Tarea N° 3- Distribuciones directas

Asignatura: Probabilidad y estadística aplicada

Alumnos:

Gonzalo Paz 4.266.825-0

Juan Pérez 4.673.899-0

Lucas Cordero 5.480.931-3

Estefany Clara 4.653.959-2

Montevideo, 03 de junio de 2023

Contenido

La Distribución Binomial:

Las distribuciones discretas son modelos matemáticos que describen la probabilidad de ocurrencia de eventos. Tienen como diferencia con las distribuciones continuas, en las cuales las variables pueden tomar cualquier valor dentro de un rango continuo, las variables discretas solo pueden asumir valores específicos y separados.

La distribución binomial es un modelo de probabilidad discreta que cuenta el número de éxitos en una secuencia de n ensayos de Bernoulli independientes entre sí, con una probabilidad fija p de ocurrencia del éxito entre los ensayos. Un experimento de Bernoulli se caracteriza por ser dicotómico, esto es, sólo son posibles dos resultados.

A uno de estos se denomina éxito y tiene una probabilidad de ocurrencia p y al otro, fracaso, con una probabilidad q = 1 - p. En la distribución binomial el anterior experimento se repite n veces, de forma independiente, y se trata de calcular la probabilidad de un determinado número de éxitos.

El experimento de lanzar una moneda al aire es un ejemplo sencillo de este caso, sólo hay dos resultados posibles cara o cruz.

En resumen, el experimento binomial tiene cinco características:

1- El experimento consiste en n intentos idénticos
2- Cada intento resulta en uno de los dos resultados.
3- La probabilidad de éxito en un solo intento es igual a p y es igual de un intento a otro. La probabilidad de fracaso es igual a *(1-p) = q*
4- Los intentos son independientes entre sí, es decir, el resultado de un experimento no afecta al resultado de los demás.
5- Lo que nos importa es el número de éxitos observados durante los n intentos.

$$P(x = k) = C^n_k p^k q^{n-k} = \frac{n!}{k!(n-k)!} p^k q^{n-k}$$

La Distribución Bernoulli

La distribución Bernoulli es un modelo de probabilidad discreta que representa el resultado de un experimento o ensayo con dos posibles resultados: éxito o fracaso. Estos resultados se pueden asociar a los valores 1 y 0, respectivamente. La distribución Bernoulli se utiliza para describir situaciones en las que solo hay dos resultados posibles y la probabilidad de éxito o fracaso es constante para cada ensayo independiente.

En términos matemáticos, la distribución Bernoulli se define mediante un parámetro p, que representa la probabilidad de éxito en un solo ensayo. La función de masa de probabilidad (PMF, por sus siglas en inglés) de la distribución Bernoulli se define de la siguiente manera:

P(X = x) = p^x * (1 - p)^(1 - x)

Donde:

P(X = x) es la probabilidad de que la variable aleatoria X tome el valor x.

x toma los valores 0 o 1 (fracaso o éxito).

p es la probabilidad de éxito en un solo ensayo.

La media (esperanza) de la distribución Bernoulli es E(X) = p y la varianza es Var(X) = p * (1 - p). Esta distribución es un caso especial de la distribución binomial, donde se realiza un solo ensayo.

La distribución Bernoulli es ampliamente utilizada en estadística y teoría de la probabilidad para modelar eventos dicotómicos, como lanzar una moneda (cara o cruz) o el éxito o fracaso de un experimento.

La Distribución Geométrica

La distribución geométrica es un modelo de probabilidad discreta que se basa en la ocurrencia de los fracasos antes de obtener el primer éxito en una sucesión de repeticiones independientes de un experimento aleatorio, que tiene únicamente dos resultados posibles, denominados éxito y fracaso, y cada repetición del experimento tiene idéntica probabilidad de éxito p. La distribución geométrica se puede aplicar a los siguientes casos:

1. El experimento consiste en repeticiones independientes de un experimento aleatorio, cada una de las cuales sólo puede tener dos resultados posibles, denominados éxito y fracaso.

2. La probabilidad de éxito en cada experimento es constante, es decir, no cambia de una repetición a otra.

3. Los experimentos son independientes entre sí, es decir, el resultado de un experimento no afecta al resultado de los demás.

La distribución de Poisson

La distribución de Poisson es un modelo de probabilidad discreta que sirve para modelar el número de sucesos que se presentan en un intervalo de tiempo o espacio. Por ejemplo, el número de llamadas que recibe un centro de atención telefónica por minuto, el número de bacterias por centímetro cuadrado en una placa de Petri, el número de fallos en un tramo de línea telefónica, etc.

En estos casos, x representa el número de eventos que ocurren en un periodo de tiempo o un espacio dado, en los cuales podemos esperar un promedio de m de estos eventos.

Las características más importantes de esta distribución son:

1. El experimento consiste en contar el número de veces que se presenta un suceso en un intervalo de tiempo o espacio.

2. Los eventos ocurren al azar y son independientes de otros.

3. La probabilidad de que ocurra un suceso en un intervalo muy pequeño es proporcional a la longitud del intervalo.

Para calcular la distribución de probabilidad de Poisson:

Sea µ el número promedio de veces que ocurre un evento en cierto tiempo o espacio. La probabilidad de k sucesos de este evento es:

$$P(x = k) = \frac{\mu^k e^{-\mu}}{k!}$$

Donde k tiene los valores : k= 0, 1, 2, 3, ….

¿Qué es un histograma?

Un histograma es una representación gráfica de una variable en forma de barras, donde la superficie de cada barra es proporcional a la frecuencia de los valores representados. Sirven para obtener una "primera vista" general, o panorama, de la distribución de la población, o de la muestra, respecto a una característica, cuantitativa y continua (como la longitud o el peso). Los histogramas son una forma gráfica de presentar una tabla de frecuencias, con la que se puede obtener información acerca de la localización de la media y la desviación típica.

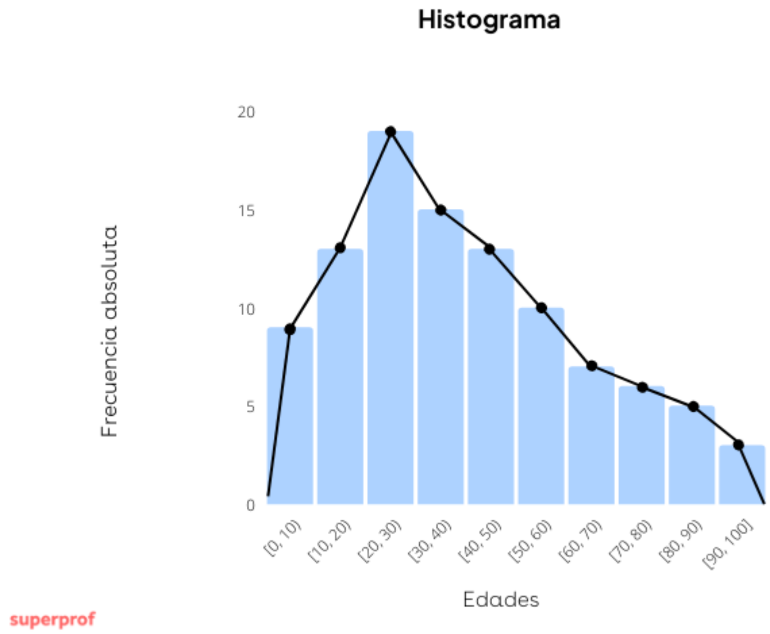


Imagen 1 Representación de un histograma.

¿Qué es un diagrama de cajas (boxplot)?

El diagrama de cajas es un tipo de gráfico que muestra un resumen de una gran cantidad de datos numéricos. Esta nos da información sobre la mediana, los cuartiles, los valores atípicos y los límites de los datos.

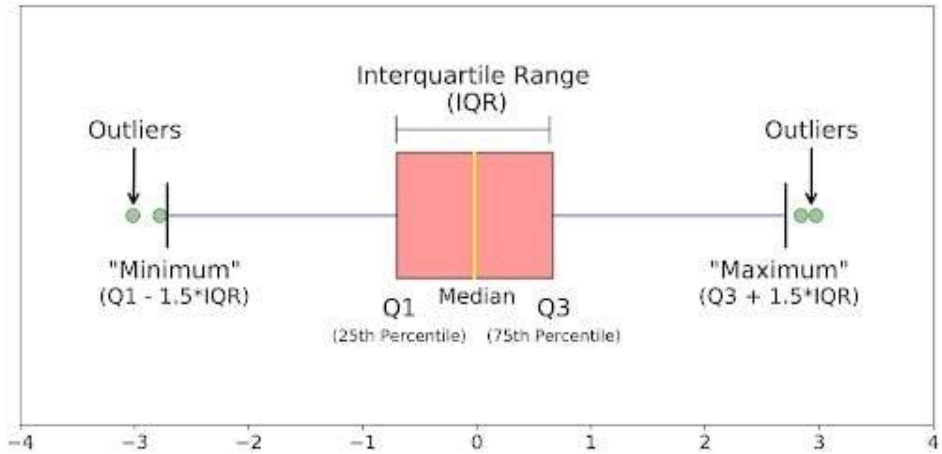


Imagen 2 Representación de un diagrama de cajas

- **Primer cuartil (Q1):** es el valor que deja a su izquierda al 25% de los datos ordenados de menos a mayor.

- **Mediana o Segundo Cuartil (Q2):** Divide en dos partes iguales la distribución. De forma que deja el 50% de los datos a su izquierda.
- **Tercer cuartil:** deja el 75% de los valores que son menores.
- **Rango Intercuartílico (RIC):** Diferencia entre el valor del tercer cuartil y el primer cuartil (RI= Q3-Q1)

Definiciones de mediana y moda.

La mediana y la moda son conceptos estadísticos utilizados para describir características de un conjunto de datos. A continuación, se presentan las definiciones de ambos términos:

La mediana es el valor que divide a un conjunto de datos en dos partes iguales, es decir, el valor central. Para calcular la mediana, se deben seguir los siguientes pasos:

Ordenar los datos de forma ascendente o descendente.

Si la cantidad de datos es impar, la mediana es el valor que ocupa la posición central.

Si la cantidad de datos es par, la mediana se obtiene calculando el promedio de los dos valores centrales.

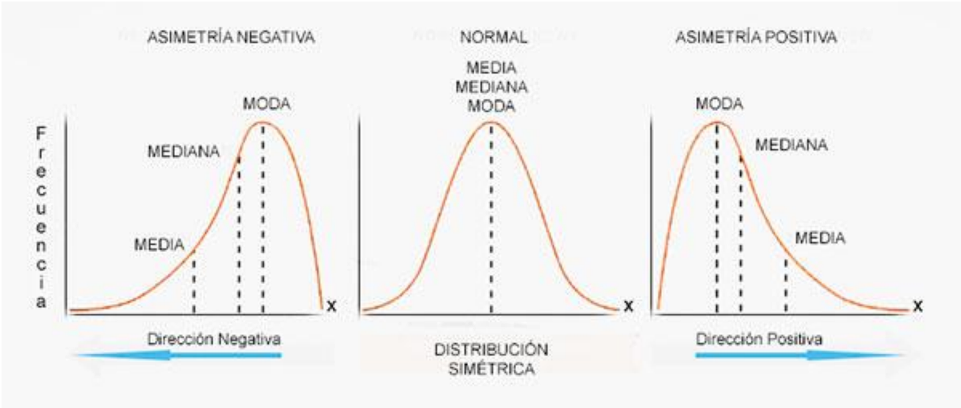
La mediana es una medida de tendencia central que no se ve afectada por valores extremos o atípicos en los datos. Se utiliza especialmente cuando los datos presentan una distribución sesgada o cuando se desea conocer el valor central representativo de un conjunto.

La moda es el valor que aparece con mayor frecuencia en un conjunto de datos. En otras palabras, es el valor que se repite con mayor frecuencia. Para calcular la moda, se deben seguir los siguientes pasos:

Identificar cuál o cuáles son los valores que se repiten con mayor frecuencia en el conjunto de datos.

Si no hay valores que se repitan, se dice que no existe una moda en el conjunto.

La moda es una medida útil para describir la frecuencia o la ocurrencia más común de un determinado valor en un conjunto de datos. A diferencia de la mediana y la media, la moda no está influenciada por los valores extremos y puede ser utilizada en cualquier tipo de distribución de datos.



¿Qué es la media empírica y la esperanza teórica?

La media empírica, también conocida como la media muestral, es una medida estadística utilizada para estimar la media de una población basándose en una muestra de datos. Se calcula sumando todos los valores de la muestra y dividiendo el resultado entre el tamaño de la muestra.

La fórmula para calcular la media empírica es la siguiente:

Media empírica = Suma de los valores de la muestra / Tamaño de la muestra

Por ejemplo, supongamos que se toma una muestra de 50 estudiantes y se registra su altura. Los valores de altura de la muestra se suman y luego se dividen entre 50 para obtener la media empírica de la altura de los estudiantes en la muestra.

Es importante destacar que la media empírica es una estimación de la media de la población y puede variar de una muestra a otra. Cuanto mayor sea el tamaño de la muestra, mayor precisión tendrá la estimación de la media poblacional.

La media empírica se utiliza ampliamente en la estadística y es una medida fundamental para resumir y describir una muestra de datos. Además, puede servir como punto de referencia para comparar con otros valores y realizar inferencias sobre la población subyacente.

La esperanza teórica, también conocida como valor esperado o esperanza matemática, es un concepto fundamental en la teoría de la probabilidad y la estadística. Representa el valor promedio que se espera obtener de una variable aleatoria en un experimento probabilístico, tomando en cuenta las probabilidades asociadas a cada posible resultado.

Matemáticamente, la esperanza teórica de una variable aleatoria discreta se calcula como la suma ponderada de los posibles valores multiplicados por sus respectivas probabilidades. En el caso de una variable aleatoria continua, se realiza una integral en lugar de una suma.

La esperanza teórica se denota generalmente como E(X), donde X es la variable aleatoria. Si X es una variable discreta con valores x1, x2, ..., xn y probabilidades correspondientes p1, p2, ..., pn, entonces la esperanza teórica se calcula de la siguiente manera:

$E(X) = x_1 * p_1 + x_2 * p_2 + ... + x_n * p_n$

En el caso de una variable aleatoria continua con función de densidad de probabilidad f(x), la esperanza teórica se calcula como:

$E(X) = \int(x * f(x))\, dx$

La esperanza teórica es un concepto importante en el análisis de datos y permite obtener una medida resumida de una variable aleatoria. Se utiliza para realizar cálculos y tomar decisiones basadas en la probabilidad y la incertidumbre asociadas a una variable aleatoria.

¿Qué es la varianza empírica y la varianza teórica?

La varianza empírica y la varianza teórica son conceptos utilizados en estadística para medir la dispersión o variabilidad de un conjunto de datos. A continuación, se presentan las definiciones de ambos términos:

La varianza empírica, también conocida como varianza muestral, es una medida de dispersión calculada a partir de un conjunto de datos observados. Se utiliza para estimar la varianza de una población más amplia cuando solo se dispone de una muestra limitada de datos. Para calcular la varianza empírica, se deben seguir los siguientes pasos:

Calcular la media aritmética de los datos.

Restar la media a cada uno de los valores individuales y elevar al cuadrado el resultado.

Sumar los resultados obtenidos en el paso anterior.

Dividir la suma entre el número de datos menos uno.

La varianza empírica es una medida que refleja la variabilidad de los datos observados y se utiliza para hacer inferencias sobre la varianza de toda la población.

La varianza teórica, también conocida como varianza poblacional, es una medida de dispersión que describe la variabilidad en una población completa o teórica. A diferencia de la varianza empírica, la varianza teórica se calcula utilizando todos los datos de la población en lugar de solo una muestra limitada. Sin embargo, en la práctica, la varianza poblacional es rara vez conocida y generalmente se estima utilizando la varianza empírica de una muestra.

La varianza teórica es una medida que proporciona información sobre la variabilidad inherente en una población y se utiliza en diversos análisis y modelos estadísticos.

Código del ejercicio:

Disponible en GitHub en https://github.com/jumpert/PyE_Aplicada_Tarea3.git

```
# Grupo: Estefany Clara, Gonzalo Paz, Juan Pérez y Lucas Cordero

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import binom
from scipy.stats import geom
from scipy.stats import poisson
from prettytable import PrettyTable
import tkinter as tk
from tkinter import messagebox
import os
ventana = tk.Tk()
ventana.withdraw()

print("Tarea 3 - Probabilidad y Estadística Aplicada\n")
mensaje = "DISTRIBUCIÓN 1: \n\nbinomial de parámetros n = 100 y p =
0,35."
messagebox.showinfo("Mensaje", mensaje)
ventana.destroy()

print(mensaje +"\n")
n, p = 100, 0.35

## Parte 1
# a) Generar 4 muestras de tamaño 100, 1000, 10000 y 100000
r2 = binom.rvs(n, p, size=100)
r3 = binom.rvs(n, p, size=1000)
r4 = binom.rvs(n, p, size=10000)
r5 = binom.rvs(n, p, size=100000)
tipos = [r2, r3, r4, r5]

# b) Boxplot de cada muestra
for tipo in tipos:
    median = np.median(tipo)
    Q1 = np.quantile(tipo, 0.25)
    Q3 = np.quantile(tipo, 0.75)
    min_val = np.min(tipo)
    max_val = np.max(tipo)
```

```
    # Rango intercuartílico
    RI = Q3 - Q1
    BII = Q1 - 1.5*RI
    BIS = Q3 + 1.5*RI
    # print('rango para detección de datos atípicos: {}, {}'.format(BII,
BIS))

    sns.boxplot(tipo)
    plt.ylabel('Valores')
    plt.title('Boxplot de {} muestras'.format(len(tipo)))
    try:
        plt.savefig('./binomial/boxplot_binomial_{}.png'.format(len(tipo)
))
    except:
        os.mkdir('./binomial')
        plt.savefig('./binomial/boxplot_binomial_{}.png'.format(len(tipo)
))
    plt.show()


# c) Histograma de cada muestra
for tipo in tipos:
    sns.histplot(tipo, kde=True)
    plt.ylabel('Frecuencia')
    plt.title('Histograma de {} muestras'.format(len(tipo)))
    plt.savefig('./binomial/hist_binomial_{}.png'.format(len(tipo)))
    plt.show()

# d) Calcular mediana y moda de cada muestra
tabla = PrettyTable()
tabla.field_names = ['Tamaño de muestra', 'Mediana', 'Moda']
for tipo in tipos:
    muestra = len(tipo)
    median = np.median(tipo)
    mode = np.argmax(np.bincount(tipo))
    tabla.add_row([muestra, round(median,2), mode])
print(tabla)
tabla.clear_rows()

# e) Hallar la media empírica de cada muestra y compararla con la
esperanza teórica de la distribución 1. ¿Qué se puede observar en las
muestras más grandes?
tabla.field_names = ['Tamaño de muestra', 'Media empírica', 'Media
teórica']
for tipo in tipos:
    muestra = len(tipo)
    media_empirica = np.mean(tipo)
    media_teorica = n*p
    tabla.add_row([muestra, round(media_empirica,2), media_teorica])
print(tabla)
tabla.clear_rows()

# f) Calcular la varianza empírica de cada muestra y compararla con la
varianza teórica de la distribución 1. ¿Qué se puede observar en las
muestras más grandes?
```

```python
tabla.field_names = ['Tamaño de muestra', 'Varianza empírica', 'Varianza teórica']
for tipo in tipos:
    muestra = len(tipo)
    varianza_empirica = np.var(tipo)
    varianza_teorica = n*p*(1-p)
    tabla.add_row([muestra, round(varianza_empirica,2), varianza_teorica])
print(tabla)
tabla.clear_rows()

print("\n==========================================================================================================\n")

## Parte 2
mensaje = "DISTRIBUCIÓN 2: \n\ngeométrica de parámetro p = 0,08."
print(mensaje)
messagebox.showinfo("Mensaje", mensaje)
p = 0.08

# a) Generar 4 muestras de tamaño 100, 1000, 10000 y 100000
r2 = geom.rvs(p, size=100)
r3 = geom.rvs(p, size=1000)
r4 = geom.rvs(p, size=10000)
r5 = geom.rvs(p, size=100000)
tipos = [r2, r3, r4, r5]

# b) Boxplot de cada muestra
for tipo in tipos:
    median = np.median(tipo)
    Q1 = np.quantile(tipo, 0.25)
    Q3 = np.quantile(tipo, 0.75)
    min_val = np.min(tipo)
    max_val = np.max(tipo)

    # Rango intercuartílico
    RI = Q3 - Q1
    BII = Q1 - 1.5*RI
    BIS = Q3 + 1.5*RI
    # print('rango para detección de datos atípicos: {}, {}'.format(BII, BIS))

    sns.boxplot(tipo)
    plt.ylabel('Valores')
    plt.title('Boxplot de {} muestras'.format(len(tipo)))
    try:
        plt.savefig('./geom/boxplot_geom_{}.png'.format(len(tipo)))
    except:
        os.mkdir('./geom')
        plt.savefig('./geom/boxplot_geom_{}.png'.format(len(tipo)))
    plt.show()

# c) Histograma de cada muestra
for tipo in tipos:
    sns.histplot(tipo, kde=True)
    plt.ylabel('Frecuencia')
```

```
    plt.title('Histograma de {} muestras'.format(len(tipo)))
    plt.savefig('./geom/hist_geom_{}.png'.format(len(tipo)))
    plt.show()

# d) Calcular mediana y moda de cada muestra
tabla.field_names = ['Tamaño de muestra', 'Mediana', 'Moda']
for tipo in tipos:
    muestra = len(tipo)
    median = np.median(tipo)
    mode = np.argmax(np.bincount(tipo))
    tabla.add_row([muestra, round(median,2), mode])
print(tabla)
tabla.clear_rows()

# e) Hallar la media empírica de cada muestra y compararla con la
esperanza teórica de la distribución 2. ¿Qué se puede observar en las
muestras más grandes?
tabla.field_names = ['Tamaño de muestra', 'Media empírica', 'Media
teórica']
for tipo in tipos:
    muestra = len(tipo)
    media_empirica = np.mean(tipo)
    media_teorica = (1-p)/p
    tabla.add_row([muestra, round(media_empirica,2), media_teorica])
print(tabla)
tabla.clear_rows()

# f) Calcular la varianza empírica de cada muestra y compararla con la
varianza teórica de la distribución 2. ¿Qué se puede observar en las
muestras más grandes?
tabla.field_names = ['Tamaño de muestra', 'Varianza empírica', 'Varianza
teórica']
for tipo in tipos:
    muestra = len(tipo)
    varianza_empirica = np.var(tipo)
    varianza_teorica = (1-p)/(p**2)
    tabla.add_row([muestra, round(varianza_empirica,2),
varianza_teorica])
print(tabla)
tabla.clear_rows()

print("\n=================================================================
=====================================\n")

## Parte 3
mensaje = "DISTRIBUCIÓN 3: \n\npoisson de parámetro λ = 30."
messagebox.showinfo("Mensaje", mensaje)
print(mensaje)
L = 30

# a) Generar 4 muestras de tamaño 100, 1000, 10000 y 100000
r2 = poisson.rvs(L, size=100)
r3 = poisson.rvs(L, size=1000)
r4 = poisson.rvs(L, size=10000)
r5 = poisson.rvs(L, size=100000)
tipos = [r2, r3, r4, r5]
```

```python
# b) Boxplot de cada muestra
for tipo in tipos:
    median = np.median(tipo)
    Q1 = np.quantile(tipo, 0.25)
    Q3 = np.quantile(tipo, 0.75)
    min_val = np.min(tipo)
    max_val = np.max(tipo)

    # Rango intercuartílico
    RI = Q3 - Q1
    BII = Q1 - 1.5*RI
    BIS = Q3 + 1.5*RI
    # print('rango para detección de datos atípicos: {}, {}'.format(BII,
BIS))

    sns.boxplot(tipo)
    plt.ylabel('Valores')
    plt.title('Boxplot de {} muestras'.format(len(tipo)))
    try:
        plt.savefig('./poisson/boxplot_pois_{}.png'.format(len(tipo)))
    except:
        os.mkdir('./poisson')
        plt.savefig('./poisson/boxplot_pois_{}.png'.format(len(tipo)))
    plt.show()

# c) Histograma de cada muestra
for tipo in tipos:
    sns.histplot(tipo, kde=True)
    plt.ylabel('Frecuencia')
    plt.title('Histograma de {} muestras'.format(len(tipo)))
    plt.savefig('./poisson/hist_pois_{}.png'.format(len(tipo)))
    plt.show()

# d) Calcular mediana y moda de cada muestra
tabla.field_names = ['Tamaño de muestra', 'Mediana', 'Moda']
for tipo in tipos:
    muestra = len(tipo)
    median = np.median(tipo)
    mode = np.argmax(np.bincount(tipo))
    tabla.add_row([muestra, round(median,2), mode])
print(tabla)
tabla.clear_rows()

# e) Hallar la media empírica de cada muestra y compararla con la
esperanza teórica de la distribución 3. ¿Qué se puede observar en las
muestras más grandes?
tabla.field_names = ['Tamaño de muestra', 'Media empírica', 'Media
teórica']
for tipo in tipos:
    muestra = len(tipo)
    media_empirica = np.mean(tipo)
    media_teorica = L
    tabla.add_row([muestra, round(media_empirica,2), media_teorica])
print(tabla)
tabla.clear_rows()
```

```
# f) Calcular la varianza empírica de cada muestra y compararla con la
varianza teórica de la distribución 3. ¿Qué se puede observar en las
muestras más grandes?
tabla.field_names = ['Tamaño de muestra', 'Varianza empírica', 'Varianza
teórica']
for tipo in tipos:
    muestra = len(tipo)
    varianza_empirica = np.var(tipo)
    varianza_teorica = L
    tabla.add_row([muestra, round(varianza_empirica,2),
varianza_teorica])
print(tabla)
tabla.clear_rows()

mensaje = "FIN DEL PROGRAMA.\n\nPodrá encontrar los gráficos en la
carpeta del programa. \nY los resultados en la consola."
messagebox.showinfo("Mensaje", mensaje)

## Solicitar al usuario que ingrese FIN para finalizar el programa, sino
seguirá ejecutándose
while True:
    fin = input("Ingrese SALIR para finalizar el programa: ").upper()
    if fin == "SALIR":
        break
    else:
        print("Ingrese FIN para finalizar el programa.")
        continue
```

Parte 1: Distribución Binomial

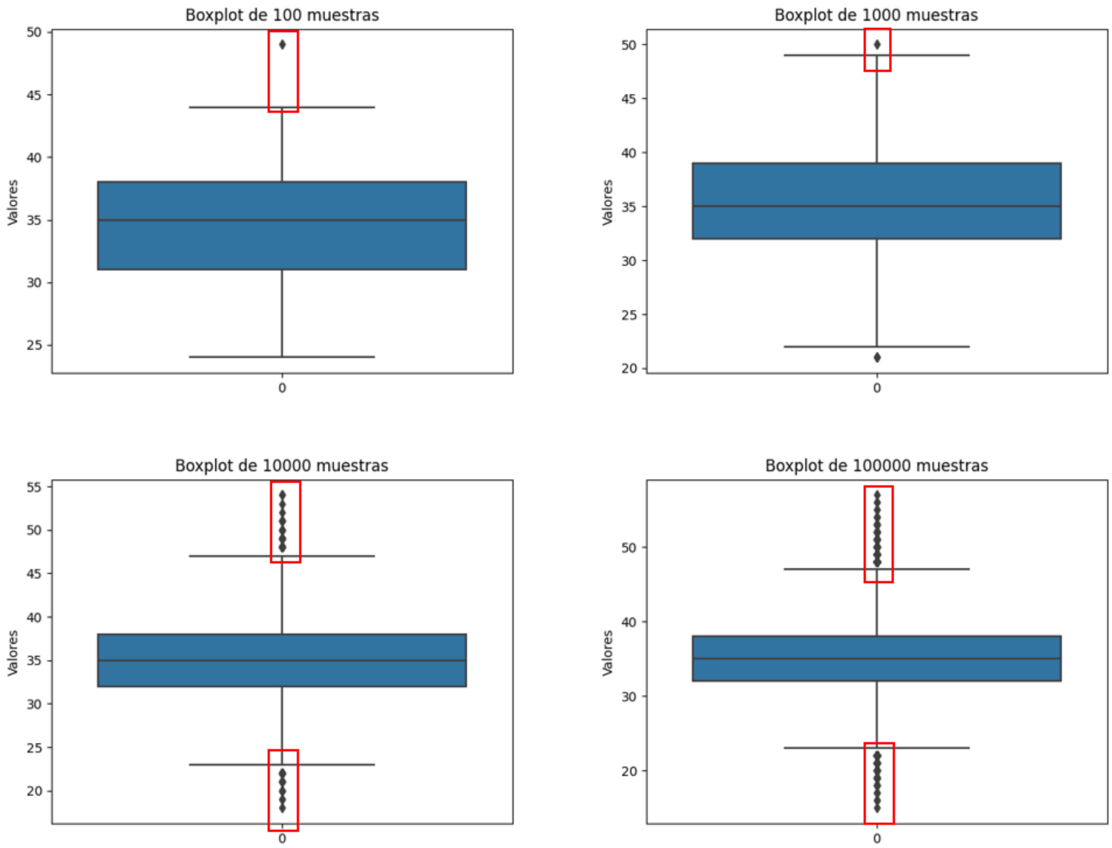
a) Generar muestras aleatorias de tamaño 10^2, 10^3, 10^4, 10^5.

```
n, p = 100, 0.35

## Parte 1
# a) Generar 4 muestras de tamaño 100, 1000, 10000 y 100000
r2 = binom.rvs(n, p, size=100)
r3 = binom.rvs(n, p, size=1000)
r4 = binom.rvs(n, p, size=10000)
r5 = binom.rvs(n, p, size=100000)
tipos = [r2, r3, r4, r5]
```

La imagen superior es la representación de cómo se generaron las muestras en el código realizado por el equipo. Se puede apreciar que ya que manejábamos datos con exponentes cada muestra tiene como nombre de variable "r" por recurso y el número del exponente asociado a cada tamaño de muestra.

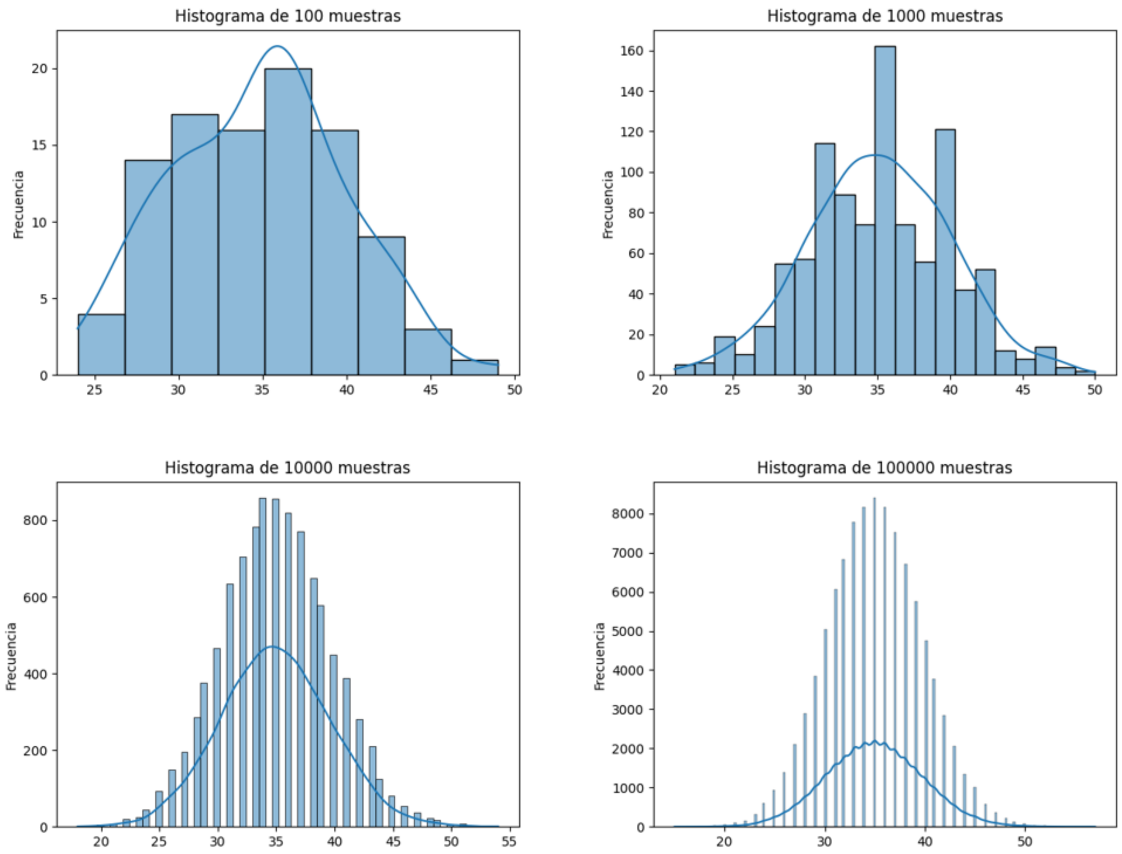
b) Hacer un diagrama de cajas para cada una de las muestras generadas en la parte anterior. ¿Existen datos atípicos en las muestras?



Imágenes del programa creado, en los recuadros rojos se marcan los datos atípicos.

Se puede observar gracias al diagrama de cajas que a medida que las muestras aumentan, también lo hacen la cantidad de datos atípicos. Existe cierta tendencia a que los valores atípicos sean en la parte superior si tomamos en cuenta que la distribución binomial tiene un coeficiente teórico de 0.5 de probabilidad de éxito en las pruebas normales.

c) Realizar un histograma de las muestras generadas.



Imágenes de la representación en Histogramas de las muestras del sistema.

d) Hallar la mediana y la moda de cada muestra.

```
binomial de parámetros n = 100 y p = 0,35.

+--------------------+---------+------+
| Tamaño de muestra | Mediana | Moda |
+--------------------+---------+------+
|        100         |  35.0   |  37  |
|        1000        |  35.0   |  35  |
|       10000        |  35.0   |  35  |
|       100000       |  35.0   |  35  |
+--------------------+---------+------+
```

e) Hallar la media empírica de cada muestra y compararla con la esperanza teórica de la distribución 1. ¿Qué se puede observar en las muestras más grandes?

Tamaño de muestra	Media empírica	Media teórica
100	35.04	35.0
1000	35.06	35.0
10000	34.99	35.0
100000	35.0	35.0

Al incrementar la cantidad de muestras generadas, se puede observar cómo la media empírica se acerca de manera más precisa hacia el valor de la media teórica. Esta convergencia se fundamenta en el principio estadístico de la ley de los grandes números, que establece que a medida que se aumenta la muestra, la media de los valores observados se aproxima a la media poblacional real. En este caso, al generar más muestras, estamos obteniendo una mejor estimación de la media teórica, lo cual nos brinda mayor confianza en los resultados obtenidos.

f) Hallar la varianza empírica de cada muestra y compararla con la varianza teórica de la distribución 2. ¿Qué se puede observar en las muestras más grandes?

Tamaño de muestra	Varianza empírica	Varianza teórica
100	17.94	22.75
1000	24.15	22.75
10000	23.27	22.75
100000	22.64	22.75

A medida que aumentamos la cantidad de muestras, se observa una aproximación de la varianza empírica hacia la varianza teórica. Sin embargo, es importante destacar que esta aproximación tiene un grado de exactitud menor.

Esta discrepancia en la exactitud se debe a la variabilidad inherente en la muestra y a posibles valores atípicos al ser generados de manera aleatoria. Aunque el aumento de muestras permite obtener una estimación más precisa de la varianza teórica, siempre existirá un margen de error debido a la aleatoriedad presente en la generación de los datos.

Es importante comprender que la varianza empírica se calcula a partir de una muestra limitada, mientras que la varianza teórica representa la variabilidad real en la población completa. Por lo tanto, aunque la varianza empírica se aproxime a la teórica al incrementar las muestras, es natural que exista una diferencia en la exactitud debido a la naturaleza limitada de la muestra.

Parte 2: Distribución Geométrica

a) Generar muestras aleatorias de tamaño 10^2, 10^3, 10^4, 10^5.

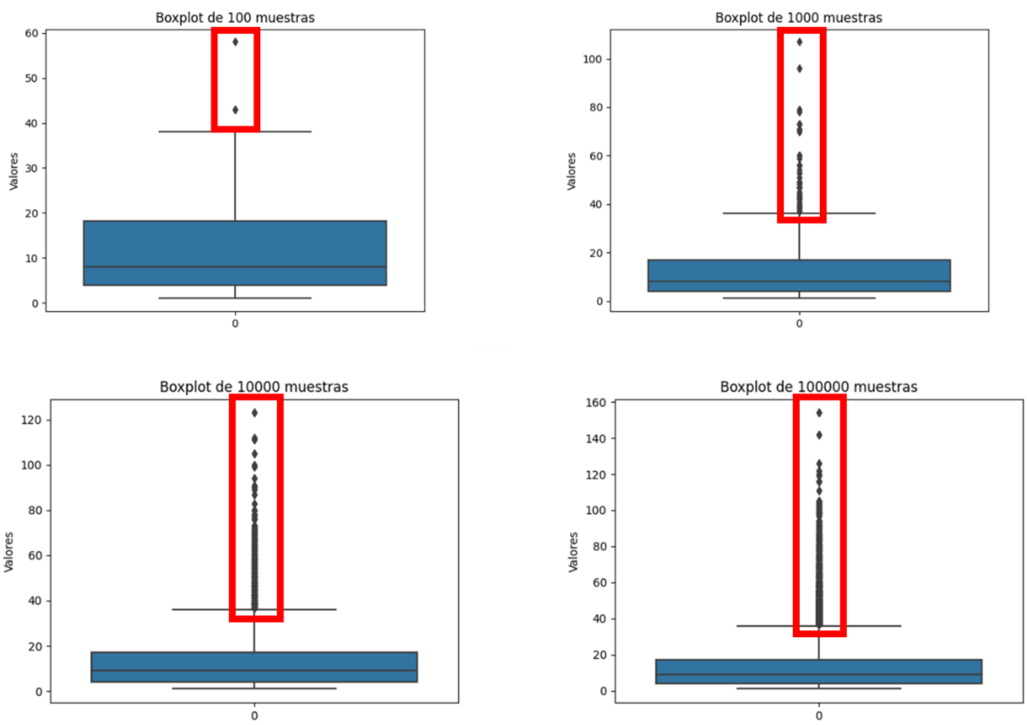
```
p = 0.08

# a) Generar 4 muestras de tamaño 100, 1000, 10000 y 100000
r2 = geom.rvs(p, size=100)
r3 = geom.rvs(p, size=1000)
r4 = geom.rvs(p, size=10000)
r5 = geom.rvs(p, size=100000)
tipos = [r2, r3, r4, r5]
```

La imagen superior es la representación de cómo se generaron las muestras en el código realizado por el equipo. Se puede apreciar que ya que manejábamos datos con exponentes cada muestra tiene como nombre de variable "r" por recurso y el número del exponente asociado a cada tamaño de muestra.

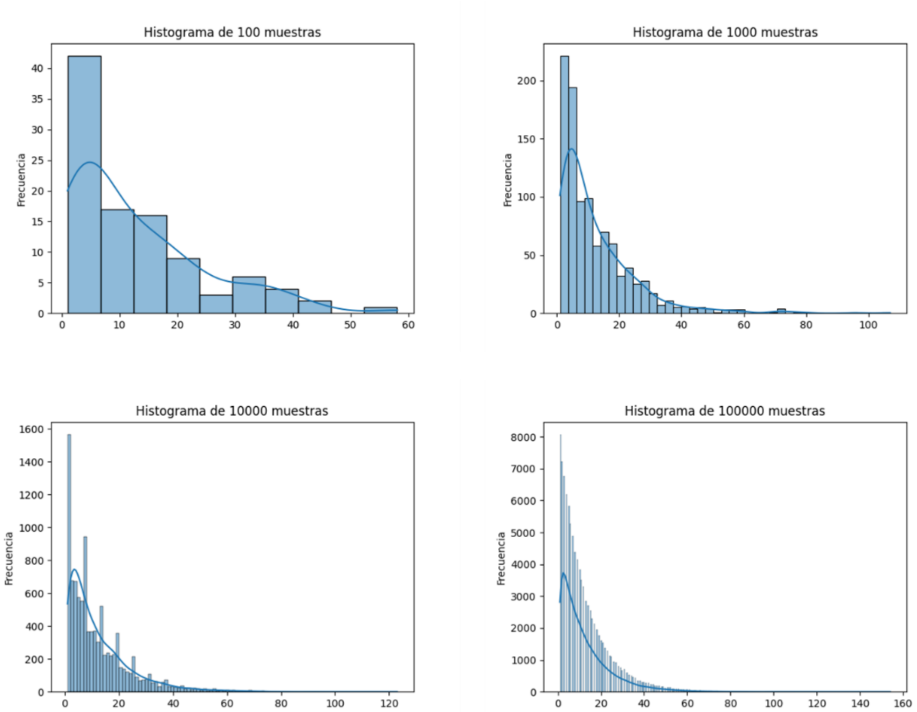
b) Hacer un diagrama de cajas para cada una de las muestras generadas en la parte anterior. ¿Existen datos atípicos en las muestras?

Podemos ver datos atípicos representados en los diagramas de cajas.



Imágenes del programa creado, en los recuadros rojos podemos ver los datos atípicos.

c) Realizar un histograma de las muestras generadas.



Capturas de pantalla de los histogramas generados por el programa

d) Hallar la mediana y moda de cada muestra

```
DISTRIBUCIÓN 2:

geométrica de parámetro p = 0,08.
+------------------+---------+------+
| Tamaño de muestra | Mediana | Moda |
+------------------+---------+------+
|        100        |   10.0  |  1   |
|        1000       |   8.0   |  1   |
|       10000       |   9.0   |  1   |
|       100000      |   9.0   |  1   |
+------------------+---------+------+
```

En esta captura del programa podemos ver los resultados de la mediana y la moda de cada uno de los tamaños de la muestra.

e) Hallar la media empírica de cada muestra y compararla con la esperanza teórica de la distribución 2. ¿Qué se puede observar en las muestras más grandes?

Tamaño de muestra	Media empírica	Media teórica
100	12.7	11.5
1000	11.91	11.5
10000	12.43	11.5
100000	12.46	11.5

Como vimos anteriormente la media empírica es una medida que se basa en la muestra de datos; cuando los valores son más grandes se tienden a aproximar con la media teórica siendo está basada en un modelo estadístico, no se utilizan los datos observados, sino que se utiliza la fórmula matemática.

f) Hallar la varianza empírica de cada muestra y compararla con la varianza teórica de la distribución 2. ¿Qué se puede observar en las muestras más grandes?

Tamaño de muestra	Varianza empírica	Varianza teórica
100	135.63	143.75
1000	139.44	143.75
10000	136.23	143.75
100000	142.99	143.75

Cómo vimos anteriormente la varianza empírica nos muestra la dispersión de los datos en relación con la media, siendo esta una estimación de la variación de los datos en la muestra.

A medida que las muestras aumentan se pueden ver que las estimaciones son más precisas y se acercan a la varianza teórica, hay una estabilidad de las estimaciones debido a que serán más similares entre sí.

Parte 3: Distribución Poisson

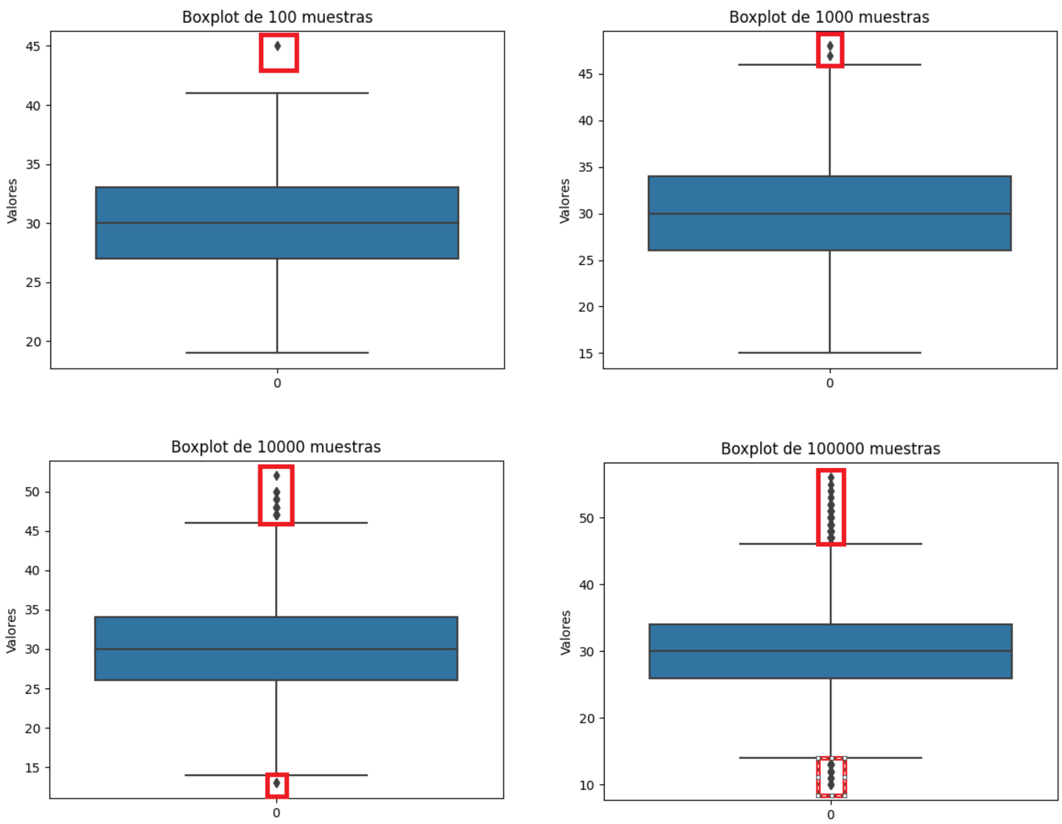
a) Generar muestras aleatorias de tamaño 10^2, 10^3, 10^4 y 10^5

```python
L = 30

# a) Generar 4 muestras de tamaño 100, 1000, 10000 y 100000
r2 = poisson.rvs(L, size=100)
r3 = poisson.rvs(L, size=1000)
r4 = poisson.rvs(L, size=10000)
r5 = poisson.rvs(L, size=100000)
tipos = [r2, r3, r4, r5]
```

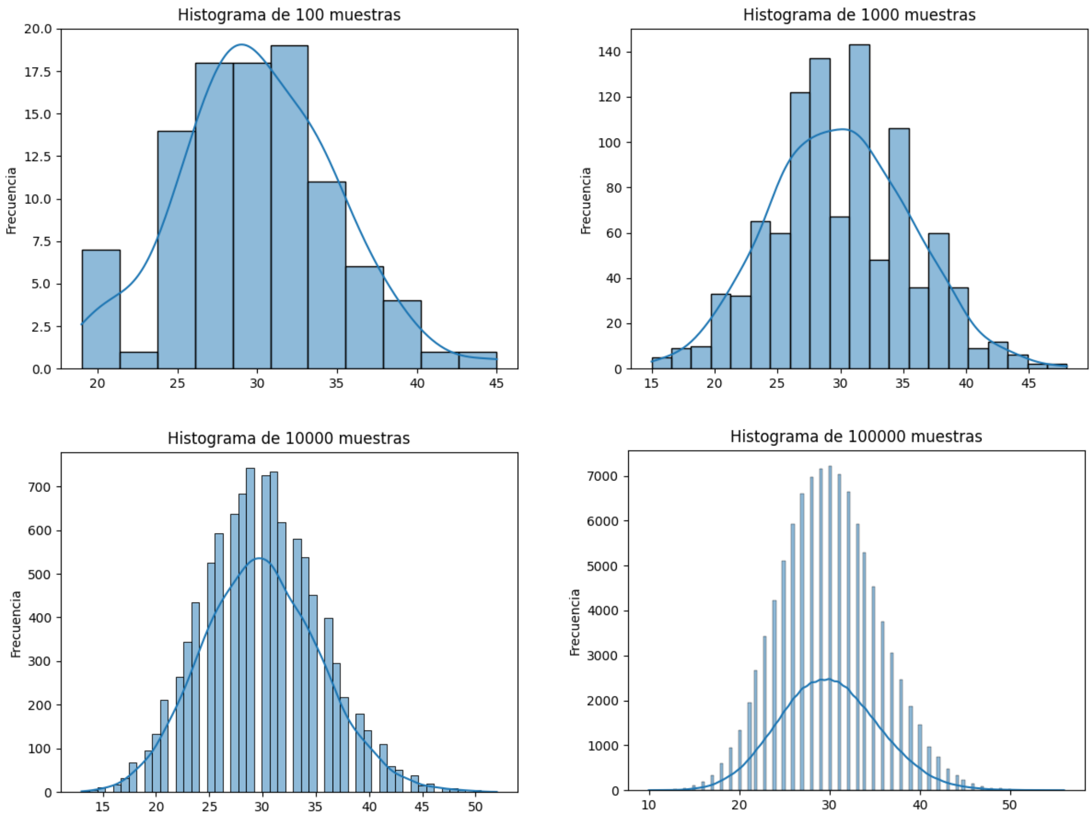
La imagen superior es la representación de cómo se generaron las muestras en el código realizado por el equipo. Se puede apreciar que ya que manejábamos datos con exponentes cada muestra tiene como nombre de variable “r” por recurso y el número del exponente asociado a cada tamaño de muestra.

b) Hacer un diagrama de cajas para cada una de las muestras generadas en la parte anterior. ¿Existen datos atípicos en las muestras?



Se puede ver la aparición de datos atípicos desde el primer diagrama de cajas con 100 muestras, también se puede observar que conforme aumenta la cantidad de muestras, también lo hace la cantidad de datos atípicos, tanto del lado superior como del lado inferior

c) Realizar un histograma de las muestras generadas.



d) Hallar la mediana y la moda de cada muestra.

poisson de parámetro λ = 30.

Tamaño de muestra	Mediana	Moda
100	30.0	28
1000	30.0	28
10000	30.0	29
100000	30.0	30

e) Hallar la media empírica de cada muestra y compararla con la esperanza teórica de la distribución 3. ¿Qué se puede observar en las muestras más grandes?

Tamaño de muestra	Media empírica	Media teórica
100	29.98	30
1000	30.0	30
10000	29.93	30
100000	29.98	30

Si bien en este ejemplo donde el caso donde la media empírica y la media teórica son exactas se puede apreciar que hay una mayor precisión con la muestra n = 100000 si lo comparamos con n = 10000

Hallar la varianza empírica de cada muestra y compararla con la varianza teórica de la distribución 3. ¿Qué se puede observar en las muestras más grandes?

Tamaño de muestra	Varianza empírica	Varianza teórica
100	23.36	30
1000	30.78	30
10000	30.15	30
100000	30.12	30

Como se mencionó anteriormente, a medida que aumenta el tamaño de la muestra, el valor de la varianza empírica se aproxima al de la varianza teórica. Como podemos observar en la tabla anterior, la diferencia entre la varianza empírica y teórica con muestra n = 100 es de casi 7 puntos, mientras que el valor se reduce a 0.12 para n = 100000.

Bibliografía:

Superprof.es. (s.f.). Histograma. Recuperado el 3 de junio de 2023, de https://www.superprof.es/apuntes/escolar/matematicas/estadistica/descriptiva/histograma.html

Builtin.com. (s.f.). Boxplot. Recuperado el 3 de junio de 2023, de https://builtin.com/data-science/boxplot

PGConocimiento.com. (s.f.). Diagrama de Caja y Bigotes (Boxplot). Recuperado el 3 de junio de 2023, de

https://www.pgconocimiento.com/diagrama-boxplot/#:~:text=El%20Diagrama%20de%20Caja%20y,valores%20at%C3%ADpicos%20y%20comparar%20distribuciones

William Mendenhall, R. J. (2010). *Introducción a la probabilidad y estadística.* México, D.F: Cenagage Learning.