Estudio de caso # 5

Universidad Externado de Colombia

Departamento de Matemáticas

Estadística 1

Juan Sosa, Ph. D.

November 14, 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 23 de noviembre 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!**

Publicidad en redes sociales

Kaggle es una comunidad en línea de científicos de datos propiedad de Google, Inc. Kaggle permite a los usuarios encontrar y publicar conjuntos de datos, explorar y construir modelos en un entorno de ciencia de datos basado en la web, trabajar con otros científicos de datos y aprendizaje automático.

Considere la base de datos Ads.txt que contiene la información de una muestra de usuarios de una red social en relación al género, la edad (años cumplidos), el salario anual en dolares, y si ha comprado algún producto on-line luego de ver un anuncio publicitario en la red social (0 = No, 1 = sí). Esta base de datos es pública y fue obtenida del siguiente enlace: `https://www.kaggle.com/rakeshrau/social-network-ads`.

El objetivo principal en este estudio es caracterizar la relación y evaluar el impacto del género, la edad y el salario sobre la eventual compra o no de productos en Internet luego de ver un anuncio publicitario. La herramienta estadística por excelencia para llevar a cabo esta tarea se conoce como regresión logística (uno de los temas de Estadística 2); pero por ahora se va a hacer un análisis descriptivo de la base de datos con el fin de estudiar la situación. A la hora de interpretar los resultados en los siguientes numerales, no olvide tener en mente el objetivo principal del estudio.

1. Clasificar las variables según su naturaleza y su escala de medición.

2. Describir numérica y gráficamente cada una de las variables las variables separadamente. Recuerde que las técnicas numéricas y gráficas dependen de la naturaleza de la variable (cualitativa o cuantitativa). Para las variables cualitativas hacer una tabla de frecuencias relativas y un diagrama de barras. Mientras que para las variables cualitativas calcular el mínimo, el máximo, los cuartiles, el promedio, el coeficiente de variación, y el coeficiente de simetría de Fisher; además hacer un histograma junto con un diagrama de caja. Presentar los resultados de las medidas estadísticas organizadamente usando tablas. Interpretar los resultados obtenidos.

3. Hacer un dispersograma del salario (eje y) frente la edad (eje x). Calcular el coeficiente de correlación lineal de Pearson. Interpretar los resultados obtenidos.

4. Hacer una tabla bidimensional de frecuencias relativas de la compra de productos (columnas) frente al género. Hacer el gráfico de barras compuesto correspondiente. Interpretar los resultados obtenidos.

5. Hacer una tabla bidimensional de frecuencias relativas de la compra de productos (columnas) frente a la edad categorizada por decadas. Hacer el gráfico de barras compuesto correspondiente. Interpretar los resultados obtenidos.

   Nota: la edad categorizada por decadas es una variable cualitativa en el que las categorias son las siguientes: categoría 1 = edades entre 10 y 19 años (inclusive); categoría 2 = edades entre 20 y 29 años (inclusive); categoría 2 = edades entre 30 y 39 años (inclusive); etc. Recuerde que en R esto se puede lograr usando la función cut.

6. Hacer una tabla bidimensional de frecuencias relativas de la compra de productos (columnas) frente al salario categorizado por cuartiles. Hacer el gráfico de barras compuesto correspondiente. Interpretar los resultados obtenidos.

Distribuciones de probabilidad

1. According to a survey conducted by TD Ameritrade, one out of four investors have exchange-traded funds in their portfolios (USA Today, January 11, 2007). Consider the number of investors who have exchange-traded funds in their portfolios in a random sample (set of individuals observed independently) of 20 investors.

   (a) What is the random variable under study? Is this random variable discrete or continuous?

   (b) What is the distribution of the random variable? Write down and plot the probability mass or density function as appropriate.

   (c) Compute the probability that exactly 4 investors have exchange-traded funds in their portfolios.

   (d) Compute the probability that at least 2 of the investors have exchange-traded funds in their portfolios.

(e) If you found that exactly 12 of the investors have exchange-traded funds in their portfolios, would you doubt the accuracy of the survey results?

(f) Compute the expected number of investors who have exchange-traded funds in their portfolios.

(g) Compute and interpret the coefficient of variation of the number of investors who have exchange-traded funds in their portfolios.

2. Everyone is familiar with waiting lines. We wait in line at banks, groceries, and fast-food restaurants. There are also waiting lines in firms where trucks wait to load and unload and on assembly lines where stations wait for new parts. Management scientists have developed mathematical models that allow managers to determine the operating characteristics of waiting lines. Some of the operating characteristics are: The probability that there are no units in the system, The average number of units in the waiting line, The average time a unit spends in the waiting line, The probability that an arriving unit must wait for service, The Poisson probability distribution is used extensively in waiting-line (also called queuing) models. Many models assume that the arrival of units for service is Poisson distributed with a specific value of λ.

   The number of users of an automatic banking machine is Poisson distributed. The mean number of users per 5-minute interval is 1.5.

   (a) What is the random variable under study? Is this random variable discrete or continuous?

   (b) What is the distribution of the random variable? Write down and plot the probability mass or density function as appropriate.

   (c) Compute the probability that no users in the next 5 minutes.

   (d) Compute the probability that three or more users in the next 10 minutes.

   (e) Compute the probability that five or fewer users in the next 15 minutes.

3. The Troubled Asset Relief Program (TARP), passed by the U.S. Congress in October 2008, provided \$700 billion in assistance for the struggling U.S. economy. Over \$200 billion was given to troubled financial institutions with the hope that there would be an increase in lending to help

jump-start the economy. But three months later, a Federal Reserve survey found that two-thirds of the banks that had received TARP funds had tightened terms for business loans (The Wall Street Journal, February 3, 2009). Of the ten banks that were the biggest recipients of TARP funds, only three had actually increased lending during this period. Increased Lending: BB&T, Sun Trust, U.S. Bancorp. Decreased Lending: Bank of America, Banks Capital One, Citigroup, Fifth Third Bancorp, J.P. Morgan Chase, Regions Financial, U.S. Bancorp.

For the purposes of this exercise, assume that you will randomly select three of these ten banks for a study that will continue to monitor bank lending practices. Let X be a random variable indicating the number of banks in the study that had increased lending.

(a) What is $f(0)$? What is your interpretation of this value?

(b) What is $f(3)$? What is your interpretation of this value?

(c) Compute $f(1)$ and $f(2)$. Show the probability distribution for the number of banks in the study that had increased lending. What value of X has the highest probability?

(d) What is the probability that the study will have at least one bank that had increased lending?

(e) Compute the expected value and coefficient of variation for the random variable.

4. Trading volume on the New York Stock Exchange is heaviest during the first half hour (early morning) and last half hour (late afternoon) of the trading day. The early morning trading volumes (millions of shares) for 13 days in January and February are shown next (Barron's, January 23, 2006; February 13, 2006; and February 27, 2006): 214, 163, 265, 194, 180, 202, 198, 212, 201, 174, 171, 211, 211. The probability distribution of trading volume is approximately normal.

(a) Compute the mean and standard deviation to use as estimates of the population mean and standard deviation.

(b) White down and plot the corresponding probability density function.

(c) What is the probability that, on a randomly selected day, the early morning trading volume will be less than 180 million shares?

(d) What is the probability that, on a randomly selected day, the early morning trading volume will exceed 230 million shares?

(e) How many shares would have to be traded for the early morning trading volume on a particular day to be among the busiest 5% of days?

5. Comcast Corporation is the largest cable television company, the second largest Internet service provider, and the fourth largest telephone service provider in the United States. Generally known for quality and reliable service, the company periodically experiences unexpected service interruptions. On January 14, 2009, such an interruption occurred for the Comcast customers living in southwest Florida. When customers called the Comcast office, a recorded message told them that the company was aware of the service outage and that it was anticipated that service would be restored in two hours. Assume that two hours is the mean time to do the repair and that the repair time has an exponential probability distribution.

(a) White down and plot the corresponding probability density function.

(b) What is the probability that the cable service will be repaired in one hour or less?

(c) What is the probability that the repair will take between one hour and two hours?

(d) For a customer who calls the Comcast office at 1:00 P.M., what is the probability that the cable service will not be repaired by 5:00 P.M.?