

ACTIVIDAD 1. Distribuciones de Probabilidad

Instituto Tecnológico y de Estudios Superiores de Mty

Grace Aviance Silva Aróstegui A01285158

Concentración TC3006C:

Carlos Alberto Sánchez Villanueva A01640495

Inteligencia Artificial Avanzada para Ciencia de Datos

Fecha, 11 de agosto del 2024.

Módulo: Estadística Prof: Ramiro Zermeño Díaz

Campus Guadalajara, Zapopan.

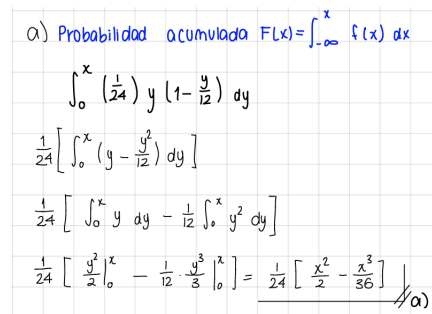
Problema 1

Una barra de 12 pulgadas que está sujeta por ambos extremos se somete a una cantidad creciente de esfuerzo hasta que se rompe. Sea Y = la distancia del extremo izquierdo al punto donde ocurre la ruptura. Suponga que Y tiene la función de densidad de probabilidad.

$$f(y) = \begin{cases} \left(\frac{1}{24}\right)y\left(1 - \frac{y}{12}\right), & \text{if } 0 \leq y \leq 12 \\ 0, & \text{otherwise} \end{cases}$$

Calcule lo siguiente:

(a) La función de distribución acumulativa de Y



a) Probabilidad acumulada $F(x) = \int_{-\infty}^x f(x) dx$

$$\int_0^x \left(\frac{1}{24}\right)y\left(1 - \frac{y}{12}\right) dy$$
$$\frac{1}{24} \left[\int_0^x \left(y - \frac{y^2}{12}\right) dy \right]$$
$$\frac{1}{24} \left[\int_0^x y dy - \frac{1}{12} \int_0^x y^2 dy \right]$$
$$\frac{1}{24} \left[\left.\frac{y^2}{2}\right|_0^x - \frac{1}{12} \left.\frac{y^3}{3}\right|_0^x \right] = \frac{1}{24} \left[\frac{x^2}{2} - \frac{x^3}{36} \right] \quad // a)$$

(b) $P(Y \leq 4)$, $P(Y > 6)$, y $P(4 \leq Y \leq 6)$

b) $P(Y \leq 4)$	$P(Y > 6)$	$P(4 \leq Y \leq 6)$
$\frac{1}{24} \left[\frac{4^2}{2} - \frac{4^3}{36} \right] = \frac{7}{24} = 0.26$	$\frac{1}{24} \left[\frac{6^2}{2} - \frac{6^3}{36} \right] = \frac{1}{2} = 0.5$	$0.5 - 0.26 = 0.24$

$$\Rightarrow P(Y \leq 4) = 0.26$$

$$\Rightarrow P(Y > 6) = 0.5$$

$$\Rightarrow P(4 \leq Y \leq 6) = 0.24$$

(c) $E(Y)$, $E(Y^2)$, $Var(Y)$

$$\Rightarrow E(Y) = 6$$

$$\Rightarrow E(Y^2) = 43.2$$

$$\Rightarrow Var(Y) = 7.2$$

$$\begin{aligned}
 c) \quad E(Y) &= \mu = \int_{-\infty}^{\infty} x \cdot f(x) dx & \text{Var}[Y] &= \underbrace{E[Y^2]}_6 - \underbrace{(E[Y])^2}_6 = 43.2 - 36 = \underline{7.2} // \\
 E(Y) &= \int_0^{12} y \cdot \left(\frac{1}{24}\right) y \left(1 - \frac{y}{12}\right) dy & \dots E(Y^2) &= \int_0^{12} y^2 \cdot \left(\frac{1}{24}\right) y \left(1 - \frac{y}{12}\right) dy \\
 &= \frac{1}{24} \int_0^{12} y^2 \left(1 - \frac{y}{12}\right) dy & &= \frac{1}{24} \int_0^{12} y^3 \left(1 - \frac{y}{12}\right) dy \\
 &= \frac{1}{24} \left[\int_0^{12} y^2 dy - \frac{1}{12} \int_0^{12} y^3 dy \right] & &= \frac{1}{24} \left[\int_0^{12} y^3 dy - \frac{1}{12} \int_0^{12} y^4 dy \right] \\
 &= \frac{1}{24} \left[\frac{y^3}{3} \Big|_0^{12} - \frac{1}{12} \cdot \frac{y^4}{4} \Big|_0^{12} \right] & &= \frac{1}{24} \left[\frac{y^4}{4} \Big|_0^{12} - \frac{1}{12} \cdot \frac{y^5}{5} \Big|_0^{12} \right] \\
 &= \frac{1}{24} [576 - 432] & &= \frac{1}{24} [5184 - 4147.2] \\
 &= \frac{1}{24} [144] = \underline{6} // & &= 43.2
 \end{aligned}$$

(d) La probabilidad de que el punto de ruptura ocurra a más de 2 pulg del punto de ruptura esperado.

$$\begin{aligned}
 d) \quad & \int_2^6 \left(\frac{1}{24}\right) y \left(1 - \frac{y}{12}\right) dy \\
 &= \frac{1}{24} \left[\int_2^6 \left(y - \frac{y^2}{12}\right) dy \right] \\
 &= \frac{1}{24} \left[\int_2^6 y dy - \frac{1}{12} \int_2^6 y^2 dy \right] \\
 &= \frac{1}{24} \left[\frac{y^2}{2} \Big|_2^6 - \frac{1}{12} \cdot \frac{y^3}{3} \Big|_2^6 \right] = \frac{1}{24} \left[16 - \frac{52}{9} \right] = \frac{23}{54} = 0.4259 \quad \therefore \underline{42.59\% \text{ de probabilidad}} //
 \end{aligned}$$

Problema 2

Sea X la temperatura, en grados centígrados, a la cual ocurre una reacción química. Suponga que X tiene una función de densidad de probabilidad:

$$f(x) = \begin{cases} \frac{1}{9}(4 - x^2), & \text{if } -1 \leq x \leq 2 \\ 0, & \text{otherwise} \end{cases}$$

(a) Corrobore que la función es una distribución válida.

$$\begin{aligned}
 a) \quad & \int_{-1}^2 \frac{1}{9} (4 - x^2) dx \\
 &= \frac{1}{9} \left[\int_{-1}^2 4 dx - \int_{-1}^2 x^2 dx \right] \\
 &= \frac{1}{9} \left[4x - \frac{x^3}{3} \right] \Big|_{-1}^2 = \frac{1}{9} \left[\frac{16}{3} - \left(-\frac{11}{3}\right) \right] = \frac{1}{9} [9] = \underline{1} // \\
 & \quad \therefore \text{Sí es una función de probabilidad}
 \end{aligned}$$

(b) Determine la función de distribución acumulativa.

$$\begin{aligned}
 \text{b) } F(x) &= \int_{-\infty}^x f(x) dx \\
 &= \int_0^x \frac{1}{9} (4-x^2) dx \\
 &= \frac{1}{9} \left[\int_0^x 4 dx - \int_0^x x^2 dx \right] \\
 &= \frac{1}{9} \left[4x - \frac{x^3}{3} \right] \Big|_0^x \\
 &= \frac{4}{9} x - \frac{1}{27} x^3 //
 \end{aligned}$$

(c) $E(Y)$, $E(Y^2)$, $Var(Y)$

$ \begin{aligned} \text{c) } E(Y) &= \mu = \int_{-\infty}^{\infty} x \cdot f(x) dx \\ &= \int_{-1}^2 x \cdot \frac{1}{9} (4-x^2) dx \\ &= \frac{1}{9} \int_{-1}^2 (4x - x^3) dx \\ &= \frac{1}{9} \left[2x^2 - \frac{x^4}{4} \right] \Big _{-1}^2 \\ &= \frac{1}{9} \left[4 - \frac{7}{4} \right] = \frac{1}{9} \left[\frac{9}{4} \right] = \frac{1}{4} // \end{aligned} $	$ \begin{aligned} E(Y^2) &= \int_{-1}^2 x^2 \cdot \frac{1}{9} (4-x^2) dx \\ &= \frac{1}{9} \int_{-1}^2 (4x^2 - x^4) dx \\ &= \frac{1}{9} \left[4 \cdot \frac{x^3}{3} - \frac{x^5}{5} \right] \Big _{-1}^2 \\ &= \frac{1}{9} \left[\frac{64}{15} - \left(-\frac{17}{15}\right) \right] = \frac{1}{9} \left[\frac{27}{5} \right] = \frac{3}{5} // \end{aligned} $	$ \begin{aligned} Var[Y] &= E[Y^2] - (E[Y])^2 \\ &= \frac{3}{5} - \left(\frac{1}{4}\right)^2 \\ &= \frac{43}{80} = 0.5375 \end{aligned} $
---	---	---

(d) La probabilidad de que la temperatura sea menor a 0°C

$$\begin{aligned}
 \text{d) } & \int_{-1}^0 \frac{1}{9} (4-x^2) dx \\
 &= \frac{1}{9} \left[\int_{-1}^0 4 dx - \int_{-1}^0 x^2 dx \right] \\
 &= \frac{1}{9} \left[4x - \frac{x^3}{3} \right] \Big|_{-1}^0 \\
 &= 0 - \frac{1}{9} \left[-4 - \left(-\frac{1}{3}\right) \right] = \frac{11}{27} = 0.4074 \\
 &= 40.74 \% //
 \end{aligned}$$

(e) La probabilidad de que la temperatura sea entre 4°C y 6°C

=> La probabilidad es 0 ya que de 4°C a 6°C está fuera del rango de la función brindada, la cual tiene como rango de -1 a 2.

Problema 3

El artículo “Computer Assisted Net Weight Control” (Quality Progress, 1983: 22-25) sugiere una distribución normal con media de 137.2 oz y una desviación estándar de 1.6 oz del contenido real de frascos de cierto tipo. El contenido declarado fue de 135 oz

- (a) ¿Cuál es la probabilidad de que un solo frasco contenga más que el contenido declarado?

```
import scipy.stats as stats

# Parámetros del problema
mu = 137.2 # Media
sigma = 1.6 # Desviación estándar
X = 135 # Valor declarado

# Z-score
z = (X - mu) / sigma

# Calcular la probabilidad P(X > 135)
probabilidad = 1 - stats.norm.cdf(z)
probabilidad

[3]
... 0.9154342776486631
```

=>La probabilidad es del 91.54 %

- (b) Suponiendo que la media permanece en 137.2, ¿a qué valor se tendría que cambiar la desviación estándar de modo que 95 % de todos los frascos contengan más que el contenido declarado?

Solución:

Primero necesitamos encontrar el z-score correspondiente al 95 %. Lo cual implica que el 5 % de los frascos tendrán un contenido igual o menor a 135 oz, por lo tanto necesitamos encontrar el z-score del 5 % de la cola izquierda de la distribución. Despejamos la fórmula de Estandarización para obtener el valor de la desviación estándar requerida:

$$\sigma = \frac{X - \mu}{z}$$

```
# Buscamos el z-score para el 5% en la cola izquierda (95% en la cola derecha)
z_95 = stats.norm.ppf(0.05)

# De la ec. de Estandarización despejada
sigma_requerida = (135 - mu) / z_95
sigma_requerida

[5] ✓ 0.0s
... 1.3375050302058846
```

=>Para que el 95 % de los frascos contengan mas que el contenido declarado, la desviación estándar tendría que tener un valor de 1.337

- (c) Entre 10 frascos seleccionados al azar, ¿cuál es la probabilidad de que por lo menos ocho contengan más que el contenido declarado?

Solución:

Para este contexto del problema, podemos utilizar la distribución binomial con los parámetros $n = 10$ (número de frascos) y $p = 0,915$ (probabilidad calculada en el inciso *a* de que un frasco contenga más de 135 oz). Queremos calcular $P(X \geq 8)$, para ello sumaremos las probabilidades de cuando $X = 8$, $X = 9$ y $X = 10$

```
from scipy.stats import binom

# Parámetros de la distribución binomial
n = 10 # Número de frascos
p = 0.915 # Probabilidad de que un frasco contenga más de 135 oz

# Probabilidad de que al menos 8 de los 10 frascos contengan más de 135 oz
prob = binom.pmf(8, n, p) + binom.pmf(9, n, p) + binom.pmf(10, n, p)
prob
```

[6] ✓ 0.0s

... 0.9532193904233863

=>La probabilidad es del 95.32 %

Problema 4

El artículo “Characterization of Room Temperature Damping in Aluminum-Idium Alloys” (Metallurgical Trans., 1993: 1611-1619) sugiere que el tamaño de grano de matriz A1 (μm) de una aleación compuesta de 2% de indio podría ser modelado con una distribución normal con valor medio de 96 y desviación estándar de 14.

- (a) ¿Cuál es la probabilidad de que el tamaño de grano exceda de 100?

Queremos calcular $P(X > 100)$. Para ello calculamos el valor Z correspondiente a un tamaño de grano de 100:

$$Z = \frac{X - \mu}{\sigma} \implies Z = \frac{100 - 96}{14} \approx 0,2857$$

Luego, buscamos la probabilidad de que Z sea mayor que 0.286 en la tabla de la distribución normal estándar o utilizando una calculadora estadística. $P(Z > 0,286)$

```
[3] 1 from scipy.stats import norm
2
3 # Parámetros de la distribución
4 mu = 96
5 sigma = 14
6
7 # a) Probabilidad de que el tamaño de grano exceda 100
8 z_a = (100 - mu) / sigma
9 prob_a = 1 - norm.cdf(z_a)
10 print("Probabilidad a: ", prob_a)
```

Executed at 2024.08.11 17:46:19 in 12ms

Probabilidad a: 0.38754848109799234

=>Por lo tanto, la probabilidad de que el tamaño de grano exceda los 100 μm es del 38.75 %

(b) ¿Cuál es la probabilidad de que el tamaño de grano sea de 50 y 80?

Para esto, calculamos los valores Z correspondientes a $X=50$ y $X=80$

```
[4] 1 # b) Probabilidad de que el tamaño de grano esté entre 50 y 80
2   z_b1 = (50 - mu) / sigma
3   z_b2 = (80 - mu) / sigma
4   prob_b = norm.cdf(z_b2) - norm.cdf(z_b1)
5   print("Probabilidad b: ", prob_b)
6
```

Executed at 2024.08.11 17:46:19 in 13ms

Probabilidad b: 0.1260403337983934

(c) ¿Qué intervalo (a, b) incluye el 90% central de todos los tamaños de grano (de modo que 5% esté por debajo de a y 5% por encima de b)?

Para encontrar el intervalo que incluye el 90% central, necesitamos encontrar los valores de Z correspondientes a las colas del 5% inferior y superior.

```
[5] 1 # c) Intervalo que incluye el 90% central de los tamaños de grano
2   z_c1 = norm.ppf(0.05)
3   z_c2 = norm.ppf(0.95)
4   a = z_c1 * sigma + mu
5   b = z_c2 * sigma + mu
6   print("Probabilidad c1: ", a)
7   print("Probabilidad c2: ", b)
```

Executed at 2024.08.11 17:46:19 in 12ms

Probabilidad c1: 72.97204922267937
Probabilidad c2: 119.02795077732061

Problema 5

Para los 3 conjuntos de datos que se proveen en el CSV:

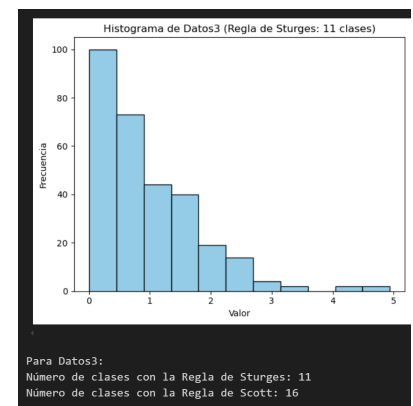
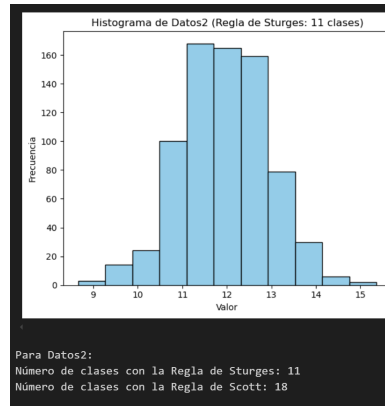
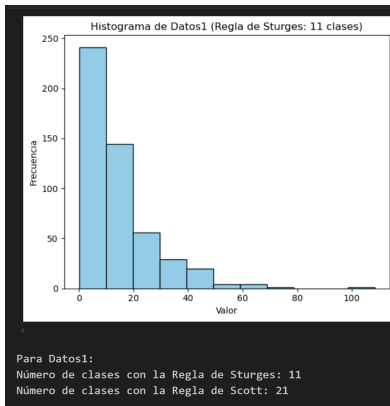
- Construye e interpreta un histograma. Utiliza la regla de Sturges para calcular el número apropiado de clases.
- Compara el número de clases con el obtenido con la regla de Scott.

Por tener 3 conjuntos de datos, tendremos 3 histogramas y mostraremos el número apropiado de clases de acuerdo con la regla de Sturges y la regla de Scott.

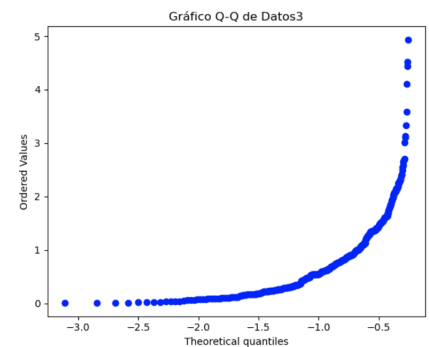
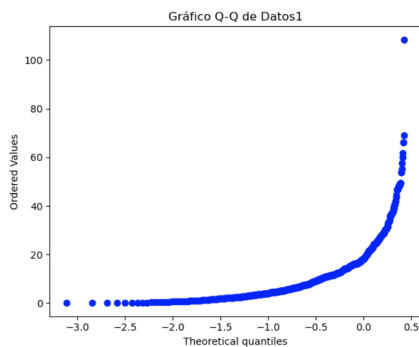
La distribución es asimétrica con una tendencia hacia la izquierda, por lo tanto la mayoría de los valores están concentrados en el rango inferior, con algunos valores atípicos más altos. Con la regla de Sturges, se generaron 11 clases, mientras que la regla de Scott sugirió 21 clases.

El histograma de Datos2 también sugiere una distribución asimétrica, pero menos marcada que Datos1. La distribución parece estar más centrada, con una ligera desviación hacia la izquierda. Con la regla de Sturges, se generaron 11 clases, mientras que la regla de Scott sugirió 18 clases.

Esta distribución tiene una asimetría pronunciada hacia la izquierda, lo que sugiere que hay varios valores extremadamente altos que influyen en la forma de la distribución, similar a los Datos1. La regla de Sturges sugiere 11 clases. La regla de Scott sugiere 16 clases.



- Construye e interpreta un gráfico Q-Q para comprobar si los datos provienen de una distribución normal. Estima los parámetros utilizando la regresión de un gráfico probabilístico.



En el primer gráfico se muestra una desviación significativa de la línea recta, especialmente en los

extremos. Esto indica que Datos1 no sigue una distribución normal. La curvatura en los extremos sugiere que los datos tienen colas más pesadas o más ligeras que una distribución normal, lo que podría implicar una distribución sesgada o con más valores extremos.

En el gráfico Q-Q para Datos2, los puntos están más alineados con la línea recta, pero aún hay desviaciones en los extremos, lo que indica que los datos no son perfectamente normales. Sin embargo, la alineación en la parte central sugiere que la distribución de Datos2 podría aproximarse más a una distribución normal en comparación con Datos1.

El gráfico Q-Q para Datos3 muestra una fuerte desviación de la línea recta, especialmente en los extremos superiores. Esto indica que Datos3 definitivamente no sigue una distribución normal. La gran curvatura sugiere que los datos tienen una distribución con una cola pesada a la derecha.

Código de Histogramas y Q-Q plot ...

```
[7] 1 import numpy as np
2 import matplotlib.pyplot as plt
3 import scipy.stats as stats
4
5 # Función para calcular el número de clases según la regla de Sturges
6 def sturges_rule(n):
7     return int(np.ceil(np.log2(n) + 1))
8
9 # Función para calcular el número de clases según la regla de Scott
10 def scott_rule(data):
11     return int(np.ceil((data.max() - data.min()) / (3.5 * np.std(data) / np.power(len(data), 1/3))))
12
13 # Función para graficar histogramas y Q-Q plot
14 def plot_histogram_and_qq(data, dataset_name):
15     # Número de clases según Sturges y Scott
16     n_sturges = sturges_rule(len(data))
17     n_scott = scott_rule(data)
18
19     # Histograma
20     plt.figure(figsize=(12, 5))
21
22     plt.subplot(1, 2, 1)
23     plt.hist(data, bins=n_sturges, color='skyblue', edgecolor='black')
24     plt.title(f'Histograma de {dataset_name} (Regla de Sturges: {n_sturges} clases)')
25     plt.xlabel('Valor')
26     plt.ylabel('Frecuencia')
27
28     # Gráfico Q-Q
29     plt.subplot(1, 2, 2)
30     stats.probplot(data, dist="norm", plot=plt)
31     plt.title(f'Gráfico Q-Q de {dataset_name}')
32
33     plt.tight_layout()
34     plt.show()
35
36     # Comparación de las reglas
37     print(f"Para {dataset_name}:")
38     print(f"Número de clases con la Regla de Sturges: {n_sturges}")
39     print(f"Número de clases con la Regla de Scott: {n_scott}")
40     print("\n")
41
42 # Aplicar la función a los tres conjuntos de datos
43 for col in data.columns:
44     plot_histogram_and_qq(data[col], col)
45
```


- (d) Utilizando Minitab o algún otro software, ¿a qué distribución es más probable que pertenezca cada conjunto de datos y cuáles serían sus respectivos parámetros?

```
... Mejor ajuste para Datos1: Exponential
Parámetros: (0.044, 13.948932000000001)
AIC: 3639.40294619679

Mejor ajuste para Datos2: Normal
Parámetros: (11.949312, 0.9809220295157681)
AIC: 2103.5143450198184

Mejor ajuste para Datos3: Exponential
Parámetros: (0.005, 0.9515633333333333)
AIC: 574.2105802650375
```

Código ...

```
from scipy.stats import norm, expon, lognorm, gamma, weibull_min
import numpy as np

# Función para ajustar distribuciones y calcular la bondad de ajuste (AIC)
def fit_distributions(data, dataset_name):
    # Limpiar los datos eliminando valores no finitos
    clean_data = data[np.isfinite(data)]

    # Lista de distribuciones a probar
    distributions = {
        'Normal': norm,
        'Exponential': expon,
        'Lognormal': lognorm,
        'Gamma': gamma,
        'Weibull': weibull_min
    }

    # Resultados de ajuste
    results = []

    for name, dist in distributions.items():
        # Ajustar la distribución a los datos
        params = dist.fit(clean_data)
        # Calcular el log-likelihood
        log_likelihood = np.sum(dist.logpdf(clean_data, *params))
        # Calcular el AIC
        aic = 2 * len(params) - 2 * log_likelihood
        results.append((name, aic, params))

    # Ordenar los resultados por AIC (menor es mejor)
    results.sort(key=lambda x: x[1])

    # Mostrar la distribución con el mejor ajuste
    best_fit = results[0]
    print(f"Mejor ajuste para {dataset_name}: {best_fit[0]}")
    print(f"Parámetros: {best_fit[2]}")
    print(f"AIC: {best_fit[1]}")
    print("\n")

# Aplicar la función a los tres conjuntos de datos
for col in data.columns:
    fit_distributions(data[col], col)
```