Estudio de caso # 4

Universidad Externado de Colombia
Departamento de Matemáticas
Estadística 1

Juan Sosa, Ph. D.

October 22, 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 **miércoles 31 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 Número de reclamaciones

El archivo `reclamaciones.txt` contiene el número de reclamaciones (por cliente) que recibe una entidad financiera acerca de un producto particular.

1. En este caso, la variable aleatoria (cuantitativa) de estudio X es el número de reclamaciones (por cliente). ¿La variable X es discreta o continua? ¿Cuál es la escala de medición? ¿Cuáles son las unidades de medición?

2. Hallar la función de masa y la función de distribución acumulada de X por medio del enfoque frecuentista de probabilidad (frecuencias relativas). Presentar los resultados en una tabla con tres columnas: x, $f(x)$, y $F(x)$.

3. Graficar la función de densidad y la función de distribución acumulada de X.

4. Calcular e interpretar el valor esperado.

5. Calcular e interpretar el coeficiente de variación.

6. Usando la tabla del numeral 2., calcular e interpretar las siguientes probabilidades: $\mathbb{P}\mathrm{r}\left[X = 0\right]$, $\mathbb{P}\mathrm{r}\left[X \leq 2\right]$, $\mathbb{P}\mathrm{r}\left[X > 3\right]$, y $\mathbb{P}\mathrm{r}\left[1 \leq X \leq 3\right]$.

7. Calcular a interpretar la probabilidad de los siguientes eventos: $|X-\mu| < k\sigma$, para $k = 1, 2, 3$, donde $\mu = \mathbb{E}\left[X\right]$ y $\sigma = \sqrt{\mathbb{V}\mathrm{ar}\left[X\right]}$. Hacer los cálculos de forma exacta usando la tabla del numeral 2., y también de modo aproximado usando la desigualdad de Chebyshev (Sección 4.9 de Sosa et al., p 79).

   Nota: Observe que $|X - \mu| < k\sigma$ es equivalente a $\mu - k\sigma < X < \mu + k\sigma$.

2 Costo de un siniestro

El archivo `seguros.txt` contiene los costos (en millones de pesos) en los cuales ha incurrido una compañía de seguros en relación a un siniestro determinado.

Estadística descriptiva

1. En este caso, la variable aleatoria (cuantitativa) de estudio X es el costo que asume la compañía cuando paga la cobertura del siniestro. ¿La variable X es discreta o continua? ¿Cuál es la escala de medición? ¿Cuáles son las unidades de medición?

2. Completar la siguiente tabla:

Variable	Mín.	Máx.	q_1	q_2	q_3	$\bar{x}$	s	CV	AF
Costo									

3. Hacer un histograma y un diagrama de caja de la variable costos.

   Nota: Hacer el histograma en `R` con `freq = FALSE` y `nclass = 50`.

4. Comentar brevemente los resultados de los numerales anteriores.

Modelo probabilístico

La distribución de Pareto en todas sus variedades ha sido ampliamente estudiada en la literatura económica y actuarial debido a su aplicabilidad. La distribución de Pareto converge a cero más lentamente que otras alternativas (e.g., distribución Gamma, distribución log-Normal), y por lo tanto resulta mucho más seguro utilizarla para determinar las primas de grandes siniestros. Esta distribución no está limitada al estudio de costos, también se utiliza frecuentemente en otras áreas para estudiar riqueza, ingresos, retornos, perdidas, etc.

Se dice que una variable aleatoria X tiene distribución de Pareto con parámetros θ y η si la función de densidad de probabilidad de X está dada por

$$f(x;\theta,\eta) = \frac{\theta\,\eta^{\theta}}{x^{\theta+1}}\,, \quad \text{para } x > \eta,$$

donde $\eta > 0$ y $\theta > 0$. En esta distribución, η se conoce como parámetro de localización (*location*), mientras que θ se denomina parámetro de forma (*shape*).

En este caso, los analistas de la compañía de seguros aseguran que, para este tipo de siniestro en particular, X tiene distribución de Pareto con parámetros

$\eta = 1$ y $\theta = 3$, esto es, la función de densidad de probabilidad de X es

$$f(x) = \begin{cases} \frac{3}{x^4}, & \text{si } x > 1; \\ 0, & \text{en otro caso.} \end{cases} \tag{1}$$

Abreviadamente, esto se escribe $X \sim \text{Pareto}(\eta = 1, \theta = 3)$, donde el símbolo "$\sim$" se lee "tiene distribución". Observe que el rango de la variable aleatoria X es $(1, \infty) = \{x \in \mathbb{R} : x > 1\}$.

1. Hacer nuevamente el histograma de los datos, y sobre este, graficar la función de densidad de X. Visualmente, ¿esta función de densidad parece ajustar bien a los datos?

   Nota: Una alternativa en `R` para graficar la función de densidad sobre el histograma consiste en usar la función `curve` con el argumento `add = TRUE`.

2. Demostrar que la función dada en la Ecuación (1) es una función de densidad autentica, es decir, demostrar que:

   $$f(x) \geq 0 \text{ para todo } x > 1 \qquad \text{y} \qquad \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1.$$

   Nota: En esta, y todas las demostraciones que siguen, incluir **todos los pasos relevantes** de la misma. Una manera sencilla de incluir ecuaciones en **Word** es mediante el **editor de ecuaciones**.

3. Demostrar que la función de distribución acumulada de X es:

   $$F(x) = \begin{cases} 1 - \frac{1}{x^3}, & \text{si } x > 1; \\ 0, & \text{en otro caso.} \end{cases} \tag{2}$$

   Graficar esta función para $1 < x < 10$.

4. De mostrar que la expresión general para calcular el percentil $100p\%$ de X es:

   $$x_p = \frac{1}{\sqrt[3]{1-p}} \tag{3}$$

   Graficar esta expresión como función de p para $0 < p < 1$.

5. Usando la función dada en la Ecuación (2), calcular e interpretar las siguientes probabilidades: $\mathbb{P}\mathrm{r}\,[X \geq 1.5]$, $\mathbb{P}\mathrm{r}\,[1.25 < X < 1.75]$, y $\mathbb{P}\mathrm{r}\,[X \leq 2]$.

6. Usando la Ecuación (3), calcular e interpretar los siguientes percentiles: $x_{0.1}$, $x_{0.5}$, y $x_{0.95}$.

7. Instalar y invocar la libreria de `R` llamada `EnvStats` usando la siguiente línea de código:
```
install.packages("EnvStats") # Instalar
library(EnvStats) # Invocar
```
   Esta libreria contiene varias funciones relacionadas con la distribución de Pareto, incluyendo la función `ppareto`, la cual permite acceder la función de distribución acumulada. Ver el enlace `https://www.rdocumentation.org/packages/EnvStats/versions/2.3.1/topics/Pareto` para más información acerca de la libreria `EnvStats`.

   Usar la función `ppareto` y la función `qpareto` para corroborar los resultados de los numerales 6. y 7.

   Nota: En `R`, solo es necesario instalar una liberia una sola vez; es decir, una vez ejecutado el código de la instalación, no es necesario hacerlo nuevamente cada vez que abra el programa.

8. Calcular e interpretar el valor esperado de X. Comparar este valor esperado con el promedio empírico. ¿Existe una diferencia sustancial entre estos valores?

9. Calcular e interpretar el coeficiente de variación de X. Comparar este coeficiente de variación con el coeficiente de variación empírico. ¿Existe una diferencia sustancial entre estos valores?

10. El tercer y cuarto momento (al rededor de la media) también miden características interesantes (pero más sutiles) de una distribución probabilística. El tercer momento mide el sesgo o la asimetría (*skewness*), la falta de simetría, mientras que el cuarto momento mide la curtosis (*kurtosis*), una medida del decaimiento (anchura) de las colas de la función de densidad. Las medidas numéricas reales de estas características se estandarizan para eliminar las unidades físicas, dividiendo por una potencia adecuada de la desviación estándar.

    Considere el sesgo, por ejemplo. El sesgo de X se define como el tercer momento (al rededor de la media) de la variable estandarizada, esto es:

$$\text{Sesgo}(X) = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right], \tag{4}$$

donde $\mu = \mathbb{E}[X]$ es el valor esperado de X y $\sigma = \sqrt{\mathbb{V}\text{ar}[X]}$ es la desviación estándar de X.

Demostrar que la Ecuación del sesgo dada en (4) es equivalente a:

$$\text{Sesgo}(X) = \frac{\mathbb{E}[X^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3}. \tag{5}$$

Utilizar la Ecuación (5) para calcular e interpretar el sesgo de X.

11. Suponga que el costo del siniestro aumenta 5%, y además sufre un incremento constante de \$750,000. Sea Y el costo transformado de acuerdo con este aumento

    (a) Observe que $Y = aX + b$. ¿Cuál es el valor de a y b?

    (b) Encontrar la función de densidad de Y.

    Nota: Usar el Teorema 8.9.3 (p. 239) de Sosa et al. (2012).

    (c) Calcular el valor esperado y el coeficiente de variación del costo transformado. ¿Hay cambios sustanciales respecto a los valores originales?

    Nota: Usar las Proposiciones 8.4.1 (p. 212) y 8.6.2 (p. 223) de Sosa et al. (2012).

    (d) Sobreponer sobre una misma gráfica las funciones de densidad de X y Y. ¿Hay diferencias importantes entre las dos funciones?

Distribución Normal

Realizar los siguientes ejercicios acerca de la Distribución Normal. Incluir **todos los pasos relevantes** para llegar a las respuestas de los mismos.

Ejercicios 8.54 y 8.57 (p. 281) de Keller et al. (2014).

Ejercicios 10.3 y 10.4 (p. 316) de Sosa et al. (2012).

Ejercicios 40, 41, y 42 (p. 259) de Anderson et al (2011).