Universidad de Antioquia
Facultad de Ciencias Exactas y Naturales
Área: Física Computacional
Docente: José David Ruiz Álvarez

Proyecto Análisis de Datos

Alexander Valencia[a], Daniel Henao[a]

[a] *Instituto de Física.*

Resumen

En términos generales, una distribución de probabilidad sirve para describir el comportamiento de una variable aleatoria. Se define como una función que asigna una probabilidad sobre la variable aleatoria, para que determinado acontecimiento ocurra. Identificando el tipo de distribución que tiene un conjunto de datos, es posible aplicar pruebas estadísticas para verificar la relación entre dos variables en situaciones hipotéticas y reales. En el presente trabajo se utilizó un conjunto de diez mil datos que representan valores de tiempo t y medidas de x y de y, como resultado de algún experimento en dos aparatos de medida. Se implementó un método basado en la calidad de un ajuste usando el estadístico χ^2 para funciones del tipo $x(t)$, $y(t)$, y también se realizó un análisis cualitativo sobre la relación entre las medidas x, y.

Palabras claves: Ajuste, Chi Cuadrado, Correlación.

I. INTRODUCCIÓN.

Hay diferentes aplicaciones asociadas con una distribución χ^2, una de ellas es la prueba de bondad de ajuste. Esta aplicación consiste en determinar si un conjunto de datos se ajusta a una distribución teórica previamente establecida, tal como una distribución normal, o a una distribución empírica que haya sido obtenida a partir de los datos de una muestra para el análisis de una población.

Una de las ventajas que tiene la prueba de bondad de ajuste es que puede ser utilizada para el análisis de datos en distribuciones discretas de variables aleatorias que tienen valores contables o finitos, y en distribuciones continuas de variables aleatorias con un conjunto de valores posibles infinito, es decir, que no se pueden contar.

Además, la prueba indica si la medida que hay entre las diferencias existentes de la distribución observada y la distribución teórica del ajuste se debe al azar, contrastando las hipótesis planteadas.

El análisis de datos es una herramienta muy útil porque en áreas como la física, permite obtener modelos aproximados de diferentes fenómenos físicos y dan paso a la descripción de estos mediante el comportamiento de las variables en forma lineal, y en casos específicos, a través del comportamiento de las variables en forma no lineal.

II. MARCO TEÓRICO.

La fórmula que define un modelo estadístico para la bondad de ajuste usando χ^2 es simplemente:

$$\chi^2 = \sum_i \frac{(O_i - T_i)^2}{T_i}$$

donde cada O_i es el valor observado de una magnitud, cada T_i es el valor teórico correspondiente (o valor esperado) y la suma se aplica sobre todo el conjunto de datos.

Para una prueba de bondad de ajuste, se define una hipótesis nula H_0, donde se asume que no hay igualdad entre las distribuciones de los valores observados y los valores teóricos. Los dos criterios para concluir la prueba son:

- A medida que el valor de χ^2 disminuye, esto es si $\chi^2 \to 0$, entonces hay un ajuste más óptimo entre ambas distribuciones.
- A medida que χ^2 aumenta, un valor mayor del estadístico implica que la hipótesis nula es viable y no se tiene una buena calidad del ajuste entre las distribuciones.

En el caso del análisis asignado, se quiere encontrar la función con las mejores condiciones para describir el comportamiento de las variables x, y, en términos del tiempo t.

Partiendo de la observación gráfica del conjunto de datos, se proponen dos formas funcionales que modelen el valor esperado y que representen a cada par de elementos $f(t)$ (f dada como x o como y según sea el caso).

En principio, se extraerán múltiples parámetros libres del ajuste χ^2 perteneciente a cada forma funcional y dentro de determinado rango de valores. A partir de los mismos, se minimizará el valor de χ^2 para elegir la mejor calidad del ajuste. Las dos hipótesis alternativas son:

- $H_1 = f_1(t)$ (la hipótesis 1 es la primera forma funcional)
- $H_2 = f_2(t)$ (la hipótesis 2 es la segunda forma funcional)

De este modo, los únicos resultados posibles corresponden a estos casos:

(i) $\quad \chi^2 \in H_1 < \chi^2 \in H_2$
(ii) $\quad \chi^2 \in H_2 < \chi^2 \in H_1$

excluyendo el caso hipotético en que $\chi^2 = 0$.

Y conforme a los criterios para la prueba de bondad de ajuste, si se cumple el caso ***(i)***, entonces se acepta la hipótesis alternativa H_1 y la primera forma funcional $f_1(t)$ es la que describe apropiadamente el comportamiento de los datos analizados, de lo contrario, en el caso ***(ii)*** se rechaza H_1 y se acepta H_2 para afirmar que la segunda forma funcional $f_2(t)$ corresponde a la del mejor ajuste.

III. RESULTADOS Y ANÁLISIS

Descripción del conjunto de datos

Para las dos variables en función del tiempo se presenta la superposición lineal de dos movimientos armónicos, representados por sus respectivas oscilaciones.

Cada una de las oscilaciones se caracteriza por tener una amplitud propia, determinada por un valor que inicialmente podría fijarse durante la evolución temporal, localizada en la región de influencia del conjunto de datos.

Pruebas de bondad del ajuste

Para aplicar cada prueba se eligió un rango de valores para las amplitudes y las frecuencias. Luego se combinaron para obtener la cantidad específica que optimizó el valor mínimo de χ^2 y así escoger el mejor ajuste entre las dos funciones.

(1) Variable x. La primera forma funcional que se propone se define mediante la superposición lineal de dos funciones coseno, dada por la expresión general:

$$x_1(t) = -A\cos(\omega_A t) + B\cos(\omega_B t)$$

Los parámetros de interés en el análisis son las amplitudes A, B, y las frecuencias de oscilación por unidad de tiempo ω_A y ω_B.

La segunda forma funcional consiste en la superposición lineal de dos funciones elípticas de Jacobi, designadas con el símbolo cn y que se pueden definir a partir de la integral

$$u = \int_0^{\varphi} \frac{d\theta}{\sqrt{1 - m\sin^2\theta}}$$

como $cn\, u = \cos\varphi$. Dado que para $m = 0$ esta función equivale al coseno, el valor usado en el ajuste ha sido $m = 0.5$, admitiendo un análisis de datos y un procedimiento con una forma funcional distinta.

Se ha usado un paquete de Scipy para calcular los valores de cada función en t. Estas funciones de Jacobi aparecen, por ejemplo, cuando se estudian sistemas físicos como el péndulo a grandes amplitudes o la peonza asimétrica. La forma propuesta es:

$$x_2(t) = -Ccn(\omega_C t, 0.5) + Dcn(\omega_D t, 0.5)$$

El resultado obtenido, en ambas formas funcionales, se resume en la Tabla 1 junto con las Figuras 1 y 2.

$x_1(t)$				$x_2(t)$			
A	1.3	ω_A	3.0	C	1.35	ω_C	3.55
B	1.8	ω_B	50.0	D	1.9	ω_D	59.0
$\chi^2 = 1271.7$				$\chi^2 = 1524.7$			

Tabla 1. Parámetros encontrados para los mejores ajustes de x vs t.

(2) Variable y. Las funciones ahora propuestas son:

$$y_1(t) = -Ecos\,(\omega_E t) - Fcos\,(\omega_F t)$$

$$y_2(t) = -Gcn(\omega_G t, 0.5) - Hcn(\omega_H t, 0.5)$$

Los parámetros que corresponden a los menores valores de χ^2 para ambas funciones se muestran en la Tabla 2. Los ajustes hechos con estos parámetros se encuentran en las Figuras 3 y 4.

$y_1(t)$				$y_2(t)$			
E	1.3	ω_E	3.0	G	1.25	ω_G	3.55
F	1.9	ω_F	50.0	H	1.9	ω_H	59.0
$\chi^2 = 1163.2$				$\chi^2 = 1385.6$			

Tabla 2. Parámetros encontrados para los mejores ajustes de y vs t.

De acuerdo con los valores de χ^2, para las dos variables se cumple el primer caso descrito para las hipótesis alternativas, aceptando así las funciones dadas por H_1 como las del mejor ajuste, las cuales se resumen así:

$$x(t) = -1.3\,cos\,(3t) + 1.8\,cos\,(50t)$$

$$y(t) = -1.3\,cos\,(3t) - 1.9\,cos\,(50t)$$

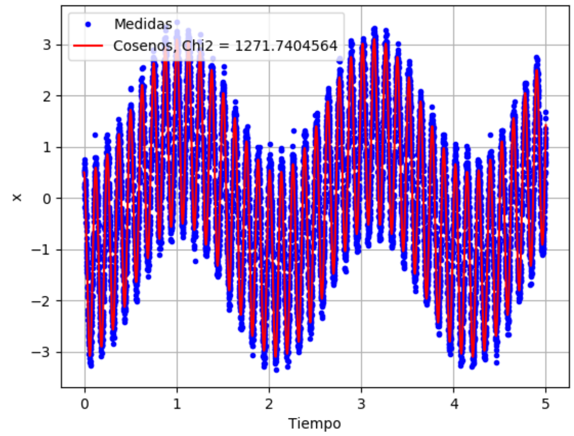

Fig 1. Ajuste mediante la función $x_1(t)$.

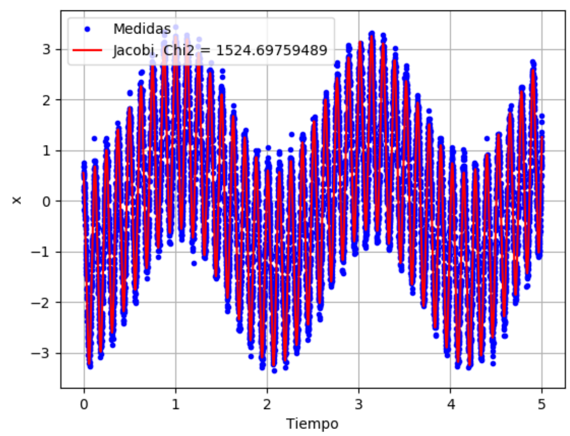

Fig 2. Ajuste mediante la función $x_2(t)$.

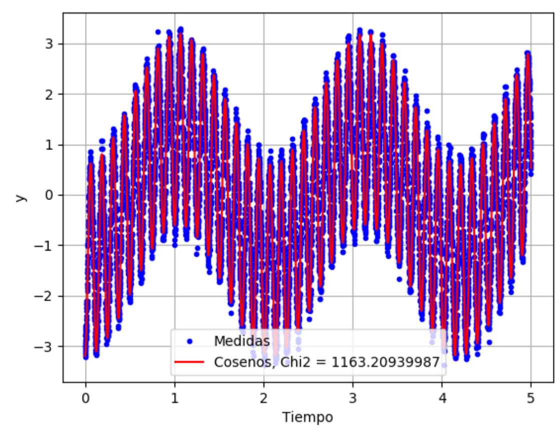

Fig 3. Ajuste mediante la función $y_1(t)$.

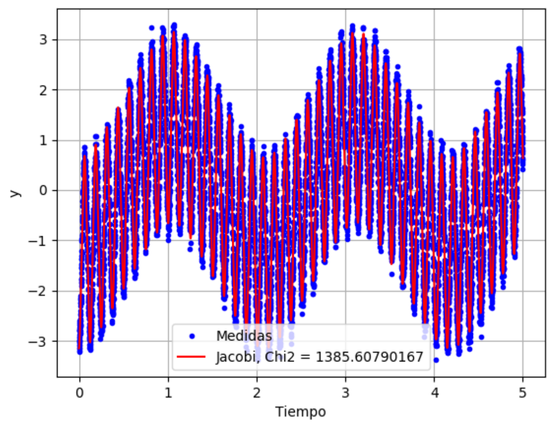

Fig 4. Ajuste mediante la función $y_2(t)$.

Correlación entre x y y

Puede verse de las formas funcionales de $x(t)$ y $y(t)$ que prácticamente la única diferencia entre ambas es el signo de uno de los cosenos, lo que aparentemente sugiere una relación clara entre ambas variables, sin embargo se consideró que este no es el caso. Una de las razones es que en la gráfica de y vs x (Figura 5), no hay regiones particulares donde los puntos se acumulen, sino que parecen distribuirse uniformemente al interior de una zona con la forma de un paralelogramo, por lo que no hay ningún patrón ni ninguna sub-zona de preferencia, que serían distintivos de una función.

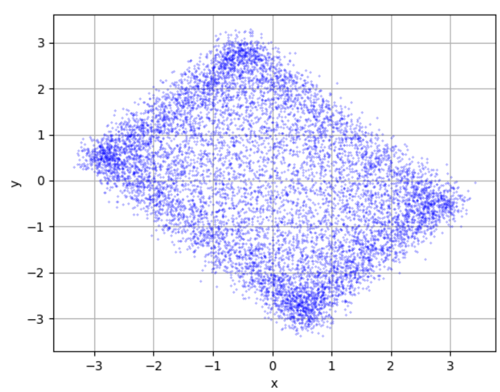

Fig 5. Gráfica de y vs x.

La otra razón es que no se encontró la manera de expresar una variable como función de la otra a partir de las formas funcionales de $x(t)$ y $y(t)$ sin prescindir de t, lo que indicaría simplemente que no se puede expresar una función $x(y)$ analíticamente.

IV. CONCLUSIONES

- Aunque no se tiene certeza del fenómeno físico particular que describen los datos, las formas funcionales del mejor ajuste confirman que se trata de un movimiento armónico complejo, tanto para $x(t)$ como para $y(t)$. La diferencia es que las curvas se encuentran desfasadas una respecto a la otra.

- Por la naturaleza armónica de las funciones que describen el conjunto de datos, no es preciso aplicar la prueba del coeficiente de correlación de Pearson para determinar una relación entre las variables x y y. Un análisis de correlación no lineal, usando el mismo método de χ^2 para calcular el coeficiente de correlación, requiere la transformación de funciones, que consecuentemente sean intrínsecamente lineales, condición que no se satisface en este caso.

REFERENCIAS.

[1] MEDWAVE, Revista Biomédica (2020). *The chi-square*, [en línea]. Disponible en: https://www.medwave.cl/link.cgi/Medwave/Series/MBE04/5266. [2020, 06 de junio].

[2] UNIVERSIDAD DE LA REPÚBLICA, URUGUAY, (2020). Distribuciones de Probabilidad, [en línea]. Disponible en: http://meteo.fisica.edu.uy/Materias/Analisis_Estadistico_de_Datos_Climaticos/teorico_AEDC/Distribuciones_Probabilidad_2011.pdf. [2020, 06 de junio].

[3] PROBABILIDADES Y ESTADÍSTICA, (2020). *Prueba de Bondad de Ajuste Chi Cuadrado*, [en línea]. Disponible en: https://www.probabilidadesyestadistica.com/prueba-de-bondad-de-ajuste-chi-cuadrado/. [2020, 06 de junio].