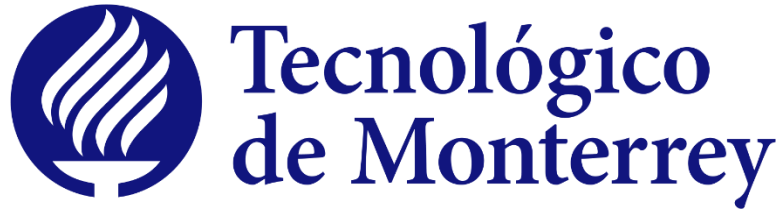


**Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Guadalajara**



Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)

M1.1 Actividad Distribuciones

Samuel Padilla Esqueda

| A01641383

Agosto 2024

Actividad 1: Distribuciones de Probabilidad

1.- Una barra de 12 pulg 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:

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

Calcule lo siguiente:

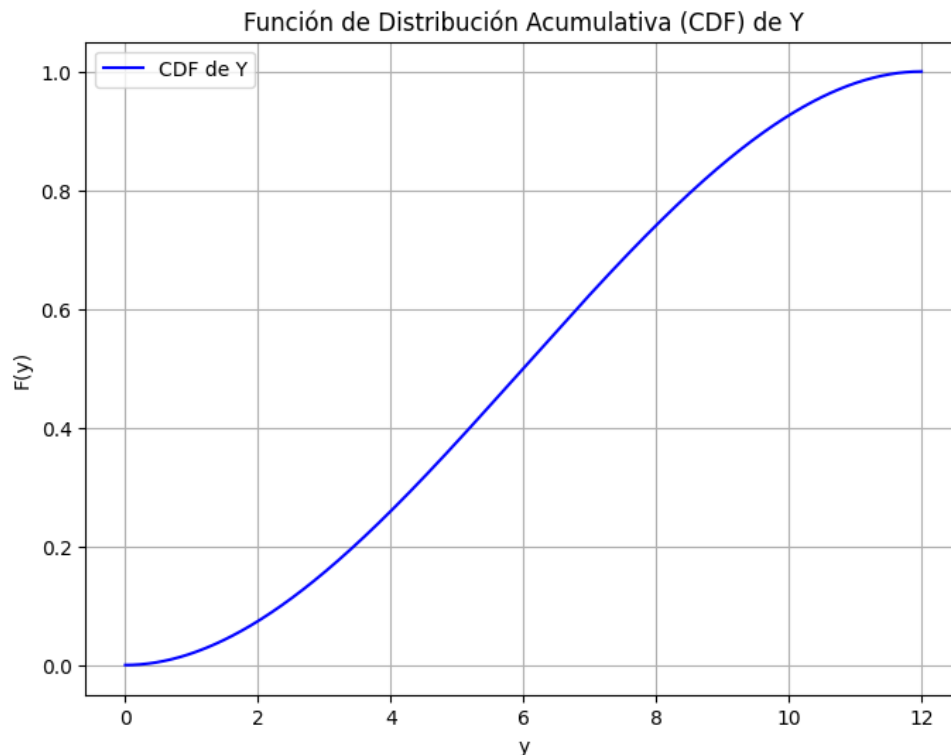
a.- La función de distribución acumulativa de Y .

Apoyándonos en Python para realizar las operaciones, sabemos que la función de distribución acumulativa se define como la integral de la función de densidad de probabilidad desde el límite inferior hasta un punto y .

```
# Definimos la función de densidad de probabilidad (PDF)
def pdf(y):
    return (1/24) * y * (1 - y/12)

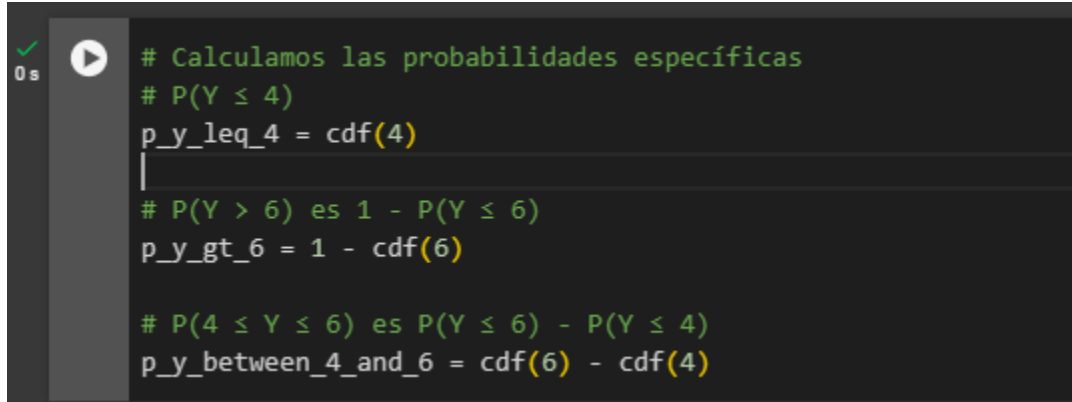
[6] # Calculamos la función de distribución acumulativa (CDF)
def cdf(y):
    # Integrar la PDF desde 0 hasta y
    result, _ = quad(pdf, 0, y)
    return result
```

Tras calcularla podemos generar valores y graficar la CDF.



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

Para calcular probabilidades específicas podemos llamar a nuestra función `cdf()` que hicimos en el ejercicio anterior, usando como parámetro la probabilidad que queremos encontrar. En el caso 1 esto es suficiente, pero para el caso 2 se nos pide el intervalo que es > 6 , para esto debemos restar $1 - \text{cdf}(6)$, debemos hacer esto también para el caso 3, pero en lugar de 1 usamos el segundo intervalo, hay que restar $\text{cdf}(6) - \text{cdf}(4)$.



```
# Calculamos las probabilidades específicas
#  $P(Y \leq 4)$ 
p_y_leq_4 = cdf(4)

#  $P(Y > 6)$  es  $1 - P(Y \leq 6)$ 
p_y_gt_6 = 1 - cdf(6)

#  $P(4 \leq Y \leq 6)$  es  $P(Y \leq 6) - P(Y \leq 4)$ 
p_y_between_4_and_6 = cdf(6) - cdf(4)
```

Obtenemos los siguientes resultados:

Probabilidad de que $(Y \leq 4)$: 25.93%

Probabilidad de que $(Y > 6)$: 50.00%

Probabilidad de que $(4 \leq Y \leq 6)$: 24.07%

c.- $E(Y)$, $E(Y^2)$ y $\text{Var}(Y)$.

Media $E(Y)$

$E(Y)$: Representa el valor promedio esperado de Y , siendo Y el punto de ruptura esperado de la barra.

$\text{Var}(Y)$: La varianza nos indica la dispersión de los valores de Y , alrededor de la media.

Esto quiere decir que nos indica que tan ampliamente están distribuidos los puntos de ruptura.

Podemos obtener los resultados con el siguiente código:

```
[13] # Esperanza E(Y)
def expected_value():
    result, _ = quad(lambda y: y * pdf(y), 0, 12)
    return result

# Esperanza del cuadrado E(Y^2)
def expected_value_squared():
    result, _ = quad(lambda y: y**2 * pdf(y), 0, 12)
    return result

# Calculamos E(Y), E(Y^2) y Var(Y)
E_Y = expected_value()
E_Y2 = expected_value_squared()
Var_Y = E_Y2 - E_Y**2

[14] # Imprimimos los resultados
print(f"E(Y): {E_Y:.4f}")
print(f"E(Y^2): {E_Y2:.4f}")
print(f"Var(Y): {Var_Y:.4f}")
```

Obtenemos los valores:

$E(Y)$: 6.0000 | Indica que, en promedio, el punto de ruptura de la barra ocurre a 6 pulgadas del extremo izquierdo.

$E(Y^2)$: 43.2000 | Valor usado para calcular la varianza.

$Var(Y)$: 7.2000 | Indica la desviación de los puntos de ruptura respecto a la media.

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

La probabilidad de que el punto de ruptura ocurra a más de 2 pulgadas del punto de ruptura esperado es calculada como $P(|Y - E(Y)| > 2)$.

Esta probabilidad nos indica cuán probable es que el punto de ruptura esté significativamente lejos del valor esperado.

Para calcular esta probabilidad establecemos un limite inferior y otro superior, calculamos usando nuestro método `cdf()`, sumando las probabilidades de que la barra se rompa a -2 pulgadas del punto esperado y a +2 pulgadas del punto esperado.

```
[15] # Calculamos las probabilidades para  $|Y - E(Y)| > 2$ 
lower_bound = E_Y - 2
upper_bound = E_Y + 2

# Probabilidad de que el punto de ruptura ocurra más de 2 pulgadas del punto de ruptura esperado
probability = cdf(lower_bound) + (1 - cdf(upper_bound))

[16] # Imprimimos el resultado como porcentaje
print(f"P( $|Y - E(Y)| > 2$ ): {probability * 100:.2f}%")
```

Obtenemos el siguiente resultado:

$P(|Y - E(Y)| > 2)$: 51.85%

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:

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

a.- Corrobore que la función es una distribución válida.

Para corroborar debemos asegurarnos de que $f(x)$ no sea negativa y que la integral de todo el espacio sea igual a 1. Para esto usamos el siguiente código:

```
[20] # Definimos la función de densidad de probabilidad (PDF)
def pdf(x):
    if -1 <= x <= 2:
        return (1/9) * (4 - x**2)
    else:
        return 0

# Verificar que la PDF sea no negativa en el intervalo
def check_non_negativity():
    x_values = np.linspace(-1, 2, 1000)
    pdf_values = np.array([pdf(x) for x in x_values])
    return np.all(pdf_values >= 0)

[22] # Calcular la integral de la PDF sobre su dominio
def verify_integral():
    integral, _ = quad(pdf, -1, 2)
    return integral

[23] # Verificar la validez de la PDF
is_non_negative = check_non_negativity()
integral_value = verify_integral()

[24] # Imprimir los resultados
print(f"La función PDF es no negativa en el intervalo [-1, 2]: {is_non_negative}")
print(f"Integral de la PDF sobre el intervalo [-1, 2]: {integral_value:.4f}")

# Verificar si la integral es aproximadamente 1
if np.isclose(integral_value, 1):
    print("La función es una distribución de probabilidad válida.")
else:
    print("La función NO es una distribución de probabilidad válida.")
```

Obtenemos el siguiente resultado:

La función PDF es no negativa en el intervalo $[-1, 2]$: True

Integral de la PDF sobre el intervalo $[-1, 2]$: 1.0000

Por lo tanto, la función es una distribución de probabilidad válida.

b.- Determine la función de distribución acumulativa.

Definimos en Python, tomamos en cuenta los límites:

```
[25] # Calculamos la función de distribución acumulativa (CDF)
def cdf(x):
    if x < -1:
        return 0
    elif -1 <= x <= 2:
        result, _ = quad(pdf, -1, x)
        return result
    else:
        return 1
```

c.- $E(Y)$, $E(Y^2)$ y $\text{Var}(Y)$.

Podemos obtener los resultados con el siguiente código:

```
[27] # Calcular E(X)
def expected_value():
    result, _ = quad(lambda x: x * pdf(x), -1, 2)
    return result

# Calcular E(X^2)
def expected_value_squared():
    result, _ = quad(lambda x: x**2 * pdf(x), -1, 2)
    return result

# Calcular E(X), E(X^2) y Var(X)
E_X = expected_value()
E_X2 = expected_value_squared()
Var_X = E_X2 - E_X**2
```

Obtenemos los siguientes resultados:

$E(X)$: 0.2500

$E(X^2)$: 0.6000

$\text{Var}(X)$: 0.5375

La media de 0.25°C indica que, en promedio, la temperatura a la cual ocurre la reacción química es de 0.25 grados centígrados.

La varianza de 0.5000 grados cuadrados mide la dispersión de las temperaturas de reacción alrededor de la media.

d.- La probabilidad de que la temperatura sea menor a 0°C

```
[30] # Calcular la probabilidad de que la temperatura sea menor a  $0^{\circ}\text{C}$ 
      probability_less_than_0 = cdf(0)

      # Imprimir el resultado como porcentaje
      print(f"P(X < 0) en porcentaje: {probability_less_than_0 * 100:.2f}%")
```

$P(X < 0)$ en porcentaje: 40.74%

Indica la posibilidad de que la temperatura sea menor a 0.

e.- La probabilidad de que la temperatura sea entre 4°C y 6°C

```
[31] # Calcular la probabilidad de que la temperatura sea entre  $4^{\circ}\text{C}$  y  $6^{\circ}\text{C}$ 
      probability_between_4_and_6 = cdf(6) - cdf(4)

      # Imprimir el resultado como porcentaje
      print(f"P( $4 \leq X \leq 6$ ) en porcentaje: {probability_between_4_and_6 * 100:.2f}%")
```

$P(4 \leq X \leq 6)$ en porcentaje: 0.00%

Indica la posibilidad de que la temperatura este entre 4 y 6, podemos ver que es 0 ya que al inicio definimos el rango entre -1 y 2 y por lo tanto es imposible estar entre estos valores.

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?

Podemos calcular esto gracias a las librerías de Python

```
[34] # Parámetros de la distribución normal
      mu = 137.2 # Media
      sigma = 1.6 # Desviación estándar

      # Contenido declarado
      contenido_declarado = 135

      # Calcular la probabilidad de que el contenido sea menor o igual al contenido declarado
      p_less_than_or_equal_135 = norm.cdf(contenido_declarado, mu, sigma)

      # Calcular la probabilidad de que el contenido sea mayor al contenido declarado
      p_greater_than_135 = 1 - p_less_than_or_equal_135

      # Imprimir el resultado como porcentaje
      print(f"P(X > {contenido_declarado} oz): {p_greater_than_135 * 100:.2f}%")
```

Obtenemos que la probabilidad es del 91.54%, por lo que la probabilidad de que 1 solo frasco contenga mas contenido que el declarado es muy elevada.

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?

```
# Parámetros dados
mu = 137.2 # Media
contenido_declarado = 135 # Contenido declarado

# Encontrar el valor crítico z correspondiente al percentil 5% (dado que queremos que el 95% esté por encima de X)
z_5_percentile = norm.ppf(0.05)

# Resolver para sigma
sigma = (contenido_declarado - mu) / z_5_percentile

# Imprimir el resultado
print(f"La desviación estándar necesaria para que el 95% de los frascos contengan más que {contenido_declarado} oz es: {sigma:.2f} oz")
```

Calculamos una nueva desviación estándar adaptada al porcentaje que necesitamos.

La desviación estándar necesaria para que el 95% de los frascos contengan más que 135 oz es: 1.34 oz

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

```
7] # Número de frascos
n = 10

# Número mínimo de frascos con éxito (contenido > 135 oz)
k_min = 8

# Probabilidad de éxito (ya calculada)
p_success = p_greater_than_135

# Calcular la probabilidad de tener al menos 8 frascos con éxito
p_at_least_8 = sum(binom.pmf(k, n, p_success) for k in range(k_min, n + 1))

# Imprimir el resultado como porcentaje
print(f"La probabilidad de que al menos 8 de 10 frascos contengan más que {contenido_declarado} oz es: {p_at_least_8 * 100:.2f}%")
```

La probabilidad de que al menos 8 de 10 frascos contengan más que 135 oz es: 95.38%

Podemos ver que es muy probable que casi siempre 8 de 10 frascos contengan mas que 135 oz.

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?

```
[38] # Parámetros de la distribución normal
mu = 96 # Media
sigma = 14 # Desviación estándar

# a. Probabilidad de que el tamaño de grano exceda de 100  $\mu\text{m}$ 
p_exceed_100 = 1 - norm.cdf(100, mu, sigma)

print(f"a. Probabilidad de que el tamaño de grano exceda de 100  $\mu\text{m}$ : {p_exceed_100 * 100:.2f}%")
```

Probabilidad de que el tamaño de grano exceda de 100 μm : 38.75%

Esto significa que la probabilidad de que el tamaño exceda 100 es poca, menos de la mitad.

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

```
# b. Probabilidad de que el tamaño de grano esté entre 50 y 80 µm
p_between_50_and_80 = norm.cdf(80, mu, sigma) - norm.cdf(50, mu, sigma)

print(f"b. Probabilidad de que el tamaño de grano esté entre 50 y 80 µm: {p_between_50_and_80 * 100:.2f}%")
```

Probabilidad de que el tamaño de grano esté entre 50 y 80 µm: 12.60%

La posibilidad de que el tamaño del grano se encuentre en este rango específico es menor todavía, casi solo el 10%.

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)?

```
# c. Encontrar el intervalo que incluye el 90% central de todos los tamaños de grano
# Usamos percentiles 5% y 95%
a = norm.ppf(0.05, mu, sigma)
b = norm.ppf(0.95, mu, sigma)

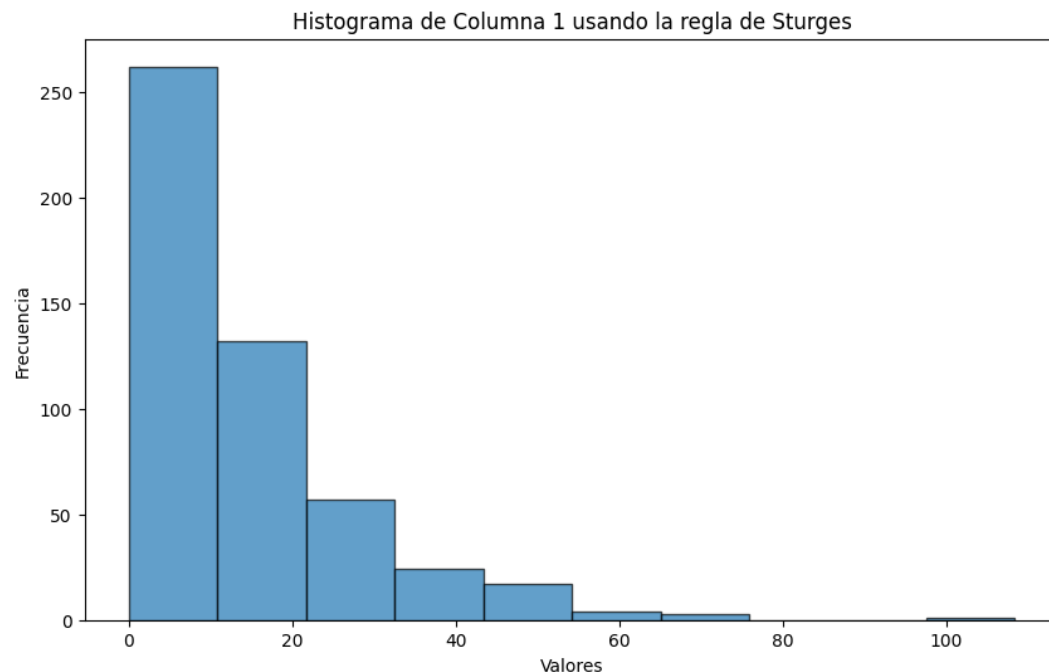
print(f"c. El intervalo que incluye el 90% central de todos los tamaños de grano es ({a:.2f}, {b:.2f}) µm")
```

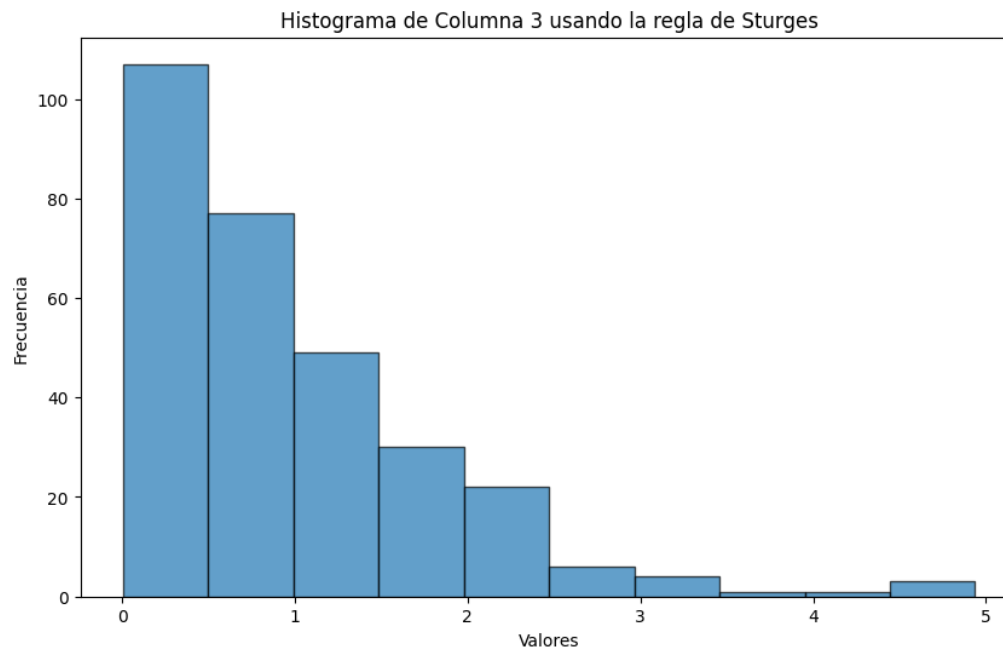
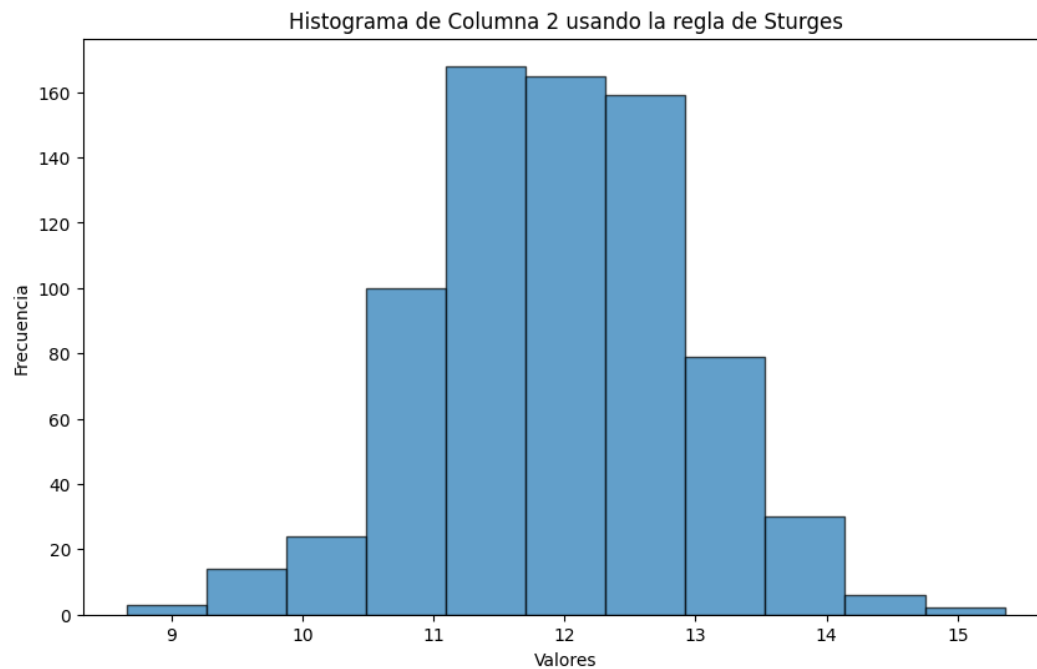
El intervalo que incluye el 90% central de todos los tamaños de grano es (72.97, 119.03) µm

Esto significa que el 90% de los tamaños de grano caen dentro de este rango, por lo que solo el 5% de grano está por debajo de 72.97 y 5% estará por encima de 119.03.

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

a.- Construye e interpreta un histograma. Utiliza la regla de Sturges para calcular el número apropiado de clases.





De acuerdo con la forma de cada histograma podemos observar hacia donde tienden los datos, gracias a su comportamiento podemos darnos cuenta de que pertenecen diferentes distribuciones. La grafica 1 y 3 se parecen mucho por lo que es posible que ambas pertenezcan a la misma distribución.

b.- Compara el número de clases con el obtenido con la regla de Scott.

Número de clases según la regla de Sturges columna 1: 10

Número de clases según la regla de Scott columna 1: 19

Número de clases según la regla de Sturges columna 2: 11

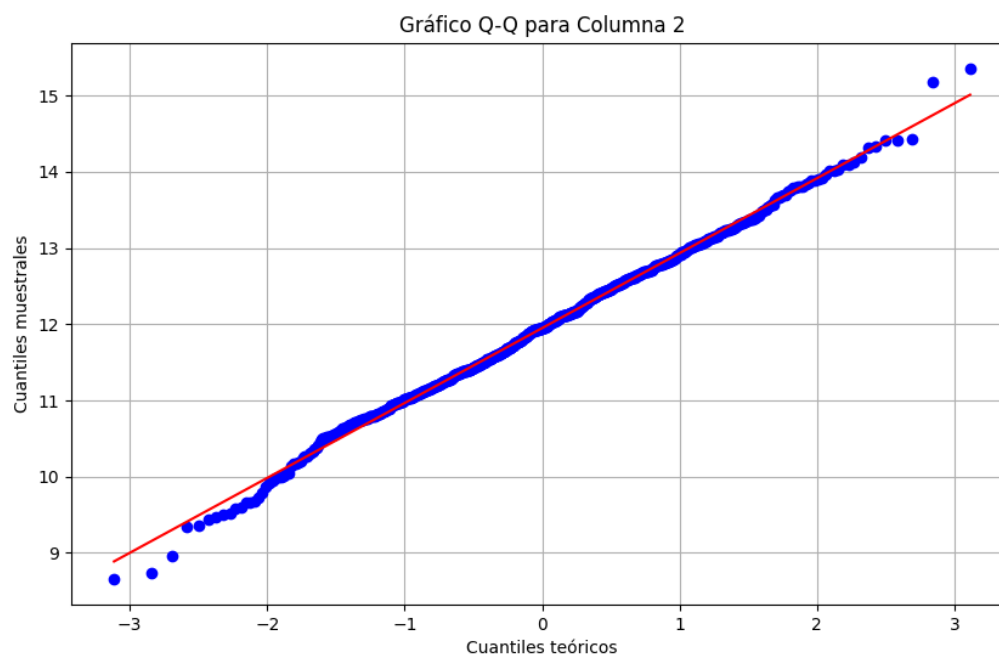
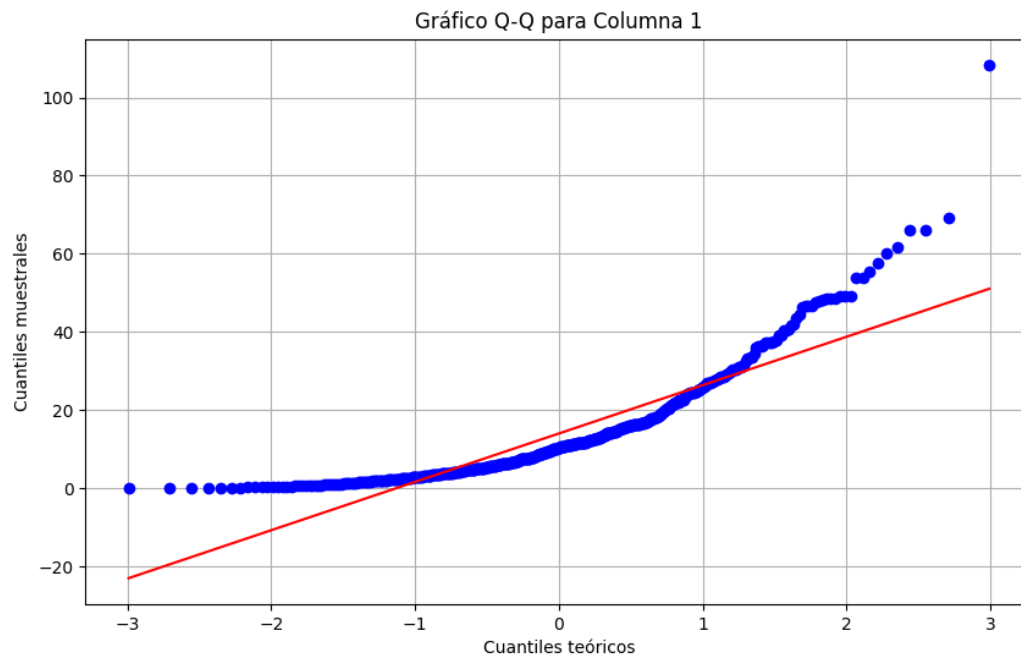
Número de clases según la regla de Scott columna 2: 18

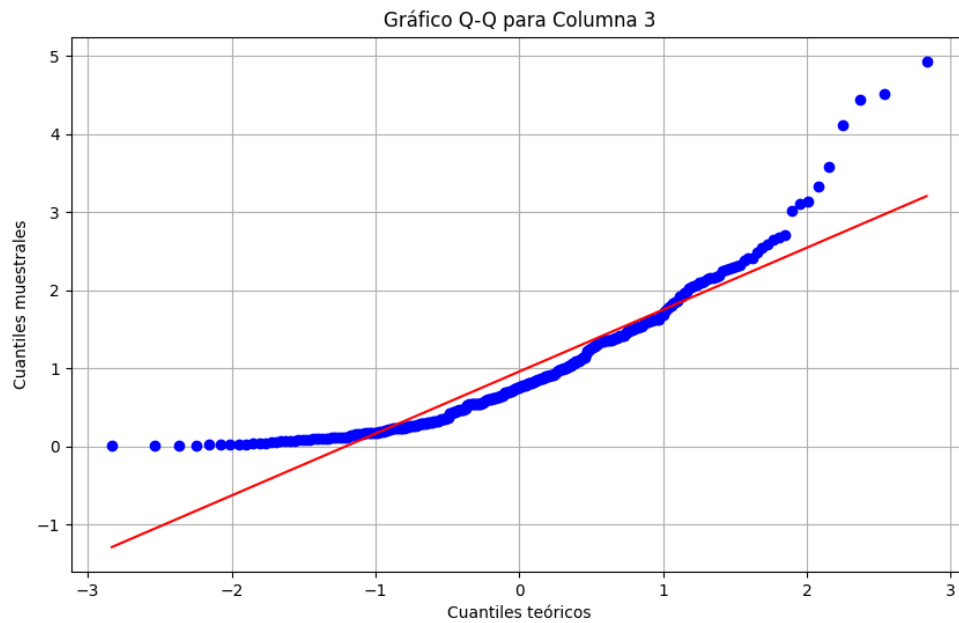
Número de clases según la regla de Sturges columna 3: 10

Número de clases según la regla de Scott columna 3: 12

Vemos que en su mayoría la regla de Scott requiere mas clases que la regla de Sturges, en algunos casos incluso casi el doble.

c.- 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.





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?

Podemos observar que en la segunda grafica los puntos se alinean casi perfectamente con la línea, por lo que creo que es una distribución normal.

La grafica 1 y 3 tienen más una curvatura, no parece que pertenezcan a la distribución normal, podría tratarse de distribuciones de cola pesada, binomial o alguna otra.