Parcial computacional sobre Test de Hipótesis: Partículas de Carga Fraccionaria

Felipe Emiliano Perez - LU: 251/17

15 de julio de 2023

Resolución

1. Para empezar genero como dice el enunciado N vectores F de números con distribución de Poisson con esperanza $\mu = 0{,}1$ usando la librería `scipy` de Python. Luego coloco un 1 en todos los elementos de mi array que tengan un valor mayor a 1. Algo a notar que va a ser relevante para el siguiente punto es que dado que la esperanza de la Poissoniana que describe el comportamiento de la V.A. detección en cada detector es $\mu = 0{,}1$ quiere decir entonces que la taza de detección es justamente $\lambda = 0{,}1$.

2. En primer lugar unifico la estadística de las N mediciones para luego calcular la V.A. distancia entre eventos (ΔX) usando (función `diff()` de la librería `NumPy`). Luego realizo el histograma de esta nueva variable aleatoria ΔX (Fig. 1):

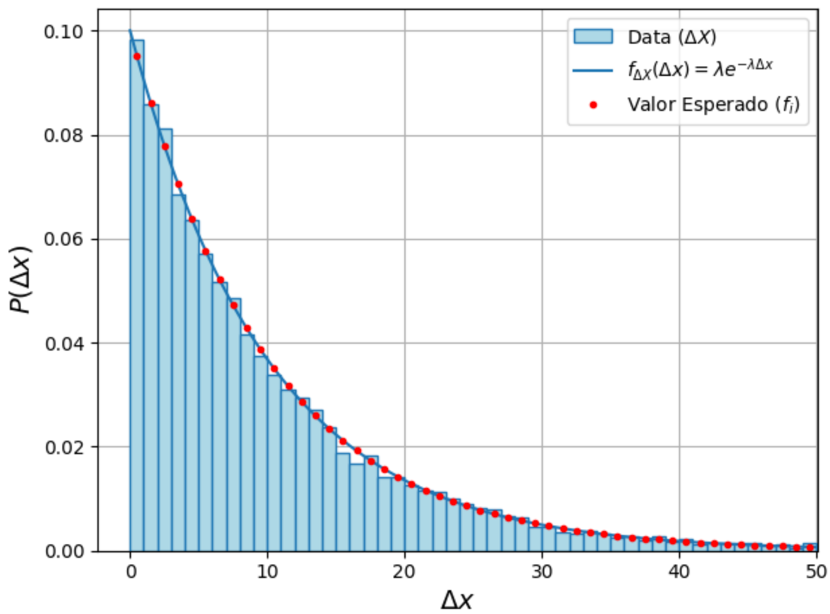

Figura 1: *Histograma normalizado de la variable aleatoria ΔX comparado con la distribución exponencial esperada, $f_{\Delta X}(\Delta x) = \lambda e^{-\lambda \Delta x}$, con $\lambda = 0{,}1$.*

Observando el histograma pareciera ser a simple vista que la distribución de esta nueva variable aleatoria coincide con la de una distribución exponencial. Haciendo un paralelismo con el problema 3 de la guía 3 (decaimiento radiactivo) que resolvió Muriel en clase es razonable intuir esto último, ya que la variable aleatoria entre eventos poissonianos sigue una distribución exponencial, en particular con el parámetro λ_F dado por el λ de la distribución de Poisson($\mu = 0{,}1$), siendo entonces $\lambda_F = 0{,}1$. Para estudiar la compatibilidad del histograma con la distribución exponencial se realiza entonces un test χ^2 tal y como vimos en la clase teórica de Ricardo, proponiendo el siguiente estadístico S:

$$\begin{aligned} S &= \sum_i^n \left(\frac{n_i - f_i}{f_i}\right)^2 \\ &= \sum_i^n \frac{(n_i - f_i)^2}{f_i} \end{aligned} \tag{1}$$

Donde n_i y f_i son el número de cuentas y el valor esperado en el bin i-ésimo. El valor de f_i es calculado a partir de la siguiente expresión, con n el número total de realizaciónes de ΔX:

$$f_i = n \int_{bin_i}^{bin_{i+1}} f_{\Delta X}(\Delta x)\,\mathrm{d}x \tag{2}$$

Dado que si H_0 es cierta, el número de cuentas en cada bin sigue una distribución de Poisson con esperanza f_i, entonces su varianza es también f_i y por lo tanto σ_i es el error en cada bin. Por lo tanto la ec. (1) del estadístico S, en la aproximación Poisson-Gaussiana ($\mu \gg 1$), dice que la distribución de esta variable aleatoria es una χ^2_{n-1} con $n-1$ grados de libertad ya que para obtener f_i usé n. En la Fig. (2) se muestra la distribución χ^2_{n-1} y la realización de S con los datos obtenidos en el item (1).

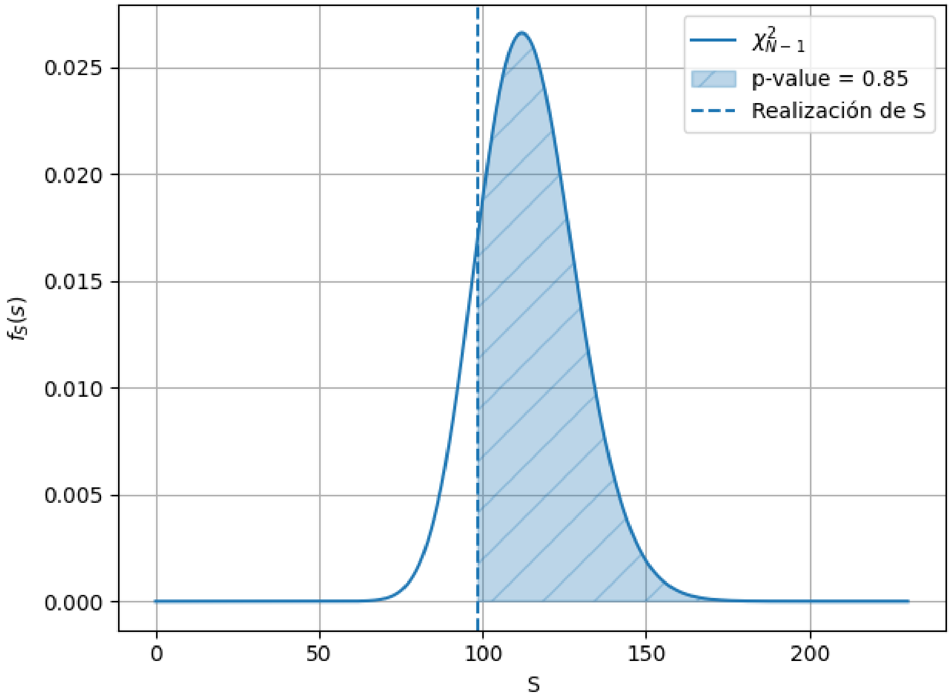

Figura 2: *Test χ^2 para estudiar la compatibilidad del histograma normalizado de ΔX y una distribución exponencial con parámetro $\lambda_F = 0{,}1$. p-value = 0.85.*

Se obtiene como resultado del test χ^2 un p-value $= 0{,}85$, en donde si se establece una significancia de $\alpha = 0{,}05$ quiere decir entonces que dada la realización de S obtenida se acepta entonces la hipótesis nula H_0.

3. En este caso para realizar el test Runs fue necesario crear una función para contar las rachas de dos conjuntos de datos y con esto poder obtener la distribución del estadístico $R(x_i, y_i)$ a partir de una simulación Monte Carlo. Asumiendo como cierta a la hipótesis nula H_0 se generan dos vectores con igual cantidad de elementos con distribución exponencial con $\lambda = 0{,}1$ para obtener el histograma de la Fig. (3) del estadístico R. Se muestra también el valor de R_{obs} obtenido a partir de los datos generados en el item (1).

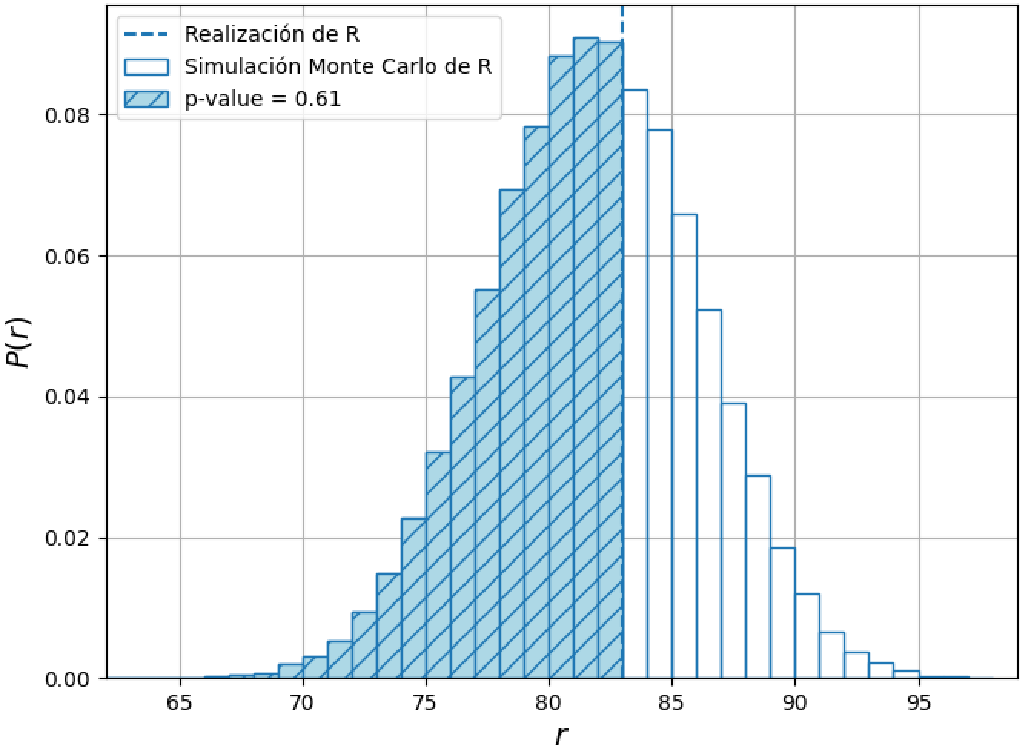

Figura 3: *Histograma del estadístico R del Test Runs. Se indica además la realización del estadístico R obtenido a partir de los datos generados en el item (1). p-value = 0.61*

Se obtiene entonces un p-value = 0.61, sumando las cuentas n_i obtenidas hasta el valor de R ontenido a partir de los datos (cola izquierda). Usando ahora el p-value del test runs y el hallado en el item (2) con el test χ^2, calculo el estadístico $\chi_4 = -2\ln(p_{\chi^2} \times p_{runs})$ obteniendo un nuevo p-value conjunto (Fig. 4).

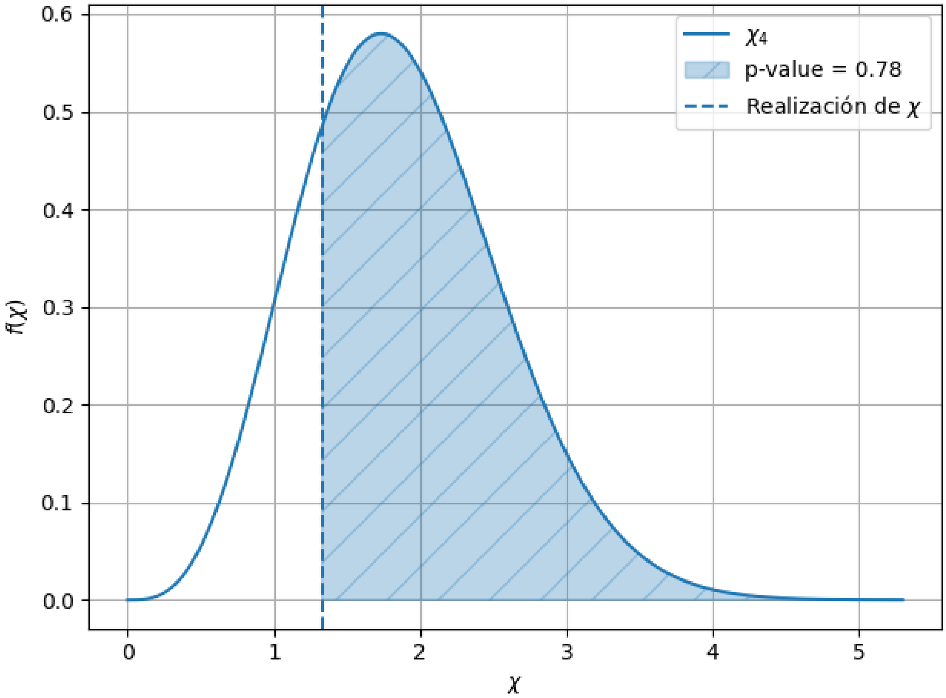

Figura 4: *Distribución del estadístico de p conjunto* $\chi_4 = -2\ln(p_{\chi^2} \times p_{runs})$*. Se indica el valor de* χ *observado. Se obtuvo un p-value=0.78*

El valor de p-value conjunto obtenido es 0.78, siendo este mayor al valor preestablecido para la significancia $\alpha = 0{,}05$, por lo tanto se acepta la hipótesis nula H_0.

4. En este caso al ser p mucho menor que n, puedo decir que es una buena aproximación que la detección de partículas en los 100 detectores sigue una distribución de Poisson con parámetro $\lambda_M = 0{,}02$. Siguiendo con la misma línea que en los items anteriores me genero entonces N vectores S con distribución de Poisson($\mu = 0{,}02$) y luego le sumo los vectores de detección de fondo F, reemplazando los valores mayores a uno por un 1 en su lugar. Al hacer esto, la distribución resultante debería ser una Poisson con $\mu = \mu_F + \mu_S$ dado que suma de Poissonianas es Poisson. Por lo tanto al calcular la variable aleatoria diferencia de distancias entre eventos, esta debería seguir una distribución exponencial con parámetro $\lambda_M = 0{,}12$ (Fig. 5).

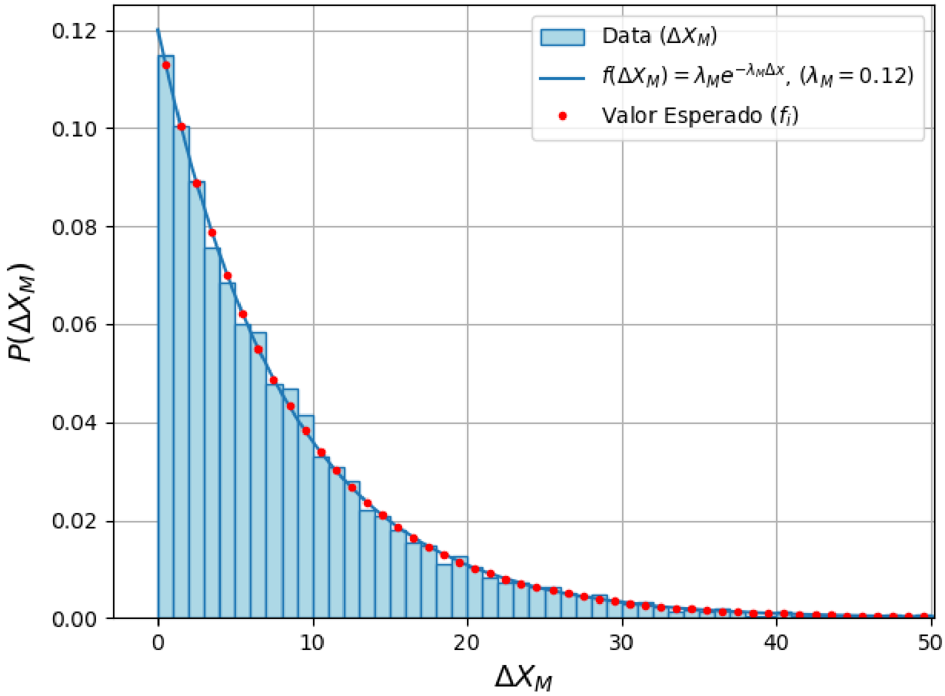

Figura 5: *Histograma normalizado de la variable aleatoria ΔX_M comparado con la distribución exponencial esperada, $f(\Delta X_M) = \lambda_M e^{-\lambda_M \Delta X_M}$, con $\lambda_M = 0{,}12$.*

Estudiando la compatibilidad con un test χ^2 se obtiene un p-value = 0.42, y por lo tanto dada una significancia de $\alpha = 0{,}05$ se acepta la hipótesis nula de que los datos provienen de una distribución exponencial de parámetro $\lambda_M = 0{,}12$.

5. Si la hipótesis H_0 es cierta entonces $\lambda = \lambda_F$, por lo tanto, para obtener el estadístico LLR puedo generar una tira de números con distribución exponencial de parámetro $\lambda_F = 0{,}1$ y hallar la distribución realizando una simulación Monte Carlo con el ratio entre las verosimilitudes de los datos cuando vale H_1 ($\lambda = \lambda_M = 0{,}12$) y cuando vale H_0 ($\lambda = \lambda_F = 0{,}1$). Para trabajar computacionalmente con el LLR abrí el logaritmo en la resta de los logaritmos de las verosimilitudes.

$$LLR(\Delta X) = -2\big(\ln L(\Delta X|H_1) - \ln L(\Delta X|H_0)\big) \tag{3}$$

Repitiendo este proceso obtuve el histograma de la Fig. (6) de la variable aleatoria LLR.

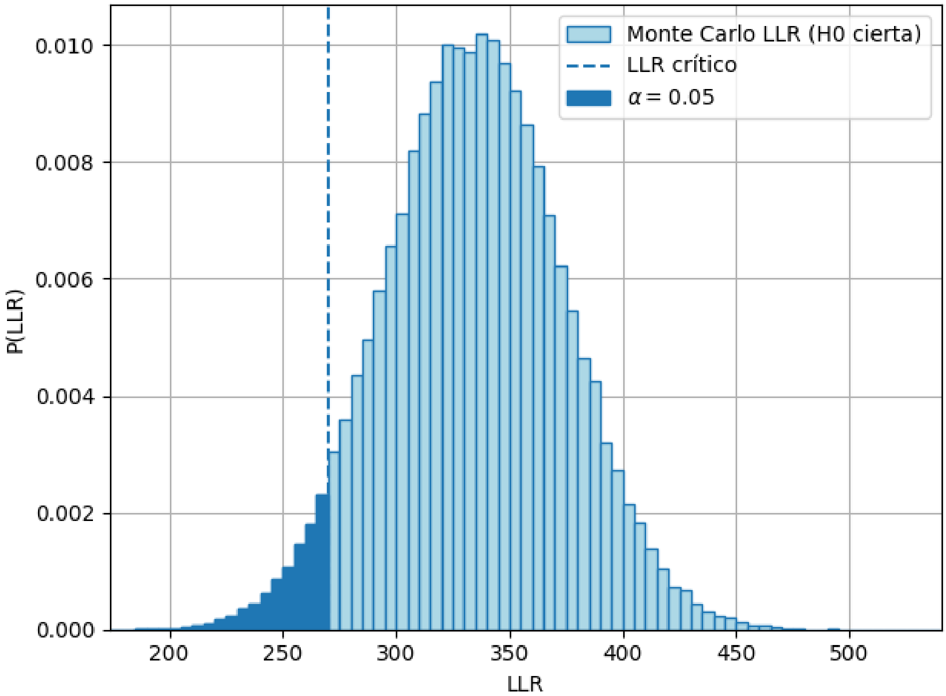

Figura 6: *Histograma normalizado de la variable aleatoria LLR cuando H_0 es cierta, obtenido por simulación Monte Carlo.*

6. Por último, para calcular la potencia es necesario primero obtener la distribución de LLR pero asumiendo como cierta ahora a la hipótesis alternativa H_1. Realizando el mismo proceso que en punto anterior pero ahora con $\lambda = 0{,}12$ obtengo el histograma del estadístico LLR cuando H_1 es cierta (Fig. 7).

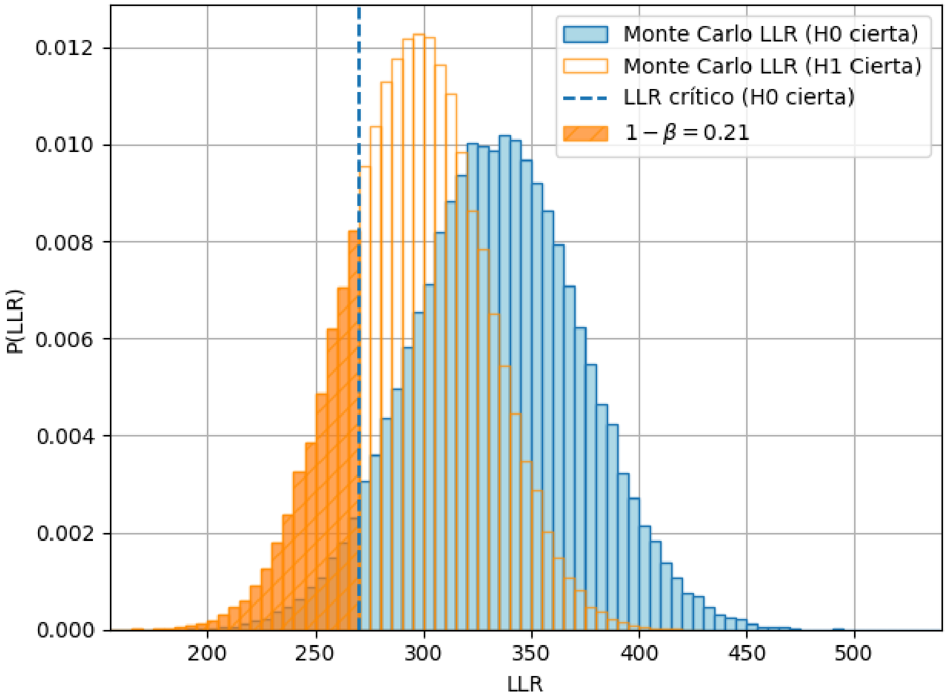

Figura 7: *Comparación de histogramas del estadístico LLR para los casos en donde vale H_0 (azul) o H_1 (naranja). La potencia del test obtenida es $1-\beta=0{,}21$.*

La potencia del test es $1-\beta = 0{,}21$, que se define según Frodesen como la probabilidad de rechazar la hipótesis nula cuando ésta es falsa que en este caso dió muy baja, o lo que es lo mismo decir que la probabilidad de aceptar la hipótesis nula H_0 cuando en realidad la hipótesis alternativa es cierta (Error Tipo II) es muy alta ($\beta = 0{,}79$). Este resultado me resulta raro dado que termino obteniendo una potencia bastante baja y se supone por lo que leí en el Frodesen que justamente este test es el que maximiza la potencia, es decir minimiza el error tipo II.