Tarea 2: Variables aleatorias

Jeaustin Sirias Chacón
GitHub: JeaustinSirias

Resumen
En este reporte académico se muestran los conceptos de probabilidad y estadística básicos para el análisis de datos números en una dinámica asistida por computación científica haciendo uso del lenguaje de programación Python3 para generar curvas de ajuste por distribución normal gausseana, histogramas, transformaciones numéricas y probabilidad de ocurrencia de sucesos intrínsecos a los datos muestreados.

Palabras claves
Python3 — probabilidad — variable aleatoria — función acumulativa — función de densidad

Escuela de Ingeniería Eléctrica, Universidad de Costa Rica
Correspondencia: JEAUSTIN.SIRIAS@ucr.ac.cr

Índice

Introducción

El estudio de la probabilidad implica de forma intrínseca la componente estadística por naturaleza. La necesidad de ordenar información de cualquier tipo de interés, interpretarla por análisis y obtener conclusiones para optimizar y/o desarrollar sistemas cotidianos, es la aplicación mas sobresaliente de los modelos probabilísticos. El establecimiento de criterios estadísticos basados en estudio y observación de patrones recursivos e iterativos es lo que adopta de forma genuina a las matemáticas en el campo.

1. Desarrollo

1.1 Nota teórica

Interpretación de histogramas

Un histograma [1] es una distribución en conjuntos y familias para un frasco de datos en particular de forma gráfica por barras. El esquema consta de dos ejes coordenados: la horizontal carga los valores independientes de acuerdo al espacio de muestras involucrado. Asimismo, el eje de las ordenadas representa la frecuencia con que se muestrean los datos en cada familia o barra. Los aspectos relevantes para la interpretación del histograma pueden destacarse como sigue:

- Observar los picos: estos representan los valores más comunes en el frasco de datos.
- Número de familias o *bins*: un dimensionamiento bajo puede ocasionar una baja distribución de datos en las barras.
- Aislamiento de barras: representan valores muy particulares del frasco de datos. Pueden influir en gran medida los resultados, interpretaciones y ajustes.

De las funciones de distribución

El ajuste de una distribución [1] de datos sobre el histograma se considera un modelo. Existe una amplia gama de distribuciones a partir de métodos matemáticos adyacentes al proceso: distribución gaussiana, Ryleigh, Burr12, Gamma... El detalle relevante de un ajuste radica en qué tan bien se moldea un modelo al comportamiento real, de modo que asemeje sus parámetros: la media, desviación estándar, varianza, kurtosis y momentos de interés en el análisis de datos.

Sobre momentos estadísticos

Entre los momentos [1] mas relevantes se exponen:

- **La media**: es una medida de tendencia central en una familia de datos y representa el valor promedio cuantitativo.
- **La varianza**: Dice qué tan dispersa es una variable aleatoria y define el riesgo de que no ocurra el suceso esperado.
- **Inclinación**: Define el grado de distorsión en términos de simetría en una curva: si el ajuste está a la izquierda o derecha se dice que hay inclinación.
- **kurtosis**: Indica el concentrado de valores del frasco de datos alrededor de la media. Gráficamente muestra qué tan prominente es el pico de una distribución.

1.2 Resultados

Para culminar las actividades propuestas en esta asignación se ha hecho uso de Python3 para acceder a librerías de computación científica. De primera entrada se solicitaba la creación de un histograma a partir de un frasco de datos aparentemente aleatorios y genéricos suministrado en formato [.csv] se importaron elementos los necesarios para el graficado y procesado de arreglos numéricos al entorno de desarrollo SublimeText3:

```
# Librerías
import seaborn as sns
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt
import csv
```

1.2.1 Histogramas, curvas y probrobabilidad

Una vez elegidas las librerías científicas a usar se emplea el objeto `plt.hist[]` con 50 divisiones por segmento (*bins*) sobre los datos suministrados para la ocasión para entonces obtener el histograma de la Figura 1.

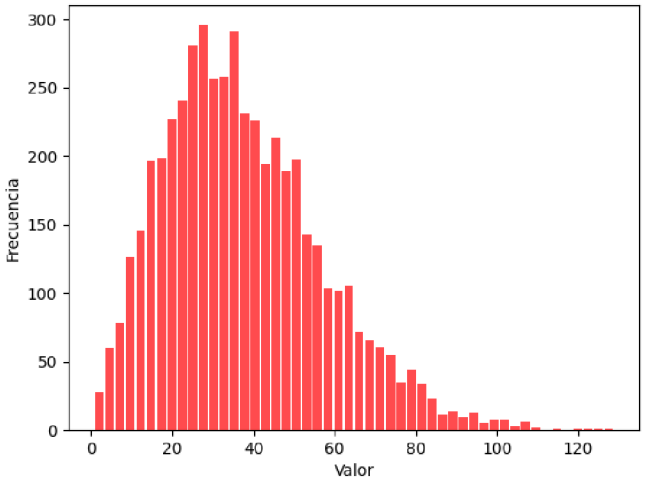

Figura 1. Histograma generado a partir de datos aleatorios.

Seguidamente empleando el ajuste por distribución de Ryleigh para generar una curva modelo empleando los parámetros de la media y la desviación estándar a partir del frasco de datos inicial a través del método `stats` de Scipy.

```
#Curva de ajuste
mu, sgm =stats.rayleigh.fit(lista)
model=stats.rayleigh(a,b)
```

Para valores de `mu=0.447` y `sgm=29.62` se ha obtenido la curva de ajuste que se muestra en la Figura 2. En aras la identificación estudiantil B66861, se procede a conocer la probabilidad de ocurrencia del suceso en el intervalo $[61, 68]$ a partir de la curva de ajuste de Ryleigh aplicando la siguiente linea en la terminal del SDK:

```
#Prob. de ocurrencia en [61, 68]
print(str(model.cdf(68)-model.cdf(61)))
```

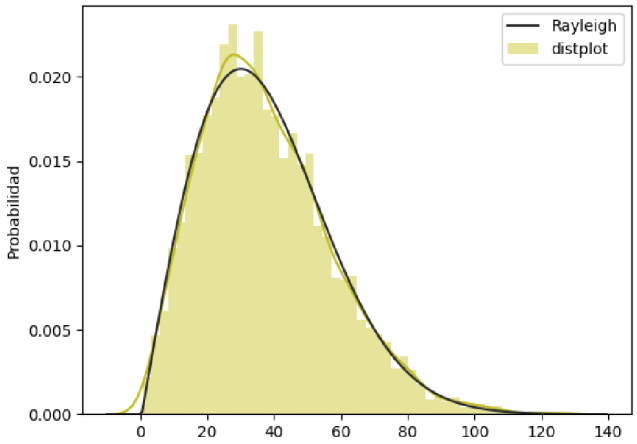

Figura 2. Curva de ajuste por distribución de Ryleigh.

Siendo entonces el resultado probabilístico $P(S_1) = 0{,}0495$ a partir del ajuste de Ryleigh, en contraste con el frasco de datos original y el concepto de frecuencia relativa, basta con programar un conteo de la cantidad de elementos abordados entre 61 y 68 ordenando de menor a mayor el arreglo de puntos flotantes:

```
#Prob. de ocurrencia en [61, 68]
sort=sorted(lista)
list=[i for i in sort if i >= 61 ...
and i <= 68]
Prob=len(list)/len(order)
```

Al final la probabilidad de ocurrencia es $P(S_2) = 0{,}0494$ con una diferencia porcentual negativa de 1 % respecto al modelo de ajuste de Ryleigh.

1.2.2 Contrastando momentos

Para este apartado los momentos se han evaluado de forma simultánea tanto para el modelo de ajuste por distribución de Ryleigh como para el frasco de datos original. Una mejor apreciación se muestra en la Tabla 1. como resumen comparativo.

Tabla 1. Momentos estadísticos relevantes

Comparativa de momentos		
	Frasco de datos	**Ajuste por Ryleigh**
Media	37.532	37.573
Asimetría	0.712	0.631
Varianza	379.354	376.613
kurtosis	0.474	0.245

Una forma de determinar estos parámetros es emplear la siguiente línea de comandos:

```
#hallando los 4 primeros momentos
m,s,v,k=model.stats(moments='msvk')
```

1.2.3 Transformación del frasco de datos

Como actividad final se solicitaba elaborar un histograma a partir de la transformación $Y = \sqrt{X}$ sobre el frasco de datos empleado desde el inicio. Introduciendo las siguientes líneas en la terminal del entorno:

```
#Transformacion
square=np.sqrt(lista)
HIST=plt.hist(square,50, ...
color='#008000', alpha=0.7, rwidth=0.85)
```

dan como resultado el histograma de la Figura 3. Obsérvese que para este caso la distribución de los elementos se ha vuelto mas uniforme en la parte central del histograma, a diferencia de la Figura 1.

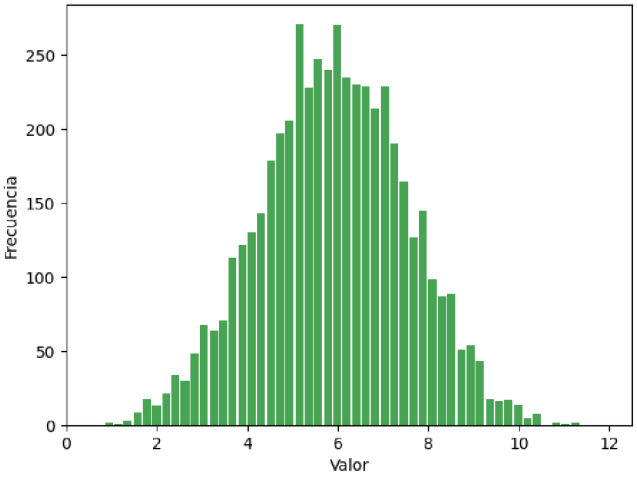

Figura 3. Curva de ajuste por distribución de Ryleigh.

2. Análisis de resultados y conclusiones

- Empezando por el histograma obtenido de la Figura 1. se observa que el conglomerado de valores mayoritariamente comunes en el frasco de datos ocurre para el intervalo [20, 43]. Asimismo por la forma de la distribución de las muestras, el histograma es naturalmente asimétrico hacia la izquierda; esta característica quiere decir que la mayoría de valores ocurren circunstancialmente rápido; por ejemplo, si dicho espacio de muestras representara el tiempo que duran los estudiantes de una clase en responder cuánto es 79+67, entonces la mayoría de los jóvenes habrán respondido antes de la mitad del total de la multitud. Adicionalmente pueden observarse la presencia de picos multimodales [?] en [20, 43]. Esto que los picos mas salientes son mas habituales que el resto.

- De forma conveniente se evaluó una curva de ajuste al histograma de la Figura 1. por distribución de Ryleigh para contrastar un modelo a partir del frasco de datos, con los mismos como se muestra en el ajuste de la Figura 2. Se ha determinado la probabilidad de ocurrencia en el intervalo [61, 68] tanto para el espacio de muestras original como para el modelo por Ryleigh, arrojado así un error relativo del 1 % entre ambos criterios; la interpretación de este este resultado radica para este caso en qué tanto ha logrado aproximar el modelo construido, la probabilidad en dicho intervalo, respecto a la solución analítica; no obstante para determinar la calidad del modelo se requiere elaborar la misma prueba para otras distribuciones comunes; sin embargo, a priori se puede decir que dicho error del 1 % es un resultado generoso por defecto.

- El contraste de los momentos estadísticos ya se tabulaban en la Tabla 1. a modo de comparación entre el ajuste de Ryleigh y el frasco original de datos: Respecto a la media ambos protagonistas muestran valores técnicamente iguales. Este parámetro está indicando que el valor promedio de la distribución y del modelo se dispone en 37.5. El uso de cifras significativas delatan el grado de rigurosidad con que se califica el la calidad del modelo. Adicionalmente obsérvese que la asimetría entrega un valor positivo en ambos casos, esto como era de esperar justifica el por qué la distribución es mas prominente hacia la izquierda en las Figuras 1. y 2; para este caso el ajuste de Ryleigh asemeja la solución analítica en un 88.63 %. Respecto a la interpretación de la varianza en 379.354 y 376.613 respaldan la evidente dispersión de valores en la muestra. Entre más arriba se encuentre este parámetro, más cerca están los valores de la media. Para este caso se arroja un error relativo de 99.27 % lo que indica que el ajuste por Ryleigh se encuentra ligeramente más lejos de la media respecto a su forma analítica. Finalmente respecto a la kurtosis, dado a que el parámetro es mayor a cero, entonces el comportamiento es leptocúrtico [1] para ambos casos; asimismo, se concluye que el modelo de Ryleigh tiene un pico menos pronunciado respecto a la distribución analítica, porque 0,474 > 0,245. Este último puede apreciarse detalladamente con las curvas de la Figura 2.

- Respecto a la transformación llevada a cabo en la sub sección 1.2.3, hay dos puntos a destacar desde la perspectiva gráfica: obsérvese el cambio de simetría entre los histogramas de la Figura 1. y 3: la inclinación de esta última se ha centrado asemejando una distribución simétrica, de modo que a priori se concluye que $\mu \approx 0$. Asimismo, gráficamente la media tiende a valores cercanos alrededor de 6. Apresuradamente se puede concluir que la radicalización uniformaliza un frasco de valores aleatorios.

Referencias

[1] Walpole et al. *Probabilidad y estadística para ingeniería y ciencias.* 9th, 2012.