Trabajo practico Modelos y Simulación

Lucía Martinez Gavier, Leonardo Luis Torres Villegas

June 2023

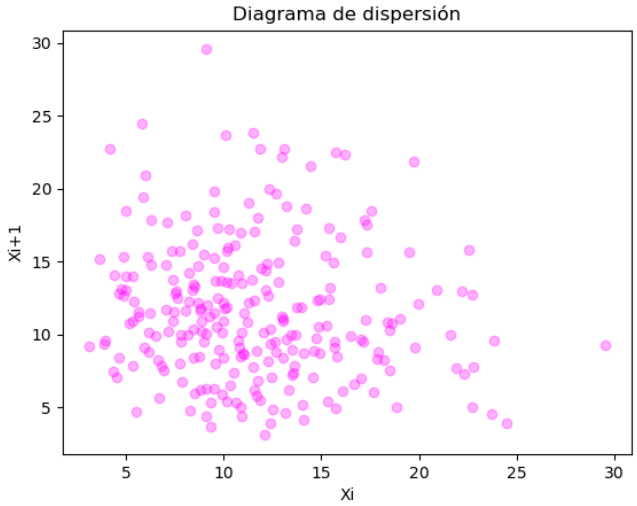
1 Introducción

A lo largo del curso hemos visto distintos métodos de análisis y validación estadística, los cuales nos permiten, a la hora de simular, seleccionar una distribución teórica y validar la calidad de nuestro modelo teórico con respecto a los datos.

En este trabajo hacemos uso de dichos métodos y herramientas para realizar un análisis estadístico de un conjunto de datos muestrales y a partir de este proponer y poner a prueba distintas hipótesis sobre la función de densidad de probabilidad teórica de dichos datos.

2 Diagrama de dispersión

A partir de los datos entregados realizamos el siguiente diagrama de dispersión:



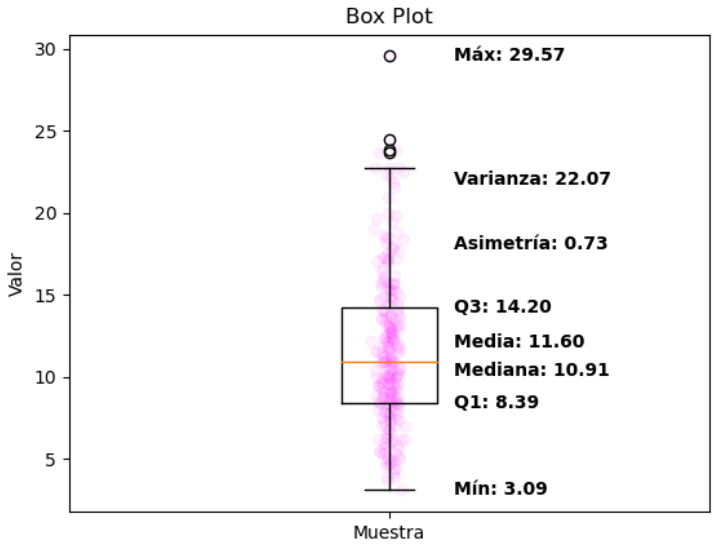
Este diagrama de dispersión surge de graficar en el plano cartesiano los pares (X_i, X_{i+1}) para $i = 1, 2, \ldots, N$, donde cada X_i es un valor de la muestra de tamaño $N = 250$.

No se observa una relación lineal clara entre las variables estudiadas. Es decir, los puntos en el gráfico no siguen un patrón lineal discernible. Esto sugiere que las variables son independientes entre sí.

3 Hipótesis sobre la familia de distribuciones

Basándonos en los datos proporcionados, calculamos las siguientes estimaciones muestrales:

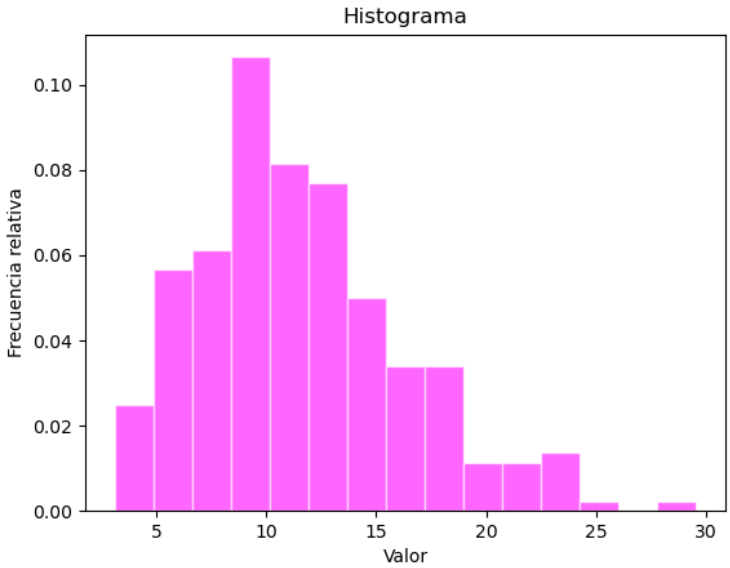
- Valor máximo: 29.567547
- Valor mínimo: 3.092925
- Media: 11.601359179999998
- Varianza: 22.066673719958665
- Skewness: 0.725423485173069
- Mediana: 10.9082725
- Cuantil 1: 8.39463
- Cuantil 3: 14.197010250000002



Podemos aseverar que la muestra sigue una distribución **asimétrica a la derecha**. Esta afirmación se basa en las siguientes observaciones:

La media es mayor que la mediana; el valor máximo se encuentra alejado tanto de la media como de la mediana; el valor del cuantil 1 se encuentra más

cerca de la mediana que del cuantil 3; y el valor de asimetría, skewness, como es un valor positivo indica una asimetría hacia la derecha.



Al analizar el histograma, podemos observar que la función de densidad propuesta no sigue una tendencia monótona. Esta falta de monotonía nos permite descartar la posibilidad de que la distribución subyacente tenga una función de densidad monótona creciente o monótona decreciente, como una exponencial, por ejemplo.

Algunas opciones posibles que satisfacen la propiedad de no contar con una tendencia monótona y tener una asimetría positiva son la distribución Gamma, Lognormal, Weibull y la distribución χ.

4 Estimación de parámetros

Basándonos en el análisis elaborado en la sección anterior, proponemos las siguientes familias de distribuciones:

- Distribución $Gamma(k, \theta)$, donde k es el parámetro de forma y θ es la escala
- Distribución $Lognormal(\mu, \sigma)$

Ahora debemos estimar los valores de los parámetros de las distribuciones propuestas. Para esto haremos uso del método de máxima verosimilitud.

4.1 Estimación log-normal

Si tenemos $X_1, \ldots, X_n$ datos observados, entonces los estimadores de máxima verosimilitud de los parámetros μ y σ se pueden estimar utilizando las siguientes fórmulas[2]:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} ln\ (X_i)}{n}$$
$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} (ln\ X_i - \hat{\mu})^2}{n}}$$

A partir de la muestra proporcionada y usando la estimación anterior obtenemos:

$$\hat{\mu} = 2.367063529069154$$
$$\hat{\sigma} = 0.4202521999350007$$

4.2 Estimación Gamma

A la hora de estimar por máxima verosimilitud los parámetros de una distribución Gamma, debemos derivar parcialmente sobre los dos parámetros θ y k.

Al derivar sobre el parámetro θ e igualar a 0 el logaritmo de la función verosimilitud obtenemos:

$$\hat{\theta} = \frac{\bar{x}}{k} \tag{1}$$

Notemos que la estimación de θ queda en función de k.

Si ahora derivamos parcialmente en función de k el logaritmo de la función de verosimilitud y reemplazamos θ con (1) obtenemos una ecuación no lineal para la cual no existe una solución cerrada para k. Es por esto que aproximaremos el parámetro k mediante la siguiente fórmula iterativa:

$$\begin{aligned} k_0 &= \frac{0.5}{log\ \bar{x} - \overline{log\ x}} \\ \frac{1}{k_{n+1}} &= \frac{1}{k_n} + \frac{\overline{log\ x} - log\ \bar{x} + log\ k - \Psi(k_n)}{k_n^2(1/k_n - \Psi'(k_n))} \end{aligned} \tag{2}$$

Donde $\bar{x}$ es la media muestral, $\overline{log\ x}$ es la media de los logaritmos de los datos y Ψ es la función digamma.

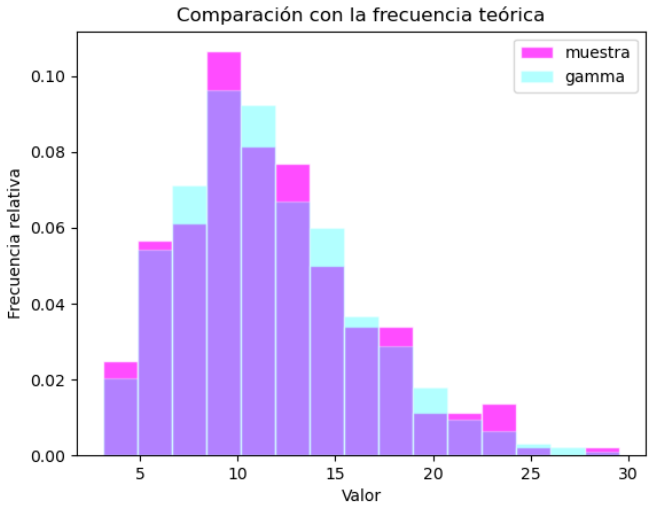
Tanto (1) como (2) fueron sacados de *Estimating a Gamma distribution*[1], en donde se encuentran explicados con más profundidad los resultados utilizados en este trabajo.

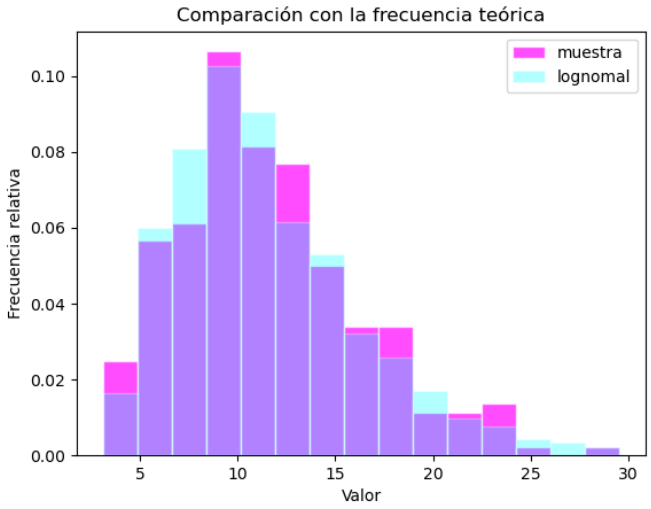
Al ejecutar dicho método con los datos brindados obtenemos los siguientes valores:

$$k = 6.11004526623543$$
$$\theta = 1.8987353897539814$$

4.3 Representación gráfica de las distribuciones estimadas

Luego de haber realizado las estimaciones por máxima verosimilitud, para comparar las frecuencias observadas y cada una de las funciones de densidad $f(x)$ propuestas para el ajuste, realizamos los gráficos que se encuentran a continuación. Superpusimos una barra adicional sobre cada barra del histograma de datos con altura $f(x)$. Esto nos permitió visualizar cómo se alinean las frecuencias esperadas de acuerdo con las funciones de densidad propuestas, con las frecuencias observadas en el histograma de datos.





5 Calidad de los ajustes

5.1 Test Chi-Cuadrado

Si bien tenemos datos de tipo continuo, aplicaremos el test chi-cuadrado discretizando los datos, es decir, agruparemos los datos otorgados en k intervalos.

Antes de realizar el test debemos determinar la cantidad de intervalos en los que vamos a categorizar nuestra muestra. Es importante tener en cuenta que elegir la cantidad y longitud de intervalos es un uno de los mayores problemas que trae el uso de este test, como bien explican Law y Kelton en "Simulation, Modeling, and Analysis"[2]:

> "El aspecto más problemático de realizar una prueba de chi-cuadrado es elegir el número y tamaño de los intervalos. Este es un problema difícil, y no se puede dar una receta definitiva que garantice buenos resultados en términos de validez (nivel real de la prueba cercano al nivel a deseado) y alto poder estadístico para todas las distribuciones alternativas y todos los tamaños de muestra."

Es por esto que hemos elegido usar la "regla de pulgar" de tomar $k = \sqrt{n}$, donde n es la cantidad de datos de la muestra, y por consiguiente tomaremos $k = 16$.

5.1.1 Distribución Gamma

Tenemos la siguiente hipótesis nula

$$H_0 = \text{Los datos provienen de la distribución continua Gamma}$$

Al calcular el estadístico de la prueba [3]:

$$T = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$$
$$T = 11.607646479132628$$

Y a partir de este, calculamos el p-valor sabiendo que si la hipótesis es cierta y n es suficientemente grande entonces el estadístico T tiene una distribución χ^2 con $k-3$ grados de libertad, ya que hemos estimado dos parámetros. Lo cual nos da $p - valor = 0.5600489712150449$

5.1.2 Distribución Lognormal

Tenemos la siguiente hipótesis nula

$$H_0 = \text{Los datos provienen de la distribución continua Lognormal}$$

Al igual que en el caso anterior calculamos el estadístico T [3]:

$$T = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$$
$$T = 16.331068205886353$$

Y nuevamente, a partir de este, calculamos el p-valor de forma análoga que en el caso anterior obteniendo $p - valor = 0.23171354246075926$

5.2 Test de Kolmogorov Smirnov

Dado que disponemos de una muestra de datos independientes, como se evidencia en el diagrama de dispersión, estamos en condiciones de aplicar el test de K-S.

Realizamos el test de K-S para las distribuciones Gamma y Lognormal, utilizando los parámetros previamente estimados.

5.2.1 Distribución Gamma

Partimos de la siguiente hipótesis nula:

$$H_0 = \text{Los datos provienen de la distribución continua } Gamma(k, \theta)$$

Y calculamos el estadístico de Kolmogorov-Smirnov [3]:

Para ello ordenamos las observaciones de menor a mayor y llamamos $Y_{(j)}$ al dato que ocupa el j-ésimo lugar lugar luego del ordenamiento.

Llamamos n al tamaño de la muestra. Y $F_e = \#\{j|Y_j \leq x\}/n$ a la distribución empírica.

$$
\begin{aligned}
D_{Gamma} &= \sup_{x \in R}(|F_e(x) - F_{Gamma}(x)|) \\
&= \max_{1 \leq j \leq n} \left\{ \frac{j}{n} - F_{Gamma}(Y_{(j)}), F_{Gamma}(Y_{(j)}) - \frac{j-1}{n} \right\} \\
&= 0.02775736785617522
\end{aligned}
$$

Una vez obtenido el estadístico, estimamos el p-valor para el cual generamos $n_{sim} = 10000$ simulaciones de muestras de variables aleatorias uniformes en $(0, 1)$ y calculamos la proporción de valores generados que exceden a D_{Gamma}. Finalmente, obtuvimos un $p - valor = 0.9868$.

5.2.2 Distribución Lognormal

En este caso, la hipótesis nula estará dada por:

$$H_0 = \text{Los datos provienen de la distribución continua } Lognormal(\mu, \sigma)$$

Calculamos el estadístico de Kolmogorov-Smirnov [3]:

Para lo cual también ordenamos las observaciones de menor a mayor, llamamos $Y_{(j)}$ al dato que ocupa el j-ésimo lugar lugar luego del ordenamiento; llamamos F_e a la distribución empírica y n al tamaño de la muestra.

$$
\begin{aligned}
D_{Lognorm} &= \sup_{x \in R}(|F_e(x) - F_{Lognorm}(x)|) \\
&= \max_{1 \leq j \leq n} \left\{ \frac{j}{n} - F_{Lognorm}(Y_{(j)}), F_{Lognorm}(Y_{(j)}) - \frac{j-1}{n} \right\} \\
&= 0.042845431863045846
\end{aligned}
$$

De la misma forma que con la Gamma, estimamos el p-valor a través de $n_{sim} = 10000$ simulaciones de muestras de variables aleatorias uniformes en $(0, 1)$ pero ahora calculando la proporción de valores generados que exceden a $D_{Lognormal}$. Y así, obtuvimos un $p - valor = 0.7294$.

5.3 Conclusiones

Si bien tanto para el test chi-cuadrado como para el test de Kolmogorov-Smirnov no se alcanza evidencia suficiente para rechazar las hipótesis, elegimos como distribución de probabilidad definitiva para el conjunto muestral otorgado la distribución Gamma, ya que los dos test coinciden en que podemos rechazar la hipótesis de que la muestra proviene de una distribución Lognormal con un nivel de confianza más alto que para la Gamma.

References

[1] Thomas P. Minka, *Estimating a Gamma distribution*.

[2] A. M. Law, W. D. Kelton, *Simulation, Modeling, and Analysis* pp. 295, 349.

[3] Dra. Patricia Kisbye, *Modelos y Simulación 2023*.