Part A – Theoretical Foundation

1. What is a Statistical Distribution?

Definition:

A **statistical distribution** describes how values of a variable are spread or distributed.

It shows:

- How often values occur
- The pattern of data
- Probability of outcomes

Types:

- **Discrete Distribution** → Countable values (0,1,2…)
- **Continuous Distribution** → Infinite possible values (real numbers)

Example:

- Transaction count per day → Discrete
- Transaction amount → Continuous

---

2. What is a Q-Q Plot?

Definition:

A **Q-Q (Quantile-Quantile) Plot** compares your dataset distribution with a theoretical distribution (like Normal).

Purpose:

- To check if data follows Normal distribution.

Interpretation:

- Points on straight line → Data is normally distributed
- Curved pattern → Not normal

---

3. Difference Between Discrete and Continuous Distribution

Feature	Discrete	Continuous
Values	Countable	Infinite
Example	Number of transactions	Transaction amount
Function	PMF	PDF

---

4. Bernoulli Distribution

Definition:

Used when there are only **two possible outcomes**.

Example:

- Success / Fail
- 1 / 0

Formula (PMF):

$$P(X = x) = p^x(1-p)^{1-x}$$

Where:

- p = probability of success
- x = 0 or 1

Example:

Transaction status:

- Success = 1
- Fail = 0

---

5. Binomial Distribution

Definition:

Used for number of successes in **n independent Bernoulli trials**.

Formula:

$$P(X = k) = \binom{n}{k} p^k(1-p)^{n-k}$$

Where:

- n = total trials
- k = successes
- p = probability of success

Example:

Number of successful transactions in a week.

---

6. Log-Normal Distribution

Definition:

If the logarithm of a variable is normally distributed, the variable follows a **Log-Normal distribution**.

Used For:

- Skewed data
- Income
- Transaction amounts

PDF Formula:

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

Example:

Transaction amount (usually right-skewed)

---

7. Power Law Distribution

Definition:

Describes data where small values are common and large values are rare but extreme.

Formula:

$$P(x) = Cx^{-\alpha}$$

Where:

- α = exponent
- C = constant

Example:

Few customers make very large purchases.

---

8. Box-Cox Transformation

Definition:

Used to stabilize variance and make data more normal.

Formula:

$$y(\lambda) = \begin{cases} \frac{x^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \ln(x), & \lambda = 0 \end{cases}$$

Why Used?

- Remove skewness
- Improve model performance

---

9. Poisson Distribution

Definition:

Models number of events occurring in fixed interval of time.

Formula:

$$P(X = k) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

Where:

- λ = average number of events

Example:

Number of transactions per day.

---

10. Z-Score Probability

Definition:

Measures how many standard deviations a value is from mean.

Formula:

$$Z = \frac{X - \mu}{\sigma}$$

Where:

- X = value
- μ = mean
- σ = standard deviation

Used For:

- Outlier detection
- Probability calculation

---

11. Difference between PDF and CDF

PDF (Probability Density Function)

- Used for continuous data.
- Area under curve = 1

CDF (Cumulative Distribution Function)

$$F(x) = P(X \leq x)$$

- Shows cumulative probability
- Increasing function from 0 to 1