Trabajo Práctico Especial de Modelos y Simulación

Junio 2023

Bas Peralta, Benjamín

Mondejar, Antonio

Introducción

Dado un conjunto de observaciones obtenidas a partir de una recolección de datos puede ser de interés determinar **cuál es la distribución de estos datos**.

Una forma de determinar si un conjunto de observaciones proviene de una distribución dada es a través de las **pruebas de bondad de ajuste**. Una prueba de bondad de ajuste es un test de hipótesis, en la cual la hipótesis nula, H_0, afirma que los datos provienen de una determinada distribución F. La hipótesis alternativa, H_1, es la negación de H0.

Según cuál sea la hipótesis se define un determinado **estadístico muestral** $T = T(X_1, X_2, \ldots, X_n)$. El estadístico es una variable aleatoria, que, bajo la hipótesis nula, tiene una distribución conocida o de la cual se saben algunas propiedades. Esto es, se conoce algo de $P_{H_0}(T \leq t)$, donde el subíndice H_0 indica que la distribución de T está basada en que las observaciones satisfacen la hipótesis H_0.

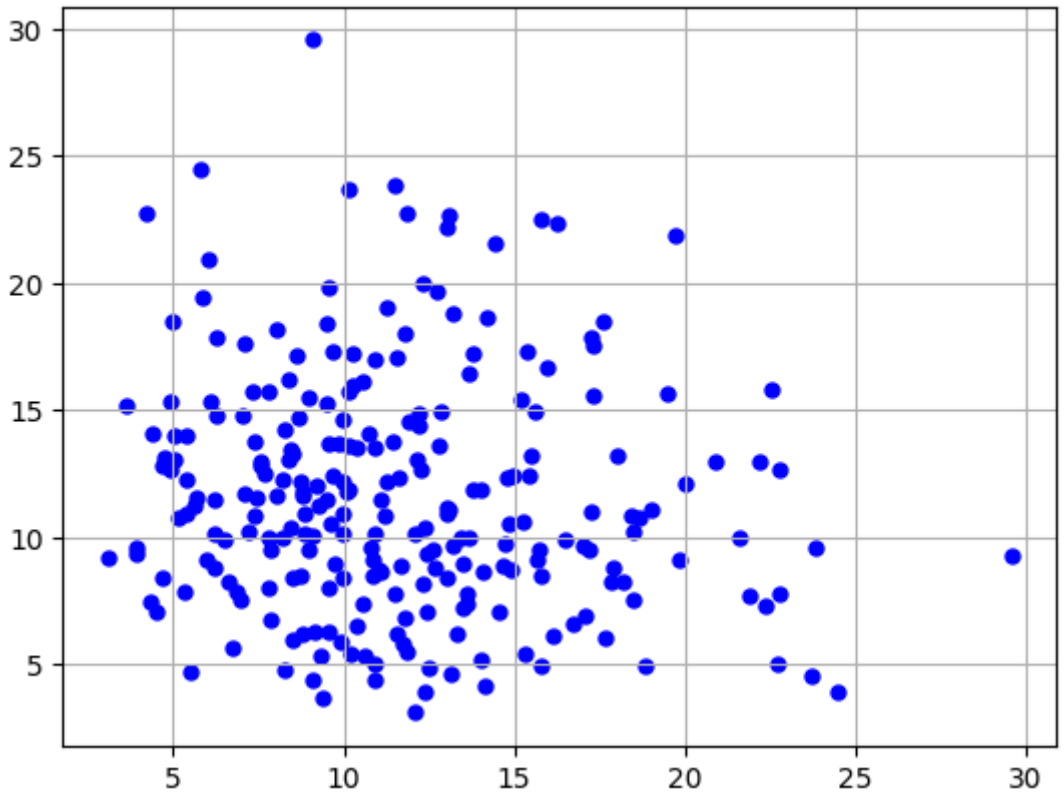
Así, dada una muestra de datos $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, se evalúa el estadístico en esta muestra y se toma una decisión de rechazar o no rechazar la hipótesis nula según "cuán probable" es haber obtenido ese valor.

Dicho todo esto, el objetivo del presente trabajo práctico es la elaboración de una hipótesis sobre la densidad de probabilidad teórica a la cual obedece un conjunto de datos muestrales.

Actividad 1: Diagrama de dispersión

Estudiar la independencia estadística de los datos de la muestra. A tal fin, construir el "diagrama de dispersión", esto es, el gráfico de los pares (X_i, X_{i+1}), con i = 1, …, n-1, donde n es el número de datos de la muestra. Interpretar el diagrama obtenido.

El diagrama de dispersión obtenido fue el siguiente:



Este diagrama nos permite estudiar la **independencia** de los datos. Si existiera una correlación entre los datos muestrales, el diagrama de dispersión mostraría que los puntos tienden a alinearse en el plano. Ahora bien, si no existiera tal correlación se esperaría que el diagrama muestre que los puntos se dispersan aleatoriamente en el plano.

Analizando el diagrama de la muestra estudiada podemos ver que lo que ocurre es lo segundo. Esto evidencia que no se puede encontrar una correlación evidente entre los distintos datos, por lo cual podemos asumir que los datos de la muestra son **independientes**.

Actividad 2: Elaboración de hipótesis

Elaboración de la hipótesis sobre la familia de distribuciones a la que pertenece la muestra. A tal fin, realizar:

a) Las estimaciones muestrales de: valores máximos y mínimos, media, varianza y "skewness" (medida de la asimetría de la distribución).

Para llevar a cabo dichas estimaciones, usamos sus estimadores, los cuales están especificados en el siguiente cuadro:

Función	Estimador muestral	Estima
Min, Max	$X_{(1)}, X_{(n)}$	rango
Media μ	$\overline{X}(n)$	Tendencia central
Mediana	$\hat{m} = \begin{cases} X_{(n+1)/2} \\ \frac{1}{2}(X_{n/2} + X_{(n/2+1)}) \end{cases}$	Tendencia central.
Varianza σ^2	$S^2(n)$	Variabilidad
c.v.=$\frac{\sigma}{\mu}$	$\hat{cv}(n) = \frac{\sqrt{S^2(n)}}{\overline{X}(n)}$	Variabilidad
τ	$\hat{\tau} = \frac{S^2(n)}{\overline{X}(n)}$	Variabilidad
Asimetría $\nu = \frac{E[(X-\mu)^3]}{(\sigma^2)^{3/2}}$	$\hat{\nu}(n) = \frac{\sum_i (X_i - \overline{X}(n))^3/n}{[S^2(n)]^{3/2}}$	Simetría

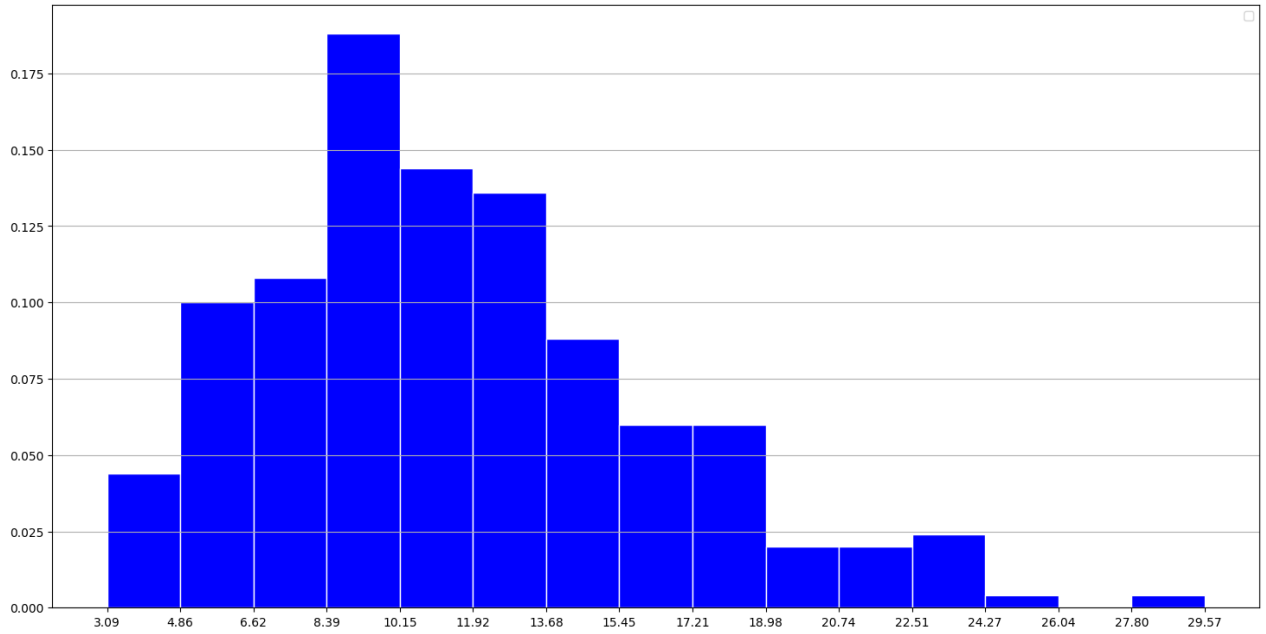
Y las estimaciones obtenidas fueron las siguientes:

Valor máximo	29.5675
Valor mínimo	3.09293
Media	11.6014
Varianza	22.0667
Asimetría	0.725423

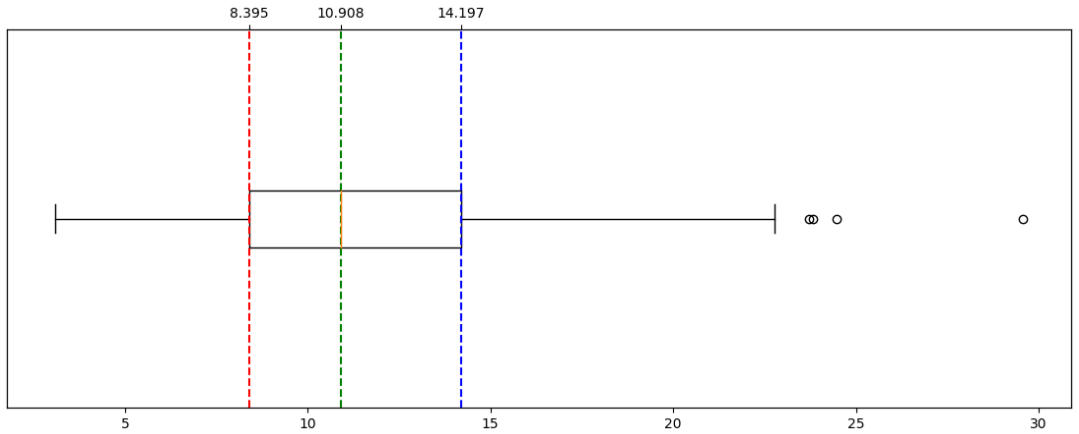
b) La confección de un histograma con los datos muestrales.

Como la muestra se trata de datos continuos, para llevar a cabo este histograma decidimos agrupar los datos en intervalos de la misma longitud. La elección de cantidad de intervalos la hicimos siguiendo la **regla de la raíz cuadrada**, que consiste en considerar k ≈ $\sqrt{n}$ cantidad de intervalos donde n es la cantidad de datos.

Dicho esto, el histograma obtenido fue el siguiente:



c) El estudio de cuantiles en la muestra y confeccionar el correspondiente "box plot".



Habiendo hecho todo este análisis sobre la muestra, pudimos sacar algunas conclusiones que consideramos nos serán útiles a la hora de elegir la distribución con la cual ajustar nuestros datos:

- Como $\overline{X}(250) = 11.6014 > 10.9083 = \widehat{x}_{0.5}$ (mediana muestral), podemos concluir que los datos son asimétricos.
- Además, como la asimetría muestral es positiva, son asimétricos hacia la derecha, es decir hay valores más separados de la media a la derecha de la misma.

Actividad 3: Proposición de familias de distribuciones de probabilidad

Proposición de al menos dos familias de distribuciones de probabilidad como modelos de ajuste de los datos. Realizar la estimación de los parámetros de las correspondientes familias de distribuciones seleccionadas, utilizando el método de máxima verosimilitud.

Para llevar a cabo esta propuesta, nos basamos en las siguientes observaciones obtenidas a partir del análisis de lo realizado en la actividad 2:

- Los datos son **continuos**.
- Todos los datos son **positivos**.
- Los datos son **asimétricos hacia la derecha**: viendo el histograma del Ejercicio 1, resulta claro que hay valores contenidos en ciertos intervalos que ocurren con más frecuencia que otros. Por ejemplo, los datos contenidos en el intervalo (7.33, 13.68) son más que los del intervalo (18.98, 24.27). Además, decimos que son asimétricos hacia la derecha porque en el boxplot del Ejercicio 3 se puede ver claramente que la caja está hacia la izquierda, lo cual indica que hay valores más separados de la mediana a la derecha.
- Los datos **atípicos** presentes en la muestra son todos **positivos**: esto se puede ver analizando el box plot del Ejercicio 2.

Teniendo todos estos factores en cuenta y guiándonos por la forma de las funciones de densidad de probabilidad de las distintas distribuciones, decidimos ajustar los datos con las siguientes distribuciones:

- Distribución **Lognormal** de parámetros $\boldsymbol{\mu}$ y $\boldsymbol{\sigma}$.
- Distribución **Gamma** de parámetros $\boldsymbol{\alpha}$ y $\boldsymbol{\beta}$.

Estimación de parámetros de las distribuciones

Distribución Lognormal

El estimador de máxima verosimilitud para **μ** es:

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

El estimador de máxima verosimilitud para **σ** es:

$$\hat{\sigma} = \left[\frac{\sum_{i=1}^{n} (ln(X_i - \hat{\mu})^2}{n} \right]^{1/2}$$

Usando estas fórmulas, obtuvimos:

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

Distribución Gamma

Los estimadores de máxima verosimilitud para **α** y **β** deben satisfacer las siguientes ecuaciones:

$$ln(\hat{\beta}) + \psi(\hat{\alpha}) = \frac{\sum_{i=1}^{n} ln(X_i)}{n} \qquad y \qquad \hat{\alpha}\hat{\beta} = \overline{X}$$

donde $\psi(x)$ es la función **digamma**, definida de la siguiente forma:

$$\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$$

donde $\Gamma(x)$ es la función **gamma**.

Siguiendo lo indicado en esta página, a partir de las ecuaciones anteriores, se puede obtener el valor de α de forma numérica utilizando el método de Newton para aproximar raíces de funciones no lineales, y una vez que se tiene este valor, utilizando la segunda ecuación de las de arriba, es fácil obtener el valor de β. Usando dicho procedimiento, obtuvimos:

- $\widehat{\alpha} = 6.110045266065787$
- $\widehat{\beta} = 1.898735389806699$

Actividad 4: Determinación de la calidad de los ajustes logrados

a) Realizar un comparación de frecuencias entre el histograma de datos y cada una de las funciones de densidad f(x) propuestas para el ajuste. A tal fin, superponer sobre cada barra del histograma de datos una barra con altura igual a Δb f(x), donde Δb corresponde al ancho de intervalo en el histograma y f(x) es cada una de las densidades propuestas.

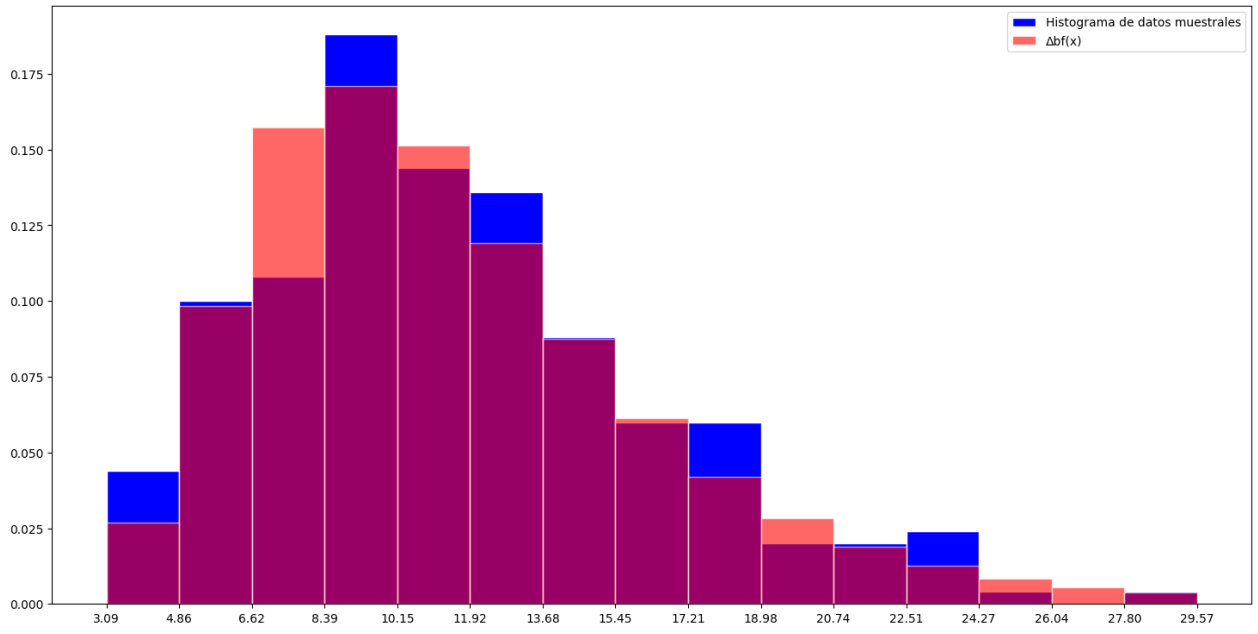
Recordemos que para hacer el histograma inicial, elegimos k ≈ $\sqrt{n}$ cantidad de intervalos. En este caso n = 250, y entonces tomamos k = 15.

Luego:

$$\Delta b = \frac{X_{(n)} - X_{(1)}}{k} = \frac{X_{(250)} - X_{(1)}}{15} = \frac{29.5675 - 3.09293}{15} \simeq 1.76497$$

Por otro lado, para graficar las nuevas barras Δb f(x) usando los mismos intervalos del histograma original, y elegimos evaluar cada f(x) en el punto medio de cada intervalo.

<u>Distribución Lognormal</u>

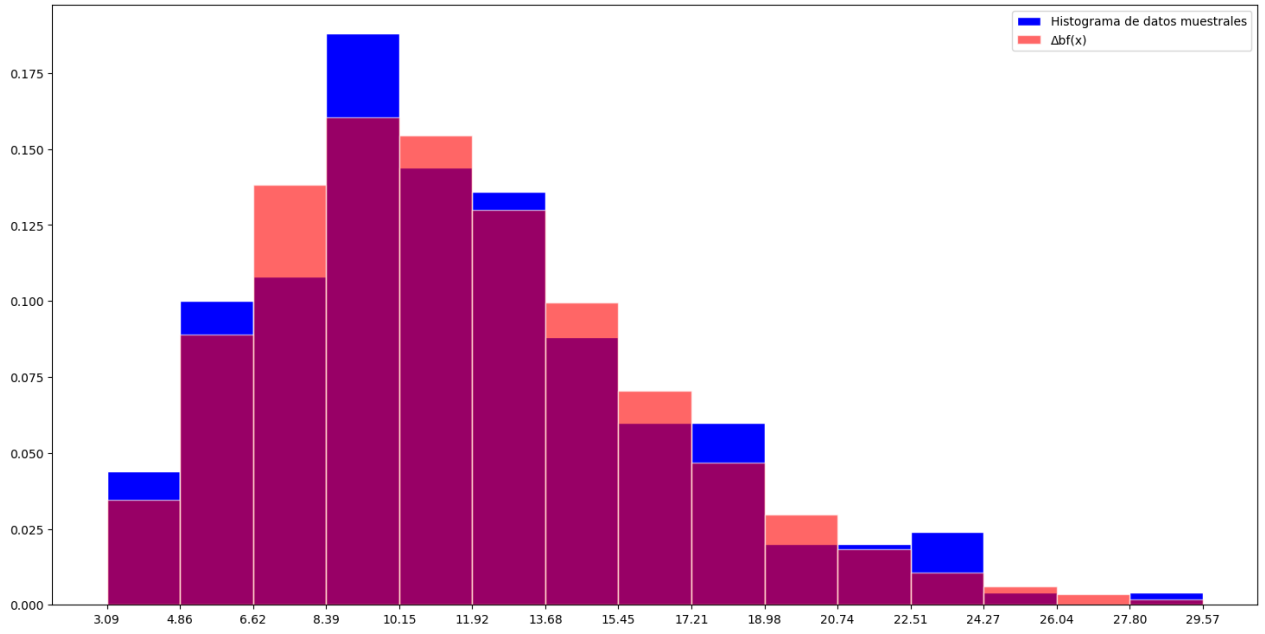


Las diferencias de frecuencias obtenidas en cada uno de los 15 intervalos fueron las siguientes (yendo de izquierda a derecha):

0.01726025, 0.00153608, 0.04936358, 0.01705337, 0.00733159, 0.01687268, 0.00061045, 0.00138549, 0.01801955, 0.00824912, 0.00116503, 0.01149848, 0.0042869, 0.00549814, 0.00034316,

sumando un total (redondeado) de 0.1604739.

Distribución Gamma



Las diferencias de frecuencias obtenidas en cada uno de los 15 intervalos fueron las siguientes (yendo de izquierda a derecha):

0.00950984, 0.01100835, 0.03023389, 0.02744346, 0.01041981, 0.00590232, 0.01136564, 0.01036779, 0.01307168, 0.00980449, 0.00182262, 0.01328609, 0.00213328, 0.00342361, 0.00213054,

sumando un total (redondeado) de 0.1619234.

b) Estimar el p-valor de la prueba de la hipótesis de que los datos provienen de las distribuciones sugeridas, utilizando la **aproximación ji-cuadrada**.

Los intervalos que usaremos para agrupar los datos serán los que se vienen usando en los histogramas, con la única diferencia de los últimos dos intervalos se juntarán en un solo, que específicamente será (26.04, ∞), y que el primer intervalo será (-∞, 4.86). Este cambio se debe a que ninguno de los datos muestrales cae en el intervalo (26.04, 27.80) (notar que en este intervalo, la columna del histograma de los datos muestrales tiene altura 0) ni en el (-∞, 3.09).

Recordemos que bajo la hipótesis nula de que los datos provienen de la distribución continua F, el estadístico para el test chi cuadrado de Pearson está dado por:

$$T = \sum_{i=1}^{k} \frac{(N_i - n\,p_i)^2}{n\,p_i}$$

donde:

- k es la cantidad de intervalos considerados.
- N_i es el número de observaciones en el intervalo i.
- n es el tamaño de la muestra.
- p_i es la probabilidad dada por la distribución F de que la variable esté en el i-ésimo intervalo.

Si la hipótesis nula es cierta y n es grande, entonces el estadístico T tiene una distribución aproximadamente χ - cuadrado con k - 1 grados de libertad: χ^2_{k-1} (de aquí el nombre de este test).

Una vez que se obtiene el valor de T observado a partir de la muestra, llamémosle t, el p-valor en este test se calcula de la siguiente forma:

$$p - valor = P_{H_0}(T \geq t) \simeq P(\chi^2_{k-1-m} \geq t)$$

donde m es la cantidad de parámetros estimados para la distribución propuesta en la hipótesis nula (en nuestro caso, m = 2 para ambas distribuciones).

Algo a notar es que si el p-valor es muy próximo al nivel de rechazo que se proponga, puede existir la duda si conviene o no rechazar la hipótesis. Una forma de ayudar a tomar esta decisión es simular muestras de tamaño n de la distribución F y para cada una de ellas calcular el estadístico T. Para un número de simulaciones suficientemente grande, la proporción de valores de T que exceden al valor T = t tomado en la muestra original es una buena estimación del p-valor.

	Lognormal	**Gamma**
Valor del estadístico T observado en la muestra	14.3710185	10.5324493
p-valor (**usando la f.d.a de χ^2**)	0.213139	0.4832218
p-valor (**con 1000 simulaciones**)	0.224	0.531

c) Estimar el p-valor de la prueba de la hipótesis de que los datos provienen de las distribuciones sugeridas, en base al **estadístico de Kolmogorov-Smirnov**.

Bajo la hipótesis nula de que los datos de la muestra provienen de la distribución continua F, el test de Kolmogorov-Smirnov esencialmente compara la distribución empírica de los datos con la distribución F, estimando la máxima distancia entre los dos gráficos.

En primer lugar, se ordenan los datos de mayor a menor: con $Y_{(j)}$ denotamos el dato que ocupa el j-ésimo lugar luego del ordenamiento. Luego, el estadístico de Kolmogorov-Smirnov está dado por:

$$D = sup_{x \in \mathbb{R}} |F_e(x) - F(x)|$$

donde

$$F_e(x) = \frac{\#\{j \mid Y_j \leq x\}}{n}$$

donde n es el tamaño de la muestra.

Luego de algunos razonamientos, se puede llegar a que:

$$D = máx_{1 \leq j \leq n}\left\{\frac{j}{n} - F(Y_{(j)}),\ F(Y_{(j)}) - \frac{j-1}{n}\right\}$$

Una vez que se obtiene el valor de D observado a partir de la muestra, llamémosle t, el p-valor en este test se calcula de la siguiente forma:

$$p - valor = P_{H_0}(D \geq d)$$

Bajo la hipótesis H_0, el estadístico D sigue una distribución de Kolmogorov, con una expresión bastante compleja. Por lo tanto, en la práctica es conveniente calcular el p-valor con simulaciones. Esto es, realizar k simulaciones de muestras de tamaño n de una variable con distribución F, calcular el correspondiente valor del estadístico $D = d_i$ para cada muestra. Finalmente, se estima el p-valor con la proporción de valor d_i que exceden al valor d, tomado en la muestra original.

	Lognormal	**Gamma**
Valor del estadístico D observado en la muestra	0.0428454	0.0277574
p-valor (**con 1000 simulaciones**)	0.34	0.924

d) Seleccionar finalmente una de las densidades de probabilidad propuestas y argumentar los motivos de dicha elección.

Para realizar esta elección, decidimos separarla según diferentes criterios:

Comparación de frecuencias (**Actividad 4**)**a**))

Observemos que la altura de cada barra del histograma inicial (el de la Actividad 2)b)) corresponde a la **proporción $\mathbf{p_j}$** de valores de la muestra que se encuentran en el intervalo $[b_{j-1}, b_j)$ correspondiente a esa barra.

Luego para cada intervalo tomamos su punto medio, esto es:

$$y_j = \frac{b_{j-1}+b_j}{2}$$

y evaluamos la función de densidad de probabilidad correspondiente a la distribución siendo analizada. Finalmente graficamos para cada intervalo j, una barra de altura $\Delta b\ f(y_j)$ siendo Δb el ancho del intervalo en cuestión y f la función de densidad de probabilidad de la distribución siendo analizada (en Actividad 4)a) se encuentra detallado cuál es su valor de Δb, que en este caso es el mismo para todos los intervalos) .

Entonces en este punto tenemos dos histogramas a comparar: uno con las proporciones p_j y el otro con las $\Delta b\ f(y_j)$. Si la distribución considerada fuera una buena representación de la verdadera distribución de los datos observados entonces ambos histogramas deberían asemejarse.

Analizando los histogramas obtenidos para ambas distribuciones que elegimos, podemos notar que las dos opciones aparentan ser una buena representación de la distribución de la muestra observada. Además, las diferencias entre los histogramas de las densidades de ambas distribuciones con el histograma de las proporciones son muy similares, por lo que consideramos que no tenemos la información suficiente para inclinarnos por una distribución sobre la otra a partir de esta métrica.

Test chi-cuadrado de Pearson (**Actividad 4**)**b**))

Teniendo en cuenta la forma en que se calcula T (explicado en la Actividad 4)b)), se lo puede considerar como una medida de distancia entre la distribución empírica de los datos y la distribución F de la hipótesis nula.

Por lo tanto, si el valor de T es grande, se considera que hay evidencias sobre que la muestra no proviene de la distribución F, y luego se **rechaza la hipótesis nula**. Por lo contrario, si el valor de T es pequeño, se dice que **no hay evidencias suficientes para rechazar la hipótesis nula**.

Ahora bien, la noción de qué "tan grande" o qué "tan chico" debe ser T para rechazar la hipótesis nula viene dada por el nivel de rechazo, denotado con la letra α, de forma tal que si el valor obtenido de T a partir de la muestra es t y

$$p - valor = P_{H_0}(T \geq t) \leq \alpha,$$

entonces se **rechaza** la hipótesis nula con un nivel de rechazo α, y un nivel de confianza (1 - α). Cuánto más grande sea el nivel de rechazo α, será más probable que se rechace la hipótesis nula H_0 pero menor será la confianza que éste rechazo nos provee. En otras palabras, la probabilidad de rechazar la hipótesis nula erróneamente será mayor.

Analizando los tests realizados, para ninguna de nuestras hipótesis existe un nivel de rechazo razonable que nos permita rechazarlas. En particular, el mínimo de nivel de rechazo que podríamos considerar para rechazar alguna de las hipótesis es 0.23.

Sin embargo, a pesar de no poder rechazar ninguna de las dos hipótesis, como el p-valor estimado bajo la hipótesis de que los datos observados provienen de una distribución Gamma es mayor que el p-valor estimado bajo la hipótesis de que los datos provienen de una distribución Lognormal (proveniente de un valor observado del estadístico menor en la distribución Gamma) entonces podemos deducir (teniendo en cuenta

lo que representa T) que la distancia entre la distribución empírica de los datos y la distribución Gamma es menor que la distancia de la distribución empírica de los datos con la distribución Lognormal, por lo cual bajo este criterio, optamos por elegir la distribución Gamma.

Test de Kolmogorov-Smirnov (**Actividad 4**)**c**))

Como los datos son de tipo continuo, el test chi cuadrado de Pearson tiene la desventaja de que al agrupar los datos en intervalos, no se considera cómo se distribuyen dentro de los mismos, por lo que el test de **Kolmogorov-Smirnov** resulta más adecuado para este tipo de datos.

Al igual que en el test Chi cuadrado de Pearson, si el valor de D es grande, se considera que hay evidencias sobre que la muestra no proviene de la distribución F, y luego se **rechaza la hipótesis nula**. Por lo contrario, si el valor de D es pequeño, se dice que **no hay evidencias suficientes para rechazar la hipótesis nula**. Si el valor obtenido de D a partir de la muestra es d, y se cumple que:

$$p - valor = P_{H_0}(D \geq d) \leq \alpha,$$

entonces se **rechaza** la hipótesis nula con un nivel de rechazo α, y un nivel de confianza (1 - α). Cuánto más grande sea el nivel de rechazo α, será más probable que se rechace la hipótesis nula H_0 pero menor será la confianza que éste rechazo nos provee. En otras palabras, la probabilidad de rechazar la hipótesis nula erróneamente será mayor.

Analizando los tests realizados, para ninguna de nuestras hipótesis existe un nivel de rechazo razonable que nos permita rechazarlas. En particular, el mínimo nivel de rechazo que podríamos considerar para rechazar alguna de las hipótesis es 0.35.

Similarmente a lo dicho para el test Chi cuadrado de Pearson, a pesar de no poder rechazar ninguna de las dos hipótesis, como el p-valor estimado bajo la hipótesis de que los datos observados provienen de una distribución Gamma es mayor que el p-valor estimado bajo la hipótesis

de que los datos provienen de una distribución Lognormal (proveniente de un valor observado del estadístico menor en la distribución Gamma) entonces podemos deducir (teniendo en cuenta lo que representa D) que la distancia máxima entre los gráficos de la distribución empírica de los datos y la distribución Gamma es menor que la distancia máxima entre los gráficos de la distribución empírica de los datos con la distribución Lognormal, por lo cual bajo este criterio, optamos por elegir la distribución Gamma.

Conclusión final

Juntando los 3 tipos de criterios mencionados anteriormente, llegamos a la conclusión de que la distribución que mejor representa la distribución verdadera de los datos observados es la Gamma por las razones previamente mencionadas.

Fuentes consultadas

- Distribuciones de probabilidad con Python
- Simulation, Modeling and Analysis - Averill M. Law - 5th Edition (Capítulo 6)
- Estimación de máxima verosimilitud de los parámetros de una distribución gamma
- Apunte de la materia Modelos y Simulación 2023 - Patricia Kisbye