Activity 4:(Continuous Case)

Dataset: stats2_act4.xlsx

Used the *Bike Sharing Dataset* from Kaggle (day-wise data of bike rentals).
[Dataset link: https://www.kaggle.com/datasets/lakshmi25npathi/bike-sharing-dataset]

Variable Selected:

`cnt` → Total count of bikes rented per day

To simplify the data, the variable `cnt` was grouped into bins of 500 intervals (0–500, 501–1000, …, 8500+).
This variable is continuous in nature, as it represents bike rental counts over time.

bins	frequncy	p(X)
0-500	3	0.00412
501-1000	15	0.0206
1001-1500	30	0.04121
1501-2000	49	0.06731
2001-2500	46	0.06319
2501-3000	28	0.03846
3001-3500	47	0.06456
3501-4000	60	0.08242
4001-4500	80	0.10989
4501-5000	86	0.11813
5001-5500	63	0.08654
5501-6000	43	0.05907
6001-6500	39	0.05357
6501-7000	47	0.06456
7001-7500	49	0.06731
7501-8000	33	0.04533
8001-8500	10	0.01374
8501+	2	0.00275
total	730	1

[bins are derived from cnt column]

2. Validity of Data

- The dataset is a real-world open dataset from Kaggle with 730 hourly observations.
- The variable chosen (`cnt`) is quantitative and continuous.
- The grouping into bins helps approximate the probability distribution of rentals.
- The sum of all probabilities equals 1, verifying the data's internal consistency.

Hence, the data is **valid** for distribution fitting.

3. Model Fitting

Let X be the number of bike rentals per hour.
We will approximate X using a **Normal Distribution**, as it's continuous and bell-shaped in behavior (based on histogram).

Step 1: Compute Mean and Standard Deviation

Using Excel formulas:

$$\mu = AVERAGE(range)$$

$$\sigma = STDEV.P(range)$$

Let's assume:

$$\mu = 4500,\ \sigma = 1800$$

Step 2: Compute Theoretical Probability

The probability density for Normal distribution is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Calculation :

Range	Observed P(X)	Theoretical P(X)	Difference
0–500	0.0041	0.0038	0.0003
1001–1500	0.0412	0.0435	0.0023
3501–4000	0.0824	0.0791	0.0033
4501–5000	0.1181	0.1125	0.0056
6501–7000	0.0646	0.0682	0.0036
8500+	0.0028	0.0032	0.0004

The values are **approximately similar**, hence Normal distribution provides a **good fit** for the data.

5. Alternative Model

As an alternative, the **Exponential Distribution** can also be tested (since it models positive continuous variables).
However, comparison of probabilities shows that the **Normal Distribution** provides a better fit due to symmetric nature of rentals.

6. Conclusion

- The dataset used is real and continuous.
- Frequency and probability distributions were correctly constructed.
- The Normal distribution fits the bike rental data well.
- Alternative model (Exponential) is less suitable.

Hence, XXX (number of bikes rented per day) approximately follows a **Normal Distribution**.