UNC Universidad Nacional de Córdoba
FAMAF Facultad de Matemática, Astronomía, Física y Computación

Selección de Distribuciones de Probabilidad

Angaramo Piñol, Facundo Nicolás

24 de Julio de 2024

Resumen

El objetivo de este trabajo es realizar un análisis exploratorio de una muestra compuesta por 300 observaciones y proponer al menos dos familias de distribuciones de probabilidad como modelos de ajuste. Con base al análisis exploratorio de los datos, se sugirieron las distribuciones Log-Normal y Gamma como posibles modelos. Se realizaron pruebas de hipótesis, tales como el Test Chi-Cuadrado de Pearson y el Test de Kolmogorov-Smirnov, para evaluar la bondad de ajuste. Los resultados indican que la distribución Gamma es una mejor candidata para modelar los datos.

1. Introducción

Dada una muestra compuesta por 300 observaciones, se tiene como objetivo realizar un análisis exploratorio de los datos y luego proponer al menos dos familias de distribuciones de probabilidad como modelos de ajuste. Posteriormente, se estimarán los p-values de la hipótesis de que los datos provienen de las distribuciones sugeridas, utilizando el Test Chi-Cuadrado de Pearson y el Test Kolmogorov-Smirnov.

2. Análisis y Visualización de Datos

Para este análisis, se utilizó una muestra de 300 observaciones de datos de tipo continuo. A continuación, se presentan las principales estimaciones muéstrales:

- mean: 3.018
- variance: 3.199
- min: 0.111
- q1: 1.729
- median: 2.836
- q3: 4.030
- max: 13.343
- skewness: 1.399

El valor de la media indica que las observaciones se centran alrededor de 3.018, mientras que la varianza sugiere una dispersión moderada en los datos. La diferencia entre la media y la mediana, así como el valor positivo de la asimetría, indican una distribución sesgada hacia la derecha.

Para visualizar mejor la distribución de los datos, se realizo un histograma [Figura 1] y un diagrama de caja [Figura 2]:

Histograma

- El histograma muestra la frecuencia relativa de las observaciones dentro de intervalos específicos. Como se puede observar se confirma la asimetría positiva, con una mayor concentración de datos en valores menores y una cola extendida hacia valores mayores.

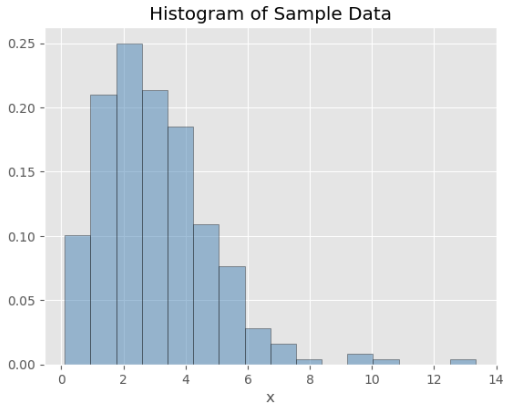

Figura 1: Histograma de los Datos Muéstrales

Box-Plot

- El diagrama de caja proporciona una representación visual de los cuantiles y los valores atípicos en la muestra. Nuevamente, podemos notar la asimetría positiva en los datos de la muestra. Además, se observan algunos valores atípicos que se alejan significativamente del resto de las observaciones.

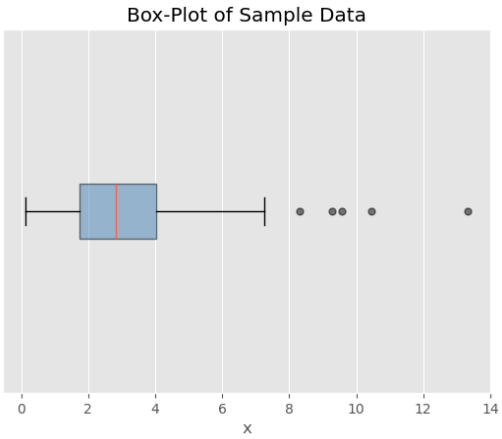

Figura 2: Box-Plot de los Datos Muéstrales

3. Propuesta de Familias de Distribuciones

Con base a los hallazgos obtenidos durante el análisis exploratorio de los datos, las dos familias de distribuciones de probabilidad propuestas como modelos de ajuste son la distribución Log-Normal y la distribución Gamma.

Distribución Log-Normal

La distribución Log-Normal es adecuada para datos que son positivos y que pueden tener una asimetría positiva. Esta distribución se obtiene cuando el logaritmo de la variable aleatoria sigue una distribución normal.

- H_0 : La muestra proviene de una v.a X con distribución Log-Normal de parámetros μ y σ desconocidos.
- H_1 : La muestra NO proviene de una v.a X con distribución Log-Normal.

Dado que los parámetros μ y σ son desconocidos se utilizo el Método de Máxima Verosimilitud para estimarlos. Los parámetros estimados fueron:

- $\hat{\mu} = 0.908$
- $\hat{\sigma} = 0.691$

En la Figura 3 se observa una comparación entre el histograma de frecuencias relativas de los datos y la función de densidad de una v.a X con distribución Log-Normal de parámetros $\hat{\mu}$ y $\hat{\sigma}$.

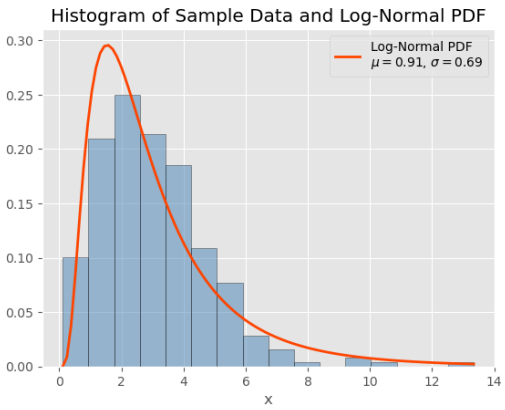

Figura 3: Función de Densidad Log-Normal

Distribución Gamma

La distribución Gamma es otra distribución que puede manejar datos positivos con asimetría positiva.

- H_0 : La muestra proviene de una v.a X con distribución Gamma de parámetros α y β desconocidos.
- H_1 : La muestra NO proviene de una v.a X con distribución Gamma.

Dado que los parámetros α y β son desconocidos se utilizó el Método de Momentos para estimarlos. Los parámetros estimados fueron:

- $\hat{\alpha} = 2.847$
- $\hat{\beta} = 1.060$

En la Figura 4 se observa una comparación entre el histograma de frecuencias relativas de los datos y la función de densidad de una v.a X con distribución Gamma de parámetros $\hat{\alpha}$ y $\hat{\beta}$

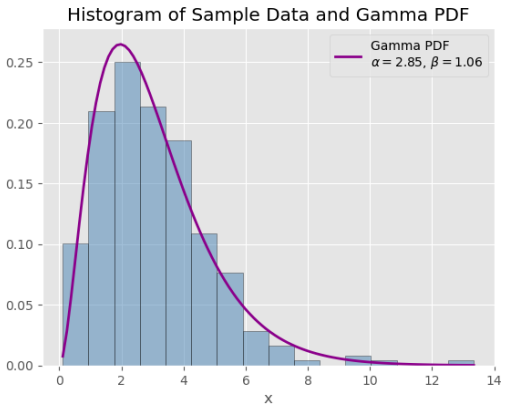

Figura 4: Función de Densidad Gamma

4. Pruebas de Hipótesis

Para evaluar la idoneidad de las distribuciones propuestas (Log-Normal y Gamma) como modelos para los datos, se realizaron dos pruebas de hipótesis: el Test Chi-Cuadrado de Pearson y el Test de Kolmogorov-Smirnov. Estas pruebas permiten determinar si la muestra de datos proviene de las distribuciones especificadas.

Test Chi-Cuadrado de Pearson

El Test Chi-Cuadrado de Pearson compara la frecuencia observada de los datos con la frecuencia esperada según una distribución teórica. Este test es útil para evaluar la bondad de ajuste de una distribución de probabilidad a un conjunto de datos.

- Distribución Log-Normal
  - Estadístico Chi-Cuadrado: 19.053
  - p-value: 0.00387
- Distribución Gamma
  - Estadístico Chi-Cuadrado: 6.093
  - p-value: 0.31031

Test de Kolmogorov-Smirnov

El Test de Kolmogorov-Smirnov (K-S) compara la función de distribución empírica de la muestra con la función de distribución acumulada teórica, estimando la distancia máxima entre los dos gráficos.

- Distribución Log-Normal
  - Estadístico K-S: 0.083
  - p-value: 0.0282
- Distribución Gamma
  - Estadístico K-S: 0.044
  - p-value: 0.58614

5. Conclusiones

Los resultados de las pruebas de hipótesis sugieren que la distribución Gamma es una mejor candidata para modelar los datos, dado que los p-values son mayores y, por lo tanto, menos indicativos de rechazo de la hipótesis nula en comparación con la distribución Log-Normal. En consecuencia, se recomienda el uso de la distribución Gamma para modelar los datos en contextos similares.

Referencias

[1] Dra. Patricia Kisbye *Modelos y Simulación.*

[2] A. M. Law y W. D. Kelton *Simulation, Modeling, and Analysis.*