Shape of Data Distribution

1. **What is a symmetric distribution in statistics? Give an example from daily life where data is symmetrically distributed.**
   A symmetric distribution has data evenly spread around the center, and measures of central tendency (mean, median, mode) are usually equal.
   Example: Heights of people.
2. **Define a skewed distribution. How might income data in a country represent skewness?**

   A **skewed distribution** is when data is not evenly spread around the mean.

   Example: Income data is usually **right-skewed** — most people earn average or low incomes, while a few earn extremely high incomes, pulling the distribution to the right.

3. **What does it mean if a data distribution is left-skewed? Provide a practical example**

   A **left-skewed distribution** means the tail is longer on the **left side**, with most values on the higher end.

   Example: Ages at retirement — most people retire around 60–65, but a few retire very early (30s or 40s), creating a left tail.

4. **What does it mean if a data distribution is right-skewed? Mention a situation where this might occur.**

   A **right-skewed distribution** means the tail is longer on the **right side**, with most values on the lower end.

   Example: Income distribution — most people earn low to average incomes, while a few earn very high incomes.

5. **How can you identify a skewed distribution by looking at a histogram of house prices in a city?**

   If the histogram of house prices has a **long tail to the right**, it's **right-skewed** (few very expensive houses).
   If it has a **long tail to the left**, it's **left-skewed** (few very cheap houses).

6. **What are the key characteristics of a normal distribution? How does it relate to human heights?**

Key characteristics of a normal distribution:

- Bell-shaped and symmetric.
- Mean = Median = Mode.
- Most data lies close to the mean.
- Follows the **68–95–99.7 rule** (about 68% within 1 SD, 95% within 2 SDs, 99.7% within 3 SDs).

Relation to human heights:
Human heights usually form a normal distribution — most people are near average height, with fewer very short or very tall individuals.

7. **Explain the empirical rule (68-95-99.7 rule) using the context of exam scores in a large class.**

The **empirical rule** says that in a normal distribution:

- **68%** of students' exam scores lie within **1 standard deviation** of the mean.
- **95%** lie within **2 standard deviations**.
- **99.7%** lie within **3 standard deviations**.

Meaning: Most students score around the class average, and very few score extremely low or high.

8. **Why is the normal distribution important in statistics and what are some real-world examples of it?**

The **normal distribution** is important because many natural and social phenomena follow it, and many statistical tests assume it. It helps in prediction, decision-making, and hypothesis testing.

Real-world examples:

- Human heights
- IQ scores
- Blood pressure
- Measurement errors
- Exam scores

9. **How does skewness affect the mean, median, and mode in distributions like employee salaries?**

Right-skewed (like salaries): Mean > Median > Mode

Left-skewed: Mean < Median < Mode

Symmetric: Mean = Median = Mode In salaries, a few very high incomes pull the mean upward, while the median better represents typical earnings.

10. **How do outliers affect the shape of a data distribution, for instance in tracking daily temperatures?**

Outliers stretch the distribution's **tails**, making it **skewed or uneven**.

Example (temperatures): If most days are 20–30°C but one day is 45°C, that outlier pulls the distribution to the **right**, making it right-skewed.

11. **How can you detect outliers in a dataset using visual tools like boxplots? Use an example of customer spending.**

In a **boxplot**, outliers appear as **points outside the "whiskers"** (usually 1.5× IQR from Q1 or Q3).

Example: Customer spending — most spend $50–$200, but a few spend $1000+; these high spenders show up as outliers above the whisker.

12. **What statistical methods can be used to handle outliers in survey data?**

Methods to handle outliers:

- Remove them if they are errors or irrelevant.
- Transform the data (log, square root) to reduce impact.
- Use robust statistics (median, IQR) instead of mean/SD.
- Cap or winsorize extreme values to a set percentile.

13. Describe a real-world example where a right-skewed distribution is observed, such as hospital visit durations.

Example: Hospital visit durations — most patients stay for a short time, but a few stay much longer, creating a **right-skewed distribution** with a long tail on the higher end.

14. How do skewed distributions affect data analysis in fields like e-commerce or finance?

Skewed distributions can **distort averages** and affect predictions:

- **E-commerce:** A few big purchases can make the average order value misleading.
- **Finance:** A few extremely high or low returns can skew risk assessments.

Analysts often use **median or transformations** to handle skewness.

15. What tools or Python libraries can be used to visualize skewness and detect outliers in large datasets?

Python tools/libraries:

- Matplotlib / Seaborn → histograms, boxplots, violin plots
- Pandas → df.describe(), df.boxplot()
- SciPy / NumPy → calculate skewness/kurtosis
- Plotly → interactive visualizations

These help see skewness and spot outliers in large datasets.