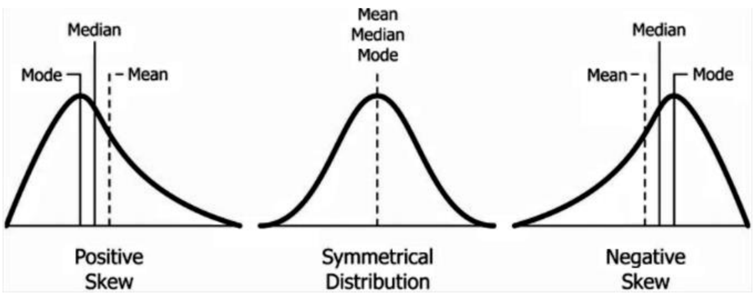
Chapter 7: Data Shape

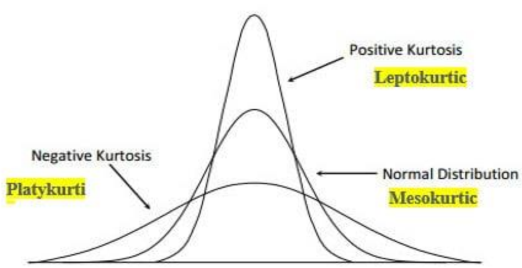
Data shape refers to the structure or arrangement of data within a dataset. Understanding the shape of data is crucial for analyzing its distribution, identifying trends, and determining suitable statistical methods for further analysis.

Key Aspects of Data Shape:

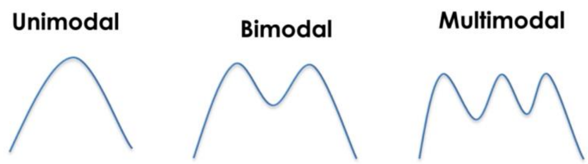
1. **Symmetry**:



   - Describes whether the data is evenly distributed around a central value.
   - Common measures include skewness.
2. **Peakedness (Kurtosis)**:



   - Refers to the height and sharpness of the peak in a data distribution.
   - Determines whether data has heavy tails (leptokurtic) or light tails (platykurtic).
3. **Spread (Variability)**:
   - Indicates how much data varies from the central value.
   - Measured by range, variance, and standard deviation.
4. **Modality**:



- Describes the number of peaks in the distribution.
  - Unimodal: One peak.
  - Bimodal: Two peaks.
  - Multimodal: More than two peaks.

Common Data Shapes:

1. **Symmetric Distribution**:
   - The left and right sides of the distribution mirror each other.
   - Example: Normal distribution (bell-shaped curve).
2. **Skewed Distribution**:
   - **Right (Positive) Skew**: The tail is longer on the right.
   - **Left (Negative) Skew**: The tail is longer on the left.
3. **Uniform Distribution**:
   - All values have approximately the same frequency.
   - Appears as a flat, rectangular shape.
4. **Exponential Distribution**:
   - A rapid rise followed by a gradual decline.
   - Often used to model waiting times or decay processes.
5. **Bimodal or Multimodal Distributions**:
   - Contains two or more peaks.
   - Often indicates a mix of two or more different populations.

Measures to Quantify Data Shape:

1. **Skewness**:

   S_k=(Mean – Mode)/ Standard Deviation

   Or

   S_k =3(Mean – Median)/ Standard Deviation

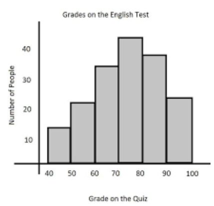
   - Quantifies the asymmetry of the data distribution.
     - Skewness >0: Right skew.
     - Skewness <0: Left skew.
     - Skewness =0: Symmetric distribution.

2. **Kurtosis**:

    - Measures the "tailedness" of the distribution.
        - Leptokurtic (>0): Heavy tails and sharp peak.
        - Platykurtic (<0): Light tails and flat peak.
        - Mesokurtic (=0): Normal distribution.
3. **Range and Variance**:
    - Provide insights into the spread of the data.

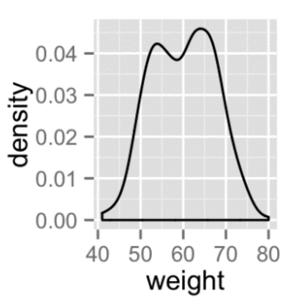
Visualizing Data Shape:

1. **Histogram**:

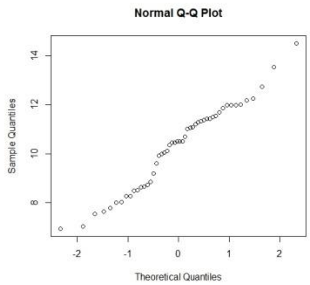


    - Shows the frequency distribution of data.
    - Helps identify symmetry, skewness, and modality.
2. **Boxplot**:

    - Highlights the spread, central tendency, and potential outliers.
3. **Density Plot**:



   - Smooth curve that estimates the probability density function of the data.
4. **QQ Plot**:



   - Compares the distribution of the data to a theoretical distribution (e.g., normal).

How to Interpret a QQ Plot

1. **Straight Line Alignment**:
   - If the data points align closely with the diagonal line, it suggests that the data follows the theoretical distribution.
2. **Deviations from the Line**:
   - Upward Curvature: Indicates heavy tails (data has more extreme values than expected, suggesting a leptokurtic distribution).
   - Downward Curvature: Indicates lighter tails (fewer extreme values than expected, suggesting a platykurtic distribution).
   - S-Shaped Curve: Suggests skewness in the data.
3. **Clustering or Gaps**:
   - Points clustering in certain regions or large gaps can indicate issues like outliers or non-uniform distribution.

Skewness Calculation

$$Sk = \frac{3(\text{Mean} - \text{Median})}{\sigma}$$

For example we have the dataset(x): **5, 6, 6, 8, 8, 8, 10, 15**

Step 1: Calculate the Mean

Mean=Sum of all values/Number of values = (5+6+6+8+8+8+10+15)/8 = 8.25

Median = (8+8) /2

Step 2: Calculate the Standard Deviation (σ)

$\sigma = \sqrt{\sum (x_i - \text{Mean})^2 / n}$

Step 3: Calculate Skewness (g1)

The formula is:

S_k = 3(Mean – Median)/ Standard Deviation

S_k = 3*(8.25-8)/ 2.5 = .75/2.5 = 0.30

Interpretation of Skewness

- Positive Skewness: A tail on the right (e.g., income distribution).
- Negative Skewness: A tail on the left (e.g., test scores with a ceiling effect).
- Skewness ~ 0: Symmetric distribution (e.g., normal distribution).

Kurtosis Calculation

$$K = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \text{Mean})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \text{Mean})^2\right)^2}$$

For example, consider the dataset(x): **2, 4, 6, 8, 10**

Step 1: Calculate the Mean

Mean = Sum of all values/ Number of values = (2+4+6+8+10)/5 = 6

Step 2: Calculate Deviations from the Mean

x_i−Mean = (−4,−2,0,2,4)

Step 3: Calculate Squared Deviations

$(xi-\text{Mean})^2$ = (16,4,0,4,16)

Step 4: Calculate Fourth Power of Deviations

$(xi-\text{Mean})^4$= (256,16,0,16,256)

Step 5: Calculate Variance

Variance = Sum of squared deviations/ n = (16+4+0+4+16)/5 = 8

Step 6: Compute Kurtosis

$$K = \frac{\frac{1}{n}\sum(x_i - \text{Mean})^4}{(\text{Variance})^2}$$

$$K = \frac{\frac{1}{5}(256 + 16 + 0 + 16 + 256)}{8^2} = \frac{\frac{1}{5}(544)}{64} = \frac{108.8}{64} \approx 1.7$$

Step 7: Adjust for Excess Kurtosis

Excess kurtosis: $K - 3 = 1.7 - 3 = -1.3$

Interpretation

- Excess Kurtosis > 0: Leptokurtic (sharp peak and heavy tails).
- Excess Kurtosis = 0: Mesokurtic (normal distribution).
- Excess Kurtosis < 0: Platykurtic (flat peak and light tails).