e-commerce

ITBA

CASO DE ESTUDIO

Segunda Entrega

Por:
Agustín Ignacio Vergniaud

Detalles técnicos:

Dataset a Analizar:
Dinero gastado anualmente en E-Commerce (U$D)
Muestra: 500 personas
Variable a Analizar: Yearly.Amount.Spent

Formato del Proyecto
Análisis Descriptivo del Dataset

Análisis Técnico:

Análisis Práctico:

Aclaración Útil:

El análisis descriptivo que el grupo realizará puede ser usado por *cualquier dataset* ya que está hecho de forma totalmente *genérica*. Con el mismo, el usuario podrá ver la modelización de sus datos respecto a la distribución normal y de esa manera evaluar el uso de este modelo.

Las imágenes serán presentadas en formato .png y las tablas en formato .csv.
Estas serán guardadas en este formato en su computadora.
Las encontrará en la carpeta default de lectura de archivos de R.
Puede ver cual es el mismo en su computadora usando getwd() y apretando Alt+Enter.
Si desea cambiarlo llame a la función getwd colocando el directorio de su agrado en el interior.
de esta manera getwd(directoriodeseado)
A modo de comienzo, si desea limpiar el environment (funciones de programas anteriores, csv anteriores, etc), puede ejecutar rm(list = ls()).
De esta manera se gastará menos memoria de ejecución en la computadora.

Análisis Técnico

Mediante el mismo explicaremos a medida de resumen lo que encontrará en el código de R adjunto.

1. Presentación de funciones

Al hacer el trabajo de forma totalmente genérica, definimos una serie de funciones pertinentes al análisis de nuestra variable y su modelización.

Funciones:

momentos <- Obtiene los momentos del dataset ingresado y los guarda en formato .csv para su posterior utilización.

comparacionnormaldensidad <- Realiza el histograma de frecuencias relativas y sobre el superpone la curva de densidad empírica y la curva teórica de la densidad de una distribución normal con el desvío y la media de la muestra. Todo lo guarda en formato .png con el nombre que el usuario desee.

comparacionnormalacumulada <- Realiza el gráfico de probabilidad acumulada de los datos en cuestión y sobre el mismo superpone la curva de probabilidad acumulada de una distribución normal teórica con el desvío y la media de la muestra.

acumu <- Construye la función de probabilidad acumulada del set de datos.

replin <- Realiza el gráfico de tendencia lineal de los datos.

criteriobox <- Construye el boxplot para ver la morfología y la simetría de los datos.

critqq <- Construye el qqplot respectivo a la comparación con la posible modelización normal de los datos.

2. Limpieza de Dataset

El R ejecuta el código de dónde obtiene del dataset original la columna Yearly.Amount.Spent qué es la variable a analizar y la guarda en un nuevo .csv con la función write.csv. El csv se encuentra adjunto en la entrega como datasetlimpio.csv.

3. Lectura de Archivos

Descarga la librería *readr* y con la función read.csv lee el dataset que el usuario quiera analizar para realizar la modelización normal de sus datos.

4. Aplicación de Criterios y Explicaciones de uso

En un análisis de estadística descriptiva, es indispensable obtener y tener en claro cuales son los valores de los respectivos **momentos** del dataset a analizar.
Es por eso que llamamos al uso de la función momentos, que exporta los valores numéricos de los mismos a un .csv.
Inicialmente sabemos que estamos analizando una muestra de 500 valores, pero no sabemos nada más acerca de la muestra.
Utilizamos el **boxplot** para ver la simetría y morfología de los datos. Esto se vé según la ubicación de la mediana en el rango intercuartil. Además, mediante el mismo podemos ver la validez de la encuesta, debido a que en el gráfico pueden aparecer outliers o valores atípicos.
Hicimos uso de un **gráfico de línea** para ver la tendencia lineal de los datos y poder ver si el trazado entre la sucesión de puntos de los datos se mantiene dentro de límites o fluctua.
Fue de utilidad realizar un **QQ Plot** para ver la similitud de los datos respecto al modelo normal. En el gráfico, la comparación de los cuantiles teóricos vs los cuantiles de la muestra del set de datos (Serie de puntos) es correcta si se asemeja mucho a la recta que propone la QQ Line.
Realizar el gráfico del **histograma de frecuencias relativas** superpuesto con la **función de densidad empírica** de la muestra y la **teórica del modelo** es una representación gráfica que el usuario puede observar para decidir acerca de la correcta modelización de sus datos respecto a la normal.
El gráfico de **funciones de probabilidad** teórica y del modelo superpuestas sirve para que el usuario pueda ver que tanto van a fluctuar las probabilidades que calcule sobre la muestra y sobre el modelo normal con el desvío y media de la muestra.

Análisis Práctico

5. Data Analysis

Debido a que la programación en R representa una modelización normal de cualquier dataset de una manera totalmente genérica, nos pareció conveniente incluirla al inicio. Ahora mostraremos su utilidad al analizar el dataset del grupo en cuestión.

Descripción de la variable a analizar

Variable: Yearly.Amount.Spent
¿Que modelamos? : La variable se obtiene de un dataset sobre un análisis de una plataforma de E-Commerce y la misma representa cuánto gastan anualmente, en dólares, personas comprando en Internet.
Tamaño de la muestra: 500 filas
¿Dónde la encuentro? : datasetlimpio.csv

Introducción del dataset al programa de modelización normal realizado

Esto nos ayuda a darnos una idea gráfica y técnica del comportamiento de nuestra variable.

Momentos

Abrimos momentosdataset.csv

La unidad de los datos provistos es el *Dólar Estadounidense (U$D)*.

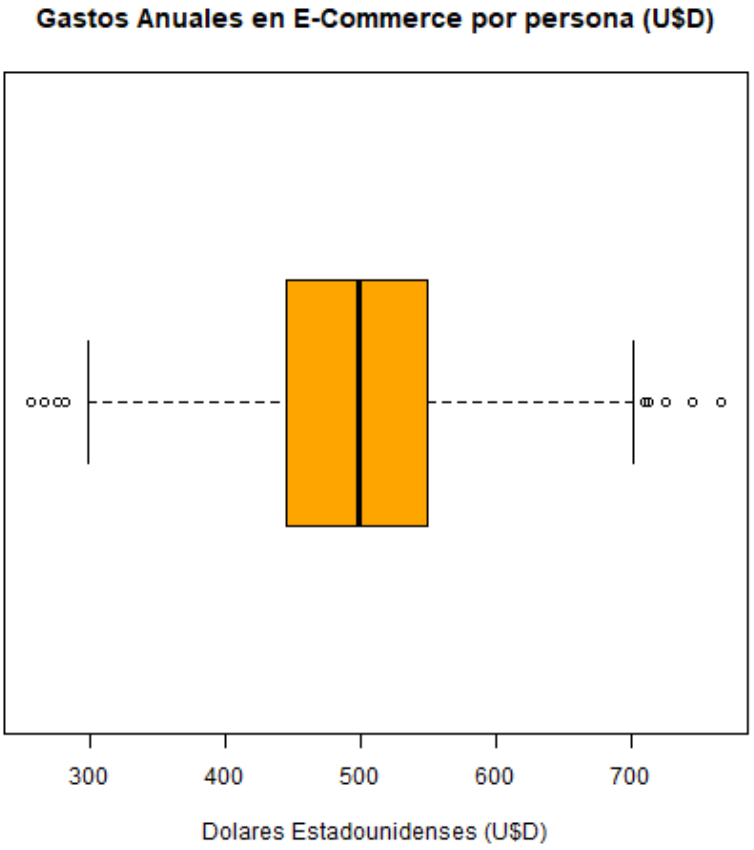
Media (μ) = 499,314038258591
Desvío Estándar (σ) = 79,3147815497068
Mediana (Q2) = 498,8878755
Varianza ($σ^2$) = 6290,834572
*Coeficiente de Asimetría (**γ**)* = 0,034685726
Curtosis (k) = 3,447372702
Rango Intercuartil (Q3-Q1) = 104,2755505

La *media* es claramente muy similar a la *mediana*. La *mediana* está levemente ubicada a la izquierda de la media. Esto refleja que la mayoría de las personas posee un gasto en E-Commerce infinitesimalmente menor al promedio de los gastos.
El hecho de que posean valores similares, es característico de una distribución normal.

Se justifica la *ubicación de la mediana respecto de la media* mediante el valor positivo del *coeficiente de asimetría*.
Esto muestra que, para esta variable, el consumo en Internet está casi simétricamente distribuido entre los límites de mayor (800U$D) y menor (300 U$D) gasto.
A su vez, su valor tan cercano a 0, muestra su clara *similitud* con la distribución normal.
La *curtosis* de una distribución normal estándar es de 3, por lo tanto la de esta se asemeja bastante a la gaussiana.
Más allá de esto, es mayor a 3 así que es *leptocúrtica*.

Boxplot

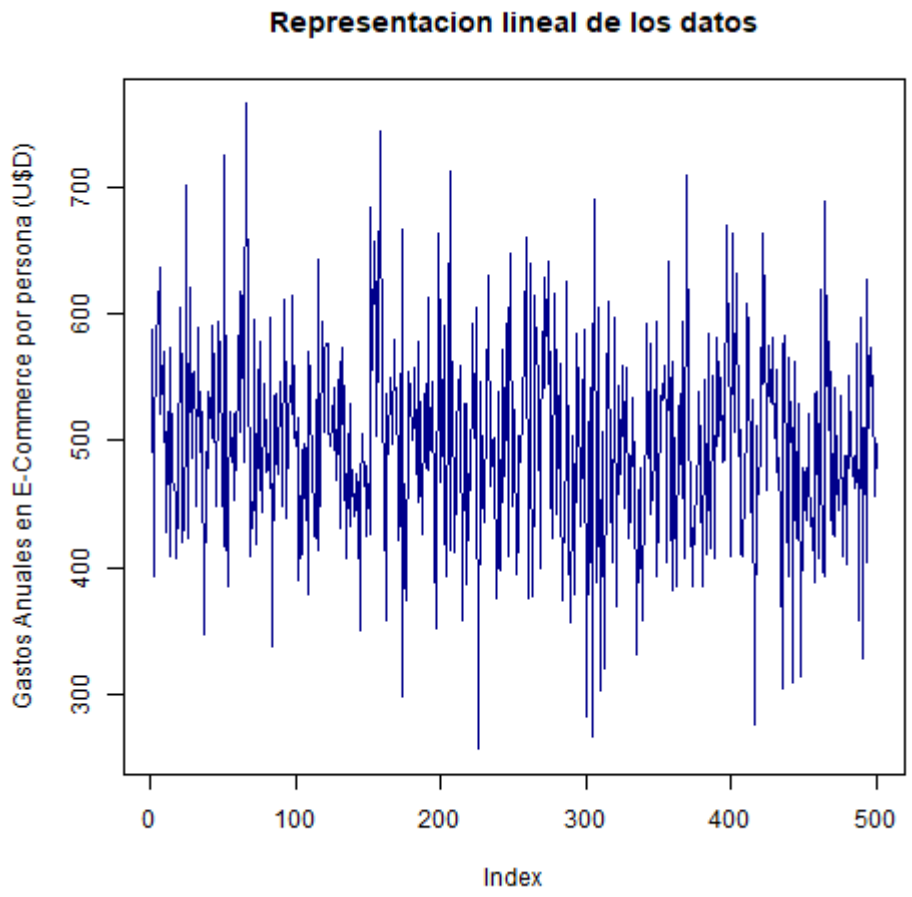
Utilizaremos el criterio del boxplot para ver la ubicación de la media con respecto al rango intercuartil.
Observaremos también la simetría y morfología de los datos.
Llamamos a nuestra función criteriobox.



Salvo algunos valores atípicos (Outliers), vemos que la mediana (Q2) está bastante

centrada en el rango intercuartil (Q3-Q1), lo que nos sigue demostrando la posible modelización normal.
No hay tantos outliers, por lo tanto se puede seguir modelizando la variable en cuanto al dataset y las conclusiones que surjan de la aproximación al modelo no serán incorrectas.

Gráfico lineal de tendencia

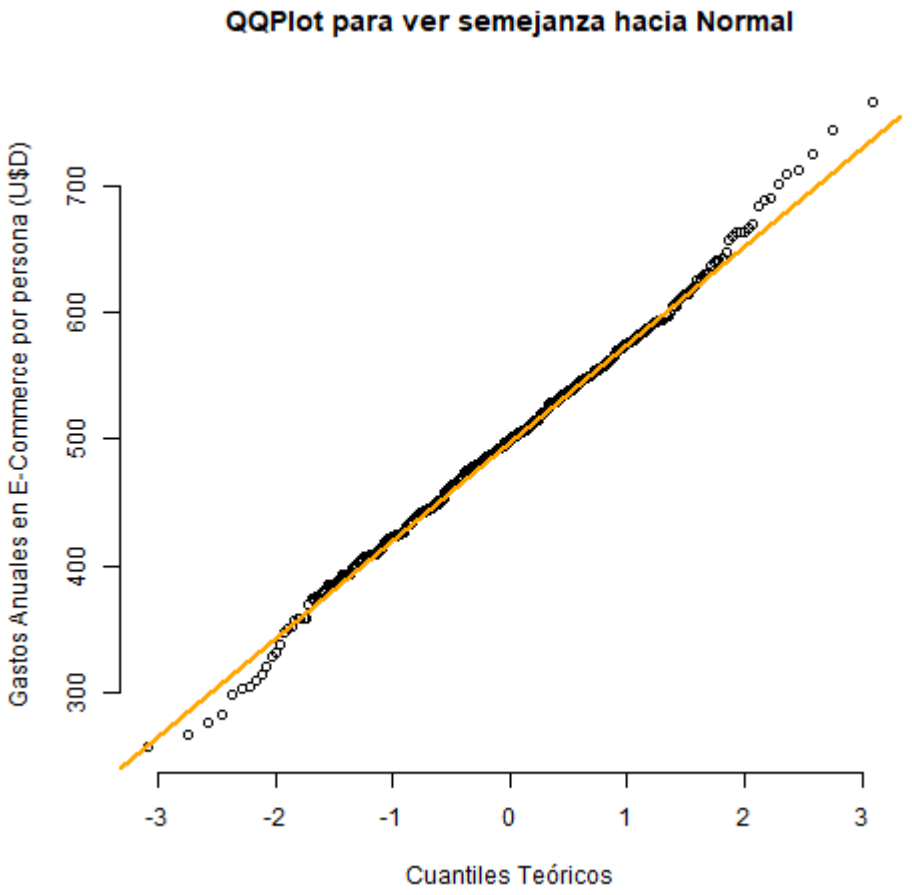


Llamamos a nuestra función *replin*.
En las *ordenadas* se coloca el dinero gastado en dólares y en las *abscisas* se indexa según el número de encuesta realizada.

Al abrir replinetapa2.png (imagen) podemos ver que en la dispersión de los datos *existen algunos picos que exceden el límite de la regularidad de los datos* y esos eran los mismos que observamos atípicos en el *boxplot*.

QQPlot

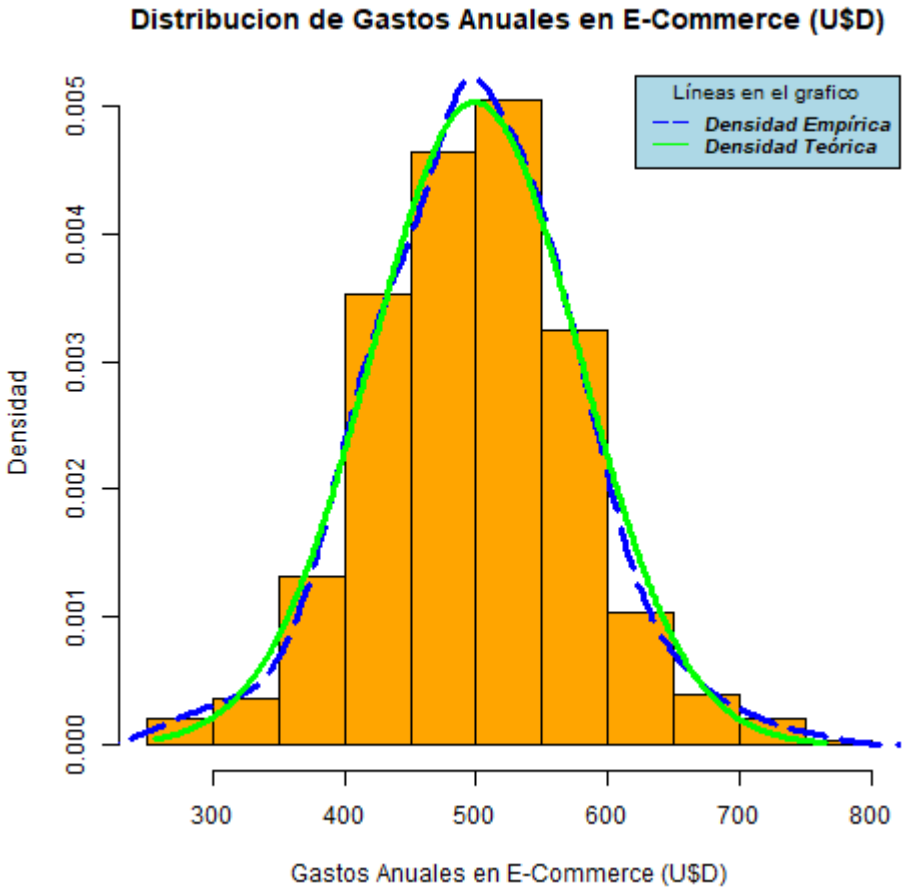
Para seguir con nuestro análisis descriptivo del dataset, nos parece pertinente realizar un QQPlot para ver su similitud al modelo que queremos aplicar.
Llamamos a nuestra función critqq.



Abrimos criterioqqetapa2.png
En el gráfico, la comparación de los *cuantiles teóricos vs los cuantiles de la muestra* del set de datos (serie de puntos) se *asemeja* mucho a la recta que propone la *QQLine*.
Esta es un criterio más que hace que podamos decir que la normal es una buena modelización de los datos.
Podemos ver que la curva es *simétrica* con *colas livianas*.

Comparación de Funciones de Densidad

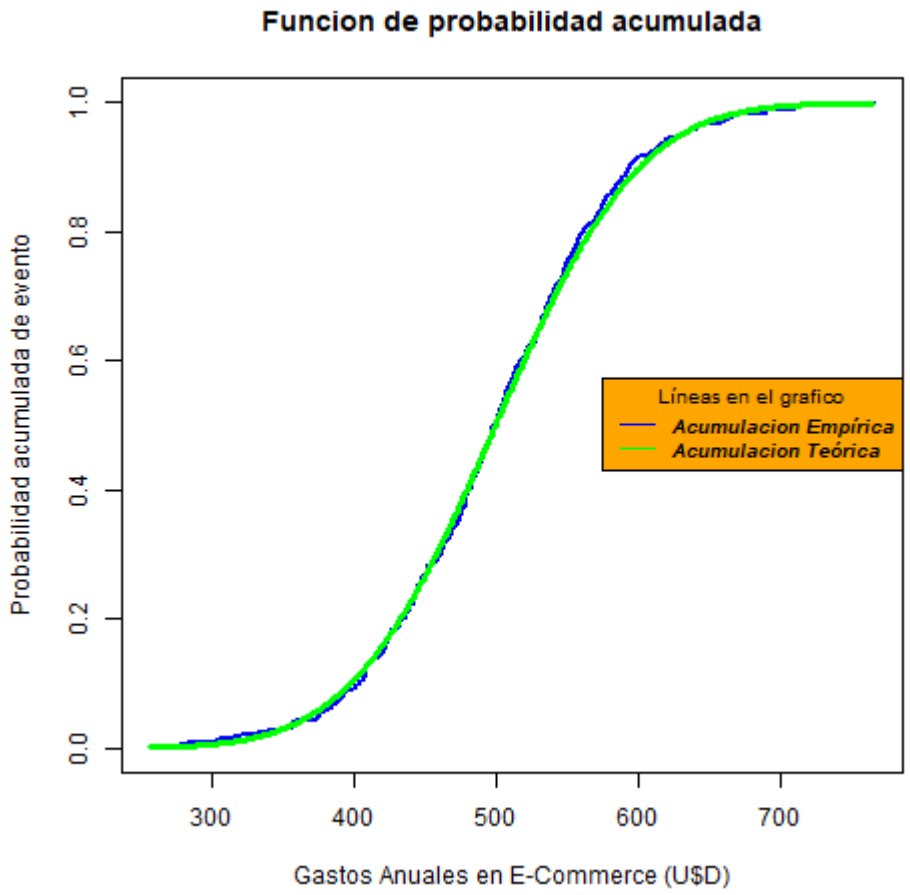
Ahora ejecutaremos la función *comparacionnormaldensidad*.
La misma grafica un *histograma de frecuencias relativas*, sobre el cual podemos estimar, según la morfología del mismo, la función de densidad.
Es por esto que en *comparacionnormal*, se grafica la *densidad empírica* de los datos, *superpuesta con el histograma* para su mejor visualización.
Además, para su comparación absoluta, se *superpone* la *función de densidad* de una *distribución normal con la media y el desvío de la muestra*.



Abrimos la imagen *graficonormal.png*.
Allí observamos el *histograma de frecuencias relativas* superpuesto con la *función de densidad empírica* del mismo y la *función de densidad de una normal con la media y el desvío de la muestra*.
Las *curvas de densidad* son casi iguales entre sí. Eso se puede ver fácilmente en el gráfico. Otra buena observación es que el histograma presenta una orientación casi Gaussiana. Todo esto sigue justificando lo que habíamos *ya visto previamente* en el QQplot y el Boxplot.

Comparación de Funciones de Probabilidad Acumulada

Ahora utilizaremos la función creada en el inicio *comparacionnormalacumulada*.
Esta *superpone* los gráficos de la *función de probabilidad acumulada* de la muestra y la *función acumulada teórica* del *modelo normal propuesto*, con la media y desvío de la muestra en cuestión.



Abrimos *graficoacumulada.png*.
Observamos que el gráfico de las dos curvas es casi idéntico.
La *probabilidad se va acumulando de manera muy similar*, eso refleja que las probabilidades que queramos obtener del modelo para evaluar el comportamiento de los datos serán *casi idénticas* a las de la muestra.

Mini propuesta de algunas situaciones de análisis posibles

Dinero gastado con el 95% de probabilidad:

Variable X - Dinero gastado en E-Commerce en 1 año.

```
quantile(setdatos,0.05) = 376.29 U$D
```

El consumo crece rápido alrededor de estos cuantiles ya que el valor mínimo es 300 U$D.

Tomando una submuestra aleatoria de 50 personas, cuál es la probabilidad que la media muestral de consumo anual supere 550 U$D.
Esto nos puede seguir para buscar predecir datos tomando muestras más pequeñas de la población.
Variable R - Media muestral de consumo anual en E-Commerce de muestra de 50 personas.

```
Calculamos nuevo desvío:
nuevosd<-sd(setdatos)/sqrt(50)
nuevosd = 11.217 U$D
La media muestral será la misma.
```

Uso pnorm que representa la probabilidad acumulada, también lo podíamos haber hecho estandarizando.

```
valorp <- (1-pnorm(550,mean(setdatos),nuevosd))
valorp = 3.110252e-06
```

Es una probabilidad extremadamente baja, lo que significa que la variación del consumo anual es muy baja en muestras más pequeñas.

Probabilidad de que los gastos anuales en 5 años sean mayores a 2400 U$D:
Es normal por propiedad de suma de normales.
Variable H - Consumo anual en E-Commerce en 5 años

```
med5y <- 5*mean(setdatos)
sd5y <- sqrt(5)*sd(setdatos)

valor5y <-(1-pnorm(2400,med5y,sd5y))
valor5y = 0.7069539
```

Es un buen valor de probabilidad para la previsión de ventas hacia el futuro de una empresa que puede operar en E-Commerce.

6. Conclusión

El dataset refleja una muestra de los gastos anuales en E-Commerce por persona, una plataforma emergente durante los últimos años.
El hecho de poseer una distribución tan cercana a la normal, que se ve muy claro en las curvas superpuestas de densidad y de distribución acumulada, muestra que los gastos de este tipo se encuentran simétricamente distribuidos entre el mínimo de 300 y 800 U$D.
El Boxplot posee pocos outliers para lo que es el tamaño de la muestra, por lo que la encuesta propone valores correctos. Mientras tanto, el QQPlot refleja una clara semejanza entre la serie de puntos y la QQLine por lo tanto nuestra previsión del modelo fue correcta.
Podemos establecer que la modelización mediante la distribución normal es correcta para evaluar situaciones como las que propusimos anteriormente.