres.
4. Repetir los númerales 2. a 3. para las mujeres.
5. Completar la siguiente tabla:

Ingresos	Mín.	Máx.	Cuartíl 1	Mediana	Cuartíl 3	Promedio	Desv. Est.	Coef. Var. (%)
Hombres								
Mujeres								

6. 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.

Modelo Lognormal

La distribución Lognormal es popular para modelar ingresos dado que con este modelo es posible producir distribuciones con muchas variedades de sesgo. Ver por ejemplo, https://en.wikipedia.org/wiki/Log-normal_distribution para más información acerca de este modelo probabilístico. Así, considere el modelo Lognormal

$$\left\{ \begin{array}{c} Y_i \overset{\text{IID}}{\sim} \text{Lognormal}(\mu, \sigma^2) \\ i = 1, \ldots, n \end{array} \right\}, \tag{1}$$

donde Y_i corresponden a los ingresos del individuo i y n es el tamaño de la muestra. Observe que en este caso $Y_i \sim \text{Lognormal}(\mu, \sigma^2)$ significa que $\log Y_i \sim \text{Normal}(\mu, \sigma^2)$ donde $-\infty < \mu < \infty$ y $\sigma^2 > 0$ son los parámetros del modelo que serán objeto de inferencia.

Para el modelo Lognormal es posible demostrar que si $Y \sim \text{Lognormal}(\mu, \sigma^2)$, entonces:

- La función de densidad de Y es

$$f(y; \mu, \sigma^2) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\log y - \mu)^2}{2\,\sigma^2}\right], \quad y > 0\,.$$

- El valor esperado de Y es

$$\mathbb{E}(Y) = \int_0^\infty y f(y; \mu, \sigma^2)\,\mathrm{d}y = \exp\left(\mu + \frac{\sigma^2}{2}\right).$$

Para obtener los estimadores de máxima verosimilitud de μ and σ^2, es posible seguir el procedimiento de maximizaición usual como sigue:

a. Obtener la función de verosimilitud:

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} f(y_i; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{y_i \sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\log y_i - \mu)^2}{2\,\sigma^2} \right] .$$

b. Obtener la función de log-verosimilitud:

$$\ell(\mu, \sigma^2) = \log L(\mu, \sigma^2) .$$

c. Derivar la función de log-verosimilitud respecto a μ y σ^2, igualar estas derivadas parciales a 0 y resolver el sistema de acuaciones para μ y σ^2.
d. Probar que las soluciones obtenidas en el numeral anterior maximizan la función de log-verosimilitud.

Este procedimiento conlleva a los siguientes estimadores de maxima de verosimilitud de μ y σ^2:

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} \log Y_i \quad \text{y} \quad \hat{\sigma}^2_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} \left(\log Y_i - \hat{\mu}_{\text{MLE}} \right)^2 . \tag{1}$$

Además, tambien es posible demostrar que la información observada de Fisher (menos la matriz de segundas derivadas parciales de la función de log-verosimilitud respecto a μ y σ^2 evaluada en $\mu = \hat{\mu}_{\text{MLE}}$ y $\sigma^2 = \hat{\sigma}^2_{\text{MLE}}$) está dada por

$$\hat{I} = \begin{pmatrix} \frac{n}{\hat{\sigma}^2_{\text{MLE}}} & 0 \\ 0 & \frac{n}{2(\hat{\sigma}^2_{\text{MLE}})^2} \end{pmatrix} ,$$

y por lo tanto, la matriz de varianza-covarianza de los estimadores de máxima verosimilitud $\hat{\mu}_{\text{MLE}}$ y $\hat{\sigma}^2_{\text{MLE}}$ es

$$\hat{I}^{-1} = \begin{pmatrix} \frac{\hat{\sigma}^2_{\text{MLE}}}{n} & 0 \\ 0 & \frac{2(\hat{\sigma}^2_{\text{MLE}})^2}{n} \end{pmatrix} . \tag{2}$$

Se quiere ajustar la distribución Lognormal para modelar tanto los ingresos de los hombres como de las mujeres de manera independiente, es decir, se asume que las mediciones de los hombres son independientes de las mediciones de las mujeres.

1. Usar las fórmulas de las ecuaciones (1) y (2) para completar la siguiente tabla:

Ingresos	$\hat{\mu}_{\text{MLE}}$	$\hat{\sigma}^2_{\text{MLE}}$	$\mathbb{V}\text{ar}(\hat{\mu}_{\text{MLE}})$	$\mathbb{V}\text{ar}(\hat{\sigma}^2_{\text{MLE}})$
Hombres				
Mujeres				

2. Tanto para hombres como muejres, hacer nuevamente el histograma de los ingresos (igual que en el análisis exploratorio de datos), esta vez superponiendo una curva de la función de densidad de la distribución Lognormal usando los valores respectivos de $\hat{\mu}_{\text{MLE}}$ y $\hat{\sigma}^2_{\text{MLE}}$; ¿la curva parece un modelo razonable de los ingresos correspondientes? Explicar brevemente.

Nota: la función de densidad del modelo Lognormal en `R` es `dlnorm`.

3. Para cada grupo, usando los valores estimados de μ y σ^2 obtenidos en el numeral 1., estimar las probabilidades de la siguiente tabla:

Ingresos	$\widehat{\mathbb{P}\text{r}}(Y \leq 1)$	$\widehat{\mathbb{P}\text{r}}(Y \leq 3)$	$\widehat{\mathbb{P}\text{r}}(Y \leq 5)$	$\widehat{\mathbb{P}\text{r}}(Y \leq 10)$
Hombres				
Mujeres				

Ingresos	$\widehat{\mathbb{P}\mathrm{r}}(Y \leq 1)$	$\widehat{\mathbb{P}\mathrm{r}}(Y \leq 3)$	$\widehat{\mathbb{P}\mathrm{r}}(Y \leq 5)$	$\widehat{\mathbb{P}\mathrm{r}}(Y \leq 10)$

Comentar brevemente los resultados obtenidos. ¿Estos resultados dan un indicio acerca de una posible brecha salarial entre hombres y mujeres?

Nota: la función de distribución acumulada del modelo Lognormal para calcular los valores de la tabla en `R` es `plnorm`.

4. Tanto para hombres como mujeres, construir un intervalo de confianza para el ingreso promedio utilizando el modelo Lognormal. Para ello, se defina θ como el valor esperado (promedio poblacional) de los ingresos, esto es,

$$\theta = \mathbb{E}(Y) = \exp\left(\mu + \frac{\sigma^2}{2}\right). \qquad (3)$$

Así, por la propiedad de invarianza funcional de los estimadores de máxima verosimilitud, se tiene que el estimador (puntual) de máxima verosimilitud de θ es

$$\hat{\theta}_{\mathrm{MLE}} = \exp\left(\hat{\mu}_{\mathrm{MLE}} + \frac{\hat{\sigma}^2_{\mathrm{MLE}}}{2}\right).$$

Además, usando el método delta, se obtiene que la varianza de $\hat{\theta}_{\mathrm{MLE}}$ es

$$\mathbb{V}\mathrm{ar}(\hat{\theta}_{\mathrm{MLE}}) = (\nabla \hat{g})^T \, \hat{I}^{-1} \, (\nabla \hat{g}) \qquad (4)$$

donde $\hat{I}$ es la información observada de Fisher de la ecuación (2) y $\nabla \hat{g}$ es el vector de derivadas parciales de $g = \exp\left(\mu + \frac{\sigma^2}{2}\right)$ respecto a μ y σ^2 evaluado en $\mu = \hat{\mu}_{\mathrm{MLE}}$ y $\sigma^2 = \hat{\sigma}^2_{\mathrm{MLE}}$, es decir,

$$\nabla \hat{g} = \left[\begin{array}{c} \frac{\partial g}{\partial \mu} \\ \frac{\partial g}{\partial \sigma^2} \end{array}\right]_{\mu = \hat{\mu}_{\mathrm{MLE}}, \sigma^2 = \hat{\sigma}^2_{\mathrm{MLE}}}$$

Usando las ecuaciones (3) y (4), completar la siguiente tabla acerca del **ingreso poblacional promedio** usando una confiabilidad del 95%:

Ingresos	Estimación	Error Estandar	Coef. de Var. (%)	Margen de Error	Intervalo de Confianza
Hombres					
Mujeres					

Interpretar los resultados obtenidos en la tabla.

5. Ahora, se quiere establacer si existen diferencias significativas entre el ingreso promedio de hombres y mujeres. Una forma de hacerlo consiste en calcular un intervalo de confianza para $\theta^{\mathrm{H}} - \theta^{\mathrm{M}}$, donde θ^{H} y θ^{M} son los ingresos promedio de hombres y mujeres, respectivamente. Este intervalo de confianza se construye de manera usual, esto es, tomando $\hat{\theta}^{\mathrm{H}}_{\mathrm{MLE}} - \hat{\theta}^{\mathrm{M}}_{\mathrm{MLE}}$ sumando y sustrayendo el margen de error ME, que en este caso esta dado por

$$\mathrm{ME} = z_{1-\alpha/2} \sqrt{\mathbb{V}\mathrm{ar}\left(\hat{\theta}^{\mathrm{H}}_{\mathrm{MLE}} - \hat{\theta}^{\mathrm{M}}_{\mathrm{MLE}}\right)}$$

donde $z_{1-\alpha/2}$ es el percentil $100(1-\alpha/2)\%$ de la distribución normal estándar. Ahhora bien, dado que los grupo de los hombres es independiente del grupo de las muejeres, se tiene que

$$\mathbb{V}\mathrm{ar}\left(\hat{\theta}^{\mathrm{H}}_{\mathrm{MLE}} - \hat{\theta}^{\mathrm{M}}_{\mathrm{MLE}}\right) = \mathbb{V}\mathrm{ar}\left(\hat{\theta}^{\mathrm{H}}_{\mathrm{MLE}}\right) + \mathbb{V}\mathrm{ar}\left(\hat{\theta}^{\mathrm{M}}_{\mathrm{MLE}}\right),$$

donde $\mathbb{V}\text{ar}\left(\hat{\theta}^{\text{H}}_{\text{MLE}}\right)$ y $\mathbb{V}\text{ar}\left(\hat{\theta}^{\text{M}}_{\text{MLE}}\right)$ corresponden a las varianzas de los estimadores de los ingresos promedios de hombres y mujeres, respectivamente, obtenidos en el numeral anterior.\ Calcular este intervalo de confianza para $\theta^{\text{H}} - \theta^{\text{M}}$ usando una confiabilidad del 95%.

Teorema del limite central

En este contexto es posible utilizar nuevamente el Teorema del Límite central para hacer inferencia sobre $\theta^{\text{H}} - \theta^{\text{M}}$. La diferencia radica en que ahora no es necesario asumir un modelo probabilístico para cada grupo. En lugar de ello, se estima el ingreso promedio θ en cada grupo utilizando la media muestral $\bar{Y}$, y se explota el hecho de que, gracias al TLC, $\bar{Y}$ tiene distribución aproximadamente normal con media $\mathbb{E}(\bar{Y}) = \theta$ y varianza (estimada) $\widehat{\mathbb{V}}\text{ar}(\bar{Y}) = s/\sqrt{n}$, donde

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

es la desviación estándar muestral de los ingresos y n es el tamaño de la muestra correspondiente. Observe que en este caso es posible utilizar el TLC dado que los tamaños de muestra son los suficintemente grandes.

1. Usando el TLC, completar la siguiente tabla acerca del **ingreso poblacional promedio** usando una confiabilidad del 95%:

Ingresos	Estimación	Error Estandar	Coef. de Var. (%)	Margen de Error	Intervalo de Confianza
Hombres					
Mujeres					

   Interpretar los resultados obtenidos en la tabla. ¿Los resultados de esta tabla parecen ser similares a los del numeral 3. de la sección anterior anterior?

2. Teniendo en cuenta las observaciones del numeral 4. de la sección del Modelo Log-normal, usando los resultados del numeral anterior, calcular el intervalo de confianza para $\theta^{\text{H}} - \theta^{\text{M}}$ con una confiabilidad del 95%.

3. Completar la siguiente tabla de resumen acerca de la inferencia sobre $\theta^{\text{H}} - \theta^{\text{M}}$:

Método	Estimación	Error Estandar	Coef. de Var. (%)	Margen de Error	Intervalo de Confianza
Modelo Lognormal (usando MLE)					
Teorema del Límite Central					

   ¿Parace haber diferencias entre los métodos? ¿Por qué?

4. Interpretar los resultados de la tala del numeral anterior. ¿Existe evidencia de una brecha salarial entre hombres y muejeres? ¿Por qué?