Machine Learning

Measures of Central Tendency: Mean, Median and Mode

Measures of Central Tendency

- A **statistic** is a characteristic or measure obtained by using the data values from a sample.
- A **parameter** is a characteristic or measure obtained by using all the data values from a specific population.

Measures of Central Tendency

- **Mean**
  - The **mean** is the sum of the values, divided by the total number of values. The symbol represents the sample mean.

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} = \frac{\sum X}{n}$$

where *n* represents the total number of values in the sample.

$$\mu = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N} = \frac{\sum X}{N}$$

- For a population, the Greek letter **μ** (mu) is used for the mean. where *N* represents the total number of values in the population.

Measures of Central Tendency

- **Mean (Contd.)**
  - The procedure for finding the mean for grouped data is given here –

A Class	B Frequency f	C Midpoint X_m	D $f \cdot X_m$
5.5–10.5	1	8	8
10.5–15.5	2	13	26
15.5–20.5	3	18	54
20.5–25.5	5	23	115
25.5–30.5	4	28	112
30.5–35.5	3	33	99
35.5–40.5	2	38	76
	$n = 20$		$\Sigma f \cdot X_m = 490$

$$\overline{X} = \frac{\Sigma f \cdot X_m}{n} = \frac{490}{20} = 24.5 \text{ miles}$$

Measures of Central Tendency

The Weighted Mean

Find the **weighted mean** of a variable *X* by multiplying each value by its corresponding weight and dividing the sum of the products by the sum of the weights.

$$\overline{X} = \frac{w_1X_1 + w_2X_2 + \cdots + w_nX_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum wX}{\sum w}$$

where $w_1, w_2, \ldots, w_n$ are the weights and $X_1, X_2, \ldots, X_n$ are the values.

Example:

Course	Credits (*w*)	Grade (X)
English composition I	3	A (4 points)
Introduction of Psychology	3	C (2 points)
Biology I	4	B (3 points)
Physical Education	2	D (1 points)

$$\overline{X} = \frac{\sum wX}{\sum w} = \frac{3 \cdot 4 + 3 \cdot 2 + 4 \cdot 3 + 2 \cdot 1}{3 + 3 + 4 + 2} = \frac{32}{12} = 2.7$$

The grade point average is 2.7.

Measures of Central Tendency

- **Median**
  - The **median** is the midpoint of the data array. The symbol for the median is MD.

- **Example:**
  - The number of cloudy days for the top 10 cloudiest cities is shown. Find the median.

    209, 223, 211, 227, 213, 240, 240, 211, 229, 212

- **Solution:**
  - Arrange the data in order.

    209, 211, 211, 212, 213, 223, 227, 229, 240, 240

    ↑

    Median

    MD = (213 + 223) / 2 = 218

Hence, the median is 218 days.

Measures of Central Tendency

- **Mode**
  - The value that occurs most often in a data set is called the **mode.**

- **Example:**
  - Find the mode of the signing bonuses of eight NFL players for a specific year.
    The bonuses in millions of dollars are
    18.0, 14.0, 34.5, 10, 11.3, 10, 12.4, 10
- **Solution:**
  - It is helpful to arrange the data in order although it is not necessary.
    10, 10, 10, 11.3, 12.4, 14.0, 18.0, 34.5
    Since $10 million occurred 3 times, a frequency larger than any other number, so the mode is $10 million.

Machine Learning

Measures of Central Tendency: Midrange

Measures of Central Tendency

- **Midrange**
  - The **midrange** is defined as the sum of the lowest and highest values in the data set, divided by 2. The symbol MR is used for the midrange.

$$\mathbf{MR} = \frac{\textbf{lowest value} + \textbf{highest value}}{\mathbf{2}}$$

- **Example:**
  - In the last two winter seasons, two cities, city-1 and city-2 reported these numbers of water-line breaks per month.
- **Find the midrange.**

  2, 3, 6, 8, 4, 1
- **Solution:**

  MR = (1 + 8)/2 = 4.5

  Hence, the midrange is 4.5

Measures of Central Tendency

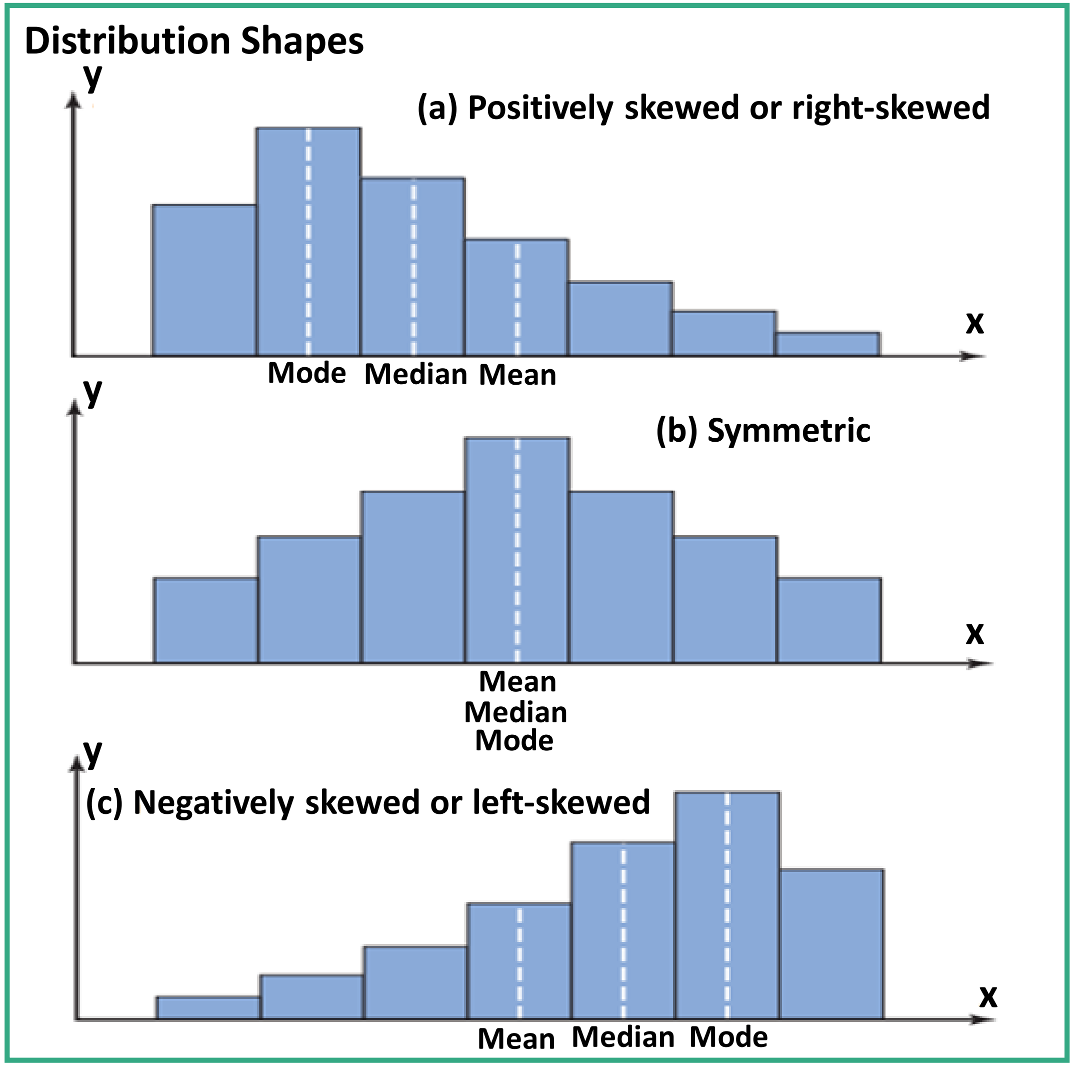
Summary of Measures of Central Tendency

Summary of Measures of Central Tendency		
Measure	**Definition**	**Symbol(s)**
Mean	Sum of value, divided by total number of values	$\mu, \overline{X}$
Median	Middle point in data set that has been ordered	MD
Mode	Most frequent data value	None
Midrange	Lowest value plus highest value, divided by 2	MR

Machine Learning

Measures of Central Tendency: Distribution Shapes

Measures of Central Tendency



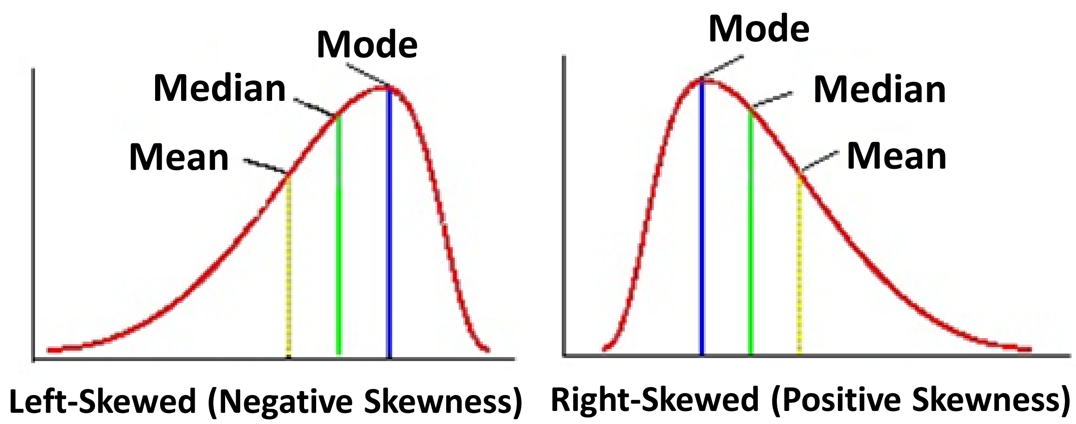
Machine Learning

Skewness & Kurtosis

Skewness – Asymmetrical Distribution

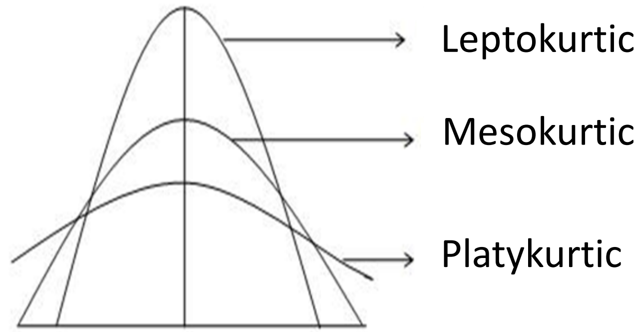
- Skewness is a measure of symmetry in a distribution. If one tail is longer than another, the distribution is skewed, e.g. Income, Populations of countries.
- Different ways to measure a skew: Pearson Mode, Bowley, Kelly's Measure, Momental.
- Which technique you use depends on what you know about your data, e.g. if you know the mean, mode (or median) and standard deviation you can use Pearson's.
- Momental skewness could be an option if you only know the mean and standard deviation for your set of data.

Skewness – Asymmetrical Distribution Contd…



- A *symmetrical distribution* has a skew of zero. A positive result means that your data is positively skewed. A negative result means that your data is negatively skewed.

Kurtosis – Sharpness of Peak of Distribution



- The degree of flatness or peakedness is measured by kurtosis. It tells us about the extent to which the distribution is flat or peak vis-a-vis the normal curve.

- The normal curve is called **Mesokurtic** curve. If the curve of a distribution is more peaked than a normal or mesokurtic curve then it is referred to as a **Leptokurtic** curve. If a curve is less peaked than a normal curve, it is called as a **Platykurtic** curve.

- Formula:

  β_2=$\mu_4\mu_2$ where $\mu_2 = \frac{\Sigma(x-\bar{x})^2}{N}$, $\mu_4 = \frac{\Sigma(x-\bar{x})^4}{N}$, $\bar{x}$ is mean

Machine Learning

Measures of Variation:
Range

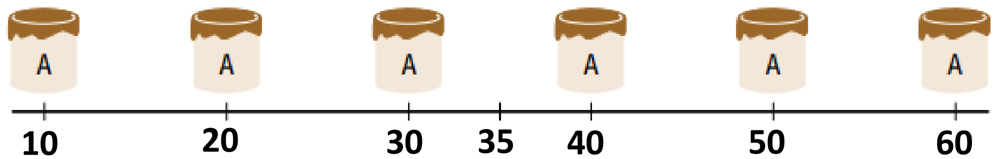
Measures of Variation

➢ **Range**

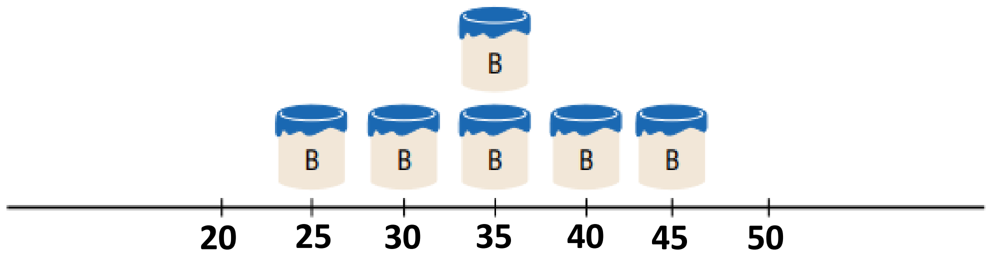
- The **range** is the highest value minus the lowest value. The symbol R is used for the range.

$$\boldsymbol{R = highest\ value - lowest\ value}$$

➢ **Example:**



(a) Brand A **Variation of paint (in months)**



(b) Brand B **Variation of paint (in months)**

For brand A, the range is $R = 60 - 10 = 50$ months

For brand B, the range is $R = 45 - 25 = 20$ months

Machine Learning

Measures of Variation: Variance & Standard Deviation

Measures of Variation

- **Variance and Standard Deviation**
  - The **variance** is the average of the squares of the distance each value is from the mean. The symbol for the population variance is $\boldsymbol{\sigma}^2$ ($\boldsymbol{\sigma}$ is the Greek lowercase letter sigma). The formula for the population variance is

$$\boldsymbol{\sigma^2 = \frac{\sum(X-\mu)^2}{N}}$$

Where, X = individual value, $\boldsymbol{\mu}$ = population mean and

N = population size

The **standard deviation** is the square root of the variance. The symbol for the population standard deviation is $\boldsymbol{\sigma}$. The corresponding formula for the population standard deviation is

$$\boldsymbol{\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(X-\mu)^2}{N}}}$$

Measures of Variation

Example: If X values are 35, 45, 30, 35, 40, 25

$$\boldsymbol{\mu = \frac{\sum X}{N} = \frac{35 + 45 + 30 + 35 + 40 + 25}{6} = \frac{210}{6} = 35}$$

A	B	C
X	$\mathbf{X - \mu}$	$\boldsymbol{(X - \mu)^2}$
35	0	0
45	10	100
30	-5	25
35	0	0
40	5	25
25	-10	100

$$\sum(\boldsymbol{X} - \boldsymbol{\mu})^{\mathbf{2}} = 0 + 100 + 25 + 0 + 25 + 100 = 250$$

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N} = \frac{250}{6} = 41.7$$

$$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}} = \sqrt{41.7} = 6.5$$

Hence, the standard deviation is 6.5.

Measures of Variation

Example:

Summary of Measures of Variation		
Measure	**Definition**	**Symbol(s)**
Range	Distance between highest value and lowest value	R
Variance	Average of the squares of the distance that each value is from the mean	σ^2, s^2
Standard deviation	Square root of the variance	σ, s

Machine Learning

Coefficient of Variation

Coefficient of Variation

The **coefficient of variation,** denoted by CVar, is the standard deviation divided by the mean. The result is expressed as a percentage.

For samples,

$$\text{CVar} = \frac{S}{\bar{X}} \cdot 100\%$$

For populations,

$$\text{CVar} = \frac{\sigma}{\mu} \cdot 100\%$$

Coefficient of Variation

- **Example:**
  - **Sales of Automobiles**
    The mean of the number of sales of cars over a 3-month period is 87, and the standard deviation is 5. The mean of the commissions is \$5225, and the standard deviation is \$773. Compare the variations of the two.
- **Solution:**
  The coefficients of variation are

$$\text{CVar} = \frac{S}{\bar{X}} = \frac{5}{87} \cdot 100\% = 5.7\% \qquad \text{sales}$$

$$\text{CVar} = \frac{773}{5225} \cdot 100\% = 14.8\% \qquad \text{commissions}$$

Since the coefficient of variation is larger for commissions, the commissions are more variable than the sales.

Machine Learning

Boxplot with Five Number Summary

Boxplot with Five Number Summary

➢ **Box Plot**

- A **boxplot** is a graph of a data set obtained by drawing a horizontal line from the minimum data value to $Q1$, drawing a horizontal line from $Q3$ to the maximum data value, and drawing a box whose vertical sides pass through $Q1$ and $Q3$ with a vertical line inside the box passing through the median or $Q2$.
- A **boxplot** can be used to graphically represent the data set. These plots involve five specific values:

  1. The lowest value of the data set (i.e., minimum)
  2. $Q1$
  3. The median
  4. $Q3$
  5. The highest value of the data set (i.e., maximum)

  These values are called a **five-number summary** of the data set.

Boxplot with Five Number Summary

- **Example of Box Plot:**
  - The data set is 89, 47, 164, 296, 30, 215, 138, 78, 48, 39. Construct a boxplot for the data.
- **Solution:**

Step 1 Arrange the data in order:
30, 39, 47, 48, 78, 89, 138, 164, 215, 296

Step 2 Find the median. Here it is Q2 = (78 + 89)/2 = 83.5

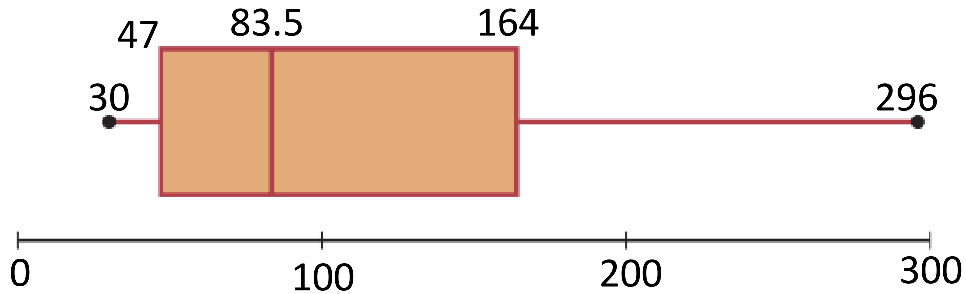
Step 3 Find *Q*1. Here it is Q1 = 47

Step 4 Find *Q*3. Here it is Q3 = 164

Step 5 Draw a scale for the data on the *x* axis.

Step 6 Located the lowest value, *Q*1, median, *Q*3, and the highest value on the scale.

Step 7 Draw a box around *Q*1 and *Q*3, draw a vertical line through the median, and connect the upper value and the lower value to the box.



Machine Learning

Measures of Positions: Standard Score & Outliers

Standard Scores

- **Standard Score or z-score**
  - A ***z* score** or **standard score** for a value is obtained by subtracting the mean from the value and dividing the result by the standard deviation. The symbol for a standard score is *z*. The formula is

$$\mathbf{z = \frac{value - mean}{standard\ deviation}}$$

  - For samples, the formula is

$$\mathbf{z = \frac{X - \overline{X}}{s}}$$

  - For populations, the formula is

$$\mathbf{z = \frac{X - \mu}{\sigma}}$$

The *z* score represents the number of standard deviations that a data value falls above or below the mean.

Outliers

- **Outliers**
  - An **outlier** is an extremely high or an extremely low data value when compared with the rest of the data values.

- **Procedure to find out Outliers:**

  Step 1: Arrange the data in order and find Q1 and Q3.
  Step 2: Find the interquartile range: IQR = Q3 - Q1.
  Step 3: Multiply the IQR by 1.5.
  Step 4: Subtract the value obtained in step 3 from Q1 and add the value to Q3.
  Step 5: Check the data set for any data value that is smaller than Q1 - 1.5*(IQR) or larger than Q3 + 1.5*(IQR)

Outliers

- **Example of Outliers:**
  Check the following data set for outliers.
  5, 6, 12, 13, 15, 18, 22, 50
- **Solution:**
  Step 1 Here $Q1$ is 9 and $Q3$ is 20.
  Step 2 So IQR = $Q3 - Q1 = 20 – 9 = 11$
  Step 3 Multiply this value by 1.5. So 1.5 * (11) = 16.5
  Step 4 So lower limit = 9 – 16.5 = –7.5 and upper limit = 20 + 16.5 = 36.5
  Step 5 Check the data set for any data values that fall outside the interval from –7.5 to 36.5. The value 50 is outside this interval; hence, it can be considered an outlier.