# Statistical Description of Data

# Statistical Description of Data

- Statistical descriptions can be used to identify properties of the data and highlight which data values should be treated as noise or outliers.

Three areas of basic statistical descriptions:

1. Measuring the Central Tendency: Mean, Median and Mode.
2. Measuring the Dispersion of Data: Range, Quartiles, Variance, Standard Deviation, and Interquartile Range
3. Graphic Displays of Basic Statistical Descriptions of Data

# Measuring the Central Tendency

- Mean: Center of set of data. The mean of this set of values is:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

- This corresponds to the built in aggregate function, average (avg()inSQL), provided in relational database systems.

# Mean

- Mean: Suppose we have the following values for salary (in thousands of dollars),shown in increasing order 30,36,47,50,52,52,56,60,63,70,70,110.

$$\bar{x} = \frac{30+36+47+50+52+52+56+60+63+70+70+110}{12} = \frac{696}{12} = 58.$$

Thus, the mean salary is \$58,000.

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1x_1 + w_2x_2 + \cdots + w_Nx_N}{w_1 + w_2 + \cdots + w_N}.$$

This is called the **weighted arithmetic mean** or the **weighted average**.

# Problems with Mean

- A major problem with the mean is its sensitivity to extreme (e.g., outlier) values.
- Even a small number of extreme values can corrupt the mean.
- For example, the mean score of a class in an exam could be pulled down quite a bit by a few very low scores.
- To offset the effect caused by a small number of extreme values, we can instead use the **trimmed mean**, which is the mean obtained after chopping off values at the high and low extremes.
- For example, we can sort the values observed for salary and remove the top and bottom 2% before computing the mean. We should avoid trimming too large a portion (such as 20%) at both ends, as this can result in the loss of valuable information.

# Median

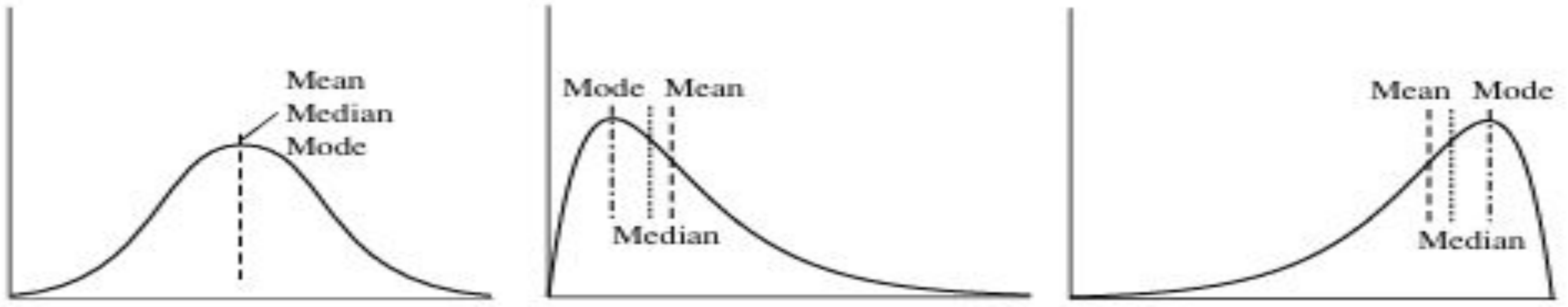
- The middle value in a set of ordered data values. It is the value that separates the higher half of a data set from the lower half.
- There is an even number of observations (i.e., 12);
- therefore, the median is not unique. It can be any value within the two middlemost values of 52 and 56 (that is, within the sixth and seventh values in the list).
- By convention, we assign the average of the two middlemost values as the median; that is, (52 + 56)/2= 108 /2 = 54.
- Thus, the median is $54,000.
- Suppose that we had only the first 11 values in the list. (odd number of values)
- This is the sixth value in this list, which has a value of **$52,000.**

# Mode

- The mode for a set of data is the value that occurs most frequently in the set.
- It can be determined for qualitative and quantitative attributes.
- Data sets with one, two, or three modes are respectively called **unimodal, bimodal, and trimodal**. In general, a data set with two or more modes is **multimodal.**
- If each data value occurs only once, **then there is no mode.**
- **For the above example** The two modes are $52,000 and $70,000. So it is **Bimodal**

# Mid Range

- The midrange of the data of Example is (30,000 +110,000)/2 = $70,000.
- In a unimodal frequency curve with perfect symmetric data distribution, the mean, median, and mode are all at the same center value, as shown in Figure
- Data in most real applications are not symmetric. They may instead be either positively skewed, where the mode occurs at a value that is smaller than the median or negatively skewed, where the mode occurs at a value greater than the median.




(a) Symmetric data (b) Positively skewed data (c) Negatively skewed data

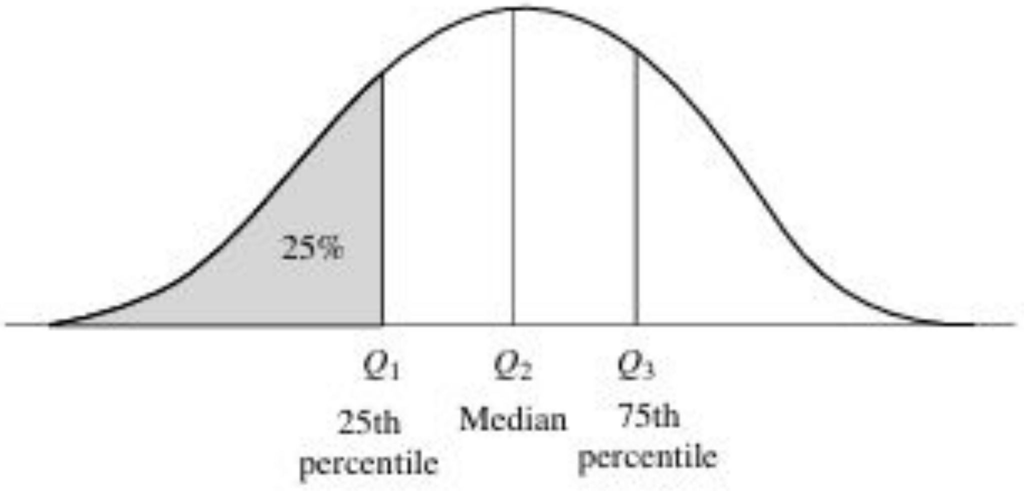
# Measuring the Dispersion of Data:

- The **range** of the set is the difference between the largest (max()) and smallest (min()) values.
- The 2-quantile is the data point dividing the lower and upper halves of the data distribution. It corresponds to the median. The 4-quantiles are the three data points that split the data distribution into four equal parts; each part represents one-fourth of the data distribution. They are more commonly referred to as **quartiles.**
- The 100-quantiles are more commonly referred to as **percentiles**; they divide the data distribution into 100 equal-sized consecutive sets.

# Measuring the Dispersion of Data

The median, quartiles, and percentiles are the most widely used forms of quantiles.

- The first quartile, denoted by Q1, is the 25th percentile. It cuts off the lowest 25% of the data.
- The third quartile, denoted by Q3, is the 75th percentile—it cuts off the lowest 75% (or highest 25%) of the data.
- The second quartile is the 50th percentile. As the median, it gives the center of the data distribution.




interquartile range (IQR) and is defined as IQR =Q3 - Q1

IQR for the dataset Q3 is median = 63 and Q1 =47 and Q2 which is median is 52

IQR=63-47= 16

# Five Point Summary, Box Plot