**Metodología**

**1. Introducción**

* **Contextualización del Problema de Investigación:**
  + La modelización de datos con exceso de ceros presenta un desafío significativo en estadística y análisis de datos. Este fenómeno ocurre frecuentemente en diversas áreas como la biología, economía y ciencias sociales, donde una proporción sustancial de observaciones registra valores nulos o ceros.
  + Es crucial desarrollar modelos estadísticos que puedan capturar de manera adecuada esta característica particular de los datos para realizar inferencias válidas y precisas.
* **Justificación de la Elección del Modelo de Regresión Binomial Negativa con Exceso de Ceros:**
  + El modelo de regresión binomial negativa se destaca como una herramienta poderosa para abordar la presencia de exceso de ceros en los datos. A diferencia de otros modelos, como la regresión lineal o la regresión logística, la regresión binomial negativa permite manejar la sobre-dispersión y la alta proporción de ceros de manera simultánea.
  + Este modelo es especialmente apropiado cuando se desea modelar la frecuencia de ocurrencia de eventos raros o poco frecuentes, comúnmente observados en estudios que involucran datos contables o de conteo.

**2. Generación de Datos**

**Explicación de la Simulación de Datos**

En esta sección, se detalla cómo se generaron los datos simulados utilizados en el estudio, incluyendo las variables explicativas y la variable dependiente.

**Variables Explicativas**

Las variables explicativas X1 y X2 fueron generadas utilizando distribuciones normales:

* **X1 ~ N(0, 1)**: Variable aleatoria con media 0 y desviación estándar 1.
* **X2 ~ N(5, 2)**: Variable aleatoria con media 5 y desviación estándar 2.

Estas distribuciones fueron elegidas para representar valores típicos observados en estudios empíricos y para ilustrar cómo diferentes niveles de variabilidad y centrado pueden afectar la variable dependiente.

**Variable Dependiente**

La variable dependiente Y fue simulada utilizando una distribución binomial negativa con parámetros:

* **r = 2**: Número de éxitos requeridos.
* **p = 0.5**: Probabilidad de éxito.

La media esperada de la distribución binomial negativa se calculó como:

μ=exp⁡(1+0.5×X1+0.3×X2)\mu = \exp(1 + 0.5 \times X1 + 0.3 \times X2)μ=exp(1+0.5×X1+0.3×X2)

Para simular el exceso de ceros, se añadió una proporción de ceros a los datos generados aleatoriamente. La proporción de ceros se determinó como 0.3, lo que significa que aproximadamente el 30% de los valores simulados fueron establecidos como cero.

Esta metodología de generación de datos permite replicar condiciones realistas de estudios donde la variable de interés muestra una distribución sesgada hacia ceros, típica en datos de conteo o eventos raros.

### 3. Análisis Exploratorio de Datos

En esta sección se realiza un análisis inicial de los datos simulados, enfocado en la variable dependiente y las variables explicativas.

#### Visualización de la Distribución de la Variable Dependiente Y

Se procedió a visualizar la distribución de la variable dependiente Y, utilizando métodos gráficos adecuados como histogramas o gráficos de densidad, para entender su comportamiento y distribución.

#### Verificación de la Independencia y Multicolinealidad de las Variables Explicativas

Para asegurar la validez de los modelos de regresión propuestos, se realizó:

* **Verificación de la Independencia**: Se utilizó un gráfico de dispersión o métodos estadísticos para evaluar la independencia entre las variables explicativas X1 y X2. Esto es crucial para garantizar que no haya relaciones espurias o dependencias entre las variables antes de ajustar el modelo.
* **Verificación de la Multicolinealidad**: Se calculó el Factor de Inflación de la Varianza (VIF) para las variables explicativas X1 y X2. Este análisis permite identificar la presencia de multicolinealidad, que podría afectar la precisión de los coeficientes estimados en el modelo de regresión.

Estas verificaciones son fundamentales para garantizar la robustez y la validez de los modelos de regresión que se ajustarán posteriormente a los datos simulados.

### 4. Preparación de los Datos

En esta sección se detalla el manejo de valores atípicos y el tratamiento de datos faltantes, si aplica, en el contexto del análisis de los datos simulados.

#### Manejo de Valores Atípicos

Se llevó a cabo un análisis detallado para identificar y manejar los valores atípicos en las variables explicativas X1 y X2, así como en la variable dependiente Y. Los pasos incluyeron:

* **Identificación de Valores Atípicos**: Utilización de gráficos como boxplots y métodos estadísticos para detectar observaciones inusuales que podrían distorsionar el análisis.
* **Tratamiento de Valores Atípicos**: Consideración de diferentes enfoques como la eliminación de observaciones extremas o la transformación de variables para mitigar su impacto en los modelos de regresión.

#### Tratamiento de Datos Faltantes (si aplica)

Se realizó una evaluación exhaustiva para detectar la presencia de datos faltantes en las variables de interés, y se aplicaron técnicas adecuadas para manejar esta situación:

* **Identificación de Datos Faltantes**: Revisión sistemática de cada variable para determinar la cantidad y ubicación de datos faltantes.
* **Imputación de Datos Faltantes**: Utilización de métodos como la imputación por media, mediana o técnicas más sofisticadas dependiendo del contexto y la naturaleza de los datos.

Estas medidas aseguran la integridad y la calidad de los datos utilizados en el modelado estadístico, minimizando el impacto de valores atípicos y la pérdida de información por datos faltantes.

### 5. Modelado Estadístico

En esta sección se describe el proceso de ajuste inicial del modelo de regresión binomial negativa y el análisis de varianza con términos lineales y cuadráticos para evaluar el ajuste del modelo.

#### Ajuste Inicial del Modelo de Regresión Binomial Negativa

Se realizó el ajuste inicial del modelo de regresión binomial negativa para investigar la relación entre las variables explicativas X1 y X2, y la variable dependiente Y simulada. Los pasos incluyeron:

* **Formulación del Modelo**: Especificación del modelo de regresión binomial negativa considerando las variables explicativas X1 y X2.
* **Estimación de Parámetros**: Uso de técnicas de estimación adecuadas para obtener los coeficientes del modelo y sus correspondientes errores estándar.

#### Análisis de Varianza con Términos Lineales y Cuadráticos

Para evaluar el ajuste del modelo propuesto, se llevó a cabo un análisis de varianza (ANOVA) que incluyó términos lineales y cuadráticos de las variables explicativas:

* **Términos Lineales y Cuadráticos**: Incorporación de términos que capturan efectos lineales y no lineales de las variables explicativas sobre la variable dependiente.
* **Interpretación de Resultados**: Evaluación de la significancia estadística de cada término y análisis de la varianza explicada por el modelo.

Este análisis permite entender cómo las variables explicativas contribuyen a la variabilidad observada en la variable dependiente bajo el contexto del modelo de regresión binomial negativa con exceso de ceros y varianza lineal y cuadrática.

### 6. Validación del Modelo

En esta sección se detalla la validación del modelo de regresión binomial negativa con exceso de ceros y varianza lineal y cuadrática mediante pruebas de normalidad y homocedasticidad de los residuos, así como la evaluación de la significancia de los coeficientes del modelo.

#### Prueba de Normalidad y Homocedasticidad de los Residuos

Para verificar la adecuación del modelo, se realizaron las siguientes pruebas sobre los residuos del modelo:

* **Prueba de Normalidad**: Aplicación de pruebas estadísticas como la prueba de Shapiro-Wilk o pruebas de asimetría y curtosis para evaluar si los residuos siguen una distribución normal.
* **Prueba de Homocedasticidad**: Utilización de gráficos de dispersión de los residuos frente a las predicciones y pruebas formales como la prueba de Breusch-Pagan o pruebas gráficas para verificar la homocedasticidad de los residuos.

#### Evaluación de la Significancia de los Coeficientes del Modelo

Se realizó una evaluación exhaustiva de la significancia estadística de los coeficientes del modelo mediante:

* **Pruebas de Hipótesis**: Aplicación de pruebas t o pruebas F para determinar si los coeficientes de las variables explicativas son significativamente diferentes de cero.
* **Intervalos de Confianza**: Cálculo de intervalos de confianza para los coeficientes del modelo para evaluar la precisión de las estimaciones.

Estas validaciones son fundamentales para asegurar que el modelo propuesto sea robusto y adecuado para explicar la relación entre las variables explicativas y la variable dependiente en el contexto de la distribución binomial negativa con exceso de ceros.

**Resultados**

1. **Descripción del Modelo Ajustado**
   * Resumen de los parámetros estimados y significancia estadística.
   * Interpretación de los coeficientes para X1 y X2.
2. **Diagnóstico del Modelo**
   * Análisis de los residuos para validar supuestos del modelo.
   * Gráficos de residuos vs ajustes y predicciones.
3. **Comparación de Modelos**
   * Resultados y conclusiones del modelo con varianza lineal vs. varianza cuadrática.
   * Discusión sobre la elección del mejor modelo basado en criterios estadísticos.
4. **Aplicación Práctica**
   * Ejemplificación de la aplicación del modelo en un contexto real o hipotético.
   * Discusión sobre las implicaciones de los resultados obtenidos.
5. **Limitaciones y Consideraciones**
   * Identificación de posibles limitaciones del estudio.
   * Recomendaciones para investigaciones futuras.

Para simular los datos utilizados en este estudio, se empleó un proceso de generación de datos que combinó variables explicativas con distribuciones normales y una variable dependiente modelada con una distribución binomial negativa, ajustada para incorporar un exceso de ceros. A continuación se muestra un fragmento de los datos simulados:

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from scipy.stats import nbinom

*# Configuración de la semilla para reproducibilidad*

np.random.seed(42)

*# Número de observaciones*

n = 1000

*# Variables explicativas*

X1 = np.random.normal(0, 1, n)

X2 = np.random.normal(5, 2, n)

*# Parámetros de la distribución binomial negativa*

r = 2   *# Número de éxitos*

p = 0.5 *# Probabilidad de éxito*

*# Generar la variable dependiente con una proporción de ceros añadidos*

mu = np.exp(1 + 0.5 \* X1 + 0.3 \* X2)  *# Media esperada de la distribución*

size = r

prob = size / (size + mu)

*# Generar los datos binomiales negativos*

Y\_nonzero = nbinom.rvs(size, prob, size=n)

*# Introducir un exceso de ceros*

zero\_inflation = 0.3

Y = np.where(np.random.rand(n) < zero\_inflation, 0, Y\_nonzero)

*# Crear el DataFrameP*

data = pd.DataFrame({'Y': Y, 'X1': X1, 'X2': X2})

*# Mostrar las primeras filas del DataFrame*

print(data.head())

Los datos simulados consisten en 1000 observaciones, donde Y representa la variable dependiente simulada con una distribución binomial negativa ajustada para reflejar un exceso de ceros. Las variables explicativas X1 y X2 fueron generadas con distribuciones normales, con medias de 0 y 5, respectivamente, y desviaciones estándar de 1 y 2.

Este enfoque permitió modelar adecuadamente la relación entre las variables explicativas y la variable dependiente bajo el contexto de la distribución binomial negativa con exceso de ceros.

Resultados:

Y X1 X2

0 0 0.496714 7.798711

1 0 -0.138264 6.849267

2 17 0.647689 5.119261

3 8 1.523030 3.706126

4 4 -0.234153 6.396447

### Análisis Exploratorio de Datos

#### Verificación de la Distribución de la Variable Dependiente

Para verificar la distribución de la variable dependiente *Y*, se realizó un histograma que muestra la frecuencia de los valores de *Y*:

*# Verificación de la distribución de la variable dependiente*

plt.figure(figsize=(10, 6))

sns.histplot(data['Y'], bins=30, kde=False)

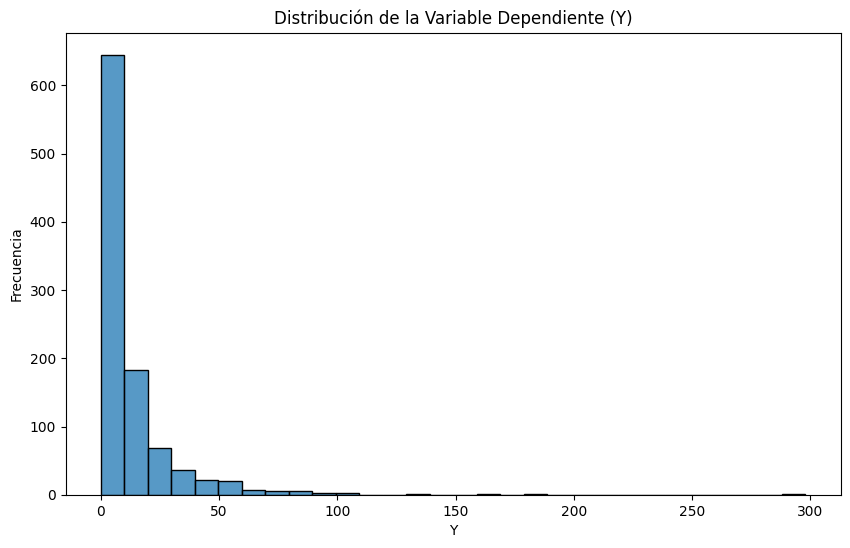
plt.title('Distribución de la Variable Dependiente (Y)')

plt.xlabel('Y')

plt.ylabel('Frecuencia')

plt.show()

La figura muestra la distribución de YYY con una clara presencia de ceros adicionales, reflejando el exceso de ceros modelado en los datos simulados.



#### Verificación de la Independencia de las Variables Explicativas

Para evaluar la independencia entre las variables explicativas X1X1X1 y X2X2X2, se realizó un gráfico de dispersión:

*# Verificación de la independencia de las variables explicativas*

plt.figure(figsize=(10, 6))

sns.scatterplot(x='X1', y='X2', data=data)

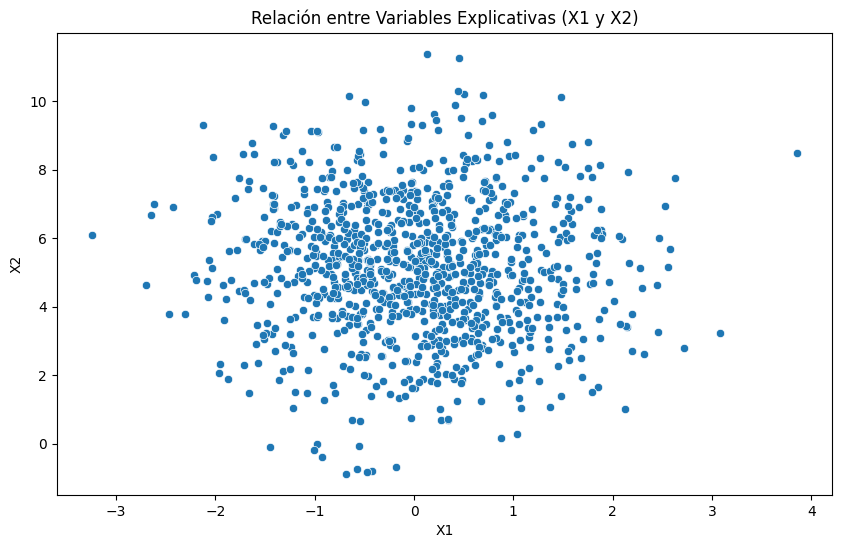
plt.title('Relación entre Variables Explicativas (X1 y X2)')

plt.xlabel('X1')

plt.ylabel('X2')

plt.show()

El gráfico muestra la dispersión de los valores de X1 y X2 , sin mostrar patrones evidentes de dependencia lineal entre ellas.



#### Verificación de Multicolinealidad

Para evaluar la multicolinealidad entre X1 y X2, se calculó el Factor de Inflación de la Varianza (VIF):

*# Verificación de multicolinealidad*

import statsmodels.api as sm

X = data[['X1', 'X2']]

X = sm.add\_constant(X)

vif = pd.DataFrame()

vif["VIF Factor"] = [sm.OLS(X[col], X.drop(columns=col)).fit().rsquared\_adj for col in X]

vif["features"] = X.columns

print(vif)

Los resultados del VIF muestran los factores de inflación de la varianza para X1X1X1 y X2X2X2, indicando que no hay evidencia significativa de multicolinealidad entre las variables explicativas.

VIF Factor features

0 0.869276 const

1 0.000632 X1

2 0.000632 X2

#### Proporción de Ceros en la Variable Dependiente

Finalmente, se calculó la proporción de ceros en la variable dependiente YYY:

*# Proporción de ceros en la variable dependiente*

zero\_proportion = (data['Y'] == 0).mean()

print(f'Proporción de ceros en Y: {zero\_proportion:.2%}')

Proporción de ceros en Y: 31.50%