Análisis de simulación de trayectorias: Caminanta Aleatoria como modelo ideal del movimiento Browniano

Ángela Morales Zamudio
Claudeth Clarissa Hernández Álvarez

Mayo 2019

Abstract

Simulando el problema del caminante borracho o random walk, se hará un análisis estadístico de 100 partículas, trazando trayectorias de 10,000 pasos cada una. El camino trazado por cada partícula fue determinado por un programa generador de números aleatorios, creando así una primera aproximación a un modelo que describa el movimiento Browniano de una partícula.

1 Introducción

El término random walk fue propuesto por Karl Pearson en 1905. Pearson pretendía conocer la distribución de posiciones de ciertos mosquitos luego de cierto tiempo. El ejemplo más habitual de caminatas aleatorias que se presenta en la actualidad es el problema del caminante o problema del borracho. Este es, en sí, uno de los casos más sencillos de la modelación de trayectorias de partículas. En este proyecto se trabajará con un conjunto de partículas (las trayectorias están dadas por un generador de números aleatorios). En las posiciones de los movimientos caóticos de las partículas en la naturaleza, éstas suelen estár sometidas a campos de fuerzas que cambian la probabilidad de las direcciones tomadas. Por lo que se comprobará si el modelo se ajusta, igual, a una distribución Gaussina al cambiar las probabilidades de tomar unas direcciones sobre otras, para así verificar qué tan bueno es para modelar este tipo de movimientos.

2 Marco Teórico

2.1 Caminatas Aleatorias (Random Walks)

Una caminata aleatoria se define como la trayectoria de una partícula en la que su posición en cierto momento sólo depende de su posición en algún instante

previo y alguna variable aleatoria que determina su subsecuente dirección y la longitud de paso. En otras palabras, la probabilidad de todos los pasos posibles estén igualmente distribuídos.
Sean X_k un conjunto de variables aleatorias idénticamente distribuídas (Distribución Uniforme, donde todos los datos tienen la misma probabilidad de ocurrir). La caminata aleaotria S_n se define como:

$$S_n = X_1 + X_2 + ... + X_n = \sum_{i=0}^{n} X_i$$

Donde el valor medio y la varianza de una caminata aleatoria están dados por:

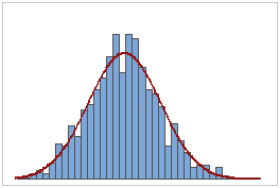
$$\bar{S_n} = \sum_{i=0}^{n} \bar{X_i} \tag{1}$$

$$var(S_n) = \sum_{i}^{n} var(X_i) + 2 \sum_{i=i}^{n} \sum_{j \neq i}^{n} cov(X_i, X_j) \tag{2}$$

Esto indica que el comportamiento de los valores medios de las caminatas aleatorias depende de las relaciones entre sus variables aleatorias. Es decir, el comportamiento de la trayectoria va a depender únicamente de cómo se realicen los pasos.
El caso más sencillo de caminata aleatoria (llamada caminata libre) ocurre cuando se tiene la misma probabilidad de avanzar que de retroceder (en cualquier eje). Esto hace que la posición del caminante también tenga valor medio nulo en cada instante.
Desde un punto de vista un poco menos formal, se está diciendo que al medir, uno observa aproximadamente la misma cantidad de pasos en una dirección y en la opuesta.
El ejemplo clásico de este tipo de movimiento es el que observó R. Brown, para una partícula de polen sobre agua [Brown, 1829]. En este caso el movimiento errático de la partícula de polen se debe a los choques que recibe ésta por las partículas del medio que la rodea.

2.2 Teorema del Límite Central

El Teorema del Límite Central dice que la distribución de probabilidad que gobierna $\bar{Y_n}$ se aproxima a una distribución Normal conforme $n \to \infty$, sin importar la distribución que gobierna a las variables aleatorias en cuestión Y_i. Más aún, la media de $\bar{Y_n}$ se aproxima a μ y la desviación estándar a $\frac{\sigma}{\sqrt{n}}$.



3 Metodología

Las simulaciones proporcionadas se basaron en un Generador de Números Aleatorios para cada partícula para determinar las parejas (X,Y) de pasos de cada una. Es por ello que se puede decir que, con sencillas simulaciones, se pueden obtener una gran cantidad de datos o archivos de datos de manera aleatoria, cada uno siguiendo su propio "dado de probabilidad" que describe su movimiento.

Para el análisis de datos se utilizó el lenguaje de programación Python, empleando las bibliotecas necesarias, entre ellas *scipy.stats*, para poder visualizar gráficas y aplicar pruebas a un conjunto de datos, los cuales fueron generados por medio de simulaciones en $ForTran$ bajo condiciones de aleatoriedad. Gracias a este programa generador de números aleatorios, se logró obtener un conjunto de datos con posiciones X y Y para cada partícula.
Se trabajaron los siguientes conjuntos de datos:

- **Primer conjunto de datos:**
  En X: La probabilidad de dar un paso a la derecha es de 0.50.
  En Y: La probabilidad de dar un paso hacia arriba es de 0.50.
- **Segundo conjunto de datos:**
  En X: La probabilidad de dar un paso a la derecha es de 0.55.
  En Y: La probabilidad de dar un paso hacia arriba es de 0.55.
- **Tercer conjunto de datos:**
  En X: La probabilidad de dar un paso a la derecha es de 0.60.
  En Y: La probabilidad de dar un paso hacia arriba es de 0.50.
- **Cuarto conjunto de datos:**
  En X: La probabilidad de dar un paso a la derecha es de 0.60.
  En Y: La probabilidad de dar un paso hacia arriba es de 0.60.

A cada conjunto de datos, haciendo uso de un análisis estadístico gracias a la bibliotéca *Scipy.stats*, se le realizó:

- Un diagrama de caja para el conjunto de datos, calculando los cuartiles con la función *describe()*.
- Un Histograma para la posición en X y otro para la posición en Y .
- Una prueba de hipótesis Shapiro-Wilks, partiendo de que se desconoce el tipo de distribución de los datos y que las observaciones en cada muestra son independientes e idénticamente distribuídas.
  Para hacer nuestra prueba de normalidad, contamos con una hipótesis nula, H_0, y una hipótesis alternativa, H_1.
  Donde:
  H_0: La muestra sigue una distribución Gaussiana.
  H_1: La muestra no sigue una distribución Gaussiana.

Si resulta ser que el p-valor es mayor al nivel de significancia, $\alpha = 0.05$ entonces se dirá que la muestra parece seguir una distribución normal (No se rechaza H_0). Por otra parte, si el p-valor es menor al nivel de significancia entonces hay evidencia de que la muestra parece no seguir una distribución normal (Se rechaza H_0).

4 Resultados

4.1 Análisis del Generador de Números Aleatorios

Se analizó la manera en que se decició la probabilidad de tomar un paso a la derecha o izquierda (en las posiciones en X), o un paso hacia arriba o abajo (en las posiciones en Y) que se utilizó. Los archivos utilizados para ello contienen dos columnas, cada una con una secuencia de números aleatorios con distribución uniforme. Debido a que la distribución es uniforme, cada vez que una partícula dio un paso, se llamó a un número aleatorio NAX para su movimiento en X, y otro número aleatorio NAY para su movimiento en Y.
Analizando tres generadores, lo que se hizo fue graficar cada pareja de puntos aleatorios para poder observar si dicha distribución era uniforme:

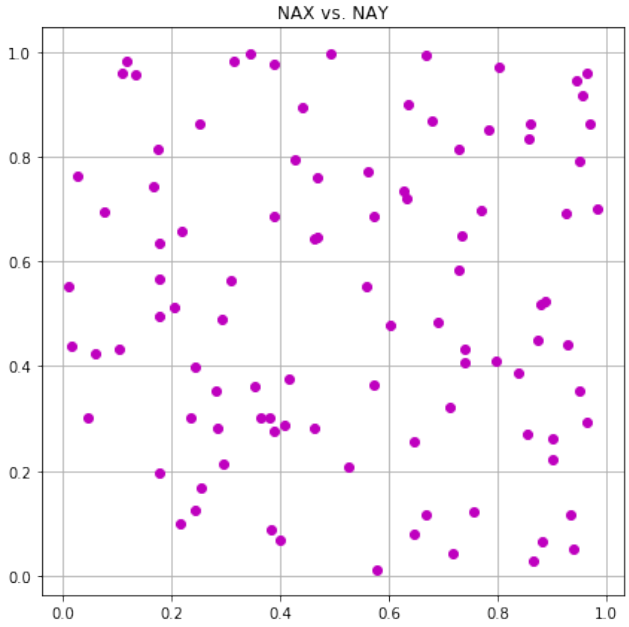


figure:Distribución de 100 puntos aleatorios

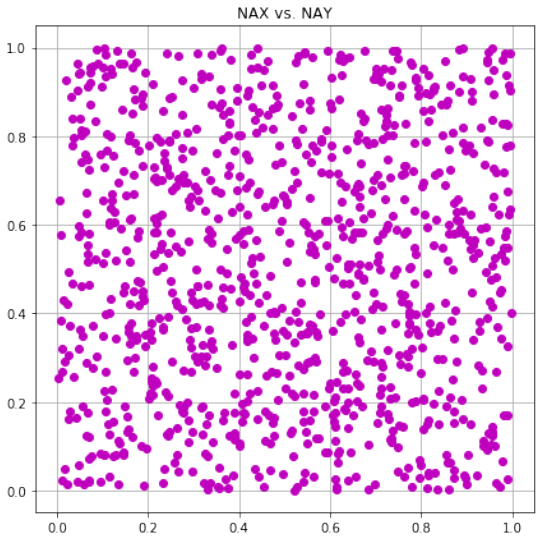


figure:Distribución de 1,000 puntos aleatorios

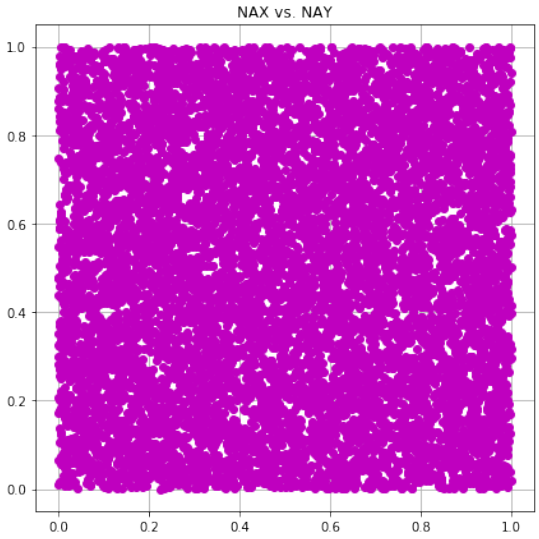


figure:Distribución de 10,00 puntos aleatorios

Gráficamente puede verse que dichos números tienen una distribución uniforme, ya que tratan de llenar todo el espacio de la gráfica de manera equitativa. En los casos analizados, se utilizó un generador de números aleatorios como estos pero cada caso con su respectiva probabilidad de movimiento en las direcciones "X" y "Y".

4.2 Primer conjunto de datos

Recordemos que para esta tabla de valores, en X, la probabilidad de dar un paso a la derecha es de 0.50, y en Y, la probabilidad de dar un paso hacia arriba es de 0.50.

4.2.1 Diagramas de Caja (Box-plot)

De manera gráfica, pueden observarse estos datos en diagramas de cajas, donde pueden visualizarse dónde están los cuartiles, la media y los datos extremos.

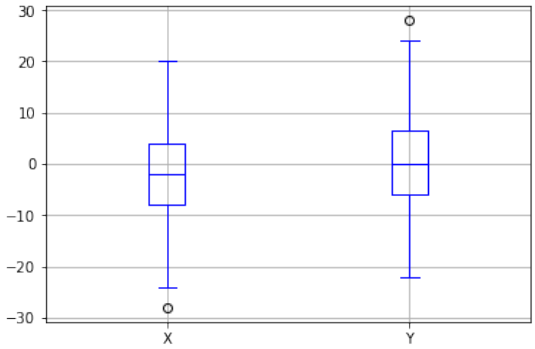


figure: Boxplots de X y Y del paso 100

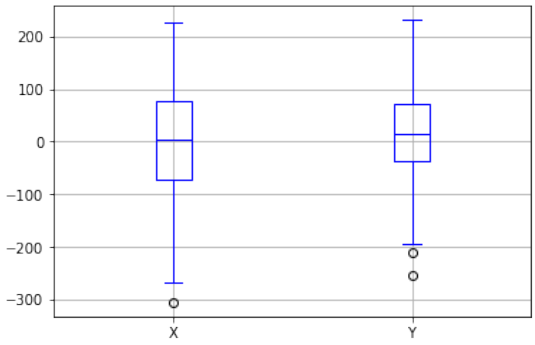
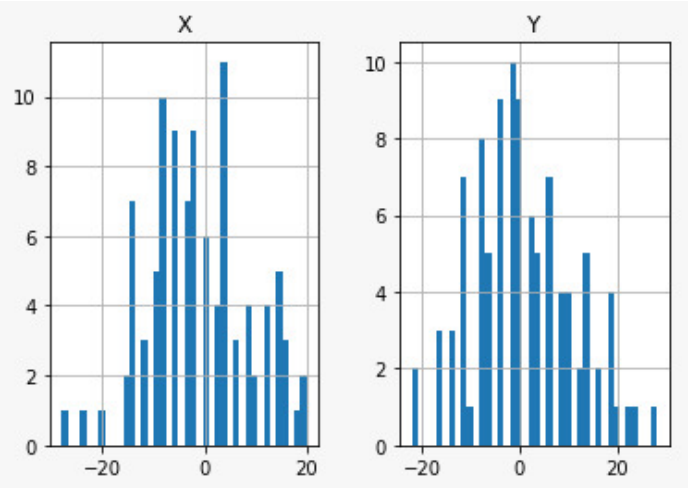


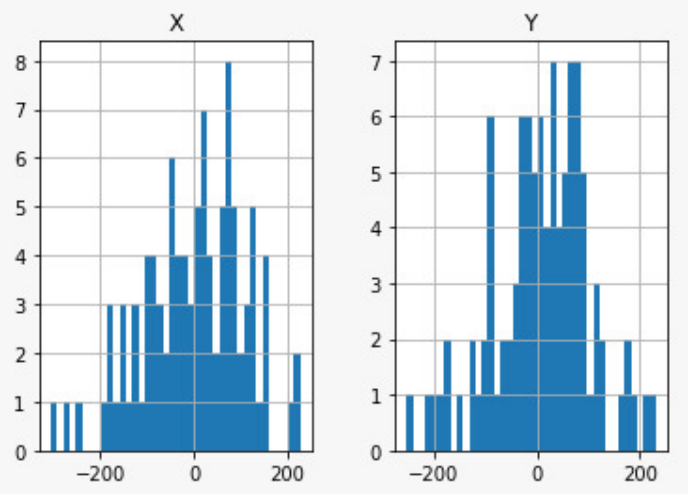
figure: Boxplots de X y Y del paso 10,000

4.2.2 Histogramas

Habiendo recolectado en dos DataFrames el paso número $10,000$ y 100 de cada una de las cien partículas, se generaron histogramas para las posiciones en X y en Y, obteniendo los siguientes diagramas:



Histogramas de las posiciones en X y en Y del paso número 100 de las 100 partículas



Histogramas de las posiciones en X y en Y del último paso de las 100 partículas

A simple vista se puede observar que parecen tener una distribución Gaussiana (Normal).

4.2.3 Prueba Shapiro-Wilk

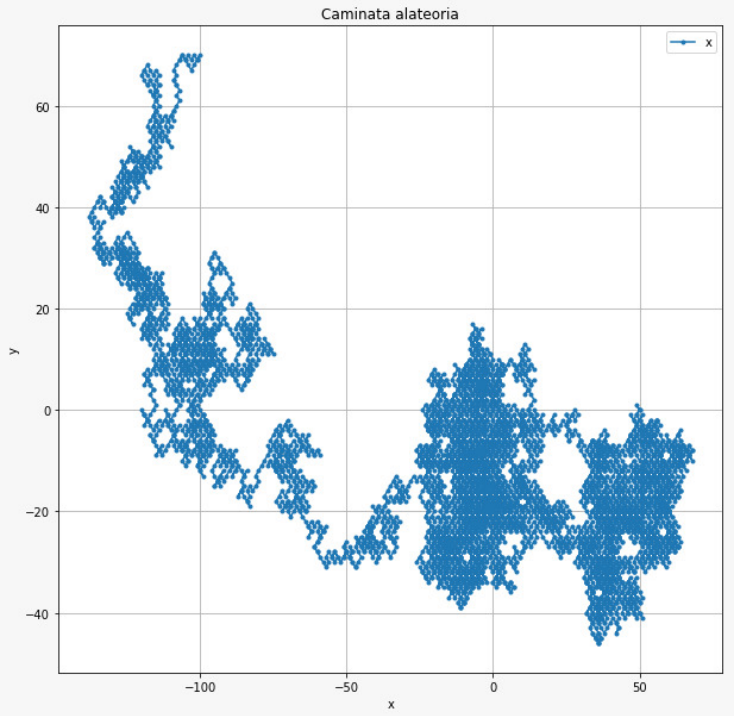
Usando una prueba de Hipótesis *Shapiro-Wilk*, dado que no se conoce en primera instancia la distribución de estos datos, se probará si siguen dicha distribución.

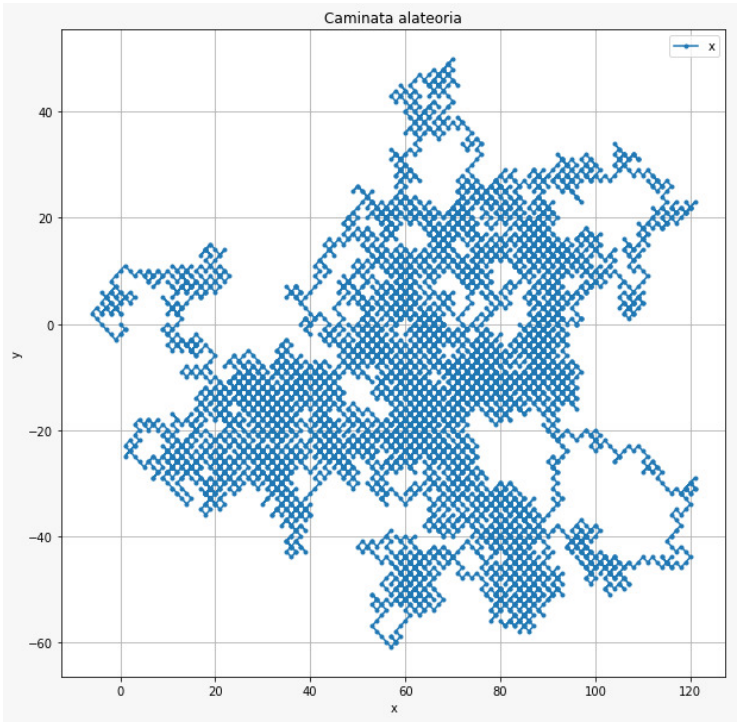
- H_0 = La muestra sigue una Distribución Gaussiana.
- H_1 = La muestra no sigue una Distribución Gaussiana.

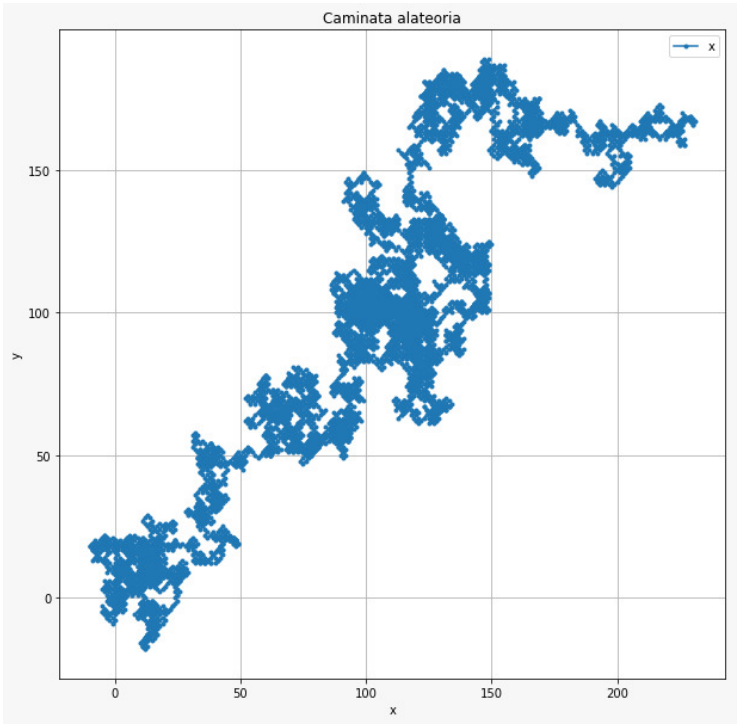
En la tabla que recopila el paso número 10,000 de cada partícula el p-valor obtenido fue de $p = 0.077$.
En la tabla que recopila el paso número 100 de cada partícula el p-valor obtenido fue de $p = 0.108$.
En ambos casos el p-valor resultó ser mayor al nivel de significancia, $\alpha = 0.05$.
En consecuencia no tenemos evidencia para no rechazar H_0. Por lo que pareciera que la muestra sigue una distribución normal.

4.2.4 Trayectorias

Se tomaron tres partículas al azar (de las 100 que teníamos registradas) y se graficaron sus trayectorias (posiciones X,Y).




Caminata alateoria
x
40
20
0
−20
−40
−60
y
0
20
40
60
80
100
120
x


Caminata alateoria
x
150
100
50
0
y
0
50
100
150
200
x

4.3 Segundo conjunto de datos

Recordemos que para esta tabla de valores, en X, la probabilidad de dar un paso a la derecha es de 0.55, y en Y, la probabilidad de dar un paso hacia arriba es de 0.55.

4.3.1 Diagramas de Caja (Box-plot)

De manera gráfica, pueden observarse estos datos en diagramas de cajas, donde pueden visualizarse dónde están los cuartiles, la media y los datos extremos.

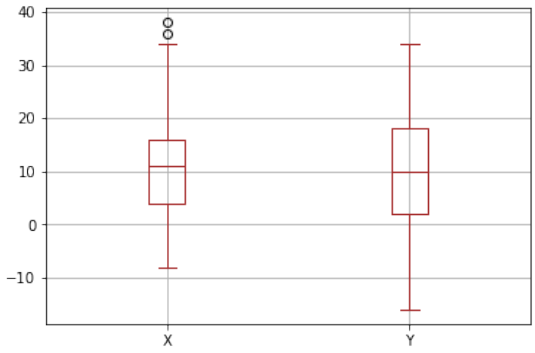


figure: Boxplots de X y Y del paso 100

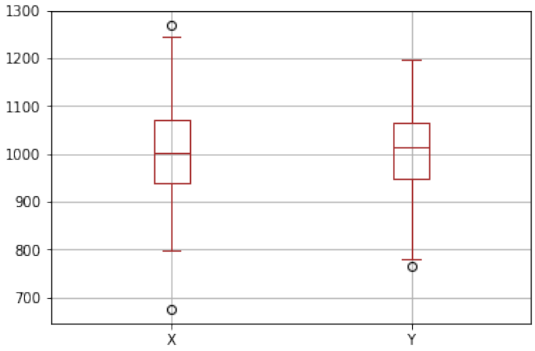
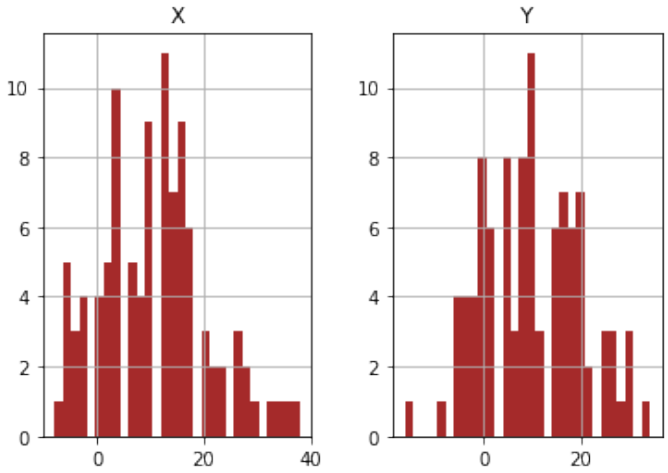


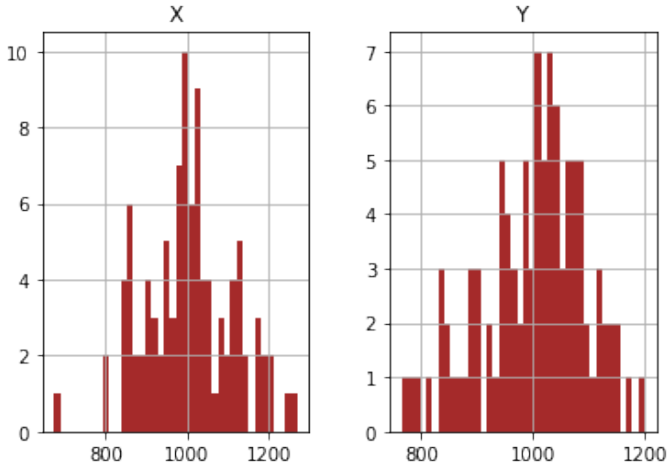
figure: Boxplots de X y Y del paso 10,000

4.3.2 Histogramas

Habiendo recolectado en dos DataFrames el paso número $10,000$ y 100 de cada una de las cien partículas, se generaron histogramas para las posiciones en X y en Y, obteniendo los siguientes diagramas:



Histogramas de las posiciones en X y en Y del paso número 100 de las 100 partículas



Histogramas de las posiciones en X y en Y del último paso de las 100 partículas

A simple vista se puede observar que parecen tener una distribución Gaussiana (Normal).

4.3.3 Prueba Shapiro-Wilk

Usando una prueba de Hipótesis *Shapiro-Wilk*, dado que no se conoce en primera instancia la distribución de estos datos, se probará si siguen dicha distribución.

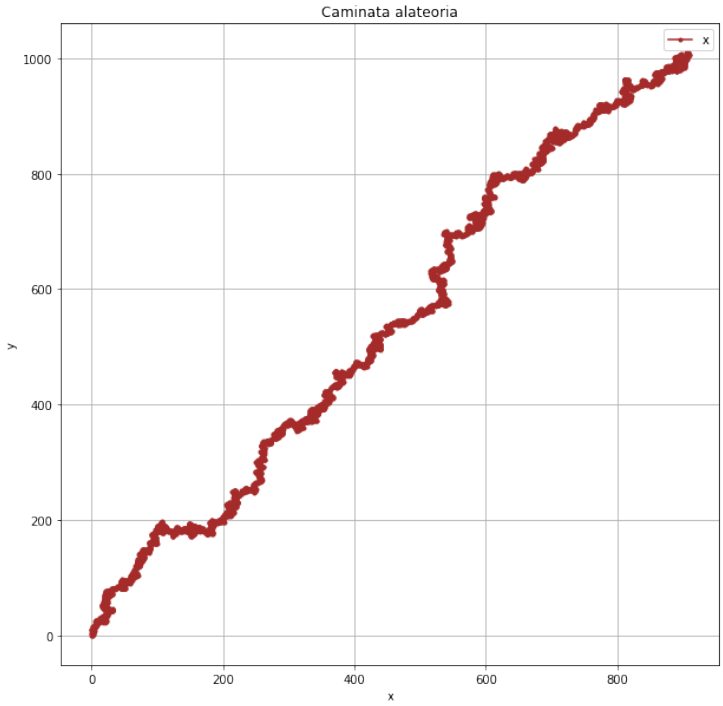
- H_0 = La muestra sigue una Distribución Gaussiana.

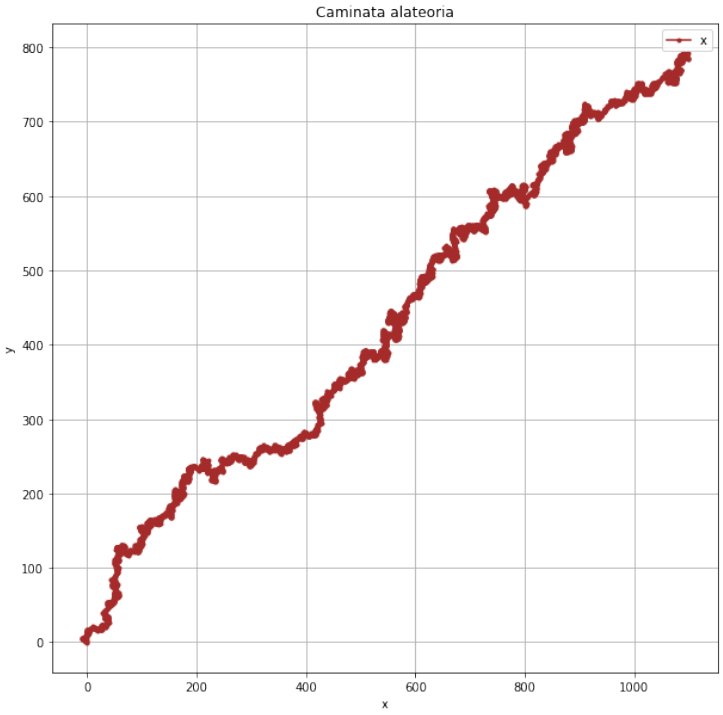
- H_1 = La muestra no sigue una Distribución Gaussiana.

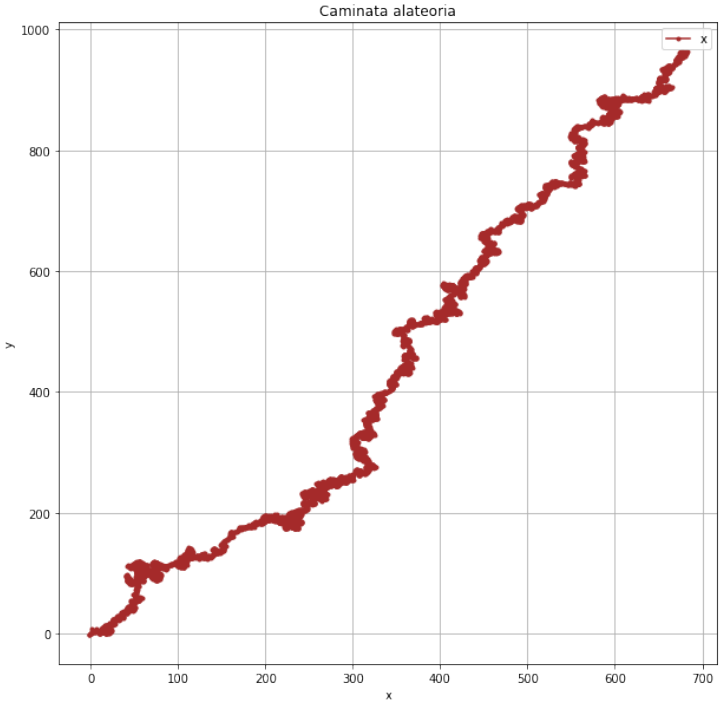
En la tabla que recopila el paso número 10,000 de cada partícula el p-valor obtenido fue de $p = 0.495$.
En la tabla que recopila el paso número 100 de cada partícula el p-valor obtenido fue de $p = 0.067$.
En ambos casos el p-valor resultó ser mayor al nivel de significancia, $\alpha = 0.05$.
En consecuencia no tenemos evidencia para no rechazar H_0. Por lo que pareciera que la muestra sigue una distribución normal.

4.3.4 Trayectorias

Se tomaron tres partículas al azar (de las 100 que teníamos registradas) y se graficaron sus trayectorias (posiciones X,Y).




Caminata alateoria
x
y
x
0
100
200
300
400
500
600
700
800
0
200
400
600
800
1000


Caminata alateoria
x
y
x
0
200
400
600
800
1000
0
100
200
300
400
500
600
700

4.4 Tercer conjunto de datos

Recordemos que para esta tabla de valores, en X, la probabilidad de dar un paso a la derecha es de 0.60, y en Y, la probabilidad de dar un paso hacia arriba es de 0.50.

4.4.1 Diagramas de Caja (Box-plot)

De manera gráfica, pueden observarse estos datos en diagramas de cajas, donde pueden visualizarse dónde están los cuartiles, la media y los datos extremos.

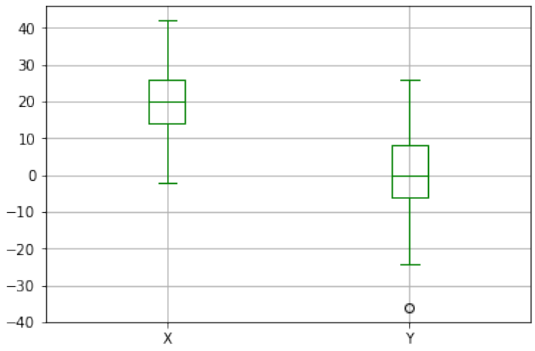


figure: Boxplots de X y Y del paso 100

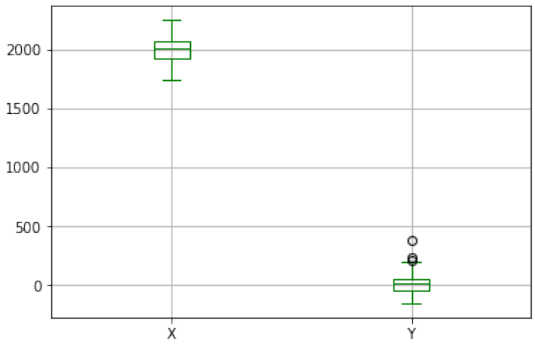
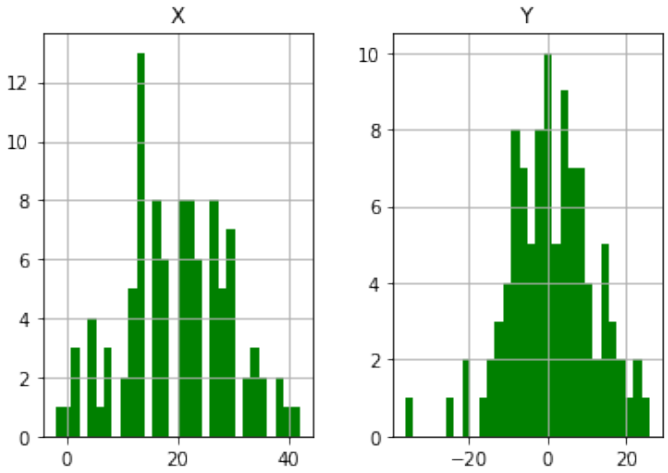


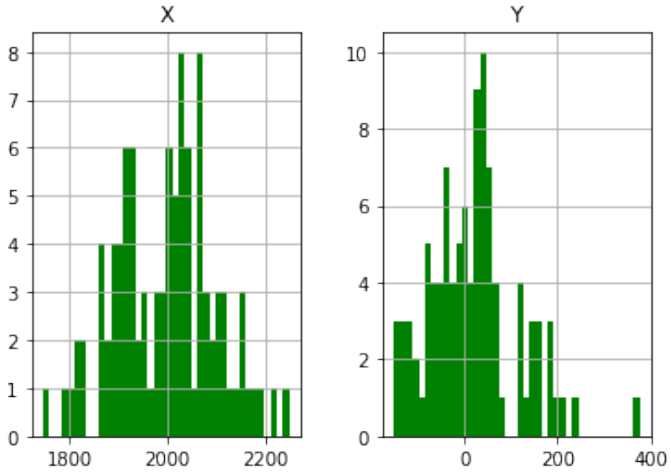
figure: Boxplots de X y Y del paso 10,000

4.4.2 Histogramas

Habiendo recolectado en dos DataFrames el paso número $10,000$ y 100 de cada una de las cien partículas, se generaron histogramas para las posiciones en X y en Y, obteniendo los siguientes diagramas:



Histogramas de las posiciones en X y en Y del paso número 100 de las 100 partículas



Histogramas de las posiciones en X y en Y del último paso de las 100 partículas

A simple vista se puede observar que parecen tener una distribución Gaussiana (Normal).

4.4.3 Prueba Shapiro-Wilk

Usando una prueba de Hipótesis *Shapiro-Wilk*, dado que no se conoce en primera instancia la distribución de estos datos, se probará si siguen dicha distribución.

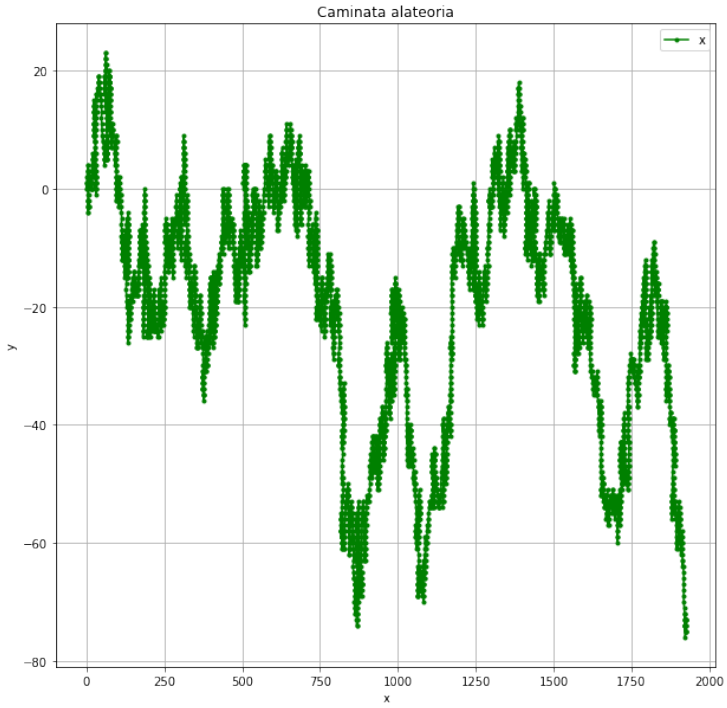
- H_0 = La muestra sigue una Distribución Gaussiana.

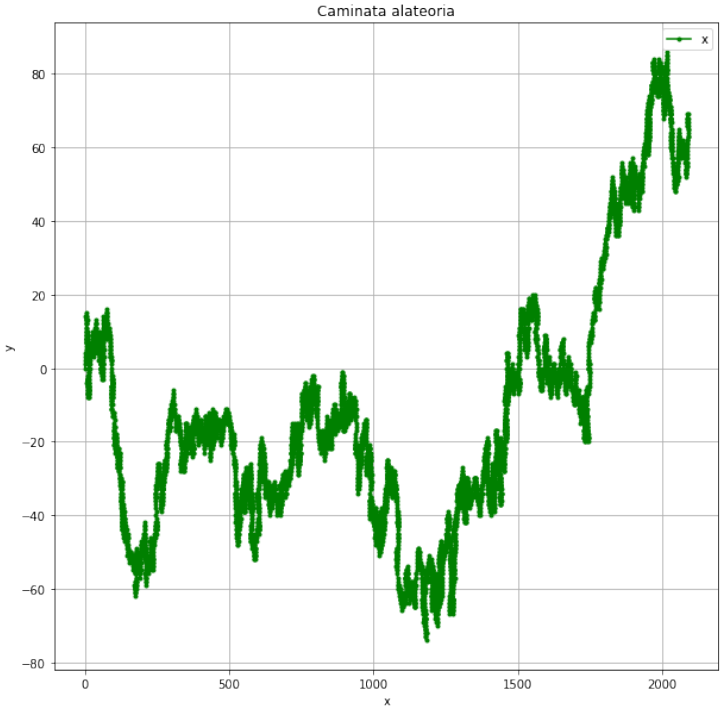
- H_1 = La muestra no sigue una Distribución Gaussiana.

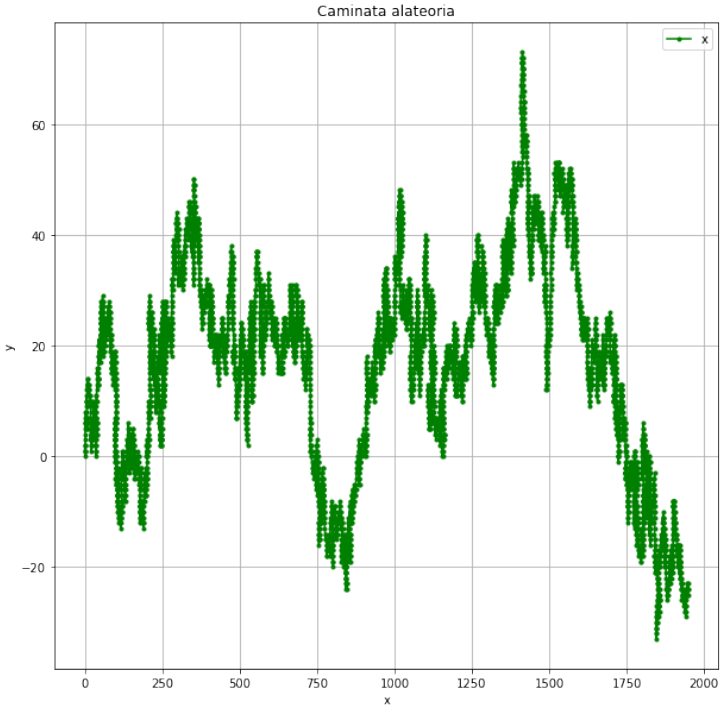
En la tabla que recopila el paso número 10,000 de cada partícula el p-valor obtenido fue de $p = 0.804$.
En la tabla que recopila el paso número 100 de cada partícula el p-valor obtenido fue de $p = 0.517$.
En las posiciones en X, el p-valor resultó ser mayor al nivel de significancia, $\alpha = 0.05$.
En consecuencia no tenemos evidencia para no rechazar H_0. Por lo que pareciera que la muestra sigue una distribución normal en el eje X.
Por otra parte, en las posiciones en Y, tanto para el paso 100 como el $10,000$, el p-valor fue menor que α, por lo que se rechaza H_0 y se tiene que hay evidencia para decir que, al parecer, no sigue una distribución Normal (puede ser el caso que se rechace cuando sí lo sea).

4.4.4 Trayectorias

Se tomaron tres partículas al azar (de las 100 que teníamos registradas) y se graficaron sus trayectorias (posiciones X,Y).




Caminata alateoria
x
y
x
80
60
40
20
0
−20
−40
−60
−80
0
500
1000
1500
2000

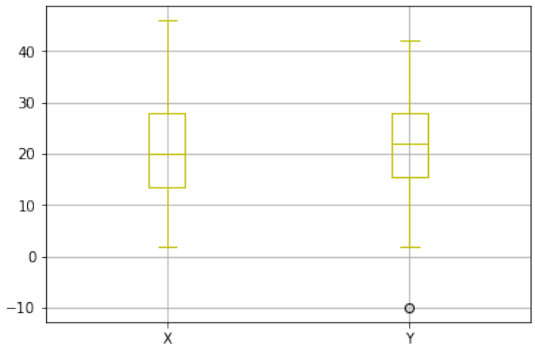

Caminata alateoria
x
y
x
60
40
20
0
−20
0
250
500
750
1000
1250
1500
1750
2000

4.5 Cuarto conjunto de datos

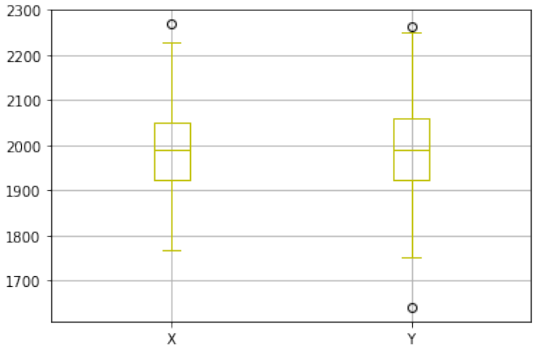
Recordemos que para esta tabla de valores, en X, la probabilidad de dar un paso a la derecha es de 0.60, y en Y, la probabilidad de dar un paso hacia arriba es de 0.60.

4.5.1 Diagramas de Caja (Box-plot)

De manera gráfica, pueden observarse estos datos en diagramas de cajas, donde pueden visualizarse dónde están los cuartiles, la media y los datos extremos.



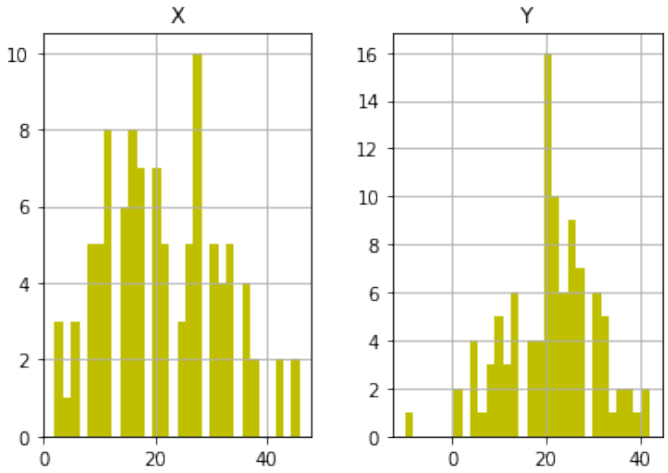
figureBoxplots de X y Y del paso 100



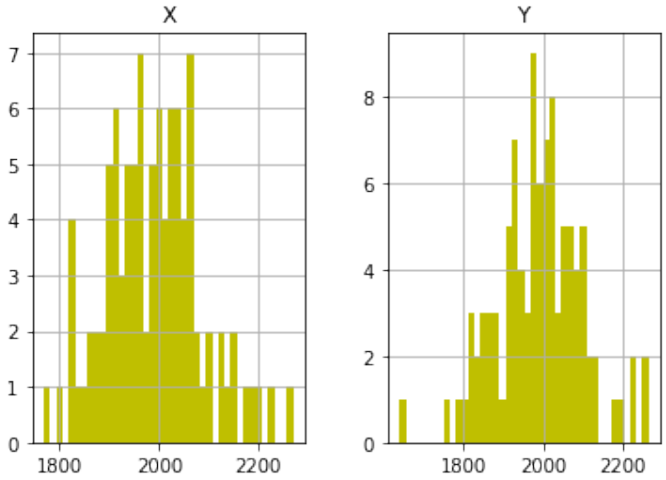
figureBoxplots de X y Y del paso 10,000

4.5.2 Histogramas

Habiendo recolectado en dos DataFrames el paso número $10,000$ y 100 de cada una de las cien partículas, se generaron histogramas para las posiciones en X y en Y, obteniendo los siguientes diagramas:



Histogramas de las posiciones en X y en Y del paso número 100 de las 100 partículas



Histogramas de las posiciones en X y en Y del último paso de las 100 partículas

A simple vista se puede observar que parecen tener una distribución Gaussiana (Normal).

4.5.3 Prueba Shapiro-Wilk

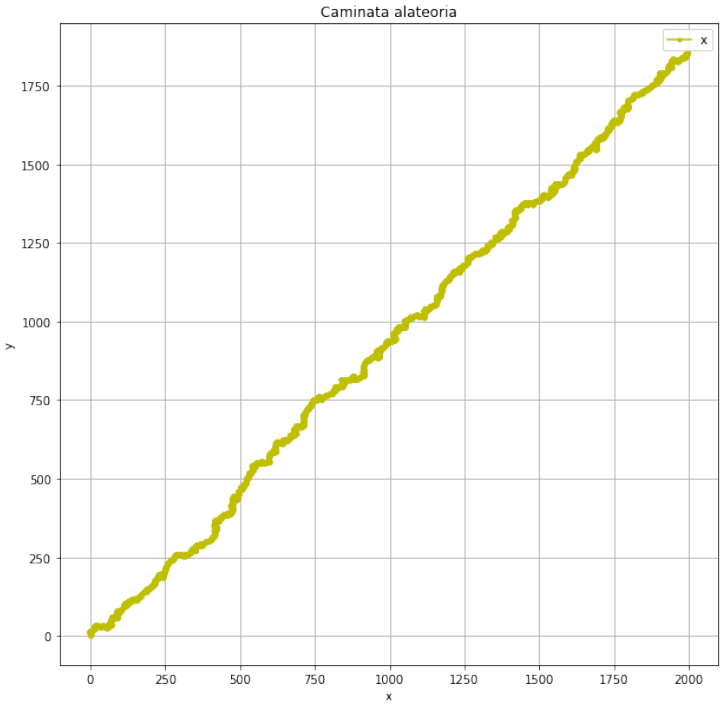
Usando una prueba de Hipótesis *Shapiro-Wilk*, dado que no se conoce en primera instancia la distribución de estos datos, se probará si siguen dicha distribución.

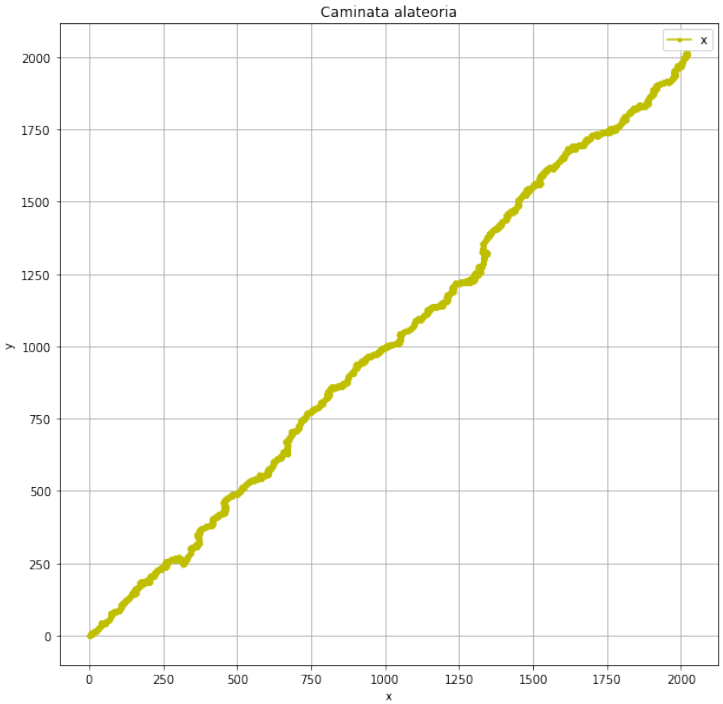
- H_0 = La muestra sigue una Distribución Gaussiana.
- H_1 = La muestra no sigue una Distribución Gaussiana.

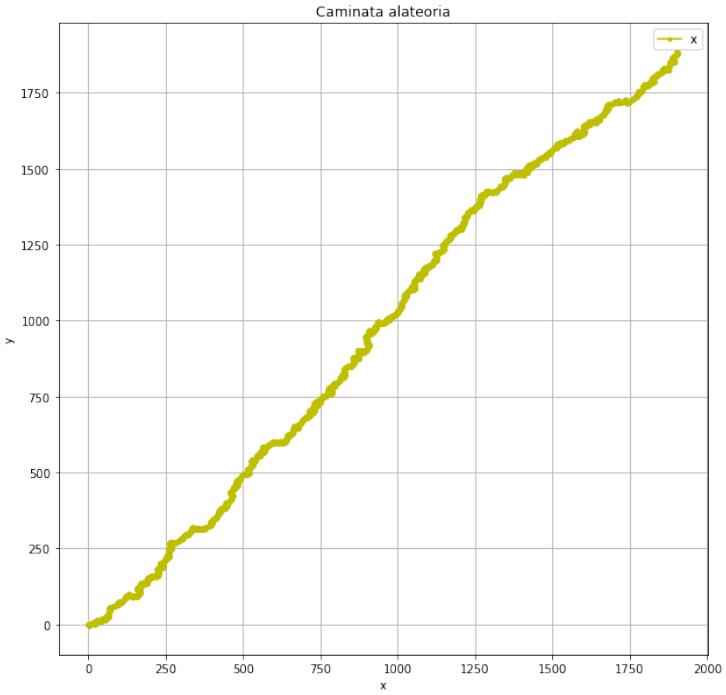
En la tabla que recopila el paso número 10,000 de cada partícula el p-valor obtenido fue de $p = 0.347$.
En la tabla que recopila el paso número 100 de cada partícula el p-valor obtenido fue de $p = 0.348$.
En ambos casos el p-valor resultó ser mayor al nivel de significancia, $\alpha = 0.05$.
En consecuencia no tenemos evidencia para no rechazar H_0. Por lo que pareciera que la muestra sigue una distribución normal.

4.5.4 Trayectorias

Se tomaron tres partículas al azar (de las 100 que teníamos registradas) y se graficaron sus trayectorias (posiciones X,Y).




Caminata alateoria
x
y
0
250
500
750
1000
1250
1500
1750
2000


Caminata alateoria
x
y
0
250
500
750
1000
1250
1500
1750
2000

5 Conclusión

Todos los datos estudiados parecieran segir una distribución normal. Esto podría deberse gracias al Teorema del Límite Central, ya que la probabilidad de que la partícula, después de una gran cantidad de pasos, esté a determinados pasos de su origen, debe corresponder a una distribución Gaussiana (Observado en los histogramas anteriores). Al cambiar la probabilidad de movimiento en cada dirección, es decir, privilegiar unas direcciones sobre otras, las gráficas mantenían la misma distribución normal con la media trasladada de lugar. Esto favorece al hecho de poder usar las caminatas aleatorias para modelar el movimiento Browniano, coincidiendo en esa característica de ajuste Gaussiano a pesar de las limitaciones probabilísticas en las direcciones tomadas por las partículas.

References

[1] Kielbowicz Valarezo, Augusto A. . (2017). Análisis estadístico y modelado numérico de trayectorias de partícula única: mecanismos de difusión y confinamiento. Mayo, 2019, de Universidad de Buenos Aires Sitio web: http://users.df.uba.ar/gpuentes/UBAThesis.pdf

[2] Gonzalez Bernal, Jesus A. . (2011). Teorema del Límite Central. Mayo, 2019, de INAOE Sitio web: https://ccc.inaoep.mx/ jagonzalez/ML/principal/node30.html

[3] Arroyo Calle, Adrián. (2018). Estadística en Python: ajustar datos a una distribución (parte VII). Mayo, 2019, de El blog de Adrián Arroyo Sitio web: https://blog.adrianistan.eu/estadistica-python-ajustar-datos-una-distribucion-parte-vii

[4] Anónimo. (2019). pandas.DataFrame.hist. Mayo, 2019, de PyData Sitio web: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.hist.html

[5] Anónimo. (2019). Camino Aleatorio. Mayo, 2019, de Wikipedia Sitio web: https://es.wikipedia.org/wiki/Camino_aleatorioDefinición

[6] Sir Maurice Kendall . (2015). ANALISIS DE TRAYECTORIA Y CONSTRUCCION DE MODELOS . Mayo, 2019, Sitio web: https://repositorio.cepal.org/bitstream/handle/11362/12592/NP14-03_es.pdf?sequence=1

[7] Anónimo. Teorema del Límite Central. Mayo, 2019, Sitio web: https://rstudio-pubs-static.s3.amazonaws.com/125985_ccf3bf4321304286a7eaa541343fa927.html

[8] Jason Brownlee. (2018). 15 Statistical Hypothesis Tests in Python . Mayo, 2019, de Machine Learning Mastery Sitio web: https://machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet/

[9] NIST/SEMATECH. (2012). Anderson-Darling and Shapiro-Wilk tests. Mayo, 2019, de Engineering Statistics Handbook Sitio web: https://www.itl.nist.gov/div898/handbook/prc/section2/prc213.htm