Trabajo Práctico Especial: Selección de Distribuciones de Probabilidad

Modelos y Simulación - 2024

Índice

Modelos y Simulación - 2024
- Índice

Integrantes

Introducción

Actividad 1
- 1.1. Consigna
  - **1.2. Item A | Estimaciones Muestrales**
  - 1.3. Item B | Histograma de la Muestra
  - 1.4. Item C | Box-Plot

Actividad 2
- 2.1. Consigna
  - **2.2. Desarrollo | Modelos de Ajuste**
  - 2.3 Planteamiento de la Hipótesis Nula.
  - 2.4. Gráficos de Ajuste

Actividad 3
- 3.1. Consigna
- 3.2. Funciones de Densidad Acumulada
- 3.3. Item A | Comparación de Frecuencias
- 3.4. Item B | Prueba de Hipótesis con Ji-Cuadrada
- 3.5. Item C | Prueba de Hipótesis con Kolmogorov-Smirnov
- 3.6. Item D | Conclusiones

Integrantes

- Juan Bratti

- Emanuel Nicolás Herrador

Introducción

En este trabajo especial, vamos a poner en práctica lo estudiado acerca del análisis de las distribuciones, el análisis estadístico y las técnicas de validación estadística para corroborar o teorizar sobre el origen de los datos de una muestra.

Más precisamente, usando una muestra de tamaño N, vamos a querer ajustarla a ciertas distribuciones propuestas por nosotros (con o sin sus parámetros correspondientes, utilizando estimadores).

El objetivo final va a ser teorizar si los datos provienen de dichas distribuciones o no. Para esto, necesitaremos identificar y evaluar una hipótesis nula H_0 y una hipótesis alternativa H_1, utilizando los tests de hipótesis de *Pearson* (una variante con intervalos para datos continuos) y de *Kolmogorov-Smirnov*.

Actividad 1

1.1. Consigna

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).
b. La confección de un histograma con los datos muestrales.
c. El estudio de cuantiles en la muestra y confeccionar el correspondiente “box plot”.

1.2. Item A | Estimaciones Muestrales

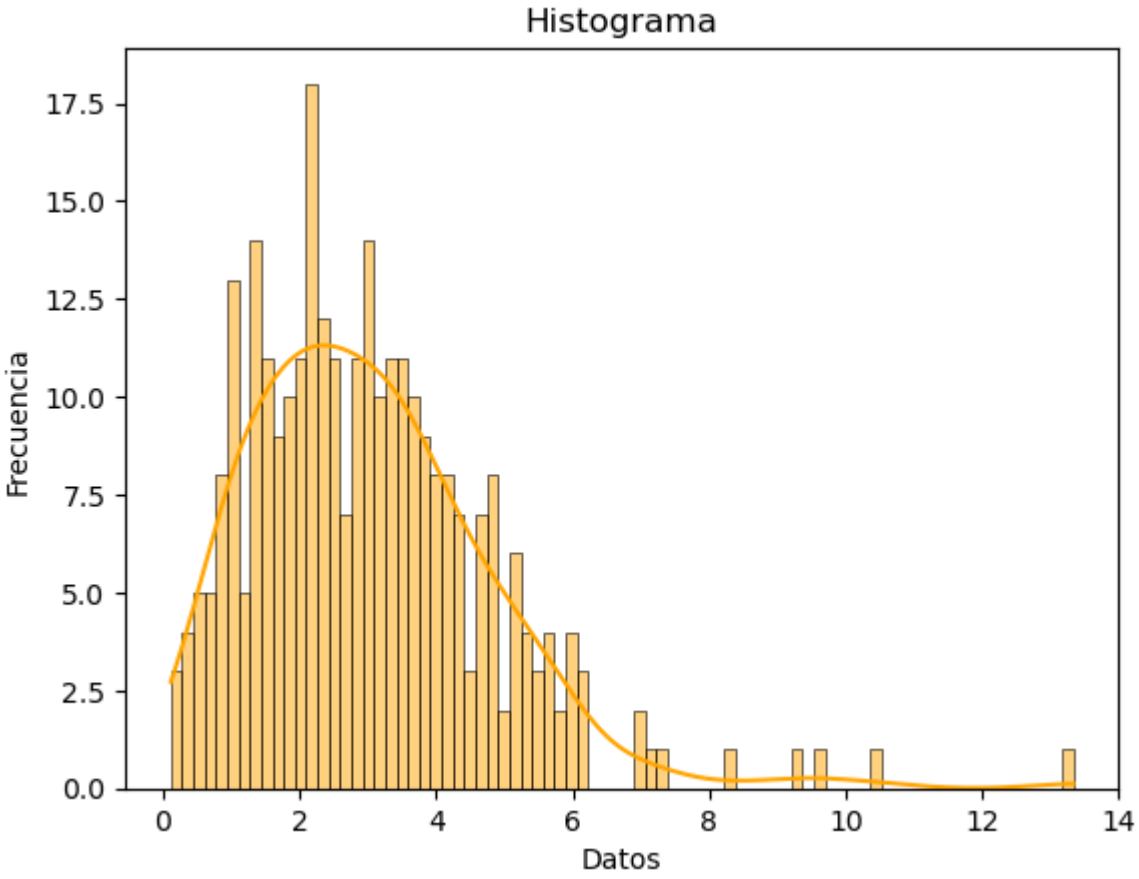
Los estimadores solicitados para nuestra muestra son los siguientes:

- **Valor máximo**: $X_{(n)} = 13.342856$
- **Valor mínimo**: $X_{(1)} = 0.111128$
- **Media muestral**: $\overline{X}(n) = \frac{\sum_1^n X_i}{n} = 3.017981$
- **Varianza muestral**: $S^2(n) = \frac{\sum_1^n (X_i - \overline{X}(n))^2}{n-1} = 3.209765$

- **Asimetría**: $\frac{\sum_1^n (X_i - \overline{X}(n))^3}{n * (S^2(n))^{3/2}} = 1.391727$

1.3. Item B | Histograma de la Muestra

El histograma de la muestra es el siguiente:



Gracias a lo cual podemos ver que tiene "forma" de *normal* con cola derecha (sin poseer valores negativos). Este es un punto muy importante que notaremos luego, dado que es un factor clave para decidir a qué distribución pertenecen los datos.

Del histograma podemos sacar la "forma" de nuestros datos y el hecho que son *todos* positivos, además de que posee valores *ouliers* extremos a derecha.

1.4. Item C | Box-Plot

Vamos a considerar el estudio de los cuantiles más importantes a la hora de realizar el gráfico de *Box-Plot*, los cuales corresponden entonces a los cuantiles 0.25, 0.5 y 0.75.

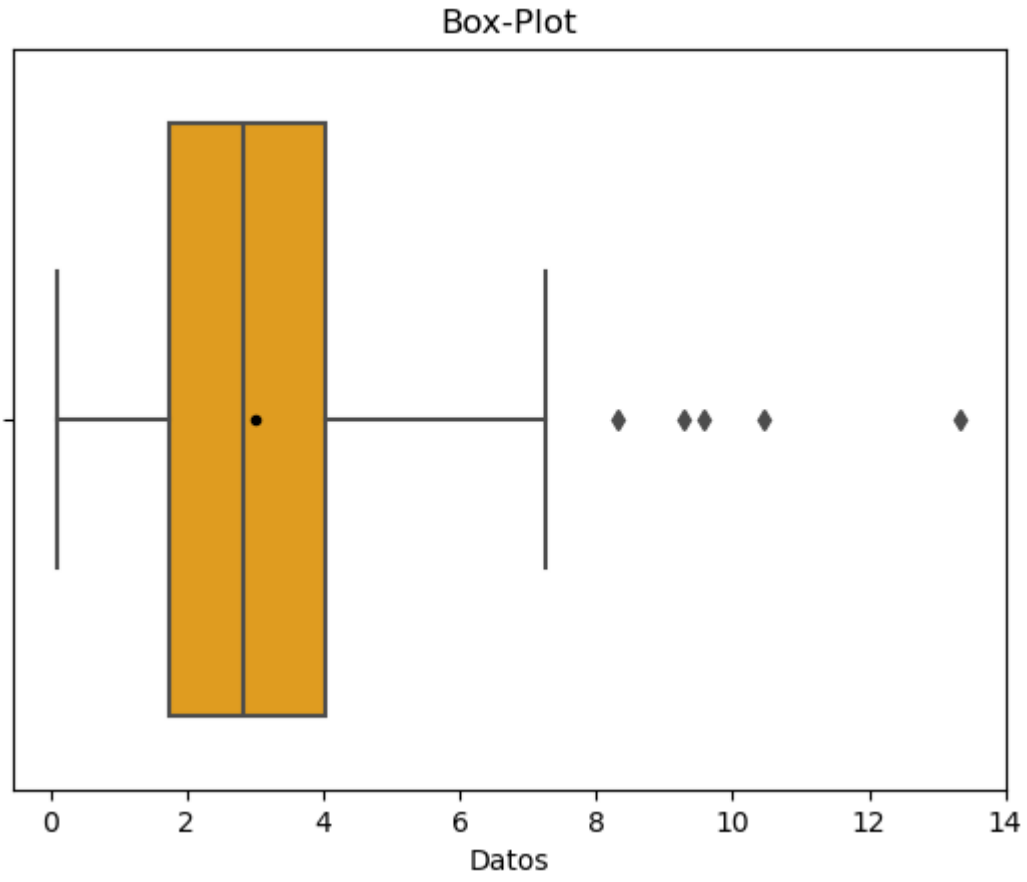
A su vez, para ver de forma correcta si existen datos *outliers*, vamos a considerar también los cuantiles 0.05 y 0.95.

A continuación, se presentan los resultados obtenidos:

$$0.05 \rightarrow 0.753190$$
$$0.25 \rightarrow 1.729471$$
$$0.50 \rightarrow 2.836248$$
$$0.75 \rightarrow 4.030335$$
$$0.95 \rightarrow 5.924371$$

Se puede notar que sí existen *outliers*, más que nada a la derecha de la distribución (lo cual puede verse fácilmente con el *histograma* anterior), porque el máximo es 13.342856 y el cuantil 0.95 es 5.924371.

Todo esto queda más claro cuando se realiza el gráfico de *Box-Plot*.



Algo curioso a destacar de este gráfico, además de los *outliers*, es que la media se encuentra muy cercana a la mediana de nuestros datos. Lo cual

también puede ser de ayuda para saber de qué distribución se trata.

Actividad 2

2.1. Consigna

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 maxima verosimilitud.

2.2. Desarrollo | Modelos de Ajuste

Visto los datos y la representación gráfica de los mismos en un histograma y en un box-plot, las distribuciones propuestas como modelos de ajuste y sus parámetros a estimar son:

- Distribución Normal:
  - Parámetro 1: media $\mu = \hat{\mu} = \overline{X}(n) \approx 3.02$
  - Parámetro 2: desviación estándar $\sigma = \hat{\sigma} = S(n) \approx 1.79$
- Distribución Gamma:
  - Parámetro 1: α donde $E[X] = \frac{\alpha}{\beta} \therefore \alpha = \hat{\alpha} = \hat{\beta} * E[X] = \overline{X}(n) * \hat{\beta} \approx 2.84$
  - Parámetro 2: β donde $Var[X] = \frac{\alpha}{\beta^2}$ y $E[X] = \frac{\alpha}{\beta} \therefore \beta = \hat{\beta} = \frac{E[X]}{Var[X]} = \frac{\overline{X}(n)}{S^2(n)} \approx 0.94$

2.3 Planteamiento de la Hipótesis Nula.

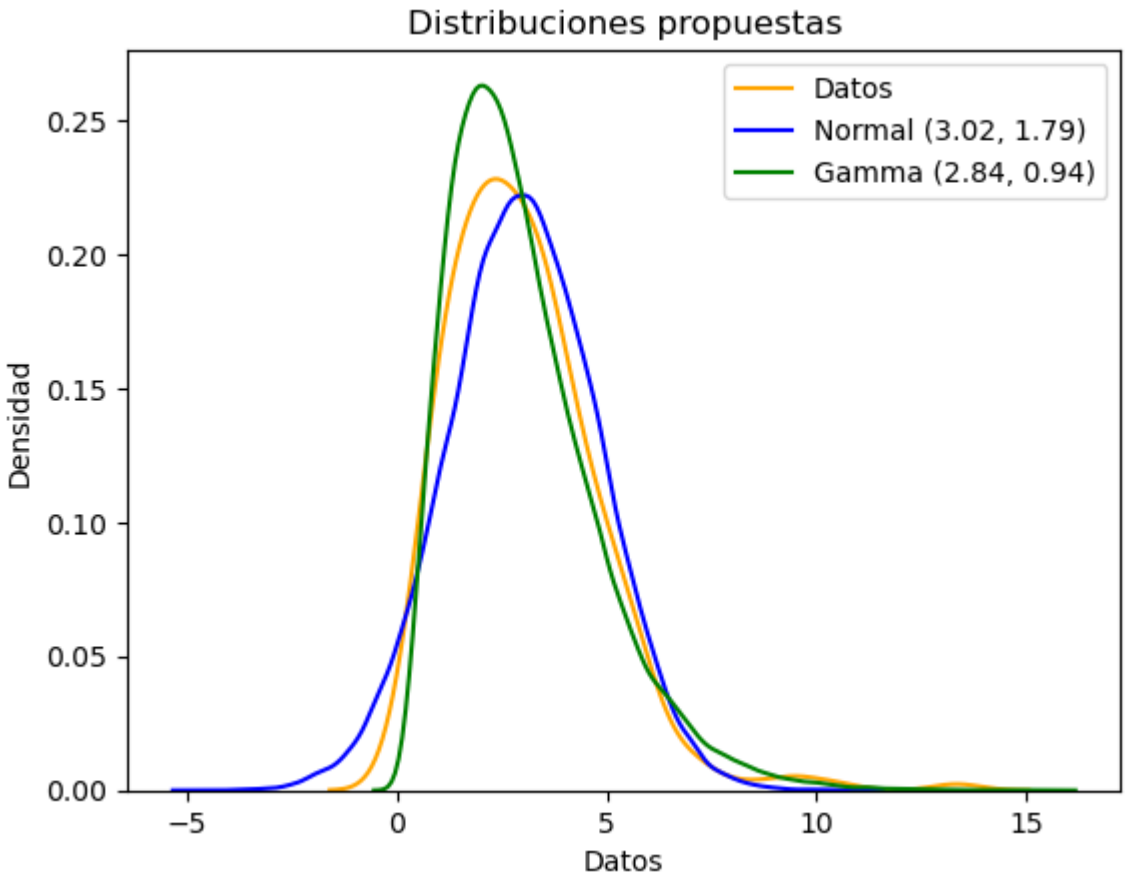
Para cada distribución propuesta, vamos a plantear H_0 y H_1

- **Distribución Normal**:
  - H_0: Los datos provienen de una distribución $Normal$ con media $\hat{\mu}$ y desviación estándar $\hat{\sigma}$.
  - H_1: Los datos no provienen de una distribución $Normal$ con media $\hat{\mu}$ y desviación estándar $\hat{\sigma}$.
- **Distribución Gamma**:

- H_0: Los datos provienen de una distribución $Gamma$ con parámetros $\hat{\alpha}$ y $\hat{\beta}$.
- H_1: Los datos no provienen de una distribución $Gamma$ con parámetros $\hat{\alpha}$ y $\hat{\beta}$.

2.4. Gráficos de Ajuste

Veamos gráficamente los modelos ajustados a los datos y los datos en sí en un mismo gráfico de densidad:



Podemos observar, que el análisis inicial siguiere que ambas distribuciones elegidas se ajustan bastante bien a los datos, con la excepción de que la $Normal$ **no** tiene en cuenta valores negativos, mientras que $Gamma$ y los datos sí, por lo que parece ser que la $Gamma$ es la más adecuada.

Para validar esta observación, desarrollaremos esta evaluación en la siguiente actividad.

Actividad 3

3.1. Consigna

Determinación de la calidad de los ajustes logrados.

a. Realizar una 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 $\Delta b f(x)$, donde Δb corresponde al ancho de intervalo en el histograma y $f(x)$ es cada una de las densidades propuestas.

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

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

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

3.2. Funciones de Densidad Acumulada

Para este punto se tiene en cuenta que la densidad acumulada de la Normal se obtiene con

$$\int_{\infty}^{x} N(0,1)(t)dt$$

la cual se calcula con la fórmula que vimos en el práctico, y para la densidad acumulada de la *Gamma*, tenemos que se puede calcular con *Monte Carlo*.

Notemos que la función de densidad de la *Gamma* es

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

Luego, la función de densidad acumulada está dada por:

$$
\begin{aligned}
F(x) &= \int_0^x f(t)dt \\
&= \int_0^x \frac{\beta^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\beta t} dt \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^x t^{\alpha-1} e^{-\beta t} dt
\end{aligned}
$$

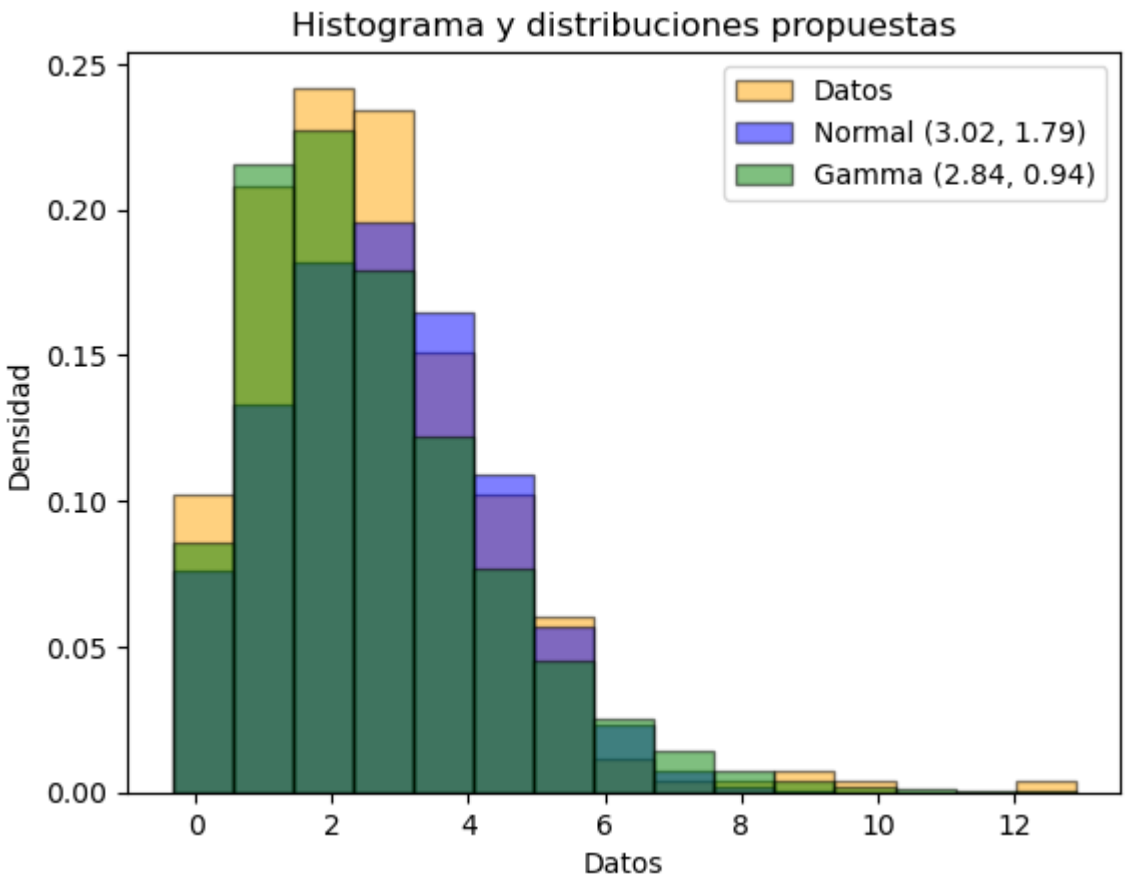
Ahora, también sabemos que para números reales, se cumple que

$$
\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt
$$

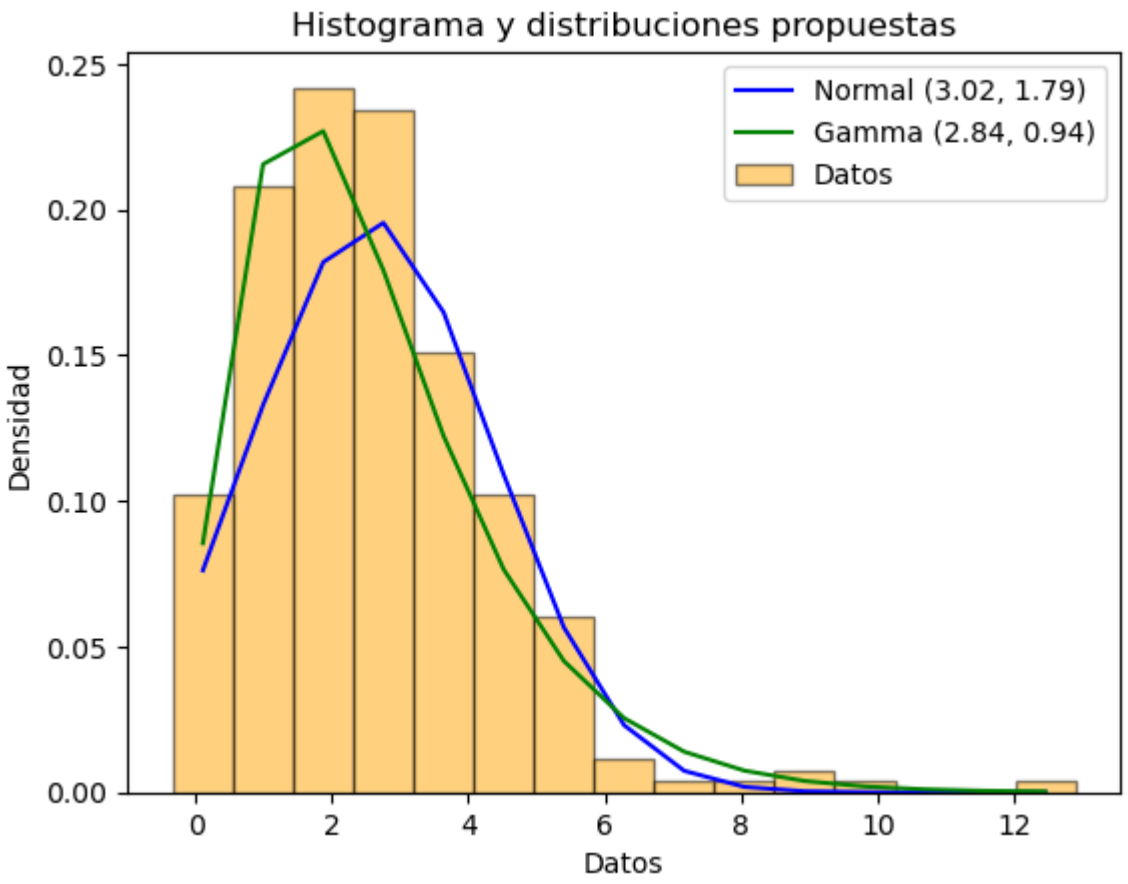
Por ello, podemos calcular las integrales con *Monte Carlo*. Sin embargo, si bien esto nos sirve para dar un pantallazo teórico, a la hora de realizar las simulaciones, convendrá **no** usar esta forma para estimar la acumulada de la función de $Gamma$ debido a la alta complejidad operacional. Por ello, hacemos uso de librerías externas (*scipy*) que están optimizadas para ello.

3.3. Item A | Comparación de Frecuencias

Si superponemos las barras de los histogramas con las funciones de densidad, obtenemos lo siguiente:



Pero, para ver mejor la representación de las distribuciones, podemos graficar a las mismas como una curvas continuas con sus gráficos de densidad:



Gracias a estos gráficos es que podemos ver que la $Gamma$ es la que mejor se ajusta a los datos, aunque solo tengamos en cuenta los valores positivos.

3.4. Item B | Prueba de Hipótesis con Ji-Cuadrada

Para este punto, si bien los datos que tenemos son continuos, podemos usar el test de χ^2 para ver si los datos provienen de una distribución $Normal$ o $Gamma$, haciendo una **discretización** de los mismos. Es decir, agrupando los datos en k intervalos consecutivos y considerando N_i como la cantidad observada en el i-ésimo intervalo, y p_i como la probabilidad teórica (dada por la distribución que estamos considerando) de que "caiga" en el intervalo.

Entonces, para calcular el estadístico usado en el test χ^2 utilizamos la siguiente fórmula:

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

donde, como dijimos, N_i es la frecuencia observada en el intervalo i-ésimo, p_i la probabilidad teórica, n el tamaño de la muestra y k la cantidad de intervalos que en este caso elegimos 21. Para ello, se eligieron uniformemente 20 valores en el rango $[0.1, 8]$ y se consideraron los extremos $(-\infty, 0.1]$ y $[8, \infty)$.

Luego, T tiene distribución χ^2 por lo que si t es el valor del estadístico calculado:

$$p - valor = P(T \geq t) = P(\chi^2_\alpha \geq t)$$

donde:

$$\alpha = k - 1 - m = 21 - 1 - 2 = 18$$

siendo m la cantidad de parámetros no especificados, que en el caso de la $Normal$ y la $Gamma$ son 2.

Finalmente, calculando los estadísticos y los p-valores correspondientes usando las fórmulas mencionadas tenemos:

Distribución	Estadístico	Grados de Libertad	P-Valor
$Normal$	60.506644	18	0.000002
$Gamma$	16.023399	18	0.590914

Si consideramos un nivel de confianza de 95%, entonces:

- En el caso de la $Normal$: se rechaza la hipótesis nula ya que $0.000002 < 0.05$. Por lo tanto, rechazamos la hipótesis de que los datos provienen de una distribución $Normal$.
- En el caso de la $Gamma$: **no** se rechaza la hipótesis nula ya que $0.590914 \not< 0.05$. Por lo tanto, no rechazamos la hipótesis de que los datos provienen de una distribución $Gamma$

Sin embargo, algo que tenemos que tener en cuenta es que el test de χ^2 no es muy bueno para distribuciones continuas, dado que no tiene en cuenta la distribución de los datos en cada intervalo, sino que solo la cantidad de datos en cada uno. Por ello, vamos a realizar el test de $Kolmogorov - Smirnov$ que nos va a brindar una mejor idea de si los datos provienen de una distribución $Normal$ o $Gamma$.

3.5. Item C | Prueba de Hipótesis con Kolmogorov-Smirnov

Para este punto, vamos a considerar la prueba de $Kolmogorov - Smirnov$ para ver si los datos provienen de una distribución $Normal$ o $Gamma$, dado que nuestros datos son *continuos*. Para ello, vamos a calcular el estadístico de $Kolmogorov - Smirnov$ y el p-valor asociado (con simulación).

Recordando un poco la teoría estudiada, como todos los valores de la muestra son distintos, vamos a tener que el estadístico tiene la siguiente forma:

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

donde si d es el valor del estadístico calculado:

$$p - valor = P(D \geq d)$$

Finalmente, calculando los estadísticos y los p-valores correspondientes usando las fórmulas mencionadas pero con sus respectivas modificaciones para simulaciones, tenemos:

Distribución	Estadístico	P-Valor
$Normal$	0.065913	0.006
$Gamma$	0.044331	0.175

Si consideramos un nivel de confianza de 95%, entonces:

- En el caso de la $Normal$: se rechaza la hipótesis nula ya que $0.006 < 0.05$. Por lo tanto, rechazamos la hipótesis de que los datos provienen de una distribución $Normal$.
- En el caso de la $Gamma$: **no** se rechaza la hipótesis nula ya que $0.175 \not< 0.05$. Por lo tanto, no rechazamos la hipótesis de que los datos provienen de una distribución $Gamma$

Una vez más, todo apunta a que la distribución que mejor se ajusta a los datos es la $Gamma$.

3.6. Item D | Conclusiones

La densidad de probabilidad que seleccionamos es la $Gamma$, dado que es la que mejor se ajusta a los datos que tenemos, ya que no se rechaza la hipótesis nula de que los datos provienen de una distribución $Gamma$.

Si bien en el gráfico de densidad de nuestros datos con muestras de las distribuciones $Normal$ y $Gamma$, parece que la $Normal$ se ajusta mejor. Esto no es así, ya que algo a considerar es que nuestros datos **no presentan** valores negativos.

Ahora, ¿esto es *tan* relevante? Sí, y para ello vamos a calcular la probabilidad de que se obtenga una muestra de 300 datos sin ningún valor negativo en el caso de una $Normal$:

→ Sea X una variable aleatoria con distribución normal y media $\hat{\mu}$ y desviación $\hat{\sigma}$:

$$\begin{aligned} P(X_1 > 0, .., X_{300} > 0) &= \prod_{i=1}^{300} P(X_i > 0) \\ &= \prod_{i=1}^{300} (1 - F(0)) \\ &= (1 - F(0))^{300} \\ &\sim 7.2303275090884 63 \times 10^{-7} \end{aligned}$$

Entonces, podemos ver que la probabilidad de que se obtenga una muestra de 300 sin ningún valor negativo en el caso de una $Normal$ es muy baja, por lo que este hecho es *muy* relevante y pesa mucho en la elección de la distribución $Gamma$.

Finalmente, debido al hecho de que la $Gamma$ respeta no tener valores negativos (y por los resultados obtenidos en los tests de hipótesis), elegimos la distribución $Gamma$ como la que mejor se ajusta a nuestros datos.

Sin embargo, una aclaración que queremos hacer es que todo este análisis considera que los datos muestrales *no están sesgados* (respecto a ser solo positivos), sino que son una muestra uniformemente obtenida de la distribución real.