Summarizing and Visualizing Data

- Priya Bhatia

Data Scientist and Mentor

Types of Data

- Numerical
  - Center
    - Mean or Median or Mode
  - Shape
    - Bell-shaped or Skewed
  - Spread
    - Range or IQR or Variance
- Categorical
  - Proportion or Count or Mode

Measures of Central Tendency

- Mean - an average of data
- Median - middle value of the ordered data
- Mode - value that occurs most often in the data

Mean vs Median

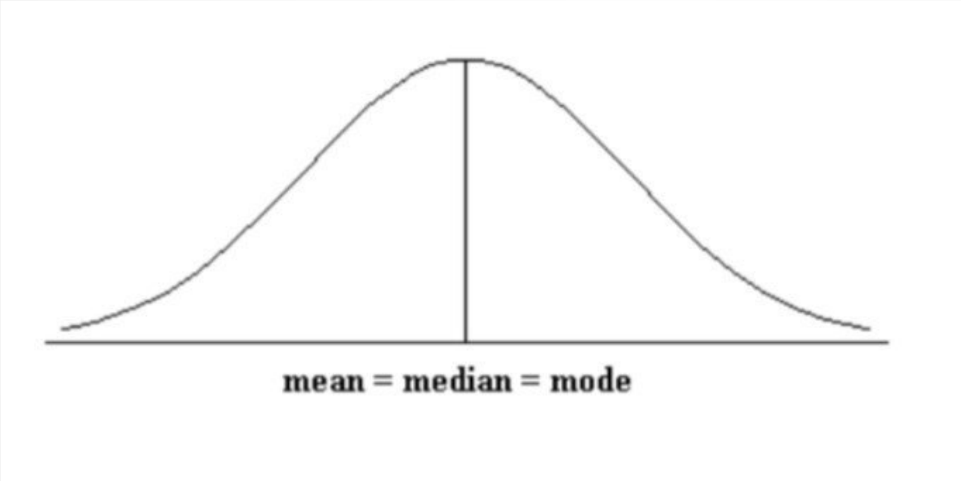
Consider seven employees' salaries as follows:

- 28,000
- 34,000
- 33,000
- 37,000
- 33,000
- 40,000
- 40,000

Question: When it is better to report the Median as compared to the Mean?

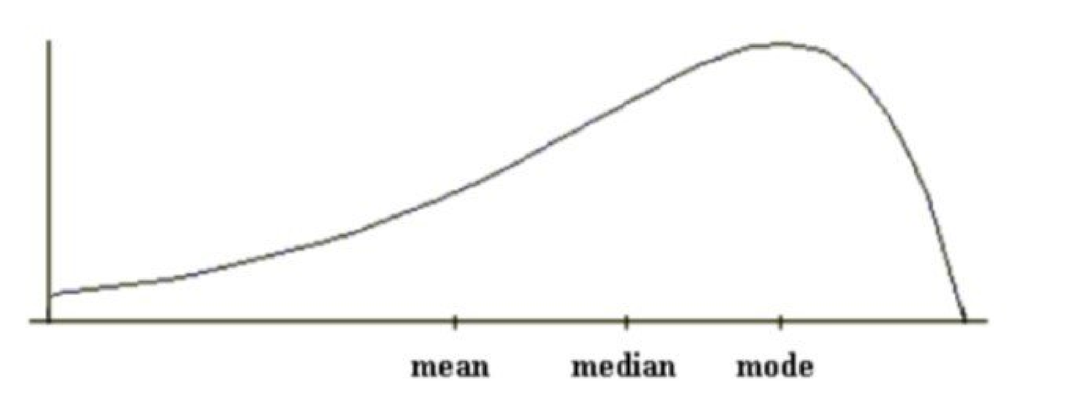
Measures of Skewness

- **Symmetric**



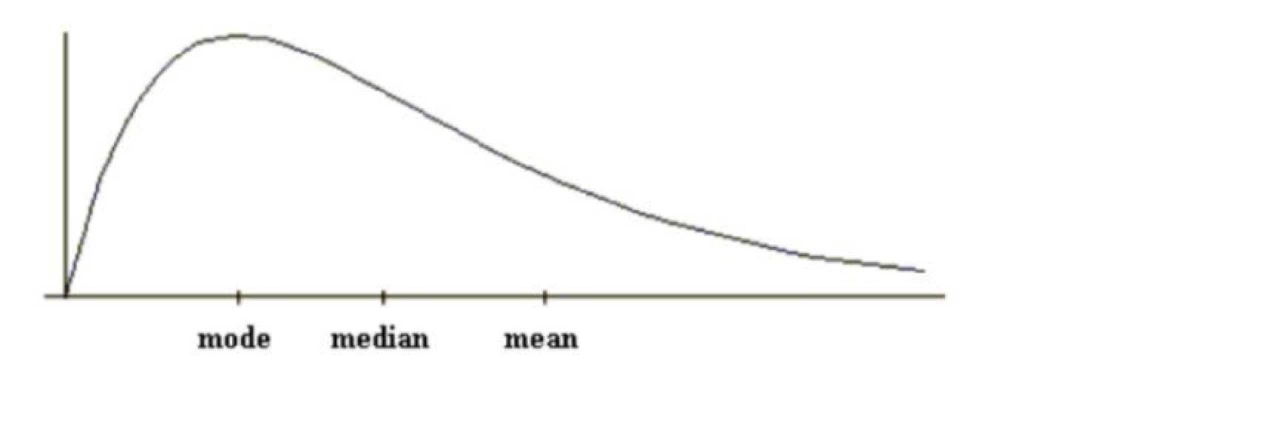
Measures of Skewness

- **Left/Negatively Skewed:** If a distribution has a long left tail, it is left-skewed (i.e., Mean < Median < Mode)
- **Example:** Retirement Age



Measures of Skewness

- **Right/Positively Skewed:** If a distribution has a long right tail, it is right-skewed (i.e., Mean > Median > Mode)
- **Example:** Salary of the employee in an organization



Measures of Dispersion

- **Range** = Maximum - Minimum
- The range is easy to calculate but is very much affected by extreme values.
- Not a robust measure of variability.

Measures of Dispersion

- 50th Percentile - 50% of the data values fall at or below the median.
- **IQR** = 75th Percentile - 25th Percentile
- Not affected by extreme values
- A robust measure of variability.

Measures of Dispersion

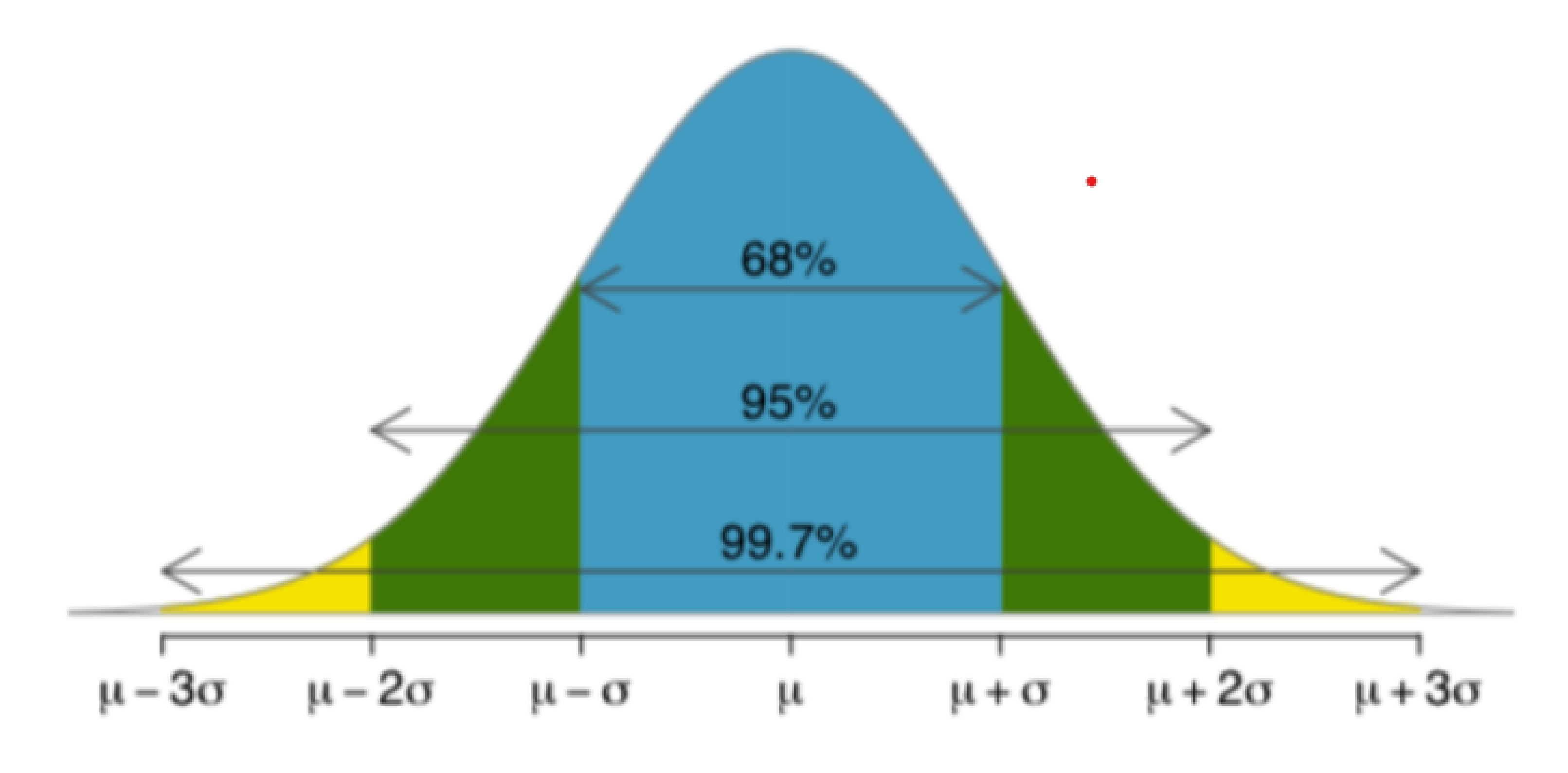
- Standard Deviation and Variance
  - Population Variance

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

-
  - Sample Variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Standard Deviation



Interview Question ??

- **Why sample variance has denominator n-1**
- **Is data closely clustered or has a wider range of values around the mean when the standard deviation is low?**
- **Test scores closely follow the normal model with a mean value of 1500 and a standard deviation as 300**
  - **At what percent of test takers score 900 to 2100**
  - **What percent score between 1500 and 2100**