Universidad Nacional de Rosario

Instituto Politécnico Superior General San Martín

Analista Universitario en Sistemas

Análisis de Variables Aleatorias

Estudio Estadístico del Dataset Olist

Alumnos:

Emanuel Duarte

Stefano Mazziotta

Candela Viola

Profesora:

Alejandra Zorzi

Cátedra de Probabilidad y Estadística

23 de julio de 2025

Índice

1. Introducción y Selección de Datos

El presente trabajo analiza el dataset "Brazilian E-Commerce Public Dataset by Olist", una base de datos real que contiene 112,650 pedidos realizados entre 2016-2018 en Brasil. Se seleccionaron dos variables para el estudio estadístico:

Variable Discreta: Cantidad de ventas por día (X)

Variable Continua: Valor total de pedidos (Y)

La elección se fundamenta en la relevancia para sistemas de información y la aplicabilidad directa en la toma de decisiones operativas de e-commerce.

2. Definiciones Formales

Variable Aleatoria Discreta X:
X = [Cantidad de ventas por día] donde $X : \Omega \to \{0, 1, 2, 3, ...\}$
$X \sim \text{Poisson}(\lambda = 180{,}06)$

Variable Aleatoria Continua Y:
Y = [Valor total del pedido (R$)] donde $Y : \Omega \to \mathbb{R}^+$
Y representa la suma del precio de productos más costos de envío.

3. Análisis de Variable Discreta

3.1. Parámetros Obtenidos

Del procesamiento de la base de datos:

- Días analizados: 612
- Ventas totales: 110,197 ítems
- $\lambda = \frac{110{,}197}{612} = 180{,}06$ ventas/día

3.2. Estadísticas Descriptivas

Para $X \sim \text{Poisson}(\lambda = 180{,}06)$:

$$E[X] = \lambda = 180{,}06 \text{ ventas/día} \tag{1}$$

$$\text{Var}(X) = \lambda = 180{,}06 \tag{2}$$

$$\sigma_X = \sqrt{\lambda} = 13{,}42 \text{ ventas/día} \tag{3}$$

$$\text{Moda} = \lfloor \lambda \rfloor = 180 \text{ ventas/día} \tag{4}$$

3.3. Tabla de Frecuencias

Rango	Frecuencia	Porcentaje
1–50	38	6.2 %
51–100	97	15.8 %
101–150	123	20.1 %
151–200	120	19.6 %
201–250	90	14.7 %
251–300	74	12.1 %
301–350	48	7.8 %
351–400	12	2.0 %
401–450	7	1.1 %
451–500	1	0.2 %
Más de 500	2	0.3 %
Total	**612**	**100.0 %**

3.4. Visualización

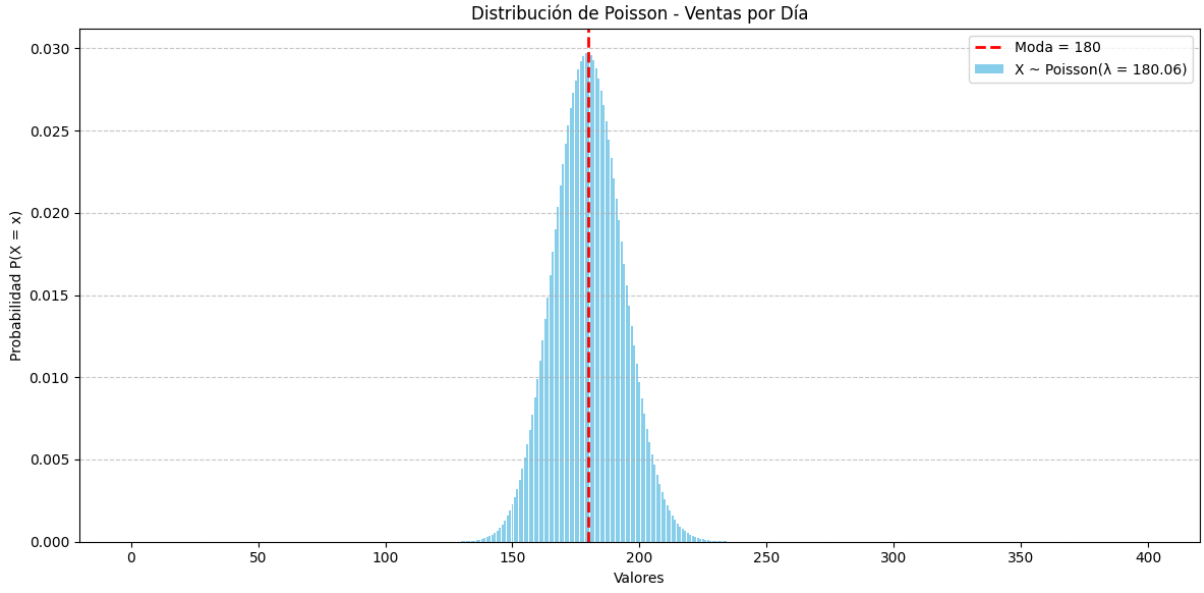

Figura 1: Distribución de Poisson: cantidad de ventas diarias con $\lambda = 180{,}06$. La línea roja punteada indica la moda en 180.

4. Análisis de Variable Continua

4.1. Tratamiento de Outliers

Se utilizó el percentil 100 % para el análisis, un total de 96,478 registros. Esto permite analizar la distribución completa incluyendo outliers, con valores que van desde R$ 9.59 hasta R$ 13,664.08.

4.2. Estadísticas Descriptivas

Para la variable completa Y:

$$\bar{Y} = 159{,}83 \text{ R\$} \tag{5}$$

$$s_Y = 218{,}79 \text{ R\$} \tag{6}$$

$$\text{Mediana} = 105{,}28 \text{ R\$} \tag{7}$$

$$\text{Rango} = [9{,}59, 13664{,}08] \text{ R\$} \tag{8}$$

$$\text{Coeficiente de Variación (CV)} = \frac{s_Y}{\bar{Y}} = 136{,}9\,\% \tag{9}$$

Percentiles:

- P25 = R$ 61.85
- P75 = R$ 176.26
- P95 = R$ 446.23

4.3. Regla de Sturges

Aplicando $k = \lceil \log_2(n) + 1 \rceil$ con $n = 96{,}478$:

$$k = \lceil \log_2(96{,}478) + 1 \rceil = 18 \text{ intervalos}$$

Ancho de intervalo: $\frac{13{,}664{,}08 - 9{,}59}{18} = 758{,}58$ R$

4.4. Visualización

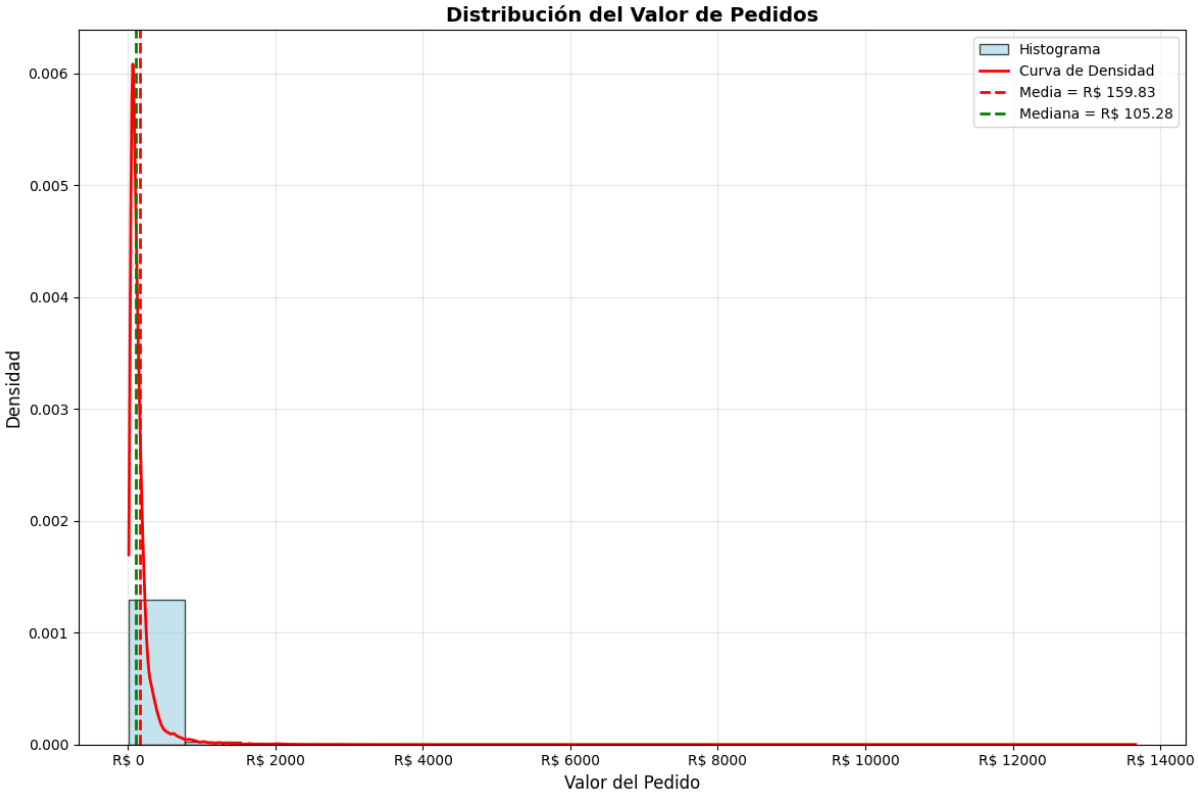

Figura 2: Distribución del valor de pedidos. Histograma normalizado con curva de densidad suavizada. Las líneas verticales muestran la media (R$ 159.83) en rojo punteado y la mediana (R$ 105.28) en verde punteado. Se observa una fuerte asimetría positiva con outliers extremos.

La Figura 2 muestra claramente la distribución asimétrica de los valores de pedidos, donde la mayoría se concentra en el rango inferior mientras que una pequeña proporción presenta valores extremadamente altos. La diferencia entre la media y la mediana evidencia la presencia de outliers que sesgan la distribución hacia la derecha.

5. Distribuciones de Probabilidad

5.1. Aproximación sin Integrar

Para la variable continua Y, la probabilidad se aproxima usando frecuencias relativas:

$$P(a \leq Y \leq b) \approx \sum_{i:[x_i,x_{i+1}[\cap[a,b]\neq\emptyset} \frac{f_i}{n} \cdot \Delta x_i$$

donde f_i es la frecuencia en el intervalo i y Δx_i el ancho del intervalo.

Ejemplo práctico:

$$P(Y \leq 500) = \frac{92,404}{96,478} \approx 0,9578$$

Esto significa que aproximadamente el 95.78 % de los pedidos tienen un valor menor o igual a R$ 500.

5.2. Comparación con Distribución Normal

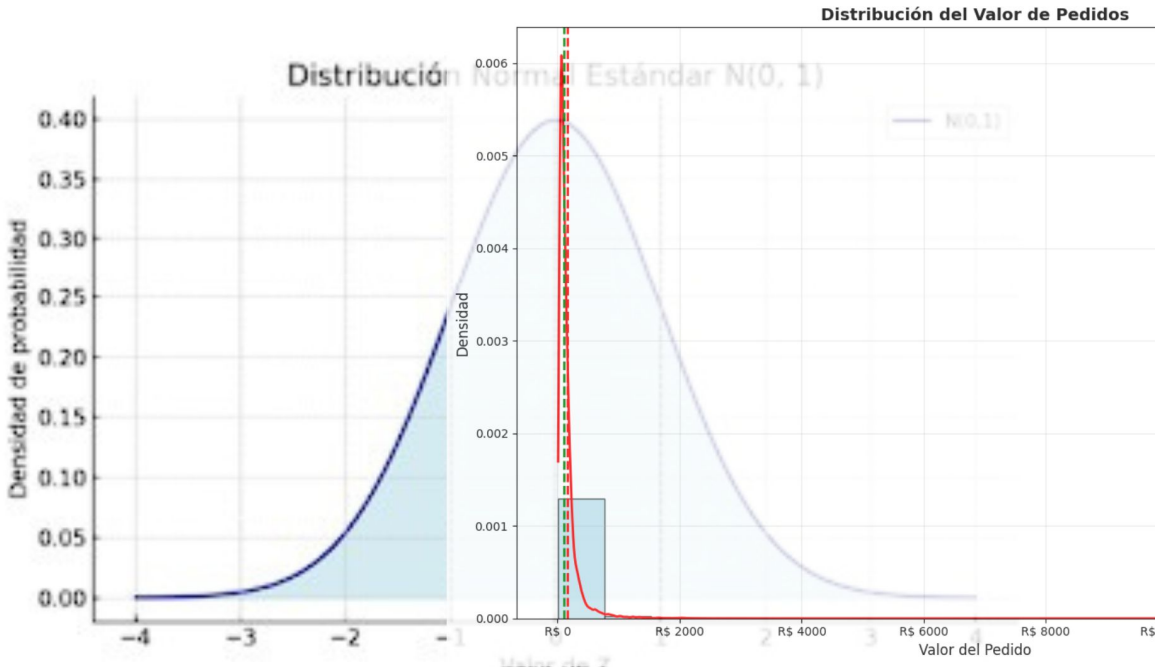

Figura 3: Comparación entre distribución normal y histograma de densidad de la variable aleatoria continua.

La distribución empírica presenta:

- Fuerte asimetría positiva (media R$ 159.83 >mediana R$ 105.28)
- Cola derecha muy extendida (outliers hasta R$ 13,664.08)

6. Conclusión

El análisis de variables discretas y continuas permitió aplicaciones prácticas como la predicción de demanda ($P(X > 200) = 1 - P(X \leq 200)$), planificación de inventarios basada en $\lambda = 180{,}06$ y dimensionamiento de recursos. En variables continuas, se aplicaron estrategias de *pricing* según la mediana (R$ 105,28), segmentación de clientes por valor de compra y análisis de rentabilidad por rangos. No obstante, hay limitaciones: los datos históricos (2016–2018) pueden no reflejar patrones actuales, existen *outliers* extremos (hasta R$ 13.664,08), el coeficiente de variación es alto (136,9 %) y no se consideran factores estacionales ni promocionales.

7. Diccionario de Conceptos

Variable Aleatoria: Función que asigna un valor numérico a cada resultado de un experimento aleatorio.

Distribución de Poisson: Distribución discreta que modela el número de eventos en un intervalo fijo de tiempo. $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$

Media (μ o $E[X]$): Valor esperado o promedio teórico de una variable aleatoria.

Varianza (σ^2 o Var(X)): Medida de dispersión que indica qué tan alejados están los valores de la media.

Desviación Estándar (σ): Raíz cuadrada de la varianza, en las mismas unidades que la variable.

Moda: Valor más probable en una distribución discreta o el punto de máxima densidad en una continua.

Mediana: Valor que divide la distribución en dos partes iguales (percentil 50).

Coeficiente de Variación (CV): $\frac{\sigma}{\mu} \times 100\,\%$. Mide la variabilidad relativa.

Percentil: Valor que deja un porcentaje específico de datos por debajo.

Regla de Sturges: Fórmula para determinar el número óptimo de intervalos en un histograma: $k = \lceil \log_2(n) + 1 \rceil$

Asimetría: Medida que indica si la distribución es simétrica o está sesgada hacia un lado.

Outliers: Valores atípicos que se alejan significativamente del patrón general de los datos.

Función de Densidad de Probabilidad (PDF): Función $f(x)$ tal que $P(a \leq X \leq b) = \int_a^b f(x)\,dx$

Distribución Normal: Distribución continua simétrica caracterizada por su media μ y desviación estándar σ.

8. Referencias

a. Olist. (2018). *Brazilian E-Commerce Public Dataset by Olist*. Disponible en Kaggle.

b. Código fuente.

c. Material de la Cátedra.

d. Documentación de SciPy: scipy.org.

e. Cheatsheets de Matplotlib: matplotlib.org/cheatsheets.

f. Documentación de Pandas: pandas.pydata.org/docs.

g. Página oficial de NumPy: numpy.org.