Estudio de caso # 2

Universidad Externado de Colombia

Departamento de Matemáticas

Estadística 1

Juan Sosa, Ph. D.

August 27, 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 5 de septiembre 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 Encuesta de Desarrollo e Innovación Tecnológica

Considere la Encuesta de Desarrollo e Innovación Tecnológica (EDIT), Industria VIII, años 2015–2016. Toda la información al respecto de la encuesta se encuentra disponible de forma gratuita en `http://microdatos.dane.gov.co/index.php/catalog/553/get_microdata`.

La EDIT es una operación estadística susceptible de constante revisión y mejora. Sin embargo, desde el punto de vista conceptual y metodológico, su diseño preserva un marco teórico fundamental que se corresponde con los principales acuerdos alcanzados por la comunidad de personas expertas, nacionales e internacionales, sobre diseño, aplicación e interpretación de encuestas nacionales de innovación. En particular, la EDIT acoge la mayoría de pautas metodológicas trazadas por la Organización de Cooperación y Desarrollo Económico (OCDE), especialmente el Manual de Oslo, y por la Red Iberoamericana de Indicadores de Ciencia y Tecnología (RICYT), en el Manual de Bogotá. La mayor parte de estas recomendaciones han sido adaptadas a las necesidades de información y restricciones técnicas identificadas para Colombia.

Siguiendo los lineamientos del Manual de Oslo (2005), la unidad estadística primaria de la EDIT es la empresa. Siguiendo el mismo lineamiento, la encuesta se encuentra diseñada según el enfoque basado en el "sujeto", el cual "trata de las actitudes y actividades innovadoras de la empresa en su conjunto. La idea es explorar los factores que influyen en el comportamiento innovador de la empresa (estrategias, incentivos y barreras a la innovación) y el ámbito de las diversas actividades de innovación, y sobre todo examinar los resultados y los efectos de la innovación" (Oslo, 2005, p. 28).

La EDIT es una operación tipo censo, ya que se toman todas las empresas industriales que cumplen los parámetros de inclusión determinados para el universo de estudio. El parámetro de inclusión corresponde a las empresas industriales que tienen establecimientos con 10 o más personas ocupadas o que en su defecto registren un valor de producción anual igual o superior a un valor que se especifica para cada año de referencia correspondiente al directorio de empresas de la Encuesta Anual Manufacturera (EAM).

La operación estadística que se desarrolla es de tipo censo, ya que se toman todas las empresas industriales que cumplen los parámetros de inclusión determinados para el universo de estudio. Por lo tanto, la EDIT comprende una cobertura geográfica del total nacional.

La base de datos completa dada en `EDIT 2015 2016.csv` contiene 7,947 registros y 638 variables. Todos los detalles acerca de las variables se encuentran en el documento `EDIT 2015 2016.pdf`, páginas 19 a 102.

a. Importar la base de datos en `R` y a partir de esta conformar una nueva base de datos que incluya únicamente los registros con la información completa de las siguientes variables:

   I1R4C2N Número total de innovaciones de bienes o servicios nuevos.

   I2R6C1 Reducción de los costos laborales.

   I2R15C1 Disminución en el pago de impuestos.

   I3R2C1 Ventas nacionales totales 2016 (Miles de pesos corrientes).

   II1R10C2 Total del monto invertido en 2016 por su empresa en actividades científicas, tecnológicas y de innovación.

   IV1R11C2 Total personal ocupado en 2016.

   V1R1C1 Departamento interno de I+D.

   Esta nueva base de datos así conformada contiene $n = 599$ registros y $p = 7$ variables. Clasificar estas variables según la naturaleza (cualitativa, cuantitativa discreta, cuantitativa continua) y la escala de medición (nominal, ordinal, intervalo, razón). Toda los detalles acerca de las variables (incluyendo la codificación y la escala de medición) se encuentra en el archivo `EDIT 2015 2016.pdf`.

   Sugerencia: una vez a sido conformada la base de datos con las siete variables y los registros completos, remueva del Global Enviroment la base de datos original para que agilizar el tratamiento de la información.

b. Hacer una tabla de frecuencias y un diagrama de barras para V1R1C1. Comentar brevemente los resultados obtenidos.

c. Hacer una tabla bidimensional de frecuencias relativas y un diagrama de barras compuesto para I2R6C1 (columnas) frente I2R15C1 (filas). Comentar brevemente los resultados obtenidos.

d. Describir numéricamente y gráficamente I1R4C2N, IV1R11C2, I3R2C1 y II1R10C2, es decir, para cada una de estas variables hacer un diagrama de caja y calcular las medidas de localización, de dispersión y de asimetría. Una forma de presentar las medidas estadísticas organizadamente es por medio de la siguiente tabla:

Medida	I1R4C2N	IV1R11C2	I3R2C1	II1R10C2
Mínimo				
Máximo				
Cuartíl 1				
Mediana				
Cuartíl 3				
Media				
Rango				
Rango Intercuartílico				
Desv. Estándar				
Coef. Variación				
Indice Yule-Bowley				
Coef. Asim. Fisher				

Comentar brevemente los resultados obtenidos.

Nota 1: otorgue nombres significativos a las variables para escribir el informe. **No** use los nombres codificados. Por ejemplo, en lugar de usar I3R2C1, use el nombre Ventas.

Nota 2: los detalles acerca las medidas de asimetría se encontran en SOSA et al., capítulo 5.

e. Hacer un dispersograma y calcular la matriz de correlación de Pearson teniendo en cuenta todas las variables del numeral anterior. Puede presentar los gráficos en un sólo prisma de dispersogramas y todos los coeficientes de correlación en una sola matriz. Comentar brevemente los resultados obtenidos.

f. Considere la variable I3R2C1:

1. Calcular el porcentaje de datos atípicos.
2. Calcular el procentaje de datos extremos.
3. Categorizar (construir intervalos) la variable usando los siguientes puntos de corte: 0M, 30 M, 60 M, 90 M, 120 M, 5,506 M, donde la letra “M” indica millones (por ejemplo, 30 M equivale a 30,000,000).

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

Nota 2: de nombres significativos a las categorias resultantes. Por ejemplo, la categoría de 0 a 30 M puede llamarse "ventas bajas".

4. Hacer una tabla de bidimensional de frecuencias relativas para I3R2C1 categorizada (columnas) frente a V1R1C1 (filas). A partir de esta tabla construir los perfiles files fila y hacer el diagrama de barras compuesto correspondiente. Comentar brevemente los resultados obtenidos.

g. ¿La distribución de las ventas de las empresas es equitativa? Con base en la información de I3R2C1, ¿qué indica el coeficiente de Gini y la curva de Lorenz acerca de tal distribución?

Nota 1: para calcular la coeficiente de Gini y graficar la curva de Lorenz puede utilizar la libreria `DescTools` de `R`. Para instalar y utilizar esta libreria solo de utilizar las siguites insintrcciones:

```
# instalar libreria
install.packages("DescTools")

# cargar libreria
library(DescTools)

# ventas
ventas <- EDIT$I3R2C1

# calculo del coeficiente de Gini
Gini(x = ventas)

# curva de Lorenz
windows()
plot(Lc(x = ventas), col = "red")
grid()
```

Nota 2: no es necesario instalar la libreria una sola vez!

h. Repetir el numeral anterior para II1R10C2.

2 Bond Funds

Case study taken from Berenson et al. (2011, p. 439).

The file **Bond Funds** contains information regarding nine variables from a sample of 184 mutual funds. These variables are Fund number: Identification number for each bond fund; Type: Bond fund type (intermediate government or short-term corporate); Assets: In millions of dollars; Fees: Sales charges (no or yes); Expense ratio: Ratio of expenses to net assets in percentage; Return 2009: Twelve-month return in 2009; Three-year return: Annualized return, 2007–2009; Five-year return: Annualized return, 2005–2009; Risk: Risk-of-loss factor of the mutual fund (below average, average, or above average).

Perform a principal component analysis on the entire data set using the quantitative variables:

1. How many principal components are enough to explain the data?
2. Report the amount of information retained by the principal components.
3. Interpret the meaning of each principal component.
4. Use a bi-plot to identify those variables and funds that excel in the data.